Extraction of Definitional Contexts using Lexical Relations

Olga Acosta
Postgraduate School of Computer Science, Universidad Nacional Autónoma de México, Mexico City, Mexico

Gerardo Sierra
Engineering Institute, Universidad Nacional Autónoma de México, Mexico City, Mexico

César Aguilar
Department of Spanish Linguistics and Literature, Catholic University of Chile, Santiago, Chile

ABSTRACT
In this paper we present a method for automatically extracting definitional contexts from restricted domains in Spanish. Definitional contexts are textual fragments where there is an implicit definition that can be identified by taking into account verbal patterns linking a term and its corresponding definition. Our interest is in definitional contexts with analytical definitions. Therefore, we focus on the extraction of textual fragments with a term and a hypernym. Then, hypernym is used for filtering non-relevant contexts by means of the occurrence frequency. It is assumed most frequent hypernyms have a higher probability of giving true definitional contexts than those less frequent hypernyms. We captured regularity in analytical definitions by means of a chunk grammar. This method achieves acceptable results in precision and recall compared with other works.

General Terms
Automatic Extraction of Conceptual Information.

Keywords
Natural language processing, information extraction, conceptual extraction, lexical relations.

1. INTRODUCTION
Concepts are a fundamental piece to all aspects of cognition. We work with concepts in our daily contact with others and the world. According to [1], concepts mirror the way that we divide the world into classes, and much of what we learn, communicate, and reason involves relations among these classes. Furthermore, one of the most important functions of concepts is the promotion of cognitive economy [2].

The classical Aristotelian vision of concepts explains the fact, that concepts are encoded in terms of necessary and sufficient conditions ([3],[4]). In specialized domains, this attempt of codification is reflected in a definition. Definitions contain relevant information about terms denoting concepts and their relations with other concepts, which helps to organize knowledge. For example, in an analytical definition:

X is a Y + differentia

A concept X is defined by a genus or superordinate, and a set of necessary and sufficient conditions that differentiates the concept from other species of the same genus. In this case X is the hyponym and Y is often considered as a hypernym [5].

Currently, semantic WEB and availability of huge textual information sources have given rise to a growing need for automatic mechanisms for extracting knowledge from textual resources [6]. In this context, analytical definitions can be extracted by several ways. For example, [7] proposed a method for extracting definitional contexts (or DCs) with analytical definitions from restricted domains by means of verbal patterns. DCs are textual fragments where there is often an implicit definition that can be identified by taking into account verbal patterns linking a term and its corresponding definition.

In this work, we present a method for automatically extracting DCs containing an implicit analytical definition. To achieve this goal, we start from the verbal patterns identified by [7] to extract candidates to DCs. Then, we use a chunk grammar and a hypernymy extractor to retrieve relevant DCs. The chunk grammar considers the structure of the DCs and the relationship among the term, the hypernym and the verbal pattern. Heuristics are applied to the potential hypernym as well as its occurrence frequency. This latter assume more frequent hypernyms have a higher probability of retrieving relevant DCs.

We applied the method to a specialized domain in Spanish. Given the ambiguity of language, until now the most success stories are found in restricted domains. The knowledge of a specific domain is limited, at least in a period of time to a set of concepts, and even ambiguous terms have very precise meanings in restricted domains. These lexical items constitute the terminology of the domain and they are useful for conceptualize it ([6],[8],[9]).

We organize our paper as follows: section 2 presents a general description of analytical definitions and their structure. Then, section 3 provides an overview about method of conceptual extraction proposed by [7], as well as most appropriate information sources in order to extract concepts. Additionally, section 4 gives an explanation about most common structure of terms in Spanish, as well as noun phrases where a hypernym can be found. In section 5 we expose our methodology. In section 6 we analyze the results and finally we present the conclusions.

2. ANALYTICAL DEFINITIONS
An analytical definition represents a very common approach for describing a concept in terms of a superordinate concept (genus), and a differentia distinguishing the concept defined from others with the same genus. According to Sowa [10], Aristotle proposed two dominant methods of classification,

which have been considered in all branches of cognitive science. The first one is the top-down method for defining concepts based on a genus or superordinate and one or more differentiae. On the other hand, the bottom-up method begins with a detailed description of individuals, classifying them in species, and grouping species in superordinates. The top-down approach was considered by Aristotle as an appropriate method for presenting the results of analysis and reasoning about them, but he recommended the bottom-up approach as a better discovery procedure for investigating a new subject. As Sowa notes, all logic-based methods are examples of an Aristotelian top-down approach.

Practically speaking, analytical definitions are a very common mechanism for defining concepts and remain as a mean of transferring knowledge in specialised domains. For example, the next definition provides a description of concept *lightning conductor* using one of the most common verbs (i.e., to be) for introducing a definition. In this case, the genus is the concept *device* while the differentia describes the function of the *lightning conductor*:

> [Lightning conductor]$_{\text{Term}}$ is a [device]$_{\text{genus}}$ [that allows to protect the electrical systems against surges of atmospheric origin.]$_{\text{Differentia}}$

According to [10], despite the evidence against the classical approach ([3],[4],[11]), for a particular application it is possible to build a top-down hierarchy of concepts by identifying genus, although this hierarchy will not be universal or globally consistent. The Sowa's perspective [10] is in line with Gruber's proposal [8] about ontologies for making ontological commitments, that is, agreements to use a vocabulary in a way that is consistent with respect to the theory specified by an ontology.

3. AUTOMATIC EXTRACTION OF CONCEPTS

3.1 Sources of Conceptual Information

If we assume that words denote concepts, then a good source to find them is textual information. Nowadays, computational lexicography and terminology are able to recognize concepts in large-text corpora ([12], [13]). To focus on this recognition, it is important to establish what the best source for obtaining relevant concepts is. In this sense, ([14], [15]) point out the value of scientific and technical literature as a source to obtain such concepts. In particular, [14] considers definitions as a linguistic representation of concepts, because definitions synthesize all the conceptual information linked to terms circumscribed to a domain-specific knowledge.

In line with these authors, we consider as conceptual information that information expressed by specialized definitions, particularly analytical definitions constituted by genus and differentia. Several authors have used this kind of definition for searching hyponymy-hypernymy relations established between terms and genus ([5], [16],[17]).

In order to recognize these relations, [5] used the IS-A operator for finding lexical syntactic patterns with a high degree of precision in MRDs. However, [16] argue that these kinds of patterns are not sufficient for describing all the possibilities to express an analytical definition in natural language. Thus, it is necessary to consider other alternative patterns capable of introducing these definitions in specialized documents.

3.2 Extraction of Conceptual Information

[7] developed a based-pattern 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:

> Sp. 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 an energetic 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. that resource that…) and the verbal pattern *se define como* (Eng. 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, [7] 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. Table 1 shows such verbs and their links considered in analytical definitions. Link element indicates an item regularly accompanying verb heads; however, there are verbs where this element can be missing ε.

Table 1. Verb predications and link elements

Infinitive verb	Link element
To be	
To characterize, to conceive, to consider, to describe, to define, to understand, to know, to refer	As
To denominate, to call, to name	ε, as

Furthermore, [18] considers two types of predicative phrases: a simple or primary predication, i.e., those predicative phrases conformed by a subject to the left of the verb, and a predicate that is located to the right of the verb:

> 1. Sp. La [conjuntivitis] $_{\text{Term}}$ es una [inflamación] $_{\text{hypernym}}$ de la conjuntiva del ojo.
>
> (Eng. [Conjunctivitis] $_{\text{Term}}$ is an [inflammation] $_{\text{hypernym}}$ of the conjunctiva of the eye).

On the other hand, a secondary predication integrates a subject in a pre-verbal position, and an object and its predicate, both after the verb. In this case, the predicate affects the object of a sentence:

2. [Alan Turing] $_{Subject}$ defines [the Turing Machine] $_{Term}$ [as a [concept] $_{hypernym}$ to describe a mechanically working mathematician.] $_{predicate}$

Other examples of secondary predications are all sentences structured in passive, where the subject is elided, for example:

3. Se define [conjuntivitis] $_{Term}$ como una [inflamación] $_{hypernym}$ de la conjuntiva del ojo.

(Eng. It is defined [conjunctivitis] $_{Term}$ as an [inflammation] $_{hypernym}$ of the conjunctiva of the eye).

In (1) and (3), we observe terms and analytical definitions linked through PrPs whose heads are the verbs to *be* and *define*. In both cases, the term *conjunctivitis* is conceived as an inflammation of the eye, for this reason the genus of these definitions is the term *inflammation*. On the other hand, example (2) shows the most canonical case of secondary predication, because this predication expressed clearly who is the author of the analytical definition (Alan Turing), what is the term defined (Turing Machine), and the hypernym linked to the term (a concept).

4. STRUCTURE OF TERMS AND HYPERNYMS

4.1 Structure of Terms

In tasks of terminological extraction for Catalan [19] and Spanish [20], patterns with preposition *de* (Eng. of) and adjectives are the most common elements for structuring compound terms in specialized domains. Table 2 shows some examples of this kind of terms.

Table 2. Terminological patterns

Pattern of term	Example
Sp. Noun + Adjective Eng. Adjective + Noun	Enfermedad cardiovascular Cardiovascular disease
Sp. Noun + Prepositional Phrase Eng. Noun + Possessive Noun + Prepositional phrase	Enfermedad de Alzheimer Alzheimer's disease Glaucoma of open angle

We only considered adjectives and preposition *de* (Eng. Of). The latter for two main reasons: it is the most recurrent ([21], [22]), at least in English and Spanish, and it is the most commonly used for construction of terms in Spanish.

4.2 Noun phrase of hypernyms

According to [5] noun definitions are normally written in such a way that one can identify the genus or hypernym of the term being defined. For instance:

1. Conjunctivitis – an inflammation of the conjunctiva of the eye.

Here *inflammation* is the hypernym of the term *Conjunctivitis* and the rest of the sentence corresponds to the *differentia*. The above heuristic seems logic and intuitive, but there are definitions where the head of noun phrase is not an appropriate hypernym for the term. For example:

2. Gemstone – any of various minerals …

3. Electron microscope is a type of microscope that uses a particle beam of electrons …

In above definitions, *any* and *type* are empty heads. These heads must be filtered with the goal of finding a hypernym related semantically with term [5]. In the case of (2), a hypernym semantically related with term is *minerals.* On the other hand, case (3), head *type* is not a hypernym, but rather it is an indicator of taxonymy [23]. In short, hypernym in (2) and (3) are found after preposition *of,* so, this situation justifies why to take into account noun phrases with prepositional phrase with head *of* as a modifier.

Finally, as [5] notes, it is necessary for a thorough study of this shadowy area in order to make optimal use of the semantic information available in MRDs. In the case of DCs, this is not an exception, empty heads and heads reflecting other kinds of relations very often occur. Given this situation, in this work we filter empty heads and heads reflecting other kinds of relations such as causal and part-whole relations.

5. METHODOLOGY

The automatic extraction of DCs is possible if occurrence of verbal patterns is considered. This work is similar in spirit to the method proposed by [7]. However, in our case, only DCs with analytical definitions implicit are considered. Further, another important difference is that our method takes into account regularity in construction of terms together with regularity in definitional patterns and hypernyms. This regularity allows us to extract candidate fragments to analytical definitions where a term and a hypernym can be located. Then, candidate hypernyms are identified and extracted in order to filter less relevant DCs. We assume most frequent hypernyms have a higher probability of being true hypernyms and hence analytical fragments and DCs where these hypernyms are present will have a high probability of being true.

A justification for our proposal is that there is enough evidence, that terms are mainly constructed by considering regular syntactic patterns. Compound terms are formed by a noun head plus modifiers such as adjectives and prepositional phrases with head *de* (Eng. of). In addition, the verbal pattern in a definition can give clues about position of the term, as well as of the hypernym within a definition. Figure 1 shows a general scheme of the proposed methodology in this work. Next subsections present a description of each step of the methodology.

5.1 Removing Parentheses

In order to clarify terminology used in an expert-beginner communicative context, it is a common situation to find information in parentheses. Information in parentheses is often related with clarifications, acronyms, etymological roots, synonyms, and even short definitions. The following examples show some of these cases found in DCs:

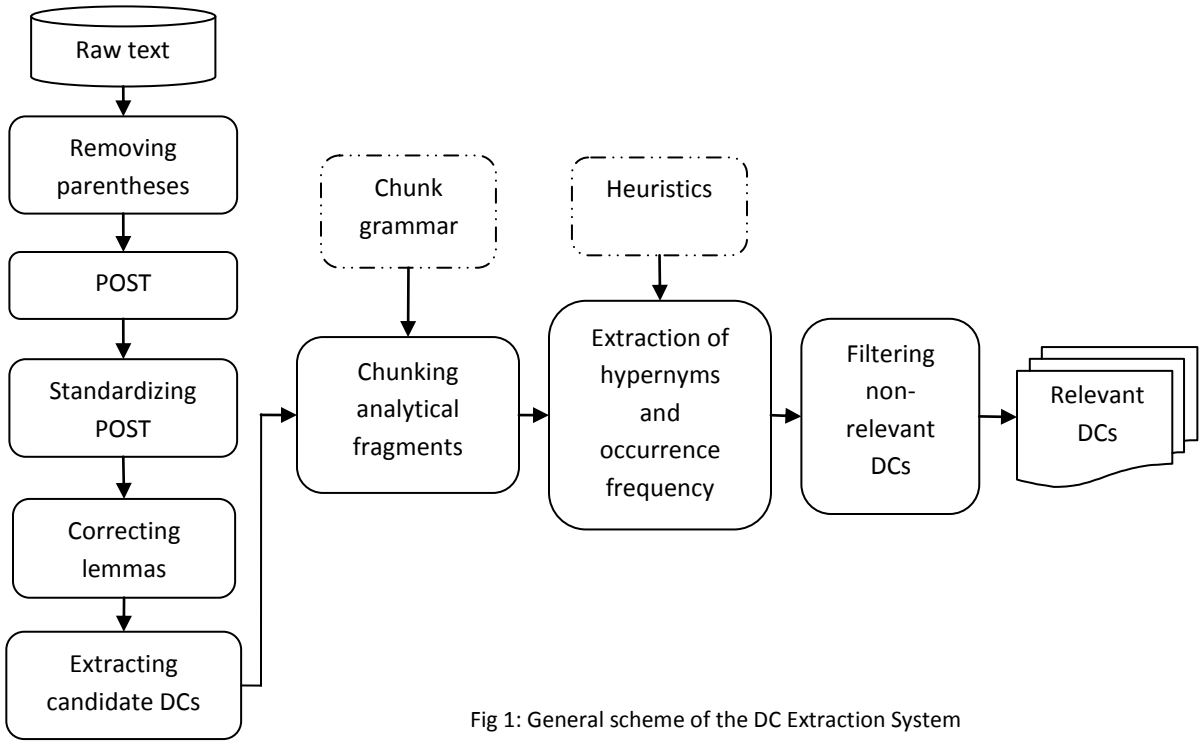

Fig 1: General scheme of the DC Extraction System

1. Electronigraphic (ERG) is a test measuring electrical activity of the retina to the light.

2. Conjunctivitis is the swelling (inflammation) or infection of the membrane covering eyelids.

Although this information is valuable, it is not considered in this work because it would require more complex regular expressions in order to capture these patterns in a chunk grammar. Hence, we automatically remove this information from text analyzed before the POST phase.

5.2 Part-of-Speech Tagging

Part-of-speech tagging (POST) is the process of assigning a grammatical category or part of speech to each word in a corpus. For instance, one POST output can provide the word form, the part-of-speech tag and the lemma with the next structure:

Word form	**Tag**	**Lemma**
Defined	VVD	Define

The next examples show a sentence in Spanish and English tagged with the TreeTagger POST tool [24]:

El/ART síntoma/NC característico/ADJ de/PDEL rojez/NC ocular/ADJ ./FS

The/DT characteristic/JJ symptom/NN of/IN ocular/JJ redness/NN ./SENT

Tags used by TreeTagger for Spanish are based on the set defined in the Penn Treebank tagset[1]. Methodology proposed in this work requires POST because we focused on capturing the regular behavior of DCs with a partial parsing called chunking.

5.3 Standardizing POST

In this work we used TreeTagger tool to tag text with POS. With the goal of narrowing the scope of rules to verbs used in analytical definitions, a step of standardizing was applied to tags of verbs shown in table 1. These verbs were tagged with a new single tag VLFIND for the cases VLFIN, VLINF, VLGER, VLADJ, with exception of tag for verb *to be* VSFIN. Table 3 presents other cases with tags modified.

Table 3. Modifications of POS tags

Item	**TreeTagger's tag**	**Modified tag**
Of Of + <art> Contraction of + <art>	PREP PREP + ART PDEL	PDEL
Or	CC	CCO
To	PREP	PA
In	PREP	PEN
.	FS	FSP

5.4 Correcting Lemmas

One of the problems presented in the POST phase was lemmatization of nouns and adjectives, especially those related with the domain analyzed. Very often POS taggers solve inflections of gender and number either assigning the token or

[1] www.ims.uni-stuttgart.de/ftp/pub/corpora/spanish-tagset.txt

leaving *unknown* lemma. For instance, cases such as *proteínas* and *proteína* (Eng. proteins-protein) can be represented by different lemmas, which can affect subsequent processes.

We implemented a phase to correct most common errors of lemmatization. One of the main heuristics considered was that word length must be at least six because it is hypothesized that the larger the word size is and coincides with the root of another word, the likelihood of being the same word but inflected is high. Table 4 shows some of the heuristics applied for regular plurals and gender inflections. Validations shown in table 4 are in Python code.

Table 4. Heuristics for lemmatization of nouns and adjectives

Validation	Type of correction and examples
len(word1)==len(word2) and word1[:-1]==word2[:-1] and word1[-1:] in vocal and word2[-1:] in vocal and word1[-1:]!=word[-1:]	Gender: Sanguíneo-sanguínea
len(word1)==len(word2) and word1[:-2]==word2[:-2] and word1[-2:] in suffixPlural and wor21[-2:] in suffixPlural and word1[-2:]!=word2[-2:]	Plural and gender: Difusos-difusas

5.5 Extracting Candidate DCs

We identified candidate DCs by considering lemma of the verbs shown in table 1. A sentence delimited by a dot must contain at least a verb commonly used in analytical definitions to be extracted. This first extraction of DCs is considered as our baseline.

5.6 Chunking

Chunking is the process of identifying and classifying segments of a sentence by means of the grouping of major parts-of-speech forming basic non-recursive phrases. A set of rules indicating how sentences should be grouped makes up a grammar called chunk grammar. The rules of a chunk grammar use tag patterns to describe sequences of tagged words, e.g. <art>?<adj>*<nc>+. Tag patterns are similar to regular expression patterns, where symbols such as "*" means zero or more occurrences, "+" means one or more occurrence and "?" represents an optional element.

In this work, we are interested in extracting DCs. These fragments follow a regular structure allowing us to capture their behavior by means of tag patterns. Table 5 shows a sentence with parts of speech grouped. For example, a noun phrase (NP) can be a single noun or a more complex structure: a determinant plus noun and a prepositional phrase modifier. Additionally, verb head of the PrP is verb *to be.*

Table 5. Chunking parts of a sentence

Conjunctivitis	is	An inflammation of the conjunctiva		
Noun (N)	V	(Det)	Noun (N)	Prepositional Phrase (PP)
NP	PrP	NP		

5.6.1 Defining a Chunk Grammar

A chunk grammar was designed by considering the syntactical behavior of 1500 definitions extracted from Wikipedia about three different knowledge areas: Biology, Medicine, and Linguistics. Definitions of Wikipedia do not follow a formal definitional pattern, and they are closer to definitions found in restricted domains. Further, Wikipedia is a resource where volunteers of the entire world can write articles about myriad topics which guarantees, in part, that the most common behavior of DCs to be present.

The following tag patterns show the structure of term, noun phrase of hypernyms and synonyms that we considered in this work. These expressions correspond to elements in the chunk grammar that were treated as constituents.

5.6.1.1 Terms (T)

According to description of subsection 4.1, a term can be:

$$T : \langle art \mid fsp \rangle \langle nc \mid np \mid alfs \mid acrnm \rangle + \langle adj \mid nc \mid np \mid alfs \mid vladj \rangle * \left(\begin{array}{l} \langle pdel \rangle \langle nc \mid np \mid adj \mid alfs \mid acrnm \rangle + \\ \langle adj \mid nc \mid np \mid alfs \mid vladj \rangle * \end{array} \right) *$$

An obligatory article <art> or dot <fsp> were included in order to reduce noise and to extract the more relevant results. On the other hand, items such as <alfs> or <acrnm> allow increasing recall of DCs to cases such as: vitamin A (<nc><alfs>), ONU (<acrnm>).

5.6.1.2 Noun phrase with hypernym (HP)

It is a relatively common situation to find empty heads (any, type, kind, and so on) in noun phrases where hypernym can be found. So, we consider noun phrases with prepositional phrases with head *of.*

$$HP1 : \langle art \mid qu \rangle ? \langle nc \mid np \rangle + \langle adj \mid nc \mid vladj \mid card \rangle * \left(\langle adv \rangle \langle adj \rangle \right) ? \left(\langle pdel \rangle \langle nc \mid np \rangle + \langle adj \rangle * \right) *$$

$$HP2 : \langle card \mid qu \rangle \left(\langle adv \rangle \langle adj \rangle \right) ? \left(\begin{array}{l} \langle pdel \rangle \langle adv \rangle ? \\ \langle nc \mid np \mid card \mid adj \rangle + \end{array} \right) +$$

$$HP3 : \langle dm \mid art \rangle \langle cque \mid rel \rangle \langle nc \rangle ? \langle adj \rangle *$$

Expressions HP1 and HP2 are evaluated in order to extract hypernym more related with the term being defined. This is done filtering either empty heads as well as indicative of other relations. The latter regular expression (HP3) includes cases where we have a demonstrative <dm> or article <art> and a relative expression where it is possible that a noun is missing:

Sp. La [energía primaria]$_{Term}$ es [aquella]$_{hypernym}$ que no ha sido afectada por alguna transformación, con la excepción de su extracción.

(Eng. The [primary energy]$_{Term}$ is [that]$_{hypernym}$ that has not been affected for any transformation, with the exception of its extraction.)

5.6.1.3 Synonyms (Syn)

Between a term and a verbal pattern other items can be introduced, for example, punctuation signs, adverbs, relative items and synonyms of the term being defined:

1. [Erythropsia]$_{Term}$ [or red vision]$_{Syn}$ is [a [condition]$_{hypernym}$ in which all objects are seen tinged with red.]$_{Definition}$

2. [Devic disease]$_{Term}$, [also known as Neuromyelitis Optica, or Devic's syndrome]$_{Syn}$ is [an inflammatory [disease]$_{hypernym}$ of the central nervous system.]$_{Definition}$

These kinds of elements were considered in our chunk grammar in order to enhance scope to fragments that deviates its behavior from the canonical:

$$Syn1: \begin{pmatrix} \langle cm \rangle ? \langle cco \rangle \langle art \rangle ? \begin{pmatrix} nc \mid np \mid alfs \\ \mid acrnm \end{pmatrix} + \\ \langle adj \mid nc \mid np \mid alfs \rangle * \\ \begin{pmatrix} \langle pdel \rangle \langle nc \mid np \mid alfs \mid acrnm \rangle + \\ \langle adj \mid nc \mid np \mid alfs \rangle * \end{pmatrix} * \end{pmatrix}$$

$$Syn2: \begin{pmatrix} \langle cm \rangle ? \langle cco \rangle ? \langle cque \rangle ? \langle adv \mid vlfind \rangle + \\ \langle csubx \rangle ? \langle art \rangle ? \langle nc \mid np \mid alfs \mid acrnm \rangle + \\ \langle adj \mid nc \mid np \mid alfs \rangle * \\ \begin{pmatrix} \langle pdel \rangle \langle nc \mid np \mid alfs \mid acrnm \rangle + \\ \langle adj \mid nc \mid np \mid alfs \rangle * \end{pmatrix} * \end{pmatrix}$$

5.7 Contextual Patterns of DCs

Verbal patterns present in DCs can give clues about position of the term and hypernym. For instance, when verb *to be* is present in DCs, generally term can be found in the left side of verb and hypernym after verb, that is, we have the next syntactical configuration: (T)<vsfin>(HP1)(Def), where term (T), verb to be <vsfin>, noun phrase of the hypernym (HP1) and rest of definition (Def). Table 6 shows some contextual patterns of analytical fragments extracted from candidate DCs. Elements within parentheses represent constituents and those in angular parentheses POS tags.

Table 6. Contextual patterns of DCs

Contextual patterns
(T) (Syn1| Syn2)?<vsfin> (HP1|HP2|HP3)(Def)
<se>?<ppc>?<vlfind><csubx>? (T) (Syn1| Syn2)?<pa|pal> (HP1|HP2|HP3) (Def)
(HP1|HP2) <vlfind> <csubx >?(T)

5.8 Heuristics for Extraction of Hypernyms

Lexical relationships represent an approach from the structural semantics for organizing a conceptual space. If we assume words represent concepts, then structural semantics proposes to organize a conceptual space by means of lexical relationships such as hyponymy-hypernymy, synonymy, meronymy-holonymy, and antonymy. Lexicons and ontologies are resources where these relationships are useful.

Once the inclusion of patterns for terms and hypernyms as well as common verbal patterns in analytical definitions is presented, we propose the extraction of hypernyms from DCs by applying the following heuristics:

1. Division of a DC fragment according to the verb pattern. For instance, if contextual pattern of DC is (T) <vsfin>(HP1), then a string of text is divided into two segments with respect to <vsfin>. For example, in Python code this is done by:

Fragm.split("ser/vsfin")

2. Filtering empty heads such as *any*, *one*, *type*, *kind* and noun heads indicative of other relations (part-whole and causal relations) in noun phrase of hypernym. Cases in table 7 represent examples. Noun heads indicative of hyponymy are:

Type, kind, subtype, class, subclass, form, specie, example, version, field, subfield, discipline, subdiscipline.

When one of these heads is found, hypernym is located after preposition *of*. However, if there is no prepositional phrase after empty head, hypernym is left empty. On the other hand, if noun heads indicative of causal or part-whole relations are found, DC is considered non-relevant. Noun heads of causal and part-whole relations are:

Part-Whole: part, piece, constituent, fragment, component, portion, segment, fraction.

Causal: cause, consequence, effect, result, product, reason, origin.

Table 7. Examples of non-hypernym heads

Fragment of DCs
1. Smog is a kind of air pollution…
2.Macular edema is the cause of loss of vision ...
3. The intestine is part of digestive system …
4. An invasive tumor is that that is extended to surrounding areas …
5.Digestive apparatus is a set of organs …
6. Tableware is a set of dishes, glasses and cups …

3. Noun heads such as *set, subset, group and family* are filtered as indicative of hyponymy, but we considered it important to analyze the rest of the sentence with the goal of discarding the presence of a list of elements. For example, in case (5) noun head *organs* can be considered as a hypernym, but in case (6), *dish* is not a hypernym

because there is a list of elements indicating another type of relation, in this case, a part-whole relation.

4. When, there is no noun head and only a determinant as *that* followed by a relative as *that* or *which*, hypernym is extracted from noun phrase of the term, that is, noun head of noun phrase of term is considered as hypernym. If noun phrase of term has only one noun, then hypernym is left empty. An example of this heuristic is case (4), here hypernym corresponds to *tumor*.

5.9 Setting Thresholds for Improving Precision

With the goal of improving precision in extraction of DCs, we proposed an additional filter phase. This phase takes into account the occurrence frequency of hypernyms obtained. We assumed that more frequent hypernym candidates have a higher probability of being true hypernyms than those less frequent. So, it is possible to determine a threshold of frequency in order to get the most relevant results.

6. RESULTS

6.1 Resources

6.1.1 Sources of textual information

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. MedlinePlus is a site with a goal to provide information about diseases, treatments, and conditions that is easy to understand..

The kind of communication used in this textual source can be considered as expert-beginner because the information is intended for patients, families, and it is created by various health institutes. Taking into account that each knowledge area has a different lexical set, this kind of communicative situation will be most explanatory using definitions where the meaning of the lexical set must be clarified.

The size of the corpus is 1.2 million of words. We chose a medical domain for reasons of availability of textual resources in digital format. Furthermore, we assume that the choice of this domain does not suppose a very strong constraint for generalization of results to other domains.

6.1.2 Computational tools

The programming language used in order to automate all tasks required was Python as well as the NLTK module. NLTK module constitutes a very valuable resource for research and development in natural language processing [25].

6.2 Analysis of Results

Chunk grammar is a method in order to extract patterns from texts. DCs are text fragments following a more or less regular structure where this kind of resources can be used with an acceptable performance. Table 8 shows results of precision, recall and F-measure obtained with our method in medicine corpus. Our baseline consists of extracting candidate DCs by considering presence of verbs shown in table 1 within sentences delimited by dot.

Results in table 8 show that chunk grammar achieved a recall of 58% and a precision of 62%. Applying first filter (filter 1) of relations causal and part-whole, a precision of 68% with a reduction of recall to 57% was obtained. On the other hand, the application of frequency thresholds of occurrence of hypernyms shows best results in precision, but as thresholds are increased, recall is significantly reduced. Therefore, if a high precision is an important issue in results, we can apply this kind of heuristics in order to get the best DCs.

Table 8. Precision, recall and F-measure

Phase	Recall	Precision	F-measure
Baseline	87%	17%	28%
Chunk Grammar	58%	62%	60%
Filter 1	57%	68%	61%
Filter 2, Freq ≥ 5	42%	80%	55%
Filter 2, Freq ≥ 6	40%	81%	54%
Filter 2, Freq ≥ 10	35%	84%	49%
Filter 2, Freq ≥ 20	27%	90%	42%

6.3 Comparing with other Works

There are significant works about task of conceptual extraction for various languages: French, English, German and Spanish. Specifically for Spanish, there are very few studies. [7] proposed a method of conceptual extraction for Spanish and it is considered a relevant work in this kind of tasks. The proposed method considers verbal patterns in order to extract DCs from texts as well as restriction rules for filtering non-relevant DCs. Verb heads used by [7] are reported in table 1 and they are the same considered in this work.

We compared both methods by applying Ecode tool and our program to the medicine corpus. Only for analytical definitions, Ecode achieved a recall of 46%, while with our method we obtained a 58%. On the other hand, our precision of 62% was higher than Ecode (41%).

Results show a significant advantage with respect to proposed method by [7], but we think it would be appropriate to apply our method to corpora of other domains in order to determine stability of our results.

7. CONCLUSIONS

We present a method to extract DCs from restricted domains. DCs are a rich source in knowledge with respect to a specific domain and they are useful in order to build from dictionaries to ontologies. Our method considers the extraction of lexical relations of hypernymy for filtering non-relevant DCs under the assumption that more frequent hypernyms have a higher probability of being true hypernyms. Hence, candidate DCs with most frequent hypernyms will have a higher probability of being true DCs.

Our results show that precision can be improved by setting frequency thresholds for more frequent hypernyms. So, if precision in results is required, heuristics as proposed in this work can give very useful results.

Despite the difficulty of defining most of concepts in terms of necessary and sufficient conditions, we considered definitions represent a point of view agreed about a concept in specialised domains, and they are useful in order to extract lexical relations and other kinds of relations (attributives, part-whole, thematic roles, and so on).

8. ACKNOWLEDGMENTS

This paper has been supported by the National Council for Science and Technology (CONACYT) of Mexico, Ref. 82050.

9. REFERENCES

[1] Smith, E. 1988. Psychology of human thought. Cambridge University Press.

[2] Rosch, E. 1978. Principles of categorization. In Rosh E. and Lloyd B. (eds.). Cognition and Cognitive Science. Elsevier.

[3] Smith, E. and Medin, D. 1981. Categories and concepts. Harvard University Press. Cambridge, Mass.

[4] Murphy, G. 2002. The big book of concepts. MIT Press. Cambridge, Mass.

[5] Wilks, Y., Slator, B., and Guthrie, L. 1995. Electric Words: dictionaries, computers and meanings. MIT Press. Cambridge, Mass.

[6] Buitelaar, P., Cimiano, P. and Magnini, B. 2005. Ontology learning from text. IOS Press. Amsterdam.

[7] Sierra, G., Alarcón, R., Medina, A. and Aguilar, C. 2003. Definitional Contexts Extraction from Specialised Texts. Proceedings: Language, Corpora and E-Learning, Frankfurt: Peter Lang Publish: 21-31.

[8] Gruber, T. R., A. 1993. Translation Approach to Portable Ontology Specifications, Knowledge Acquisition 5(2), 1993:199-220.

[9] Velardi, P., Fabriani, P., and Missikoff. 2001. Using Text Processing Techniques to Automatically enrich a Domain Ontology. In Proceedings of the ACM International Conference on Formal Ontology in Information Systems.

[10] Sowa, J. 2006. Categorization in Cognitive Computer Science. Handbook of Categorization in Cognitive Science, Elsevier.

[11] Laurence, S. and Margolis, E. 1999. Concepts and Cognitive Science. Concepts: Core Readings, Cambridge, Mass.: MIT Press.

[12] Malaisé, V., Zweigenbaum, P., and Bachimont, B. 2005. Mining Defining Contexts to Help Structuring Differential Ontologies. Terminology 11 (1). 21-53.

[13] Klavans, J., and Muresan, D. 2001. Evaluation of DEFINDER: A System to Mine Definitions from Consumer-Oriented Medical Texts. In Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01).

[14] Sager, J. 1990. A Practical Course in Terminology Processing. Amsterdam/Philadelphia. John Benjamins.

[15] Smith, B. 2003. Ontology. In Floridi, L. (ed.). Blackwell Guide to the Philosophy of Computing and Information. Blackwell. Oxford: 155–166.

[16] Sierra, G., Alarcón, R., Aguilar C. and Bach, C. 2010. Definitional verbal patterns for semantic relation extraction. In Auger, A. and Barrière, C. (eds.). Probing Semantic Relations: Exploration and Identification in Specialized exts. Amsterdam/Philadelphia. John Benjamins, 73-96.

[17] Ortega, R., Montes, M., and Villaseñor, L. 2007. Using Lexical Patterns for Extracting Hyponyms from the Web. In: MICAI 2007. Advances in Artificial Intelligence. LNCS, Vol. 4827, pp.904-911. Springer, Berlin.

[18] Aguilar, C. 2009 Análisis Lingüístico de Definiciones en Contextos Definitorios. Tesis de doctorado, UNAM.

[19] Estopà, R. 2003. Extracció de terminologia: elements per a la construcció d'un SEACUSE. Doctoral Thesis. IULA-UPF. Barcelona, Spain.

[20] Vivaldi, J. 2004. Extracción de candidatos a términos mediante la combinación de estrategias heterogéneas. Doctoral Thesis. IULA-UPF. Barcelona, Spain.

[21] Litkowski, K. 2002. Digraph Analysis of Dictionary Prepositions Definitions. Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia. Association for Computational Linguistics.

[22] Jurafsky, D. and Martin, J. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall. New Jersey.

[23] Croft, W. and Cruse, A. 2004. Cognitive Linguistics. Cambridge University Press.

[24] Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference of New Methods in Language. WEB Site: www.ims.uni-stuttgart.de~schmid.TreeTagger

[25] Bird, S., Klein, E. and Loper, E. 2009. Natural Language Processing whit Python. O'Reilly, Sebastropol, Cal.