Towards the Extraction of Conceptual Information from Corpora

Gerardo Sierra, Alfonso Medina, Rodrigo Alarcón, César A. Aguilar
Instituto de Ingeniería, UNAM. México 04510, D. F.
Fax: (015255) 56 22 81 37
{GSierraM, AMedinaU, RAlarconM, CAguilar}@iingen.unam.mx
http://iling.torreingenieria.unam.mx

1 Introduction

To obtain the terminology of a specific domain, as well as the corresponding definitions of every term, terminologists could either consult specialists, or could consult texts in the area. In order to simplify this last option, computational terminology has developed some tools to facilitate textual analysis for terminographical purposes based on specialised corpora.

From a computational linguistics point of view, specifically related to information extraction, terminology uses statistical and rule-based methods [Cabré et. al. 2001] to extract terms from specialised texts. Furthermore, terminology needs to identify the corresponding definitions of a specific term. Often, when an author introduces a new term, which is not well known to the readers, he/she provides the definition emphasising the new concept with a set of syntactic and typographical features. We thus call *definitional context* the structure consisting of the term, the definition and the emphatic features in a specialised text.

In order to develop a tool capable of extracting definitional contexts from annotated corpora, an inventory of recurrent patterns used by authors to introduce concepts is necessary, as well as a computational linguistic technique capable of identifying concepts from specialised texts.

One of the objectives of the Language Engineering Group, from the Instituto de Ingeniería, UNAM (National Autonomous University of Mexico) is to extract conceptual information from corpora. For this reason, the group is working on the study of recurrent patterns in definitional contexts, the elaboration of an annotated corpus on engineering, and the development of the tools required for conceptual information extraction.

In this paper we present the process of developing a definitional context extraction tool. First, we present the minimal elements of a definitional context. Second, we define a typology of all the recurrent patterns found in definitional contexts of Spanish specialised texts. Third, we present the kinds of tagging necessary. And fourth, we briefly mention the characteristics that a search engine must have.

2 Definitional contexts

Our investigation is based on previous efforts such as the systematic identification of definitions based on lexical and metalinguistic patterns [Pearson, 1998], the analysis of *Explicit Metalinguistic Operations* [Rodríguez, 1999] and the analysis of *Knowledge-Rich Contexts* [Meyer, 2001].

As a result of these efforts, we have established that a *definitional context* is a structure in a specialised text consisting of two minimal elements, the term and the definition, and the emphatic features which accompany them.

2.1 Term

A term is a linguistic sign (i.e., one word or one set of words), that makes reference to a specialised concept (Cabré, 1999; Estopà, 2001). The most relevant features of a term are:

- Syntactic structure. In Spanish a term may be constituted by a noun phrase, sometimes followed by one or more prepositional phrases. In some cases, a term can be a verbal phrase (infinitive verb functioning as a noun), and an optional set of prepositional phrases [Cardero, 2001].
- Highlighting elements. There are some cases where typographical marks (italic, bold, underline, capital letters) are employed to highlight the term within the definitional context.
- Anaphoric relationship. Terms do not necessarily appear explicitly in a definitional context. That is, they do not appear next to their definition; instead, there is an anaphoric relation to the term within the definitional context. For example, a term can be presented as the theme of a section in

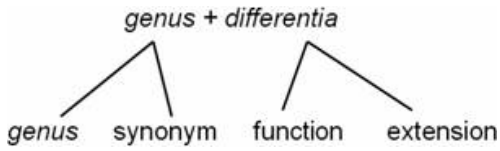
a text, given by the title, and explained within the section without being mentioned. That is, the anaphoric reference between the term and the definition may be the only clue as to their association.

2.2 Definition

A definition is the description of a concept represented by a term. This description establishes relationships with other terms, in order to delimitate the meaning of the concept. The most relevant features to understand a definition are:

- Typology. The starting point to identify the different types of definitions is the Aristotelian analytic definition [Sager/Ndi-Kimbi, 1995; Wilks/Slator/Guthrie, 1996], as represented in figure 1:
  - Analytic: genus + differentia.
  - Single genus: no further description of the differentia.
  - Synonymic: a strong semantic relationship with the genus.
  - Functional: differentia gives the function of the concept.
  - Extensional: differentia enumerates the parts composing the concept.

Figure 1. Analytic definition



- Syntactic structure. Most definitions start with a noun phrase, although verbal phrases are common in functional ones.
  - Noun phrases can start with a quantifier, determiner, or demonstratives.
  - The genus may consist of a set of prepositional phrases after the initial noun phrase.
  - The differentia may be introduced by subordinated sentences composed by noun, adjective and prepositional phrases.
- Highlighting elements. Some highlighting features emphasise the presence of definitions.
  - Quotation marks are one of the most common typographical features.
  - Authoring references usually accompany definitions.

3 Recurrent patterns in definitional contexts

We have seen that a definitional context is composed of two minimal elements: the term and the definition. Also, we have observed that there are some explicit characteristics that function as visual effects or grammatical features to help readers to identify the presence of an important concept. All these elements will be called ***patterns***. We use the following symbols to represent them: **T** (term)**, D** (definition), **tm** (typographical mark), **VP** (verbal predication) and **PP** (pragmatic predication).

Among possible sequences of these patterns, **T** could be connected with the definition by a **VP** or a **tm**. At the same time, a **tm** could characterise a **T** or a **D**. Here we will use the "+" sign to represent the sequences of those elements, while we concatenate the simultaneous combination of **tm** with **T** or **D**.

To study these patterns systematically we propose grouping them in three different sets: typographical, syntactic and mixed patterns.

3.1 Typographical patterns

We will call ***typographical patterns*** those sequences of terms and definitions connected by punctuation marks and commonly highlighted with typographical features (see examples in table 1).

- These patterns are the simplest forms found in specialised texts. They resemble the kind of definition commonly found in dictionaries.
- There are not verbs connecting terms and definitions; instead, punctuation marks appear, i.e. period, colon, coma, end of paragraph, etc.
- Typographical features are visual effects which emphasise the presence of either or both terms and definitions (bold, capital letters, quotation marks, etc.).

Table 1. Examples of typographical patterns

Pattern	Definitional context
T tm + tm + D tm	**Diseño:** *Desarrollo de configuraciones para la resolución de algún problema en base y sujetándose a sus restricciones.*
T tm + tm + D tm	DESASTRE. *Perturbación de la actividad normal que ocasiona pérdidas o daños extensos o graves*.
T tm + tm + D	"Impactos agregados sociales" ¶ Los que impactan a la sociedad, produciendo, por ejemplo, la perturbación de las relaciones familiares

3.2 Syntactic patterns

Another kind of definitional contexts is that where syntactic forms are used either to connect the term and the definition, or to provide some additional information about the concept. These patterns are called ***syntactic patterns*** and use pragmatic or verbal predications.

Table 2. Examples of syntactic patterns

Pattern	Definitional Context
PV + T + D	Se considera como protección civil a la actividad solidaria de los diversos sectores que integran a la sociedad…
T + PP + PV + D	Un soporte logístico de plataforma, de manera general, se define como un territorio equipado para el desarrollo de actividades logísticas…
PP + T + PV + D	De acuerdo con esta conceptualización, los daños probables se definen como el riesgo que corre el SA por ser expuesto al…

- **Pragmatic predications**. These syntactic forms give us information about usage or treatment of the term. Also, they give us some clues to understand a concept in the context it appears.
  - Pragmatic predications include adverbial phrases, e.g. *generalmente* (generally), prepositional phrases, e.g. *en términos generales* (in general terms) and simple words, e.g. *concepto* (concept).
  - The structure of pragmatic predications relies on the different styles each author uses to write a specialised document.
- **Verbal predications**. These forms use a verb to connect a term with a definition. The verbs used in definitional contexts are commonly called ***metalinguistic verbs***. Generally, metalinguistic verbs are employed to refer to language itself. In Spanish, some of the most important metalinguistic verbs are *definir*, *describir*, *denominar* (to define, to describe, to denominate). Because of their structure, verbal predications could be classified in two groups:
  - **Simple forms** use a verb or a verbal periphrasis that could appear with a grammatical particle, e.g. *afirma que* (he/she/it asserts that). Normally, the name of the author who defines the term, or the theory or textual reference where it is defined, appears within the definitional context.
  - **Complex forms** use the pronoun ***se*** plus a verb or a verbal periphrasis; e.g. ***se*** *define como* (it is defined as). This pronoun is the only difference with respect to simple forms. There is a semantic distance between the author and the definition he or she provides. This semantic distance is often provided implicitly by the impersonal meaning of the pronoun ***se*** along with the name of the author.

3.3 Mixed patterns

There is a kind of pattern that combines both typographical and syntactic elements. These patterns will be called ***mixed patterns***.

- They represent a more robust structure than the non-mixed patterns, because they emphasize typographically (punctuation and typographical marks) and syntactically the presence of definitional contexts.
- Mixed patterns are the most recurrent structure: they are present in 60% of the definitional contexts of our analysed corpus.
- As described above, typographical marks emphasise both the term and/or definition, while verbal predications are used to connect the terms with the definitions. Some pragmatic predications may also appear.

Table 3. Examples of mixed patterns

Pattern	Definitional Context
T tm + VP + D	**a. Canal de comercialización** es el conjunto de actores y actividades que interactúan para que un bien producido…
T tm + VP + D	- *Las actividades* se definen como los elementos principales de una acción…
PP + T + VP + D tm	Según G. Malagón (1996, p.18) un hospital se define como: "**una parte integrante de la organización médica, cuya función es la de proporcionar a la población**…

4 Corpus tagging

All of the above points to the idea of annotating corpora in order to extract the conceptual information contained in them. There are at least two levels at which the annotation schemes may be conceived; namely, that of visual effects or typographical tagging, and that of linguistic structure, consisting of both, part of speech tagging and syntactic or parsing structure tagging.

The analysis of our corpus resulted in the following model, which represents the possible combinations of the patterns discussed above:

(PP/VP) + T/D (tm) + (PP/VP/tm) + D/T (tm) + (PP)

We can see from this model that a definitional context could be identified by determining the presence of certain combinations of these constitutive elements. The automatisation of this implies the elaboration of an annotation scheme, which would necessarily include typographical and syntactic tags

4.1 Typographical tagging

Even when typographical tagging is not very extended in corpus-based research, for conceptual information extraction it is very important to retain the marked typography from the original. As we have seen, terms and definitions are usually emphasised because of the intrinsic linguistic interest given by the author. Therefore, marking typographical features is relevant.

Our preliminary analysis of definitional contexts let us identify the required tags to highlight the presence of concepts introduced in a printed text. The ways an author uses to emphasise the concepts differ from other authors and go beyond the most known tags in corpora. A deep study of the patterns shows this variety of ways:

- Terms appearing as the title of a text section and only mentioned paragraphs below in an anaphoric relation with its definition.
- Text superimposed outside the main text (pull quotes and footnotes) giving the definition for technical terms.
- Bullets displaying the term followed by a comma and the definition.
- Terms (and sometimes definitions) highlighted through word spacing, different font size, type or color, embedded within the text.

- Quotation within the same paragraph distinguished only by quotation marks or reference to an author.

Therefore, our corpus is rich in typographical encoding. Some of our tags include: different font, point type and color; word spacing; capital letters, small caps, subscripts and superscripts; division head elements such as titles, bylines, bullets, etc.; footnotes, endnotes, quotes and pull quotes.

Tagging punctuation is unnecessary as it can be searched without any encoding and the cost to encoding it is too high. However, it is convenient that the search engine clusters similar punctuation elements for conceptual extraction, such as: parentheses (including square brackets and braces); points (either full stop, comma, period and dash, colon, etc.); quotation marks (single or double).

4.2 POST

From the discussion presented above, it can be seen that the application of traditional POS tags to corpora can help when subsequent conceptual information extraction procedures are applied.
POS tagging is necessary to determine the internal structure of the prepositional, noun, verbal and adverbial phrases which build terms, definitions and, not less importantly, verbal and pragmatic predications.

EAGLES tag set provides special codes to specify attributes associated to each POS tag, which may or may not be relevant to information extraction. For instance, Spanish adjectives exhibit several attributes such as gender (feminine/masculine) and number (singular/plural).

However, for our purposes, according to the patterns presented above, most of the attributes typically associated to the Spanish part of speech categories are not needed for conceptual information extraction. First and second attributes of EAGLES encoding (category and type) will be enough for our purposes.

- It does not matter whether the verb forms exhibit inflexion morphemes corresponding to the future, present or past tenses, or to the particular modes (indicative, subjunctive, infinitive, gerundive, etc.).
- Whether or not the verb is an auxiliary verb seems to be very relevant.
- Grammatical person can be expected to be important: in our data, conceptual information is predominantly introduced by means of third person verb forms.

4.3 Parsing

After considering the structure of terms, definitions, and syntactic and pragmatic predications, as well as the syntactic relations among all of these structures, it should not be surprising that parsing annotation may also be of help when it comes to automatic conceptual information extraction.

- Noun phrases. They may function as terms and are important components of definitions.
  - As specified above, a term may consist of both a noun phrase and a prepositional phrase, or combinations of both [Cabré *et al*, 2001].
  - Definitions may be composed of at least one well formed sentence, which is likely to contain any of the syntactic structures mentioned above.
- Pragmatic predications. Their nature is more related to author's style or specialised jargon idiosyncrasies. Nevertheless, these are also constituted by syntactic phrases; *.e.g.* the prepositional phrase en *términos generales* (in general terms), the noun phrase *la característica principal* (the main characteristic) and the adverbial phrase *tradicionalmente* (traditionally). There might even be some overlapping, *e.g.* prepositional phrases may have an adverbial function and constitute pragmatic predications: *de manera general* (in a general manner), *según G. Malagón* (according to G. Malagón).
- Verbal phrases. Metalinguistic verbs are important clues to detect definitional contexts, as specified above. However, we have observed that non-metalinguistic verbs must also be considered; namely, those constituting verbal predications such as *se visualiza como* (it is visualized as), *se basa en* (it is based on).
- Other structures. Certain collocations have to be contemplated; for instance, those structures formed by a verb and a noun, where the verb has suffered some semantic erosion: *tiene la finalidad de* (it has the aim of), *tiene la función de* (it has the function of), etc. In a few words, we should consider the structure and variation of verbal predications in order to determine whether or not a verb functions as a connecting element within a definitional context.

Tagging all of these structures will help a search engine to determine what sequences of syntactic patterns occur that could signal the presence of a definitional context and to discriminate between what belongs to the definition and what does not.

5 Minimal requirements for a search engine

A conceptual information extraction system must allow the user to retrieve the elements of a definitional context from a specialised corpus: terms, definitions, verbal and pragmatic predications, and the typographical features involved. This implies the search for concordances and collocations of these elements.

- Term extraction. Need to search simple and lemmatised words, syntactic structures and typographical features.
- Verbal and pragmatic predication extraction. Need to search simple and lemmatised words, and syntactic structures.
- Definition extraction. Need to search syntactic structures and typographical features, as well as the retrieval of all of the above elements.

To retrieve definitional contexts, the system will permit the search for combinations of the elements presented here.

6 Conclusions and future work

In this paper, we have examined some recurrent patterns that characterise definitional contexts, namely syntactic and typographical. The latter are the visual effects that authors may apply to texts in order to emphasise important information such as definitional contexts. The former refer to the linguistic structure typically found to codify conceptual information and can be embodied by syntactic and pragmatic predications. On the one hand, what we have called syntactic predications are connectors among terms and definitions and typically contain verbs called metalinguistic because they refer to language itself. On the other hand, pragmatic predications give us information about author's style and usage or treatment of the term. More importantly, they give us relevant clues to understand concepts in their contexts. All of this is a motivation to annotate corpora for automatic extraction of conceptual information. It is clear that our annotation scheme is to consider both typographical and linguistic information. In this sense, it will include at least typographical, part of speech, and parsing tagging.

Our investigation can certainly be extended in different directions, some of which are in progress.

- Definitions. Further studies of their structure should be permanently undertaken given the complexity of what a definition is. It will let us delimitate the boundaries of a definition in a definitional context.
- Verbal predications. We need to identify more verbs that can be employed to define terms. Also, we need to evaluate their accuracy.
- Pragmatic predications. It is important to study the structure of pragmatic patterns in order to expand our initial paradigm.
- Verbs and definitions. We consider necessary to explore how a specific type of definition relies upon a particular verbal predication. L'Homme [2002] and Estopà, Lorente and Foguerà [2002] have explored this topic.
- Compound patterns. Although we have not described them here, there are other patterns of definitional contexts. These patterns are sequences where more than one term is defined. Pragmatic predications become essential in these sequences.
- Anaphora. There are anaphoric relations between terms and definitions. They are mechanisms of textual cohesion (Botley & McEnery, 1999). Researching these may be of help for the localisation of definitional contexts and their boundaries.

7 References

Botley S & McEnery A M 1999. “Discourse Anaphora. The Need of Synthesis”. In Botley & McEnery (edits.), *Corpus-based and Computational Approaches to Discourse Anaphora*. Amsterdam/Philadelphia, John Benjamins Publish.

Cabré, M T 1999. *La terminología. Representación y comunicación*. Barcelona, IULA-UPF.

Cardero, A 2001. *El procesamiento de una terminología. Referencia especial a la terminología de control de satélites en el área de las telecomunicaciones en México*. PhD Thesis, Mexico, UNAM.

Estopà, Rosa 2001. “Elementos lingüísticos de las unidades terminológicas para su extracción automática”. In Cabré T, Feliu J (eds), *La terminología científico-técnica*. Barcelona, IULA-UPF.

L'Homme, M.C. 2002. "What can Verb and Adjectives tell us about Terms?" In *Terminology and Knowledge Engineering, TKE 2002. Proceedings*, August 28-30 2002, Nancy (France), pp. 65-70.

Rodríguez, C. 1999. *Operaciones Metalingüísticas Explícitas en Textos de Especialidad*. Treball de Recerca. M.A. Thesis. Barcelona, IULA-UPF.

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

Wilks Y, Slator B, Guthrie L, 1996 *Electric Words. Dictionaries, Computers and Meaning*. Cambridge, Mass. MIT Press.