# What Do We Mean When We Speak About Named Entities?

Oriol Borrega, Mariona Taulé
and M. Antònia Martí[1]

## 1. Introduction

The concept of Named Entity (NE) has its origin in the Named Entity Recognition and Classification (NERC) tasks, an offspring of Information Retrieval systems, and became one of the main interest points in the Sixth and Seventh Message Understanding Conference (MUC-6, MUC-7) competitions, held back in 1995 and 1998 (Cinchor, 1997; Black, 1998). From then on, most competitions and programs include at least one task related to it. That is the case, for instance, of the Automatic Content Extraction (ACE) program, which began in 1999, and whose aim is to develop technologies in order to automatically infer from human languages the *entities* being mentioned, their *relations* and the *events* in which they participate (Doddington et al., 2004). So, the concept of NE appeared in an environment of NLP applications and is far from being linguistically clear and settled. Although there are some disciplines like linguistics and language analysis which study some particular items that share some characteristics with it, they do not fill in the needs to properly set solid criteria to handle NERC tasks.

On the one hand, the classic grammatical approach to Proper Noun (PN) analysis is surely insufficient to deal with the problems NERC poses. In one of the reference contributions on the matter in Spanish (Fernández Leborans, 1999), PNs are defined in their prototypical uses in regard of a series of common but not unique features, such as capitalization, lack of inflection, lack of lexical meaning, lack of determiner, lack of translation, monoreferenciality, or incompatibility with restrictive complementation. As can be seen, that approach cares only about PNs in the strict linguistic sense, so it leaves a vast amount of cases out of consideration (such as alphanumerical expressions or temporal expressions) which are relevant when talking about NEs. This proposal has proven to be too narrow to be used in NLP, especially in the case of weak named entity detection and classification, for weak named entities usually do not even contain a PN.

On the other hand, the philosophical approach to language and reference, via the concepts of Singular Terms and Definite Descriptions, is far too wide. Summarizing a lot of bibliography in a few words (Fernández Moreno, 2006), Singular Terms include Proper Names, singular Indexical Terms and singular Definite Descriptions. Definite Descriptions are expressions introduced by a singular definite article which predicate a property possessed by a single individual. Singular Indexical Terms include paradigmatically both pronouns (personal and demonstrative) and some adverbs (place and time deictics). Finally, PNs are defined by two features: their reference does not depend on the context of emission, and it is not dictated by the internal structure of the name itself. This approach deals with a number of linguistic phenomena which should not be part of a linguistically-based approach to NE (at least in most of the cases). In fact, most Reference Theories treat Proper Names, most pronouns and any definite noun phrase as a definite expression, whereas NERC systems care only about a subgroup of

---

[1] Department of Linguistics, University of Barcelona
*e-mail*: oriol.borrega@thera-clic.com, mtaule@ub.edu, amarti@ub.edu

these: definite noun phrases with single referents.

We should assume therefore that the concept of Named Entity, from a computational linguistic approach, is somewhere in the cross-section of these two points of view. But, furthermore, NERC is a task which strongly depends on the particular applications it is being developed for. That means the consideration of a particular string of language production may vary from a competition to another or from a NLP system to another, as often happens with terminology, numerical expressions and other elements which are not strict Proper Names.

The set for our proposal to NE is the CESS-ECE corpus. Our approach is motivated by the fact that we wanted to annotate this corpus at different levels in order to make it a fitting testing ground to prove certain linguistic hypotheses, and a fitting basis which would allow the further development of several Machine Learning systems. Thus, we had to define a set of general annotation criteria for NE (as well as for some other tasks) which would be interesting further on for a variety of processes and applications.

Our interest of defining the concept of NE relies in the need to give NEs a good linguistic basis which would set an accurate frame for the subsequent computational tasks. Up to the moment, most approaches to NEs brought under attention mainly the computational problems and difficulties to solve in order to improve the scoring and the performance of each algorithm. We wanted to settle from the beginning a series of clear criteria to tell the noun phrases which are a NE from the ones that are not. Also, these criteria would be followed by different annotators to avoid inconsistencies, thus improving the overall quality of the annotation task.

The aim of this paper is to present the criteria established in order to identify what is a NE, and to define their formal and semantic boundaries. In section 2, CESS-ECE corpus will be presented; in section 3, we will discuss the general criteria followed in the annotation, and the labels we have used (section 3.1); in section 4, we will further discuss these general criteria in detail, beginning with our approach to Trigger Words (section 4.1) and undergoing then a formal characterization of NEs (section 4.2) and their semantic characterization (section 4.3); in section 5 we will explain some semantic changes which may occur in NEs; finally, section 6 contains the conclusions of the paper.

## 2. CESS-ECE Corpus

The corpus we work on is still under construction. It is a subset 200.000 words for Catalan and 185.000 words for Spanish taken from CESS-ECE corpus. CESS-ECE corpus is a multilingual Treebank enriched with different kinds of semantic information. It consists on 500.000 words for each language (Spanish and Catalan) taken mainly from newspaper’s corpora (Figure 1). In the moment of annotating it, the corpus had already undergone an automatic process of morphological annotation and syntactic chunking, and a manual process of deep syntactic annotation including constituents and functions.

The annotation with NEs was being done in parallel with the semantic tagging of roles. Both tasks were manually done and are currently completed for the subset mentioned above (approximately 200.000 words for each language).

Each level of annotation is the basis for the next one: the whole is a modular process. Therefore, our input, being the output of the manual deep syntactic annotation, had already morphological and syntactic labels (constituents and functions).

| **Corpus** | **Sources**[2] | **Notation** | **Process** |
|---|---|---|---|
| CESS-ECE-CAT (Catalan) | EFE (75.000)<br>ACN (225.000)<br>El Periódico (200.000) | Morphological | Automatic |
| | | Superficial syntax | Automatic |
| | | Deep syntax | Manual |
| CESS-ECE-ESP (Spanish) | Lexesp (85.000)<br>EFE (225.000)<br>El Periódico (200.000) | Morphological | Automatic |
| | | Superficial syntax | Automatic |
| | | Deep syntax | Manual |

**Figure 1**: Composition of the corpus

As a result, when notating the corpus with NEs, the annotators were forced to bring under consideration and follow the syntactic patterns of the text. Notation with NEs was made at two different levels: PoS for Strong Named Entities (SNEs) and syntax for Weak Named Entities (WNEs). The combination of these two facts, which would seem logical and positive in principle, has in fact some major advantages and drawbacks.

As advantages go, it is obvious that, as only noun phrases were notated, syntactic analysis makes it easier to detect and label only the pertinent nodes of the syntactic tree, without need for determining where each phrase begins and ends. Furthermore, Proper Nouns (PNs), numerical expressions and temporal expressions had already specific morphological labels, helping both their detection and annotation.

But syntactic trees were sometimes found to be too rigid, a condition which led us to notate some really long noun phrases including all complements of a determined NE, whereas, sometimes, we would have rather notated only the core of the phrase, or the main phrase without its dependent complements. In some occasions, the guidelines followed in the syntactic notation have posed some problems. That is the case, particularly, of the way coordination was treated. Coordinate structures of several nominal elements, in Spanish and Catalan, may or may not repeat the article in each one of its instances. When the article is not repeated, the coordination was not made at phrase level, but only at noun-group level, a level below noun phrase. In that case, if the elements coordinated were NEs, we could not notate them at syntactical level, for they were not complete noun phrases (Figure 2).

---

[2] EFE is the Spanish news agency. ACN (Agència Catalana de Notícies) is the Catalan news agency. The subgroup of 200.000 words from the newspaper “El Periódico” consists in the same news both for Catalan and Spanish, from January to December 2000. Lexesp is a balanced corpus of 6.000.000 words for Spanish.

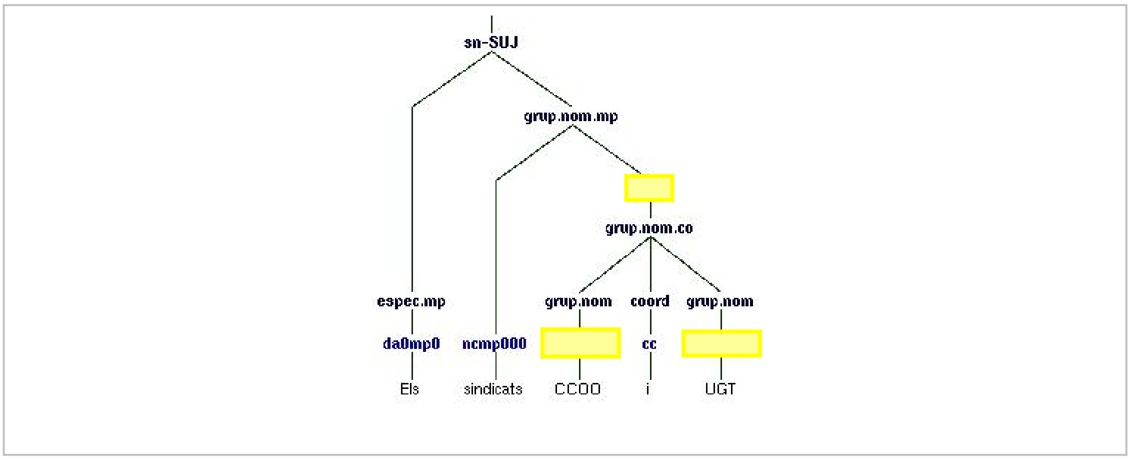

**Figure 2**: Coordination at noun-group level (Trade Unions CCOO and UGT)

## 3. General criteria: What's a NE?

The classical definition of what NEs are would read in the lines of NEs are the whole of the relevant entities in a particular dominion or application. Even though we agree on the usefulness and motivations of such definition, we think it is not complete. The guidelines we discuss in this paper were formulated in order to face different aspects of the task we had to fulfil, and are an attempt at completing the definition above.

There are two underlying golden rules in the definition of NE we propose. On one hand, only a noun or a noun phrase can be a NE. On the other hand, the referent of any NE must be unique and unambiguous. As we will see further on, these rules need to be specified and exactly determined by means of a complete set of more concrete and accurate guidelines in order to be applied to the annotation of the corpus. For instance, there are a number of exceptions to them that need be examined and discussed: even though it is clear that a non-definite noun phrase will never be a NE according to our criteria, there are some cases of plural definite noun phrases that we have chosen to consider and annotate as NEs.

First of all, we distinguish between Strong Named Entities (SNE) and Weak Named Entities (WNE). SNE is a morphosyntactical category, whereas WNE is a syntactical one. That means SNEs are a particular type of word (as are common nouns, verbs, adjectives and so on), while WNEs are a particular type of Phrase (comparable to noun phrases, adjective phrases, etc.). Later on we will discuss the different labels used to annotate each one of these elements.

We define SNEs as being formed by a word, a number, a date, or, in some cases, a string of words referring to a single individual entity in the real world, as is the case, for instance, of a person's name and surnames (*John Kennedy Toole*), a book's title (*A Confederacy of Dunces*) or some names of geographical features, countries, etc. (*New Orleans*). In these cases, we consider, analyze and annotate the whole string as a single element with its corresponding PoS tag.

In our approach, WNEs consist on a noun phrase, be it simple or complex. Therefore, they are syntactic elements. Note that a single SNE may make a complete noun phrase and, therefore, in some occasions, the linguistic form of a particular SNE and a particular WNE may match up perfectly (see Figure 5 below). Otherwise, as we will see in further detail on section 4 of this paper, it is not necessary that WNEs have a

SNE in them as a constituent. Some definite noun phrases whose core is a common noun may become a WNE because of syntactic, semantic and pragmatic reasons. Furthermore, in the case of numerical expressions, a SNE may not form or be part of a WNE.

### 3.1 Labels used

In CESS-ECE corpus, each level of annotation (PoS, Syntax, Semantic Roles, etc.) has its own labels. In Figure 3, we refer the labels concerning NEs. Syntactic labels occur at phrase level, whereas morphological labels occur at PoS level.

| | **Label** | **Meaning** |
|---|---|---|
| **Morphological** | np0000p | Person SNE |
| | np0000o | Organization SNE |
| | np0000l | Place SNE |
| | np0000a | Other SNE |
| | Z | Alphanumerical SNE (Numbers) |
| | Zm | Alphanumerical SNE (Coins) |
| | Zp | Alphanumerical SNE (Percentages) |
| | W | Temporal SNE |
| **Syntactical** | snp | Person WNE |
| | sno | Organization WNE |
| | snl | Place WNE |
| | sna | Other WNE |
| | snn | Alphanumerical WNE |
| | snd | Temporal WNE |

**Figure 3**: Morphological and syntactical labels

The label *np** (proper noun) opposes to *nc** (common noun). *Np0000p* stands for *"nom propi de persona"* (person proper noun). *Np0000o* stands for *"nom propi d'organització"* (organization proper noun) and, following the grid, we find *"nom propi de lloc"* (place proper noun) and *"nom propi d'altres"* (Other proper nouns). *Snp* stands for *"sintagma nominal de persona"* (person noun phrase). The following stand for *"sintagma nominal d'organització"* (organization noun phrase), *"sintagma nominal de lloc"* (place noun phrase), *"sintagma nominal d'altres"* (other noun phrase), *"sintagma nominal numeral"* (numerical noun phrase) and *"sintagma nominal de data"* (temporal noun phrase).

## 4. Formal characterization of NEs

In this section we will record and discuss the main formal criteria we have followed in the detection, delimitation and classification of both SNEs and WNEs. Firstly, we will discuss our concept of Trigger Word (section 4.1), a capital element for the analysis and guidelines; secondly, the criteria to detect and delimitate NEs (section 4.2), including which PoS elements may be considered a part of a NE; and, finally, the semantic criteria to classify NEs (section 4.3). All subsections will be complemented with examples taken from the corpus itself.

### 4.1 Trigger Words (TWs)

Trigger Words (TWs) are a central element in the approach to NERC we have followed when annotating the corpus. They are the basis of great part of the answers we have given to NE classification and detection problems, especially when it comes to WNEs. Therefore, it is convenient to define how we consider and treat them.

In this approach, TWs are common nouns (as opposed to PNs) with some syntactic, pragmatic and semantic particularities. Not every common noun is capable of becoming a TW in a given environment, although it should be pointed here that different applications may require different semantic definitions of TW.

From a syntactical point of view, TWs are always the core of a noun phrase containing a PN or a similar element. That is to say, if a given noun is not the core of its phrase, it will not be considered a TW. Otherwise, that same noun may be considered a TW in other conditions. Compare *president Clinton said […]*, where *president* is a TW, with *Bill Clinton was appointed president*, where it is not.

From a pragmatic point of view, TWs may be considered as discourse markers which anticipate and point at the apparition of a NE. They are often complemented either by a PN, by a Prepositional Phrase containing a PN or by some other similar structure. Furthermore, TWs are a central element in most WNEs, defining their semantic classification and other features.

From a semantic point of view, TWs are always related to the NE categories we have defined beforehand. To put it in a graphic way, common nouns capable of becoming a TW in a system which would be used in an application in the field of Arts (such as *painter*, *opus*, *picture*, *etc.*) would not be the same as if that same system would be used in the field of mathematics (which should rather be the like of *theorem*, *law* or *corollary*). That means it is an application-dependant concept, semantically.

Linguistically speaking, there are two main problems in the use of TWs: quantity and variability.

It is impossible to make an inventory of all existent TWs, even of all existent TWs of a particular area of knowledge. Every day a new TW in that area may pop up, a new use of a determined name may turn it into a possible TW, etc.

All that said, we have decided to use TWs, despite their limitations, because they are a very powerful tool in the handling of WNEs. We use a big database of 4.000 semantically classified TWs extracted from the chunking of a representative corpus of news from different media (newspapers mostly). The semantic categories are *person*, *organization*, *place*, *coin* and *others*.

### 4.2 Formal guidelines: detection and delimitation

#### 4.2.1 Strong Named Entities

PNs, dates and alphanumeric expressions (numbers, coin names and percentages) are considered SNEs, and are likewise annotated (Figure 4).

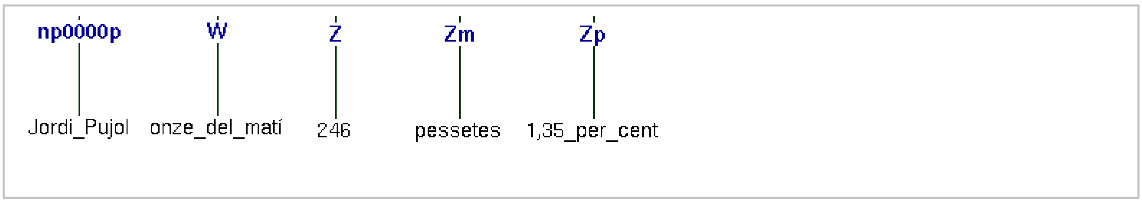

**Figure 4**: SNEs. ( Jordi Pujol // eleven o'clock in the morning // 246 // pesetas // 1,35 percent).

Most of them are identified automatically by de morphological analyzer and there is little manual intervention in their annotation, having only to correct some minor mistakes. The annotation tag is the PoS label.

WNE, on the other hand, are syntactic structures and they has to be annotated manually because there is no any syntactic analyzer of Catalan and Spanish able to delimitate and classify them.

### 4.2.2 Singular Noun Phrases containing Proper Nouns

All definite singular noun phrases containing a PN are considered WNEs, even if the PN is the only element of the noun phrase (Figure 5) or if it is not the core of the phrase (Figure 6). In the later case, the core should be a TW, complemented by a PN, a prepositional phrase containing a PN or some other similar structure. It is useful to bring under consideration that, especially in Spanish, many PN do not need a determiner and are, therefore, self-definite expressions.

In some occasions, it has not been possible to annotate WNE formed by a single PN in the syntactical level due to the limitations of the previous syntactic analysis of the corpus.

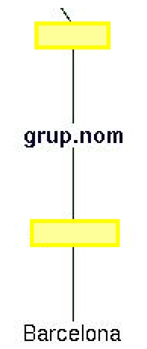

**Figure 5**: Match between SNE and WNE. The Proper Noun is the only element of its phrase.

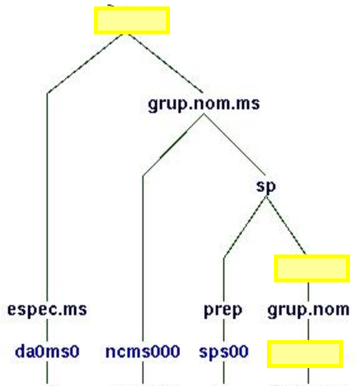

**Figure 6**: WNE with a PN which is not the core of its phrase (the village of Escaldes).

### 4.2.3 Dates, Coins, Percentages and Numbers

All definite noun phrases containing a nominal element with the numerical (Z, Figure 7), percentage (Zp, Figure 8), date (W, Figure 9) or coin (Zm, Figure 10) label in the PoS are considered WNEs. We consider percentages to be always definite expressions, even though in a strict linguistic sense they may be introduced by a non definite determiner. Note also that numerical expressions (numbers, Z label) may be nouns (as in *the number 6*), pronouns (as in *of all students, only six passed the exam*) or determiners (as in *only six students passed the exams*). In the latter case, even though we have chosen to consider the number a SNE and to annotate it accordingly, we have also decided that a determiner would not form a WNE, for it introduces not a definite singular noun phrase. Nevertheless, if a given NP should be annotated as a WNE otherwise because of other considerations, it has been done so, as is the case of coins (amounts of money).

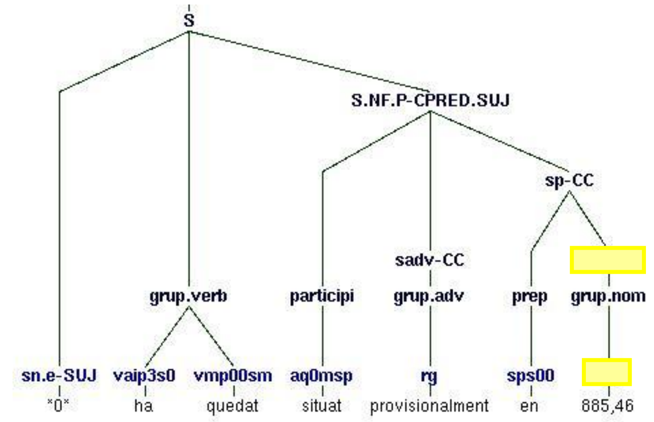

**Figure 7**: WNE with a numerical element in it (Ø has been set provisionally at 885.46)

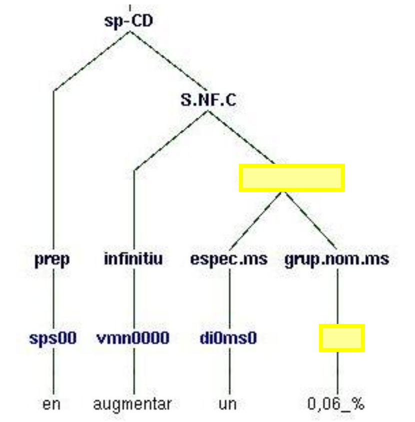

**Figure 8**: WNE with a percentage element in it (increasing by 0.06 percent)

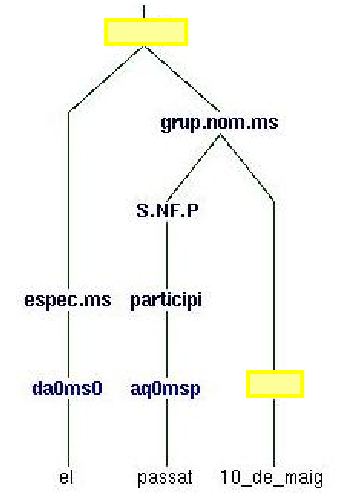

**Figure 9**: WNE with a date element in it (last May 10th).

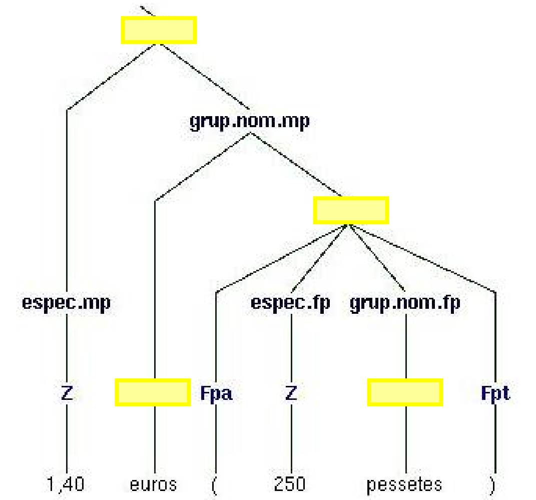

**Figure 10**: WNE with a coin element in it (1.40 euros –250 pesetas).

### 4.2.4 Singular Noun Phrases without Proper Nouns

All definite noun phrases whose core is a TW complemented either by a national adjective (Figure 11) or by a relational adjective derived from a PN (Figure 12) are considered WNEs. Usage is essential to decide which relational adjectives are capable of turning their noun phrase into a WNE and which are not. *Republican*, for instance, may be ambiguous, for it might mean “belonging to the Republican Party” as well as “one who is in favour of avoiding monarchy”. In such cases, we decided not to annotate the NP as a WNE. Note that in some other cases, there is a difference between relational adjectives and other adjectives, as in *barcelonista* (which means “belonging to the F.C.Barcelona”) and *barceloní* (inhabitant of Barcelona).

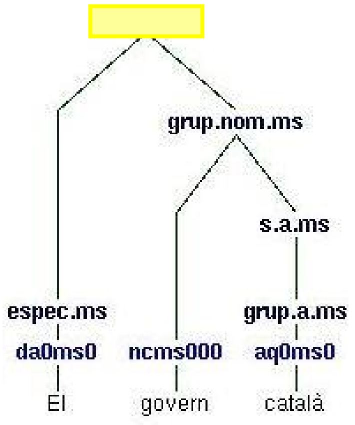

**Figure 11**: WNE without a PN, formed by a TW complemented by a national adjective (the Catalan government).

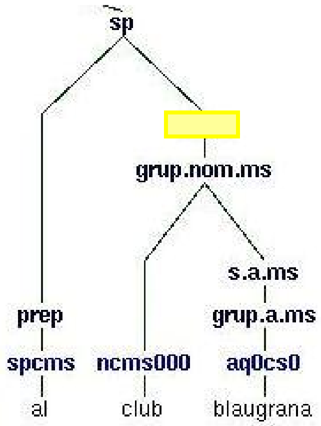

**Figure 12**: WNE without a PN, formed by a TW complemented by a relational adjective (the blue and scarlet club –namely, Football Club Barcelona).

### 4.2.5 Magnitudes

Noun phrases which contain a magnitude name determined by a numeral are considered WNEs (Figure 13). A list of magnitude names is being created to facilitate the automatic annotation of this kind of Entities, which linguistically have many common features with coin names. Once the list is created, a morphological label will be assigned to these elements (similar to the Zm of coins).

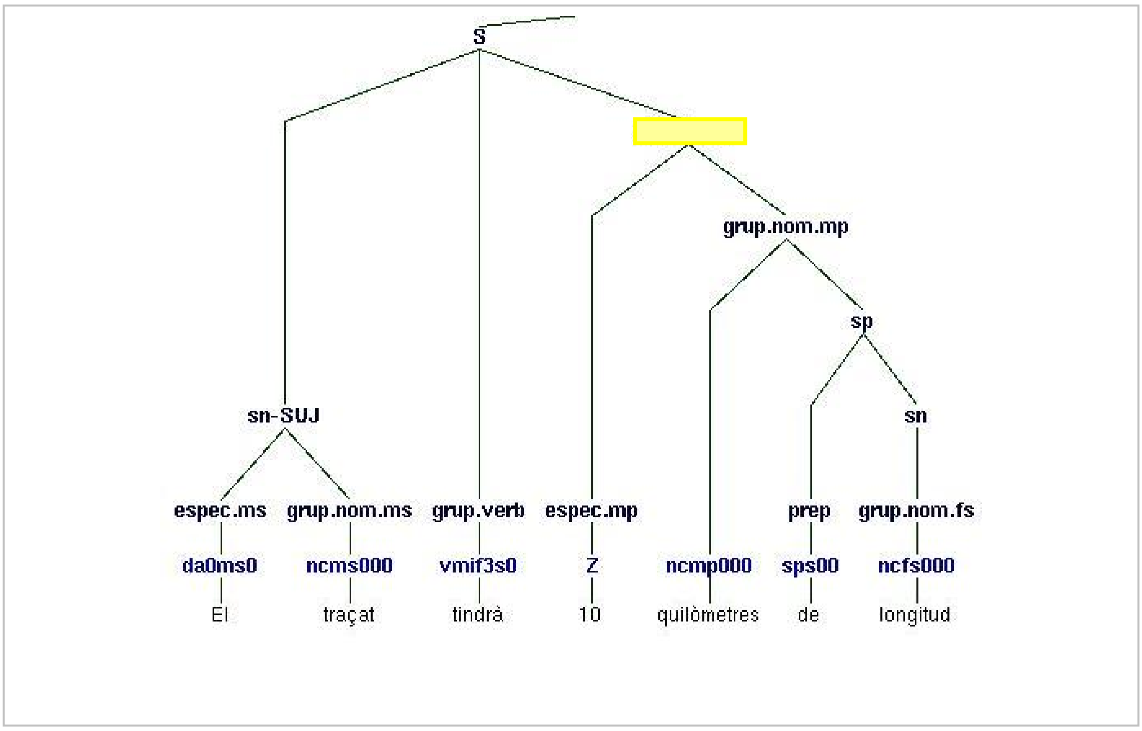

**Figure 13**: WNE formed by a magnitude name specified by a numeral. Notice the lack of a PoS tag for magnitude elements (The route will be 10 kilometres long).

### 4.3 Semantic guidelines: classification

The semantics of a NE are normally determined by the core noun of the phrase. Consequently, the classification proposed here is lexical-centred. We use six entity types: person, organization, place, temporal, alphanumerical, and others. As each one of these types poses particular problems, we will analyze them separately.

#### 4.3.1 Person NEs (syntactic label: snp; morphosyntactic label: np0000p)

Under this category we classify all singular definite expressions referring either to real or imaginary people or to entities which look human or have anthropomorphic representations (religious, mythological, folkloric or fiction characters). For instance: *Josep-LLuís_Carod-Rovira* (proper name), *l'alcalde de Súria* (the major of Súria), *el rei Carnestoltes* (King Carnival, a folkloric character).

Most Person NEs follow consistently the formal guidelines exposed above. However, there are two specific cases which demand deeper analysis: plurals and coordination.

*a) Person NEs and plurals*

We have stated that we do not consider plural NPs to be NEs: usually, an amount of individuals does not form a superior entity, but remains a group of independent individuals. Nevertheless, there are a few plural expressions relating to well defined human groups and, thus, capable of referring to a superior entity which should be treated and analyzed as a NE (Figure 14). We have detected two instances of this kind of expressions: ethnic groups, cultures, nations, etc.; and concrete demographic groups.

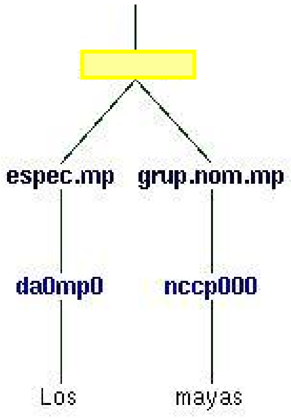

**Figure 14**: An example of plural person NE: ethnic groups (The

### b) *Person NEs and coordination*

From a semantic point of view, two coordinate Person NEs do not form a new NE a single entity, but remain two single entities, referring to two different individuals. Therefore, they will be annotated separately, leaving the union node untouched. But, in some cases, two coordinate NPs may refer to the same person (Figure 15). This happens when they predicate two different properties or qualities of that particular individual. In that case, the syntactic node of union will be annotated consequently as snp.

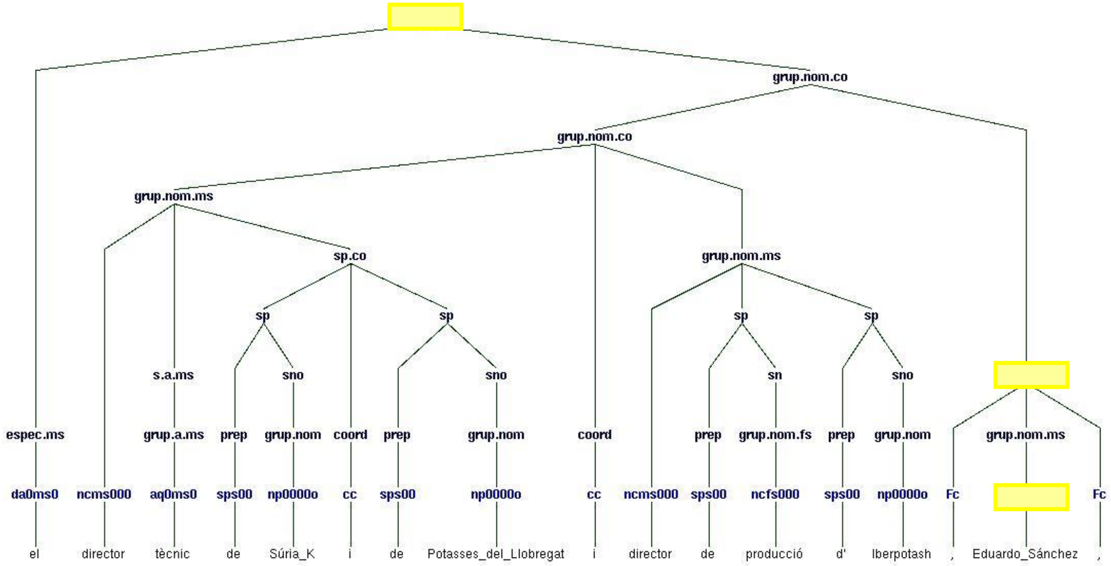

**Figure 15**: An example of two different NPs referring to the same person (the Epithet "Eduardo_Sánchez"). Notice the label "snp.x" on top, which indicates the existence of coordination (the technical manager of Súria_K and Potasses_del_Llobregat and Production Manager of Iberpotash, Eduardo_Sánchez,).

### 4.3.2 Organization NEs (syntactic label: sno; morphosyntactic label: np0000o)

Under this category we classify all singular definite expressions referring to stable human groups which have a well-defined structure: societies and enterprises, political and social agents, institutions, associations, sport teams, music bands, etc. In general, this kind of human groups has a PN to name it. For instance: *General_Motors* (proper name), *l'Orquestra_Simfònica_del_Vallès* (the Vallès symphonic orchestra), *el ministeri d'Afers_Exteriors* (the foreign office).

Even though we cannot reject the possibility of finding parallel cases to those described in the preceding point of this paper, our experience has shown us that it is highly improbable in the usage of the language. In fact, this is the category which resembles the most the ideal definition of NE we propose.

### 4.3.3 Place NEs (syntactic label: snl; morphosyntactic label: np0000l)

Under this category we classify all singular definite expressions referring to real or imaginary physical spaces: geographical features, political divisions of territory, places for human activities, imaginary places, mythological places, etc. For instance: *Catalunya* (Catalonia, proper name), *la regió metropolitana de Barcelona* (Barcelona's metropolitan area). Again, we must pay special attention to plurals and coordination.

a) *Place NEs and plurals*

There are particular places whose name is always plural (like *the United States of America* or *the Alps*). Needless to say, those are always considered NEs. But, furthermore, in some occasions we may refer to geographical spaces by means of a plural NP other than their PN (Figure 16). As always, the ultimate criterion to decide whether we annotate or not a particular phrase is the singularity of its referent – in this particular case, whether it refers to a single space or not. This phenomenon is more usual when we speak about political divisions of territories.

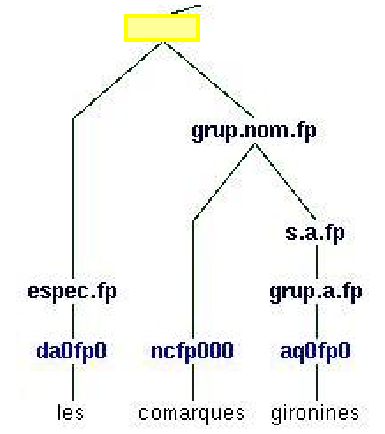

**Figure 16**: plural place NE (the counties in Girona)

### *b) Place NEs and coordination*

Because of the strategies we use to organize space in Spanish and Catalan, usually based in a series of reference points, we may talk about places using coordinate expressions. If the referent of such expressions is unique, they will be annotated as NEs (Figure 17). This may happen, for instance, when we talk about a single point situated in the crossing of two or more lines, or when we talk about a space situated between two or more distant points.

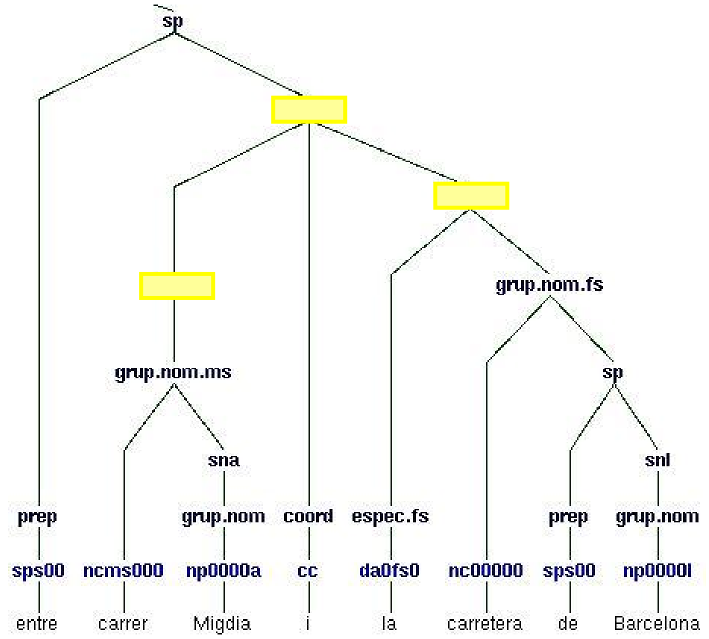

**Figure 17**: an example of coordinate place NE (in the crossing of Migdia Street and Barcelona Road).

### 4.3.4 Temporal NEs (syntactic label: snd; morphosyntactic label: W)

Under this category we classify all singular definite expressions referring to a particular moment or time period. It includes not only dates in the strict sense, but also hours, years, periods of history, etc. For instance: *dilluns_4_de_juny* (Monday, June the 4th), *any_1991* (year 1991), *les 15:55_hores* (15:55 hours).

Input has a great importance when annotating Temporal NEs. We have chosen to annotate only the phrases in which there is an element with the W label in the PoS as a result of the morphological analysis. In many occasions, days of the week or historic periods are annotated as common nouns, preventing us to label them anew and annotate the syntactic level with NEs. Some minor improvements will be made in the tokenizer to improve the quality of the annotation of these kinds of elements.

Syntactically speaking, the ways we talk about time in Spanish and Catalan are somehow similar to the expressions we use to refer to space, using points of reference. Therefore, we might find some coordinate temporal expressions which form a single Temporal NE.

### 4.3.5 Alphanumeric NEs (syntactic label: snn; morphosyntactic labels: Z, Zm, Zp)

This is the most heterogeneous of all categories. Under it, there is a variety of expressions whose common characteristic is that they all designate quantities or numbers. Otherwise, their syntactic behaviour is quite different one from each other. A reflex of this is that they do not share their morphosyntactic labels. Instances of each of these elements are:

Numbers: *220.000*, *3,6*.
Percentages: *40_per_cent*, *46_%*.
Coins: *10 milions de pessetes* (10 millions pesetas), *37.500 euros*.
Magnitudes: *10 hectàrees* (10 hectares), *46,52 nusos* (46.52 knots).

*a) Numbers (morphological label Z)*

Numerals may be nouns, pronouns or determiners. In all three cases, the analyzer annotates them with a Z label in the PoS. Therefore, they are all considered SNEs. But, in order to determine which numerals may be treated also as part of a WNE, we have to consider that syntactic difference.

Whenever the numeral is a noun or a pronoun, the noun phrase it is into will be considered a numerical WNE if its core is either the numeral itself (Figure 18) or a numeral TW (Figure 19). In terms of usage, these are the less frequent cases for numerals.

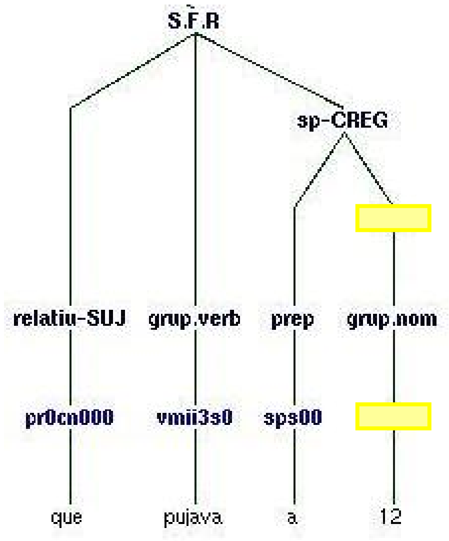

**Figure 18**: numerical NE with a numeric element as the core of the phrase (which increased to 12)

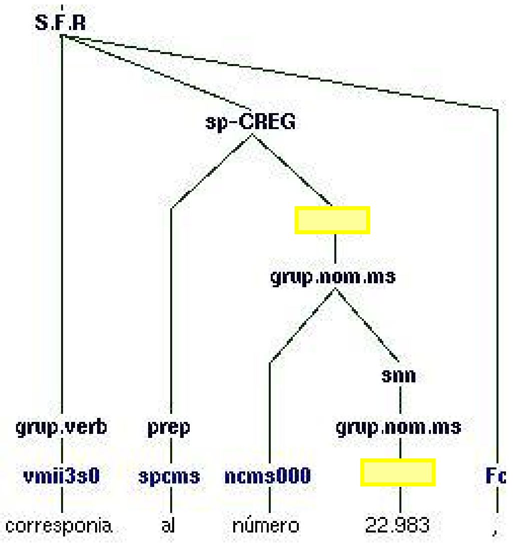

**Figure 19**: numerical NE with a numerical TW complemented by a numerical apposition (corresponding to the number

On the other hand, if the numeral is a determiner, the phrase it introduces will not be considered a WNE, for it will not be a definite singular noun phrase. But, as we have already seen, and will see in the following points of this paper, there are cases in which WNEs may be syntactically plural. If the numeral is the determiner of a plural noun phrase which would be considered a WNE following the guidelines we have explained, that expression will be consequently annotated. In this case, the semantic category of the WNE will be determined by the core noun of the phrase.

b) *Percentages (morphological label Zp)*

All noun phrases, whose core is a percentage with the Zp label in the PoS are considered an Alphanumerical WNE, regardless of whether they are syntactically definite or not. That is to say, we consider that percentages are always definite expressions *per se*, even if they are preceded by an indefinite determiner. There are many different ways to write down a percentage: using numbers, words, phrases, etc. (Figure 20). Whenever the morphological analyzer allocates the Zp label in a particular element, it will be considered a percentage, and annotated likewise in the syntactic level.

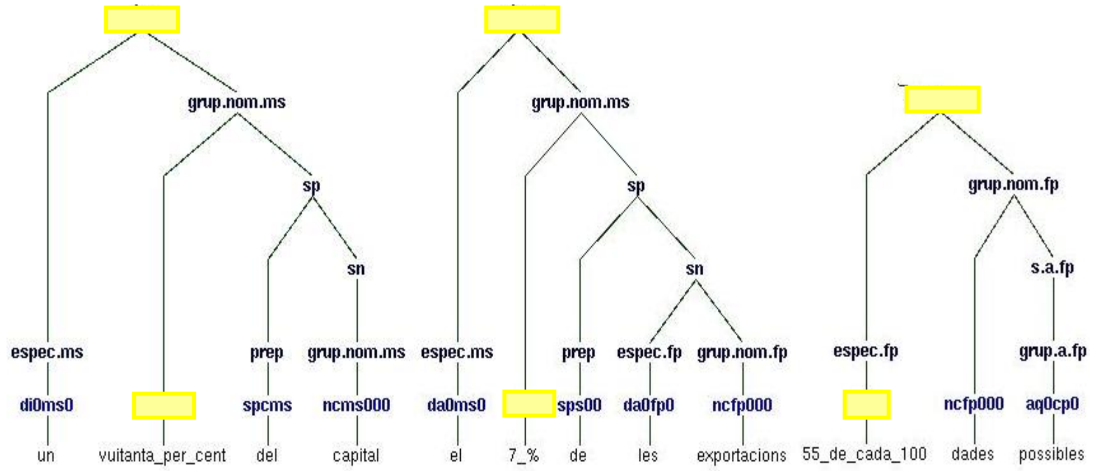

**Figure 20**: three instances of percentage NEs (eighty_per_cent of the capital // 7_percent of the exports // 55_of_every_100 possible data).

Again, we must point out that coordination is a possible way of referring to a percentage or to an interval between two percentile points (Figure 21). As always, if the referent of the expression is unique, it will be annotated.

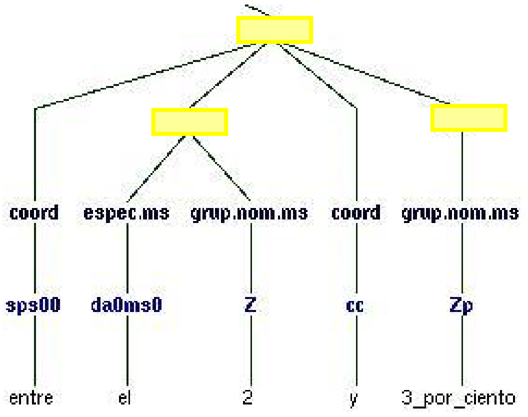

**Figure 21**: coordinate percentage NE (between 2 and 3_percent)

### *c) Coins (morphological label Zm)*

The behaviour of coin names is similar to that of some PNs. For that reason, and for the relevance of these kinds of expressions in most environments, we have decided to consider WNEs all expressions with a Zm label in the PoS which refer to an amount of money (Figure 22). In most occasions, the core of these expressions is not the element with the Zm label, but a numeral TW (such as million, thousand, etc., Figure 23). Again, most of the time these expressions are introduced by a numeral determiner (with a Z morphological label) and are, therefore, plural phrases.

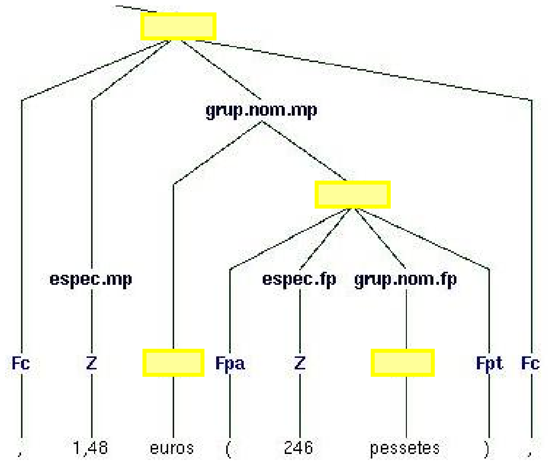

**Figure 22**: coin NEs (, 1.46 euros -246 pesetas-,).

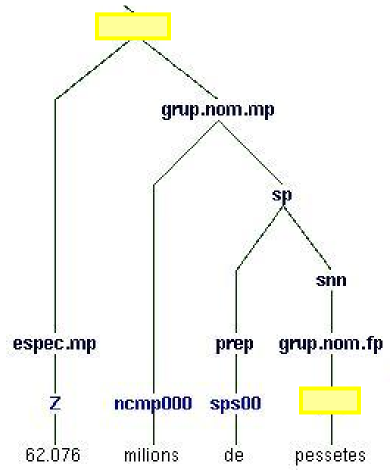

**Figure 23**: coin NE formed by a numeric TW complemented by a phrase containing a coin name (62,076 millions pesetas).

### *d) Magnitudes (without specific morphological label)*

Under this category we classify all expressions which refer to distances, capacity, volumes, temperatures, etc. by means of the names of the units used to measure them (*meters*, *degrees*, *bites*, etc., Figure 24).

These expressions have many features in common with coin names: they are usually plural, they are introduced by a numeral determiner, and their core may not be the magnitude name itself, but a numeral TW.

In the moment of beginning the annotation of the corpus, we had not a specific morphological label for these words. Therefore, magnitude NEs are currently only a syntactic category, identified by the label snn. A list of magnitude names is being created in order to give them a morphological label which will allow our system to also annotate them at PoS level. Once it is done so, the annotated corpus will be revised.

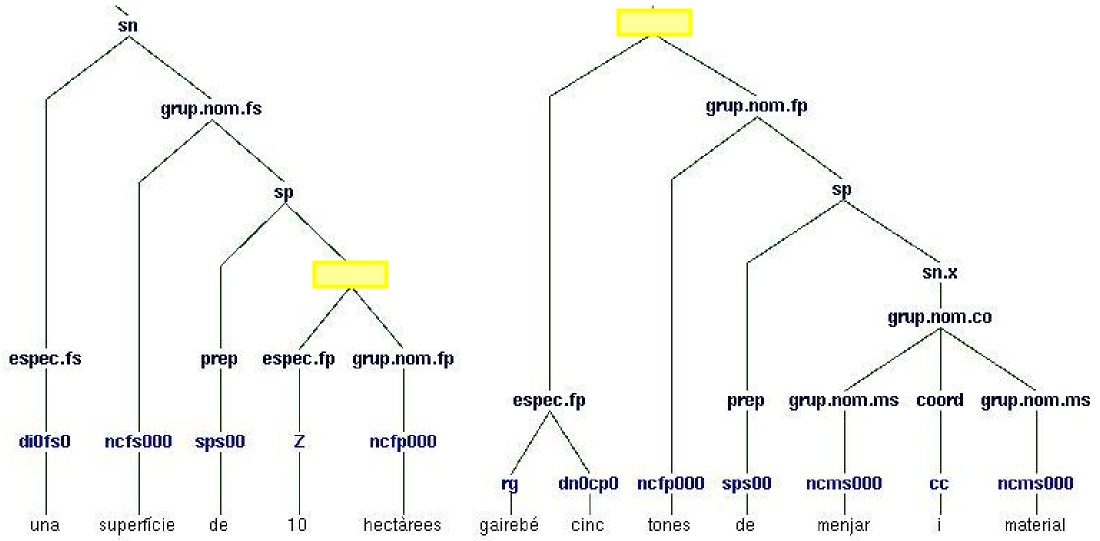

**Figure 24**: magnitude WNEs (a surface of 10 hectares // almost five tons of aliments and materials).

### 4.3.6 Others NE (syntactic label: sna; morphosyntactic label: np0000a)

Under this category we classify all singular definite expressions referring to single, univocal entities which cannot be classified under any of the categories above. Even though this category may look a bit of a ragbag, the fact is that we can establish a series of sub-categories which cover most instances of WNEs classified as “others”:

a) Animal individuals (real or fictional) and fiction beings without anthropomorphical representations: *Tom i Jerry*, etc.
b) Publications and art works: pictures, books, journals, internet addresses, etc.: *www.oasi.org*, *La_Vanguardia*, *El_sopar_dels_idiotes* (The Dinner Game), etc. (Figures 25 and 26).
c) Laws and regulations: *Pla_Hidrològic_Nacional* (National Hydrologic Policy), etc. (Figure 27)
d) Theories, scientific or philosophic laws, ideologies, etc.: *Estado_del_Bienestar* (Welfare State), *kantismo* (kantism), etc.
e) Historical events: *la Guerra_Cívil* (Spanish Civil War), *Any_Dalí* (Dali’s Year), etc. (Figure 28).
f) Artefacts, inventions, machines, etc.: *TGV* (High-Speed Train), *Internet*, etc. (Figure 29).
g) Competitions, prices, decorations: *Premi_dels_Escriptors_Catalans_2003* (Catalan Writer’s Prize, 2003), *Champion’s League*, etc.
h) University degrees and careers: *Administració_i_Direcció_d’_Empreses* (Management and Direction), *Càtedra_d’_Estudis_Marítms* (Sea Studies Cathedra), etc.
i) Heavenly bodies: *el Sistema_Solar* (The Solar System), etc.

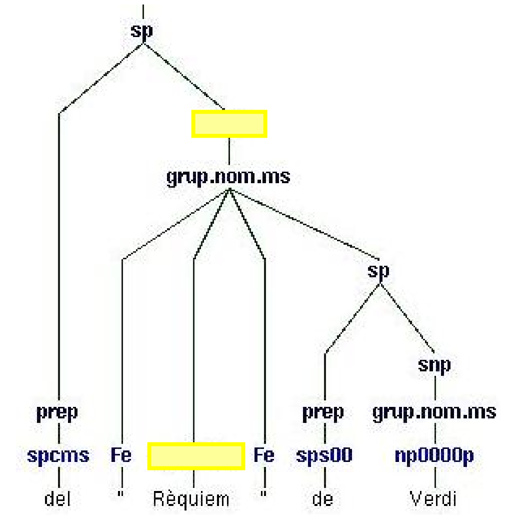

**Figure 25**: Others WNE, artistic opus (Verdi's "Requiem").

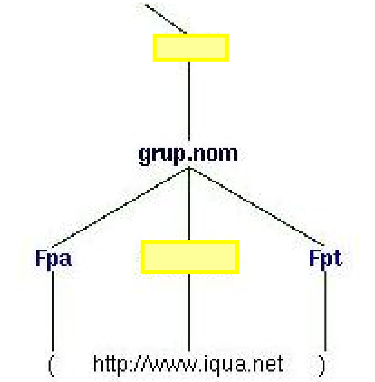

**Figure 26**: Others WNE, internet address

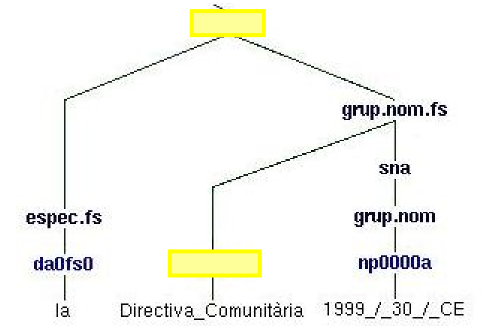

**Figure 27**:Others WNE, laws and regulations (the Communitarian Policy 1999 / 30 / CE).

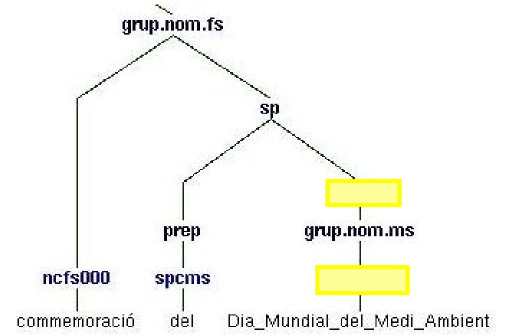

**Figure 28**: Others WNE, historical events (commemoration of the Global Day of Environment).

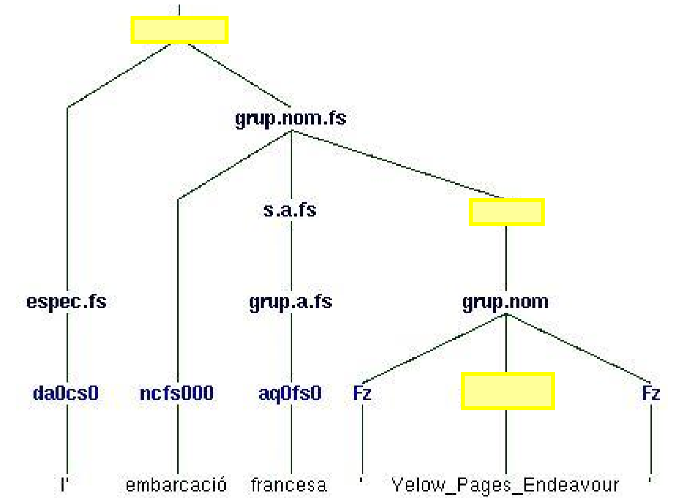

**Figure 29**: Others WNE, artefacts, inventions and machines (the French ship 'Yelow_Pages_Endeavour').

The annotation of NEs in two different levels (morphosyntactic: PoS, SNEs; syntactic: phrase level, WNEs) allows us to reflect changes in the semantics of NEs motivated by the apparition of a TW or by other contextual or pragmatic causes.

As we have already seen when talking about TWs, the semantic classification of the TW and any SNE occurring in a given phrase may or may not match up. If they do not match up, the semantic characterization of any WNE is determined by the core noun of the phrase, not by any determined SNEs which may occur in it. And, usually, core nouns of WNEs are TWs (Figure 30).

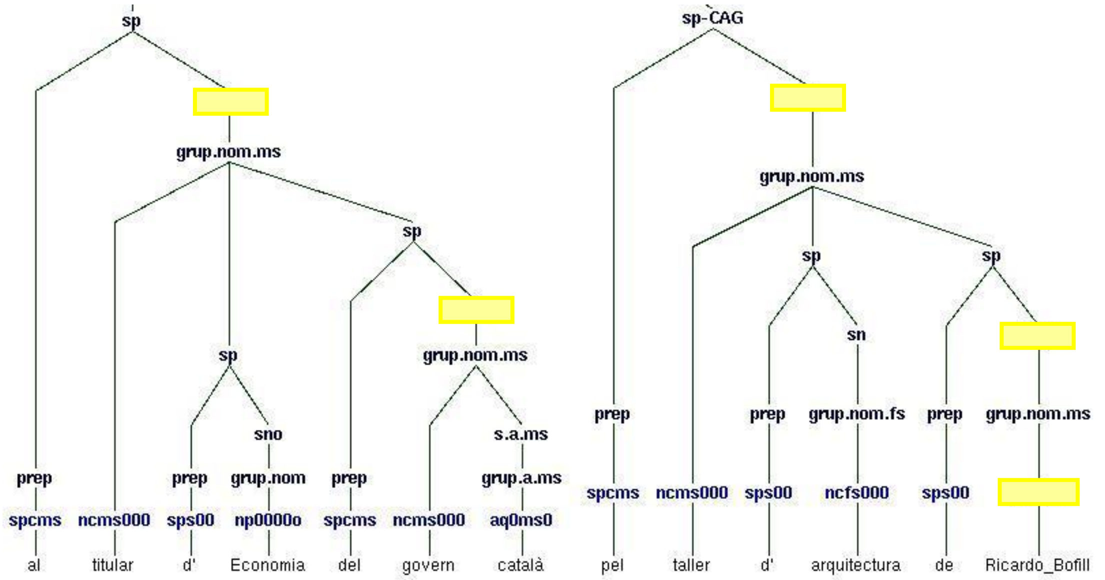

**Figure 30**: Two instances of semantic changes in WNE classification motivated by the apparition of TWs (the head of Economy of the Catalan government // by the Architecture Workshop of Ricardo_Bofill).

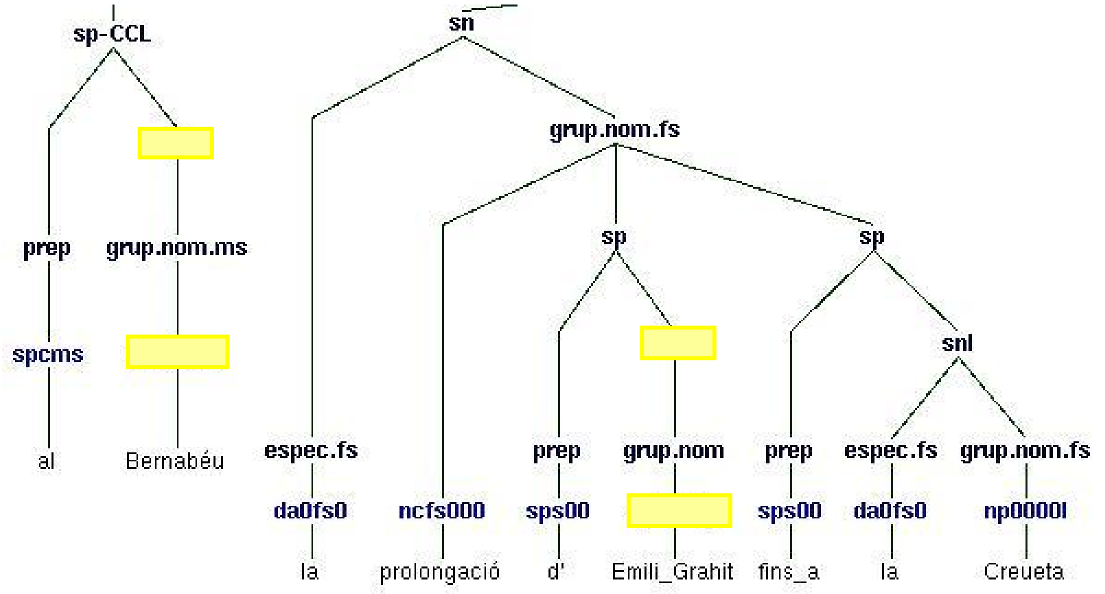

**Figure 31**: Two instances of semantic changes in WNE classification without the apparition of TWs. In both cases, a person SNE becomes a place WNE (in the Bernabéu –Real Madrid's Football Stadium // the prolongation of Emili_Grahit to the Creueta –both elements are streets).

Finally, it must be pointed that some NEs may shift their semantic classification depending on the context, without any change in their formal structures. That is so because we might use the same expression to refer to two different objects. This is very common in the case of State names, enterprise names and other organisation NEs, which may be used to talk about the physical space they occupy (their territory, their headquarters, etc.), or about the actions they undertake as organisations (Figure 32).

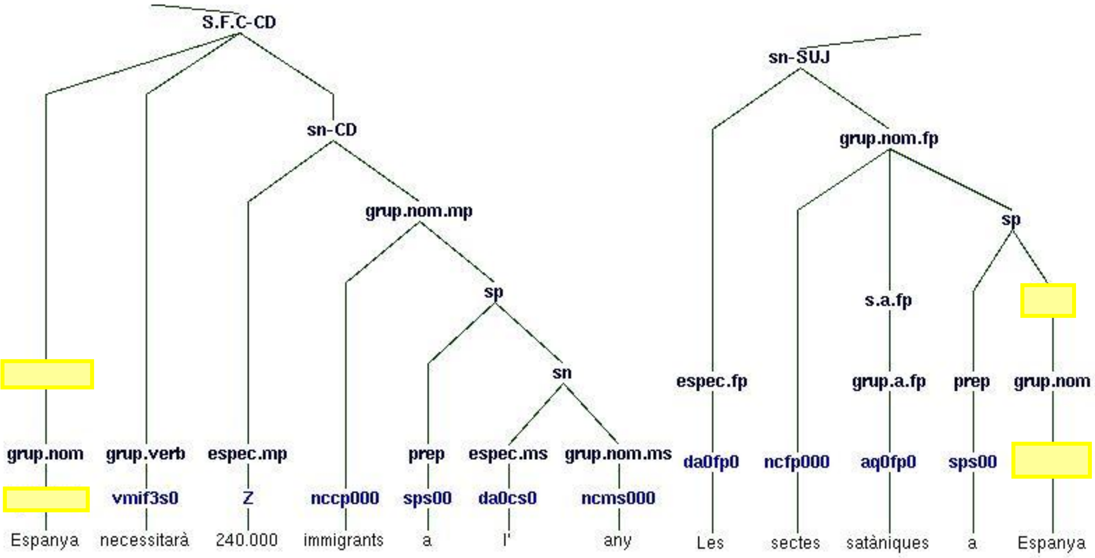

**Figure 32**: Two instances of the NE Spain working as an organization NE and as a place NE (Spain will need 240,000 immigrants each year //The satanic cults in Spain).

## 6. Conclusion

In this paper, we have presented the criteria followed in the manual annotation with Named Entities of a corpus consisting in 200.000 words for Catalan and 200.000 words for Spanish. The definition of this set of guidelines aims at tackling the problems posed by NERC systems from a linguistic point of view, rather than from a merely computational one.

We have decided to follow a division of NEs into Strong Named Entities and Weak Named Entities. We consider Strong NEs a morphological category, annotated at PoS level, and Weak NEs a syntactic category, annotated at phrase level. Our proposal of NE characterization and definition is based on formal features (morphological and syntactical) and semantic features, mostly relying on Trigger Words. We have settled criteria both to recognize and delimitate Entities, and to semantically classify them. The two different kinds of activities (recognition and classification) are interdependent, and based in two capital elements: previous PoS tagging and syntactic analysis, and Trigger Words.

Trigger Words have served us to detect NEs and to semantically classify them. We have considered TWs the central element of the NE in most occasions, for they are the core of the noun phrases we have annotated and, thus, they leak many of their features to the Entity involved. We have used a database of over 4.000 TWs in the process.

The main objective when annotating the corpus is to make it a valid and consistent tool for further research in NLP. Also, it has been used in the 9th task of SemEval (Multilevel Semantic Annotation of Catalan and Spanish, Màrquez et al., 2007). Shortly, we will annotate another 300.000 for Catalan and 300.000 more words for Spanish in order to reach the 500.000 words annotated corpus for each language.

Our aim is that the guidelines herein proposed be useful to researchers regardless of the language they may work with.

## Acknowledgements

We would like to thank Maria Jesús Díaz for her helpful advice and suggestions. This paper has been supported by FI-2006 grant from the Generalitat de Catalunya and the European Social Fund. It is based on work supported by the CESS-ECE (HUM2004-21127), Lang2World (TIN2006-15265-C06-06), and Praxem (HUM2006-27378-E) projects.

## Bibliography

Arévalo, M. (2001) Gramática para la detección y clasificación de las entidades con nombre (DEA Report). Barcelona: Departament of Linguistics, University of Barcelona.

Arévalo, M., X. Carreras, L. Márquez, M.A. Martí, L. Padró and M.J. Simón (2002) A Proposal for Wide-Coverage Spanish Named Entity Recognition, in Procesamiento del Lenguaje Natural, 28, pp. 63–80.

Arévalo, M., M. Civit and M.A. Martí (2004) MICE: A Module for Named Entites Recognition and Classification, in International Journal of Corpus Linguistics, vol.9, num. 1, pp. 53–68. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Black, W.J., F. Rinaldi and D. Mowatt (1998) FACILE: Description of the NE system used for MUC-7, in Proceedings of the Seventh Message Understanding Conference. Virginia: Fairfax.

Cinchor, N. (1997) MUC-7 Named Entity Task Definition, version 4.5. Available from http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html

Doddington, G., A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, R. Weischedel (2004) The Automatic Content Extraction (ACE) Program, Task, Data and Evaluation. Available from http://papers.ldc.upenn.edu/LREC2004/ACE.pdf

Fernández Leborans, M.J. (1999) El nombre propio, in Bosque, I. and V. Demonte (ed.) Gramática descriptiva de la lengua española, pp. 77–128. Madrid: Espasa.

Fernández Moreno, L. (2006) La referencia de los nombres propios. Madrid:Trotta

Màrquez, L., L. Villarejo, M. Taulé, M.A. Martí (2007) ‘Task 9: Multilevel Semantic Annotation of Catalan and Spanish’. *SemEval’07 Workshop ACL*, pp. 42–47. Praga: ACL.

Martí, M.A., M. Taulé, L. Màrquez, and M. Bertrán. (to appear) CESS-ECE: a multilingual and multilevel annotated corpus. Available online from http://www.lsi.upc.edu/~mbertran/cess-ece/publications.