Design Principles for a Spanish Treebank

Montserrat Civit Torruella
TALP Research Center – Universitat Politècnica de Catalunya
Jordi Girona 1-3 C6 D212
08034 Barcelona
civit@talp.upc.es
M. Antònia Martí Antonín
Centre de Llenguatge i Computació – Departament de Lingüística
Universitat de Barcelona
Gran Via de les Corts Catalanes 585
08007 Barcelona
amarti@fil.ub.es

July 16, 2002

Abstract

Treebanks are widely recognised as a necessary source of information in NLP as well as in Linguistics studies. In this paper we present and justify methodological principles and syntactic criteria to build a Treebank for Spanish: annotating only explicit information, constituents and syntactic functions and being theory independent. Previous work is also presented in order to account for taken decisions. The annotation process will be done in different steps so that each one of them is the input of the next. We present the basic guidelines of syntactic annotation and the boundaries of the work to be done in a first step: annotation of low constituents and surface functions. Moreover, some semantic information (subject type) is likely to be included.

1 The need

It is widely admitted that Treebanks constitute a crucial resource both to develop NLP applications and to acquire linguistic knowledge about how a language is used.

So far, there is not a freely available Treebank for Spanish, in spite of some references to a 1,000-sentence corpus ([25], [26], [27]). It is our aim to build up a 100,000-word Spanish Treebank which will be enriched in the future with semantic as well as pragmatic information, and which will be free for research purposes. Two groups are mainly involved in this project: CLiC, from the Universitat de Barcelona and TALP, from the Universitat Politècnica de Catalunya[1].

[1] This research has been partially funded by the Spanish Research Department: X-Tract PB98-1226; PETRA (TIC2000-1735-C02-02); PROFIT (FIT-150500-2002-244); MCT's grant (AP98 39864555) and by the Catalan Research Department (CIRIT's consolidated research group 1999SGR-150).

2 Methodological principles

In order to build up this Spanish Treebank, the annotation criteria of the most significant existing corpora of English ([22], [32], [36], [30]) and other languages ([2], [6], [7], [9], [10], [11], [12], [13], [14], [19], [20], [24], [28], [34], [35], [37]) have been consulted. Bearing in mind these works, a set of parameters has been defined, so as to establish the main theoretical and methodological principles for building a Treebank.

These principles are the following:

1. Implicit *versus* explicit: which elements should be annotated?

   As it will be explained later (see section 5), the annotation process will be done in different steps. In the current phase only explicit elements will be annotated. That means that elliptical words/constituents, movement, anaphoric or coreferential phenomena will not be taken into account.

2. Constituency *versus* dependency

   There is an open discussion about the annotation scheme to be assumed when building a treebank. On one hand, some papers claim that dependency annotation is more suitable if it is free-word-order language ([11], [14], [28]), while others make their choice on the basis of the application required [30]. Finally, in some cases, the annotation system follows the linguistic tradition [9].

   On the other hand, constituency is usually employed to annotate languages like English in which there is a fixed constituent order. Moreover, in this case, there is an almost exact matching between constituents and functions, that is, the position of a given constituent corresponds to one concrete syntactic function (for instance, in canonical declarative sentences, any noun phrase immediately preceding a verb is usually the subject).

   Spanish is a free constituent order language, although the word order cannot be altered within a constituent. For instance, there are three ways to say *John came this morning*

   > *Juan ha venido esta mañana*
   > *Esta mañana Juan ha venido*
   > *Esta mañana ha venido Juan*

   which are not exactly equivalent in their meaning. The focused element varies in each sentence, so, the pragmatic meaning is different. Furthermore, there are two noun phrases and both can precede the verb, so that it is impossible to know which the subject is (unless you have semantic information). At this stage, constituent annotation is convenient for Spanish as a previous step for the annotation of syntactic functions.

   Even though syntactic characteristics of Spanish are important, there are other reasons for our choice, such as the software available in our group. The idea is to take the maximum advantage of the tools we have developed so far: apart from a morphological analyser and a tagger, we have a chunker that parses sentences, as it is explained in section 4.

3. Annotating syntactic functions

   Constituents will receive two labels: the constituent tag (NP, PP, VP, etc.), and another one standing for syntactic functions (Subject, Object, Indirect Object, etc.). It should be mentioned here that only main functions will be annotated, that is, only verbal complements (including the subject) and adjunts are going to be labelled, while the labelling of noun complements is left over. Information provided by functions will be really useful to further develop a deep parser.

4. Maintaining surface word order

   According to the previous points, no word order alterations are going to be made in the annotation process. The strategy now is quite conservative. However, we are not disregarding this possibility in further developments of the treebank.

5. Being theory independent

   Linguistic theories give solutions for some specific problems but they lack coverage, that is, they work with a hypothetical model of language that does not face problems arising from corpora. Besides, theory deals with very specific (even rare) phenomena which hardly ever appear in corpus (see [31]).

   In the literature about treebanks, two positions about theory foundations arise: treebanks which are theoretically founded and treebanks that are theory independent. Among treebanks that are annotated according to one theory, two cases should be mentioned: treebanks annotated following the GB framework, like the PennTreeBank, and those annotated according to the HPSG theory. The PennTreeBank ([22] and [36]) is annotated with the principles of the X-bar theory, even though there is not a full application of all the theoretical issues. Some difficulties arise, for instance, with the need of distinguishing arguments and adjuncts, and with the PP-attachment, as stated in [36] and [21]. [20] and [34] follow the HPSG theory. The former justifies the choice on the premise that it facilitates the evaluation of an HPSG grammar; it provides a uniform way to represent different types of linguistic information; and, it is widely used in computational linguistics. The latter claims that HPSG allows to simultaneously represent constituents as well as dependency relations; that theory permits a consistent description of linguistic facts; that it enables translation to other formalisms; and, finally, that it can be used to support annotators' work. It is worth noticing that these treebanks are constitued by chosen sentences instead of by large texts more or less randomly selected. So, even if the number of sentences is high, they do not deal with what is largely understood by *real text*, that is, text reflecting any kind of linguistic phenomena.

   As for annotation systems which do not follow any theory, it should be said (as in [1]) that this option allows to adopt solutions equally profitable for linguists, computer scientists, psycholinguists etc. Following this proposal, we do not wish an application of one or another linguistic theory, but to fix a standard of constituency and functional annotation, neutral enough to be used for any research about Spanish and easy to translate into other formalisms.

   We think that the more neutral the annotation scheme, the more suitable for NLP purposes as well as for linguistic research. In fact, nowadays there is no theory about language use, and to build one, it seems necessary to know previous relevant facts about language use. Neutral, shallow annotations give 100 % coverage, even though they imply a loss in depth.

   Other reasons for taking this decision were that there is no previous experience in syntactic annotation in our group and that we do not have any deep parser. Simpler annotation seems a better starting point, because it is always possible to add new fine grained annotation levels over a first shallow one.

3 The CLiC–TALP Corpus

The CLiC-TALP corpus will consist on 1 million words. On one hand, it contains 500,000 words from LexEsp [33][2], and, on the other hand, 500,000 words from a Spanish newspaper (*La Vanguardia*), which covers a variety of authors and domains. The treebank will be built up from a set of randomly selected sentences (100,000 words) from the first part of the Corpus. Our aim is to syntactically annotate the whole CLiC-TALP corpus. The first set of 100,000 words will be used as a test-bed to evaluate the consistency and accuracy of the syntactic annotation outlined here.

4 Previous work

We have developed several resources for corpus analysis in Spanish: MACO, a morphological analyser; RELAX, a morphosyntactic tagger and TACAT, a syntactic chunker. These tools are organised in a pipeline-like process.

Figure 1 shows the pipeline process.

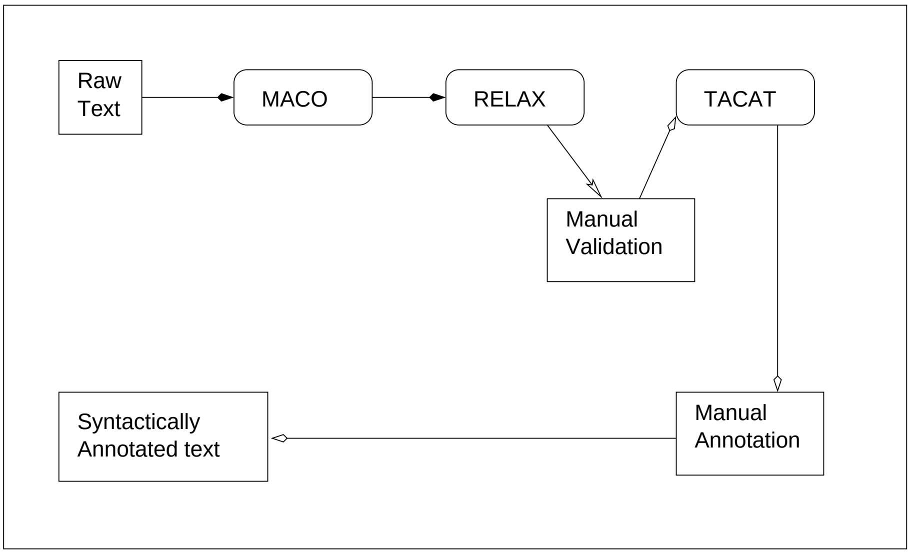

Figure 1: Pipeline

[2] LexEsp is a 6-million-word corpus of extracts from novels, scientific and weekly magazines, newspapers and sports papers ranging from 1978 to 1995. All texts were originally written in Spanish (both from Spain and South-America).

4.1 Morphological Analyser

MACO ([15]) is a Morphological Analyser for Spanish (and Catalan) which provides both lemma(s) and POS-tag(s) for each word and whose output has the following form:

$$word \quad lemma_1 - tag_1 \quad ... \quad lemma_n - tag_n$$

Tags codify 13 part-of-speech categories (noun, verb, adjective, adverb, pronoun, determiner, preposition, conjunction, interjection, dates, punctuation, numbers and abbreviations) as well as subcategories and morphological features ([17]), as it is proposed by Eagles [23]. The total amount of tags is 285[3]. The following table summarises the tagset.

Category	**Cat-Gloss**	**Subcategory**	**Features**
N	noun	proper, common	gender; number
V	verb	main, auxiliary, semiauxiliary	tense, mood, person, number, gender
A	adjective	qualifying, ordinal	gender; number
R	adverb	general, negative	–
P	pronoun	personal, demonstrative, possessive, numeral, indefinite, relative, interrogative	gender; number; person
D	determiner	article,demonstrative, possessive numeral, indefinite, relative, interrogative	gender; number; person
S	preposition	simple-complex	gender; number
C	conjunction	subordinating, coordinating	–
I	interjection	–	–
W	date	–	–
F	ponctuation	–	–
Z	number	–	–
Y	abbreviation	–	–

The morphological analysis of the sentence
Pese a que es rentable publicitariamente, el ciclismo español tiene un gran problema.
Although is profitable commercially, the Spanish cycling has a big problem.
'Although commercially profitable, Spanish cycling is in big trouble.'
is as follows:

Pese_a_que pese_a_que CS
es e NCFP000 ser VSIP3S0
rentable rentable AQ0CS0
publicitariamente publicitariamente RG
, , Fc
el el DA0MS0
ciclismo ciclismo NCMS000
español español AQ0MS0 español NCMS000
tiene tener VMIP3S0
un uno DI0MS00 uno DN0MS0 uno PN0MS000
gran gran AQ0CS0

[3] An extented explanation of the tagset can be found in [17], which is the guideline for the annotators of the corpus, and in [18] (both in Spanish).

problema problema NCMS000
. . Fp

As it can be seen, three of the words in the previous sentece are ambiguous (*es*, *español*, *un*), that is, they receive more than one $lemma - tag$ pair. There are different kinds of ambiguity. (a) For POS: *es* can be a noun or a verb and *español* an adjective or a noun. (b) For subcategory: *un* can be a numeral or an indefinite determiner. (c) For inflection (there is no example here, but, for instance, *cólera* is a noun which can be masculine (then, it means 'cholera') or femenine (meaning 'anger, rage')).

The amount of ambiguity in the analyser is shown in table 1.

# of words	# of tags
728,892	1
163,597	2
16,939	3
2,118	4
225	5
149	6
16	7
2	8
# of word forms	**# of interpretations**
911,938	1,117,522

Table 1: Ambiguity in the Analyser

4.2 Morphological Tagger

As for morphological desambiguation, RELAX [29] is a constraint-based probabilistic tagger which allows the introduction of manually written constraints. The accuracy of the output varies between 94-96%.

Once the tagger has been applied, the resulting analysis is:

Pese_a_que pese_a_que CS
es ser VSIP3S0
rentable rentable AQ0CS0
publicitariamente publicitariamente RG
, , Fc
el el DA0MS0
ciclismo ciclismo NCMS000
español español AQ0MS0
tiene tener VMIP3S0
un uno DI0MS0
gran gran AQ0CS0
problema problema NCMS000
. . Fp

4.3 Manual validation of the CLiC-TALP corpus

After the application of automatic analysis and tagging, a manual validation phase started. The task consists on verifying and correcting (if necessary) the output of the tagger. Three validation tasks were done at the same time: compounds, lemmas and tags. The main sources of errors were:

1. Compounds. Sequences which could be misinterpreted as compounds or non-compounds. For instance, *esto es* is a compound in exemple-1 but a non-compound in exemple-2:
   ex-1:
   probablemente zamba, esto es, mestiza de negra e india
   'probably *zamba*, that is, mixed race of black and Indian'

   ex-2:
   Esto es lo que se llama aprendizaje hebbiano
   'That is what is called hebbian learning'

   This is not the case with *pese_a_que* in the sentence, because it is an unambiguous compound which appears in the database of the morphological analyser as a unique entity.

   It should be mentioned that compounds are annotated with the same tagset as the non-compounds, and that we do not consider any tag for each of the parts[4].

2. Inter-categorial ambiguity. It mainly concerns differences between determiners and pronouns, and between nouns and adjectives, because these are the most frequent ambiguity classes.

3. Intra-categorial ambiguity. Since the tagger only works with the first two digits (category and subcategory) of the morphological tag, it did not always solve gender, number, or person ambiguity. Some of these cases have been solved by introducing some hand-written rules and results improved significantly.

4. Particularly ambiguous words. Two of them can be mentioned: *que*, *se*. *Que* may be tagged as conjunction or as relative pronoun; sometimes both tags are possible in the same context although never simultaneously. As for *se*, it is one of the most difficult words to disambiguate in Spanish. It has three possible tags: reflexive pronoun, mark of pronominal verb and mark of impersonal or passive sentence.

At this stage, as the goal is to provide a useful linguistic annotation, some more fine-grained details were introduced that do not appear in the morphological analyser. For instance, it is possible, and admitted, in Spanish, to use the dative pronominal form *le* instead of the accusative one *lo* if the referent is a masculine person, so we introduced the accusative case for *le*. It is also the case of adverbialised adjectives. It is quite usual in Spanish to use adjectives (in their masculine singular form and without inflection) to express manner (for instance, *hablar alto, hablar claro*[5]). The adverbial tag for these forms has not been introduced in the analyser in order to avoid ambiguity. These are the only two cases in which MACO, RELAX, on one hand, and annotators, on the other, work with a different tagset.

Up to now, 100,000 words have been manually validated. The following process will be to re-train the tagger in order to improve its output, and continue the validation process. Because of a lack of funding, there was only one part-time person working on it. This project has been recently funded

[4] Contrary to the annotation system of the French Treebank [3].

[5] Literally *speak loud, speak clear*; 'speak loudly, speak clearly'.

and we hope to accelerate the validation process of the rest of the CLiC-TALP corpus. Each text was validated once and there was a coordination task including writing the documentation, updating the analyser, post-checking the annotator's work and weekly meeting with her to discuss the details of the annotation.

4.4 Chunker

TACAT [8] is a chart parser which works left-right and bottom-up. It produces, with the help of a context-free grammar, the chunking of the text, taking as input the output of the tagger.

The grammar is hand-written. It contains about 1,500 rules. The following examples correspond to noun-phrase rules (*sn*)[6]:

```
sn ==> espec-ms, grup-nom-ms.
sn ==> espec-mp, grup-nom-mp.
sn ==> espec-fs, grup-nom-fs.
sn ==> espec-fp, grup-nom-fp.
```

As it is said in [5] *chunks [...] can be recovered quite reliably [...]. Resolving attachments generally requieres information about lexical association between heads.* [4] defines *chunk* in terms of 'major heads'. A 'major head' *is any content word that does not appear between a function word* f *and the content word* f *selects, OR a pronoun selected by a preposition.* As example, *proud* is a 'major head' in *a man proud of his son* but not in *a proud man.*

The idea of chunks adopted in our annotation system is a bit different from Abney's, because of the syntactic characteristics of Spanish. It is noteworthy that adjectives usually appear after a noun, and not before. To take Abney's proposals literally we should consider *orgulloso* ('proud') 'major head' in *un hombre orgulloso de su hijo*[7] but not in *un hombre orgulloso*[8]. As it is well known, PP-attachment is one on the main issues in NLP, so what we consider is that adjectives are always attached to nouns (i.e. they belong to the nominal chunk) and PPs form a separate chunk except in *de*-PP (see below).

In this case, the chunker produces this segmentation:

[un hombre orgulloso]$_{NP}$ [de su hijo]$_{PP}$

Noun-phrases do not contain prepositional ones, except for one case: when the PP is introduced by the preposition *de* ('of') immediately following the noun. To take this decision an experiment has been done consisting on retrieving $[noun +' de']$ sequences from corpus. The starting hypothesis was that *de*-PPs were attached to the previous noun. Out of 210 sentences with 237 examples, 230 proved the hypothesis; 3 cases had ambiguous attachments and 4 were modifiers of other elements. So *de*-PPs were included in NP if, and only if, preposition *de* was next to the noun. The rule is:

```
grup-nom-ms ==> n-ms, sp-de.
```

A similar experiment was done with other prepositions and with $[noun + adjective +' de']$ sequencies. In both cases, results were not so satisfactory: about 50 % of the cases were wrongly analysed and no rules were introduced, therefore there was no attachment.

[6] *espec* stands for any determiner(s) preceeding the noun; *grup-nom* stands for 'noun group', that is, nouns or nouns plus adjectives. Both elements are rewritten in other rules in order to consider all the possibilities. Elements after the hyphen guarantee the agreement between the elements.

[7] 'A man proud of his son'.

[8] A man proud; 'A proud man'.

Recognised chunks are: NP, PP, AjP and AvP. As for VP, the grammar recognises verbal groups, that is, simple and complex verbal forms, like *es* ('is'), *ha sido* ('has been'), *debería haber sido* ('should have been'), *tiene que ser* ('has to be') in all their forms. Clitics and other particles (such as negative adverbs) are not included in the verbal-group.

Other elements recognised by the grammar, such as relative pronouns, subordinating conjunctions, clitics, etc. are left as *orphan nodes* in the tree.

As for coordination, this grammar only deals with the simplest case: a coordinated chunk is made if, and only if, coordination happens between two (or more) single lexical items. For instance: a coordinated nominal chunk is built for *una lección de* **poderío y clase** ('one lesson of might and class') but it is not for *la debilidad sentimental, la resignación y el miedo* ('the sentimental weakness, the resignation and the fear'[9]) because of the articles.

Finally, it should be mentioned that there is no analysis of clauses. It is possible to know where they start (because the complementiser is mandatory in Spanish) but it is impossible to know where they finish, so they will be built manually.

The result of the chunker is shown in figure 2[10].

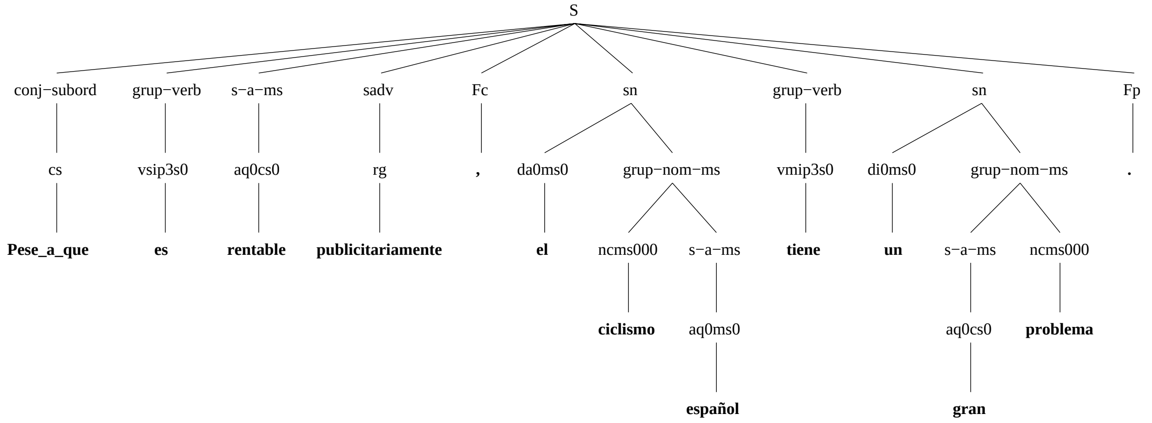

Figure 2: Chunker output

The main idea was to produce 'correct', even though incomplete, analysis because we rely on the correctness of the parser output to start the manual syntactic validation.

5 Syntactic Annotation

Taking chunks as starting point, what needs to be done in this phase is to build complex phrases and assign syntactic functions to them. The syntactic annotation process will be accomplished in two steps: first, constituents will be bracketed and labelled. In a second step, syntactic functions will be added to the constituents.

Methodologically speaking, we distinguish at least three phases in each step. Firstly, the basic principles of the syntactic annotation (for constituents as well as for functions) will be set. These principles

[9] Articles are more widely used in Spanish than in English.

[10] Labels have the following meanings: *conj-subord* means subordinating conjunction; *grup-verb* verbal group; *s-a-ms* is a masculine singular adjectival phrase; *sadv* stands for adverbial phrase; *sn* means noun phrase. Word-forms are marked in bold and their mother nodes are the morphological tags.

are related to syntactic phenomena to be dealt with (see below). Secondly, several people will annotate 200 sentences each, in order to find out problems and to determine the basic principles. Weekly meetings will take place to discuss these problems and reach solutions and. These discussions are believed to help the setting-up of the guidelines as well as to improve the first annotation, which could change if necessary. Thirdly, the rest of the corpus will be annotated while the first and second phases might need to be reconsidered if further problematic cases are encountered.

The basic principles of syntactic annotation are:

1. considering separately the verb and its complements.

   It is traditionally said that any sentence has two main constituents: subject and predicate, the second one including the verb, its arguments and its adjuncts. As it is well-known, the relationship between verb and arguments is closer than that between verb and adjuncts. Since Spanish is a free constituent order language, establishing a predicate node could mean having to alter the surface order of the elements in the sentence. Hence, it has been decided not to deal with a predicate constituent. Instead, each phrase of the sentence will constitute a node of the tree. The exception to this rule are both finite and non-finite clauses. Clauses will form a unique node, like sentences. Within the clauses, any phrase will be represented as a node.

2. not changing the superficial order of the elements in the sentence.

   Related to the previous point, no change on the constituent order will be done. The problem appears also with discontinuous elements. No decision has been taken yet about how to deal with comparative structures in which the two elements usually occur separately[11] or with movement phenomena in interrogative sentences, especially when a *wh*-word comes from an embeded clause like *qué te gustaría ser* (what would you like to be). If there is no change in the constituent order, some convention has to be adopted to mark that *qué* is the attribute of *ser* and not of *gustaría*.

3. annotating surface functions of the sentence.

   We first decided to establish a hierarchy of functions, in a similar way to [16] for dependency relations. This would have allowed us to underspecify them if any function was uncertain. However, reviewing some examples and also taking into account what is said in [36] and [21], it seemed that the distinction between arguments and adjuncts is lexical rather than syntactical. For instance, locative elements might be considered adjunts if they appear with verbs like *decir (speak), querer (want, love)* but arguments with verbs of movement like *ir (go)* o *venir (come)*. The only way to know that is having a lexicon containing exhaustive information about subcategorisation. So, it has been decided to annotate only the surface function (subject, object, indirect object, predicative, locative, etc.). One of the goals of the treebank will be to retrieve information about subcategorisation and argumental structure of Spanish verbs[12] in order to build up a lexicon that would contain such information.

4. not considering the possibility of double functions.

   Let us consider these two sentences:
   (a) *¿Qué te gustaría ser?*
   What you would like to be?
   'What would you like to be?'

[11] Like in English: **more** *understandable* **than**.

[12] So far, a lexicon of 2,000 verbal senses has been developed but the information it contains is not enough to deal with real text.

(b) *La señora Aguirre quiere castigar la suciedad de Madrid*
The Mrs Aguirre wants punish the dirt of Madrid
'Mrs Aguirre wants to punish the dirt of Madrid'

In (a), pronoun *te* ('you') is the indirect object of *gustaría* (´would like´) and the semantic subject of *ser* (´to be´). In (b), *La señora Aguirre* is the subject of *quiere* and *castigar*. *Gustar* and *querer* are verbs of control: the first is an object control verb and the second a subject control one. Following the principle *the simpler the better*, only surface functions will be annotated now (i.e. *te* will be tagged as an indirect object in (a) *La señora Aguirre* as a subject of *quiere* in (b)).

5. dealing with non-finite verbal phrases.

   This point is related to the previous one. Non-finite verbal phrases will not receive a *PRO* subject.

6. distinguishing different kinds of (semantic) subjects.

   As it has been said before, we will not deal with empty categories(see section 2). Although Spanish is a pro-drop language and subjects do not usually appear explicitly in sentences, we plan to include some semantic tags refering to semantic roles (like *agent, patient, cause, experiencer*) for explicit subjects. This point is now under discussion.

7. dealing with ambiguous attachments.

   A human annotator can usually take decisions about attachments, because context helps him/her to solve the ambiguity. However, there are some remaining ambiguities for which a decision is needed. Let us consider sentence (c):
   (c) *la facultad de aprender y reaccionar ante nuevas situaciones*
   the faculty of learn and react in front of new situations
   'the capacity to learn and react when facing new situations'

   In this sentence, *ante nuevas situaciones* is a PP which may depend on *reaccionar* or on both infinitives *aprender y reaccionar* and there is no way to know the appropriate attachment because context does not give us enough information. Defining a default attachment seems to be a better solution, such as the closer (*reaccionar*) or the higher (coordination) node. However, even if there is no error when generalizing in this case, it is possible to make a mistake in other cases. This will happen in the well-known case of *I saw a man with a telescope*. Thus, another possibility is to choose the higher (or the nearer) attachment but mark (in some way) that there is another option, and that there is an ambiguity.

6 Samples of syntactic annotation

Although software annotation has not been developed yet, we present here a first approach about the kind on work to be done and how syntactic tagging will look like.

Sample 1:

Indignos de la civilización que les cobija
Undeserving of the civilisation that them shelters
'Undeserving of the civilisation that shelters them'

This is an adjectival sentence where:

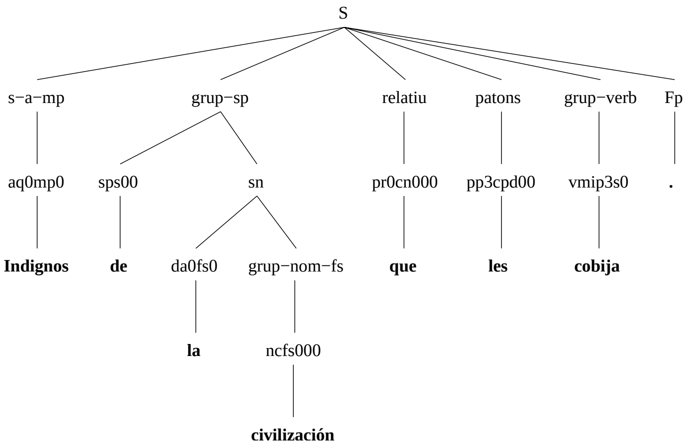

Figure 3: Sample 1 (chunker output)

1. the relative clause (*que les cobija*) needs to be built. This implies the addition of a node (subordinate-clause-REL). Within the clause, syntactic function labels need to be added, too: SUBJ (subject) for the relative pronoun (*que*), and OD (direct object) for the clitic (*le*).
2. the clause needs to be moved down and to become a daughter node of the GRUP-NOM-FS.
3. the PP (GRUP-SP) (*de la civilización que les cobija*) is a complement of the adjective (*indignos*), so it has to be moved to the S-A-MP-daugther position.

The result is shown in figure 4:

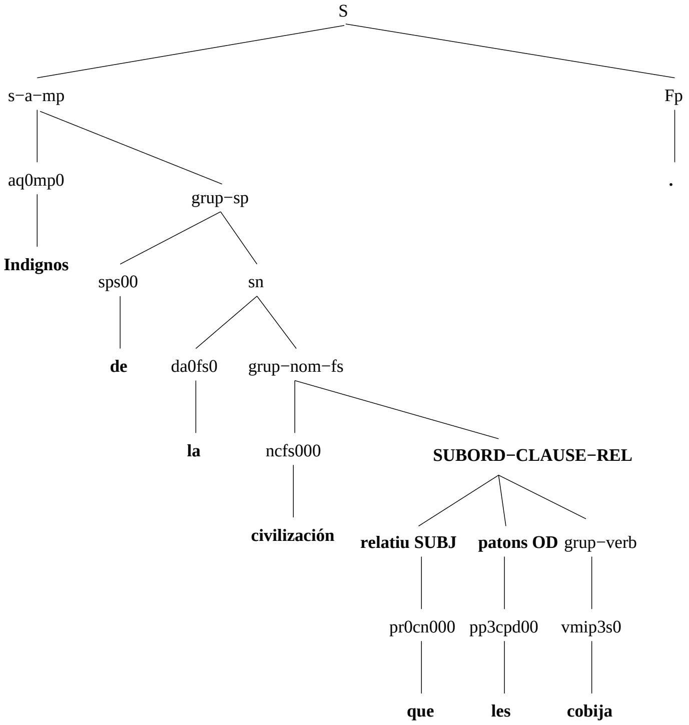

Figure 4: Sample 1 (manual annotation)

Sample 2:

La Iglesia habla del problema del Mal en el mundo
the church talks of-the problem of-the Evil in the world
’The Church talks about the problem of Evil in the world’

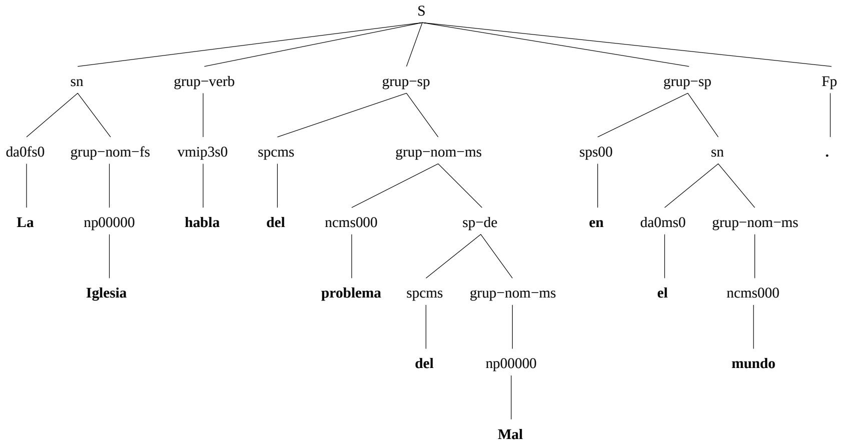

Figure 5: Sample 2 (chunker output)

Modifications to be applied to the chunker output (shown in figure 5) are the following:

1. the last PP (*en el mundo*) needs to be moved to the sister-position of the noun *problema*.
2. syntactic functions should be added to the NP (SN) (*la Iglesia*), which is the subject, and to the PP (GRUP-SP) (*del problema del Mal en el mundo*), which is a prepositional object (CRV).

Figure 6 shows the result.

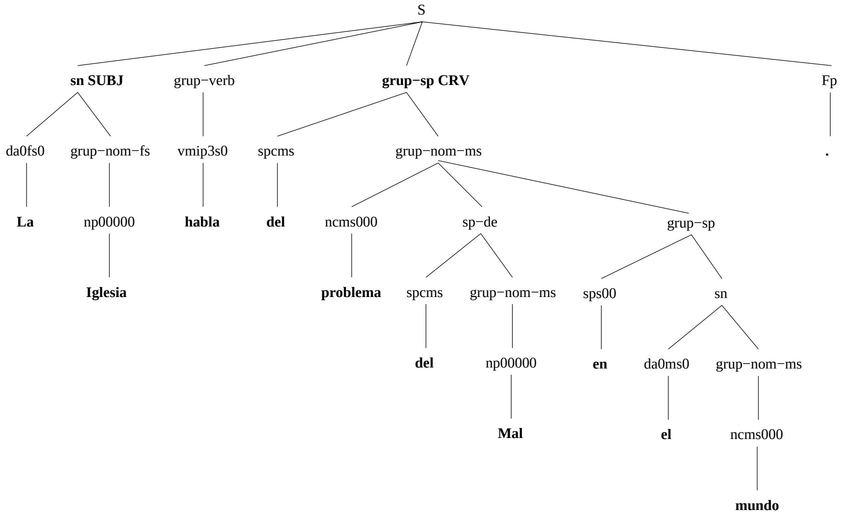

Figure 6: Sample 2 (manual annotation)

Sample 3:

Pese a que es rentable publicitariamente, el ciclismo español tiene un gran problema.
Although is profitable commercially, the Spanish cycling has a big problem.
‘Although commercially profitable, Spanish cycling is in big trouble.’

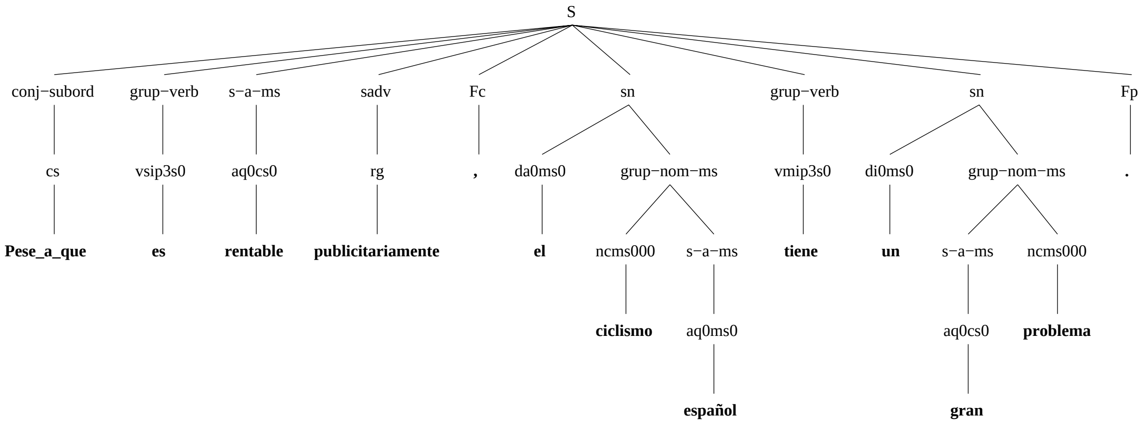

Figure 7: Sample 3 (chunker output)

In this last example, modifications are:

1. adding an extra-node for the subordinate clause (subordinate-clause-ADV) (*Pese_a_que es rentable publicitariamente*). Within the clause, *publicitariamente* needs to be moved down, and the AP (S-A-MS) node needs to be labelled with its function tag ’attribute’ (ATR).
2. adding functional tags to the main nodes: ADJ-CONC (concessive adjunct) to the clause; SUBJ (subject) to *el ciclismo español* and OD (direct object) to *un gran problema*.

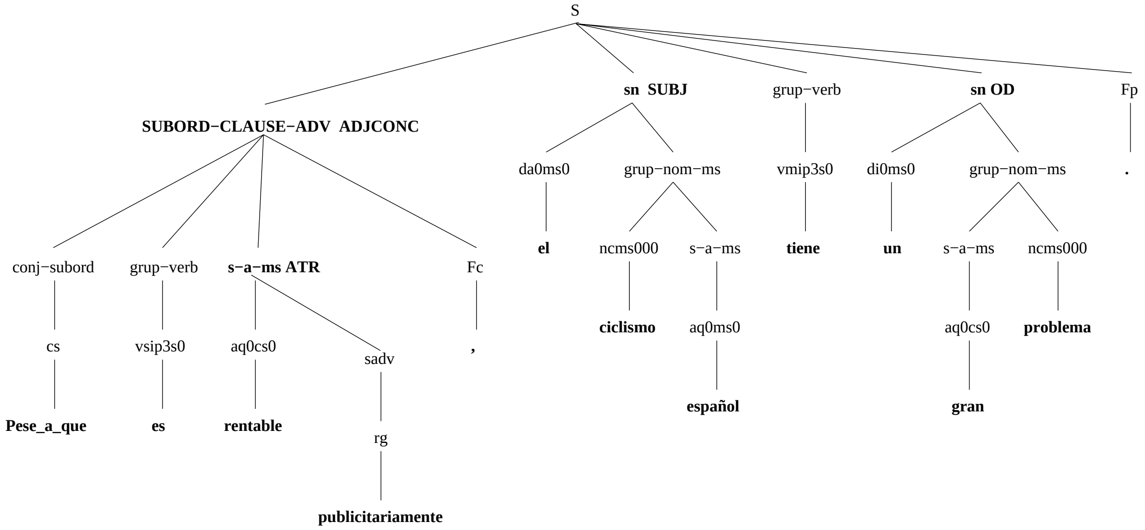

Figure 8: Sample 3 (manual annotation)

As it can be seen, none of the movements alters the surface word order in the sentences.

7 Conclusions

In this paper we have presented the main criteria to build a treebank of Spanish. Basic methodological principles as well as general syntactic annotation criteria have been presented. The principles and criteria adopted have been stated taking into account both syntactic characteristics of Spanish and works already done in this field. We have adopted a step-by-step methodology in order to prioritise coverage instead of deep analysis. Since the project has been recently funded, the annotation process has started at the beginning of July, and some smpes have been given.

References

[1] A. Abeillé, F. Toussenel, and M. Chéradame. Corpus le monde. annotations en constituants. guide pour les correcteurs. Technical report, LLF, UFRL, 2001. dernière mise à jour: 12-oct-2001.

[2] A. Abeillé, L. Clément, and A. Kinyon. Building a treebank for French. In *Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, pages 87–94, Athens, Greece, 2000.

[3] A. Abeillé, L. Clément, and A. Kinyon. *Building and Using syntactically annotated corpora*, chapter Building a treebank for French. Language and Speech. Kluwer, Dordrecht, 2001.

[4] S. Abney. Parsing by Chunks. In R. Berwick, S. Abney, and C. Tenny, editors, *Principle-Based Parsing*. Kluwer Academic, 1991. available: http://www.sfs.nphil.uni-tuebingen.de/˜ abney/.

[5] S. Abney. Part-of-Speech Tagging and Partial Parsing. In K. Church, S. Young, and G. Bloothooft, editors, *Corpus-Based Methods in Language and Speech*. Kluwer Academic, 1996. available: http://citeseer.nj.nec.com/abney96partspeech.html.

[6] I. Aduriz, I. Aldezabal, M. Aranzabe, B. Arrieta, J.M. Arriola, A. Atutxa, A. Diaz de Ilarraza, K. Gojenola, M. Oronoz, and K. Sarasola. Construcción de un corpus etiquetado sintácticamente para el euskara. In *XVIII Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural*, 2002. Forthcoming.

[7] S. Afonso, E. Bick, R. Haber, and D. Santos. 'Floresta Sintá(c)tica': a Treebank for Portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*, pages 1698–1703, Las Palmas de Gran Canaria, Spain, May 2002.

[8] J. Atserias and H. Rodríguez. Tacat: Tagged corpus text analyzer. Technical report, Software Department (LSI). Technical University of Catalonia (UPC), 1998.

[9] A. Bemova, J. Hajic, B. Hladka, and J. Panevova. Morphological and Syntactic Tagging of The Prague Dependency Treebank. Journées Atala, Corpus annotés pour la syntaxe, Paris, June 1999. available: http://talana.linguist.jussieu.fr/treebanks99/.

[10] A. Böhmova and E. Hajicova. How Much of the Underlying Syntactic Structure can be Tagged Automatically. Journées Atala, Corpus annotés pour la syntaxe, Paris, June 1999. available: http://talana.linguist.jussieu.fr/treebanks99/.

[11] I. Boguslavsky, I. Chardin, S. Grigorieva, N. Grigoriev, L. Iomdin, L. Kreidlin, and N. Frid. Development of a Dependency Treebank for Russian and its possible Applications in NLP. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*, pages 852–856, Las Palmas de Gran Canaria, Spain, May 2002.

[12] C. Bosco, V. Lombardo, D. Vassallo, and L. Lesmo. Building a treebank for Italian: a Data-driven Annotation Schema. In *Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, pages 99–105, Athens, Greece, 2000.

[13] T. Brants, W. Skut, and H. Uszkoreit. Syntactic Annotation of a German Newspaper Corpus. Journées Atala, Corpus annotés pour la syntaxe, Paris, June 1999. available: http://talana.linguist.jussieu.fr/treebanks99/.

[14] T. Brants, W. Skut, and H. Uszkoreit. *Building and Using syntactically annotated corpora*, chapter Syntactic Annotation of a German Newspaper Corpus. Language and Speech. Kluwer, Dordrecht, 2001. available: http://treebank.linguist.jussieu.fr/toc.html.

[15] J. Carmona, S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of the First Conference on Language Resources and Avaluation. LREC'98*, pages 915–922, Granada, 1998.

[16] J. Carroll, T. Briscoe, and A. Sanfilippo. Parser Evaluation: a Survey and a New proposal. In *Proceedings of the First Conference on Language Resources and Avaluation. LREC'98*, pages 447–454, Granada, 1998.

[17] M. Civit. Guía para la anotación morfológica de corpus. Technical Report X-Tract WP-00/06, Universitat de Barcelona, 2000. available: http://www.lsi.upc.es/˜ civit/publicacions.html.

[18] M. Civit, I. Castellón, and M.A. Martí. Joven periodista triste busca casa frente al mar o la ambigüedad en la anotación de corpus. Congreso Internacional sobre nuevas tendencias de la lingüística, Granada, November 2001. available: http://www.lsi.upc.es/˜ civit/publicacions.html.

[19] J. Hajic. Building a Syntactically Annotated Corpus: the Prague dependency Treebank. *Issues of Valency and meaning*, pages 106–132, 1998.

[20] M. Marciniak, A. Mykowiecka, A. Przepiórkowski, and A. Kupsc. *Building and Using syntactically annotated corpora*, chapter Construction of an HPSG Treebank for Polish. Language and Speech. Kluwer, Dordrecht, 2001. available: http://treebank.linguist.jussieu.fr/toc.html.

[21] M. Marcus, G. Kim, M.A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, New Jersey, March 1994.

[22] M. Marcus, B. Santorini, and M.A. Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 1993. available: http://www.cis.upenn.edu/˜ treebank/home.html.

[23] M. Monachini and N. Calzolari. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. a common proposal and applications to european languages. Technical report, EAGLES, 1996. available: http://www.ilc.pi.cnr.it/EAGLES96/browse.html.

[24] S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M.T. Pazienza, D.Saracino, F. Zanzotto, N. Mana, F. Pianesi, and R. Delmonte. *Building and Using syntactically annotated corpora*, chapter Building the Italian Syntactic-Semantic Treebank. Language and Speech. Kluwer, Dordrecht, 2001. available: http://treebank.linguist.jussieu.fr/toc.html.

[25] A. Moreno, R. Grishman, S. López, F. Sánchez, and S. Sekine. A Treebank of Spanish and its Application to Parsing. In *Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, pages 107–111, Athens, Greece, 2000.

[26] A. Moreno and S. López. Developing a Spanish TreeBank. Journées Atala, Corpus annotés pour la syntaxe, Paris, June 1999. available: http://talana.linguist.jussieu.fr/treebanks99/.

[27] A. Moreno, S. López, F. Sánchez, and R. Grishman. *Building and Using syntactically annotated corpora*, chapter Developing a Spanish Treebank. Language and Speech. Kluwer, Dordrecht, 2001. available: http://treebank.linguist.jussieu.fr/toc.html.

[28] K. Oflazer, B. Say, D.Z. Hakkani-Tür, and G. Tür. *Building and Using syntactically annotated corpora*, chapter Building a Turkish Treebank. Language and Speech. Kluwer, Dordrecht, 2001. available: http://treebank.linguist.jussieu.fr/toc.html.

[29] Lluís Padró. *A Hybrib Environment for Syntax-Semantic Tagging*. PhD thesis, Software Department (LSI). Technical University of Catalonia (UPC), 1997.

[30] O. Rambow, C. Crecwell, R. Szekely, H. Taber, and M. Walker. A Dependency Treebank for English. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*, pages 857–863, Las Palmas de Gran Canaria, Spain, May 2002.

[31] G. Sampson. Probabilistic models of analysis. In R. Garside, G. Leech, and G. Sampson, editors, *The Computational Analysis of English*, chapter 1, pages 16–29. Longman, 1987.

[32] G. Sampson. *English for the Computer. The SUSANNE corpus and Analytic Scheme*. Clarendon Press, Oxford, 1995.

[33] N. Sebastián, M.A. Martí, M.F. Carreiras, and F. Cuetos. *LEXESP: Léxico Informatizado del Español*. Edicions de la Universitat de Barcelona, 2000.

[34] K. Simov, P. Osenova, M. Slavcheva, S. Kolkovska, E. Balabanova, D. Doikoff, K. Ivanova, A. Simov, and M. Kouylekov. Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*, pages 1729–1736, Las Palmas de Gran Canaria, Spain, May 2002.

[35] M. Tadic. Building the Croatian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*, pages 441–446, Las Palmas de Gran Canaria, Spain, May 2002.

[36] A. Taylor, M. Marcus, and B. Santorini. *Building and Using syntactically annotated corpora*, chapter The Penn Treebank: an overview. Language and Speech. Kluwer, Dordrecht, 2001. available: http://treebank.linguist.jussieu.fr/toc.html.

[37] T. Váradi. The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*, pages 385–396, Las Palmas de Gran Canaria, Spain, May 2002.