Evaluation of the XIP Spanish Grammar

(December 2006, E. López García)

1. Introduction

The purpose of this document is to show and comment the results of the evaluation of the XIP Spanish grammar after having analyzed 400 sentences (12433 words) with the general Spanish grammar.

The texts of the analyzed corpus are newspaper articles from the Spanish newspaper 'El Mundo' and coming from different resorts. They were taken from the on-line edition around mid-August.

Precision and *Recall* were calculated for the following dependencies:

- **SUBJ** (subject of the verb, binary relation),
- **OBJ** (direct object, binary relation),
- **MOD** (noun modifier, binary relation).

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database (expressed as a percentage).

Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved (expressed as a percentage).

2. Evaluation

	Subject	***Direct Object***	***Noun Modifier***
Precision	0.78	0.77	0.45
Recall	0.67	0.84	0.86

3. Analysis

3.1. Subject

We have identified six major causes for failure (by failure we mean subjects that have not been located):

- The first and most important one is the difficulty of discriminating between the **relative pronoun** "*que*" in a relative sentence, since it can either be the subject or the direct object. In Spanish, there are very few contexts where you can really distinguish between both of these syntactic relationships for the pronoun. If we take into account the total number of errors in subject extraction, we get to the percentage of **36%**.

- The second big problem with an average of **17%** is **postponed subjects**. In Spanish, the subject can be either before or after the verb (and even not appear at all). In most cases, if it is present, it appears before the verb. Therefore, when it is after the verb, it is hard to find a context in which we can make a clear difference between the subject and the object.

- The third problem that we encountered has to do with **coordinated subjects**. It is very difficult to locate coordinated items because many times the distance between them is too big and the contexts are too varied to be able to make any rules. Thus, when the subject of a verb consists of two or more coordinated elements we have the typical problems that we find when extracting coordination. The average of this kind of error is **12%**.

- Other major problem was the fact that there is not a module that recognizes **named entities** and therefore, proper names, dates, company or institution names are not well disambiguated and not well analyzed. The fact that proper names (like newspapers) sometimes appear surrounded by quotes also complicates the task. This problem represents **8%**.

- **8%** of the errors were due to a **bad disambiguation**.

- An incorrect recognition of **appositions** provoked **4%** of the errors. Usually, a noun that has apposition behind it is the real subject of a verb. But when a noun phrase that constitutes the apposition is not recognized as it, it will be taken as the subject of the verb.

- The remaining **15%** corresponds to other miscellaneous problems.

3.2. Direct Object:

Firstly, we have to make a precision, the direct objects that we were looking for, were only nominal objects and not sentences. For example, in a sentence like “*Le dije que no viniera*” (I told him not to come), we won’t extract the dependency between “*dije*” and “*viniera*”. Although the whole subordinated sentence is the object of the main verb, the Spanish grammar is not developed enough to take this kind of object into account. Therefore, we leave this kind of objects out of our evaluation.

- The biggest difficulty in finding the direct objects was to distinguish direct objects from indirect objects. In Spanish, the mark of an indirect object is the preposition "*a*", but this mark is shared with direct **objects that are animated**. Therefore, to create rules, we would have needed semantic information such us "animated", which we did not have. Thus, **27%** of errors were due to this reason; if NPs had contained this feature, our grammar would have been able to recognize them.

- The second big problem is about the **relative pronoun** "*que*"; the same problem that we see for subject extraction. **23%** of the errors were due to relative sentences.

- Another problem that we already found for the subject is the **coordination** of elements (**12%**).

- Problems with **disambiguation** provoked **8%** of the errors and a **wrong chunking** provoked almost **5%** of them.

- The extraction of direct objects needs some lexical information of verbs which we did not have at the beginning. We needed to know if a verb is transitive or intransitive or if it can admit two objects. We added this information to our grammar as a **feature ("trans")**. Nevertheless, some verbs lacked this information and this caused almost **4%** of the errors.

- The fact that the objects sometimes were inside **quotation marks** affected their correct location. This problem represents **3%** of the errors.

- The remaining **18%** corresponds to miscellaneous causes.

3.3. Noun Modifier:

The biggest difficulty when extracting this dependency is not the location of the modifier but to attach it to the right modified noun. Indeed, the bad results in terms of precision coincide with the vast difficulty in deciding to which noun group a modifier is attached.
Nevertheless, we should distinguish between three types of modifiers: adjectives, prepositional phrases and appositions.

- Adjectives: no particular problems, they are in most of the cases found and well attached.

- Prepositional phrases: this kind of modifier presents the biggest number of errors. We have almost no structure that can tell us which name a prepositional phrase is modifying. The lack of named entities recognition module also contributes to an increase of errors. For example, when an entity like "*Ministerio de Trabajo y de Industria*" is not recognized as one entity, we will have two modifiers for *Ministerio,* and one for *Trabajo* where we should not have any.

- Appositions: they can be surrounded by brackets, quotation marks, commas or even nothing. The contexts are varied and very often there are proper nouns implied. That

> is the reason why this problem is also much related to the lack of named entities module. As we indicated before, the named entities recognition would allow to spot proper names of persons, locations and companies, which would be very useful for a correct disambiguation, a better segmentation and therefore, a better dependencies' extraction.

If we look at this example:

"*Algunos de los participantes llevaban banderas del grupo islámico prohibido Lashkar-e-Toiba y muchos gritaron lemas contra Estados Unidos y varios países europeos por la publicación de las viñetas de Mahoma.*"

(Some of the participants carried flags of the Islamic and forbidden group *Lakar-e-Toiba* and a lot of them yelled mottos against the United States and some other European countries because of the publication of Mahomet's caricatures)

The grammar will be able to extract all the modifiers -except for the one that is an apposition of a noun *MOD(grupo, Lashkar-e-Toiba)*- but we will have some noises like:

MOD(publicación, Mahoma),
MOD (países, publicación)
MOD (países, viñetas)
MOD (países, Mahoma)

All of these cases are prepositional phrases that in principle could be attached to any of those nouns. Nothing but semantics (however, not always!) could help us to improve our rules to extract more precise noun modifiers.

In general, we could state that the style of newspapers is often characterized by long and complex sentences that would need a deeper and more detailed context of the rules.