"Syntactic Parsing Of Unrestricted Spanish Text".

Irene Castellón*
Montse Civit *
Jordi Atserias **

*Laboratori de Lingüística Computacional.
Dept. Filologia Romànica.Universitat de Barcelona.
castel@lingua.fil.ub.es
civit@lingua.fil.ub.es
**Dept Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
batalla@lsi.upc.es

Abstract

This research focusses on the syntactical parsing of morphologycal tagged corpora. A proposal for a corpus oriented Spanish grammar is presented in this document. This work has been developed in the framework of the ITEM project and its main goal is to provide multilingual background for information extraction and retrieval tasks. The main goal of Tacat analyser is to provide a way of obtaining large amounts of bracketed and parsed corpora, both general and specific domain. Tacat uses context free grammars and has as input following categories of Parole specification.The incremental methodology that we use allows us to recognise different levels of complexity in the analysis and to produce compatible outputs of all the grammars.

1. Introduction

In this paper we present a proposal for a corpus oriented Spanish grammar developed in the framework of the ITEM[1] project. This work has been developed in order to provide multilingual background for information extraction and retrieval tasks. Information extraction and retrieval is an area that deals with many levels of information (Atserias et al.'98); the present of research focusses on the syntactical parsing of morphologycal tagged corpora.
We have used Tacat as an analyser that has as its main goal to provide a way of obtaining, at a moderate human cost, large amounts of bracketed and parsed corpora, both general and domain specific, for several purposes. The grammars developed are context free whose terminal categories following the parole specification.
The goal of the grammars is to get groups of the main constituents of sentences in Spanish. The syntactic work, presently, consists of the development of three grammars. Each one produces a complete analysis, but with different granularity about its structure. The incremental methodology that we use allows us to recognise different levels of complexity in the analysis and to produce compatible outputs of all the grammars.

Section 2 describes the TACAT Analyzer whereas section 3 deals with the Spanish grammars developed.

[1] Item TIC96-1243-C03-02

2. TACAT

This section describes TACAT, a tool for syntactically analysing tagged corpora. This parser allows partial parsing, parsing in several steps and modification of the parse tree structure on the fly. A more detailed description of the tool and its environment can be found in (Atserias & Rodríguez 98)

To implement the CFG parser we use the well known Bottom Up Chart Parser Algorithm with some modifications. These modifications and the main characteristics of the parser are described in the sections below.

2.1 The Input

The texts to be analysed must be previously pos-tagged (one or more POS tags per word). The tagset used for tagging can be freely defined by the user (the input corpus has to be previously tagged according to this tag set). But not only pos-tagged corpora can be used as input, a partially or fully analysed corpora can also be used as input.
TACAT accepts as input any tagged text (with a free format) as the corpus to be analysed or a partially analysed text in order to try to complete the analysis starting with the pre-analysed parts. It also handles incomplete analysis. So the TACAT parser can proceed with any grammar selected by the user, and the process can imply the performance of several parsing steps. Each one uses as input the result of the previous step for obtaining a more precise analysis.

2.2 Modifying the tree structure on the fly

Our aim is to allow linguists to write more human readable grammars but keep , as far as possible, the right structure of the parse tree. To avoid some of the problems that arise when using CFG we modify, on the fly, the structure on the parse tree according to the following directives.

- **Flat Categories:** These categories will not appear in the output if the category immediately above is the same. This is useful for avoiding the effect of the

successive application of recursive rules.

- **Hidden Categories:** These categories will not appear in the output analysed. This will allow the user to write human readable grammars more easily.
- **Group Categories:** These categories will appear in the output analysed only if they are the top node in the analysis tree (or a direct son of the Unkown node in the case of Partial Analysis). Some categories may not be relevant for the full analysis but they are relevant for grouping when the analysis is partial.
- **Notop Categories:** These categories will appear in the output analysed provided they are not the top node in the analysis tree (or a direct son of the Unknown node in case of Partial Analysis). Some categories may be added as a way to generate a more complex structure (e.g. subject) but must be considered inappropriated if there are no further results.

These categories also make it easier to adapt the grammar to the user/application needs, as those directives allow to reduction/modification the set of labels that appear in the parse-tree without, in fact, touching the grammar.

2.3 Choosing the best parse-tree

When there are some complete analyses for the whole sentence we first choose randomly (as we don't know the grammar's initial category) an inactive edge. The heuristic for consequently choosing the best analysis is to get the shortest rule that can be applied to obtaining this inactive edge in each step.

2.4 Robust Parsing

When there is no complete analysis for the whole sentence we proceed from left to right choosing the longest inactive edge and then using the same heuristic as in the complete analysis. All these sub-trees are joined as the son of a node labelled Unknown (as "??").

2.5 Time Optimisation

In order to speed up the algorithm TACAT handles the Epsilon production (rules with no category on the right side) in a special way. The Epsilon production is not stored as rules, instead the head symbol of these rules are marked as *anullable* in order to avoid the unnecessary trigging of this kinds of rules. So, the composition-edge chart method has been modified to add as a fact an anullable symbol if it's needed for the application of another rule.

To avoid the problem of anullable categories at the beginning of the right side of the rule, an index trigs the rules which have this fact as the first non anullable symbol. This index is build when the grammar is loaded. This approach seems to increases the parsing speed.

2.6 Portability, reusability and applications using TACAT

Portability:
At present, we have compiled and used successfully the TACAT parser under Linux, Unix and DOS.

Graphical Interfaces to TACAT:
As TACAT is implemented on C++ it has been easily integrated as a set of Tcl[2] /tk[3] commands for parsing and manipulating the grammars. Also a small graphical user interface has been developed for viewing the grammars, the symbol tables and so on. Our future goal is to extend this graphical environment to make it as powerful as the windows interface of TACAT. This will make the entire system fully portable between the different platforms (mainly UNIX/Linux and DOS/Windows).

Tacat and GATE:
TACAT has been coupled with GATE (Cunningham et al. 95) as part of the ITEM's integration task. GATE (General Architecture for Text Encoding) is a graphical environment developed by the University of Sheffield for integrating different tools used in Natural Language Engineering . The system is implemented in tcl/tk and C++.

Reusability:
As the TACAT parser is implemented in an Object Oriented Language (C++), it's easy to derive new classes to change the behaviour of the chart parser. For example, the output format can be easily changed by over-writing the tree virtual functions that outputs the best parse-tree. The modification of these tree functions allow one to write the best parse-tree in pre-order or post-order. So it was possible to couple TACAT to the tree drawing utilities[4] used to generate the graphics that appear in this document without , in fact, touching the original code.

3. Spanish Grammar for Corpora

3.1. Gramesp

GRAMESP (Civit &Castellón 98) is composed of three context free grammars and its application produces increasingly more refined analises. The main objective is to obtain different levels of analysis in order to always produce a parsing result. Basically, the syntactical phenomena that Gramesp recognizes at this moment are:

1) nominal (sn), adjectival (sa), verbal (sv), prepositional (sp) and adverbial phrases (sadv).

2) lexical and syntactic coordination (with some ambiguities)

[2] Tcl is a simple scripting language whose interpreter is a library of C procedures that can easily be extended.
[3] Tk improves tcl with commands for building graphical interfaces.
[4] Written by Joseph Rosenzweig.

3) subordination marks

4) Checking agreement within nominal phrases

3.1.1.Grammar 1

The first grammar (G1) has as input a pos-tagged text and the output is a bracketed corpus indicating the largest interpretation. G1 contains 381 rules and operates on morphological categories of Parole specification, whereby the first process consists in grouping these categories (a total of 339) in morphosyntactic ones (44).

E.g.: The adjective node can be formed by five types of parole categories.

a ==> aq0cs0. %alegre
a ==> aq0ms0. %bueno
a ==> aq0fs0. %bonita
a ==> aq0mp0. %baratos
a ==> aq0fp0. %guapas

Also, there are some rules that group several morphosyntactic categories with the same distribution in order to reduce their variety.
E.g.: pronouns can be:

pron ==> psubj.
pron ==> patons.
pron ==> pdem.

The rest of the rules have been designed for grouping phrases. G1, in general, recognizes groups such as nominal phrases formed by adjectival complements but neither those with prepositional phrases nor relative clauses. Prepositional phrases, adjective phrases and complex verbal forms are also recognized, so we can see the rule concerning to the infinitive groups, as can be seen in the example in figure 1.

grup-inf ==> infaux, parti.

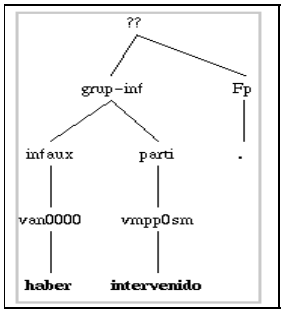

Figure 1: 'to have intervened'

Another syntactical phenomena recognized by G1 is lexical coordination. Figure 2 shows an analysis of lexical coordination :

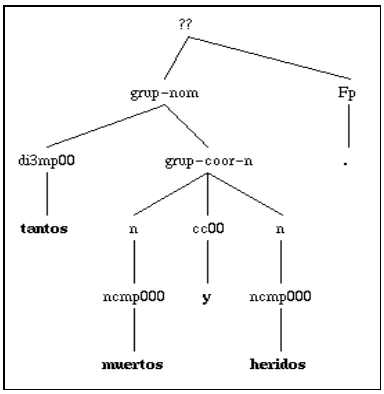

Figure 2:'so many dead and wounded people'

Moreover G1 forms phrases testing the constituency but not agreement restrictions. Therefore the groups are not exactly well formed.

*{{estos_di3mp00}_spec uchacho_ncmp000}_grup-nom

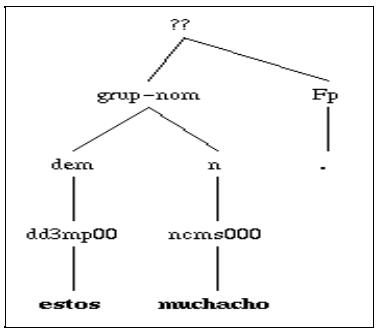

Figure 3: (*)'these boy'

As we have mentioned in section 2, the output structure can be modified by linguists in order to hide some categories. In the first grammar, the categories that have been hidden are the heads of the phrase and the categories that are obtained as a result of the application of recursive rules.

3.1.2.Grammar 2

The second grammar developed, G2 (537 rules), is practically equivalent to G1 but its analysis is more restrictive. This grammar checks the agreement of number in nominal and adjective phrases so it works with morphological information. This forces modification of the rules for adding morphological information.

as ==> aq0cs0. %alegre
as ==> aq0ms0. %bueno
as ==> aq0fs0. %bonita
ap ==> aq0mp0. %baratos
ap ==> aq0fp0. %guapas
ap ==> aq0cp0. %alegres

In some cases this modification has produced the increase of rules , that is the case, for instance, of pronouns:

pron3s ==> psubj3s.

pron3s ==> pdems.
pron3s ==> pinterrogs

In the following example (Fig. 4), we can see how agreement restrictions are applied.

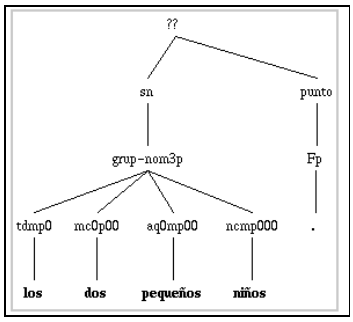

Figure 4: 'the two little boys'

To deal with coordination, the enlargement of the number rules has been considerable as the result of three factors:

(i) The intervention of morphological categories of person and number in the analysis.
(ii) The recursivity that allows the coordination of two or more elements.
(iii) The combination of different persons (1,2,3) in coordination rules.
Thus the coordination rules applied to the sentence 'Él, tú y yo comemos.' produces the analysis shown in figure 5:

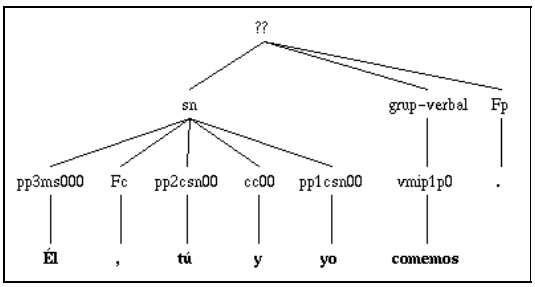

Figure 5: 'He, you and me eat'

In addition to nominal coordination, G2 solves the coordination of some adjective and verbal phrases.

grup-nom ==> grup-nom1s.
grup-verbal ==> grup-verb1s.
s-adj ==> grup-adjs.

The main groups of G2 are: grup-nom (nominal phrases), grup-adj (adjective phrases) , grup-verb (verbs and complex verbal forms) and sp (prepositional phrases).

In the output structure the hidden directives avoid bracketing the different nominal or adjective groups that have information about agreement of number and person (as grup-nom3p, grup-nom1s, etc.). The different directives also allow the output of grammar 1 and 2 to be equivalent.

3.1.3. Grammar 3

The third grammar (G3) has been defined according to genre. At the moment we have developed a version for parsing the CPirapides. It determines the boundaries of verbal phrases and sentences.

G3 groups the main constituents of G2 to form the verb phrase. This verb phrase includes verbal complements provided they appear behind the verb. The next rules deal with this issue.

sv ==> grup-verbal.
sv ==> grup-verbal, sn.
sv ==> grup-verbal, sp, sn
…

Another goal of G3 is the analysis of subordinate clauses, in fact G2 marks the subordination particle but not the boundaries of clauses. These are the corresponding rules in G3.

prop ==> subord, sv.
prop ==> prel, sv.
prop ==> cuyos, grup-nom3s, sv.

Moreover the last rules G3 propose a simple sentence structure.

frase ==> sn, sv.
frase ==> sv.

An example of their application can be seen in the figure 6.

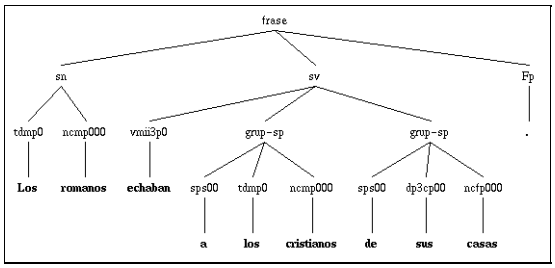

Figure 6: Sentence analysis

3.2. Analyzing Corpora

The analysis of unrestricted text is a task that requires paying special attention to the methodology. The set of productions of real language (Lexesp) has structures that are not easy to predict. The length of the sentences is variable (but long sentences are usual), subordinate clauses appear frequently and there are several types of text such dialogues or narrations, that have different sentences structures.
In this way the analysis of corpora has been carried out increasingly to CPirápides and Lexesp that have different properties. CPirápides is a reduced corpus of about 4.900

sentences corresponding to prototypical examples of verbal senses. These senses are linked with Wordnet and EuroWordNet (Miller 91) (Vossen 98) moreover all the verbal senses are related to Levin verbal classes (Levin 93) (Dorr et al 97). From a syntactical point of view, Cpirápides presents very simple sentences, where adjunct rarely appear elements. On the contrary, Lexesp[5] is a real corpus of about 5 million words. It includes several types of text , and its structures are more complex than CPirápides. The syntactical difficulties that arise in the analysis of Lexesp is to determine the boundaries of sentences.
Tacat allows application of the grammars in an accumulative way. This functionality makes the grammar able to always provide a parsing increasingly refined. The order of application of grammars is very flexible. It's possible to apply the grammars developed in different order o independently. The possible combinations could be :

at sentence level:
1) G2 < G3
2) G1< G3
3) G2 < G1 < G3

at phrase level:
4) G2 < G1
5)G2

The steps of analysis that have been proposed as a long-term job are:

a) Parenthization of Cpirapides at phrase level with G2-1
b) Parenthization of Cpirapides at sentence level with G3
c) Modification by hand of Cpirapides at sentence level in order to create a syntactic Corpus (treebank) that is necessary for many learning processes related to Natural Language knowledge sources.
d) Parenthization of Lexesp at phrase level with G2-1
e) Parenthization of Lexesp at sentence level

At the moment we are beginning the third step, and the evaluation that we will present shows partial results because the error index has been detected manually. The evaluation has been carried out with a sample of 250 sentences of CPirapides. An initial parsing of Lexesp has been carried out, but there are not significative results yet. The chosen performance sequence is G2 < G3. It could be applied G1 between G2 and G3 but the secondary effects of this analysis have not been tested.

As we have mentioned above, the error index has been detected manually with a sample of about 5% of the corpus (250 sentences). This first sample has served us to establish the evaluation criteria.

[5] Lexesp (Lexesp II Acción especial APC96-0125)

We must consider different kinds of errors:
a) morphological tagger errors
b) errors that provide of the grammar:
    2. the lack of structure or lexical category in the rules
    3. the selection of a parser tree when there is/are other(s) more correct, in general produced because of structural ambiguities

The first type of error, that was found around 5% (Padró 97), has been corrected manually in Cpirapides in order to identify the parser errors. The second error (b.1) has a simple solution, adding the structure in the grammar rules, the consecuence of this enlargement could be an increase of the ambiguity error.

The third error is very difficult to correct by means of a purely syntactical grammar, so, in order to remove this error of Cpirapides , we are now correcting the data manually. Basically we find two problems in the phrase level parsing: first, the linear order, and second, the ambiguity of structures. Spanish has a free order in verb complement, now we are begining an study of the different combinations in reference to the order of apparition of these complements. The consequence of this study will be the enlargement of the relative rules to verb phrases and therefore it seems obvious that the ambiguity problem in the analysis will grow.

The ambiguity of grouping arises when we are in front of equivalent sentences that have different dependence relationships, as is the case of prepositional phrases or coordination. As an example, let us observe the following sentences:

a.1 Anula las reservas de avión
(He canceles the fligth reservation)
b.1 Borraron las pruebas del suelo
(They remove the evidences off floor)
In both sentences we have the same sequence:

a.2 v sn sp
b.2 v sn sp

The first grammar (G2) produces as an analysis the sequence of three constituents for both sentences:

a.3

{Anula_vmip3s0**}-grup-verb**
{{las_tdfp0}_spec reservas_ncfp000**}_sn**
{ de_sps00 { avión_ncms000 }sn **}_grup-sp**

b.3
{ Borraron_vmis3p0**}-grup-verb**
{{ las_tdfp0}_spec pruebas_ncfp000**} _sn**
{ del_spcms { suelo_ncms000 }_sn **}_grup-sp**

If we suppose that for a.1 the correct analysis is not a.3 , an acceptable analysis could be:

a.3'
{Anula_vmip3s0**}-grup-verb**
{{las_tdfp0}_spec reservas_ncfp000{de_sps00 {
avión_ncms000 }sn }_grup-sp **}_sn**

G3 poses the same problem that G1 and G2, the assignment of the prepositional phrases, and it does not seems possible to arrange it with a context free grammar. Lets observe next example, corresponding to figure 7, b.4 produced by G3:

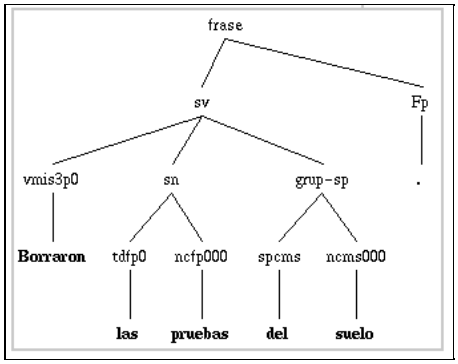

Figure 7:The prepositional phrase as verb complement

b.4
{Borraron_vmis3p0
{{las_tdfp0}_spec pruebas_ncfp000 **} _sn**
{ del_spcms suelo_ncms000 **}_sp }_sv**

Although this analysis is correct, it does not happen the same with the following (Fig. 8) :

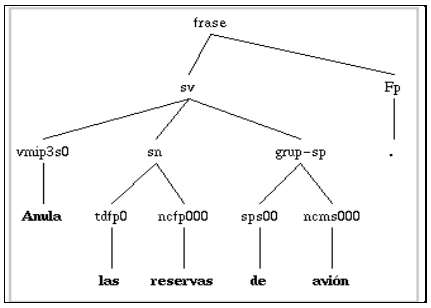

Figure 8: The ambiguity of prepositional phrase

a.4*
{Anula_vmip3s0
{{las_tdfp0 reservas_ncfp000 **}_sn**
{ de_sps00 { avión_ncms000 }_sn **}_grup-sp }_sv** ._Fp

This kind of mistake has not been taken into account in the following table in G2 because it is an structure that does not appear in the rules. On the contrary, it has been taken into account in the errors of G3, because it provides a sentence structure.

The percentage of error has been obtained in two ways , first evaluation corresponding to the number of errors in relation with the number of groups produced in the sample (1), and the second evaluation in relation with the number of sentences of corpus(2)

CPIRAPIDES
sample: 250 sentences

	G2	G3
E b.1	0'9%	0'7%
E b.2	0%	3'9%

Table 1 . Evaluation 1 of parsing results

	G2	G3
E b.1	3%	3%
E b.2	0%	16%

Table 2 . Evaluation 2 of parsing results

The results show that the lack of structures is the same in both grammars, but ambiguity (b.2) errors grow in G3. It is due to the fact that G3 enriches the parsing with nodes such as verbal phrases (VP), subordinate clauses and sentence .

The reason for the decline of b.1 in the first evaluation between G2 and G3 is the difference in the number of groups of each output, we can observe that in the second evaluation this index of errors is the same.

4. Future work

In relation to the analyzer, we want to modify the Flat directive in order to define Flat rules or path instead of FLAT categories.
Another task could be to extend TACAT to be contextual and allowing the user to write his/her own semantic functions in C++ by using dynamic libraries.

We are considering to use TACAT in the developing of a multilingual IE system.

Concerning to the grammars, the first objetive is to evaluate the analysis with a larger sample than the data presented in this paper.
The next task that we will do is to correct manually the parsing of G3 of Cpirapides in order to create a treebank . Once complete it, this corpus could be a possible source for automatic parsing
with statistical methods. The subsequent task is the parsing of Lexesp and therefore the estension of the grammars.

The sources and documentation (related papers and an html documentation of the classes) of TACAT and also the Spanish grammars are avalaible throught the web at http://www-lsi.upc.es/batalla/~batalla/research/tacat.html

5. References

Atserias J y H. Rodríguez (1998) "TACAT: TAgged Corpus Text Analyzer" Technical Report LSI-UPC RT-2-98.

Atserias J., Català N., Castell N., Rodríguez H., Turmo J.

"Del texto a la información" to appear in *Novatica.* 1998.

Carmona J., Màrquez,Ll., Ll.Padró, S.Cervell, M.A.Martí, R.Placer, M.Taulé (1998) "An Environment for Morphosyntactic Processing Of Unrestricted Spanish Text" submitted in The Meeting of European Language Resources and Evaluation.

M. Civit y I. Castellón (1998)"Gramesp: una gramática de corpus para el español" AESLA 1998

Cunningham, H., R.J. Gaizauskas, Y. Wilks (1995) "A general Architecture for Text Enginneering (GATE) - a new approach to Language Engineering." R&D CS-95-21 Technical Report.University of Sheffield.

Dorr,B., M.A.Martí, I.Castellón (1997) "Spanish EuroWordNet and LCS-Based Interlingual MT" (submitted in The Meeting of European Language Resources and Evaluation)

Levin, B. (1993) *Towards a lexical organization of English verbs*, University of Chicago Press, Chicago.

Miller, G. (1991) 'WordNet: a dictionary browser', en *Proceedings of the First International Conference on Information Data*. Waterloo, Ontario: University of Waterloo Centre for the New Oxford English Dictionary.

Padró,L.(1998) "A Hybrid Environment for Syntax-Semantic Tagging".PhD Thesis. Dept. Llenguatges i Sistemes Informatics.Universitat Politecnica de Catalunya. Barcelona.

Vossen, P. EuroWordNet *Computer & Humanities* (monographical volume) (Forthcoming)