# Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text

**J. Atserias*, J. Carmona*, I. Castellón†, S. Cervell†, M. Civit†, L. Màrquez*, M.A. Martí†, L. Padró*, R. Placer†, H. Rodríguez*, M. Taulé†, J. Turmo*.**

* Software Department – Universitat Politècnica de Catalunya
c/ Jordi Girona 1–3, 08034 Barcelona, Catalonia.

† Computational Linguistics Laboratory – Universitat de Barcelona
Gran Via 585, 08007 Barcelona, Catalonia.

**Abstract**

This on–line demonstration is about an environment for massive processing of unrestricted Spanish text.

The system consists of three stages: morphological analysis, POS disambiguation and parsing. The output of each can be pipelined into the next. The first two phases are described in (Carmona *et al.*, 1998) and the third is described in (Atserias *et al.*, 1998), both published in this conference.

The execution may be performed inside the GATE environment, which enables visualization and analysis of intermediate results, or either in background, if higher efficiency is required for massive text processing.

**Keywords:** Morphological analysis, corpus linguistics, POS tagging, linguistic resources.

## 1 Introduction and Motivation

This demonstration is related to the systems for unrestricted Spanish text processing described in (Carmona *et al.*, 1998) and (Atserias *et al.*, 1998). The former describes the morphological analysis and parsing modules, and the later explains the syntactic parser.

Those systems are being used inside the ITEM and LexEsp projects as basic modules for Spanish text processing as described below.

### 1.1 ITEM Project

ITEM is a project funded by Spanish Research Department (CICYT) consisting basically of integrating different existing NLP tools and resources in a unique environment, in order to enable and ease the construction of multilingual information extraction and retrieval systems.

The environment includes tools for NLP of Catalan, Basque and Spanish. The integrated tools include basic NL tasks (tokenizers, morphological analyzers, taggers, parsers, etc.) as well as higher level tools oriented to information extraction. New tools and resources are also being developed, and existing tools are improved and integrated.

The integration environment also contains several lexical resources such as corpus, machine–readable dictionaries (MRDs), lexicons, taxonomies, grammars, etc.

All the integrated tools and resources are documented, available and transportable. The software used to support this integration is GATE[1] (Cunningham *et al.*, 1996).

Partners in this project are the Computational Linguistics Group from the University of Barcelona (http://www.ub.es/ling/labcat.htm), the NLP research group from the Technical University of Catalonia (http://www.lsi.upc.es/ acquilex/nlrg.html), the NLP group from the Basque Country University (http://www.ji.si.ehu.es/Groups/IXA/), and the NLP group from the Spanish Open University, UNED (http://sensei.ieec.uned.es/item/grupoLN.htm).

### 1.2 LexEsp Project

The LexEsp Project is a multi–disciplinary effort impulsed by the Psychology Department from the University of Oviedo. It aims to create a large database of language usage in order to enable and potentiate research activities in a wide range of fields, from linguistics to medicine, through psychology and artificial intelligence, among others.

One of the main issues of that database of linguistic resources is the LexEsp corpus, which contains 5.5 Mw of written material, including general news, sports news, literature, scientific articles, etc.

This paper is organized as follows: Section 2 describes the morphological analyzer which constitutes the first analysis step. The following stage –POS tagging– is described in section 3, and the parsing process is explained in section 4. Finally, section 5 outlines the contents of the on–line demonstration.

## 2 Morphological Analysis

The construction of the MACO+ morphological analyzer consisted of two steps: First, the old MACO version was used to generate all possible Spanish forms following the criteria described in section 2.1, and they were stored in a dictionary. Second, as described in

[1] GATE (General Architecture for Text Encoding) is a graphical environment developed in Sheffield to integrate different NL Engineering tools. This integration is reached by sharing a unique common syntactic format based on the TIPSTER Architecture.

section 2.2, an efficient look-up procedure and other specific modules were written to exploit the data.

### 2.1 Form Generation Linguistic Model

Linguistic knowledge is organized in root classes and inflectional paradigms. Roots are classified in terms of the inflectional paradigms they accept. Words are considered orthographically. In this sense, linguistic regular forms as 'cazar/caces', 'apagar/apague' has been considered as having two roots: 'caz/cac' and 'apag/apagu'. For each kind of ortographical irregularity there is an inflectional paradigm and a class of roots.

The set of roots was collected from MRDs and corpora and semi-automatically assigned to a root class. All the inflected forms corresponding to the stored roots were generated by exhaustively applying the inflectional rules. Overgeneration was avoided, thus, all generated forms are correct forms in Spanish.

Derivational processes allow to reduce significantly the number of roots, but derivation implies many difficulties in the lemma assignment, for this reason we have dealed only with inflection.

In order to complete and refine the obtained set of forms, the root set is periodically enlarged with new roots. Current dictionary contains over 800,000 forms corresponding to some 90,000 lemmas. For more details on the linguistic model see (Carmona *et al.*, 1998).

### 2.2 Architecture of MACO+

The architecture of the morphological analyzer is a modular pipeline of specialized recognizers. Modules can be activated or deactivated for each particular analysis. The existing modules recognize each of the following items: simple date patterns, abbreviations, proper nouns, compounds, numbers, punctuation, words and *clíticos* (suffixed pronouns). The word module is the real analyzer, while the others are specific heuristics to identify the listed special items. Obviously, heuristics in each module can be improved independently.

The implementation of MACO+ is UNIX-perl based. This makes it easily transportable and overcomes the first flaw of the first version.

The second drawback was overcome by using all the roots and inflectional paradigms to generate all possible forms. The forms were stored in a dictionary and the word analysis critical module is implemented as a dictionary look-up procedure using sophisticated caching and indexing techniques. In addition, this module is able to auto-configure in order to best exploit the particular hardware it is running on. See (Carmona *et al.*, 1998) for details.

### 2.3 Results

MACO+ has been tested on a fresh unrestricted corpus of 100,000 words. For each word, all possible interpretations are obtained. Each interpretation contains the lemma and a PAROLE compliant morphological tag describing information such as category, subcategory, gender, number, person, mode, etc.

Working with all modules, the current version of the analyzer has a speed of over 600 words/second on a SUN Ultra Sparc architecture, and over 200 words/second running under Linux on a Pentium-120 processor. Previous figures include input/output processing time.

The resulting coverage is about 99.5%, on the 5.5Mw LexEsp corpus. The analyzed corpus presents a 39.26% of ambiguous words and an average ambiguity ratio of 2.63 tags/word for ambiguoug words, 1.64 overall. The estimated recall (words that have the correct tag among those proposed) is 99.3%.

## 3 Morphosyntactic Disambiguation

The results produced by the morphological analyzer described so far can be pipelined into a morphological disambiguator –POS tagger– to obtain the appropriate reading in the given context.

In the framework of the ITEM and LexEsp projects, two different POS taggers are being used to annotate a Spanish corpus of over 5Mw. First, a decision–tree based tagger (Màrquez & Rodríguez, 1997), which learns a language model from a tagged corpus, as well as prediction rules for the possible readings for words not found in the dictionary. Second, a relaxation labelling based tagger (Padró, 1996), which can use and combine information from different sources (n–gram, decision trees, manually written, etc.) provided it is put in the form of context constraints.

In addition, it is being studied whether it is possible to take advantage of their collaboration to produce better disambiguation, and use it to enlarge the training corpora keeping the noise to a minimum.

## 4 Syntactic Parsing

This section describes the TACAT parser as well as a proposal of general grammar for Spanish.

The main goal of TACAT is to provide a way of obtaining, at a moderate human labour cost, large amounts of bracketted and parsed corpora, both general and domain specific. The goal of the grammar is to get groups of the main constituents of sentences in Spanish.

### 4.1 The Parser

TACAT is a tool for syntactically analysing tagged corpora. This parser allows partial parsing, several parsings steps and parse tree structure modification. The system uses a bottom up chart parser with the following characteristics:

1. Handling lambda productions in a special way: the empty productions are not stored as rules. Instead, the head symbol of this rules is marked as nullifyable in order to avoid its unnecessary trigging. So, the composition-edge chart method has

been modified to add as a fact a nullifyable symbol when it is needed for the application of another rule. To avoid the problem of nullifyable categories at the beginning of the right hand side of the rule, when a new fact (inactive edge) is added, an index (that is build when the grammar is loaded) triggers the rules which have this fact as the first non nullifyable symbol. This increases the parsing speed.

2. The Input: The texts to be analysed must be previously POS tagged The tagset used for tagging can be freely defined by the user (the input corpus has to be previously tagged according to this tagset). But not only tagged corpora can be used as input, also a partially or fully parsed corpora can be used as input. The system also handles incomplete analyses. So the TACAT parser can proceed with a grammar selected by the user, and the process can imply the performance of several parsing steps. Each one using as input the result of the previous step for obtaining a more precise analysis.

3. Modifying the tree structure: Our aim is to allow linguists to write a more human-readable grammars but keeping as much as posible 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.

   - The Flat Categories will not appear in the output if the immediately category above is the same.
   - The Hidden Categories will not appear in the output analysed.
   - The Group Categories will appear in the output analysed only if they are the top node in the analysis tree.
   - The Notop Categories will appear in the output analysed only if they are not the top node in the analysis tree.

4. Choosing the best parse-tree: When there are some complete analysis for the whole sentence we first we choose randomly (as we don't know the grammar's initial category) an inactive edge and then the Heuristic for choosing the best analysis is to get the shortest rule that can be applied to obtaining this inactive edge in each step.

5. Partial Parsing: When there is no complete analysis for the whole sentence we proceed from left to right choosing the longest inactive edge and then use the heuristic for the complete analysis.

TACAT, implemented on C++, has been integrated inside GATE as part of the ITEM's integration task.

### 4.2 Spanish Grammar

Three grammars have been developed and its succesive applications produce analysis increasingly refined.

The first grammar (G1)[2] has 381 rules and operates with morphological categories of Parole specification, whereby the first process consists in grouping these categories (a total of 339) in morphosyntactic ones (44). G1 recognizes simple groups as nominal, prepositional and adjective phrases, periphrastic verbs and lexical coordination. The input of G1 is a tagged corpus and the output is a bracket corpus indicating the longest interpretation.

The second grammar developed, G2 (537 rules), is practically equivalent to G1 but its analysis is more strict. This grammar works with morphological information in the syntactic level because it checks the number's concordance of nominal and adjective phrases. G2 also solves the coordination of some nominal and verbal phrases. The outputs of G2 and G1 are the input of G3.

The third grammar (G3) has been defined depending of text type (genre). We have also developed an extension (G3pir) to parse the Pirapides Corpus[3], it determines the boundaries of verbal phrases and sentences. Now we are working in the extension (G3lex) to parse the LexEsp corpus.

## 5 Demonstration Content

The on–line demonstration will show the performance of the system in two different working environments:

- Inside the GATE environment, using its powerful visualization capabilities. Figure 1 shows the pipeline of the three modules inside GATE. GATE enables to redirect the output of any module to the input of any other –provided the kinds of information they use are compatible–. This allows great processing flexibility to easily build high level NL processes.

- In background mode, to demonstrate the system speed in massive processing when no visualization is required. The background mode runs as a UNIX pipeline, with no user interface, oriented to the background processing of massive text.

In both cases, all three stages of the system – morphological analyzer, POS tagger and parser– will be demonstrated separately, showing their intermadiate results. A visualization of those intermediate results is shown in the following figures: Figure 2

[2] G1 is an augmented version of the grammar developed in (Climent, 1997).

[3] In Pirapides project we are developing a verbal lexicon that will be applied in the syntactic and semantic anotation of corpora. Pirapides is a Linguistic Project of the Computational Linguistics Laboratory, which includes a corpus with 4006 sentences created in Computational Linguistic Laboratory (UB).

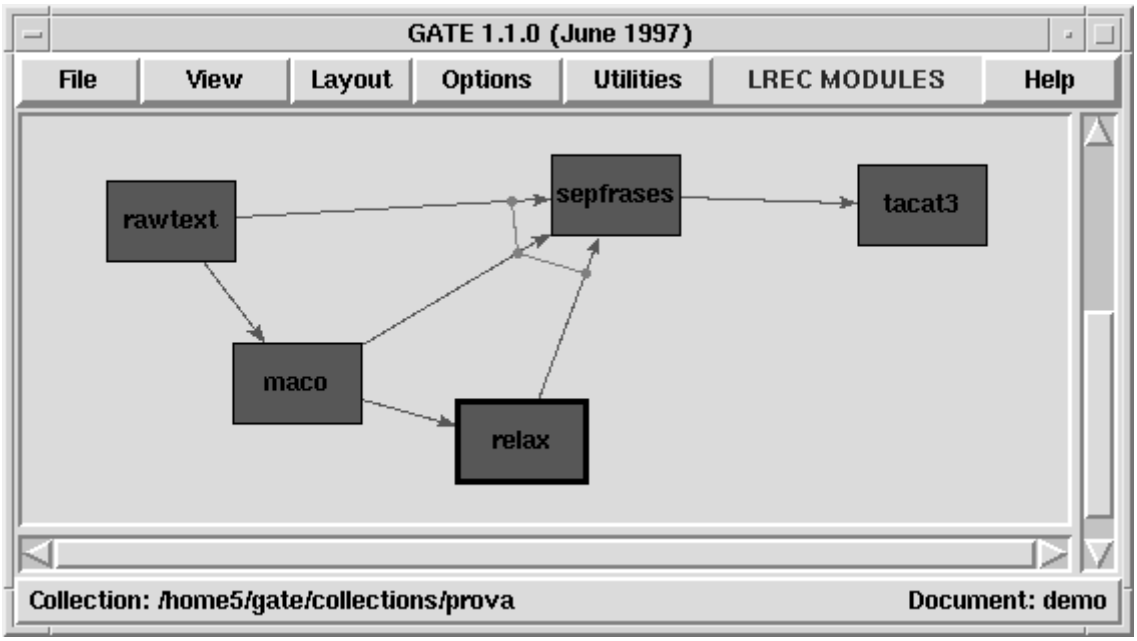


Figure 1: A pipelined system built inside GATE

shows the morphological analyzer output for a simple sentence, figure 3 shows the relaxation labelling based disambiguation on the MACO+ output for the same sentence, and figure f-tacat shows a partial parsing produced using grammar G2, which identifies –in this example– noun phrases, verb phrases and prepositional phrases.

Sample results from morphological analysis are shown in figure 2, the disambiguation results after POS tagging can be found in figure 3, and the tree resulting from the parsing process appears in figure 4.

## 6 Acknowledgments

This research has been partially funded by the Spanish Research Department (CICYT's ITEM project TIC96–1243–C03–02), by the EU Commission (EuroWordNet LE4003) and by the Catalan Research Department (CIRIT's quality research group 1995SGR 00566).

## References

Acebo, S.; Ageno, A.; Climent, S.; Farreres, X.; Padró, L.; Ribas, F.; Rodríguez, H. & Soler, O. (1994). MACO: Morphological Analyzer Corpus-Oriented. ESPRIT BRA-7315 Acquilex II, Working Paper #31.

Atserias J; Castellón I & Civit M. (1998). Syntactic Parsing of Unrestricted Spanish Text. In *Proceedings of 1st International Conference on Language Resources and Evaluation*, LREC'98. Granada, Spain.

Cardie, C. (1994). *Domain Specific Knowledge Acquisition for Conceptual Sentence Analysis.* PhD Thesis, University of Massachusets, Amherst, MA.

Carmona J.; Cervell S.; Màrquez L.; Martí M.A.; Padró L.; Placer R.; Rodríguez H.; Taulé M. & Turmo J. (1998). An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of 1st International Conference on Language Resources and Evaluation*, LREC'98. Granada, Spain.

Climent, S. (1997) *CHAOS (Chunk Analyser of Spanish).* Report UB-LG.1997-1 Secci de Lingstica General, Universitat de Barcelona.

Cunningham, H.; Wilks, Y. & Gaizauskas, R. (1996). GATE - a General Architecture for Text Engineering. In *Proceedings of 16th International Conference on Computational Linguistics*, COLING '96. Copenhagen, Denmark.

Daelemans, W.; Zavrel, J.; Berck, P. & Gillis, S. (1996). MTB: A Memory–Based Part–of–Speech Tagger Generator. In *Proceedings of 4th Workshop on Very Large Corpora, Copenhagen.*

Elworthy, D. (1993). Part–of–Speech and Phrasal Tagging. Technical Report, ESPRIT BRA–7315 Acquilex II, WP #10.

Karlsson, F.; Voutilainen, A.; Heikkilä, J. & Anttila, A. (1995). *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter.

Larrosa, J. & Meseguer, P. (1995) Constraint Satisfaction as Global Optimization. In *Proceedings of 14th International Joint Conference on Artificial Intelligence*, IJCAI '95.

Magerman, M. (1996). Learning Grammatical Structure Using Statistical Decision–Trees. In *Proceedings of the 3rd International Colloquium on Grammatical Inference, ICGI '96.* Lecture Notes in Artificial Intelligence 1147.

Màrquez, L. & Padró, L. (1997). A Flexible POS Tagger Using an Automatically Acquired Language Model. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, E/ACL '97. Madrid, Spain.

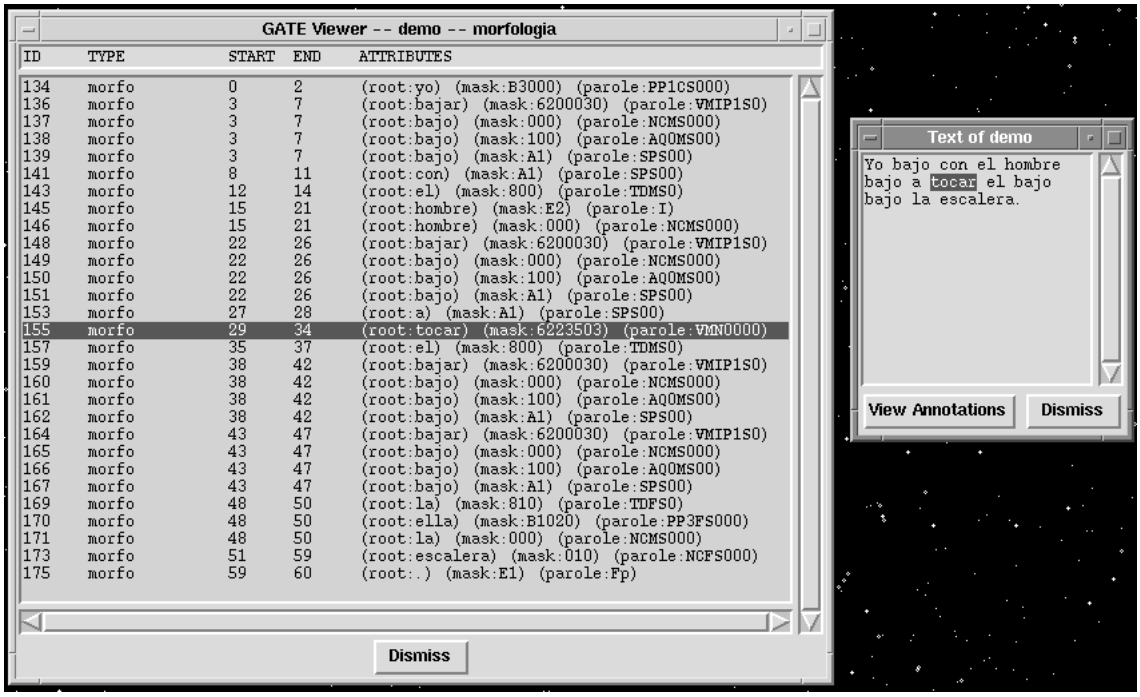


Figure 2: Morphological analyzer results


Màrquez, L. & Rodríguez, H. (1997). Automatically Acquiring a Language Model for POS Tagging Using Decision Trees. In *Proceedings of the Second Conference on Recent Advances in Natural Language Processing*, RANLP '97. Tzigov Chark, Bulgaria.

Màrquez, L. & Rodríguez, H. (1998). Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the 10th European Conference on Machine Learning*, ECML'98. Chemnitz, Germany.

McCarthy, J. & Lehnert, W. (1997). Using Decision Trees for Coreference Resolution. In *Proceedings of 14th International Joint Conference on Artificial Intelligence*, IJCAI '95.

Mooney, R. (1996). Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning. In *Proceedings of Conference on Empirical Methods in NLP*, EMNLP '96.

Padró, L. (1996). POS Tagging Using Relaxation Labelling. In *Proceedings of 16th International Conference on Computational Linguistics*, COLING '96. Copenhagen, Denmark.

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

Pelillo, M. & Refice, M. (1994). Learning Compatibility Coefficients for Relaxation Labeling Processes. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.16, n.9.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann. San Mateo, CA.

Rosenfeld, R.; Hummel, R. & Zucker, S. (1976). Scene labelling by relaxation operations. In *IEEE Transactions on Systems, Man and Cybernetics*, Vol.6, n.6.

Schmid, H. (1994). Probabilistic Part–of–Speech Tagging Using Decision Trees. In *Proceedings of the Conference on New Methods in Language Processing.* Manchester, UK.

Southwell, R. (1940). *Relaxation Methods in Engineering Science.* Clarendon.

Voutilainen, A. & Padró, L. (1997). Developing a Hybrid NP parser. In *Proceedings 5th ACL Conference on Applied Natural Language Processing*, ANLP '97, Washington DC.

Waltz, D. (1975). Understanding line drawings of scenes with shadows. *Psycology of Computer Vision.* P. Winston, New York: McGraw-Hill.

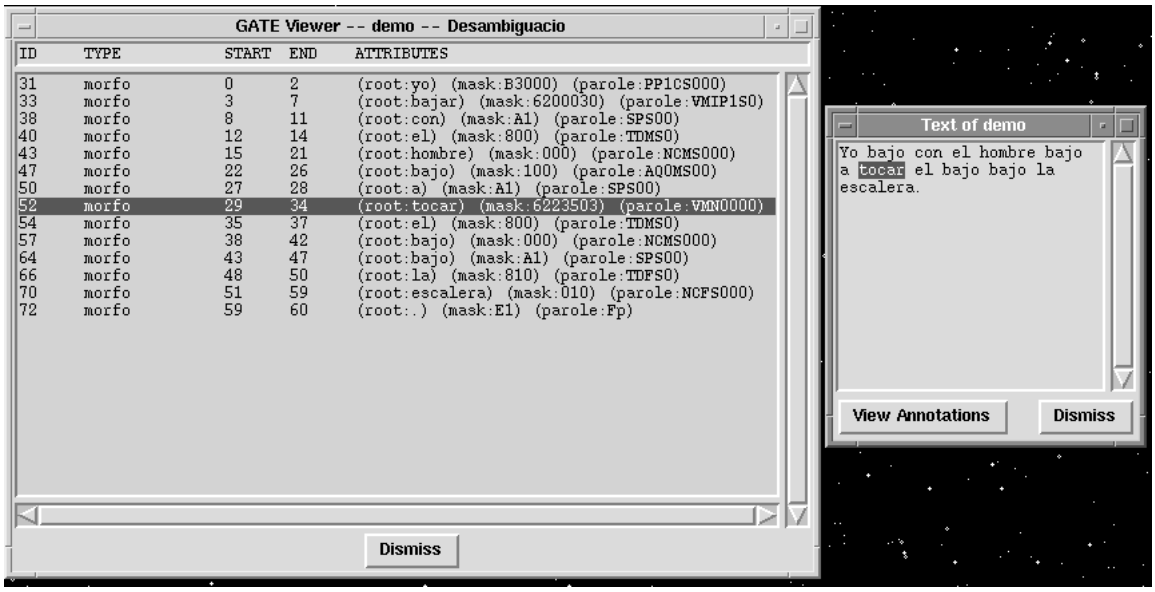


Figure 3: POS tagger results

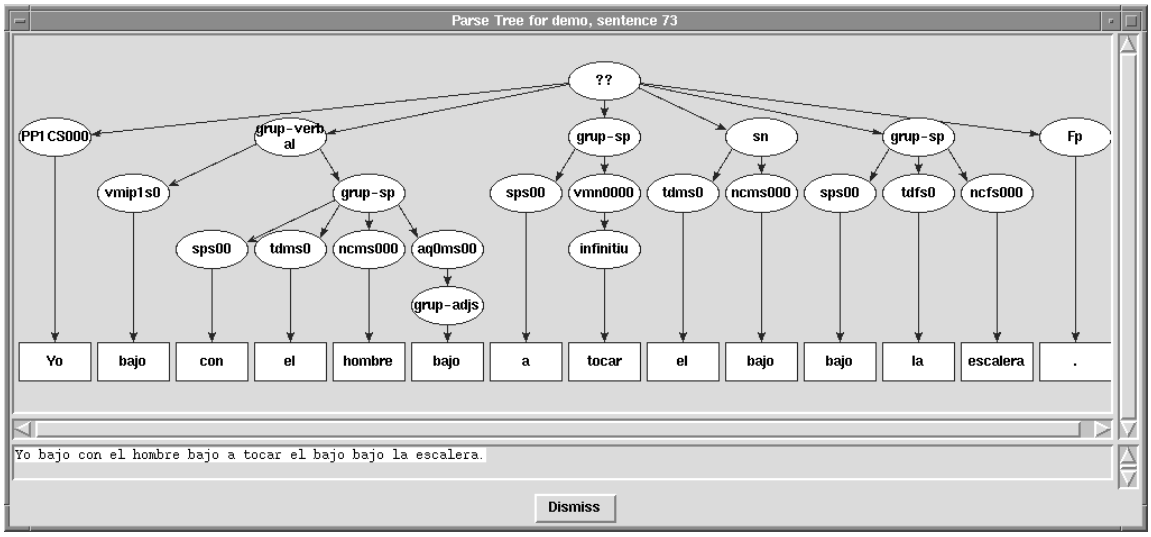


Figure 4: Parser results