A Treebank of Spanish and its Application to Parsing

Antonio Moreno[*], **Ralph Grishman**[†], **Susana López**[*],
Fernando Sánchez[*], **Satoshi Sekine**[†]

[*]Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid, Spain
{sandoval, susana, fernando}@maria.lllf.uam.es

[†] Department of Computer Science
New York University, U.S.A
{grishman, sekine}@cs.nyu.edu

Abstract

This paper presents joint research between a Spanish team and an American one on the development and exploitation of a Spanish treebank. Such treebanks for other languages have proven valuable for the development of high-quality parsers and for a wide variety of language studies. However, when the project started, at the end of 1997, there was no syntactically annotated corpus for Spanish. This paper describes the design of such a treebank and its initial application to parser construction.

1. Constructing a Spanish treebank

1.1. Preliminary considerations

This paper presents joint research between a Spanish team and an American one on the development and exploitation of a Spanish treebank. Such treebanks for other languages have proven valuable for the development of high-quality parsers and for a wide variety of language studies. As there was no previous experience in building a syntactically annotated corpus for Spanish, the first effort consisted necessarily in writing a set of annotation guidelines. The starting point was the existing documentation at that time, especially the Penn Treebank project (Marcus, Santorini and Marcinkiewicz, 1993; Bies et al., 1995), the EAGLES preliminary recommendations (EAGLES, 1996), and the Negra corpus (Skut et al., 1997).

Our experience in developing Spanish NLP systems told us that a pure phrase structure annotation (typical of the English treebanks) would not be enough for inducing relevant rules for Spanish. At the least, information about agreement and syntactic functions is necessary for Spanish, and we wanted to incorporate that information in our trees in the form of features.

The treebank has been created mostly by hand, although some automatic pre-tagging of the data is performed, as described below, to speed treebank creation.

1.2. Data selection

As of September 1999, the corpus consisted of 1,500 annotated sentences, with a total of 22,695 words and an average of 15.13 words/sentence. The sentences were taken from two different sources, a newspaper on-line edition (*El País Digital*, `http://www.elpais.es/`) and a consumer association magazine (*Compra Maestra*). The selection was made by the human annotators, taking as guidelines their qualitative and subjective knowledge of the complexities of Spanish. In accordance with this, the first 500 sentences were chosen by difficulty criteria such as discontinuous constituents, ambiguity, several embedded clauses in one sentence, anaphora and discourse markers, etc. The idea was to attack the problematic issues from the beginning.

In the current phase, we use an automatic selector of sentences from the html version of the sources. The program allows us to specify the range of sentence length (e.g. from 10 to 35) that the randomly chosen sentences should have, although we have not used this restriction in creating our current treebank. The idea now is to avoid human bias in the selection.

1.3. The annotation guidelines

The 88-page annotation guidelines includes a typed inventory of categories and features (a small fragment is shown in Table 1) , the annotation scheme, and specific directions for a great variety of Spanish phenomena[1]. The trees are encoded in a nested, parenthesized structure, with the elements at each level including the (part of speech or phrasal) category, the (syntactic and semantic) features, and the constituent nodes.

” The lexical categories are represented as

```
(CAT "<string>" "<lexeme>" FEATURE1
FEATURE2 ...  FEATUREn)
```

and the non-terminal constituents as

```
(CAT1 ...
  (CAT2 ...  )
  ( ...  )
  (CATn ...  ))
```

The structure closely reflects the surface syntax: in particular, we have been very cautious about empty categories. Only null subjects (very frequent in Spanish, not only in non-finite clauses like raising and control, but in finite verb sentences) and elided material (just in conjunctions, not in comparatives or traces) are annotated. In tagging, we follow the Penn Treebank schema, with * for null subjects and

[1] The annotation guidelines are available upon request to sandoval@maria.lllf.uam.es.

Categories	Features and values
N	{MASC ; FEM ; NEUT} {SG ; PL} {ACRNM ; PCENT ; TRATM} {PROPER} {MEASURE} {FOREIGN-WORD} {IDIOM} {TIME} {LOCATIVE} {COMPARATIVE} {COORDINATED}
NP	{SUBJ ; OBJ1 ; OBJ2 ; OBL ; ATTR ; TIME ; LOCATIVE ; MEASURE ; APPOS} {REF} {ID} {COMPARATIVE ; COMPARATIVE-1 ; COMPARATIVE-2} {MASC ; FEM ; NEUT} {SG ; PL} {P1 ; P2 ; P3} {IDIOM} {POLITE} {COORDINATED}

Table 1: Feature specification for N and NP. Braces enclose mutually exclusive sets of values.

? for elided elements. We use the features REF and ID to index together the expressed element and the elided one. Figure 1 shows an example of subject control in our annotation. Figure 5 provides an example of coordinated sentences with elided elements.

```
(S
  (NP SUBJ ID-1 SG P3
    (ART "<El>" "el" DEF MASC SG)
    (N "<Gobierno>" "Gobierno" SG P3))
  (VP TENSED PRES IND SG P3
    (V "<quiere>" "querer" TENSED PRES IND SG P3)
    (CL INFINITIVE OBJ1
      (NP * SUBJ REF-1)
      (VP UNTENSED INFINITE
        (V "<subir>" "subir" UNTENSED INFINITE)
        (NP OBJ1
          (ART "<los>" "el" INDEF MASC PL)
          (N "<impuestos>" "impuesto" MASC PL))))))
```

Figure 1: Subject control: ”El Gobierno quiere subir los impuestos” *The Government wants to raise taxes*.

In cases where there was uncertainty about how to assign a given feature, it has been omitted. The guidelines (Moreno, López and Sánchez, 1999) provide directions on how to annotate ambiguity in constituent attachment, null elements, and complex constituents. For a linear phrase structure representation, multiword constituents (i.e. several words forming a single unit) are problematic. Verbal periphrasis (i.e. compounds of AUX + (PARTICLE) + MAIN VERB), phrasal compounds, lexicalisation (such as the so-called ”support verbs”), or portmanteau words are instances of asymmetric relations between the surface strings and their semantic meaning as a whole. In order to express both levels, one element of the feature description is used for the surface representation (marked by `"<...>"`) and the other for the lexical meaning (marked only by `"..."`). Figure 2 shows examples of periphrasis (when several surface words, `"<tiene que ir>"`, are mapped into a unique lexical item, `"ir"`, plus the MODAL feature) and also portmanteau or amalgam (when a single word, `"<al>"`, represents two different lexical units, `"a"` and `"el"`).

The main sections of the specifications are dedicated to Spanish-specific phenomena such as the ”se”-constructions or clitics[2] (examples in Figures 3 and 4), as well as other relevant and frequent topics in corpora (idioms, date and hours, measures, etc.). Experience tells us that those elements show a great variation in patterns, and they need a unified treatment that is not always easy to define. In addition, they are very frequent in newspapers texts. An important effort in the project has been devoted to the definition of annotation guidelines for those topics, and still we are not satisfied with the treatment, since we have not been able yet to provide clear directions for some cases, leaving the final decision to the annotators.

```
(S
  (NP SUBJ MASC SG P3
    (N "<Manuel>" "Manuel" PROPER))
  (VP TENSED PRES MODAL SG P3
    (V "<tiene que ir>" "ir" TENSED PRES MODAL SG P3
      (AUX "tener_que" TENSED PRES SG P3)
      (V "ir" UNTENSED INFINITE))
    (PP A LOCATIVE
      (PREP "<al>" "a")
      (NP
        (ART "el" DEF MASC SG)
        (N "<dentista>" "dentista" MASC SG)))
    (NP TIME
      (ART "<el>" "el" DEF MASC SG)
      (N "<viernes>" "viernes" MASC SG))))
```

Figure 2: Multi-term and portmanteau: ”Manuel tiene que ir al dentista.” *Manuel has to go to the dentist.*

```
(S
  (NP * SUBJ SG P3)
  (VP TENSED PAST IND SG P3
    (NP OBJ2
      (P "<Se>" "le" PERS DAT SG P3))
    (NP OBJ1
      (P "<lo>" "lo" PERS ACUS SG P3))
    (V "<dio>" "dar" TENSED PAST IND SG P3))))
```

Figure 3: Pre-clitics: ”Se lo dio.” *Someone gave it to him/her.*

1.4. Tools

Although the corpus is basically hand-coded, the annotators make use of a combination of tools and resources either developed by the group or obtained from the public domain. These tools and resources can be divided into:

[2] Note that for postclitics, we use the same strategy as in parsing amalgams like ”al” or ”del”, where the surface string is only annotated in the main constituent (preposition or verb).

```
(VP UNTENSED INFINITE
  (V "<dárselo>" "dar"...  #CLITIC ID-1
    (NP
      (P "se" PERS P3 SG DISCONTINUOUS REF-1))
    (NP OBJ1
      (P "lo" PERS P3 SG DISCONTINUOUS REF-1))))
```

Figure 4: Post-clitics: ”Dárselo.” *To give it to someone.*

1. Annotation tools:

   (a) A **statistical POS tagger**, which provides the most frequent category and inflectional features for each word. This tool reduces substantially the effort of the annotator at the lexical level, and helps to control errors in feature assignment. This morphosyntactic tagger is described in Sánchez, Ramírez and Declerck (1999). As a lexical resource, we use a 50,000 lemma lexicon accessed by a generating inflectional morphological component. Disambiguation is performed by means of a reductionist grammar. We are currently reusing the grammar developed within the Constraint Grammar formalism (Karlsson et al., 1995). The approach to disambiguation favors recall rather than precision, so some disambiguation work is still left to the human post-editor, but the system rarely promotes an incorrect analysis.

   (b) A **chunker** that recognizes base NP, ADJP, VP, ADVP, and PP phrases. These phrases can be recursively identified, so nesting of phrases is allowed up to a given maximum level. The chunker is applied after post-editing the output from the annotation system.

2. Debugging tools:

   (a) A **graphical tree-drawer** for the annotated sentences. We use a publically released program, CLIG (Computational Linguistics Interactive Grapher), developed by K. Konrad, at the Department of Computational Linguistics in Saarbruecken, Germany[3]. It helps us to visualize the trees and to check errors. CLIG allows the definition of clickable tree nodes which favors the drawing of nodes with just basic information so as to allow for rapid inspection of constituent structure, and then, by clicking on the relevant node(s) check node feature values. A program is used to automatically produce the forest of CLIG objects from treebank native syntax.

   (b) A **feature checker** that verifies the proper assignment of features.

   (c) A **phrase structure rule generator**, which is used to detect possible incorrect annotations.

The debugging tools are used both for correcting wrong tagging during the annotation and for evaluating the results. In the former case, the graphical tree-drawer is used for inspecting the annotated sentence. CLIG not only shows the branches and the categories, but also the features for each node. The tree-drawer provides another way of approaching the corpus, very useful for the human annotator. Figure 6 shows a sample tree generated by CLIG.

The other tools are used for evaluation. The feature checker is a small program that shows the feature assignment in every tree, ordered by categories. It allows us to detect problems with respect to our feature specifications. In a early evaluation of the first 500 sentences, we could detect what the most common errors were (mainly lack of features and improper assignment of features) and which phrases are the most prone to error with respect to our feature annotation scheme (NPs and ADVPs). This information were used to improve the feature assingment in the next 1000 sentences.

Finally, the phrase structure rules generator is used to detect ”strange” combinations of constituents. This tool provides a different point of view to the coder, since it presents the results of the annotation. The PS rules generator has been useful for detecting some inconsistencies.

We are currently involved in a complete evaluation of the 1,500 sentences, whose results probably will suggest some changes in the guidelines.

2. The experiment in grammar rule induction

We have used the treebank to train a statistical parser, the ”Apple Pie Parser” (Sekine, 1995). This parser provides an efficient system for finding the most likely analysis, given a probabilistic context-free grammar and probabilistic information about the part-of-speech of lexical items. The parser also incorporates a number of special features, such as rule chunking, rule specialization for specific lexical items, and lexical semantic preferences, which are not currently used because of the small size of the current treebank.

We set aside forty sentences for testing, and used the remainder of the treebank to derive a stochastic context-free grammar, with the probabilities of productions based on the relative frequency with which a non-terminal expands to different structures in the treebank. Because the treebank is still too small to provide adequate lexical coverage, we used the MULTEXT Spanish Dictionary (as distributed by ELRA), converting its classification to be compatible with our treebank.

We estimate the probabilities of the different parts of speech for a lexical item based on its occurrence with these parts of speech in the treebank, and counting each part of speech assignment in MULTEXT as equivalent to a single instance in the treebank.

The text to be parsed is preprocessed to split off postclitics from infinitives and participles, and to split amalgams (”del” and ”al”) into their constituent preposition and article. In addition, idioms are identified and combined into single tokens.

[3] CLIG is a grapher for visualizing linguistic data structures and it is free for scientific purposes. The home page is `http://www.ags.uni-sb.de/ konrad/clig.html.`

For evaluation, the parses produced for the 40 test sentences were compared against the treebank, counting as correct only brackets for which the constituent label and the start and end words were correct. On this basis, recall was 73.6% and precision was 74.1%. While this is lower than results on large treebanks, it is an encouraging result for such a small training corpus and simple grammar creation strategy. These results are better than those Sekine (1998) reports for a comparable-sized English corpus, taken from the Wall Street Journal; this is probably related to the fact that our test sentences are, on average, shorter (16.2 words/sentence) than Wall Street Journal sentences (over 20 words/sentence). The accuracy of part-of-speech tagging (performed as part of parsing) was 96.3%; 45% of the sentences had no crossing brackets between parser output and treebank.

A manual review of some of the parser output suggests that the largest contributor to the parsing error rate is incorrect attachment of right modifiers (particularly since a single misplaced modifier can give rise to several incorrect brackets). These errors could best be corrected by learning lexical dependency patterns, either from a larger treebank or through separate training from unannotated corpora. Another significant source of error is incorrect part-of-speech assignment of lexical items not in the MULTEXT Dictionary (all such items are currently treated as nouns; the system does not presently incorporate any morphology statistics). Although we have not done a formal evaluation, it appears that restricting ourselves to the context-free skeleton in place of the richer feature structures provided by the treebank has played a lesser role in terms of the number of bracketing errors.

3. Future work

We are planning to conduct three different experiments:

1. To enlarge a fragment of the treebank only with categorial and lexical information, that is, supressing the other features, in order to test whether the richer information is really relevant for rule induction in Spanish in the newspapers domain.

2. To incorporate feature information selectively, e.g. agreement or syntactic functions, into the corpus-trained grammar in order to test the parser performance. In this case, experiments will be needed to determine just which features will be helpful in parsing. We are continuing to enlarge another fragment of the treebank only with this basic information. Once the parser is tuned, we plan to use it for further (partial) pre-annotation of the text in order to speed treebank creation.

3. To apply the current 1500 feature-rich annotated sentences to a process of deriving a probabilistic LFG grammar, as reported in van Genabith, Way and Sadler (1999).

On the other hand, we continue working on the guidelines, either for dealing with new phenomena or for refining the current directions.

With respect to the availability of the corpus, the idea is to put it in the public domain when the corpus reaches an acceptable size (about 5,000 sentences), it has been exhaustively debugged, and the annotation guidelines have been consolidated. For those interested in following the process, the project page is `http://www.lllf.uam.es/ sandoval/UAMTreebank.html`.

4. References

Bies, A., M. Ferguson, K. Katz, and R. Macintyre, 1995. *Bracketing Guidelines for Treebank II Style Penn Tree bank Project*.

EAGLES, 1996: *Preliminary Recommendations for the Syntactic Annotation of Corpora*.

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila, 1995. *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*. Berlin: Mouton de Gruyter.

Marcus, M., B. Santorini, and M.A. Marcinkiewicz, 1993. Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, 19 (2): 313-330.

Moreno, A., and S. López, 1999. Developing a Spanish Tree Bank. In *Proc. Journées ATALA, Corpus annotés pour la syntaxe*. Paris, 18-19 June 1999.

Moreno, A., S. López, and F. Sánchez, 1999. *Spanish Tree Bank: Specifications*. Laboratorio de Lingüística Informática, UAM. Version 4, 30 August 1999.

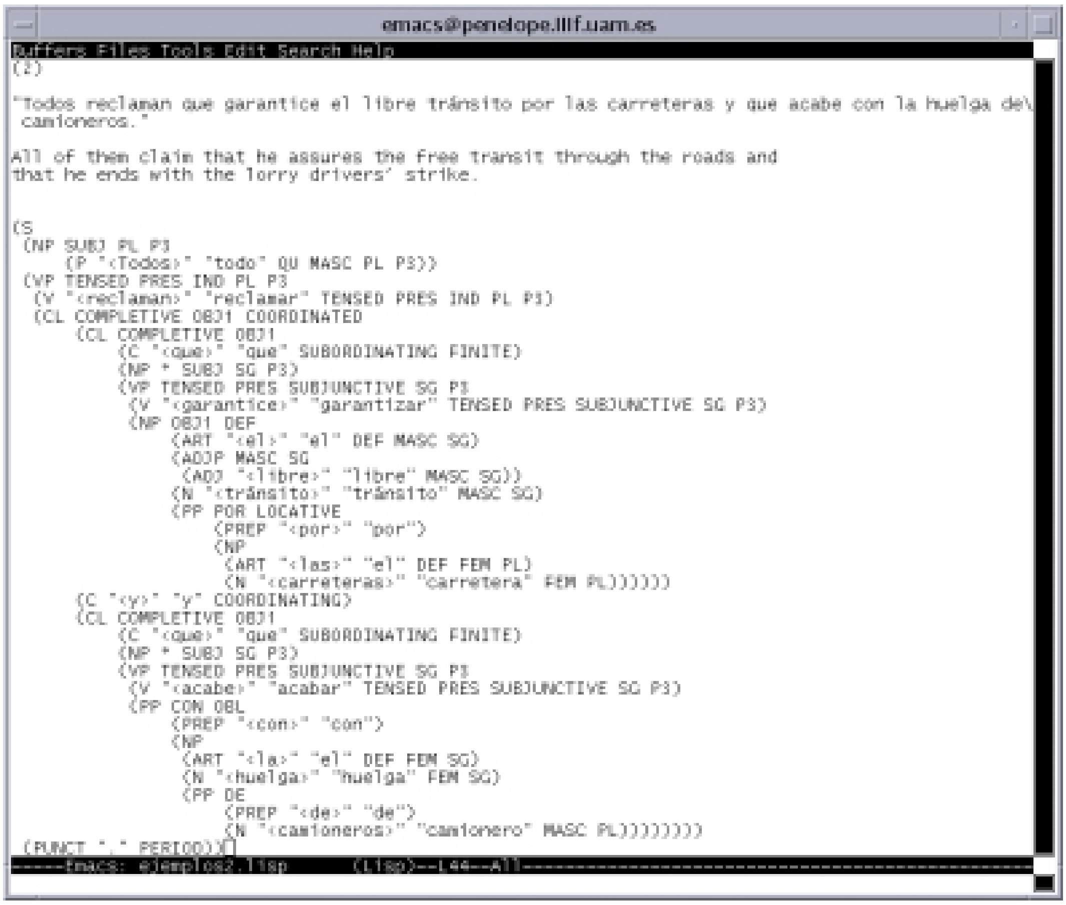
Sánchez, F., F. Ramírez and Th. Declerck, 1999. Integrated set of tools for robust text processing. In *Proc. of the VEXTAL Conference*. Venice, November 1999.

Sekine, S. 1995. A Corpus-based Probabilistic Grammar with Only Two Non-terminals. In *Proc. Fourth Int'l Workshop on Parsing Technologies*.

Sekine, S. 1998. *Corpus-based Parsing and Sublanguage Studies*. Ph.D. Dissertation, Department of Computer Science, New York University.

Skut, W., B. Krenn, T. Brants, and H. Uszkoreit, 1997. An Annotation Scheme for Free Word Order Languages. In *Proc. of the Fifth Conference on Applied Natural Language Processing (ANLP)*, Washington, D.C.

van Genabith, J., A. Way, and L. Sadler, 1999. Semi-automatic generation of f-structures from Tree Banks. In *Proc. of the LFG99 Conference*. Stanford, CSLI.


emacs@penelope.lllf.uam.es
Buffers Files Tools Edit Search Help

```
(2)

"Todos reclaman que garantice el libre tránsito por las carreteras y que acabe con la huelga de\
 camioneros."

All of them claim that he assures the free transit through the roads and
that he ends with the lorry drivers' strike.


(S
 (NP SUBJ PL P3
      (P "<Todos>" "todo" QU MASC PL P3))
 (VP TENSED PRES IND PL P3
  (V "<reclaman>" "reclamar" TENSED PRES IND PL P3)
  (CL COMPLETIVE OBJ1 COORDINATED
      (CL COMPLETIVE OBJ1
          (C "<que>" "que" SUBORDINATING FINITE)
          (NP * SUBJ SG P3)
          (VP TENSED PRES SUBJUNCTIVE SG P3
           (V "<garantice>" "garantizar" TENSED PRES SUBJUNCTIVE SG P3)
           (NP OBJ1 DEF
               (ART "<el>" "el" DEF MASC SG)
               (ADJP MASC SG
                (ADJ "<libre>" "libre" MASC SG))
               (N "<tránsito>" "tránsito" MASC SG)
               (PP POR LOCATIVE
                   (PREP "<por>" "por")
                   (NP
                    (ART "<las>" "el" DEF FEM PL)
                    (N "<carreteras>" "carretera" FEM PL))))))
      (C "<y>" "y" COORDINATING)
      (CL COMPLETIVE OBJ1
          (C "<que>" "que" SUBORDINATING FINITE)
          (NP * SUBJ SG P3)
          (VP TENSED PRES SUBJUNCTIVE SG P3
           (V "<acabe>" "acabar" TENSED PRES SUBJUNCTIVE SG P3)
           (PP CON OBL
               (PREP "<con>" "con")
               (NP
                (ART "<la>" "el" DEF FEM SG)
                (N "<huelga>" "huelga" FEM SG)
                (PP DE
                    (PREP "<de>" "de")
                    (N "<camioneros>" "camionero" MASC PL))))))))
 (PUNCT "." PERIOD))
-----Emacs: ejemplos2.lisp      (Lisp)--L44--All-------------------------------------------
```

Figure 5: Coordination and ellipsis

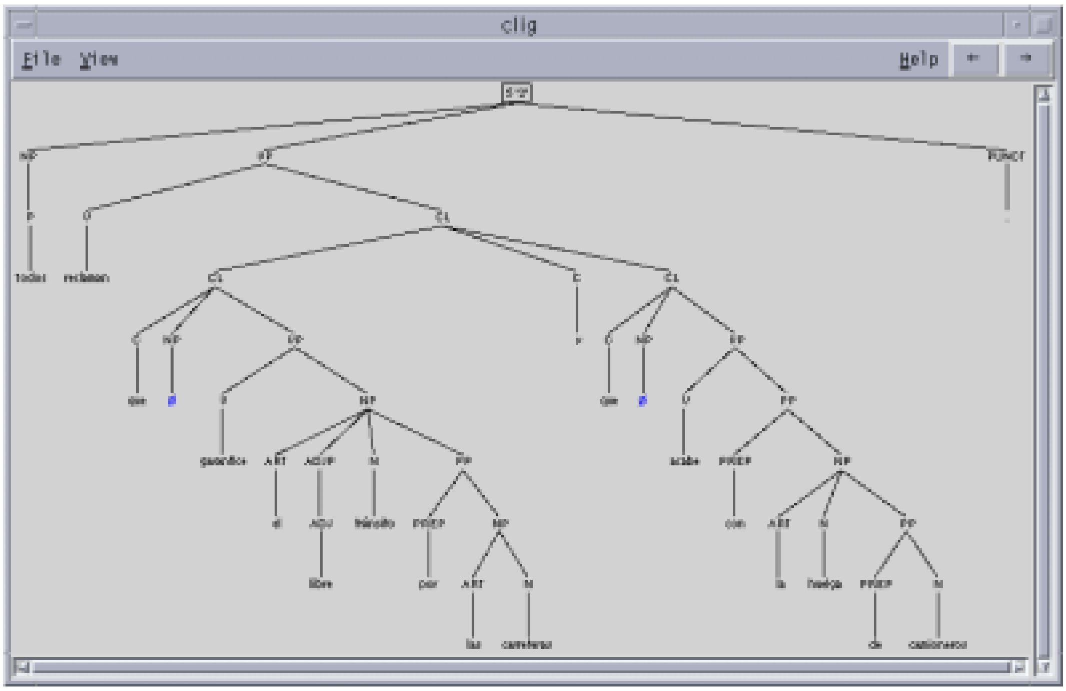


Figure 6: Sample tree of the above sentence generated by CLIG