CINTIL TreeBank Handbook: Design options for the representation of syntactic constituency

António Branco, João Silva, Francisco Costa and Sérgio Castro

DI–FCUL–TR–2011–02

DOI:10455/6746

(http://hdl.handle.net/10455/6746)

July 2011

CINTIL TreeBank Handbook:

Design options for the representation of syntactic constituency

António Branco, João Silva, Francisco Costa and Sérgio Castro

University of Lisbon

January 2011

1 Introduction

Treebanks are data sets of utmost importance for the study of natural languages and for their computational processing. They permit the training and evaluation of different processing tools, including taggers, chunkers, parsers, deep linguistic grammars, etc.

A treebank is an annotated corpus. It is a data set consisting of a collection of individual writen utterances associated to the representation of their linguistic structure, which can be set to capture different degrees of linguistic information.

CINTIL Treebank is a corpus of Portuguese utterances annotated with the representation of constituency relations. It is being developed and maintained at the University of Lisbon.

This document aims at supporting the utilization and exploitation of the CINTIL Treebank. It presents its major design options in what concerns the representation of syntactic relations.

The adopted design options were informed by advanced linguistic theorizing. The reader is referred to the literature for a thorough discussion and justification of them.

For the source of the utterances in this corpus, for its composition and for the annotation methodology used see (Barreto *et al.*, 2006).

The CINTIL Treebank has two versions. There is a reference version for human users, and there is a variant for training probabilistic parsers. Where the latter differs from the reference version, that is indicated below by text between square brackets starting by "Prob Parser:".

1.1 Concordancer

The CINTIL DepBank can be searched through a concordancer online at http://lxcenter.di.fc.ul.pt/services/en/LXServicesSearcher.html

The example graphs displayed below are associated to its identifier in the corpus. These sentences can be recovered in this concordancer with these identifiers.

2 Constituency relations

2.1 constituency in a nutshell

In a sequence of lexemes w1 w2 w3, if the subsequence w1 w2 has a higher level of aggregation than the subsequences w1 w2 w3 or w2 w3, the sequence w1 w2 is considered to form a constituent of w1 w2 w3, of which w1 and w2 are themselves constituents.

The contrasting levels of aggregations are determined through the application of empirical tests which rely on grammatical intuitions or judgments on syntactic well-formedness. These empirical tests are based on judiciously designed minimal pairs of sequences. To test a putative constituent, these minimal pairs are constructed, for instance, by means of the insertion of a parenthetical element inside it, by displacing it to a non canonical word order in the sentence, by replacing it by an anaphoric expression, or by coordinating it with other known constituents, etc.

A constituent is represented by enclosing the relevant sequence in square brackets (e.g. [w1 w2] w3), or in an alternative, but equivalent notation, by forming a one level depth tree whose leaves are w1 and w2 and the top node stands for the whole constituent.

A syntactic category is a set of constituents with identical syntactic distribution, that is constituents whose replacement by each other preserves the syntactic well-formedeness of larger expressions they are constituents of (provided some other key grammatical relations are not affected by that replacement, such that morphological agreement, subcategorization, etc.).

The categorization of constituents is represented by decorating the nodes of the constituency trees with tags signaling the appropriate categories. These tags are usually acronyms of the categories they correspond to. For instance, NP stands for Noun Phrase, S for Sentence, etc. See section "4. Tag Set" below for the list of categories in use in the treebank.

2.2 minimal constituents

A lexeme is a terminal node and its category is represented in the immediately dominating, pre-terminal node. They form a unary branching tree.

2.3 syntactic predication

The constituency relations are intertwined with other grammatical relations, determining and being determined by them. Syntactic predication is one such relation of interest.

A syntactic predication is organized around a predicate and its complements, possibly extended with modifiers and specifiers.

To integrate a well formed utterance, a predicate requires that a number of other expressions (zero or more), of certain syntactic category or categories, co-occur with it. A predicate and its complements form a constituent.

2.4 head

Lexemes of categories N, V, A, P, ADV, CONJ and C may be syntactic predicates.

A syntactic predicate of category *X* is a special constituent (termed head) of their phrase, of category *X*P. In that constituency tree, the path from *X* to *X*P only contains (zero or more) intermediate nodes of category *X*'. That node *X*P, as well as the intermediate nodes *X*', are said to be projected by that head *X*.

In general, the head of a phrase XP is its single constituent of lexical category *X*, thus immediately dominated by a pre-terminal *X* node, except for multi-words, whose individual items project several pre-terminals *X*s immediately dominated by a node *X'*.

In the treebank, in general, for major categories, a head *X* is represented as projecting an *X*P when this is a constituent having complements or modifiers of *X* as subconstituents (see also section "7.3 Comparatives").

Given their specific or ambivalent nature in categorial terms, this schema is adapted for the following items:

2.4.1 personal pronouns

A personal pronoun has category PRS. It is the head of an NP.

2.4.2 clitic pronouns

A clitic pronoun has category CL. It is the head of an NP.

2.4.3 participles

A past participle has category V. It is the head of an AP in attributive and predicative constructions.

2.5 complements

A complement of a predicate *X* is a constituent of the projected *X*P, immediately dominated by *X*P or by an intermediate *X*'. Such nodes of the phrase are said to be (internal) complements of the head.

Given its specific nature, verbal predicates may also have an external complement, not occuring inside the VP they project (see also section "5.1 Sentences").

Given its specific nature, nominal heads project an NP even when no complement exists or is realized (see also section "5.2 Nominals" for more details on NPs).

Complements are of the following categories: NP, PP, AP, ADVP, CP (see also section "7.3 Comparatives").

2.6 specifiers

Inside the NPs, besides the head, complements and modifiers, other expressions may occur, that are specifiers.

A specifier of an NP projected by a head N is a constituent of that NP, immediately dominated by it or by an intermediate N', provided all other dominating N's are also dominated specifiers (see also section "5.2 Nominals" for more details on NPs).

Specifiers are of the following categories: QNT, ART, D, DEM, POSS, CARD.

2.6.1 bare NPs

Given the key semantic function of specifiers, it is considered that NPs without a phonetically realized specifier (bare NPs) still undergo some process of specification. As a result, the NP node of bare NPs has a unary branch to the immediately dominated node.

The exception is to be found in Proper Names that modify a common noun, whose category is N' if it is a multi-word proper name (e.g. *o actor Artur Semedo*), else is N (e.g. *o rio Jadar*).

2.7 modifiers

The event described by a predicate and its complements can be further characterized by other cooccurring lexemes or phrases, that are modifiers.

A modifier *Y* is in an adjunction position to the modified constituent *Z*, that is it is a sister node of that *Z*, and both are dominated by a node also of category *Z*.

Modifiers are of the following categories: ORD, ADV, ADVP, A, AP, PP, CONJP, NP, CP, VP.

3 Non-constituency relations

Trees are aimed at depicting constituency relations. In the CINTIL treebank, they are further decorated with information tags relevant also for two types of grammatical relations that are of a non-constituency nature, namely grammatical dependency relations and semantic role relations.

Such information tags encode, respectively, grammatical functions and semantic functions of the corresponding nodes. They are displayed in accordance to the pattern *Z-GF-SF*, where *Z* is a constituency category, *GF* is a grammatical function, and *SF* is a semantic function (e.g. NP-SJ-ARG1) in in nodes where this applies.

A grammatical function results from an abstraction over complements and modifiers of different predicates. It permits to categorize complements, or modifiers, with similar syntactic constraints on their realization, such as category, case, agreement, canonical word order, inflection paradigm, etc.

A semantic function, or semantic role, is also an abstraction over complements and modifiers of various syntactic predicates, but along a different, semantic, dimension. It permits to categorize complements, or modifiers, according to similar semantic constraints on their denotation, that is in terms of the similar contribution that the extra-linguistic elements they may denote bring for the characterization of the event being described. Given the semantic roles are much more elusive than grammatical functions, following common practice with respect to the creation of PropBanks (see also section 3.2 below on the CINTIL PropBank), the option here was to adopt a set of roles for complements that primarily permits to semantically distinguish complements of the same predicate among each other.

The possible values of grammatical functions are listed in section 4.3 and for semantic functions are listed in section 4.4.

3.1 CINTIL DepBank

Grammatical functions are a necessary but not sufficient element to characterize grammatical dependencies. Grammatical dependency relations can be depicted as graphs whose nodes are words and whose directed arcs establish a connection from a lexeme to its subordinate lexemes.

An arc represents the dependency of the subordinate item to the head. These dependencies can be of a number of different types, which are mostly the grammatical functions, and with whose tags the arcs are decorated.

Corpora annotated with grammatical dependency graphs are known as Dependency Banks. The CINTIL Treebank is aligned to a dependency bank, the CINTIL DepBank. The bridging elements are the grammatical function tags decorating the nodes, in the treebank, and the arcs, in the dependency bank.

For the Handbook of the CINTIL DepBank see (Branco *et al.*, 2011).

3.2 CINTIL PropBank

Treebanks encoding constituency relations which are extended to encode also semantic functions, or semantic roles, of elements of syntactic predications have been termed as PropBanks in the literature. Given the nodes of the CINTIL Treebank are decorated with semantic functions, this annotated corpora can be taken as being also the CINTIL PropBank.

It is worth noting that in so called PropBanks, the semantic relation signaled by the tag on a given constituent indicates a semantic relation between that constituent and a predicator in the utterance. Hence, that relation being signaled over a single constituent is not fully identified in an explicit way as one of the terms is not indicated.

Nonetheless, usually the relevant predicate is the closest predicate in the tree, which belongs to the same minimal predication as the tagged constituent does.

The cases where this does not hold are typical cases of complex predicates, formed by means of several chained verbs, e.g. modals, auxiliaries and raising verbs. In such cases the tag used to code the semantic function is sufixed with "cp" (standing for "complex predicate") in order to help the search and concordancing of the treebank (for more details see the section 4.4 below)

For an annotated corpus with fully fledged representation of semantic relations, see the CINTILLogicalFormBank.

4 Tag set

4.1 lexical categories

A	Adjective
ADV	Adverb
ART	Article
C	Complementizer
CARD	Cardinal
CL	Clitic
CONJ	Conjunction

D	Determiner
DEM	Demonstrative
ITJ	Interjection
N	Noun
ORD	Ordinal
P	Preposition
PERCENT	Percentage
PNT	Punctuation
POSS	Possessive
PRS	Personal pronoun
QNT	Quantifier
REL	Relative pronoun
V	Verb

4.2 non-lexical categories

A'	Adjective sub-phrase constituent
ADV'	Adverb sub-phrase constituent
ADVP	Adverb phrase
AP	Adjective phrase
CARD'	Cardinal sub-phrase constituent
CONJ'	Conjunction sub-phrase constituent
CONJP	Cardinal sub-phrase constituent
CP	Complementizer phrase
ITJ	Interjection
N'	Nominal sub-phrase constituent
NP	Noun phrase
PERCENTP	Percentage phrase
POSS'	Possessive sub-phrase constituent
PP	Preposition phrase
QNT'	Quantifier sub-phrase constituent

S	Sentence
V'	Verb sub-phrase constituent
VP	Verb phrase

4.3 grammatical functions

SJ	Subject
SJac	Subject of an anticausative
SJcp	Subject of complex predicate
DO	Direct Object
IO	Indirect Object
OBL	Oblique Object
M	Modifier
PRD	Predicate
C	Complement
SP	Specifier

4.4 semantic functions

ARG1	Argument 1
ARG11	Argument 1 of subordinating predicator and Argument 1 of the subordinate clause (semantic function of Subjects of so called Subject Control predicators)
ARG21	Argument 2 of subordinating predicator and Argument 1 of the subordinate clause (semantic function of Subjects in so called Direct Object Control predicators)
ARG31	Argument 3 of subordinating predicator and Argument 1 of the subordinate clause (semantic function of Subjects in so called Indirect Object Control predicators)
ARG2	Argument 2
ARG3	Argument 3
ARG*n*cp	Argument *n* in complex predicate constructions
ARG*n*ac	Argument *n* of anticausative readings
LOC	Location
EXT	Extension
ADV	Adverbial

CAU	Cause
TMP	Temporal
PNC	Purpose, goal
MNR	Manner
DIR	Direction
PRD	Predication
POV	Point of view
M	Modifiers (other cases)

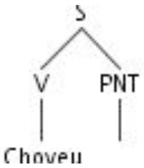
5 Specific phrases

Phrases of category S and NP have specific constituency format (see also section 7.3 on comparatives).

5.1 Sentences

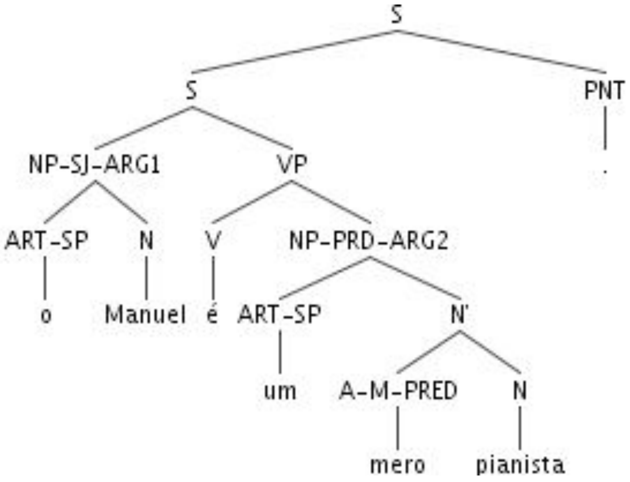
Phrases of category S have no S' daugthers or S-categorial head.

An S is projected out of a V in case this is an impersonal or intransitive verb, or a verb with no realized complement other than the Subject, with no modifiers:



#Id:a001/1[1]

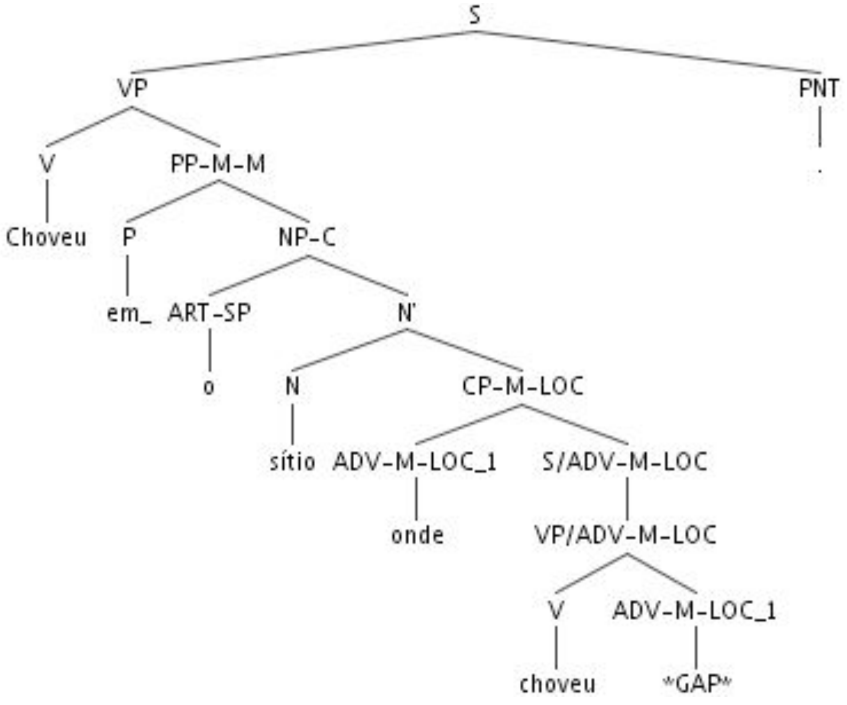
An S is projected out of a VP in case this VP's head has an internal complement:



[1] This is the identifier code of this sentence in the corpus.

#Id:b001/42

Or is modified:

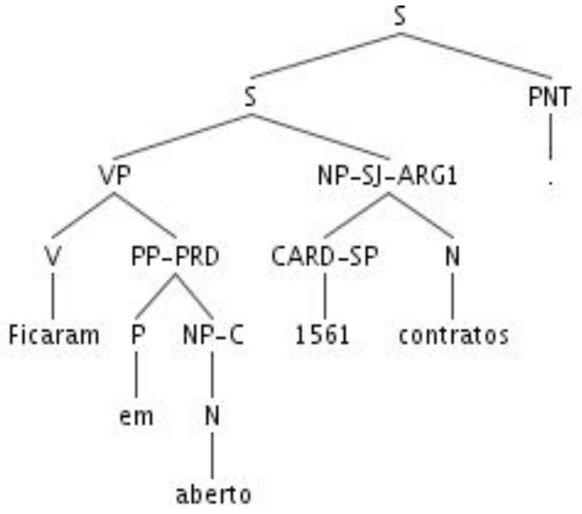


#Id:a012/569

For S not projected out of a verbal head, see section 10.2 on non verbal utterances.

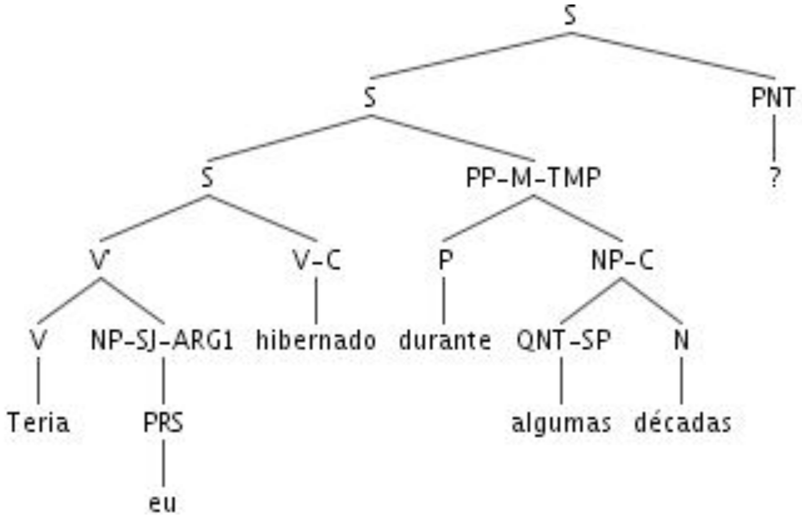
In canonical SVO word order, the Subject is immediately dominated by S and a sister node of the projecting V or VP.

In VOS word order, extraposed Subjects are sister nodes of its V or VP and immediately dominated by S:



#Id:b051/3036

In VSO order, when the extraposed Subject intervenes between the Verb and its internal complement, the extraposed Subject is dominated by V':



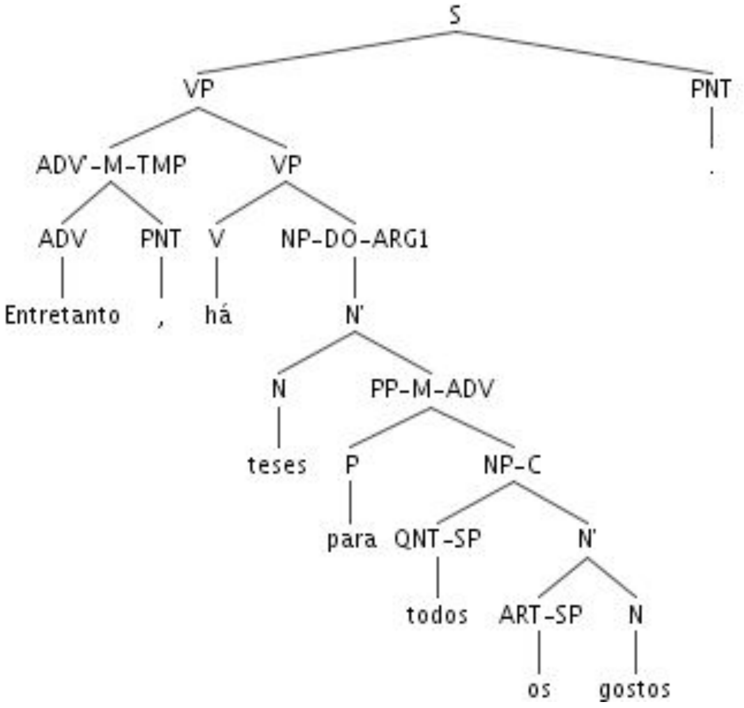
#Id:b012/782

Quantifiers floating to a post-verbal position, as in *Os jogadores viram todos isso*, are in adjunction to a projection of the verb. Those floating to an immediatley pos-nominal position, as in *Os jogadores todos viram isso*, are in adjunction position to their NP (see example #Id:b092/5911, in section 6.1 below).

5.2 Nominal

Phrases of category NP may have specifier daugthers. In general, these are left-branching nodes.

Bare NPs, with no realized specifier, are characterized by having a unary branch immediately below the NP node projected by its nominal head:



#Id:e000369/56038

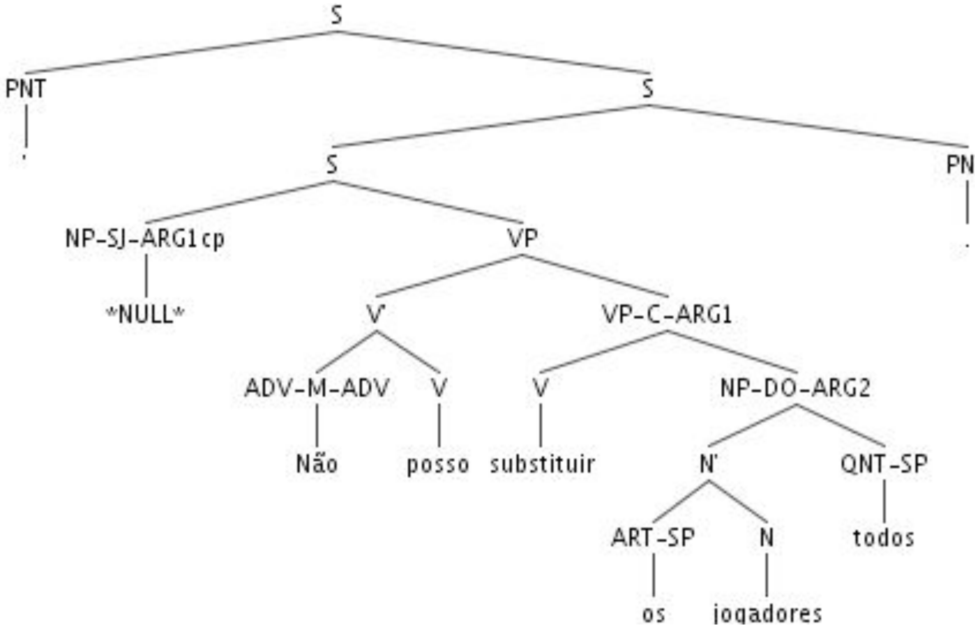
6 Phonetically null items

Phonetically null items mark positions in the tree related to other positons in the tree (in case of traces), or mark ellided elements whose context is rich enough to support the recovery of their interpretation (in case of null subjects or null heads).

[Prob Parser: Phonetically null items are removed from the tree and represented by means of appropriated tree configuration or appropriate relabelling of the relevant nodes.]

6.1 null subjects

Null subjects are marked by *NULL* and are immediately dominated by the S node projected by a V or a VP:

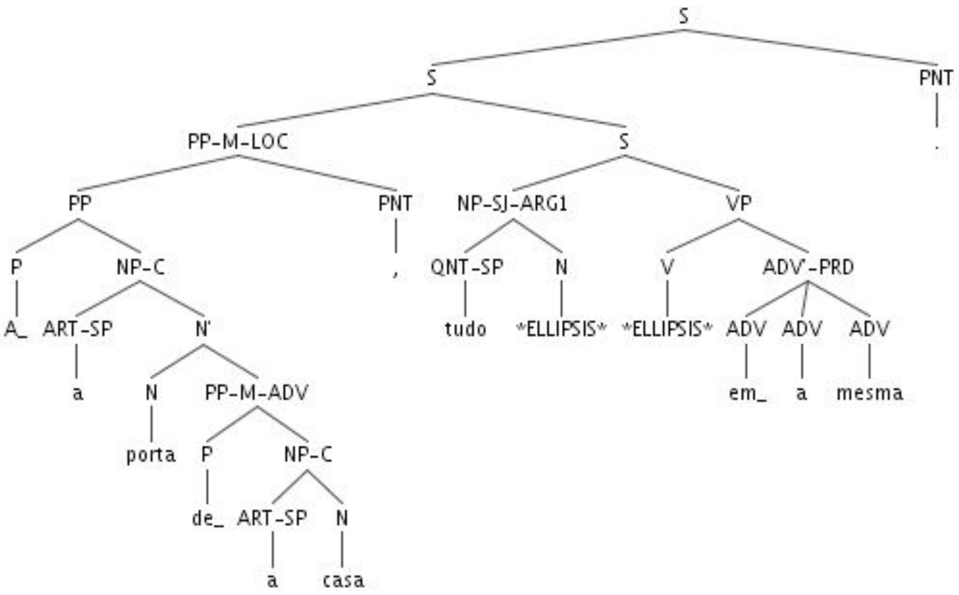


#Id:b092/5911

[Prob Parser: Null subjects are represented by unary branching below S.]

6.2 null heads

Null heads may be nominal or verbal. They are marked by *ELLIPSIS*:

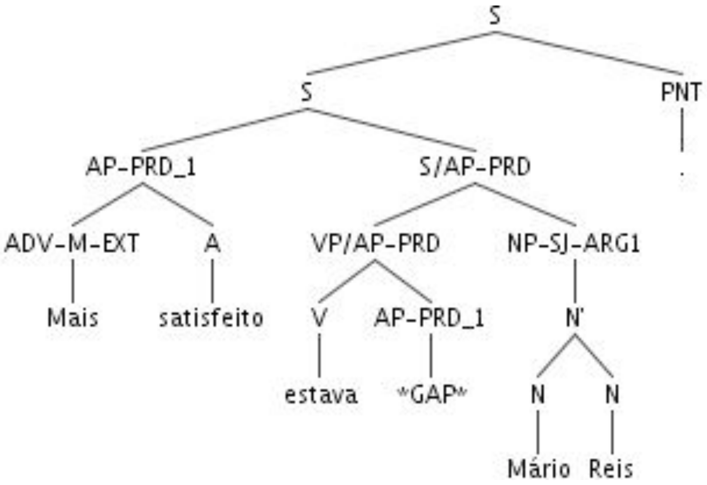


#Id:b001/11

[Prob Parser: Null heads are represented by *XP* or *X′* with no descendant *X*.]

6.3 traces

Traces of constituents that are marked by *GAP* followed by _*n* where *n* is a natural number. The category of the "displaced" node is coindexed with the trace and thus also followed by _*n*:

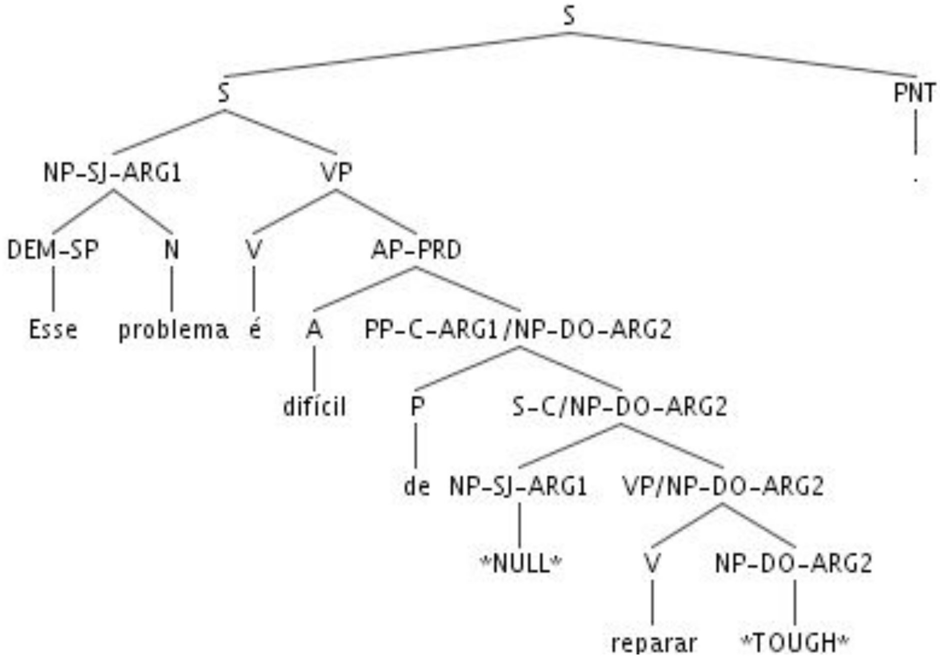


#Id:b094/6024

[Prob Parser: Any node *Z* in the tree path between the trace and the gap filler of category *W* is relabeled as *Z*/*W*.]

6.4 "though" null objects

Null direct objects specifically licensed by "though" constructions are marked by *THOUGH*.



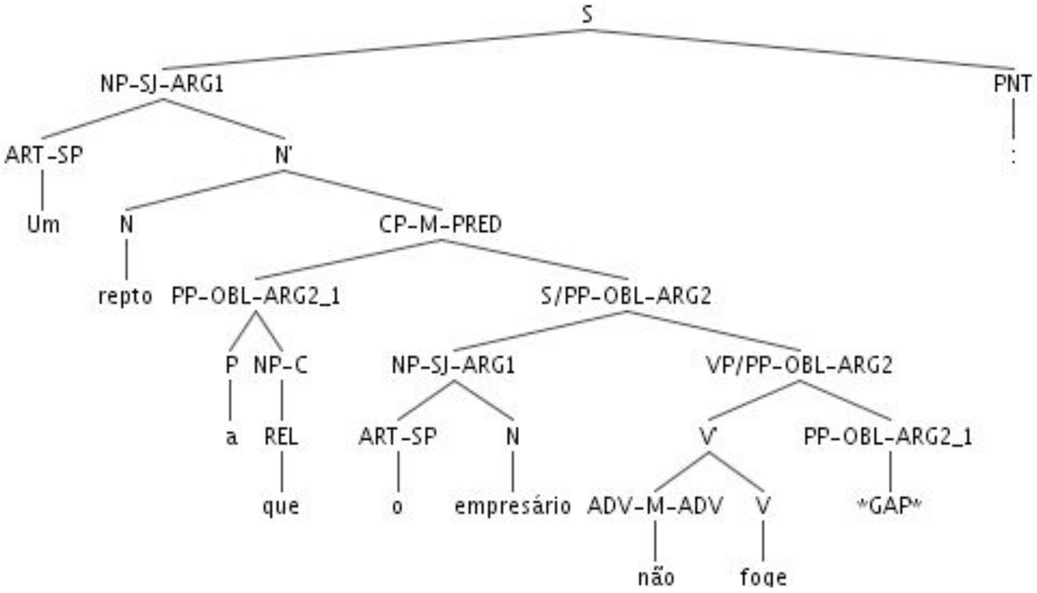
#Id:a012/591

[Prob Parser: "Though" objects are represented by unary branching below VP.]

7 Specific constructions

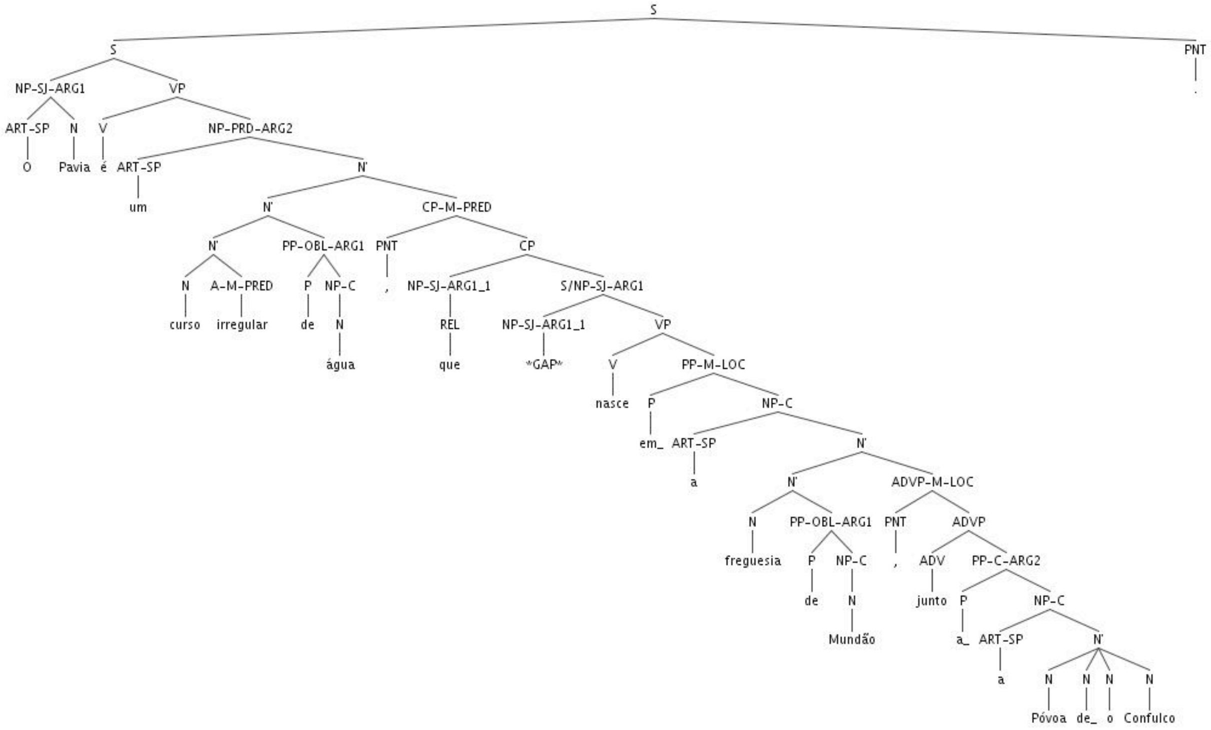
7.1 relatives

A modifying relative clause is dominated by N'. It is of category CP, with two immediate constituents, an *X*P projected by a relative pronoun REL and a clause S. It has grammatical function M and semantic role PRED:



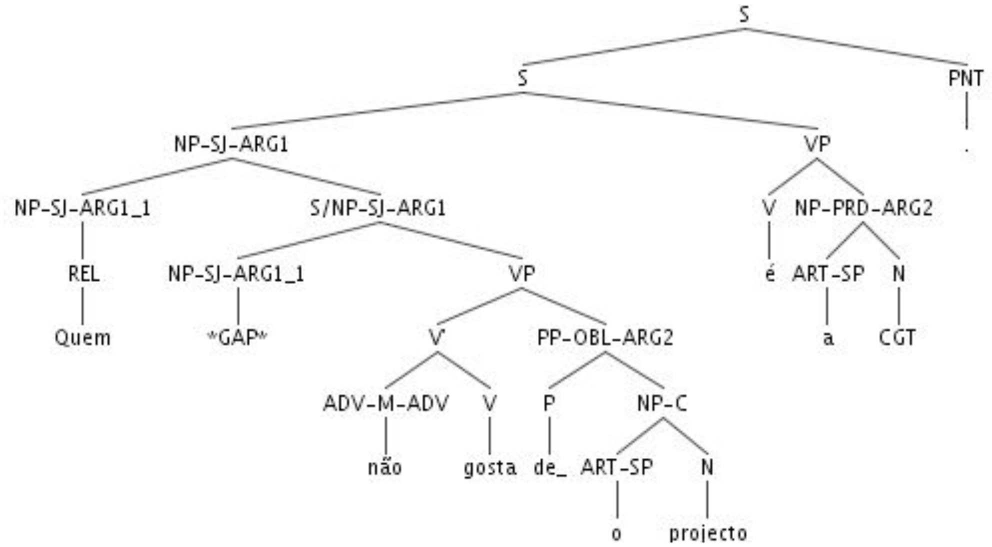
#Id:b080/4682

An appositive relative clause is preceded by a comma ',', which is in adjunction to the CP constituent and forms another CP constituent with it.



#Id:b219/16012

A free relative clause is of category NP, which is projected out of the relative:

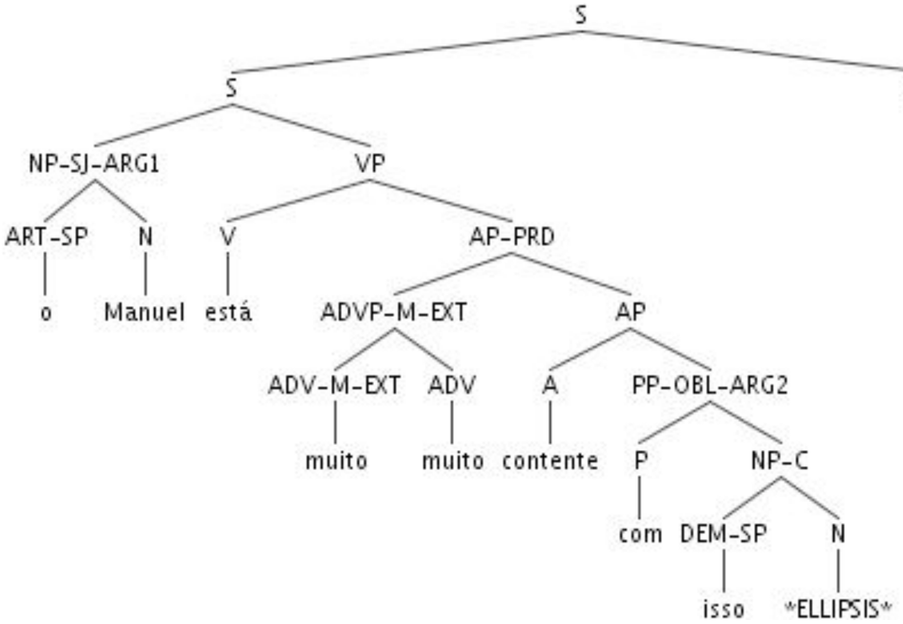


#Id:e000149/38916

For the representation of the phonetically null trace, in correspondence to the relativizer *X*P, see section 6.3 on Traces (see also section "8 Long-distance relations").

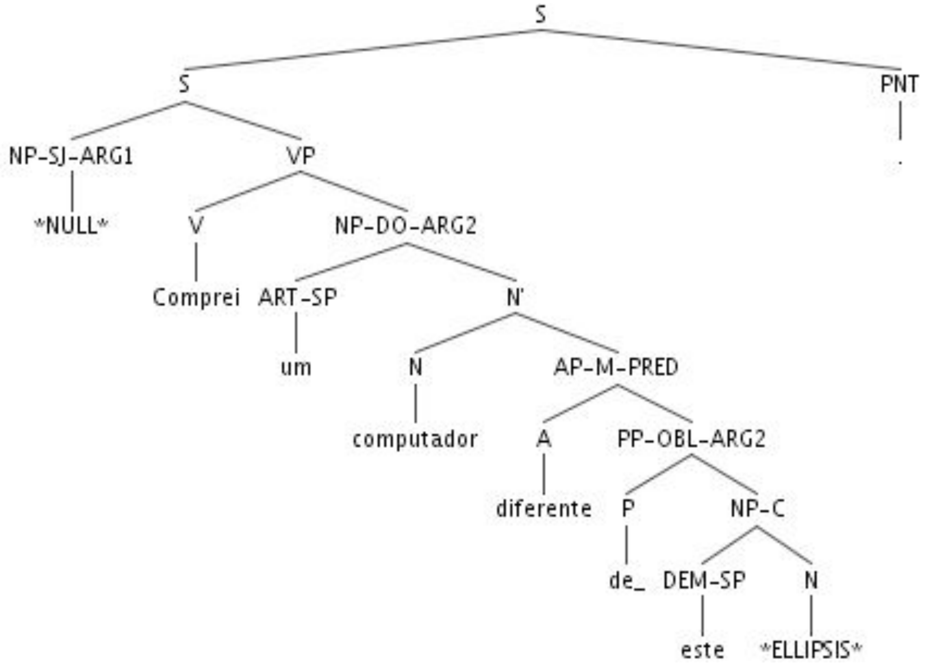
7.2 adjectives: predicative and attributive

In predicative constructions, the Subject is ARG1 of the copula verb, and the corrsponding logical form it shows up as ARG1 of the adjective.



#id:a003/102

The same holds with respect the head noun in attributive constructions. That is the case of the noun *diferente* in this example:

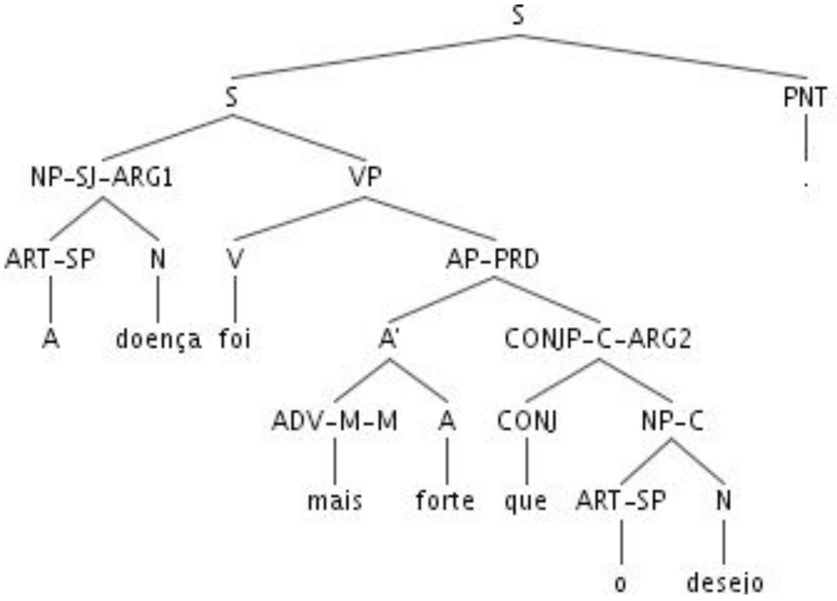


#id:a001/34

Accordingly, any further arguments of the adjective surfacing in the tree are tagged with ARG*n* with *n*>=2. That is the case of *deste* in the example above.

7.3 comparatives

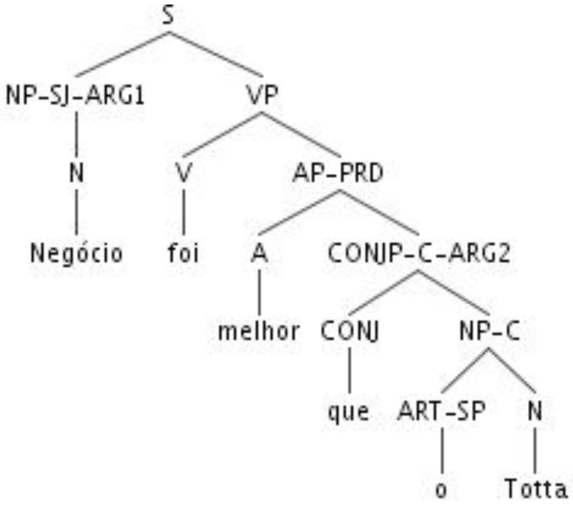
A comparative construction is typically built around an adjective by two constituents, an adverbial of degree and a CONJP phrase:



#Id:e000282/49262

(some adverbs may also support comparative constructions, as with *perto* in the exampel *mais perto do que a Maria*)

The exception happens with adjectives like *maior*, *menor*, *melhor*, *pior*, which also express the comparison, in which case the comparative construction is built around the adjective and the CONJP phrase.



#Id:e000481/64969

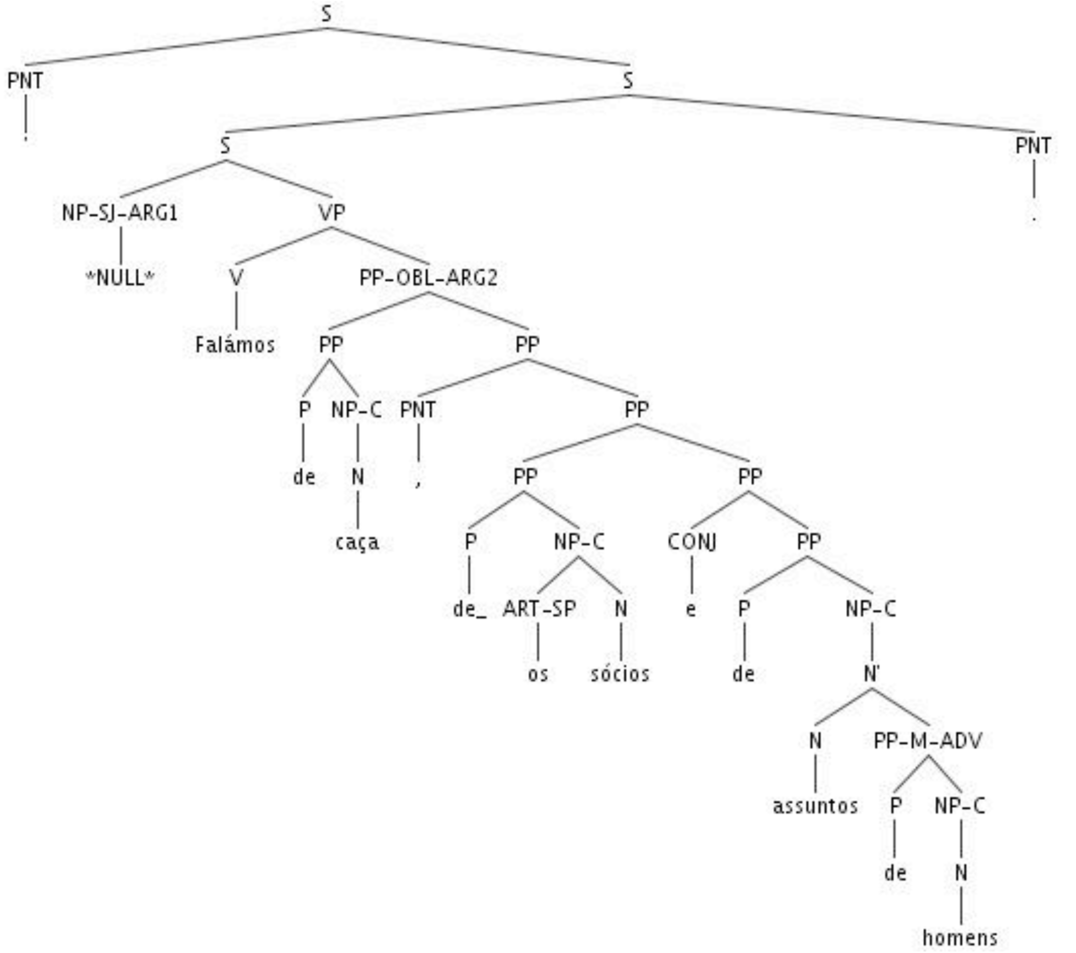
The adverbial of degree (e.g. *mais*, *menos*, *tão*) is sister of the adjective, dominated by an A' node. It is superficially tagged as A-M-M, that is as modifier, but note that in logical form the adjective shows up as the ARG1 of this adverb.

The CONJP phrase is a sister node of that node A'. It is projected by one of the conjunction expressions for comparatives *que*, *de que*, *de_ o que*, *como*, *quanto*. It is a complement of the adverbial of degree. Hence this adverbial happen not to project an ADVP. This phrase is tagged as CONJP-C-ARG2, indicating that it is the complement and ARG2 of the adverb.

The CONJP may be absent of the comparative construction. In such case, though it can be semantically recovered from the context, there is no phonetically null item inserted in the tree to mark it.

7.4 coordination

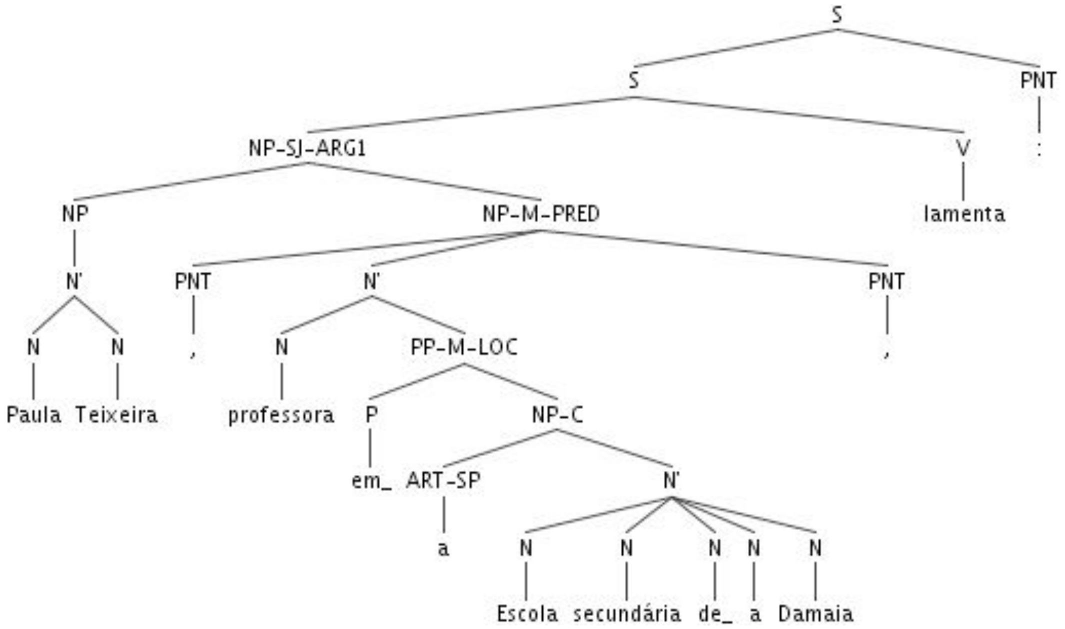
Coordination of two constituents A and B by means of a coordinative conjunction Conj (either a lexical item, such as *e*, or a comma) are a cascade of adjunctions [A [Conj [B]]].



#Id:b001/30

7.5 appositions

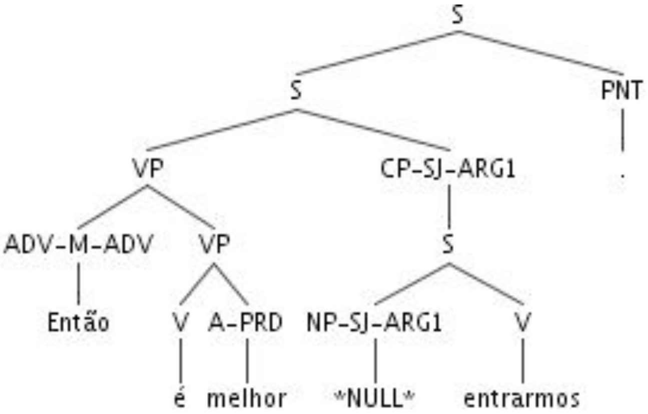
Appositions are adjoined to NPs:



#Id:b005/254

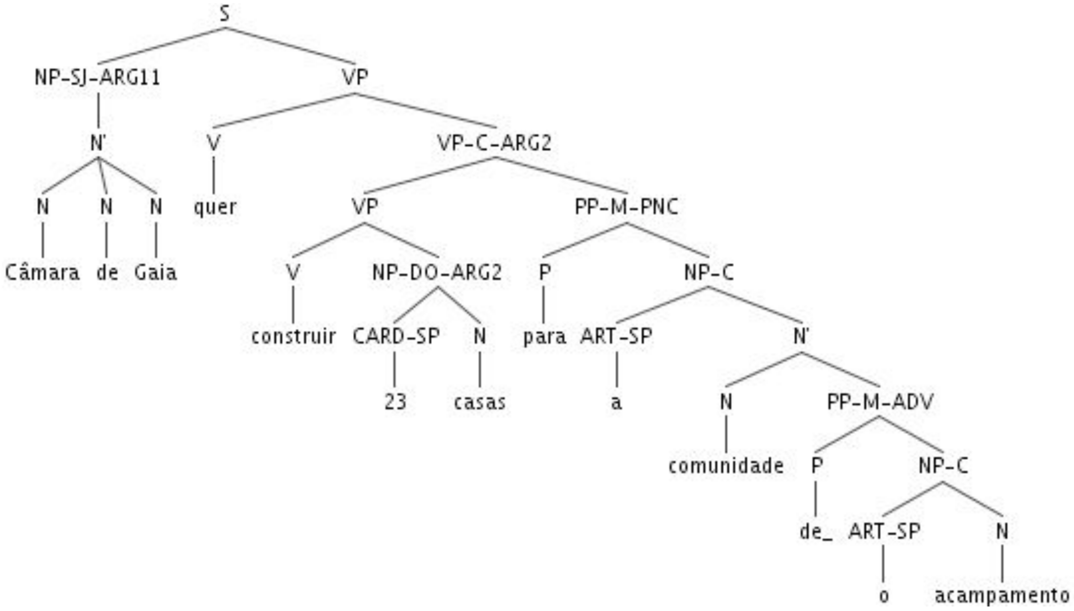
7.6 infinitives: inflected and non inflected

An inflected infinitive projects an S:



#Id:c031/23222

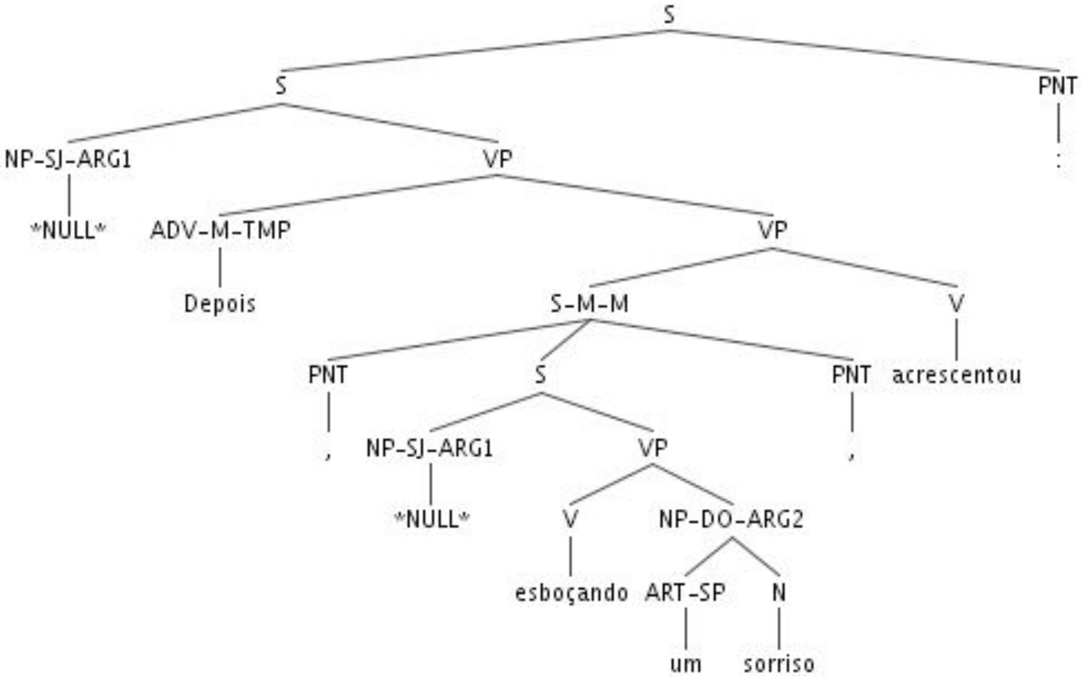
And a non inflected one projects a VP:



#Id:b076/4469

7.7 gerunds

When in complex predicate constructions, preceded by an auxiliary verb, a gerund projects a VP. Otherwise, a gerund projects an adverbial sentence with a null subject:



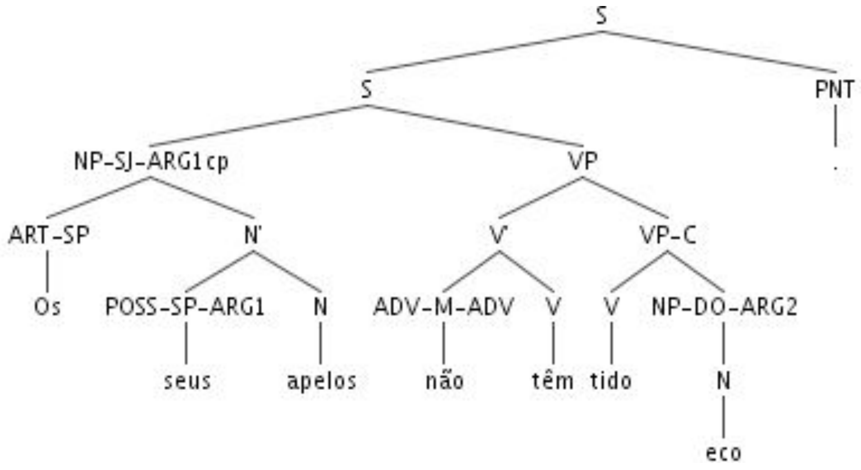
#Id:c020/22209

7.8 complex predicates: auxiliary, raising and modal verbs

Auxiliary, modal and raising verbs select for a VP.

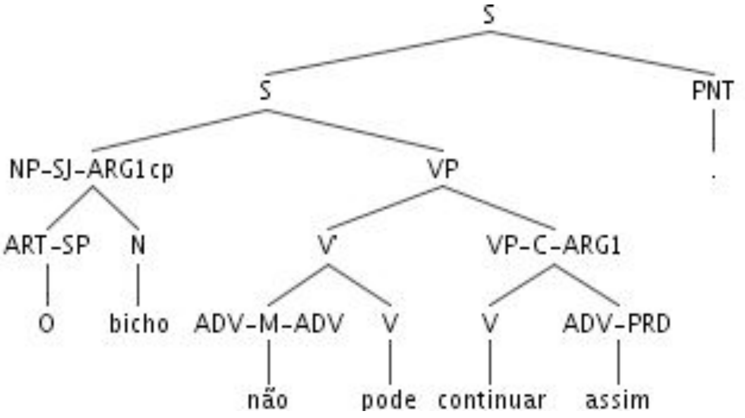
Different status of this VP is signaled by its grammatical function and/or semantic role.

Auxiliaries syntactically select for a complement VP, and are thus sister nodes of VP-C:



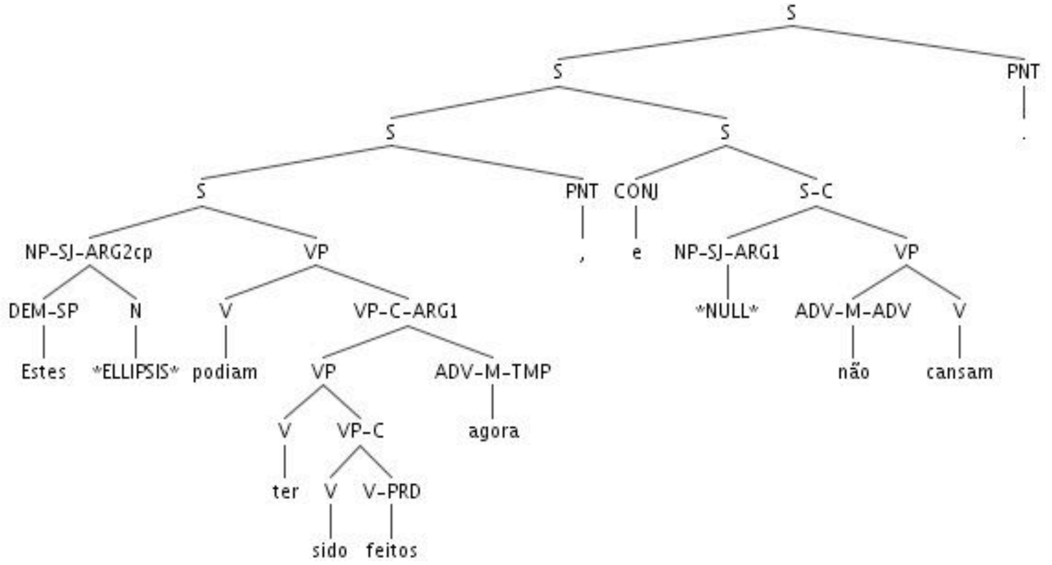
#Id:e000585/73183

Modal and raising verbs select not only syntactically but also semantically for a complement VP. They are thus sister nodes of VP-C-ARG1:



#Id:c003/20534

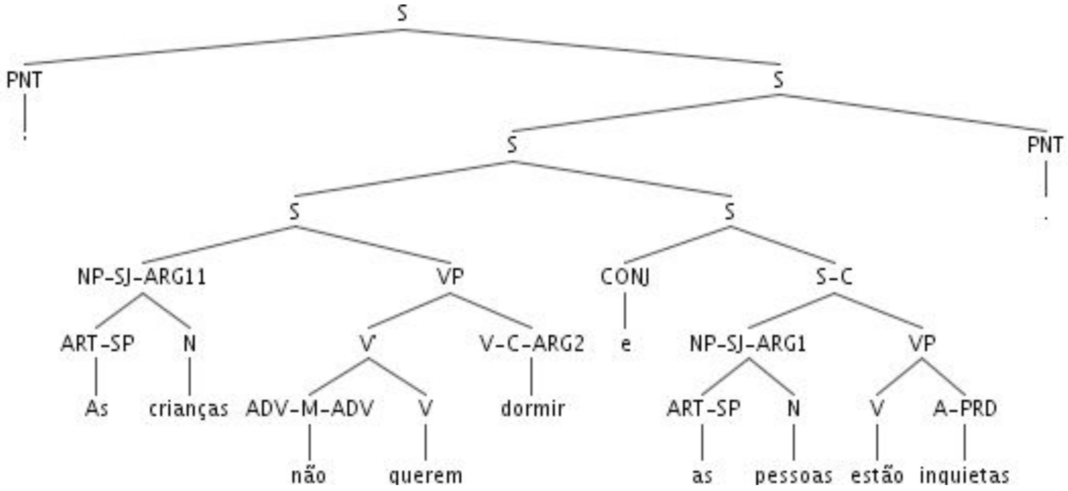
In a complex predicate, formed by any sequence of auxiliary, raising or modal verbs, its Subject is marked as NP-SJ-ARG*n*cp, signaling that it is the Subject of the topmost verb (viz. -SJ-) and the ARG*n* of some verb down below in the complex predicate:



#Id:b134/8372

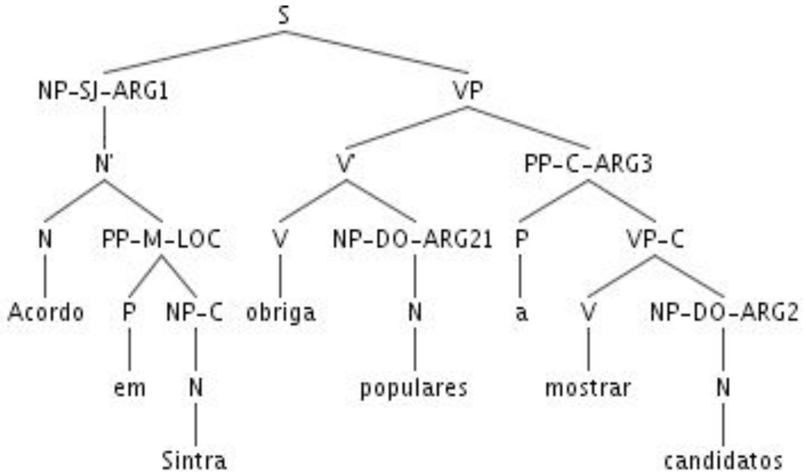
7.9 control verbs

Subject control verbs (e.g. *querer*) select for a Subject NP-SJ-ARG11, signaling that it is both the subject of the control verb and in the clause occurring as direct object of the later:



#Id:b001/34

Object control verbs (e.g. *obrigar*) select for a Direct Object NP-DO-ARG21, signaling that it is both the object of the control verb and the subject in the clause occurring as the other internal argument of the control verb.

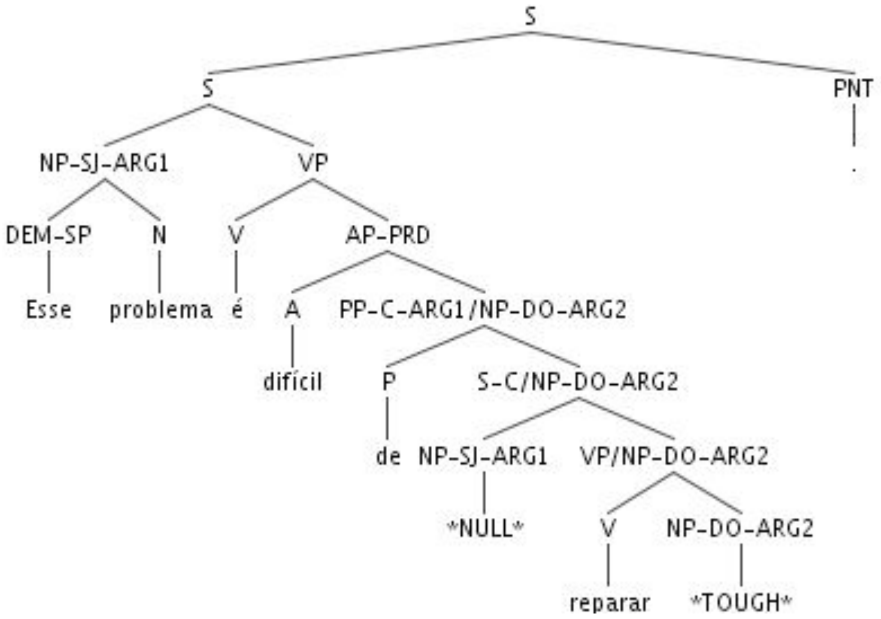


#Id:e000660/79129

Indirect object control verbs (e.g. *pedir*) select for a Indirect Object PP-IO-ARG31, signaling that it is both the indirect object of the control verb and the subject in the clause occurring as the other internal argument of the control verb.

7.10 "though" constructions

In "though" constructions, the sentential complement of the adjective, introduced by the preposition *de*, and projected by an inflected infinitive, has a phonetically null object marked with *TOUGH*:



#Id:a012/591

For more details, see also section 7.6 on infinitives and 6.4 on "though" null objects.

7.11 clitics

Clitics project NPs. In terms of constituency, they enter the same positions as any NP projected from an N with similar grammatical function (see also section 10.4

on the tokenization of clitics).

8 Long-distance relations

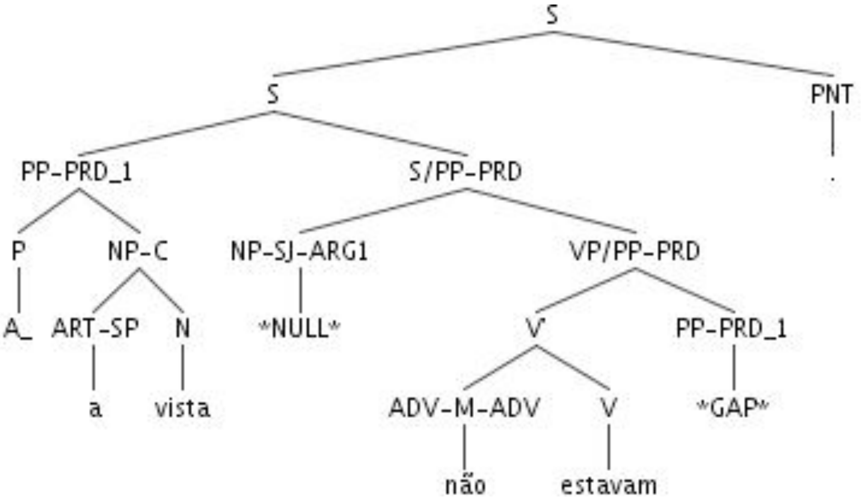
Long distance relations are established between a constituent and a right downwards position in the tree where this constituent typically occurs in (declarative) counterparts with cannonical SVO word order. Constructions with long-distance relations include topicalizations, interrogatives and relatives.

The cannonical position is marked by a phonetically null item *GAP* which is coindexed with the constituent with which it establishes a long-distance relation.

[Prob Parser: The long distance dependency is represented by decorating every node in the path inside the tree connecting the node immediately dominating the putative gap and the sister node of the "displaced" constituent. These nodes in that path are decorated by concatenating to their category tags a slash "/" followed by the triple CAT-GF-SR of that "displaced" constituent, where CAT is its category, GF is its grammatical function, and SR is its semantic role.]

8.1 topicalization

The topicalized constituent is in adjunction to the constituent from which it was topicalized:



#idc049/24856

8.2 relatives

The relative phrase projects a CP immediately dominating the constituent S from which it was relativized.

See also section 7.1 on Relatives.

8.3 interrogatives

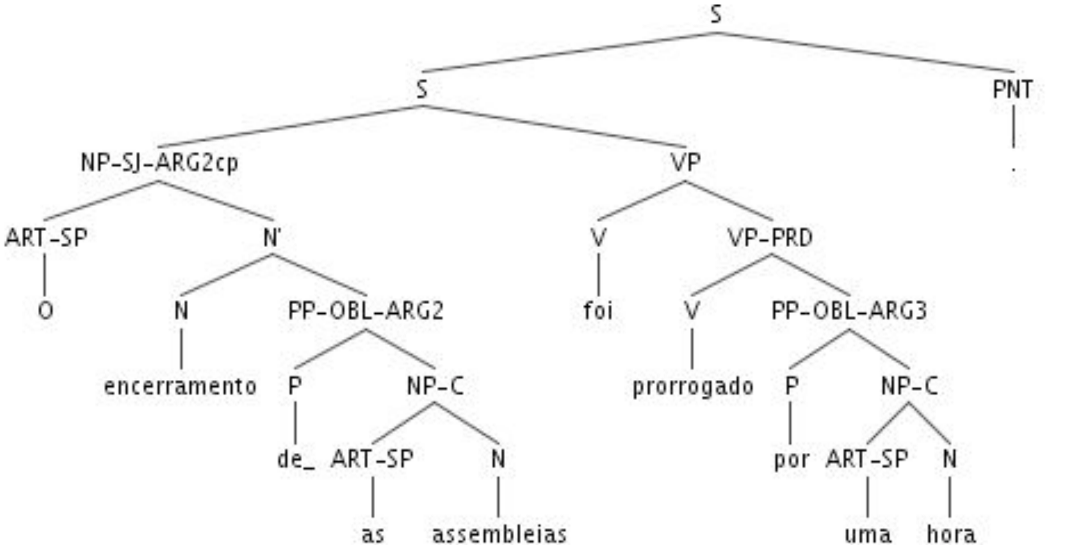
In its current version, the corpus does not contain yet interrogatives with long distance relations.

9 Valency alternations

9.1 passives

The *by*-phrase is an internal complement of the past participle verb form, with grammatical function OBL and semantic role ARG1.

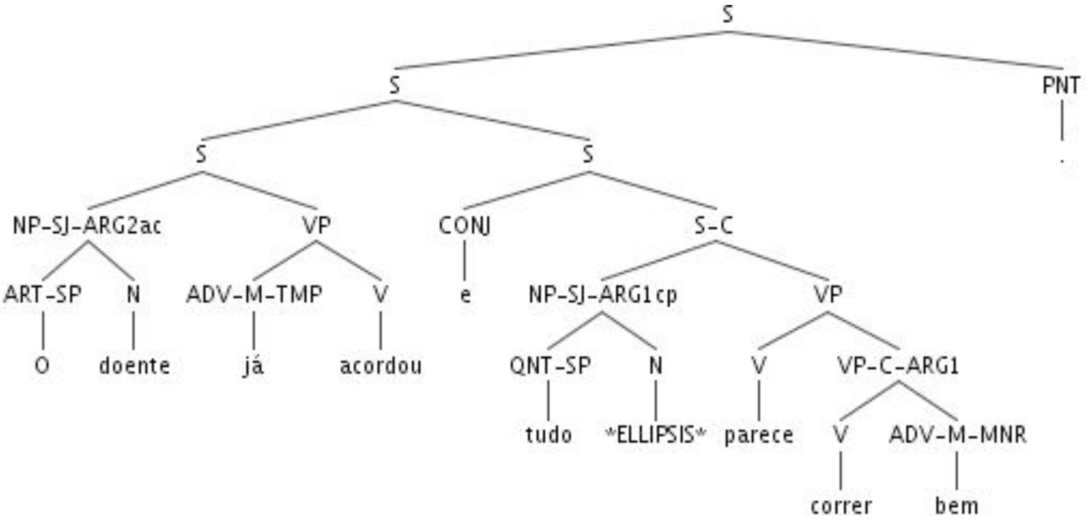
The corresponding Subject bears ARG2cp (see also section 7.8 on complex predicates above):



#Id:b179/11830

9.2 anti-causatives

The Subject of an anticausative verb is ARG2ac:



#ide000530/68760

In a predication supported by the transitive counterpart of a possible anticausative verb, the Subject is ARG1, as with other transitive verbs. As expected, in its passive alternation, the Subject is ARG2.

10 Tokenization

10.1 sentence spliting

Sentences are splited at the expected points. It is worth of mention the case of utterances involving colon ":", which will be split into two separate entry sentences in the treebank, one preceding the colon and the other following it.

10.2 non verbal utterances

Titles of newspaper articles, stretches around colons, etc. are cases of possible utterances in the corpus wich are not projected by a corresponding verbal head. In any case, every entry utterance in the corpus is dominated by an S node.

10.3 contractions

Contractions are expanded. The first element of an expanded contraction is marked with an "_" (underscore) symbol, for instance *do* → |*de_*|*o*|.

10.4 clitics

Clitics are detached from the verb. The detached clitic is marked with a "-" (hyphen) symbol, as for instance *dá-se-lho* → |*dá*|*-se*|*-lhe*|*-o*|

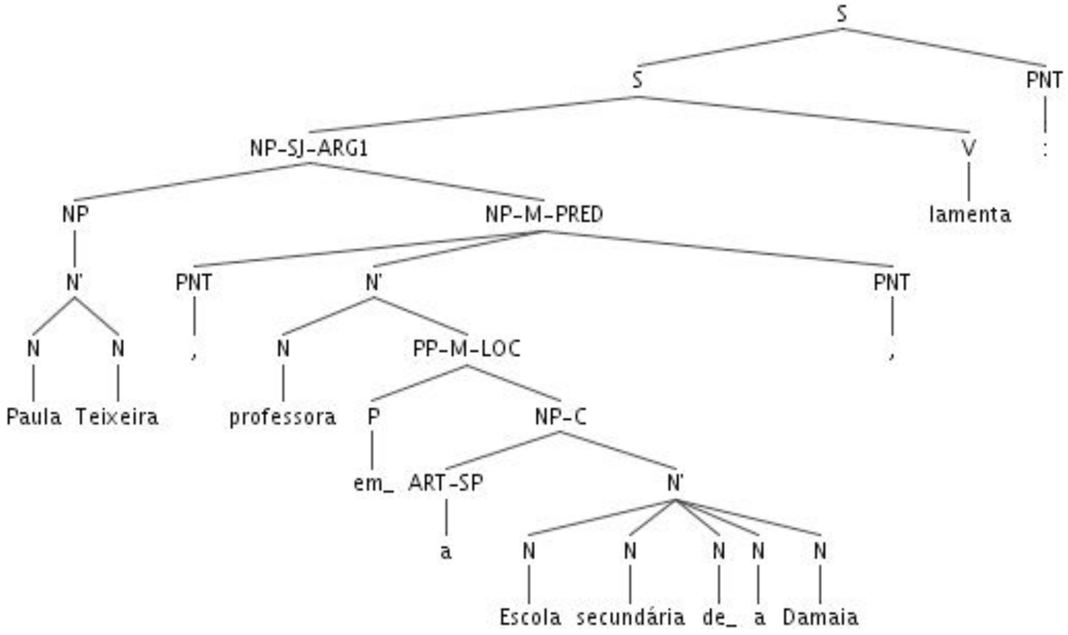
When in mesoclisis, a "-CL-" mark is used to signal the original position of the detached clitic: *afirmar-se-ia* → |*afirmar-CL-ia*|*-se*|

Possible vocalic alterations of the verb form are marked with "#" (hash) symbol, as for instance in *vê-las* → |*vê#*|*-las*|.

11 Multi-word expressions

11.1 Proper names

Multi-word proper names form a flat constituent where every word is sister of each other, is of category N, and is dominated by a single common N' node. This head projects an NP except when it is a modifier of a common noun:

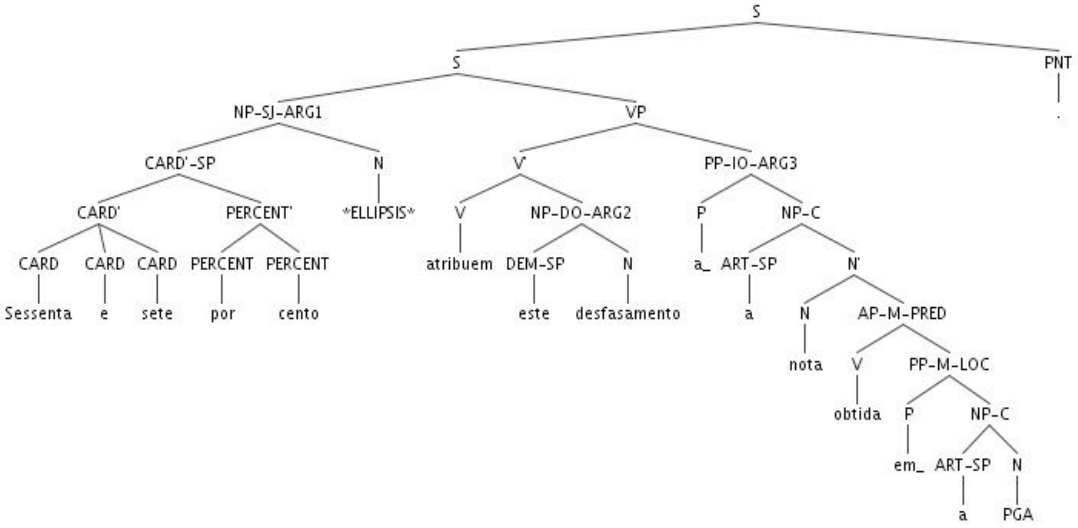


#b005/254

See also sections 2.6.1 on bare NPs, and 5.2 on NPs.

11.2 cardinals

Complex cardinals have a flat structure like a multi-word named-entity.



#ide000650/78330

12 Textual marking

12.1 punctuation

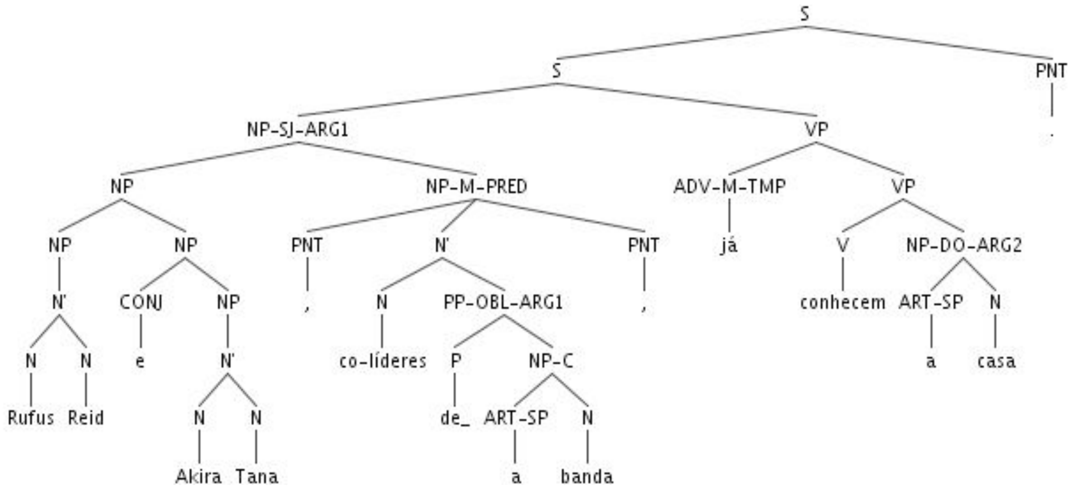
Each punctuation marks is a constituent of category PNT.

End of sentence markers are in top most adjunction.

12.2 comma

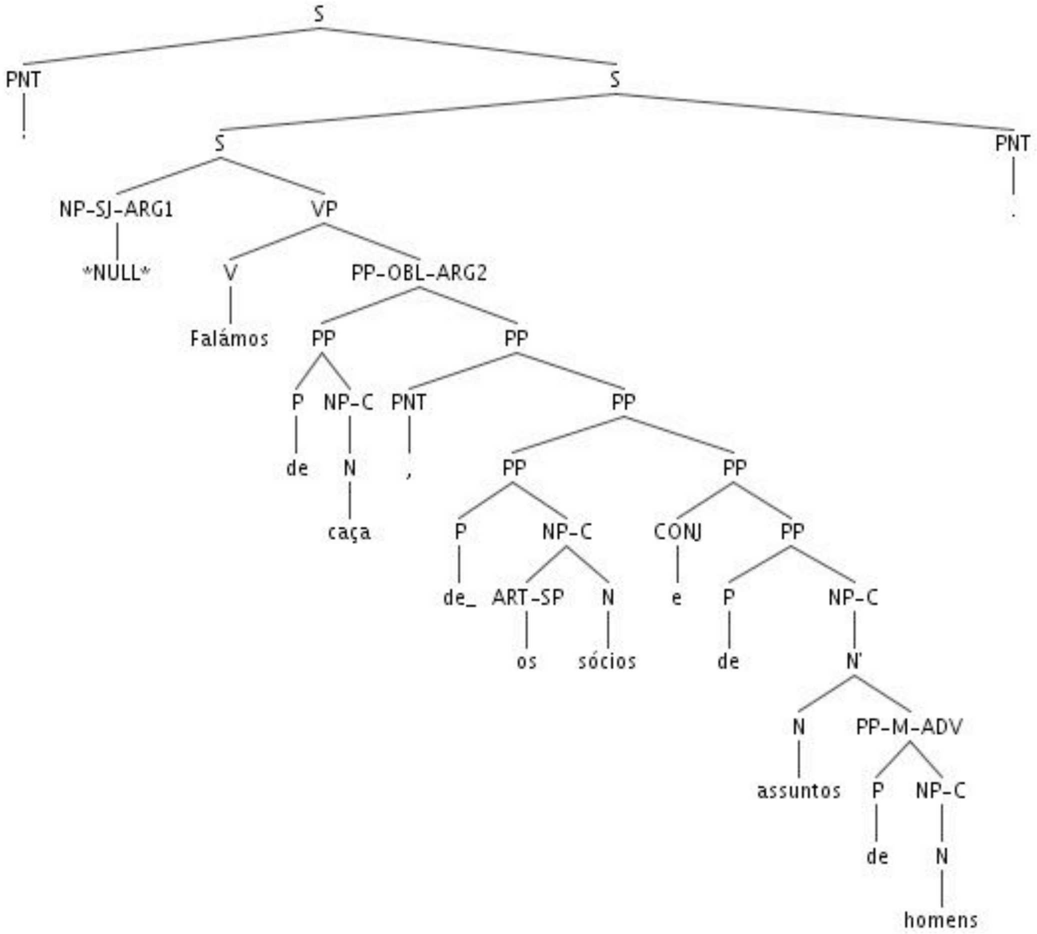
Commas separating left periphery constituents are right adjoined to these constituents.

Commas surrounding appositions are top most constituents of the appositive constituent.



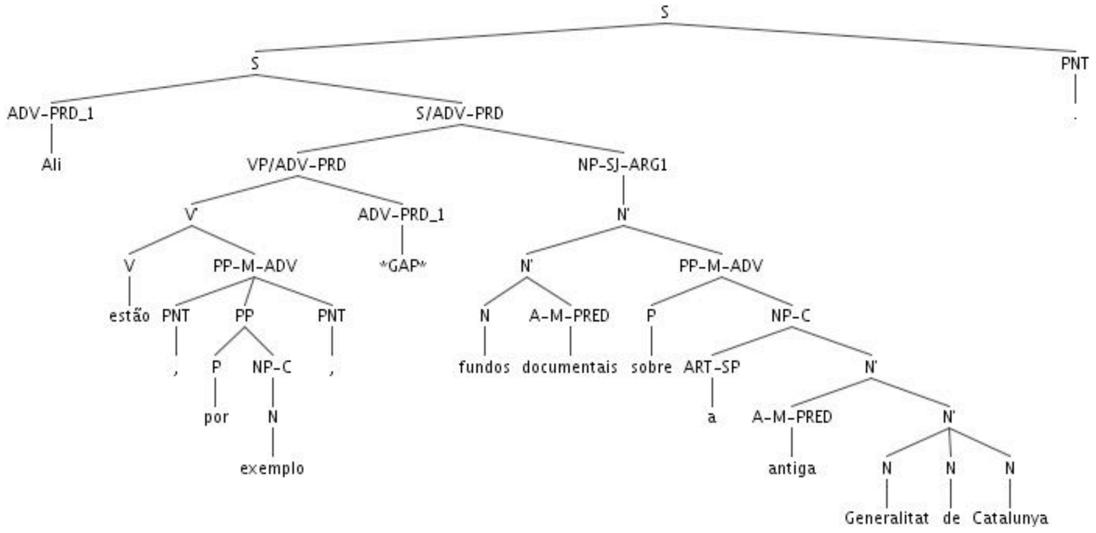
#id:b029/1761

Commas with coordinative value are represented like lexical coordinative conjuctions are (for further details, see section 7.4 on coordination):



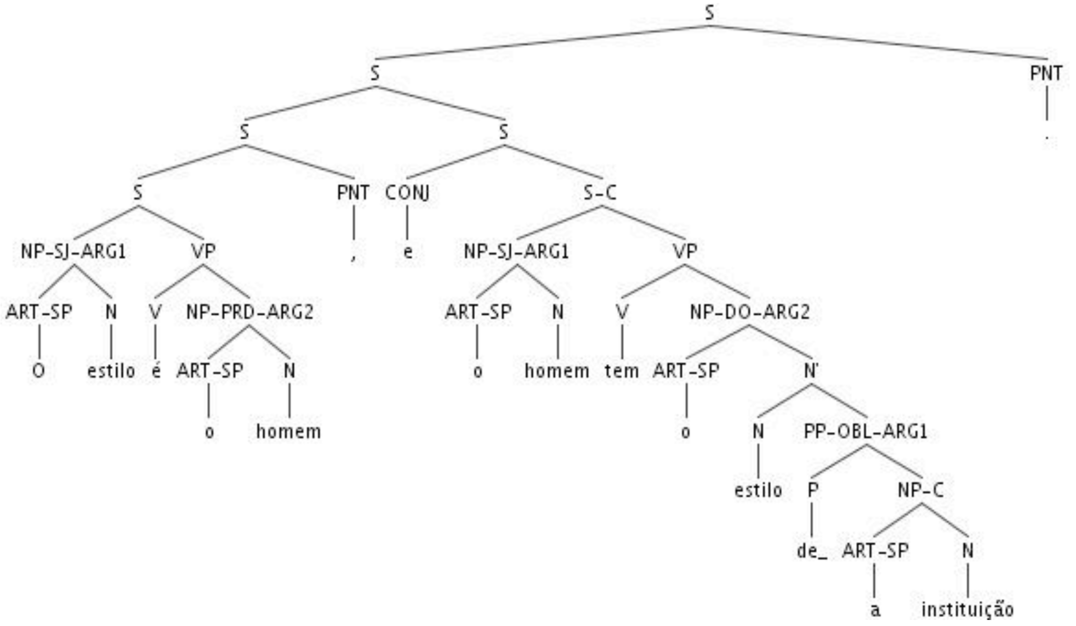
#ida001/30

Commas surrounding parentheticals are adjoined to the surrounded constituent. With several parentheticals in sequence, the first one is surrounded, the following ones have a single comma at its right:



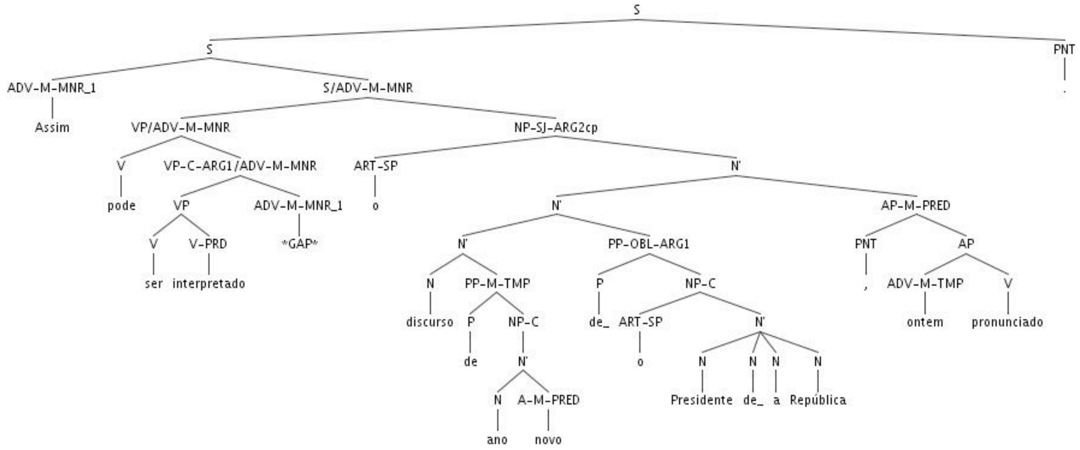
#idb227/16800

Commas emphasizing conjunctions, thus immediately preceding them, are right adjoined to the left coordinated constituent:



#idb184/12279

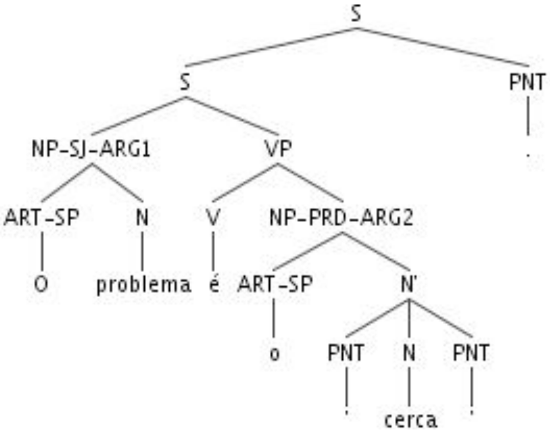
Other "pause" commas are left adjoined:



#idb128/8002

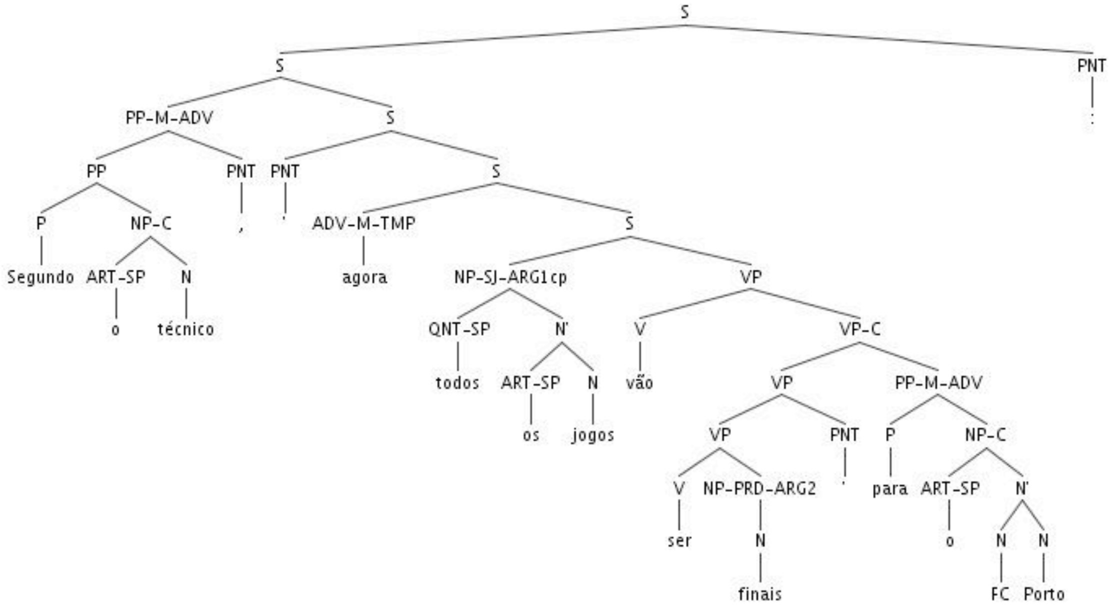
12.3 quotation marks

Quotation marks surrounding constituents are adjoined to them. When they surround a lexical item of category *X*, they are dominated by an *X'* node together with X:



#Id:b010/654

Quotation marks surrounding strings not forming a constituent are adjoined to the highest possible node:



#Id:b091/5858.

13 References

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Nascimento, Filipe Nunes and João Silva, 2006, "Open Resources and Tools for the Shallow Processing of Portuguese", *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC2006), Genoa, Italy.

Branco António, Sérgio Castro, João Silva, Francisco Costa, 2011, *CINTIL DepBank Handbook: Design options for the representation of grammatical dependencies*. Department of Informatics, University of Lisbon, Technical Reports series, nb. di-fcul-tr-11-03, http://hdl.handle.net/10455/6747 .