CINTIL DepBank Handbook: Design options for the representation of grammatical dependencies

António Branco, Sérgio Castro, João Silva and Francisco Costa

DI–FCUL–TR–2011–03

DOI:10455/6747

(http://hdl.handle.net/10455/6747)

July 2011

Published at Docs.DI (http://docs.di.fc.ul.pt/), the repository of the Department of Informatics of the University of Lisbon, Faculty of Sciences.
CINTIL DepBank Handbook:
Design options for the representation of grammatical dependencies
António Branco, Sérgio Castro, João Silva, Francisco Costa

University of Lisbon

January 2011

1 INTRODUCTION

1.1 Concordancer

2 DEPENDENCY IN A NUTSHELL

2.1 Extension: semantic relations

2.2 CINTIL Treebank

3 TAG SET

3.1 lexical categories

3.2 grammatical functions

3.3 semantic functions

4 DEPENDENCY RELATIONS

5 PHONETICALLY NULL ITEMS

5.1 null subjects

5.2 null governors

5.3 traces
5.4 "though" null objects

6 SPECIFIC CONSTRUCTIONS

6.1 comparatives

6.2 coordination

6.3 complex predicates: auxiliary, raising and modal verbs

6.4 "though" constructions

6.5 clitics

7 LONG-DISTANCE RELATIONS

7.1 topicalization

7.2 relatives

7.3 interrogatives

8 VALENCY ALTERNATIONS

8.1 passives

8.2 anticausatives

9 TOKENIZATION

9.1 sentence splitting

9.2 non verbal utterances

9.3 contractions

9.4 clitics

10 MULTI-WORD EXPRESSIONS

10.1 Proper names

10.2 cardinals

11 TEXTUAL MARKING

11.1 punctuation

11.2 comma

11.3 quotation marks

12 REFERENCES
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 written utterances associated to the representation of their linguistic structure, which can be set to capture different degrees of linguistic information.

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

This document aims at supporting the utilization and exploitation of the CINTIL DepBank. 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 DepBank 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:".

The present document has a companion, which is the Handbook for the CINTIL TreeBank (Branco et al., 2011). Many of the issues addressed here may have received a more thorough consideration there.

1.1 Concordancer

The CINTIL DepBank can be searched through a concordancer online at http://lxcenter/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 Dependency in a nutshell

In an utterance, a lexeme B depends on a lexeme A when the occurrence of B in its specific position is made possible by the occurrence of A. In such case, it is considered to exist a grammatical dependency relation from the lexeme B, the dependent element, to the lexeme A, the governor element of the dependency.

Dependency relations can be depicted as graphs whose nodes are lexemes and
whose directed arcs establish a connection from a governor to its dependent lexemes.

In the CINTIL DepBank, individual lexemes are further annotated with a feature bundle containing, where appropriate, information on lexical category, lemma and inflection.

Dependency relations can be of a number of different types, which are mostly the usual grammatical functions, and with whose tags the arcs are decorated.

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.

The possible values of grammatical functions are listed in section 3.2 below.

2.1 Extension: semantic relations
The CINTIL DepBank was extended so that besides the tags for the different dependency relations, the arcs are further decorated with tags indicating the semantic relation at stake.

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 lack of isomorphism between grammatical relations and semantic relations in what concerns argumental roles, the extension of the CINTIL dependency bank with semantic relations is restricted to non argumental ones.

The possible values of semantic functions are listed in section 3.3 below.

For a fully fledged representation of semantic relations, see the CINTIL LogicalForm Bank.

2.2 CINTIL Treebank
Corpora annotated with grammatical constituency trees are known as TreeBanks stricto sensu. The CINTIL DepBank is aligned to a constituency bank, the CINTIL TreeBank. The key 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 Treeank see (Btanco et al., 2011).

3 Tag set

3.1 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

3.2 V Verbgrammatical functions
SJ      Subject
SJjac   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
3.3 Semantic functions

LOC Location
EXT Extension
ADV Adverbia
CAU Cause
TMP Temporal
PNC Purpose, goal
MNR Manner
DIR Direction
PRD Predication
POV Point of view

[Prob Parser: Tags for argumental semantic roles are imported from the CINTIL TreeBank and kept in the version of the CINTIL DepBank for the probabilistic parser. Given the lack of complete isomorphism between grammatical relations and semantic relations in this respect, even though they decorate arcs here, like what happens in the CINTIL TreeBank, these argumental tags refer only to one of the terms of the putative semantic relation, which in this case is the dominated node, pointed to by the arc. For a fully fledged representation of semantic relations, see the CINTIL LogicalFormBank]

4 Dependency relations

The head of a constituent that is a complement, a modifier or a specifier in a given predication, is dependent of the corresponding syntactic predicator.

The dependency relation of a specifier is SP, and of a Modifiers is M.

The dependency relation of a complement is its grammatical function.

In the case of a complement of a Preposition, its complement is dependent under the grammatical relation of C (standing just for "Complement").
5  **Phonetically null items**

Phonetically null items signal positions in the graph related to other positons in the graph (in the case of traces), or signal 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 their arcs are removed from the grap.]

5.1  **null subjects**

Null subjects are marked by *NULL*:

```
DSP
  Sjcp
    *NULL* Não posso substituir os jogadores todos.
```

[Prob Parser: leaf nodes representing null subjects are removed, together with the associated arc.]

5.2  **null governors**

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

```
M-LDC
  /\DSP
  | C  N-ADV C SP SP
  |_\            
  A_ a porta da_ a casa , tudo *ELLIPSIS* *NULL* em_ a mesma .
```

[Prob Parser: the representation of the null governor by a phonetically null category is removed, and the domination arcs getting out of the elided governor will be getting out of the lowest cousin head in the constituency tree.]

5.3  **traces**

Differently from a constituency treebank, traces of constituents displaced are not represented here by a phonetically null category. Instead, the head of the displaced node is dominated by its governor.

```
M PRD PUNCT
  \  /  
  Mais satisfeito estava Mário Reis .
```

[Prob Parser: the representation of the null governor by a phonetically null category is removed, and the domination arcs getting out of the elided governor will be getting out of the lowest cousin head in the constituency tree.]

5.4  **"though" null objects**

Null direct objects specifically licensed by "though" constructions are represented by *THOUGH*.
[Prob Parser: leaf nodes representing "though" null objects are removed, together with the associated arc.]

6 Specific constructions

6.1 comparatives

A comparative construction is typically built around an adjective by two constituents, an adverbial of degree and a CONJP phrase, which have the following rendering in terms of dependency relations:

(some adverbs may also support comparative constructions, as with *perto* in the example *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:

The adverbial of degree (e.g. *mais, menos, tão*) is dependent on the adjective.

The head of the CONJP phrase is dependent of the adverbial of degree, of which is a complement.

The CONJP may be absent of the comparative construction. In this case, tough it can be semantically recovered from the context, there is no phonetically null item inserted in the graph.

6.2 coordination

Given a constituent A formed by coordination, in a given predication, the head of the first conjunct (from left to right) is a dependent of the corresponding predicator. The dependency relation type is the grammatical function of A.

Each conjunct is dependent on the immediately preceding conjunct under a
dependency relation of type COORD.

The commas or conjunctions are dependent on the subsequent conjunct under a dependency relation of type CONJ.

6.3 complex predicates: auxiliary, raising and modal verbs

In a complex predicate, the Subject relation is established with the leftmost verb. That is the case with auxiliary verbs:

With modal verbs,

With any sequence of auxiliary, raising or modal verbs,
6.4 "though" constructions

The sentential complement of the adjective, introduced by the preposition de, and projected by an inflected infinitive, has a phonetically null object:

```
  PUNCT
  /   \
SP  Sj  PRD  C  C  Sj  DO
  |   |   |   |   |   |   |
Esse problema é difícil de *NULL* reparar *TOUGH*. 
```

#Id:a012/591

For more details, see also section 5.4 on "though" null objects.

6.5 Clitics

Clitics enter the same dependencies as any N with similar grammatical function.

See also section 9.4 on the tokenization of clitics.

7 Long-distance relations

Long distance relations of dependency are established between a lexeme and a governor of the minimal predication where it typically occurs in (declarative) counterparts with canonical SVO word order. Constructions with long-distance relations include topicalizations, interrogatives, and relatives.

7.1 Topicalization

The head of the topicalized constituent is dependent on the governor of the minimal predication from which it was topicalized.

```
  PRD
  /   \
C   Sj  M  PUNCT
  |     |
A_ a vista *NULL* não estavam .
```

#idc049/24856

7.2 Relatives

The head of a relative phrase is dependent on the governor of the minimal predication from which it was relativized. Be it a so-called restrictive relative:

```
  PUNCT
  /   \
M   OBL  Sj  PRD  C  Sj  SF
  |     |   |     |   |   |   |
Um repto a que o empresário não foge :
```

#Id:4682

Be it an appositive relative clause:
Or be it a free relative clause:

```
Quem não gosta de_ o projecto é a CGT .
```

### 7.3 interrogatives

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

### 8 Valency alternations

#### 8.1 passives

The by-phrase has grammatical function OBL (see also section 7.8 on complex predicates above)

```
O encerramento de_ as assembleias foi prorrogado por uma hora .
```

#### 8.2 anticausatives

The Subject of an anticausative os SJac:

```
O doente já acordou e tudo *ELLIPSIS* parece correr bem .
```

### 9 Tokenization

#### 9.1 sentence splitting

Sentences are split 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 another following it.

#### 9.2 non verbal utterances

Titles of newspaper articles, stretches around colons, etc. are cases of possible
utterances in the corpus which are not structured around a corresponding verbal governor.

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

9.4 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 mesoclis, 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|.

10 Multi-word expressions
10.1 Proper names
The first element (left to right) of a multi-word proper name is dependent on the governor subcategorizing for that proper name. The remaining lexemes of the multi-word proper name are dependent on that first element under dependency relations tagged with N.

10.2 cardinals
Complex cardinals have a graph representation like a multi-word named-entity, whose arcs are labelled with CARD.

11 Textual marking
An end of sentence full stop is dependent under a dependency relation tagged with PUNCT. Every other textual mark is dependent under a dependency relation tagged with DEP.
11.1 punctuation
End of sentence markers are dependent from the main syntactic predicate of the utterance.

11.2 comma
Commas separating left periphery constituents are dependent on the head of these constituents.

Commas surrounding appositions are dependent on the head of the NP being modified by the apposition.

Commas with coordinative value are represented like lexical coordinative conjunctions are: for details, see section 6.2 on Coordination.

Commas surrounding parenthetics are dependent on the head of constituent being modified by the parenthetical.

A comma emphasizing a conjunction, thus immediately preceding it, is dependent on the same governor as that conjunction, but under the dependency tagged with DEP.

Other "pause" commas are dependent on the main predicate of the constituent immediately to its right.
11.3 quotation marks
Quotation marks surrounding a constituent are dependent on the head of that constituent.

References