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 **3**

1.1 Concordancer **3**

2 DEPENDENCY IN A NUTSHELL **3**

2.1 Extension: semantic relations **4**

2.2 CINTIL Treebank **4**

3 TAG SET **4**

3.1 lexical categories **4**

3.2 grammatical functions **4**

3.3 semantic functions **6**

4 DEPENDENCY RELATIONS **6**

5 PHONETICALLY NULL ITEMS **7**

5.1 null subjects **7**

5.2 null governors **7**

5.3 traces **7**

5.4 "though" null objects 7

6 SPECIFIC CONSTRUCTIONS 8

6.1 comparatives 8

6.2 coordination###REVER no corpus/conversor 8

6.3 complex predicates: auxiliary, raising and modal verbs 9

6.4 "though" constructions 10

6.5 clitics 10

7 LONG-DISTANCE RELATIONS 10

7.1 topicalization 10

7.2 relatives 10

7.3 interrogatives 11

8 VALENCY ALTERNATIONS 11

8.1 passives 11

8.2 anticausatives 11

9 TOKENIZATION 11

9.1 sentence spliting 11

9.2 non verbal utterances 11

9.3 contractions 12

9.4 clitics 12

10 MULTI-WORD EXPRESSIONS 12

10.1 Proper names 12

10.2 cardinals 12

11 TEXTUAL MARKING 12

11.1 punctuation 13

11.2 comma 13

11.3 quotation marks 14

12 REFERENCES 14

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 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 considration 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
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

COORD	Coordination
CONJ	Conjunction
N	Name in multi-word proper names
CARD	Cardinal in multi-word cardinals
PUNCT	Punctuation

DEP - Generic dependency

3.3 semantic functions

LOC	Location
EXT	Extension
ADV	Adverbial
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").

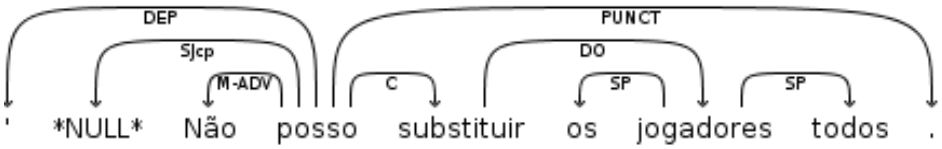
5 Phonetically null items

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*:

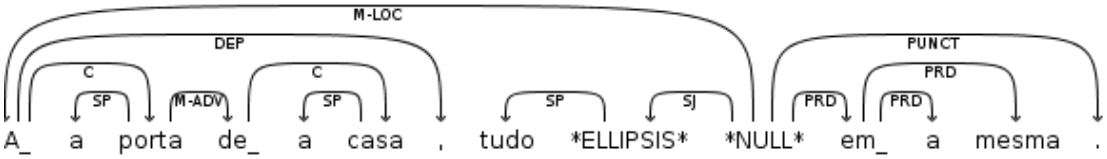


#Id:b092/5911

[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*:



#Id:b001/11

[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.]

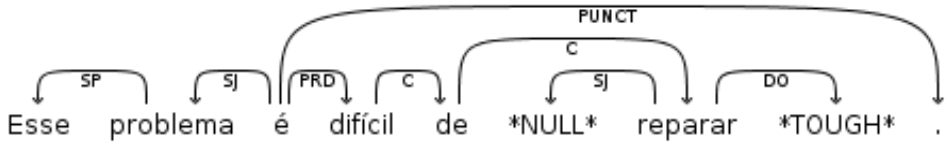
5.3 traces

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.

#Id:b094/6024

5.4 "though" null objects

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



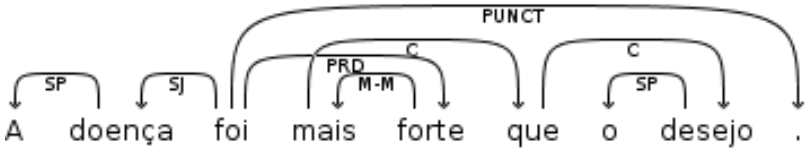
#Id:a012/591

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

6 Specific constructions

6.1 comparatives

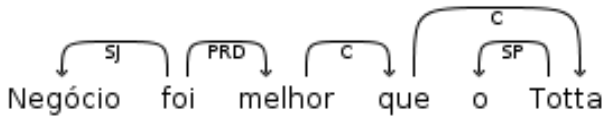
A comparative construction is typically built around an adjective by two constituents, an adverbial of degree and a CONJP phrase, which have the following rendering in terms of dependency relations:



#Id:e000282/49262

(some adverbs may also support comparative constructiosn, 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 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.

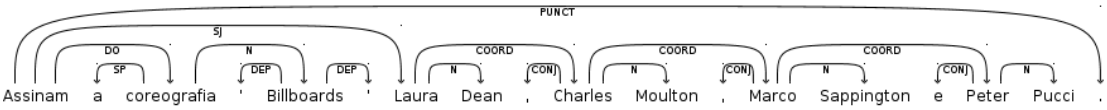
6.2 coordination

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 dependecy relation type is the grammatical function of A.

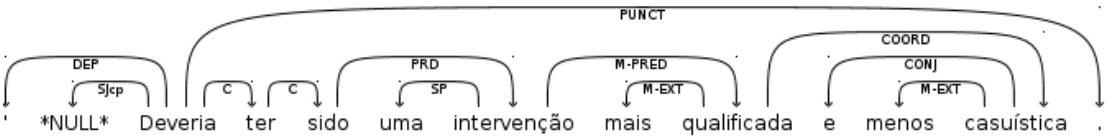
Each conjunct is dependent on the immediately preceding conjunct under a

dependency relation of type COORD.

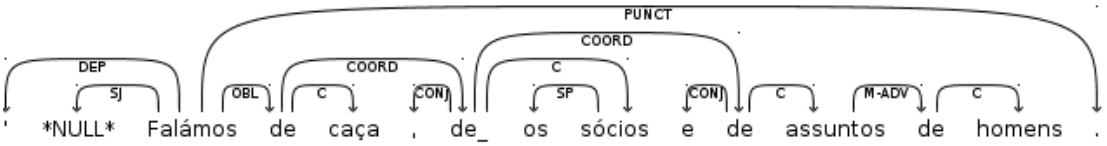
The commas or conjunctions are dependent on the subsequent conjunct under a dependency relation of type CONJ.



#Id:b00067/3939



#Id:b00062/3668



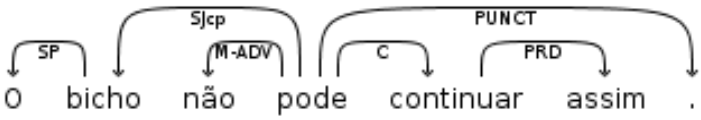
#Id:b00001/30

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:

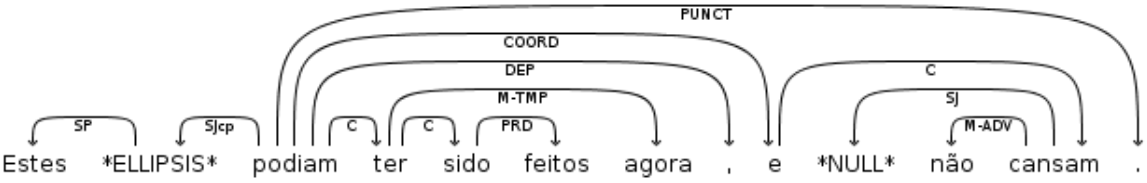
#Id:e000585/73183

With modal verbs,



#Id:c003/20534

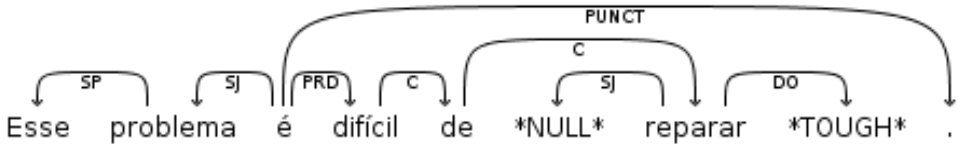
With any sequence of auxiliary, raising or modal verbs,



#Id:b134/8372

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:



#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.

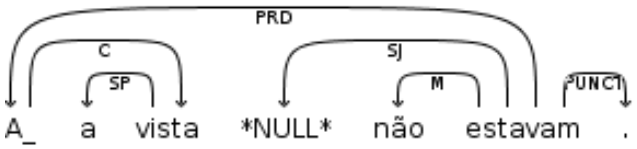
See also section 9.4 on the tokenization of clitics.

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

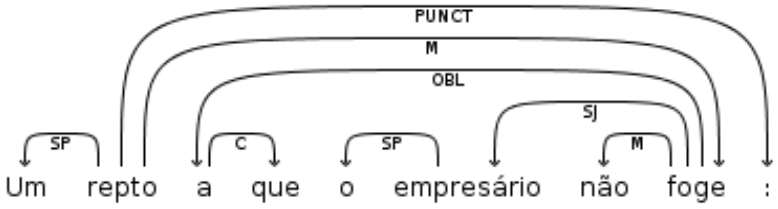
The head of the topicalized constituent is dependent on the governor of the minimal predication from which it was topicalized.



#idc049/24856

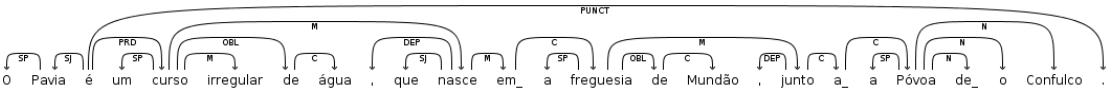
7.2 relatives

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:



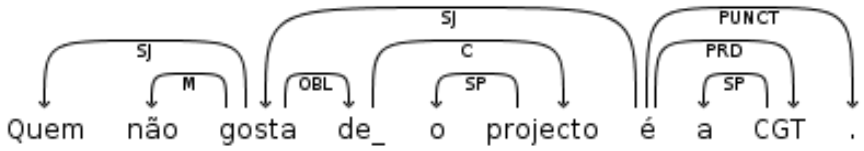
#Id:4682

Be it an appositive relative clause:



#Id:16012

Or be it a free relative clause:



#Id:38916

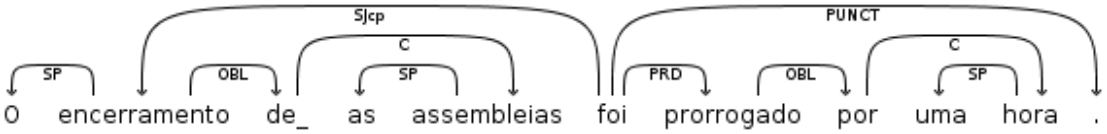
7.3 interrogatives

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)



#Id:b179/11830

8.2 anticausatives

The Subject of an anticausative os SJac:

#ide000530/68760

9 Tokenization

9.1 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 another following it.

9.2 non verbal utterances

Titles of newspaper articles, stretches around colons, etc. are cases of possible

utterances in the corpus wich 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*_|*o*|.

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*|

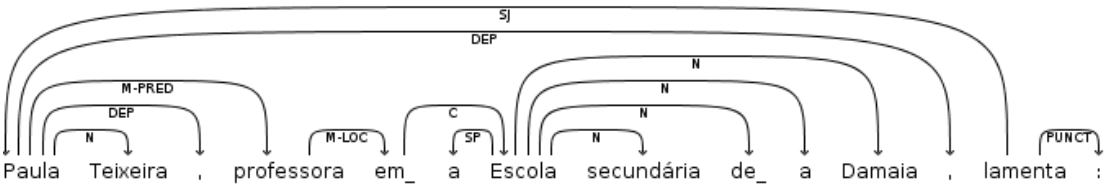
When in mesoclisis, a "-CL-" mark is used to signal the original position of the detached clitic: *afirmar-se-ia* → |*afirmar-CL-ia*|*-se*|

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

10 Multi-word expressions

10.1 Proper names

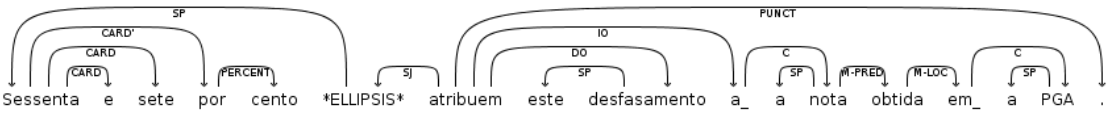
The first element (left to right) of a muti-word proper name is dependent on the governor subcategorizing for that proper name. The remaining lexemes of the multi-word proper name are dependent on that first element under dependency relations tageed with N.



#idb254

10.2 cardinals

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



#ide000650/78330

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.

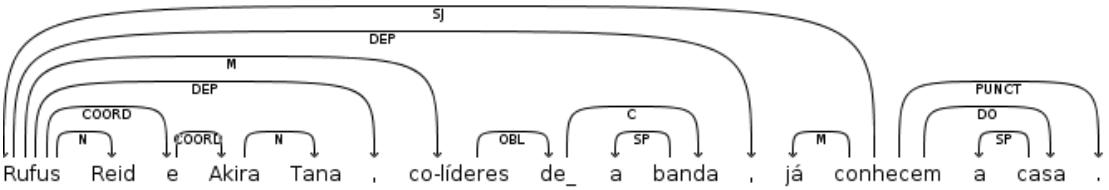
11.1 punctuation

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.

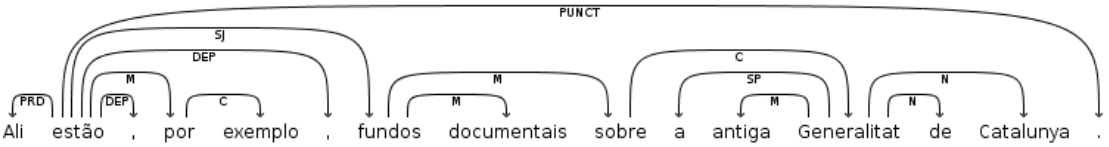
Commas surrounding appositions are dependent on the head of the NP being modified by the apposition.



#id:b029/1761

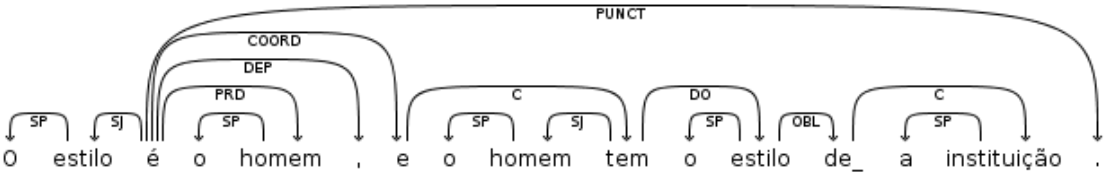
Commas with coordinative value are represented like lexical coordinative conjuctions are: for details, see section 6.2 on Coordination.

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



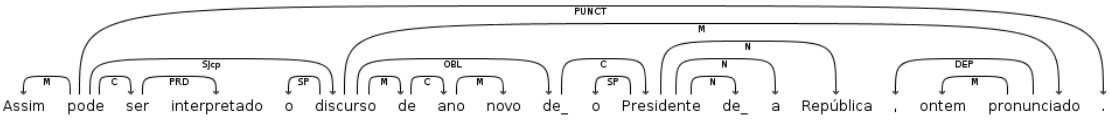
#idb227/16800

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.



#idb184/12279

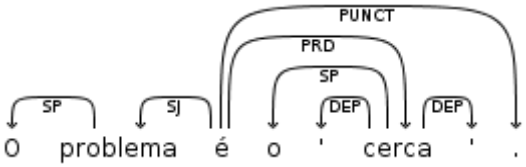
Other "pause" commas are dependent on the main predicate of the constituent immediately to its right.



#idb128/8002

11.3 quotation marks

Quotation marks surrounding a constituent are dependent on the head of that constituent.



#Id:b010/654

12 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 TreeBank Handbook: Design options for the representation of syntactic constituency*. Department of Informatics, University of Lisbon, Technical Reports series, nb. di-fcul-tp-11-02, http://hdl.handle.net/10455/6746 .