When CORDIAL Becomes Friendly: Endowing the CORDIAL Corpus with a Syntactic Annotation Layer

Catarina Magro
Center of Linguistics of University of Lisbon
cmm@clul.ul.pt

Abstract

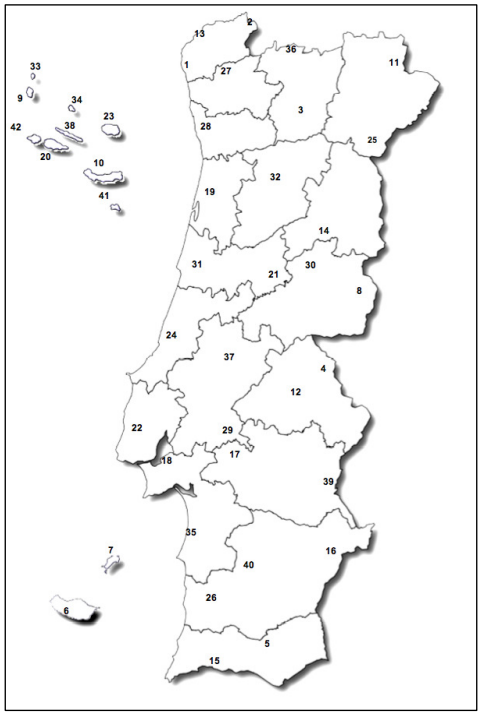
This paper reports on the syntactic annotation of a previously compiled and tagged corpus of European Portuguese (EP) dialects – *The Syntax-oriented Corpus of Portuguese Dialects* (CORDIAL-SIN). The parsed version of CORDIAL-SIN is intended to be a more efficient resource for the purpose of studying dialect syntax by allowing automated searches for various syntactic constructions of interest. To achieve this goal we adopted a rich annotation system (the UPenn corpora annotation system) which codifies syntactic information of high relevance. The annotation produces tree representations, in form of labelled parenthesis, that are integrally searchable with *CorpusSearch*, a search engine for parsed corpora (Randall, 2005-2007). The present paper focuses on CORDIAL-SIN annotation issues, namely it presents the general principles and guidelines of the adopted annotation system and describes the methodology for constructing the parsed version of the corpus and for searching it (tools and procedures). Last section addresses the question of how an annotation system originally designed for Middle English can be adapted to meet the particular needs of a Portuguese corpus of dialectal speech.

1. The CORDIAL-SIN corpus

The Syntax-oriented Corpus of Portuguese Dialects (CORDIAL-SIN) is being built up at the Linguistics Center of University of Lisbon (CLUL) within the scope of a research project aimed at promoting the study of European Portuguese dialect syntax by means (among other things) of the implementation of an online linguistic resource fulfilling the empirical demands of dialect syntax inquiry[1].

CORDIAL-SIN is a corpus of spoken dialectal EP that collects a geographically representative body of excerpts of spontaneous and semi-directed speech, selected from the oral interviews gathered by the Linguistic Variation Team at CLUL in the course of several Dialect Geography projects (ALEPG; ALEAç; ALLP; BA). At its current state (the final state, in terms of extent), the corpus covers 42 locations within the (continental and insular) territory of Portugal and it compiles about 600 000 words. Map 1 shows the geographical distribution of the CORDIAL-SIN locations.

The corpus is available online, on the CORDIAL-SIN website, under three different formats[2]: (i) verbatim orthographic transcripts (which include phonetic and morphological variants and also general spoken language phenomena), (ii) normalized orthographic transcripts (which eliminate phonetic transcriptions of variants and the marked up spoken language phenomena) and (iii) morphologically tagged texts (automatically tagged using the morphological tagger created by M. Finger for the *Tycho Brahe Corpus of Historical Portuguese*; cf. Finger, 1998, 2000).



Map 1: Geographical distribution of CORDIAL-SIN locations

CORDIAL-SIN was compiled and tagged between 1999 and 2007; the corpus syntactic annotation is implemented over POS tagged texts and is currently being carried out.

2. The CORDIAL-SIN syntactic annotation

2.1. The annotation system

Presently the CORDIAL-SIN team main goal is to make available a more efficient resource for the purpose of studying (dialect) syntax, namely a parsed version of the corpus that allows searching not only for words or word

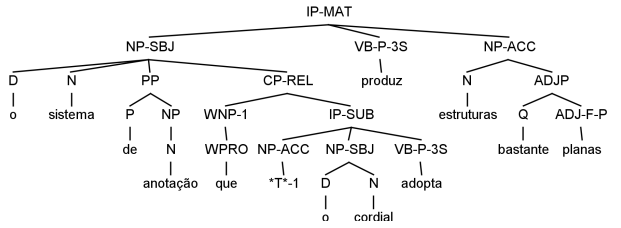
[1] The CORDIAL-SIN project is supported by national funding (PRAXIS XXI/P/PLP/13046/1998; POSI/1999/PLP/33275; POCTI/LIN/46980/2002; PTDC/LIN/71559/2006).

[2] CORDIAL-SIN is part of a European network of dialect syntax, promoted by the ESF-funded project *Edisyn*, and, in the near future will be also searchable (and interoperable with other dialectal corpora/databases) via the *Edisyn Search Engine*.

sequences but also for syntactic structure. To this end, we make use of a rich annotation system, based on the *Penn Parsed Corpora of Historical English* syntactic annotation set-up (Kroch & Taylor, 2000; Kroch, Santorini & Delfs, 2004; Kroch, Santorini & Diertani, 2010), which is integrally searchable with *CorpusSearch*, a query program for parsed corpora (Randall, 2005-2007)[3].

The Penn annotation system is designed to facilitate automated searches for various syntactic constructions of interest and not to associate every sentence with a correct structural description. This strategy leads to quite flat and sometimes linguistically uncommitted syntactic objects: multiple branching nodes, some word level nodes (e.g. verbs, negation, sentence focus particles), omission of undecidable information (e.g. VP boundaries) and subtle distinctions (e.g. argument vs adjunct PPs), use of default rules (in w.r.t. location of wh-traces and structural ambiguity, among others). The representation in (1) illustrates a typical CORDIAL-SIN syntactic structure.

(1)



the system of annotation that the cordial adopts produces structures quiet flat

The annotation system that CORDIAL adopts produces quiet flat structures

In spite of this, the Penn system is a really rich annotation system which provides the marking up of information of high relevance, such as constituent boundaries, phrase and clause dependencies, categorial information (e.g. NP, PP, ADVP), grammatical relations (e.g. SBJ, ACC, DAT), discoursive functions (e.g. left dislocation, pragmatic marker), sentence and clause types (e.g. EXL, CMP, QUE), some null constituents and certain transformational relations.

Syntactic annotation is represented as labeled bracketing over morphologically tagged texts. At the word level, POS tags are preserved[4]. Phrase and clause main labels are category labels and extended labels provide information concerning 'sub-category', 'grammatical relation' or 'discoursive function'. Table 1 presents the core set of labels and extended labels allowed by the original system and (2) and (3) illustrate how CORDIAL-SIN tagged and parsed texts look like (in the labeled bracketing structures, depth of indenting corresponds to depth of structural embedding).

Phrase labels	
NP	Noun Phrase
NP-SBJ	Noun Phrase (subject)
NP-ACC	Noun Phrase (DO, nominal predicate)
NP-ADV	Noun Phrase (adverbial)
NP-VOC	Noun Phrase (vocative)
NP-DAT	Noun Phrase (dative)
NP-GEN	Noun Phrase (dative of possession)
PP	Prepositional Phrase
PP-ACC	Prepositional Phrase (partitive object)
ADVP	Adverbial Phrase
ADJP	Adjective Phrase
NUMP	Numeral Phrase
INTJP	Interjection Phrase
QP	Quantifier Phrase
WXP	Wh-Phrase (e.g. WNP, WPP)
Clause labels	
IP-MAT	Independent or conjoined declarative IP
IP-IND	Independent, non-declarative IP
IP-SUB	Subordinate IP (under CP)
IP-ADV	Adverbial IP
IP-INF	Infinitival clause
IP-GER	Gerund clause
IP-PPL	Participial clause
IP-SMC	Small clause
CP-THT	*That* clause
CP-REL	Relative
CP-FRL	Free Relative
CP-CLF	Cleft
CP-ADV	Adverbial clause
CP-DEG	Degree clause
CP-CMP	Comparative clause
CP-EXL	Exclamative
CP-IMP	Imperative
CP-QUE	Question

Table 1: Labels and extended labels

(2)

```
e/CONJ andávamos/VB-D-1P com/P as/D-F-P redes/N-P    @de/P @o/D
and    were1PL          with the     fishing_net of    the

badejo/N ,/, que/WPRO são/SR-P-3P mais/ADV-R baixas/ADJ-F-P .../.
whiting  ,   that     are3PL      more      deep

and we were with the whiting fishing net that are deeper.
```

[3] The Penn annotation scheme is equally used on the *Tycho Brahe* corpus (a parsed corpus of historical portuguese) and on the *Canadian Parsed Corpus of Historical French* – both of them currently under construction too. The idea is that all these corpora, by using the same standards with respect to data annotation, constitute a corpora network suitable for research on comparative syntax.

[4] The format of the POS tags and the basics of the tagset essentially stem from the Tycho Brahe POS annotation system. Main tags include morpho-syntactic tags, word specific tags and punctuation tags; subtags codify inflectional information or specify in more detail morpho-syntactic information (on the CORDIAL-SIN POS annotation system, cf. Magro & Morgado (2008)).

(3)

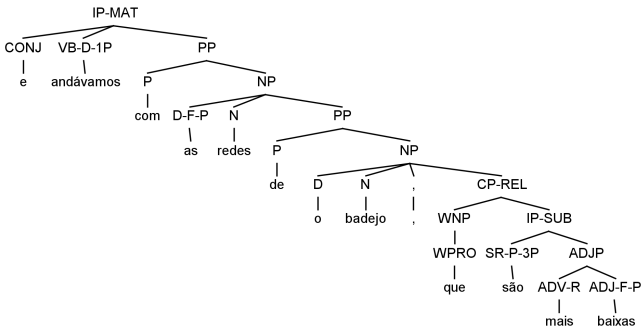
```
(IP-MAT  (CONJ e)
         (NP-SBJ *pro*)
         (VB-D-1P andávamos)
         (PP  (P com)
              (NP  (D-F-P as)
                   (N-P redes)
                   (PP  (P @de)
                        (NP  (D @o)
                             (N badejo)
                   (, ,)
                   (CP-REL (WNP-1  (WPRO que))
                           (IP-SUB (NP-SBJ *T*-1)
                                   (SR-P-3P são)
                                   (ADJP  (ADV-R mais)
                                          (ADJ-F-P baixas))))))))
         (. ...))  [CORDIAL-SIN, VPA07]
```

2.2. The annotation process

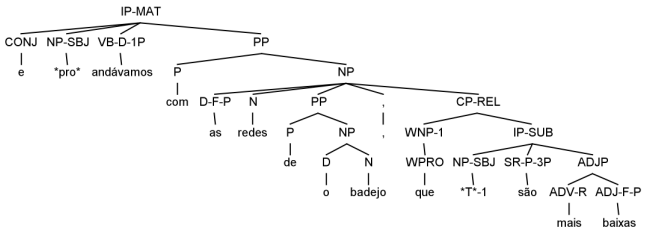
The implementation of CORDIAL-SIN syntactic annotation takes advantage of the tools developed by the Penn corpora team. The parsed texts result from a two steps process. At the first stage, a version of Mike Collins and Dan Bikel's statistical parser (Collins, 1999; Bikel, 2004), modified for treebank construction by Seth Kulick, runs over the POS tagged texts. At the second stage, the parser output is hand corrected with the help of *CorpusDraw*, an editing annotation tool[5, 6].

The parser output is a regular text file (ASCII file) that contains parsed, labelled sentences. *CorpusDraw* takes this file as a source file, displays the tree structures assigned to the sentences in the parsed corpus and allows the annotator to edit these trees. The *CorpusDraw* graphical user interface contains a row of editing buttons that enables the annotator to change syntactic labels, to break up run-on sentences or to consolidate fragments, to add subcategory information, to change attachment level, to add empty categories and to coindex related elements[7]. *CorpusDraw* will not permit the annotator to accidently change the order of words in the sentence or to delete any. (4) and (5) show the same parsed sentence pre and post editing respectively (see (2) for the gloss and translation). Notice that, in (4), both clauses lack a subject position, the relative clause is wrongly attached (it is modifying the noun *badejo/ whiting* and not the noun *redes/ fishing net*, as it should be) and the wh-phrase is not projected nor co-indexed with the subject of the relative clause. In (5) these adjustments are already made and the sentence is properly annotated.

(4)



(5)



At the end of this process, we get a parsed version of the corpus in a format that allows to retrieve automatically several syntactic configurations of interest using *CorpusSearch*, a dedicated search engine for parsed corpora. *CorpusSearch* operates in a friendly manner, making use of a basic query language plus linguistically intuitive search functions and providing very versatile searching options.

A query file of *CorpusSearch* contains a *node* (which gives *CorpusSearch* a node boundary within which to search) and a *query* (which instructs *CorpusSearch* as to what action to carry out). A basic query is composed by a search function call (each search function looks for one basic structural relationship) and the respective arguments (which correspond to the nodes being searched for). Search function arguments may take the form of an or-list, may include wild cards, and may be negated. Any number of search-function calls may be combined into more complex queries using the logical operators AND, OR, and NOT.

CorpusSearch creates a text output file, containing the sentences that match the condition(s) in the query (this output file is itself searchable).

Table 2 lists the main search functions of *CorpusSearch*. (6) shows a query sample that searches for nodes labelled ADJP (adjective phrase) that immediately dominate nodes labelled CP-CMP (comparative clause) and (7) displays a

[5] *CorpusDraw* and *CorpusSearch* (to be presented bellow in the text) are components of *CorpusSearch2* – a java program developed by Beth Randall that supports research in corpus linguistics (Randall, 2005-2007). *CorpusSearch2* is useful both for the construction of syntactically parsed corpora and for searching them. It runs under any Java-supported operating system (requires Java 2, version 1.5 or later) and expects labeled bracketing (Penn Treebank style).

[6] We are currently training the Penn parser with EP hand corrected annotated data. Details about the automatic parser performance have not been available yet. Nonetheless, such a rich annotation system will always require a considerable amount of human editing. In the face of an unsatisfactory parser performance it will be possible to make use of *CorpusSearch* corpus-revision feature (which allows to make automatic changes to an entire corpus) to assign syntactic structure to the corpus.

[7] The actions controlled by the editing buttons can also be triggered by the use of shortcuts, both keystrokes and mouse clicks, which facilitates the time-consuming task of human editing of the parser output.

sentence found by the query in (6).

Search function	Description
CCommands	neither x nor y dominates the other and the first branching node dominating x does dominate y
Dominates	y is contained in the sub-tree dominated by x
iDominates	y is a child (exactly one generation apart) of x
iDomsFirst	y is a first immediate child of x
iDomsLast	y is a last immediate child of x
iDomsOnly	y is the only child of x
iDomsTotal	counts the number of nodes immediately dominated by x
iDomsMod	x dominates y, and the only nodes intervening on the path from x to y (if any) are members of z
iDomsViaTrace	x immediately dominates t and t is co-indexed with another node z
HasSister	x and y have the same mother
Precedes	x comes before y in the tree but x does not dominate y
iPrecedes	x comes immediately before y in the tree but x does not dominate y
SameIndex	x has the same index as y
IsRoot	searches for the argument label at the root of the tree of the parsed token
Exists	searches for label or text anywhere in the sentence

Table 2: CorpusSearch main search functions

(6)

```
node: IP*
query: (ADJP iDominates CP-CMP)
```

(7)

```
(IP-MAT (NP-SBJ (PRO Ele))
        (SR-P-3S é)
        (ADJP (ADV-R mais)
              (ADJ velho)
              (CP-CMP (WADJP-29 0)
                      (C que)
                      (IP-SUB (ADJP *T*-29)
                              (NP-SBJ (PRO eu)))))
        (, ,))  [CORDIAL-SIN, VPA14]

Ele é  mais velho que  eu
he  is more old   than me

He is older than me.
```

2.3. Adapting the Penn system

The adopted annotation system – that was originally designed for Middle English – has to be adjusted to meet the particular needs of a Portuguese corpus of dialectal speech. In this domain, our team's work is that of adapting the existing system and creating the solutions required by those grammatical features where Portuguese and English differ or by other aspects that are specific of our dialectal data, namely, microvariation within the same construction and syntactically-relevant discourse phenomena.

The original annotation schemes and label set are preserved or slightly adapted wherever is possible but, in certain cases, new solutions are required. For consistency and for the ease of the process, the Portuguese additions are defined within the standards already operative in the Penn corpora and readable by the automatic parser.

Typical examples of the solutions we have found for Portuguese specifics are presented in the following subsections (for a comprehensive description of the CORDIAL-SIN syntactic annotation system, see Carrilho & Magro, 2009)

2.3.1. The label set

To set apart some syntactic units that abound in spoken texts, a small group of new extended labels was added on to the original label set, namely the labels -ANS, -POL, -TAG and -PRG. To avoid adding extra complexity for the annotation, no empty categories are codified inside the units identified with these extended labels.

Answers to yes/no and wh-questions are always annotated as IP-ANS.

(8)

```
(INQ E trazia-as já feitas?)
(IP-ANS (VB-D-1S Trazia)
        (. .))  [CORDIAL, PFT12]

INQ E trazia-as já feitas?
INQ And did you bring them already done?

Trazia
brought

Yes.
```

(9)

```
(INQ E tinha umas coisas para respirar?)
(IP-ANS (ADVP(ADV-NEG Não_senhora))
        (. .))  [CORDIAL, PFT40]

INQ E tinha umas coisas para respirar?
INQ And did it have any thing to breathe through?

Não senhora
no madam

No.
```

(10)

```
(INQ Passavam por sítios onde sabia que não havia guarda, não é?)
(IP-ANS (ADVP (ADV Pois))
        (. .))  [CORDIAL, AAL66]

INQ Passavam por sítios onde sabia que não havia guarda, não é?
INQ You went through places without customs officers, didn't you?

Pois.
sure

Sure.
```

(11)

```
(INQ Então porque é que as eiras eram normalmente em cima dos cabeços?)
(IP-ANS (PP (P Por)
            (NP (N causa)
                (PP (P @de)
                    (NP (D @o)
                        (N vento)))))
        (. .)) [CORDIAL, AAL10]
```

```
INQ Então porque é que as eiras eram normalmente em cima dos cabeços?
INQ Why the threshing-floors were usually on the top of the hill?

Por causa de o   vento
because   of the wind

Because of the wind.
```

IP-POL labels utterances which reinforce the truth value of a previous assertion. Just as in IP-ANS, no empty category is annotated inside IP-POL.

(12)

```
(IP-MAT (CONJ Porque)
        (NP-SBJ *pro*)
        (SR-P-3S é)
        (ADJP (ADJ branco))
        (. .))
(IP-POL (SR-P-3S É)
        (. .))  [CORDIAL, VPA24]

Porque  é  branco. É
because is white.  is

Because it's white.
```

(13)

```
(IP-MAT (NP-SBJ *pro*)
        (NEG não)
        (SR-P-3S é)
        (PP(P @de)
           (NP (D @este)
               (N género)))
        (, ,))
(IP-POL (NEG não)
        (. .))  [CORDIAL, VPA07]

Não é  de este género, não.
not is of this kind,   not

It's not of this kind.
```

Question-tags, another recurring element of spoken language, were also marked with a distinct extended label –TAG. CP-QUE-TAGs include only overt elements, without additional structure.

(14)

```
(IP-MAT (NP-SBJ *pro*)
        (TR-P-3P têm)
        (NP-ACC (D-UM um)
                (N bocadinho)
                (PP (P de)
                    (NP (N ferrugem))))
        (, ,)
        (CP-QUE-TAG (NEG não)
                    (TR-P-3P têm))
        (. ?))  [CORDIAL, VPA36]

têm  um bocadinho de ferrugem, não têm?
have a  little    of rust,     not have

They are a bit rusty, aren't they?
```

(15)

```
(IP-MAT (NP-SBJ *pro*)
        (VB-D-1S @Amostrei)
        (NP-DAT (CL @lhe))
        (NP-ACC *ICH*-252)
        (ADVP (ADV ontem))
        (DS -)
        (CP-QUE-TAG (NEG não)
                    (VB-D-1S amostrei)
                    (. ?))
        (DS -)
        (, ,)
        (NP-252 (D-F a)
                (N lula))
        (. .))  [CORDIAL, VPA36]

Amostrei-lhe ontem     – não amostrei? – a   lula
showed   you yesterday – not showed?   – the squid

I showed you the squid yesterday – didn't I?
```

Lastly, we created the generic extended label –PRG (pragmatic), which may apply to different constituents acting as pragmatic markers (Fraser, 1996), such as discourse connectives or parallel markers (e.g. *Well* / *You know?* / *Isn't it?* / *Look!*, among others). As in the previous cases, inside constituents labelled -PRG, empty categories are never annotated.

(16)

```
(IP-MAT (NP-SBJ *pro*)
        (ADVP-PRG (ADV bom))
        (, ,)
        (VB-P-3P fazem)
        (ADVP (ADV assim))
        (. .))  [CORDIAL, PST01]

bom,  fazem assim
well, do    like.this

Well, they do like this
```

(17)

```
(IP-MAT (NP-LFD (D-F A)
                (N pesca))
        (, ,)
        (CP-IMP-PRG (VB-SP-3S olhe))
        (, ,)
        (VB-P-3S @larga)
        (NP-SE-14 CL @se))
        (NP-SBJ-14 (D-F a)
                   (N rede))
        (PP (P por)
            (NP (D-F a)
                (N borda)))
        (. .)) [CORDIAL, VPA05]

A pesca,     olhe, larga-se  a   rede        por a   borda
the fishing, look, throw SE_CL the fishing.net by  the board

Look. The fishing net is thrown overboard.
```

(18)

```
(CP-D (C Porque)
      (IP-IND (, ,)
              (ADVP (ADV agora))
              (, ,)
              (NP-SBJ (D-P esses)
                      (N-P barcos))
              (TR-P-3P têm)
              (NP-ACC (D-F-P essas)
                      (N-P sondas)
                      (PP (P de)
                          (NP (N-P choques))))
              (, ,)
              (CP-QUE-PRG (ET-P-3S está)
                          (PP (P a)
                              (IP-INF (VB perceber)))
                  (. ?)))  [CORDIAL, VPA26]

Porque,  agora, esses barcos têm  essas sondas de choques,
because, now,   those boats  have those probes of shocks

está a perceber?
is   understanding

Because, now, those boats have electric fishing probes, you see?
```

2.3.2. The annotation schemes

The annotation of finite clause complements represents a case in which the original scheme defined for English could be applied to the Portuguese data in a straightforward way.

The embedded clause is labelled CP-THT. The complementizer (C) and the subordinate IP (IP-SUB) are annotated as CP-THT immediate constituents:

(19)

```
(IP-MAT-PRN (NP-SBJ *exp*)
            (VB-P-3S parece)
            (CP-THT (C que)
                    (IP-SUB (NP-SBJ *pro*)
                            (SR-D-3S era)
                            (NP-PRD (N pintor)))))
            [...] [CORDIAL, AAL04]

Parece que  era pintor
seems  that was painter

It seems that he was a painter.
```

The non-standard phenomenon of recomplementation (or complementizer doubling) could easily find a codification along the same lines. Sentences involving a doubly filled complementizer position are annotated as two instances of CP-THT; the phrase positioned between the two complementizers is immediately dominated by the higher CP-THT and co-indexed with an empty *ICH* (for "interpret constituent here"):

(20)

```
(IP-MAT (NP-SBJ *exp*)
        (VB-P-3S Parece)
        (CP-THT (C que)
                (PP-1 (P @em)
                      (NP (D-F @a)
                          (NPR Suíça)))
                (CP-THT (C que)
                        (IP-SUB (PP *ICH*-1)
                                (NP-SBJ *pro*)
                                (VB-P-3P dão)
                                (NP-ACC (Q-F muita)
                                        (N importância))
                                (PP (P a)
                                    (NP (D-F-P essas)
                                        (N-P coisas))))))
        (. .))  [CORDIAL, AAL04]

Parece que  em a  Suíça        que  dão
seems  that in the Switzerland that give

muito importância a  essas coisas
very  importance  to those things

It seems that they attach great importance to those things in
Switzerland.
```

For the treatment of clitic climbing (a nonexistent phenomenon in English) we had to slightly adapt the original system, making use of an already available option. We applied the annotation scheme of A-movement to the annotation of clitic climbing.

A-movement traces are annotated when the extraction involves more than a single IP. This sort of trace is codified using the symbol *, which is co-indexed with the moved phrase (see (21)). In clitic climbing contexts, the annotation marks the connection between the climbed clitic and the clause where it originates by means of the same co-indexed null element * (see (22)).

(21)

```
(IP-MAT (NP-SBJ-1 (DEM Isso))
        (SR-P-3S é)
        (VB-AN considerado)
        (IP-SMC (NP-SBJ *-1)
                (NP-ACC (N crime))))

Isso é considerado crime
this is considered crime

This is considered a crime.
```

(22)

```
(IP-MAT (NP-SBJ *pro*)
        (ADV também)
        (NP-25 (CL o))
        (VB-P-1S vou)
        (IP-INF (NP-ACC *-25)
                (VB levar))
        (. ...))  [CORDIAL, VPA14]

também o  vou levar
also   it go  take

I will take it too.
```

Finally, let's see an example of how we expanded a defined annotation scheme to cover the different types of dialectal relative clauses.

Following the Penn system, a relative clause with an antecedent is CP-REL, which immediatelly dominates WXP (the projection of the relative pronoun) and IP-SUB. The WXP is co-indexed with a wh-movement trace (*T*) in the first position inside IP-SUB:

(23)

```
(IP-MAT [...]
        (NP-SBJ (D-P os)
                (N-P coelhos)
                (CP-REL (WNP-1 (WPRO que))
                        (IP-SUB (NP-SBJ *T*-1)
                                (VB-P-3P comem)
                                (NP-ACC (DEM isto)))))
        (VB-P-3P morrem)
        (. .))  [CORDIAL, AAL01]

os  coelhos que  comem isto morrem
the rabbits that eat   this die

The rabbits that eat this die.
```

The same applies to the annotation of free-relatives, except for their external label, which is now CP-FRL:

(24)

```
(IP-MAT (ADVP (ADV Também))
        (NP-SBJ *exp*)
        (HV-P-3S há)
        (NP-ACC (CP-FRL (WNP-1 (WPRO quem))
                        (IP-SUB (NP-SBJ *T*-1)
                                (VB-SP-3S faça)
                                (NP-ACC (D-F essa)
                                        (N coisa)))))
        (. .))  [CORDIAL, AAL02]

Também há       quem faça essa coisa
also   there.is who  does that thing

There is also who does that.
```

For relatives where a pied-piped preposition has not been produced (chopping relatives), we adopted an annotation scheme that was already defined for the deletion of stranded prepositions in English: the omitted preposition is annotated in curly brackets and labelled CODE:

(25)

```
(IP-MAT (NP-SBJ-2 *exp*)
        (SR-D-3S Era)
        (NP-2 (D esse)
              (N alqueive)
              (CP-REL (WPP-4 (P (CODE {em}))
                             (WNP (WPRO que)))
                      (IP-SUB (PP *T*-4)
                              (NP-SBJ (D-F a)
                                      (N gente))
                              (VB-P-3S lavra))))
        [...]  [CORDIAL, AAL14]

Era esse alqueive    {em} que  a   gente  lavra
was that fallow.land {in} that the people plough

It was that fallow land that we plough.
```

To deal with the particular case of resumptive relatives we had to create a new variant upon the Penn solution. This new scheme was conceived in order to remain consistent with the standard annotation of relatives. Thus, the word *que* is still represented inside a WNP. However, no trace of movement is annotated. Instead, the resumptive pronoun is marked with the extended label -RSP (for resumptive).

(26)

```
(IP-MAT (CONJ e)
        (VB-P-3S fica)
        (NP-SBJ (D-F aquela)
                (N coisa)
                (CP-REL (WNP (WPRO que))
                        (IP-SUB (NP-SBJ-1 *exp*)
                                (ADV-NEG nunca)
                                (NP-SE-1 (CL se))
                                (NP-OBL-RSP (CL lhe))
                                (VB-P-3S mexe))))
        (. .))  [CORDIAL, AAL15]

e fica aquela coisa que nunca se lhe mexe
and stays that thing that never SE_CL it touches

And that thing stays there and is never touched.
```

3. References

ALEAç – Atlas Linguístico e Etnográfico dos Açores (J. Saramago, coord.) [http://www.clul.ul.pt/english/sectores/variacao/projecto_aleac.php]

ALLP – Atlas Linguístico do Litoral Português (G. Vitorino, coord.) [http://www.clul.ul.pt/english/sectores/variacao/projecto_allp.php]

ALEPG – Atlas Linguístico-Etnográfico de Portugal e da Galiza (J. Saramago, coord) [http://www.clul.ul.pt/english/sectores/variacao/projecto_alepg.php]

BA – Segura, M. L. (1987). A Fronteira Dialectal do Barlavento do Algarve. Diss. PhD. University of Lisbon.

Bikel, D. (2004). On the Parameter Space of Generative Lexicalized Statistical Parsing Models. Diss. PhD. University of Pennsylvania.

Carrilho, E. & C. Magro (2009). CORDIAL-SIN – Syntactic Annotation System Manual. [http://www.clul.ul.pt/english/sectores/variacao/cordialsin/Syntactic%20annotation%20manual.html]

Collins, M. (1999). Head-Driven Statistical Models for Natural Language Processing. Diss. PhD. University of Pennylvania.

CORDIAL-SIN – The Syntax-oriented Corpus of Portuguese Dialects (A. M. Martins, coord.) [http://www.clul.ul.pt/english/sectores/variacao/cordialsin/projecto_cordialsin_corpus.php]

EDISYN – European Dialect Syntax Project (S. Barbiers, coord) [http://www.meertens.knaw.nl/edisyn]

Finger, M. (1998). Tagging a morphologically rich language. In *Proceedings of the First Workshop on Text, Speech and Dialogue (TSD'98),* pp. 39--44.

Finger, M. (2000). Técnicas de Otimização da precisão empregadas no etiquetador Tycho Brahe. In *Proceedings of V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR 2000),* pp. 141-154.

FRASER, B. (1996). Pragmatic Markers. *Pragmatics*. 6(2), pp. 167--190.

Kroch, A. & A. Taylor (2000). Penn-Helsinki Parsed Corpus of Middle English, second edition. [http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html]

Kroch, A., B. Santorini & L. Delfs (2004). Penn-Helsinki Parsed Corpus of Early Modern English. [http://www.ling.upenn.edu/emodeng]

Kroch, A. & B. Santorini & A. Diertani (2010). Penn Parsed Corpus of Modern British English. [http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html]

Magro, C. & C. Morgado (orgs.) (2008). CORDIAL-SIN – POS Annotation Manual. [http://www.clul.ul.pt/english/sectores/variacao/cordialsin/pos_annotation_manual.pdf]

Randall, B. (2005-2007). CorpusSearch 2. [http://corpussearch.sourceforge.net]

Tycho Brahe Parsed Corpus of Historical Portuguese (C. Galves, coord.). [http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html]