

Morphological Tagging and Syntactic Annotation of a Dialectal European Portuguese *Corpus*

Ernestina Carrilho*, Catarina Magro†, Sandra Pereira†

* Faculdade de Letras de Lisboa / Centro de Linguística da Universidade de Lisboa

† Centro de Linguística da Universidade de Lisboa
Av. Gama Pinto, 2 – 1649-003 Lisboa
{e.carrilho, cmm, spereira}@clul.ul.pt

Abstract

This presentation reports on an ongoing project of morphologically tagged and syntactically annotated *corpus* of spoken nonstandard European Portuguese. Issues pertaining to the tagging and the annotation processes will be addressed from a linguistic perspective, focused on the structure and application of the tagsets used for annotating this *corpus*.

1. Introduction

The *Syntactically Annotated Corpus of Portuguese Dialects* (CORDIAL-SIN, from the Portuguese name *Corpus Dialectal com Anotação Sintáctica*) is an ongoing project of annotated *corpus* of spoken dialectal European Portuguese (henceforth EP). It started in September 1999 as a first year pilot-study (funded by FCT - PRAXIS XXI/P/PLP/13046/1998), further developed as a three years project (POSI/1999/PLP/33275) by a team of five linguists, under the coordination of Ana Maria Martins at the Centro de Linguística da Universidade de Lisboa (CLUL).

The project main goal is to build up a major resource for linguistic research on dialects. It aims at providing optimal access to precise morphological and syntactic information, ultimately enhancing the study of dialect syntax, a field with no tradition in the Portuguese domain.

The *corpus* consists of a geographically representative body of selected excerpts of spontaneous and semi-directed speech. These materials were drawn from an independently existing rich collection of speech which had been recorded within the scope of several projects of the Variation Research Team of the CLUL, namely, the *Atlas Linguístico-Etnográfico de Portugal e da Galiza* (ALEPG); the *Atlas Linguístico do Litoral Português* (ALLP); the *Atlas Linguístico e Etnográfico dos Açores* (ALEAç); and the *Frenteira Dialectal do Barlavento Algarvio* (BA).

At the current state, the excerpts of dialectal speech selected for the *corpus* cover 24 localities within the continental and insular territory of Portugal, amounting to about 300,000 words. The *corpus* is available via internet (http://www.clul.ul.pt/sectores/cordialsin/projecto_cordial_sin.html), under different formats: (i) *verbatim* orthographic transcripts; (ii) normalized orthographic transcripts; (iii) morphologically tagged versions of the normalized transcripts; (iv) syntactically annotated texts built on the morphologically tagged versions.

Verbatim orthographic transcripts include the marking up of phonetic and morphological variants, and of generalized spoken language phenomena, such as hesitations, filled and empty pauses, repetitions, rephrased segments, false starts, truncated words, speech overlappings, unclear productions, etc. From these *verbatim* transcripts, normalized orthographic transcripts

are automatically obtained by eliminating the marked up features of spoken language and phonetic transcriptions. The tagging and the syntactic annotation apply over the normalized transcripts.

Verbatim transcripts, normalized orthographic transcripts and morphologically tagged texts are gradually made available online as the *corpus* building up proceeds. Since the syntactic annotation guidelines may not be completely established before the end of the annotation process, the syntactically annotated transcripts will not become available until the project is concluded.

In this paper, we will focus on the tagging and annotation phases of this *corpus*, which are greatly inspired by the systems used by the *Penn-Helsinki Parsed Corpus of Middle English, second edition* (henceforth PPCME2, see <http://www.ling.upenn.edu/mideng>) (Kroch & Taylor, 2000) and the *Tycho Brahe Parsed Corpus of Historical Portuguese* (henceforth TB, see <http://www.ime.usp.br/~tycho/corpus>). Collaborative work with the teams developing these *corpora* has permitted the tuning of already available tagging and annotation tools, in such a way that they could satisfactorily apply to dialectal EP and serve the purposes of the CORDIAL-SIN. Besides accelerating the tagging and annotation phases, this cooperation ensures the ease of linguistic information retrieval (a query tool operating on the annotation system in use is already available – cf. PPCME2 web page).

In the following sections we describe the main guidelines of the tagging and annotation systems adopted from the TB and the PPCME2, emphasizing on the structure and application of the tagsets as developed within the scope of the CORDIAL-SIN.

2. CORDIAL-SIN Morphological Tagging

2.1. The tagging process

The morphological tagging operation is to a great extent facilitated by the use of an automated morphological tagger, created by M. Finger for tagging the TB *corpus* of Portuguese texts (written by Portuguese authors born from the sixteenth to the nineteenth centuries). After training over a sample of 30,000 hand corrected words of the dialectal *corpus*, the rate of accuracy of this tagger proved to be satisfactory enough to encourage the use of its output as the basis for a hand

refined (and corrected) tagged version of the *corpus*. An additional TB tool designed for verifying the tags corrected by hand is used after manual refinement and correction to ensure the precise format of the tags. Thus, CORDIAL-SIN's morphologically tagged transcripts result from a three steps process involving: (i) automatic tagging by the TB tagger; (ii) manual tag correction and refinement using the CORDIAL-SIN's morphological annotation system; (iii) automatic verification of the corrected tags.

2.2. The morphological annotation system

The format of the morphological tags and the basics of the tagset of the CORDIAL-SIN essentially stem from the system designed for the TB automatic tagger (cf. Galves & Britto, 1999, Britto et al., 1999, and *The TB Morphological Annotation System* www.ime.usp.br/~tycho/corpus/manual/tags.html).

Tags have an internal structure consisting of an ever-present main tag (e.g. D, for determiner), and, in certain cases, sub-tags (e.g. F for feminine, P for plural), diacritics attaching different main tags (“+”, “!”) or main tags to sub-tags (“.”), and figures indicating clusters (see Table 1 for overview).

Tag	Application	Ex.
/D	singular masculine determiner	<i>o/D</i>
/D-P	plural masculine determiner	<i>os/D-P</i>
/D-F-P	plural feminine determiner	<i>as/D-F-P</i>
/P+D-F	preposition plus singular feminine determiner contraction	<i>da/P+D-F</i>
/VB+CL	verb (infinitive) plus enclitic pronoun	<i>dar-lhe/VB+CL</i>
/VB-R-1S!CL	verb (future) plus mesoclitic pronoun	<i>dar-te-ei/VB-R-1S!CL</i>
/P31	first element of a triple prepositional cluster	<i>por/P31</i> <i>mor/P32</i> <i>de/P33</i>

Table 1: Morphological tags' internal structure

Such structured tags straightforwardly allow for detailed morphological information, which is a highly appealing option when tagging a morphologically rich language such as EP¹. Indeed, for a number of possible structured tags as high as 1115, the CORDIAL-SIN tagset reduces to 39 main tags plus a smaller set of 25 sub-tags.

¹ On the architecture of the TB tagger, especially designed with such a tag system, and on how it permits to increase the degree of accuracy of Brill's (1993, 1995) tagging method on a morphologically rich language, see Finger (1998, 2000).

Main tags include POS tags and punctuation tags. The complete CORDIAL-SIN main tagset is given in Table 2.

Main Tag	Application
SR	verb SER
HV	verb ESTAR
ET	verb HAVER
TR	verb TER
VB	all other verbs
N	common nouns
NPR	proper nouns
PRO	personal pronouns
PRO\$	possessive pronouns
CL	clitics in general
SE	clitic SE
D	definite determiner and inflected demonstratives
DEM	invariable demonstratives
ADJ	general adjectives and ordinal numbers
ADV	adverbs and speech connectives
Q	quantifiers
CONJ	coordinating conjunctions
CONJS	subordinating conjunctions
C	complementizer
WPRO	Wh-pronouns
WPRO\$	possessive Wh-pronouns
WADV	Wh-adverbs
WD	Wh-determiners
P	prepositions
FP	focus particles
NUM	cardinal numbers
NEG	negative particle
INTJ	interjections and onomatopoeias
OUTRO	the word <i>outro/a</i> (all cases)
SENÃO	the word <i>senão</i> (all cases)
COISO	the word <i>coiso/a</i> (when replacing a word of any category)
MESMO	the word <i>mesmo/a</i> (with a determiner and no name)
TAL	the word <i>tal</i> (with a determiner and no name)
MAL	the word <i>mal</i> (in predicative / transitive constructions, alternating with the adjective or the DO)
BEM	the word <i>bem</i> (in predicative / transitive constructions, alternating with the adjective or the DO)
.	final punctuation
,	non-final punctuation
QT	quotation marks
DS	dash

Table 2: Main tagset

The set of sub-tags codifies inflectional information – tense/mood and person/number for verbs or gender and number for nominal categories. It also specifies in more detail some morpho-syntactic information (e.g. the -NEG sub-tag to identify negative adverbs, quantifiers, prepositions, focus particles or conjunctions).

The system also allows main tags attachment for contractions or cliticizations and tags and figures combination for multiple words behaving as clusters.

For a detailed description of the tagset and its application, see *CORDIAL-SIN — Manual de Anotação Morfológica* (www.clul.ul.pt/sectores/cordialsin/manual_annotacao_morfologica.pdf).

The enhancements introduced by the CORDIAL-SIN project on the original *TB* tagset are the addition of (i) new word specific main tags; (ii) new person/number inflectional sub-tag for verbs; and (iii) a new NEG sub-tag for negative words. The project also makes a more extensive use of clusters and sub-tag distribution and endorses a wider application of multi-tagging strategy.

This refinement of the initial system, implemented during the phase of manual correction of tags, serves a twofold purpose. Above all, it helps disambiguating morphological information relevant for queries on the current annotated version of the *corpus*. On the other hand, such specific information gives a richer input to the syntactic annotation phase.

3. CORDIAL-SIN Syntactic Annotation

3.1. The syntactic annotation process

Differently from the morphological annotation phase, the process of syntactic annotation is entirely developed by hand. The option for such a time-consuming task is plainly justified by the nature of the CORDIAL-SIN *data* and by the type of rich annotation aimed at.

Manual syntactic annotation is introduced over morphologically annotated texts, with the aid of an annotation tool working in ambient Linux (the tool actually used by the PPCME2 for correcting the output of an automated parser)².

As already pointed out, the CORDIAL-SIN syntactic annotation system is highly inspired by the PPCME2 system (see <http://www.ling.upenn.edu/~ataylor/ppcme-lite.htm>). The adoption of this type of rich annotation system for a Portuguese *corpus* required the adaptation of the existing system to a grammar which differs from Middle English in many respects. Accordingly, the initial phase of the CORDIAL-SIN syntactic annotation process has been devoted to the tuning of the basic annotation system, a task which was carried out in strict collaboration with the PPCME2 and the TB teams.³ Hand annotation of a 10,000 words sample of the *corpus* has served to define and consolidate the main guidelines of the system so as it could apply to Portuguese texts.

² This tool consists of a task-specific mouse-based package, which is embedded in the GNU Emacs editor.

³ In particular, with Anthony Kroch and Helena Britto, respectively. A first proposal of the Portuguese system was discussed with A. Kroch in December 2000, and a further extended version of the system was established with H. Britto in April 2002.

As is well known, real *data* annotation itself is usually a very complex task. In the present case, additional complexity was expected, given the spoken and dialectal nature of the *corpus*. Sentences that call for detailed consideration are frequent, even though the basic lines of the system are already defined. Difficult annotations are decided upon after discussion by the whole team, and each new difficult example is added to the annotator's manual, in order to assure consistency. Thus, it is expected that the syntactic annotation guidelines will be progressively enriched during the whole course of the annotation phase, as more data are analysed and as new difficult sentences arise. (See http://www.clul.ul.pt/english/sectores/cordialsin/manual_syntactic_annotation_system.pdf, for the current version of the *Syntactic Annotation Manual*).

3.2. The annotation system

3.2.1. Main guidelines

The CORDIAL-SIN syntactically annotated transcripts are built on previously tagged texts. The syntactic annotation produces a tree representation in the form of labeled brackets.

```
(IP-MAT (CONJ e)
  (NP-SBJ *pro*)
  (VB-D-1P andávamos)
  (PP (P com)
    (NP (D-F-P as)
      (N-P redes)
      (PP (P+D do)
        (NP (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))))))
  (. ...)) [VPA07]
```

Figure 1. CORDIAL-SIN syntactically annotated sentence

As in the PPCME2, the annotation represents quite flat trees, allowing for multiple branching nodes and for some words projecting only a word-level node (e.g. inflected verbs, negation, sentence focus particles).

In addition to constituent boundaries and phrase and clause dependencies, the annotation marks up grammatical relations, clause-types, some empty categories and some transformational relations. At the word level, morphological labels are preserved. Phrase and clause labels indicate category, often specified by an extended label indicating syntactic function (e.g. subject, direct object), clause type (e.g. relative, adverbial, interrogative), or other relevant information (e.g. left dislocation, pragmatic marker).

3.2.2. Labels and extended labels

Even though most labels and extended labels come originally from the PPCME2 system, a restricted number of additional labels were introduced for the EP annotation. In particular, some new extended labels were created for the CORDIAL-SIN use, especially adapted to spoken data

annotation (e.g. -CON for pragmatic markers, and -ANS, -POL, -TAG, cf. Table 4). Tables 3 and 4 show the main label set used in the CORDIAL-SIN syntactic annotation. (The complete set is available online, see *Syntactic Annotation Manual*).

Label	Category (and function)
NP	Noun Phrase
NP-SBJ	Noun Phrase (Subject)
NP-ACC	Noun Phrase (Direct Object or 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)

Table 3: CORDIAL-SIN phrase labels

Label	Category (and function)
IP-MAT	Independent or conjoined declarative IP
IP-IND	Independent, non-declarative IP
IP-SUB	Subordinate IP
IP-ADV	Adverbial IP
IP-INF	Infinitival clause
IP-GER	Gerund clause
IP-PPL	Participial clause
IP-SMC	Small clause
IP-ANS	Answer
IP-POL	Reinforcement of an assertion
CP-EXL	Exclamative
CP-IMP	Imperative
CP-QUE	Question
CP-QUE-TAG	Question-tag
CP-INF	Infinitive introduced by <i>que</i>
CP-THT	<i>That</i> clause
CP-REL	Relative
CP-FRL	Free Relative
CP-CLF	Cleft
CP-ADV	Adverbial clause
CP-DEG	Degree clause
CP-CMP	Comparative clause

Table 4: CORDIAL-SIN clause labels

3.2.3. Adapting the PPCME2 system to EP

Besides the addition of some new extended labels, the adaptation of the PPCME2 annotation system to EP *corpora* essentially required the conception of additional ways of codifying new syntactic constructions, within the possibilities offered by the system (and, consequently, by the annotation tool). For instance, the CORDIAL-SIN/TB system includes unambiguous codification for most clitics, adding information on clitic climbing or exceptional case marking contexts, which was not required for the PPCME2 annotation. Also, the codification of certain types of constructions (such as clefts and topicalization/left-dislocation) implied, for the EP *corpora*, the creation of new variants upon the PPCME2 solutions, given the diversity of related constructions allowed by EP.

The annotation system so designed for the CORDIAL-SIN is thus compatible with CorpusSearch, a linguistically intuitive query tool, especially developed by Beth Randall for use with the PPCME2⁴, which ultimately permits fast and massive information retrieving on relevant aspects of the syntax of the CORDIAL-SIN *data*.

4. References

- Brill, Eric, 1993. *A Corpus-Based Approach to language Learning*. PhD thesis, University of Pennsylvania.
- Brill, Eric, 1995. Transformation-based error-driven learning and Natural Language Processing: a case study in part-of-speech tagging. *Computational Linguistics* 21: 543-565.
- Britto, Helena, Charlotte Galves, Ilza Ribeiro, Marina Augusto, and Ana Paula Scher, 1999. Morphological annotation system for automatic tagging of electronic textual *corpora*: from English to Romance languages. In *Proceedings of the 6th International Symposium of Social Communication*, Santiago de Cuba. Editorial Oriente. 582-589.
- Finger, Marcelo, 1998. Tagging a Morphologically Rich Language. In *Proceedings of the First Workshop on Text, Speech and Dialogue (TSD'98)*. Brno, Czech Republic. 39-44.
- Finger, Marcelo, 2000. Técnicas de Otimização Empregadas no Etiquetador Tycho Brahe. In *Proceedings of V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR 2000)*. Atibaia, Brazil.
- Galves, Charlotte and Helena Britto, 1999. A construção do *Corpus Anotado do Português Histórico Tycho Brahe*: o sistema de anotação morfológica. In I. Rodrigues and P. Quaresma (eds.) *Proceedings of the IV PROPOR*. Évora. Universidade de Évora. 55-67.
- Kroch, Anthony S. and Ann Taylor, 2000. *The Penn-Helsinki Parsed Corpus of Middle English, Second Edition*. Department of Linguistics, University of Pennsylvania.

⁴ On this tool, see <http://www.ling.upenn.edu/mideng/CS-manual.pdf>.