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

Ernestina Carrilho[1], Catarina Magro[2], Sandra Pereira[2]

[1] Faculdade de Letras de Lisboa / Centro de Linguística da Universidade de Lisboa
[2] Centro de Linguística da Universidade de Lisboa
Av. Gama Pinto, 2 – 1649-003 Lisboa
{ecarrilho, cmm, spereira}@clul.ul.pt

Abstract. This presentation reports on an ongoing project of morphologically tagged and syntactically annotated *corpus* of spoken non-standard 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) under development at Centro de Linguística da Universidade de Lisboa (CLUL). 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, coordinated by Ana Maria Martins, with Anthony Kroch, Charlotte Galves and João Saramago as consultants.

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 *Fronteira Dialectal do Barlavento Algarvio* (BA).

At the current state, the excerpts of dialectal speech selected for the *corpus* cover 24 locations 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/english/sectores/cordialsin/projecto_cordialsin.html), under different formats: (i) *verbatim* orthographic transcripts; (ii) normalized orthographic tran

scripts; (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 ASCII versions of the normalized transcripts are the input for the tagging and the syntactic annotation. These ASCII versions include some delimiters that identify expressions to be ignored by the annotation tools, such as headings, codes for supressions, etc.

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 [7]) 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 [6], Britto et al. [3], 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, subtags (e.g. F for feminine, P for plural), diacritics attaching different main tags ("+", "!") or main tags to subtags ("-"), and figures indicating clusters (see Table 1 for overview).

Table 1. Morphological tags' internal structure

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

Such structured tags straightforwardly allow for detailed morphological information, which is a highly appealing option when tagging a morphologically rich language such as EP.[1] 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 subtags.

[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 tagging method (cf. Brill [1] & [2]) on a morphologically rich language, see Finger [4] & [5].

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

Table 2. Main tagset

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	coord. conjunctions
CONJS	subord. conjunctions
C	complementizer

Tag	Application
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 *outro/a*
SENÃO	the word *senão*
COISO	the word *coiso/a* (when replacing a word of any category)
MESMO	the word *mesmo/a* (with a determiner and no name)
TAL	the word *tal* (with a determiner and no name)
MAL	the word *mal* (in predicative / transitive constructions, alternating with the adjective or the DO)
BEM	the word *bem* (in predicative / transitive constructions, alternating with the adjective or the DO)
.	final punctuation
,	non-final punctuation
QT	quotation marks
DS	dash

The set of subtags 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 -R and the -S subtags, indicating the comparative and superlative values of adjectives).

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

The enhancements introduced by the CORDIAL-SIN project on the original *TB* tagset are:

i. the development of new word specific main tags, such as COISO, TAL ou BEM, enlarging the strategy used by TB for the words *outro* and *senão*;

(1) Eu ando assim coisa/COISO porque aqui não o há.
(2) Depois de crescer, tira-se a manta, tira-se a tal/TAL, põe-se no tabuleiro.
(3) Ah, bem/BEM aos olhos faz ele tudo.

ii. the development of new verbal inflectional subtags (person/number);

(4) Agora, muitas vezes, é entregarmo-nos/VB-F-1P+CL só às mãos do doutor.
(5) Calai/VB-I-2P, cala-te/VB-I-2S+CL Agatão.

iii. the development of a new negation subtag. This -NEG subtag can attach to different maintags — ADV, P, CONJ, Q or FP — codifying the negative value of these words;

(6) Não há peixe que se ponha-se ao sol sem/P-NEG salgar.
(7) Não queria saber nem/CONJ-NEG de igrejas, nem/CONJ-NEG de coisa nenhuma/Q-NEG-F.
(8) É agora cá/FP-NEG uma louva-a-Deus e não mexe!

iv. the application of existing subtags to new contexts, such as the -F verbal inflectional subtag, created by TB for the inflected infinitive, which is also used by CORDIAL-SIN for inflected gerund (a particular phenomenon of verbal morphology of dialectal EP);

(9) Em sendem/SR-G-F-3P muitos, já se lhe chama uma vara.

v. the wider application of the 'word as unit' strategy. Besides prepositional locutions, CORDIAL-SIN applies this treatment to complex proper nouns, complex numerals, adverbial locutions and conjunctional locutions;

(10) E essa casa era alugada à senhora/NPR31 dona/NPR32 Agrícia/NPR33.
(11) E houve aqui um barco que já caçou cento/NUM31 e/NUM32 tal/NUM33 corvinas.
(12) É uma manta mas, afinal/ADV31 de/ADV32 contas/ADV33, ele há um pano que não tem o nome de manta.
(13) E de/CONJ31 maneira/CONJ32 que/CONJ33 semeei lá uns nabos e uns rabanetes.
(14) Mesmo até os rebentos que arrebentem pela cepa acima dá cachos, ao/CONJS31 passo/CONJ32 que/CONJS33 estes já não são assim.

vi. the extension of multi-tagging strategy, exploring the possibility of associating a single word to multiple tags.

(15) Está certo, Sr Enfermeiro. Se é para meu bem/N, bote.
(16) …para ver se eles estão a falar bem/ADV ou se estão a falar mal.
(17) Iam assim até ralas, para ficarem bem/BEM no pisão.
(18) E é espesso quase como uma enxó. Não é bem/FP a enxó, que é mais direita.
(19) Ora, mas eles, ainda/ADV31 bem/ADV32 não/ADV33, lá chegavam ao pé da gente. Aquilo, ainda/ADV31 bem/ADV32 não/ADV33, era uma derrota.
(20) Está calor, se/CONJS31 bem/CONJS32 que/CONJS33 corra uma aragem.

These refinements of the initial system, implemented during the phase of manual correction of tags, serve several purposes: 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 and adequates the existent tagset to spoken non-standard data.

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* (spoken and non-standard) and by the kind of 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

[2] This tool consists of a task-specific mouse-based package, which is embedded in the GNU Emacs editor. It enables the annotator to add constituent boundaries, pre-defined labels, some empty categories and co-indexing, and to perform any kind of correction on the annotation without affecting the transcripts.

[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.

define and consolidate the main guidelines of the system so as it could apply to Portuguese texts. These general guidelines resulted in the first version of the annotator's manual.

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))))))
        (. ...))
```

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 and sentence type, some empty categories (such as null subject and null object, among others) and some transformational relations (such as wh-movement in relatives and questions). 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 ob

ject), clause type (e.g. relative, adverbial, interrogative), or other relevant information (e.g. left dislocation, pragmatic marker).

3.2.2 Labels and Extended Labels

Most labels and extended labels originally come from the PPCME2 system. Table 3 shows the main label set used in the CORDIAL-SIN syntactic annotation. (The complete set is available online, see *Syntactic Annotation Manual*).

Table 3. CORDIAL-SIN phrase and clause labels

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)

Label	Category (and type)
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 *que*
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

3.2.3 Adapting the PPCME2 system to EP

Concerning the label set, it must be added that, besides the original PPCME2 labels, a restricted number of additional labels were introduced for the CORDIAL-SIN annotation purposes. In particular, some new extended labels were created for the CORDIAL-SIN use, such as -CON, -ANS, -POL, -TAG (cf. Table 3 above). Such labels were particularly needed to set apart some syntactic units that abound in spoken texts, and whose internal structure is not represented (to avoid adding extra complexity to the annotation).

The extended label -CON adjoins to different main labels (IP, CP, ADVP) to mark up different kinds of pragmatic markers (for instance, markers used to gain the hearer attention, as in the following example):

```
(21)
(IP-MAT   (CONJ porque)
          (NP-SBJ(D aquele)
                 (N senhor))
          (SR-P-3S é)
          (PP (P+D-F da)
              (NP (N religião)
                  (PP (P+D-F da)
                      (NP (N verdade)))))
          (, ,)
          (CP-QUE-CON (VB-P-3S sabe))
          (. ?))
```

Answers to both yes-no questions and wh-questions are marked up as IP-ANS:

```
(22)
(CODE <inq> INQ1 E trazia-as já feitas?  </inq>)
(CODE <inf> INF </inf>)
(IP-ANS   (VB-D-1S Trazia)
          (. .))

(23)
(CODE <inq> INQ2 Tinha coisas para respirar? </inq>)
(CODE <inf> INF </inf>)
(IP-ANS   (ADVP (ADV-NEG21 Não)
                (ADV-NEG22 senhora))
          (. .))
```

The extended label -POL adjoins to the label IP to annotate polarity items that occur as after-thought or appositive elements that reinforce the polarity value of an assertion:

```
(24)
(IP-MAT   (NP-SBJ *pro*)
          (NEG não)
          (SR-P-3S é)
          (PP (P+D deste)
              (NP (N género)))
          (, ,))
(IP-POL    (NEG não)
           (. .))
```

Finally, the extendend label -TAG combined with the label CP-QUE identifies question-tags:

```
(25)
(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))
          (. ?))
```

Besides the addition of new extended labels, the adaptation of the PPCME2 annotation system to EP *corpora* essentially required the conception of additional ways of codifying different syntactic constructions. In certain cases, non-standard EP could easily find a codification along the lines proposed for Middle English - this was the case, for instance, of double complementatizer (*que…que*) structures such as in the following example:

```
(26)
(IP-MAT (NP-SBJ(PRO Eu))
        (VB-P-1S sei)
        (CP-THT (C que)
                (NP-2 (DEM aquilo))
                (CP-THT(C que)
                       (IP-SUB (NP-SBJ *ICH*-2)
                               (NEG não)
                               (SR-P-3S é)
                               (PP (P por)
                                   (ADVP (ADV mal))))))
        (, ,)
        (CP-QUE-CON (VB-P-3S sabe))
        (. ?))
```

In most cases, however, new codifications had to be conceived, within the possibilities offered by the original 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, relatives and topic constructions) implied, for the EP *corpora*, the creation of new variants upon the PPCME2 solutions, given the diversity of related constructions allowed by EP. Thus, although the CORDIAL-SIN annotation for standard relatives (see ex. (27)) and for free relatives (see ex. (28)) takes up the annotation schema already proposed in the PPCME2 system, other types of relative clauses, such as resumptive ones (see ex. (29)), chopping relatives (see ex. (30)) and *é que* relatives (see ex. (31)), required detailed consideration leading to new annotations:

```
(27)
(NP-ACC   (D-UM-F uma)
```

```
          (N senhora)
          (CP-REL(WNP-1 (WPRO que))
                 (IP-SUB(NP-SBJ *T*-1)
                        (TR-D-3S tinha)
                        (NP-ACC(NUM sete)
                               (N-P filhas)))))
(28)
(IP-MAT   (NP-SBJ *exp*)
          (HV-P-3S há)
          (NP-ACC (CP-FRL (WNP-1 (WPRO quem))
                          (IP-SUB(NP-SBJ *T*-1)
                                 (VB-SP-3S largue)
                                 (NP-ACC(D-F a)
                                        (N rede))
                                 (PP (P por)
                                     (NP (D-F a)
                                         (N popa))))))
          (. .))
(29)
(IP-MAT   (CONJ E)
          (ADVP (ADV depois))
          (ADVP (ADV lá))
          (VB-D-3S foi)
          (NP-SBJ(D o)
                 (N barco)
                 (CP-REL(WPP-167 (P (CODE {em}))
                                 (WNP (WPRO que)))
                        (IP-SUB  (PP *T*-167)
                                 (NP-SBJ (PRO eu))
                                 (VB-D-1S andava))))
          (, ,))
(30)
(NP (D-F Essa)
    (ADJ-G tal)
    (N feiticeira)
    (CP-REL  (WNP (WPRO que))
             (IP-SUB(NP-SBJ *pro*)
                    (NP-DAT-RSP-28 (CL lhe))
                    (VB-P-3P chamam)
                    (IP-SMC(PP-SBJ *ICH*-28)
                           (NP-ACC(NP(D-F a)
                                     (N pata-roxa)))))))
(31)
(PP  (P+D-F na)
     (NP (N casa)
         (CP-REL(WADVP-272 (WADV onde))
                (C (SR-P-3S é)
                   (C que))
                (IP-SUB(ADVP *T*-272)
                       (NP-SBJ *pro*)
                       (VB-P-3S dorme))))))
```

The annotation system so designed for the CORDIAL-SIN is compatible with CorpusSearch, a linguistically intuitive query tool, especially developed by Beth Randall for use with the PPCME2[4], which ultimately permits fast and massive information retrieving on relevant aspects of the syntax of the CORDIAL-SIN *data*.

References

1. Brill, Eric, 1993. *A Corpus-Based Approach to language Learning*. PhD thesis, University of Pennsylvania.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.