Helena Britto (Campinas), Marcelo Finger (São Paulo),
Charlotte Galves (Campinas)

Computational and linguistic aspects of the construction of the Tycho Brahe Parsed Corpus of Historical Portuguese

Le Corpus Annoté du Portugais Historique Tycho Brahe est un corpus électronique de libre accès par Internet, composé de textes d'auteurs portugais nés entre les XVIe et XIXe siècles, ayant pour objectif la disponibilisation de 2.000.000 mots annotés morphologiquement et syntaxiquement. Cet article présente les principaux aspects linguistiques et informatiques impliqués dans la construction de ce corpus, en particulier l'élaboration et l'application des systèmes de notation morphologique et syntaxique du point de vue linguistique et le développement d'un outil d'étiquetage automatique créé spécifiquement pour ce corpus.

1. Introduction

Created to be a tool in the study of European Portuguese (EP) in its Classical and Modern periods (Galves et al. 1998), the Tycho Brahe Parsed Corpus of Historical Portuguese is an electronic corpus containing Portuguese prose, written by native speakers of EP born between the 16th and the 19th centuries.[1] The corpus is freely available on the web since 1999 (<http://www.ime.usp.br/~tycho/corpus>), in agreement with the view that scientific data must be public. Its texts are always extracted from reliable philological editions and represented by their first 50.000 words whenever possible.[2] Each text is accessible in three different formats: orthographic transcript (1),[3] morphologically tagged (2), and syntactically annotated (3).

(1)
```
<P_03>
<comment> [DEDICATÓRIA] </comment>
<heading> Senhor: </heading>
Ofereço a Vossa Majestade (...)
```

(2)
```
<P_03>
<comment> [DEDICATÓRIA] </comment>
<heading> Senhor/NPR :/. </heading>
Ofereço/VB-P a/P Vossa/PRO$-F Majestade/NPR (...)
```

(3)
```
(CODE (P_03))
(CODE (<comment>) ([DEDICATÓRIA]) (</comment>))

(IP (DP-VOC (NPR Senhor)) (. :)
    (DP-SBJ *pro*)
    (VB-P Ofereço)
    (PP       (P a)
              (DP (PRO$-F Vossa)
                  (NPR Majestade))) (...)
```

(in Matias Aires (1705-1763))[4]

In the present paper, linguistic and computational aspects involved in the construction of the Tycho Brahe Corpus will be focused. Particularly the linguistic guidelines behind the morphological and syntactic annotation systems and their computational adequacy will be discussed. Special attention will be drawn towards the computational implementation of these systems, especially towards the morphological system, which already presents the Tycho Brahe tagger, a rule-based part-of-speech tagger, as its most visible result.

2. The Tycho Brahe morphological annotation system

Strongly inspired by the Penn-Helsinki Parsed Corpus of Middle English (PPCME) methodology (Taylor / Kroch 1998),[5] the main goal of the Tycho Brahe Corpus is to make categorical and structural information about EP largely accessible. Based on this approach, the text tagging represents the first step of the annotation process. The tagging process treats lexical items as members of a certain part of speech class, selected under syntactic distributional criteria, and provides the necessary basis for the subsequent syntactic coding, in the sense that from a tagged

sequence .../P .../N,[6] a prepositional phrase (PP) is projected.
The following requirements have guided the Tycho Brahe tag set construction and application. First, the morphological tag set should be descriptively adequate, i.e. it should properly represent and discriminate the categories necessary to the description of the language in general, and the corpus in particular. This adequacy necessity has obliged us to modify the tag set proposed for Middle English, in order to accurately capture the morphological richness Portuguese exhibits, a feature shared by Romance languages in general. Second, the set must also be compatible with a computational treatment of the corpus files, since they are automatically tagged. To solve the tension between computational complexity and Portuguese morphological richness, a system of tags with internal structure has been proposed (Finger 1998). The tag always consists of a primary part indicating the part of speech class to which the lexical item belongs. This part can be accompanied or not by a secondary part, that specifies a clear-cut subgroup in a certain part-of-speech class, or several morphological inflectional [+marked] features carried by the lexical items (Finger 1998; Britto et al. 1999; Britto / Finger 1999). The diacritic "-" connects primary and secondary parts with each other, while "+" combines part of speech tags when more than one applies, as for contractions.

(4) Perto/ADV da/P+D-F Cidade/NPR principal/ADJ-G da/P+D-F Lusitânia/NPR está/ET-P uma/D-UM-F graciosa/ADJ-F Aldeia/NPR

Near to the main city of Lusitania is a graceful village[7]

(in Rodrigues Lobo (1579-1621))

These structured tags allow that distinct steps be applied on the Tycho Brahe tagger training, keeping the computational complexity generated by the increase of morphological tags under control (see section 3).
The tag set composition and application must guarantee complete and unambiguous searches. Motivated by syntactic distributional criteria, two or more different tags can be applied to one and the same word. In cases where the decision between the application of one or another tag is non-trivial, the system provides default tags to help keeping data homogeneously tagged, and therefore reliable from the linguistic point of view as well as from the computational one, since the Tycho Brahe tagger is trained with the support of manually checked tagged files. For example, according to our morphological annotation system, for cases involving a decision between a relative que, tagged /WPRO, and an explicative que, tagged /CONJ, the tag /WPRO should be applied, assuming the linguistic point of view that it can bring fruitful results to studies about relative construction strategies in the history of EP.[8]
Besides these constraints, our morphological annotation system also includes grammatical aspects that we shall briefly sketch out now.
Two main groups of verbs are considered: full verbs, i.e. verbs that assign theta-roles to their arguments, are tagged /VB; different tags are applied to ser/SR (be-individual level), estar/ET (be-stage level), ter/TR (have) and haver/HV (there to be), given that diachronically they seem to oscillate between a full verb and an auxiliary verb behavior.[9] Inflectional tags, indicating visible verbal morphology, are added to the primary coding in appropriate cases. Nouns and pronouns are tagged differently from each other, not only because they belong to different part of speech classes traditionally, but also because they display a peculiar diachronic behavior in EP.[10] Our system, as well as PPCME's, uses distinct tags for common /N and proper /NPR nouns, as well as for strong /PRO and possessive /PRO$ pronouns. Except for PRO, the main tags can be associated with inflectional tags for gender and number.
For clitics, two tags are proposed: /SE and /CL. The clitic se receives a special coding in view of its many syntactic functions (as reflexive, passive or indefinite particle) as well as of its idiosyncratic morpho-syntactic properties, as the one illustrated in (5).

(5) Ele mo deu. (mo = me+o) versus * Ele so deu. (so = se+o)
He gave me it He gave it to himself

With respect to determiners, not only the elements traditionally called definite articles but also inflected demonstrative pronouns are tagged /D. This is due to the fact that they display the same syntactic behavior during the whole history of the language (i.e they can be followed or not by a noun). For non-inflected demonstrative pronouns, however, a different tag, /DEM, is applied, since these elements function as neutral pronouns. Indefinite articles are also tagged in a different way: /D-UM, i.e. the primary tag /D always associated with a secondary tag -UM (plus other sub-tags for gender and number if necessary). Distinct tags for definite and indefinite determiners are justified by the fact that indefinites are grammatically interesting elements of their own. These determiners can be interpreted as a referential or as a quantificational element (definite ones are always [+referential]). Since this difference has syntactic repercussions, easily retrieving indefinites from the corpus facilitates research. That is what justifies a distinct tag for them.
As a last example, the lexical items that, from an interpretive point of view, quantify over entities or events receive tag Q, which can be associated with gender or number inflectional tags. In Modern Portuguese, the quantificational property applied to entities can be neutralized depending on the position of the item that expresses it in the nominal phrase. What follows is that items that are generally classified as quantifiers in pre-nominal position can be interpreted as adjectives when in post-nominal position. In the present annotation system, developed both for Classic and Modern Portuguese, this distinction is not considered. Thus, quantifying items are always tagged as Q. With respect to adjectives proper, besides gender and number inflectional tags which might apply to them, one can also identify the comparative and superlative adjective forms, using tags R and S,

respectively. Finally, as far as adverbs are concerned, the so-called intensity adverbs (muito, pouco, and others) have been classified as event quantifiers in the system proposed here, in parallel with the entity quantifiers. The tag ADV, therefore, has only been used to code time, place and manner adverbs as well as the forms mais and menos. The sub-tags R and S apply to adverbs in the same way as to adjectives.[11]

3. The Tycho Brahe tagger and refiner

Building large annotated corpora, such as the Tycho Brahe Corpus of Historical Portuguese, is only feasible if we use automatic methods for such tasks as part-of-speech tagging. As we wish to reach 2.000.000 words in the corpus, automated methods for morphological tagging and syntactic parsing had to be developed for Portuguese.
One of the most successful part of speech taggers according to the literature is that of Brill (1995), which is reported to correctly tag 97% of the words in English. (One should not be over-impressed with such a figure, for the probability of a 100-word text being tagged correctly at a rate of 97% per-word correctness is smaller than 4.75%). According to Brill's method, the tagger has first to be "trained" and then may be applied to texts. The training part takes much longer than the application part, so it deserves more attention.
Due to the morphological richness of Portuguese, the Tycho Brahe Corpus makes use of 154 different tags, as opposed to the 36 tags used by Brill. Finger (1998) estimated that – supposing we had Brill's algorithm being trained in Portuguese and English on the same computer with a corpus of the same size – the Portuguese program would run 319 times slower than the English one. If, as reported in the documentation of the Penn Treebank Corpus, the training of Brill's tagger on a manually tagged 500.000-word corpus took one day, the Portuguese tagger running on a similar machine would take 319 days, more than 10 months!
Since this is unacceptable, we devised an alternative approach, developing the Tycho Brahe tagger through the following steps.

3.1. Taming the complexity of tagging morphologically rich languages

The basic idea is to abandon the idea of tags as basic atomic entities and to start considering some 'internal structure' in them. What we do is to separate in a tag the basic component from its complementary part. Roughly speaking, the basic component can be thought as a category and the complement as a set of features.
With the division between basic component and suffix of tags, we can divide the training part of the tagger in two phases: the learning of the basic components and the refinement of the suffixes.
As we saw above, thematic verbs are tagged with basic tag VB, but four special verbs are tagged separately, namely ser, estar, haver and ter. This decision was taken due to the frequency of occurrence of these verbs and the prominent distinct grammatical roles that they can play (see section 2). This is a very expensive decision, since these four verbs and their inflected forms (without counting the addition of clitics) contribute with 50 tags to the tag set. However, if we consider only the basic tags, they amount only to five.
Applying this method to the complete set of tags, we end up with only 37 distinct basic components, almost the same number as all of Brill's English tags, which are 36. This enabled us to reduce the training of the corpus to a complexity comparable to that of English.

3.2. The two-step learning phase

The adoption of tags with internal structure allows us now to refine Brill's rule-based tagging method. The learning of tag transformation rules will be done in two steps, one that deals only with the basic component of tags, and another that deals with its complements.
The two phases of this method are the following:

(a) Use Brill's method to obtain a simplified tagger using the basic tags only, ignoring their internal structure.

As a result of that phase, the program will have learned transformation rules that deal only with the 37 basic tags. Thus, this step has the same complexity as the English tagger.

(b) Refine the tags obtained in the initial step, taking into consideration features such as gender and number agreement, tense inflection, etc.

Step (b) uses explicit linguistic (morphological) knowledge, as opposed to step (a), which is basically a generate-and-test search process. In this respect, step (b) itself can be divided into two sub-steps. The first one is a morphological inspection of the words, together with other agreement verification; no learning is involved in this step as it uses built-in linguistic knowledge. The second step uses contextual information to refine the basic tag.
Furthermore, we found out that such contextual information could be used not only to refine the output of the coarse tagger, but also to correct it. Indeed, two modules were added to the program,

one that does correction prior to refinement, and another that does it afterwards. The refinement is done by hand-made rules, and no learning techniques are used for this task.
At the present stage, we have obtained a precision of 95.45%. This has been achieved by training the corpus with 130.000 annotated words and testing them with an independent set of texts containing 45.000 annotated words.

4. The Tycho Brahe and the PPCME syntactic annotation systems

Following the methodological steps applied to the construction of the Tycho Brahe tagger, which generates our morphologically tagged files, some of our tagged files will be syntactically annotated manually, in order to construct and train the Tycho Brahe parser, a statistical parser in development. This initial stage is a very fruitful period for checking to what extent the annotation system proposed is powerful enough for an adequate treatment of the corpus data.
Strongly and largely inspired by the PPCME syntactic system, the annotation system proposed here as parsing scheme must conciliate computational and linguistic demands, since our main goal is to minimize computational complexities involved on the syntactic parsing and the parser training, providing the maximal accuracy in the treatment of linguistic issues.
This computational-linguistic articulation on the syntactic level can be exemplified as follows.
Given that the syntactic annotation is applied to morphologically tagged data, which drive the syntactic parsing, the derivation of syntactic projections is computationally costly, because these are not immediately equivalent to any lexical item previously tagged. In order to reduce computational operations, in the Tycho Brahe syntactic annotation system, as in the PPCME's, intermediate projections, in the sense of generative X' model, are not coded. Although this decision implies to generate structures not expected by current proposals in syntactic theory, for the purposes of a parsing scheme these are appropriate configurations. From the computational point of view, they are usually the less complex ones, since for each of the previously tagged elements at most one projection is derived. From the linguistic point of view, they properly preserve relevant information, as dominance, c-command, etc.
Purely linguistic motivation also justifies the absence of certain projections in the parsing scheme. According to Taylor / Kroch (1998: 99), "some phrases were omitted because their boundaries are too difficult to define. This is the case for VP, which, especially in Early Middle English where the order of the verb and its complements is still in flux (at least on the surface), cannot easily be included". In EP the same occurs. As indicated in the specialized literature, not only diachronically but also synchronically, the exact position of EP's internal arguments in visible syntax is a controversial matter. Thus, as already proposed by the PPCME's team, in the present system VP level is also not coded.
Finally, clitics always project an accusative, dative, oblique or genitive DP. For search purposes, this is the most convenient decision, since it allows an unified search of pronominal versus non-pronominal complements. Clitic climbing is also properly coded, since this is an important research field in Romance linguistics.
(6) illustrates the output of the parsing of a sentence from Matias Aires (1705-1763))

(6)
```
((IP-MAT (DP-SBJ (Vossa/PRO$-F)
      (Majestade/NPR))
  (PP (só/FP)
    (neste/P+D)
    (DP (livro/N)))
  (DP-10 (a/CL)
  (pode/VB-P)
  (IP-INF  (DP-ACC *ICH*-10)
       (VB (sentir/VB)
         (e/CONJ)
         (ver/VB)))
```

An interesting question is raised by determiners. In EP, these elements are followed not only by common nouns, but also by proper nouns (7), possessive phrases (8), adverbs (9) or infinitives even when the verbs are accompanied by their arguments (10):

(7) Mas/CONJ êste/D espectáculo/N (...) evitou-se/VB-D+SE com/P a/D-F gritaria/N que/WPRO fêz/VB-D o/D Tamagnini/NPR (...)
but this scandal was inhibited by the vociferation Tamagnini produced
(in Marquesa d'Alorna (1750-1839))

(8) (...) que/C a/D-F nossa/PRO$-F Côrte/NPR está/ET-P cheia/ADJ-F (...)
that (the) our Court is full
(idem ibidem)

(9) O/D mais/ADV-R (...) não/NEG passa/VB-P de/P frásis/N-P naturais/ADJ-G-P
the more is nothing more than natural sentences
(in Francisco Manuel Melo (1608-1666))

(10) E assim a êle se deve, depois de Deus, o/D conservar/VB as fazendas

and therefore to him is due, after God, the saving of the farms

(in António Vieira's (1608-1697) private correspondence)

Since the automatic parser uses tag information to automatically project syntactic phrases, the presence of D in all these cases largely motivates the decision of coding these sequences as DP, and not as NP as defended by Taylor / Kroch (1998) for English. From the corpus-linguistic point of view, this decision seems to be preferable, since this important information about EP syntax, i.e. the range of phrases D can take as its complement (Zamparelli 1996), can be directly retrieved from the corpus.[12]

5. Conclusions

A large part of the Tycho Brahe Corpus has already been morphologically tagged, and the results are very satisfactory. From the linguistic point of view, the morphological tag set proposed is strongly suitable for Portuguese prose from different periods. It has also proved efficient in allowing non-ambiguous searches. From the computational point of view, and with respect to the precision of the automated tagger, the good results obtained have helped us to have great part of the corpus already built. Concerning the syntactic parsing, the initial results also seem to indicate that the proposed annotation system is satisfactory. With regard to the automated parser, we hope to be able to obtain the first version of an efficient tool in a near future.

references

Brill, Eric 1995: Transformation-based error driven learning and natural language: a cases study in part of speech tagging. Computational Linguistics 21, 543-565.

Britto, Helena / Finger, Marcelo 1999: Constructing a parsed corpus of historical Portuguese; in: Proceedings of the ACH/ALLC International Humanities Computing Conference '99 (<http://www.iath.virgina.edu/ach-allc.99/proceedings/ britto.html>).

— et al. 1999: Morphological annotation system for automatic tagging of electronic textual corpora: from English to Romance languages; in: Centro de Lingüística Aplicada (ed) 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: P. Sojka, V. Matousek, K. Pala, and I. Kopecek (ed) Proceedings of the 1st Workshop on Text, Speech and Dialogue TDS'98. Brno: Masaryk University, 39-44.

Fodor, Janet / Sag, Ivan 1982: Referential and quantificational quantifiers. Linguistics and Philosophy 5, 355-398.

Galves, Charlotte / Britto, Helena 1999: A construção do Corpus Anotado do Português Histórico Tycho Brahe: o sistema de anotação morfológica; in: I. Rodrigues and P. Quaresma (ed). Proceedings of the IV PROPOR. Evora: University of Evora, 55-67.

— et al. 1998: Rhythmic Patters, Parametric Settings and Linguistic Change (unpublished manuscript; <http://www. ime.usp.br/~ tycho/presentation>).

Taylor, Ann / Kroch, Anthony 1998: The Penn-Helsinki Parsed Corpus of Middle English II. Pittsburg, University of Pennsylvania (unpublished manuscript).

Torres-Morais, M. Aparecida 1995: Do Português Clássico ao Português Europeu Moderno: Um Estudo da Cliticização e do Movimento do Verbo. Campinas: State University of Campinas (Unicamp) (unpublished PhD thesis).

Zamparelli, Robeto 1996: Layers in DP: the basic idea (PhD thesis available at <http://ww.cogsci.ed.ac.uk/~roberto/ layers/basic.html>).

---

[1] The Tycho Brahe Corpus is being developed within the scope of the research project 'Rhythmic Pattern, Parameters Settings and Linguistic Change', coordinated by Charlotte Galves and supported by FAPESP (Grant # 98/03382-0) (<http://www.ime.usp.br/~tycho>; <http://www.fapesp.br>).

[2] Maria do Céu's (1658-1753) text is an example of the fact that, according to their importance for the historical study of the language, even texts under 50.000 words can be included in the Tycho Brahe Corpus.

[3] Orthographic transcript files will not be subject to detailed presentation or discussion in this paper. It is important to notice, however, that they are systematically submitted to an annotated system for extra-linguistic material codification that encapsulates information such as text edition, editor's comment in the interior of the text, researcher's comment, headings, page number, and the end of (complex) sentences when not indicated by the editor (Britto / Finger 1999). Notice also that editor's footnotes, prefaces, introduction, appendix or annexes are never included in the Tycho Brahe corpus files.

[4] For more information about the authors and texts from where our examples were extracted, see <http://www.ime.usp.br/~tycho/corpus/show-texts>.

[5] For access to PPCME, see <http://www.ling.upenn.edu/mideng>.

[6] P = preposition; N = common noun.

[7] ADV = adverb; P+D-F = contraction between the preposition de and the feminine singular definite determiner a; NPR =

proper noun; ADJ-G = neutral singular adjective; ET-P = Simple Present tense of to be in its stage level form; D-UM-F = feminine singular indefinite determiner; ADJ-F = feminine singular adjective.

[8] A more detailed discussion of the requirements and criteria exposed above can be found in Galves / Britto (1999).

[9] Last but not least: differently from what has been proposed in the system for Middle English, modal verbs don't receive any specific primary tag in our codification system. As Ruwet (1968) points out, no verbs in French exhibit any of the properties that English modals do (specific forms for the past or the impossibility of double negation). The same seems to apply for Romance languages in general and Portuguese in particular.

[10] In the 19th century, period in which the linguistic change that rules out the order 'X cl V', X [+referential] (in the sense of Fodor and Sag 1982) had already taken placed, examples as 'Pronoun cl verb' are still observed, while 'Noun cl verb' disappear (cf. Torres-Morais 1995).

[11] For detailed information on our morphological annotation system, see <http://www.ime.usp.br/~tycho/corpus/manual/tags.html>.

[12] For more information on our syntactic annotation system, see <http://www.ime.usp.br/~tycho/corpus/manual/tags.html>.