MORPHOLOGICAL ANNOTATION SYSTEM FOR AUTOMATED TAGGING OF ELECTRONIC TEXTUAL CORPORA: FROM ENGLISH TO ROMANCE LANGUAGES

HELENA BRITTO
State University of Campinas (UNICAMP)– Campinas – SP – Brazil
CHARLOTTE GALVES
State University of Campinas (UNICAMP)– Campinas – SP – Brazil
ILZA RIBEIRO
UNIFACS/State University of Feira de Santana – Salvador/F. de Santana – BA – Brazil
MARINA AUGUSTO
State University of Campinas (UNICAMP)* – Campinas – SP – Brazil
ANA PAULA SCHER
University of São Paulo (USP)/ State University of Campinas (UNICAMP)* – SP – Brazil

I Introduction

Based on the Penn-Helsinki Parsed Corpus of Middle English[1], the Tycho Brahe Parsed Corpus of Historical Portuguese[2] consists of an electronic annotated corpus composed of prose, originally written in Portuguese by native speakers of European Portuguese (henceforth EP) born between the 16th and 19th centuries. The present annotation system to be applied to Portuguese has been developed in the lines of the system designed for English by Taylor & Kroch (1994) and aims at codifying this corpus morphologically.

For the elaboration of this system, the following goals were pursued. Firstly, to facilitate automatic searches of morphologically classified lexical items, we envisaged a system which would be broad enough so as to be applied to Portuguese texts from different historical periods as well as strict enough so as to allow unambiguous searches. On the other hand, we tried to accommodate such linguistic needs to computational demands derived from the implementation of the automated morphological system, which imposes that a minimum possible number of tags and sub-tags be assumed. Secondly, our ultimate aim is to provide an adequate basis for the automated syntactic parser to be developed. Thus, the selection of morphological tags (specifically the ones related to parts of speech) was based on syntactic distributional criteria.

2 Overall Design

This morphological annotation system comprises the following basic groups of tags: **Parts of Speech tags**; **inflectional tags**; **diacritics** and **punctuation tags.**

As has been proposed for English, **Parts of Speech tags,** as the name implies**,** are used for the classification of the word according to the part of speech it belongs to. Nevertheless, differently from the system designed for English, the one developed for Portuguese assumes a group of **inflectional tags** to be associated to the **Parts of Speech tags** in order to capture the morphological richness Portuguese exhibits, a feature shared by romance languages in general. Not only are such tags used to indicate which lexical items compose a clear-cut subgroup in a certain part-of-speech class[3], but also to indicate several morphological visible inflectional [+ marked] features the lexical items carry. Considering the diacritics, Taylor & Kroch (op.cit.) proposed the following signals: "+", to combine tags when more than one applies as for contractions or even written junctures used for distinct lexical items; "/", to separate the lexical item from its tag(s); and numbers, to refer to separate parts of words, where the first number indicates the number of parts and the second number, which part the part is[4]. In addition to these signals, we assume "!", to refer to the medial position of clitics (see Section 3.2) and "-", to combine **Parts of Speech tags** to **inflectional tags**. Thus, (1a) is an example of codification our system assumes in which N is the main tag used for nouns, indicating the part of speech the word belongs to, and P is the secondary inflectional tag, standing for plural. (1b), on the other hand, shows the codification obtained through the use of the tagging system proposed for English:

(1) a. **casa/N**; **casas/N-P**
b. **house/N**; **houses/NP**

This kind of tagging allows the gradual introduction of tags, from the **Parts of Speech** ones to the more inclusive **inflectional** ones, in the computational training of the automated morphological parser. Thus, in cases such as (2), the automated training starts from the **Parts of Speech tag** ADJ, taken as the primary one, to be followed by the introduction of F, the **inflectional tag** for feminine, taken as the secondary one, and, subsequently, of P, the **inflectional tag** for plural, taken as the tertiary one.

(2) **bonitas/ADJ-F-P** 'beautiful '

As for **punctuation tags**, which refer to punctuation signals, the system proposed herein basically assumes the same pattern adopted for English.

3 System Outline

3.1 Nouns

Common and proper nouns have been attributed the **Parts of Speech tags** N and NPR respectively, since they behave differently from a syntactic point of view (cf. Chomsky 1981, among others). As for the inflectional features related to nouns, Portuguese exhibits visible morphology

concerning gender and number. Nevertheless, common and proper nouns have only been attributed number **inflectional tags**: singular is identified by the absence of a particular tag whereas plural receives tag P.

(3) **João/NPR**, sua **casa/N** e seus **bichos/N-P** 'João, his house and his pets'

Since the syntactic context is irrelevant for gender attribution to nouns, we decided not to contemplate this part of speech class with a gender **inflectional tag**, differently from what happens to determinants and adjectives. This idiosyncratic noun property is clearly shown in a list to be established of all the nouns contained in the corpus. This list will compose a dictionary and serve as reference for the automated morphological parser.

3.2. Pronouns

Taking into consideration syntactic differences among (i) strong pronouns in subject or object positions; (ii) null pronouns and (iii) weak complement-pronouns (or clitics), commonly found in null subject languages, like European Portuguese and some other Romance languages (see Cardinaletti & Starke 1994, among others), different primary tags were attributed to each of the lexical items of the tripartite division outlined above[5]. Tag PRO was attributed to strong pronouns in subject or object positions:

(4) a. **Ela/PRO** encontrou-lhe 'She met-him '
b. Maria comprou o carro para **mim/PRO** 'Maria bought the car to me'

Regarding the clitics, two tags were proposed: SE for clitic **se** in any of its use; and for all of the remaining clitics, the tag CL which may yet be associated to gender and number **inflectional tags**. Clitic **se** receives a particular tag on the basis of its peculiar behavior concerning not only the fact of performing various syntactic functions (reflexive, passivizing and indetermining particles), and showing idiosyncratic morphological behavior (it never allows contraction[6]), but above all because it proves to be, as has been largely noted in the literature, a topic on its own of great synchronic and diachronic interest (cf. Nunes, 1990, 1995; Cyrino, 1993). Differently from the system designed for English, it is also worth pointing out that in what concerns the clitics, mesoclisis[7], which is largely productive in the history of Portuguese, was attributed the diacritic signal "!" as previously mentioned:

(5) a. João **se/SE** machucou 'João himself hurt '
b. Maria não **lhas/CL+CL-F-P** entregou 'Maria did not her-them give '
c. **dar-te-ei/VB-R!CL** 'give-you-future/1ps '

This **Parts of Speech** class deserved a third and final tag, PRO$, which was attributed to possessive pronouns:

(6) a. **Meu/PRO$** irmão morreu 'My brother died '
b. **Vossa/PRO$-F** Majestade chegou 'Your Highness arrived '

3.3 Determiners

Tag D was applied not only to the elements traditionally classified as definite articles but also to the inflected demonstratives, all of which also receive tags for number and gender (see (7)). This is due to the fact that, as can be seen in the history of Portuguese, articles and inflected demonstratives share the same syntactic distribution[8](see (8)). Additionally (9) shows elements receiving tag D, followed by the **inflectional tag** G, which stands for neutral gender.

(7) Eu encontrei **o/D** rapaz e **esta/D-F** moça 'I met the boy and this girl '
(8) **O** rapaz; **este** rapaz; ***o este** rapaz; ***este o** rapaz 'The boy; this boy... '
(9) **Tal/D-G** problema; **tais/D-G-P** senhoras 'Such problem; such women '

Concerning the case of 'um' ('a'), we opted to attribute it tag D, associated to the **inflectional tag** -UM, so that its automatic identification with the other members of the determiners class would be facilitated[9]. The associated inflectional tag –UM has the advantage of differentiating this element from the other articles, which is desirable due to its property of being either [+referencial] or [+quantifiedl] (cf. Fodor & Sag 1982; Diesing 1992):

(10) Eu encontrei **um/D-UM** rapaz 'I met a boy '

3.4. Verbs

Two different groups of verbs were contemplated: on one hand, under the tag VB, are all the verbs that, besides presenting verbal agreement and finite features, assign thematic roles to their arguments, being considered full verbs in the sense of Pollock 1989, Chomsky & Lasnik 1993; Chomsky 1993. On the other hand, separate tags were attributed to ser ('be-individual level') (SR), estar ('be-stage level'), (ET), ter ('have') (TR) and haver ('there to be') (HV), of which features seem to oscillate diachronically so that sometimes they may be employed as full verbs, and sometimes function as mere auxiliary verbs which are restricted to carrying inflectional information. Associated to these main **Parts of Speech tags**, **inflectional tags**, indicating visible verbal morphology, may be added to the primary codification: F, for inflected infinitive; I, for imperative; P, for present; SP, for present subjunctive; D, for (perfect and imperfect) past; RA, past perfect; SD,

for past subjunctive; R, for future, SR, for future subjunctive; and, finally, G, for gerund; PP, for perfect participle[10] e AN, for passive participle:

(11) Ele **tinha/TR-D saído/VB-PP** 'he had left'

Last but not least, differently from what has been proposed in the system for English, modal verbs were not attributed specific tags under our codification system. As Ruwet (1968) points out, no verbs in French exhibit properties modals do for English (specific forms for past or the impossibility of double negation). The same seems to apply to Romance languages in general, specifically Portuguese:

(12) a. receive/received; can/*canned; can/could
b. receber/recebia; poder/podia
(13) a. I cannot *don't/*not/*no make it.
b. Eu não posso não fazer isto.

3.5 Quantifiers, adjectives and adverbs

The lexical items that, from an interpretive point of view, quantify over entities or events receive tag Q, which can be associated to gender or number **inflectional tags**.

(14) **Alguém/Q-G**; **toda/Q-F** criança 'Somebody;every child'

In 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 tagged as quantifiers in pre-nominal position, might be tagged as adjectives when in post-nominal position:

(15) a. **Qualquer/Q-G** homem **bonito/ADJ** 'Any man handsome'
b. Um homem **qualquer/ADJ-G bonito/ADJ** 'A man any handsome'

Besides gender and number **inflectional tags** that might apply to adjectives, one can also identify the comparative and superlative adjective forms, using tags R and S, respectively:

(16) moça **lindíssima/ADJ-S-F** 'girl beautiful-sup'

Finally, as far as adverbs are concerned, the so-called intensity adverbs have been classified as event quantifiers in the presently proposed system, in parallel with the entity quantifiers presented above (cf. Peres 1991; Mattos e Silva 1993). The tag ADV, therefore, has only been used to codify time, place and manner adverbs[11].

(17) a. Event quantifiers:
Ele trabalha **muito/Q** 'He works a lot'
b. Time, place and manner adverbs:
Amanhã/**ADV** volto aqui/**ADV** rapidamente/**ADV** 'Tomorrow return-1ps here quickly' **[12]**

The variation tags R and S apply to adverbs in the same way as to adjectives:

(18) Saiu lentissimamente/ADV-S 'Left-3ps slowly-sup'

3.6 Prepositions and prepositional phrases

Tag P is used to identify prepositions. In cases of lexical contraction, tags D or DEM can be associated to it.

(19) **do/P+D** texto 'of+the-masc text'

As for the so-called prepositional phrases, the history of Portuguese reveals that these groups of words result, in general, of various grammaticalization processes. From the diachronic point of view, this means that their components migrate systematically from a class of words to another. A possible consequence, then, of categorial tagging of each of the items that form these groups of words separately might be the impossibility of automatic selection of these data from the corpus. Thus, aiming at avoiding such problem, these groups of words are tagged P with the association of number tags, the first one indicating the number of elements that form the group and the others, the position in which they occur in it:

(20) **apesar/P-21 de/P-22** você 'in spite of you'

3.7 Conjunctions and derived phrases

As well as suggesting tags for the two main conjunction groups - coordinating (CONJ) and subordinating (complementizers (C) and other subordinating (CONJS)) (see(21a))- the system suggests tagging the so-called conjunctional phrases (21b) in a way that is similar to the one proposed for the prepositional phrases:

(21) a. Ele saiu **e/CONJ** disse **que/C** virá **quando/CONJS** puder. 'He left and said that come-fut when can-fut/subj'
b. **no/CONJ-21 entanto/CONJ-22** ' however'

Some observations should be made for tags CONJ and CONJS. Firstly, differently from what was proposed for English, our system tags not only e (and) and mas (but), but also porém, entretanto, todavia (however, nevertheless, etc) and others as coordinating conjunctions (CONJ), and not as adverbs or prepositions. This is proposed in spite of the different diachronic characteristics of the latter as opposed to the former ones. For example, as was the case for French (cf. Adams 1987; Dufresne 1993), only e and mas might not count as occupying the first syntactic position in the historical periods of Portuguese in which the V2 effect was evident (cf. Ribeiro 1995; Torres Morais 1995). However, due to the lack of systematic researches about the connectives porém, entretanto, todavia or others in Portuguese, we initially went for the traditional classification, according to which all the above mentioned connectives are equally called coordinating conjunctions. Secondly, tag CONJS is another innovation of our system, if compared to the one proposed for English, since, in Portuguese, these elements do not obviously behave as prepositions or adverbs[13].

3.8 Relative elements

To relative elements, we apply tags WPRO or WPRO$

(22) a. A moça **que/WPRO** está aqui (...) 'The girl that is here (...)'
b. A moça **cuja/WPRO$-F** filha está aqui (...) 'The lady whose daughter is here (...)'

3.9 Interrogative elements

To interrogative elements, the following tags apply: WPRO, WADV and WQ for direct and indirect interrogatives, and WD for interrogative determiners.

(23) a.**Quando/WADV** você comprou **o que/WPRO**? 'When you bought what'
b. Ele perguntou **se/WQ** você vem. 'He asked if you come'
c. **Que/WD livro/N** foi comprado ? 'Which book was bought'

3.10 Negative Particles

Negative particles are tagged NEG:

(24) Ele **não/NEG** me encontrou. 'He not me find-past'

3.11 Numerals

Since ordinal numbers behave as adjectives and, therefore, are tagged ADJ, tag NUM applies to cardinal numbers only.

(25) A **terceira/ADJ-F** nau chegou em **1513/NUM** 'The third ship arrived in 1513'

3.12 Interjections

Tag INTJ applies to items that belong to the interjections class.

3.13 Foreign and unknown words and punctuation signals

Since foreign and unknown words show no specific morphological features associated to particular languages, the same **Parts of Speech tags** proposed in the system designed for English were assumed for Portuguese: FW and XX, respectively. The same is valid for punctuation signals since written western tradition shares a uniform punctuation system.

4. Conclusion

The preliminary results of the application of this annotated system suggest that it is strongly suitable to Portuguese prose from different periods (cf. Augusto, Britto & Scher 1998). It has also proved efficient in blocking ambiguous searches of various classes of words. Furthermore, it seems to be highly applicable for other Romance languages too, since they all share most of the morphological properties predicted in the system proposed. Finally, as Finger (1998) points out, the present system structuring of tags in levels may guarantee satisfactory results with respect to its implementation for automatic morphological tagging.

References

ADAMS, M. (1987) **Old French, Null Subjects and Verb Second Phenomena**. Ph.D Thesis, Los Angeles:UCLA.

AUGUSTO, M., H. BRITTO & A.P. SCHER (1998) "Morphological tagging for different periods of Portuguese prose". Ms. Campinas:Unicamp.

CARDINALETTI, A. & M. STARKE (1994) "The typology of structural deficiency: on the three grammatical classes". Working Papers in Linguistics **4**(2): 41-109. University of Venice.

CHOMSKY, N. (1981) **Lectures on Government and Binding**. Dordrecht: Foris.

_____ (1993) "A minimalist program for linguistic theory" In: K. Hale & S.J. Keyser (eds) **The View from Building 20:** Cambridge: MIT Press.

CHOMSKY, N. & H. LASNIK (1993) "The theory of Principles and Parameters". In: J. Jakobs, A. Stechow, W. Sternefeld, and T. Vennemann, eds. **Syntax: an international handbook of contemporary research**. Berlin: the Gruyter.

CINQUE, G. (1997) **Adverbs and Functional Heads: a cross-linguistic perspective**. To be published by Oxford University Press.

CYRINO, S. (1993) "Observações sobre a mudança diacrônica no português do Brasil: objuto nulo e clíticos" In: I. Roberts & M. Kato (orgs) **Português Brasileiro: uma viagem dicrônica**. Campinas: Ed. da Unicamp.

DIESING, (1992) **Indefinites**. Cambridge: MIT Press.
DUFRESNE, (1993) **L' Articulation Syntatique et Phonologique de la Cliticisation: le cas des pronoms sujets en moyen français**. Ph.D Thesis. Montreal: University of Quebec
FINGER, M. (1998) "Tagging a morphologically rich language: the construction of the Tycho Brahe Parsed Corpus of Historical Portuguese". Ms. São Paulo: IME-USP.
FODOR, J. & I. SAG (1982) "Referential and quantificational indefinites". Linguistics and Philosophy **5**: 355-398.
MATTOS e SILVA, R.V. (1993) **O Português Arcaico**. São Paulo:Contexto.
NUNES, J. (1990) **O Famigerado SE: uma análise sincrônica e diacrônica das construções com SE apassivador e indeterminador.** MA Dissertation. Campinas: Unicamp.
_____ (1995) "Ainda o famigerado SE". DELTA **11**(2): 201-240.
PERES. J. (1991) "Basic Aspects of the Theory of Generalized Quantifiers" In: F ilgueiras, M. et alli (orgs.) **Natural Language Processing: Proceedings of 2nd Advanced School in Artificial Intelligence,** Berlin: Springer-Verlag, pp. 141-157
POLLOCK, (1989) ' Verb movement, UG and the structure of IP" , L.I. **20**: 365-424.
RIBEIRO, I. (1995) **A Sintaxe da Ordem no Português Arcaico: o efeito V2.** Ph.D Thesis. Campinas: Unicamp.
RUWET, N. (1968) **Introduction à la Grammaire Générative**. Paris: Plon.
TAYLOR, A. & A. KROCH (1994)
TORRES MORAIS, M.A. (1995) **Do Português Clássico ao Português Moderno: um estudo da cliticização e do movimento do verbo.** Ph.D Thesis. Campinas: Unicamp.

---

* Doctorate Student

[1] The *Penn-Helsinki Parsed Corpus of Middle English* is available at: http://www.ling.upenn.edu/mideng

[2] *Tycho Brahe Parsed Corpus of Historical Portuguese* is available at: http://www.ime.usp.br/~tycho

[3] For more details see **inflectional tags** in Section 3.3 – Determiners.

[4] Used for the codification of prepositional phrases and some conjunctions (see Sections 3.6 and 3.7).

[5] Regarding null pronouns, there is no explicit codification associated to them in terms of morphological tagging, although we preview its syntactic codification, since they are extremely important as far as Romance languages are concerned.

[6] Such morphological behavior is expressed by the asymmetry below:
(i) ele mo deu. '*he me-it gave' vs.* ele *so deu 'he **himself-it gave'*

[7] The clitic splits the verb into its base form and the inflectional parts.

[8] Since uninflected demonstratives show actual pronominal behavior (not a determiner one), the tag DEM was reserved to them:
(i) **Isto/DEM** acontece '*This happens'.*

[9] Data of the kind presented in (i) below is also attributed the tag D-UM:
(i) Um/D-UM e outro.... '*one and another*'
(ii) um/D-UM lápis e dois livros '*one pencil and two books*'

[10] Although past participle receives a specific tag under the present annotation system, the original Latin forms for present participle and future participle did not survive in Portuguese, not even in its archaic period. Therefore the set of tags proposed don't contemplate them.

[11] The literature on the topic presents no consensus for the classification of *time*, *place* and *manner* adverbs as real adverbs or as event quantifiers (cf. Cinque 1997 vs. Loebner 1987 *apud* Peres 1991). Thus, the tags being proposed here may be modified in future versions of the present annotation system, depending on the degree of suitability of one theoretical position or another for the diachronic data of Portuguese.

[12] In spite of its morpho-sybtactic properties, the so-called *adverbial phrases* are classified as separate words: no adverb-related classifying tags have been proposed for such phrases.
(i) com/P certeza/N '*for sure'*

[13] Depending on the diachronic analysis of the syntactic behavior of the Portuguese so-called *subordinating conjunctions* as adverbs or prepositions, the application of tag CONJS in such cases may be reconsidered.