Reusing Available Resources for Tagging a Spoken Portuguese Corpus

Amália Mendes, Raquel Amaro, M. Fernanda Bacelar do Nascimento

Centro de Linguística da Universidade de Lisboa
Complexo Interdisciplinar, Av. Prof.Gama Pinto, nº 2, 1649-003 Lisboa
amalia.mendes@clul.ul.pt; ramaro@clul.ul.pt; fbacelar.nascimento@clul.ul.pt

Abstract

This paper discusses the experience of reusing annotation tools developed for written corpora to tag a spoken corpus with POS information. Eric Brill's tagger, initially trained over a written and tagged corpus of 250.000 words, is being used to tag the C-ORAL-ROM spoken corpus, of 300.000 words. First, we address issues related with the tagset definition as well as the tagger performance over the written corpus. We discuss important options concerning the spoken corpus transcription, with direct impact on the tagging task, as well as the additional tags required. Transcription options allow in some cases for automatic tag identification and replacement, through a post-tagger process. Other cases, like the annotation of discourse markers, are more complex and require manual revision (and eventual listening). Since the final annotation will not only include the POS tag but also the wordform lemma, the paper also addresses issues related to the lemmatisation task. The positive results obtained show that the process of tagging and lemmatising a spoken Portuguese corpus through the reuse of already available resources may constitute an example of how to minimize the costs of such a task, without compromising the results. Finally, we discuss some possible developments to improve the tagger's performance.

1. Introduction

Tagging a spoken corpus with part-of-speech (POS) information presents certain specificities not found in the annotation of written corpora. However, our experience shows that it is possible to attain satisfactory results in spoken texts POS tagging by reusing and adapting resources developed for a written corpus.

The spoken corpus that is actually being tagged has been developed under the project *C-ORAL-ROM: Integrated Reference Corpus for Spoken Romance Languages*[1] – a project of the European Commission addressing spoken speech. This corpus is about 300.000 words and covers several registers: informal, formal, media and phone conversations. Our objective is not only to tag the corpus with POS information, but also to lemmatise the data – increasing the complexity of our task – reusing, whenever possible, already available resources.

We proceeded first by considering the already developed tagset and the training of Eric Brill's tagger over a written and tagged corpus of 250.000 words for the project *Recursos Linguísticos para o Português: um corpus e instrumentospara a sua consulta e análise*[2]. The use of a previously developed resource *Léxico Multifuncional Computorizado do Português Contemporâneo*[3] – LMCPC, a frequency lexicon based on a written 16M words corpus, proved helpful for the lemmatising task, and, hopefully, will also prove to be valuable for the improvement of the tagging task.

2. Tagging a written corpus

We used Eric Brill's tagger (Brill 1993) trained over a written Portuguese corpus of 250.000 words, morphosyntactically annotated and manually revised. Several genres compose this corpus: newspaper (65%), books (20%), magazines (5%) and varia (10%).

The morphosyntactic annotation covered the main POS categories (Noun, Verb, Adjective, etc.) and secondary ones (tense, conjunction type, proper noun and common noun, variable *vs.* invariable pronouns, etc.), but person, gender and number categories were not included, due to limits in time and human resources.

2.1. Some aspects of the tagset definition

The difficult and time-consuming task of deciding between ambiguous categories was avoided by the use of portmanteau tags. Therefore, distinctions such as the one between the indefinite article and numeral for the annotation of the form *um, uma*, the one between inflected or non-inflected infinitive verb forms, and the one between some common or proper nouns, were solved by the portmanteau tags /ARTi:NUMc, for the first case, /VB:VBf, for the second, and /Np:Nc for the last.

Some functional distinctions between categories were added when it seemed important for future research. It is the case, for instance, of the distinction between the past participle in compound tenses (/VPP) and the past participle in other contexts (/PPA):

[1] *C-ORAL-ROM: Integrated Reference Corpus for Spoken Romance Languages* is being developed by CLUL, under M. Fernanda Bacelar do Nascimento supervising. National C-ORAL-ROM corpora will be distributes by ELDA.

[2] *Recursos Linguísticos para o Português: um corpus e instrumentos para a sua consulta e análise* was developed by CLUL, 2001-2003, under M. Fernanda Bacelar do Nascimento supervising. Corpus available for on-line queries at http://www.clul.ul.pt/sectores/projecto_rld1.html.

[3] The *Léxico Multifuncional Computorizado do Português Contemporâneo* was developed by CLUL, 1997-2000, under M. Fernanda Bacelar do Nascimento supervising. Lexicon available for download at http://www.clul.ul.pt/sectores/projecto_lmcpc.html.

*ele/PES tinha/VAii comprado/**VPP** um/ARTi:NUMc livro/Nc*
*olhos/Nc fechados/**PPA***

The prepositional, conjunctional, pronominal and adverbial locutions were also tagged, resulting in the following information for each tagged element of the locution: category, element position number and identification number (for cross-reference in an appended list of locutions). The locution identification number is inserted after the tagging process to avoid multiplying the tagset length.

num/LADV1_ 117 instante/LADV1_117
logo/LCONJ1_47 que/LCONJ2_47
à/LPPREP1_003 beira/LPREP2_003 de/LPREP3_003
o/LPRON1_07 qual/LPRON2_07

Since some words sequence constituting a locution may also occur freely, a manual revision was considered necessary for obtaining a maximum success annotation.

The contractions of two lemma were annotated by joining two tags through the sign '+' (dos/**PREP+ARTd**), and the wordforms connected by hyphen received two tags also connected by hyphen (disse-me/**Vppi-CL**). These two tagging options have the effect of expanding the total tagset into an indefinite number (from a minimum of 54 tags, to a maximum of more than 204), by combining several tags that are recognised by the tagger as a new single one.

2.2. Some comments on the results

After the tagger training and after the automatic tagging of a written corpus, the results show two aspects that will have to be considered in the future: first, some difficulties in the automatic tagging of the locutions and, second, the lack of identification of certain words.

In order to respond to the locution tagging problems, two solutions are being studied: on one hand, the inclusion of the locution identification number in the tagset, bearing in mind all the subsequent problems derived from the hudge tagset length increase (note that the current prepositional locutions list alone exceeds 484 locutions); on the other hand, the conception of a post-tagging tool for the locutions annotation is being considered.

In order to respond to the second case, the future development will be to insert in the annotation process the LMCPC, extracted from a 16 million words corpus (considerably larger than the used training corpus – 250.000 words), in which a large set of wordforms occurs (around 140.000).

3. Tagging and lemmatising a spoken corpus

The spoken corpus was tagged with the tool described in the previous section. In spite of having been trained over a written corpus, and surprisingly against our expectations, the results achieved were very satisfactory, with a success rate of 91,5%.

Nevertheless, some post-tagging adaptations had to be made in order to achieve the established spoken corpus annotation.

3.1. Specific spoken language phenomena

Some characteristic spoken language phenomena, such as word repetition and truncated words don't seem to affect the tagger performance, either statistically, either in terms of contextual rules.

However, the tagger identifies and tags the prosodic marks (question marks, slashes, and so on) as punctuation, making it necessary to automatically remove these tags in a following stage.

Besides the previous phenomena mentioned, and due to the specific transcription guidelines used in the C-ORAL-ROM project, there are several other phenomena that required tagset adaptations:

a) extra-linguistic elements;
transcription: hhh; Tag: **EL**
b) fragmented words;
transcription: beginning with &; Tag: **FRAG**
c) words impossible to transcribe (impossible to hear, for example);
transcription: xxx; Tag: **Pimp**
d) paralinguistic elements, such as *hum*, *hã* and onomatopoeias.
Tag: **PL**
e) discourse markers, such as *pá*, *portanto*, *pronto*;
Tag: **MD**
f) discoursive locutions, such as *sei lá*, *estás a ver*, *quer dizer*, *quer-se dizer*.
Tag: **LD**

In the cases described in (a), (b) and (c), the adopted specific transcription allows for automatic tag identification and replacement, through a post-tagger process. The same process is applied in the cases described in (d), since there is a finite list of symbols representing paralinguistic elements.

The discourse markers (cf. (e) an (f)) present a more difficult case, since they correspond to forms that also belong to other word categories: for instance, *não sei* is automatically tagged as *não/ADV sei/Vpi*, making a manual revision (and eventual listening) necessary in order to decide whether the form is a discourse marker or not.

The tagging of proper nouns is, on the contrary, simplified in the spoken corpus tagging process, since proper nouns are the only forms transcribed with initial capital letter.

The development stage that follows consists in the Eric Brill tagger's training over a manually revised spoken corpus, as well as in the exploitation of the tagger contextual rules in order to optimize its performance. Amongst other things, we aim at improving the locution tagging process since locutions account for an increase of around 2% of the error rate.

3.2. Lemmatisation

The final format of the spoken corpus annotation includes, for each form, not only the POS tag, but also the correspondent lemma:

word\LEMMA\tag.

The lemma is extracted automatically from the LMCPC: the form is searched for in the lexicon, the correspondent(s) lemma(s) is(are) found and placed near the form, in a process parallel to the automatic tagging. So, it is possible for a wordform to be attributed several lemma, requiring thus manual lemma selection.

In the future, with the foreseen improvement of the success rate, we expect to be able to trust the automatic POS tagging in order to cross information with the lexicon POS data and select the proper lemma for each wordform.

In the cases from (a) to (d) described above, and in the proper nouns case, the lemma is considered empty, since it is clear that there is no lemma for that expressions.

In the case of locutions, since the lemma is the locution set, there is no need for the locution identification number. Locution lemmatisation made it necessary to develop a tool to automatically compose the desired lemma format:

o\O_QUAL\LPRON qual\O_QUAL\LPRON

We present next a tagged and lemmatised extract from one of the conversations of the corpus:

```
*FER: e\E\CONJc ela\ELA\PES / <
como\COMO\ADV reagiu\REAGIR\Vppi > ?$
*BEN: [<] < &eh\-\FRAG / &hum\-\FRAG > /
reagiu\REAGIR\Vppi muito\MUITO\ADV
bem\BEM\ADV //$ < hhh\-\EL >$
*FER: [<] < hhh\-\EL >$
*BEN: / reagiu\REAGIR\Vppi muito\MUITO\ADV
bem\BEM\ADV //$ começa\COMEÇAR\Vpi
na\EM+A\PREP+ARTd segunda-feira\SEGUNDA-
FEIRA\Nc //$ entretanto\ENTRETANTO\ADV /
eu\EU\PES disse-lhe\DIZER-LHE\Vppi-CL
que\QUE\CONJs /$
*AUG: < então\ENTÃO\ADV e\E\CONJc as\A\ARTd
férias\FÉRIA\Nc > ?$
*BEN: / [<] < que\QUE\CONJs não\NÃO\ADV
sabia\SABER\Vii > se\SE\CONJs / eu\EU\PES
podia\PODER\Vii na\EM+A\PREP+ARTd segunda-
feira\SEGUNDA-FEIRA\Nc / por\POR\PREP
ter\TER\VB os\O\ARTd conselhos\CONSELHO\Nc
//$ de\DE_MANEIRA_QUE\LCONJ
maneira\DE_MANEIRA_QUE\LCONJ
que\DE_MANEIRA_QUE\LCONJ / a\A\ARTd
&senho\-\FRAG / ficou\FICAR\Vppi
então\ENTÃO\ADV combinado\COMBINAR\PPA /
eu\EU\PES telefonar\TELEFONAR\VB:VBf
para\PARA\PREP lá\LÁ\ADV /$
```

4. Final comments

The development of tagged corpora is, definitely, a human resources and time-consuming task.

The process of tagging and lemmatising a spoken Portuguese corpus through the reuse of already available resources here presented may constitute an example of how to minimize the costs of such a task, without compromising the results.

Summing up, this complex process, besides the spoken corpus constitution and transcription, has consisted in:

i) the definition of a suitable tagset and tagging options;

ii) the adaptation of a written tagged corpus to the desired tagset;

iii) the training of Eric Brill's tagger;

iv) the automatic replacement and/or withdraw of the tags, according to the specific spoken language phenomena transcription;

v) the creation of a tool for the automatic lemmatisation of the corpus, using an already existent lexicon;

vi) the creation of a tool for the automatic lemmatisation of the locutions elements;

vii) and, at last, the manual revision of the final result.

The following stage will consist in the Eric Brill's tagger training over the resulting spoken corpus, manually revised.

We hope to achieve significant improvements regarding the performance of spoken corpus automatic tagging.

Acknowledgements

We want to thank several colleagues for their help in preparing and revising this paper: João Santos, Rita Veloso, Florbela Barreto, Sandra Antunes and Luísa Alice Santos Pereira.

References

Bacelar do Nascimento, M. F. (2001) "Um novo léxico de frequências do português" in Biblos, vol. de *Homenagem ao Professor Herculano de Carvalho* (no prelo).

Brill, E. (1993) *A corpus-based approach to Language Learning*, PhD thesis, University of Pennsylvania, Departement CIS.

Cresti, E., *et al*. (2002) "The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus LREC", in M. C. Rodrigues & C. Suarez Araujo (a cura di), Proceedings of the *Third International Conference on Language Resources and Evaluation*, Paris: ELRA, vol. 1, pp. 2-10.

Moreno, A. & J. M. Guirao (2003) "Tagging a spontaneous speech corpus of Spanish" in *Proceedings of RANLP-2003 – Recent Advances in Natural Language Processing* , (forthcoming).

Van Eynde, F., J. Zavrel & W. Daelemans (2000) "Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus", in Gavrilidou, M. et al. (eds.) *Proceedings of the Second International Conference on Language Resources and Evaluation. European Language Resources Association*, Paris, 1427-1433.