Morphological Tagging of a Spoken Portuguese Corpus Using Available Resources

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 Lisbon
amalia.mendes@clul.ul.pt, ramaro@clul.ul.pt, fbacelar.nascimento@clul.ul.pt

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 Portuguese 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

Annotating a spoken corpus with part-of-speech (POS) information presents certain specificities not found in the annotation of written corpora. Spoken discourse shows, as it is known, many characteristic phenomena that are not found in the written speech or, better still, that when found are disregarded as errors. However, our experience shows that it is possible to attain satisfactory results in spoken texts POS tagging by reusing and adapting resources developed for 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 language in four romance languages: Spanish, Portuguese, French and Italian. This corpus contains 320.000 words and covers several registers: informal (private and public) and formal (natural context, media and phone conversations). One of the main goals of the C-ORAL-ROM is the publication of grammar essays about the spoken language. It was in this context that we faced the necessity of proceeding with the corpus POS tagging and wordform lemmatisation considering always the equation time/effort – accuracy. The objective of not only tagging the corpus with POS information, but also lemmatising its wordforms, increased the complexity of our task and led us to reuse, whenever possible, already available resources.

We proceeded first by revising the tagset used for tagging the PAROLE Portuguese corpus, a 250.000 words written corpus. The experience with the PAROLE tagging process showed that several categories needed to be revised and these observations were used for the new tagset definition. An already executable version of Eric Brill's tagger was also available and had been trained over the tagged and revised PAROLE Portuguese corpus (under the project *Recursos Linguísticos para o Português: um corpus e instrumentos para a sua consulta e análise*[2]). Later, the use of a previously developed resource, a frequency lexicon based on a written 16M word corpus (*Léxico Multifuncional Computorizado do Português Contemporâneo*[3] – LMCPC), proved helpful for the lemmatising task. This resource, hopefully, will also prove to be valuable for the improvement of the tagging task.

The organization of this paper reflects the proceedings of our work. The first part will address some issues related to the tagging process of the written corpus, namely the tagset definition and the achieved results. Secondly, we will describe the tagging process of the spoken corpus in what concerns the specific spoken language phenomena and the tagger's performance. The lemmatisation process will be discussed in the following point, focusing on the adaptation of the

---

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.

existent frequency lexicon and on the limitations of the developed tool. Finally, we will present the achieved results and make some comments on further possible developments.

2. Tagging a written corpus

2.1. Some aspects of tagset definition

The morphosyntactic annotation of the Portuguese PAROLE corpus had already required for a tagset to be defined. However, the team hands-on experience had shown that some of its categories needed to be revised, even in what concerned written texts. The new tagset was defined according to, whenever possible, category definitions of conventional grammar, in order to reach as many potential users as possible.
The established morphosyntactic annotation system covers the main POS categories (Noun, Verb, Adjective, etc.) and the 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.
The prime decisions here presented may be seen as controversial. However, tagset options were made in order to minimize possible inconsistencies due to the reviewer's subjective judgments as well as errors due to the automatic tagging process.
The difficult and time-consuming task of deciding between ambiguous categories was in some cases avoided by the use of portmanteau tags. Therefore, distinctions between

(i) Indefinite Article and Numeral for the annotation of the form *um*, *uma*,

ex: Ele comeu a maçã. (definite article)
Ele comeu duas maçãs. (numeral)
Ele comeu uma maçã. (indefinite article or numeral?)

(ii) the inflected or non-inflected infinitive verb forms,

ex: Os ministros reuniram-se para debaterem o problema. (inflected infinitive)
Os ministros reuniram-se para debater o problema. (non-inflected infinite)
O conselho reuniu-se para debater o problema. (inflected or non-inflected infinitive?)

(iii) some common or proper nouns,

ex: O Presidente morreu ontem. (proper noun)
Mário Soares já não é o presidente. (common noun)
O presidente foi recebido com honras de Estado. (common or proper noun?)

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):

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

Note that this distinction does not increase the error rate in the tagging process since the past participle in compound tenses occurs always in a well-defined context: the past participle is always preceded by an auxiliary verb.
We decided to tag some multi-word expressions, namely prepositional, conjunctional, pronominal and adverbial expressions. However, it was necessary to impose some constraints on what expressions to consider, since, for instance, the adverbial ones are an open class. Therefore, these expressions were only annotated as such when attested in dictionaries. Still, the multi-word expressions set is, unfortunately, an indefinite set.
The information for each tagged element of the multi-word expression includes: category, element position number and identification number (for cross-reference in an appended list of multi-word expressions). The identification number is inserted after the tagging process to avoid multiplying the tagset length.

ex:
num/LADV1_ 117 instante/LADV2_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 multi-word expression may also occur freely, a manual revision was necessary to assure maximum accuracy.
The contracted wordforms were not separated. Instead, they were annotated by joining two tags through the sign '+'.

ex: dos/PREP+ARTd

A similar option was adopted in the case of wordforms connected by hyphen: they received two tags also connected by hyphen.

ex: disse-me/Vppi-CL

These last 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 allowing the combination of several tags that are recognized by the tagger as a new single one (See complete tagset in Annex 1). Although tagset enlargement will most probably increase the error rate of the automatic tagging process, these options were taken in order to avoid texts manipulation and modification.

2.2. Tagging procedure

The 250.000 words PAROLE Portuguese corpus, tagged and manually revised, was used as a training corpus, with a revised tagset. This written corpus covers several genres: newspaper (65%), books (20%), magazines (5%) and varia (10%). Since several changes have been introduced in the PAROLE tagset (see above), the corpus had to be manually revised in order to reflect the new categories and tags.
Eric Brill's tagger (Brill 1993) was then trained over this revised tagged corpus. The tagging process was executed according to the author's instructions, without any modification. We divided the corpus in two halves, one part being used for building the tagger dictionary and lexical rules, and the other part to build the contextual rules file. We decided to use the entire 250.000 word corpus to train the tagger as an attempt to cover all of the tags defined within our tagset. As explained in the previous section, some linguistic options have important consequences on the tagset dimension, since they lead to a high increase of categories. This also reflects on some lemma frequency since some homonymic words seldom occur with a given POS category, causing some distortion in the statistical calculus made by the tagger. So, the use of the entire corpus enabled us to test the tagger's performance regarding two main aspects: the ability to deal with an extremely large tagset, on one hand, and the capacity for extracting lexical and contextual rules concerning rare words, on another. To enlarge the tagger's word recognition capacity, we rebuilt the tagger dictionary using, this time, the entire training corpus.

2.3. Evaluating results

Although the tagset resulted in an indefinite set (since complex tags were introduced), the tagger achieved a 93% success rate, tested over new manually revised fiction texts. At first hand, the success rate can seem inferior to what should be expected, but we consider it satisfactory since we were able to keep our tagset, without having to reduce it and to give up some linguistically motivated tags for reasons of computational feasibility.
After automatically tagging the written corpus, results show that there are two aspects that need attention in the future: first, some difficulties in the automatic tagging of multi-word expressions and, second, errors on certain rare words identification.
Although multi-word expressions are fixed sequences, the tagger does not seem to be able to acquire the necessary information that enables correct annotation. It is true that most of the words included in these expressions also occur as independent elements, what might constitute a problem for a statistically-based tagger. However, we were expecting for the tagger to acquire contextual rules that would deal with these multi-word expressions, and this was not the case. In order to respond to this specific problem, there seems to be two possible solutions. The first one would be the inclusion of the expression identification number in the tagset. However, this is an improbable solution, since it would imply a huge tagset length increase (note that the current prepositional multi-word expressions list alone exceeds 484 elements). The second, and most feasible, solution would be the conception of a post-tagging tool for the multi-word expressions annotation.

As to the identification and correct tagging of rare words, the future development will be to complement the tagger lexicon dictionary with the LMCPC lexicon, extracted from a 16 million words corpus (considerably larger than the training corpus – 250.000 words) and describing a large set of wordforms (around 140.000).

3. Tagging 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%. This success rate was calculated after some automatic post-tagger adaptations since the spoken texts format is very different from the written one, especially in what concerns punctuation and some specific transcription marks.
It was also necessary to work on other post-tagging adaptations that would contemplate the "visible" specific spoken language phenomena, i.e., easily identifiable phenomena directly related to spoken language.
We will now describe and comment on the processes undertaken to achieve the spoken corpus final annotation.

3.1 Specific spoken language phenomena

The characteristic aspects of spoken texts to be treated can be divided in two types: format type (directly related to transcription options), and use type (directly related to the spoken language characteristic phenomena).
The format aspects comprise the prosodic marks tags and the speaker identification label. As it is obvious, these marks were previously accorded and are systematic. However, the tagger identifies and tags the prosodic marks (question marks, slashes, and so on) as punctuation or symbols. Nevertheless, the tag erasure was a simple and automatic process. We present bellow the list of marks treated.

Tagger treatment	Final format	
//O	/	(non-terminal break)
//O //O	//	(terminal break)
?/O	?	(interrogation terminal break)
./O ./O ./O	...	(suspension terminal break)
$/SIMB	$	(alignment mark)
</SIMB >/SIMB	<>	(overlapping marks)
[/O </SIMB]/O	[<]	(overlapping marks)
*INF/SIGL :/O	*INF:	(speaker identification label)

In what concerns the second type of characteristics mentioned above, the required tagset adaptations comprised:

a) extra-linguistic elements; transcription: hhh;	Tag: **EL**
b) fragmented words; transcription: beginning with &;	Tag: **FRAG**
c) words or sequences impossible to transcribe (impossible to hear, for example); transcription: xxx; yyyy	Tag: **Pimp / Simp**
d) paralinguistic elements, such as *hum*, *hã* and onomatopoeias.	Tag: **PL**
e) discourse markers, such as *pá*, *portanto*, *pronto*;	Tag: **MD**
f) discoursive multi-word expressions, such as *sei lá*, *estás a ver*, *quer dizer*, *quer-se dizer*.	Tag: **LD**
g) non-classifiable forms, for words whose context does not allow an accurate classification	Tag: **SC**

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 most of the cases described in (d) since there is a predictable finite list of symbols representing paralinguistic elements. However, onomatopoeias need manual revision, since they can vary from speaker to speaker.
Discourse markers (in (e) and (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*, since there are no frequent (if any) discourse markers in the written language. This fact requires a manual revision (and even listening) in order to decide whether the form is a discourse marker or not. In some cases, the context does not give enough information to decide on which morphological category to assign and so the word receives the tag /SC, meaning that it is a non-classifiable form.
There are two major differences in the tagset definitions for the spoken corpus annotation: the treatment of proper nouns and the treatment of multi-word expressions.
The tagging of proper nouns is simplified in the spoken corpus tagging process, since proper nouns are the only forms transcribed with initial capital letter. Since the decision to consider a word as a proper noun was made during the transcription process, there is no need for the portmanteau tag Np:Nc,. In order to simplify this procedure and to prevent subjective judgments, the only proper nouns considered were toponyms and antroponyms. For instance:

a\ARTd Ana\Np trabalha\Vpi na\PREP+ARTd celbi\Nc com\PREP sede\Nc em\PREP Coimbra\Np

The multi-word expressions tagging process was also simplified since lemma were included in the annotation (see section 4). This way, neither the element position number, nor the expression identification number was required.

3.2 Comments on the results

Some characteristic spoken language phenomena that are not as clearly visible as the ones previously mentioned seem to affect the tagger's performance: it's the case of word repetition and split sentences. This results from the inherent properties of these two distinct types of language. For instance, in the written discourse a sequence of two identical wordforms is an indication that the two forms have different POS category, whereas in the spoken discourse the repetition of two identical wordforms with the same POS is a very frequent phenomena:

é mais fácil **se se** for de carro – se/CONJs se/CL (written text)
não conheço **a a** Maria – a\ARTd a\ARTd (spoken text)

As a consequence, further development will consist in Eric Brill's tagger training over a manually revised spoken corpus, as well as in the exploitation of the tagger contextual rules in order to optimize its performance. Among other things, we aim at improving the multi-word expression tagging process – inherited from the written corpus tagging process – since multi-word expressions account for an increase of around 2% of the error rate.

4. Lemmatisation of the spoken corpus

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.

In order to accomplish this task in a simple, efficient and economical way, we decided to use a corpus-based frequency lexicon of Portuguese (LMCPC) as the source for a lemmatisation tool.

4.1 The LMCPC

The LMCPC is a 26.443 lemma frequency lexicon with 140.315 wordforms, with a minimum lemma frequency of 6, extracted from a 16 million word corpus of contemporary Portuguese. The lemma and its correspondent forms (including inflected forms and compounds) are followed by morphosyntactic and frequency information.
The lemma and wordforms are labeled concerning main POS categories, as N (noun), V (verb), A (adjective), or other, namely F (foreign word), G (acronym/sigla), X (abbreviation). The wordforms with non-canonical orthography were also included under their rightful lemma. Regarding quantitative information, the frequencies were extracted from the POS tagging information and from some calculations (for the problematic forms) based on manually revised data.
Although not used in its full potential, the LMCPC proved to be a very useful resource for our purposes. The .txt format in which it can be displayed made its manipulation extremely easy.

4.2 Lemmatisation process

The lemmatisation of the spoken corpus comprised two major tasks: the formatting of the LMCPC data and the construction of a tool to extract the lemma from the lexicon.
The adaptations required to LMCPC format were due to the different POS tagset adopted in each of the projects: in the LMCPC project we used a main POS category classification, whereas in the Recursos Linguísticos para o Português project as well as in the C-ORAL-ROM project we also used subcategories classification. Unfortunately, we were not able to use, at this stage, the POS information present in the LMCPC to improve the lemma selection process. Therefore, the formatting procedure reduced the LMCPC data to a list of lemma and correspondent wordforms, one per line, being the lemma the first character sequence of the line.
The lemmatisation tool developed turned out to be very simple. It consists in a Perl script that extracts from the LMCPC data file the lemma for each token of the corpus: each form of the corpus is searched for in the lexicon and the

correspondent lemma is(are) found and placed near the form. However, it is possible for a wordform to receive several lemma, requiring thus manual lemma selection.
In the case of multi-word expressions, since the lemma is the entire set of elements, there was no correspondence between the result wanted and the LMCPC data. Therefore, it was necessary to develop a tool to automatically compose the desired lemma format from a given list of expressions. The final format of the lemmatisation of a multi-word expression is given bellow:

o\O_QUAL\LPRON qual\O_QUAL\LPRON

This option made it unnecessary to include both the expression identification and the element position number.

Some forms of the spoken corpus are not lemmatised. These are, obviously, the cases of the extra-linguistic elements, paralinguistic elements, fragmented words, words and sequences impossible to transcribe and proper nouns. In these cases, the empty lemma is represented by a '-':

&frag\-\FRAG

With the foreseen improvement of the tagger's success rate and the adaptation of the POS information present in the LMCPC, we expect to be able to trust the automatic POS tagging and, consequently, to select from the lexicon the proper lemma for each wordform, avoiding the attribution of several lemma.

5. Results

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

```
Text identification: Pfamcv04

*PAU: por\POR_ACASO\LADV acaso\POR_ACASO\LADV / eu\EU\PES sugeri\SUGERIR\Vppi
uma\UMA\ARTi:NUMc coisa\COISA\Nc / lá\LÁ\ADV no\EM+O\PREP+ARTd big\BIG\ESTR
brother\BROTHER\ESTR / que\QUE\RELi era\SER\Vii //$ aquilo\AQUILO\DEMi
lá\LÁ\ADV / ao\AO_PÉ_DE\LPREP pé\AO_PÉ_DE\LPREP da\AO_PÉ_DE+A\LPREP+ARTd
casa\CASA\Nc do\DE+O\PREP+ARTd big\BIG\ESTR brother\BROTHER\ESTR /
na\EM+A\PREP+ARTd parte\PARTE\Nc de\DE\PREP trás\TRÁS\PREP / há\HAVER\Vpi
assim\ASSIM\ADV um\UM\ARTi:NUMc / monte\MONTE\Nc //$ e\E\CONJc /
aquilo\AQUILO\DEMi +$

*NUN: fazer\FAZER\VB turismo\TURISMO\Nc de\DE\PREP guerra\GUERRA\Nc //$
lá\LÁ\ADV ?$

*PAU: < não\NÃO\ADV //$ era\SER\Vii dar\DAR\VB umas\UMA\ARTi > +$

*AMA: [<] < hhh\-\EL / sniper\SNIPER\ESTR > //$

*PAU: não\NÃO\ADV / não\NÃO\ADV //$ < era\SER\Vii dar\DAR\VB > +$

*NUN: [<] < yyyy\-\Simp > a\A\ARTd cabecita\CABEÇA\Nc de\DE_FORA_DE\LPREP
fora\DE_FORA_DE\LPREP da\DE_FORA_DE+A\LPREP+ARTd casa\CASA\Nc //$ < tá\-\PL /
hhh\-\EL > //$

*AMA: [<] < hhh\-\EL >$

*PAU: [<] < não\NÃO\ADV / não\NÃO\ADV / é\SER\Vpi o\O\ARTd
contrário\CONTRÁRIO\Nc > //$ era\SER\Vii / é\SER\Vpi / é\SER\Vpi / é\SER\Vpi
assim\ASSIM\ADV //$ e\E\MD eles\ELE\PES quando\QUANDO\CONJs há\HAVER\Vpi
```

```
muitos\MUITO\INDv / muitas\MUITA\INDv pessoas\PESSOA\Nc a\A\PREP
espreitar\ESPREITAR\VB /$

*AMA: hum\-\PL < hum\-\PL > //$

*PAU: / [<] < o\O\ARTd &mon\-\FRAG > / no\EM+O\PREP+ARTd monte\MONTE\Nc /
eles\ELE\PES não\NÃO\ADV podem\PODER\Vpi vir\VIR\VB cá\CÁ\ADV para\PARA\PREP
fora\FORA\ADV / para\PARA\PREP não\NÃO\ADV comunicarem\COMUNICAR\VBf //$ hhh\-
\EL / lá\LÁ\ADV / os\O\ARTd gajos\GAJO\Nc da\DE+A\PREP+ARTd / da\DE+A\PREP+ARTd
produção\PRODUÇÃO\Nc / mandam-nos\MANDAR-O\Vpi-CL para\PARA\PREP / mandam-
nos\MANDAR-O\Vpi-CL para\PARA\PREP &den\-\FRAG / para\PARA\PREP
dentro\DENTRO\ADV //$ e\E\CONJc a\A\ARTd / a\A\ARTd / solução\SOLUÇÃO\Nc
que\QUE\RELi eu\EU\PES sugeri\SUGERIR\Vppi / foi\SER\Vppi / darem-lhes\DAR-
LHE\Vpi-CL pressões\PRESSÃO\Nc de\DE\PREP ar\AR\Nc //$ < hhh\-\EL >$

*AMA: [<] < hhh\-\EL >$

*PAU: / e\E\CONJc disparar\DISPARAR\VB / sobre\SOBRE\PREP os\O\ARTd
voyeurs\VOYEUR\ESTR / < hhh\-\EL > //$
```

This sample exemplifies many of the issues here addressed. Besides all the previously mentioned aspects, the sample shows that word repetition, sentence overlapping, sentence interruption and sentence reconstruction are recurrent phenomena in the spoken language.
For all these reasons, the development of accurately tagged corpora, especially spoken corpora, is definitely a human resources and time-consuming task.
Nevertheless, 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, aside from 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 lemma construction for the multi-word expression elements;
vii) and, at last, the manual revision of the final result.

6. Further developments

Some further developments concern both the written and spoken tagging processes, namely the automatic tagging of the multi-word expressions and the improvement of the wordform identification process. In order to do so, we will consider the development of a post-tagging tool for the multi-word expressions annotation, on one hand, and we are preparing the introduction of the LMCPC data in the tagger dictionary file, on the other.
Regarding the spoken corpus annotation process, specifically, it is necessary to improve the tagger's performance by training the Eric Brill's tagger over the resulting spoken corpus, manually revised. Depending on this procedure, we also hope to improve the lemmatisation process, through the use of the LMCPC POS information.

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, Department 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.

Annex

The revised tagset used for tagging the spoken corpus includes the following tags:

Categories		**TAGS**
Main Class	**VERB** **AUXILIARY VERB**	**V** **VAUX**
Specifications	Presente do Indicativo	**pi**
	Pretérito Perfeito do Indicativo	**ppi**
	Pretérito Imperfeito do Indicativo	**ii**
	Pretérito Mais que Perfeito do Indicativo	**mpi**
	Futuro do Indicativo	**fi**
	Condicional	**c**
	Presente do Conjuntivo	**pc**
	Pretérito Imperfeito do Conjuntivo	**ic**
	Futuro do Conjuntivo	**fc**
	Infinitivo	**B**
	Infinitivo flexionado	**Bf**
	Gerúndio	**G**
	Imperativo	**imp**
	Particípios passados em Tempos Compostos	**VPP**
	Particípios passados adjectivais	**PPA**
Main Class	**NOUN**	**N**
Specifications	Proper Noun	**p**
	Common Noun	**c**
Main Class	**ADJECTIVE**	**ADJ**
Main Class	**PREPOSITION**	**PREP**
Main Class	**ADVERB**	**ADV**
Main Class	**CONJUNCTION**	**CONJ**
Specifications	Coordenative	**c**
	Subordinative	**s**
Main Class	**NUMERAL**	**NUM**
Specifications	Cardinal	**c**
	Ordinal	**o**
Main Class	**CLITIC**	**CL**
Main Class	**PERSONAL PRONOUN**	**PES**

Main Class	**ARTICLE**	**ART**
Specifications	Indefinite	**i**
	Definite	**d**
Main Class	**DEMONSTRATIVE**	**DEM**
Specifications	Inflected	**i**
	Non-inflected	**v**

Main Class	**INDEFINITE**	**IND**
Specifications	Inflected	**i**
	Non-inflected	**v**
Main Class	**POSSESSIVE**	**POS**
Specifications	Inflected	**i**
	Non-inflected	**v**
Main Class	**RELATIVE/ INTERROGATIVE/ EXCLAMATIVE**	**REL**
Specifications	Inflected	**i**
	Non-inflected	**v**
Main Class	**ADVERBIAL LOCUTION**	**LADV**
Main Class	**CONJUNCTIONAL LOCUTION**	**LCONJ**
Main Class	**PREPOSITIONAL LOCUTION**	**LPREP**
Main Class	**PRONOMINAL LOCUTION**	**LPRON**
Main Class	**INTERJECTION**	**INT**
Main Class	**ENFATIC**	**ENF**
Main Class	**FOREIGN WORD**	**ESTR**
Main Class	**ACRONYMOUS**	**SIGL**
Main Class	**EXTRA-LINGUISTIC**	**EL**
Main Class	**PARA-LINGUISTIC**	**PL**
Main Class	**FRAGMENTED WORD**	**FRAG**
Main Class	**DISCOURSE MARKER**	**MD**
Main Class	**DISCURSIVE LOCUTION**	**LD**
Main Class	**WORD IMPOSSIBLE TO TRANSCRIBE**	**Pimp**
Main Class	**SEQUENCE IMPOSSIBLE TO TRANSCRIBE**	**Simp**
Main Class	**NON-CLASSIFIABLE FORM**	**SC**

SUB-TAGS	TAG
Ambiguous form (ex.: *um\ARTi**:**NUMc*)	**:**
Contracted forms (ex.: *da\PREP+ARTd*)	**+**
Hyphenated forms (ex.: *viu-se\Vppi-CL*)	**-**