Open Resources and Tools for the Shallow Processing of Portuguese

Florbela Barreto,[©] António Branco,[⑦] Eduardo Ferreira,[⑦] Amália Mendes,[©]
Maria Fernanda Bacelar do Nascimento,[©] Filipe Nunes,[⑦] João Ricardo Silva[⑦]

University of Lisbon
[⑦]Department of Informatics, NLX – Natural Language and Speech Group
{ahb,eferreira,fnunes,jsilva}@di.fc.ul.pt
[©]Center of Linguistics, Corpus Linguistics Group
{florbela.barreto,amendes,fbacelar.nascimento}@clul.ul.pt

Abstract

This paper presents the linguistic resources and tools for the shallow processing of Portuguese developed in the scope of a research intitiative at the University of Lisbon. These resources include a 1 million token corpus that has been accurately hand annotated with a variety of linguistic information, as well as several state-of-the-art shallow processing tools capable of automatically producing that type of annotation. At present, the linguistic annotations in the corpus are sentence and paragraph boundaries, token boundaries, morphosyntactic POS categories, values of inflection features, lemmas and named-entities. Hence, the set of tools comprise a sentence chunker, a tokenizer, a POS tagger, nominal and verbal analyzers and lemmatizers, a verbal conjugator, a nominal "inflector", and a named-entity recognizer, some of which underline several on-line services.

1. Introduction

This reports on a research initiative by the Department of Informatics, NLX-Natural Language and Speech Group (coord.) and the Center of Linguistics, Corpus Linguistics Group, both of the University of Lisbon, funded by a research grants of the Portuguese Ministry of Science.

The main objective of this project is to develop linguistic resources and tools for the computational processing of Portuguese. Both the resources and the tools are geared towards shallow morphosyntactic processing.

In Section 2, the TagShare Corpus is presented, with details being provided about the corpus composition, the tagset and the linguistic information that has been included.

In Section 3, an overview of each processing tool is given.

Finally, in Section 4, some final remarks and possibilities for future work are presented.

2. TagShare Corpus

The main linguistic resource developed is a corpus of Portuguese with 1 million tokens that is linguistically interpreted with high quality, accurately hand checked information that is relevant for linguistic research, in general, and for developing and evaluating shallow processing tools, in particular. One of its most important characteristics is the transcribed spoken materials, which correspond to ca. 1/3 of the total corpus.

Before the start of the current research initiative, the Center of Linguistics compiled several spoken materials - within national and international projects, such as the C-ORAL-ROM project (Bacelar do Nascimento et al., 2005) and the Português Fundamental project (Bacelar do Nascimento et al., 1987) - that are now included in the TagShare spoken subcorpus. These materials range from formal to informal registers and display several communicative situations, namely phone calls, media broadcasts, conversations, monologues, formal exposition, etc.

The written subcorpus is partly composed of materials previously gathered, also by the Center of Linguistics, in the PAROLE project (Marrafa et al., 1999), ca. 260 000 tokens. The written texts belong to several genres: newspapers, books, magazines, journals and miscellaneous (proceedings, dissertations, pamphlets, etc.).

A detailed overview of the corpus composition can be obtained from the table below:

Spoken 318,593 tokens	Informal	52.1%	165 838
	Formal	20.8%	66 274
	Media	19.5%	62 116
	Phone	7.6%	24 365
Written 711,135 tokens	Newspaper	60.8%	432 232
	Book	25.7%	182 890
	Magazine	8.5%	60 482
	Misc.	5.0%	35 531
Total			1 029 728

Table 1: Corpus constitution

2.1. Linguistic Information

The TagShare corpus contains linguistic information of different nature and from different levels of sophistication. This information is encoded under the usual format of tags, checked for their accuracy by trained linguists, covering five main levels of information.

Segmentation: The boundaries of each sentence and paragraph are tagged and every token is circumscribed by blanks. Contractions are expanded, clitics in enclisis and mesoclisis are detached into autonomous tokens, and punctuation is associated with explicit information concerning the blanks surrounding them in the raw

version. Multi-word expressions from closed POS classes (e.g. Conjunctions, Prepositions, etc) are identified as forming a lexical unit.

Part-of-Speech: By means of POS tags, each token is associated with an indication of its morphosyntactic category.

Inflection: Information concerning inflectional morphology: every token is associated with explicit information encoding their values for Mood, Tense, Person and Number, if they are from verbal classes, or Number and Gender if they are nominals. Nominals include also information about their degree, namely superlative for Adjectives, and diminutive for both Adjectives and Nouns.

Lemma: Every nominal and verbal token is associated to its lemma, which corresponds to the entry that one would be found in a dictionary for that word. Following the Portuguese lexicographic tradition, for nominal tokens this is usually the masculine singular form of the word when it exists while for verbal tokens the lemma corresponds to the infinitive form of the verb.

MWU for NER: Delimitation and classification of multi-word expressions for named-entity recognition following the usual IOB tagging schema for NER (Ramshaw and Marcus, 1995), and the typical classes of Number, Date, Person, Location, etc. (Chinchor, 1997).

2.2. Tagset

A major goal underlying the TagShare corpus is the construction of a high quality dataset to help develop and evaluate processing tools. Accordingly, the tagset was designed aiming at avoiding as much as possible the data sparseness bottleneck, yet without losing significant linguistic contrast and information.

The final tagset includes the following type of tags:

Major POS: definite article (DA), common noun (CN), Verb (V), etc. These tags are attached to the token, separated by a '/' symbol.

```
o/DA rapaz/CN comeu/V o/DA bolo/CN
The  boy     ate    the  cake
```

Distinguished verb forms: infinitive (INF), gerund (GER), past participle in compound tenses (PPT), and other participles (PPA).

```
ser/INF casado/PPA
to be   married
comendo/GER
eating
tenho/VAUX sido/PPT
I have     been
```

Auxiliary verbs: auxiliary verb (VAUX), infinitive auxiliary verb (INFAUX), gerund auxiliary verb (GERAUX).

```
tenho/VAUX sido/PPT
I have     been
ter/INFAUX sido/PPT
to have    been
tendo/GERAUX sido/PPT
having       been
```

Speech specific elements: discourse marker (DM), extra-linguistic elements (EL), paralinguistic elements (PL), fragment (FRAG), etc.

```
pois/DM, pronto/DM
right    well
hhh/EL
hhh
a/DA chuva/CN faz/V ping/PL ping/PL
the  rain     makes ping    ping
po/FRAG po/FRAG porta/CN
do      do      door
```

MWU: adverbials (LADV*n*), prepositions (LPREP*n*), etc.: each POS tag prefixed with 'L' is extended with information regarding the position (*n*) of the corresponding token inside the multi-word expression.

```
a/LADV1 pé/LADV2
on      foot
ao/LPREP1 contrário/LPREP2 de/LPREP3
on        contrary         of
```

Inflectional feature values: gender: feminine (f), masculine (m) or underspecified (g); number: singular (s), plural (p) or neutral (n); person: first (1), second (2) or third (3); degree: diminutive (dim), superlative (sup) and comparative (comp); mood: indicative (i), subjunctive (c), etc. These tags are attached to the POS tag, separated by a '#' symbol.

```
florzinha/CN#fs-dim
small flower
lindíssimos/ADJ#mp-sup
very beautiful
casou/V#ppi-3s
he/she married
simples/ADJ#gn
simple
```

Lemmas: the lemmas for tokens from nominal and verbal open classes. The lemmas are inserted between the word form and the POS tag, using '/' as the delimiter.

```
flores/FLOR/CN#fs
flowers, FLOWER
lindíssimo/LINDO/ADJ#ms-sup
very beautiful, BEAUTIFUL
casou/CASAR/V#ppi-3s
he/she married, TO MARRY
```

Components of NE expressions: denominators of fractions (DFR), Part of Address (PADR), Social Title (STT), etc.: these tags identify major components of expressions for Named Entities.

```
centésima/DFR parte/CN
hundredth     part
Avenida/PADR João/PNM XXI/DGTR
Avenue       João     XXI
Dr./STT Pedro/PNM Silva/PNM
Dr.     Pedro     Silva
```

NE expressions: Every token in the corpus is marked with one of three tags in accordance with the MUC guidelines under the IOB schema:

'O' (outside): This tag indicates that the corresponding token is not part of a named-entity.

'B' (begin): This tag indicates that the corresponding token is the first token in a named-entity.

'I' (inside): This tag indicates that the corresponding token is part of a named-entity (but it is not the first token).

Additionally, a suffix indicating the type of the named-entity is appended to the tags: person (PER), organization (ORG), location (LOC), work[1] (WRK), event (EVT) and others (MSC).

```
Guarda/B-ORD Nacional/I-ORG Republicana/I-ORG
Guard        National       Republican
Presidente/B-PER Jorge/I-PER Sampaio/I-PER
President        Jorge       Sampaio
O/B-WRK Código/I-WRK Da/I-WRK Vinci/I-WRK
The Da Vinci Code
```

3. Software Resources

The tools that have been developed are geared towards shallow morphosyntactic processing, meaning that lexemes are associated with basic linguistic information that can be uncovered by means of computationally efficient procedures building upon the word structure and/or upon a limited amount of context.

The tools work in a pipeline scheme, where each tool takes as input the result of the previous step. Seen as a black box, the whole pipeline appears as a single tool that takes raw text and delivers that text with the annotations mentioned above. This modular architecture has the advantage of being easily extendable, as new functionality can be easily added by appending a new tool to the pipeline.

Each tool will be described in greater detail in the following sections.

3.1. Sentence and Paragraph Chunker

This tool marks sentence (<s>) and paragraph (<p>) boundaries, unwrapping sentences that are split over different lines. It uses rules that are encoded in the form of a finite state automaton. This allowed to fine-tune the tool for cases that are specific to Portuguese orthography. In particular, the tool properly segments dialog, using the paragraph mark to delimit turn-taking (i.e. the change of speaker).

The tool has a recall of 99.94% and a precision of 99.93%. These results were obtained when testing the chunker on a 12 000 sentence corpus that has been accurately hand tagged with respect to sentence and paragraph boundaries.

For more details, see (Branco and Silva, 2004).

3.2. Tokenizer

This is a rule-based tool that segments text into lexically relevant tokens. Besides the obvious tokenization cases, driven by the whitespace characters already present in the text, this tool handles several Portuguese-specific non-trivial issues. These include such cases as the following:[2]

Expansion of contractions (e.g. between prepositions and other words)

pela • |por|a|

Detachment of clitics in enclisis or mesoclisis

viu-o • |viu|-o|

More important, this tool also handles the cases of those ambiguous strings that can be tokenized in more than one way as for instance deste, which depending on its particular occurrence, can be tokenized as the single token |deste| (if occurring as a verb) or as the two tokens |de|este| (if occurring as the contraction of the preposition de and the demonstrative este).

In Portuguese, there are only 13 such strings, but they amount to 2% of the 260 000 token corpus upon which the tool was tested. Consequently, this is not a minor issue that can be overlooked, and errors at such an early stage of processing will have a considerable negative impact on the subsequent processing steps.

The tokenizer has 99.44% precision for the ambiguous strings, and full precision for the remaining cases.

For more details, see (Branco and Silva, 2003).

3.3. POS Tagger

This tool assigns a single POS tag to each token. For this task, a statistical tagger was trained using the TnT software package, a fast and efficient tagger whose underlying algorithm is based on second-order Hidden Markov Models with back off and suffix analysis (Brants, 2000).

For training, a 260 000 corpus was accurately hand tagged by using a tagset with ca. 60 tags. After training over 90% of that corpus, the tagger achieves 96.87% accuracy, measured with a single run over the 10% of the corpus not used during training.

For more details, see (Branco and Silva, 2004).

3.4. Nominal Featurizer

This tool associates Gender (masculine or feminine) and Number (singular or plural) feature values to words from the nominal classes (Common Noun, Adjective and Past Participle).

Usually, this task is performed by a POS tagger using a tagset that has been extended with feature values. However, this increase in the number of tags typically leads to a lower tagging precision due to the data-sparseness problem. With this tool we explore the approach of taking nominal featurization as a dedicated task.

For the construction of this tool, we build upon the regularities found in morphology and assign feature values based on the termination of a word. More specifically, a set of rules (terminations and their corresponding default inflection values) is built from a reverse dictionary. Any exceptions to these rules can be easily found in machine-readable dictionaries by searching for words with the

1Any kind of artistic production: movies, books, songs, paintings, sculptures, etc.

2In these examples, the '|' symbol will be used to mark the token boundaries more clearly.

designated termination but with feature values that differ from the default. For example:

Rule: words ending in `-são` are feminine singular
Exceptions: `artesão` (*Eng.: craftsman*), `brasão` (*Eng.: crest*), ... which are masculine singular

However, by using only rules and exceptions it is not always possible to assign feature values based solely on the word form. This is mainly due to the so-called *uniform* words, which are lexically ambiguous with respect to their feature values. For instance, the word `ermita` (*Eng.: hermit*) can be masculine or feminine, depending on its particular occurrence.

To tackle this problem, an algorithm was implemented that explores the fact that, in nominal phrases, there is agreement between nominal lexemes from open classes and lexemes from closed classes (e.g. Articles, Demonstratives, Ordinals, Quantifiers, etc.), and that the items of closed classes can be fully listed and associated with their corresponding feature-value tags. This makes it possible to assign tags by "propagating" them from the items of the closed classes.

For instance, when occurring within a NP, the features values for the uniform word `ermita` (*Eng.: hermit*) can be determined by propagating them from the preceding determiner:[3]

`o#ms` • `ermita#ms` (*Eng.: the [masc.] hermit*)
vs.
`a#fs` • `ermita#fs` (*Eng.: the [fem.] hermit*)

When evaluated over a 260 000 token corpus, this tool achieves 99.05% precision while leaving ca. 5% of the tokens with underspecified feature values, yielding a recall score of ca. 95%.

A statistical approach can be used to disambiguate the remaining cases, leading to full recall at 98.38% precision.

Alternatively, many of these underspecified cases can be resolved by a syntactic analyzer that takes advantage of NP-external agreement. An extrapolation of 113 such cases indicates that the syntactic analyzer is able to resolve 84% of them, leading to a great increase in recall. In fact, this increase is large enough to yield an F-measure that surpasses that of the traditional stochastic approach of assigning POS tags extended with inflection information.

For more details, see (Branco and Silva, 2005a).

3.5. Nominal Lemmatizer

This tool assigns to each nominal lexeme in the corpus its lemma, i.e. a canonical, inflectionally normalized form. This form typically corresponds to the masculine singular form, when it exists. For example, the various inflected word forms listed below will receive the same lemma:

`altos` (masculine, plural)
`altinha` (feminine, singular, diminutive)
`altíssimas` (feminine, plural, superlative)
Lemma: `alto` (masculine, singular) (*Eng.: tall*)

For the construction of this tool we again build upon the morphological regularities of words by creating a set of rewriting rules that, depending on the termination of a word, apply a default transformation to that ending. In this way, a single transformation rule allows the proper lemmatization of a large amount of words.

Rule: rewrite `-ta` into `-to`
`aberta` • `aberto` (*Eng.: open*)
`adulta` • `adulto` (*Eng.: adult*)
`alta` • `alto` (*Eng.: tall*)

Naturally, these rules must be supplemented by a list of exceptions. For example:

`carta` • `carta` (*Eng.: letter*)
`porta` • `porta` (*Eng.: door*)

Note that there are some cases where the lemma does not depend solely on the word form. The most problematic of these cases is when the lemma depends on the sense of the token at stake. For instance, the word `copas` may refer to the suit of playing cards (*Eng.: hearts*) or it might be the plural form of the word `copa` (*Eng.: cupboard*). This kind of sense ambiguity cannot be successfully resolved without a previous step of word sense disambiguation. Consequently, this presents an inevitable upper bound for the shallow lemmatization process, preventing it for ever achieving total coverage of the targeted word form.

The tool was evaluated over a list of ca. 10,500 Adjectives and Common Nouns in the vocabulary of the TagShare Corpus. In this list there are 19 tokens that are lexically ambiguous with respect to lemmatization. As these cases cannot be resolved, on can infer a recall score of 99.82%. Over the remaining tokens, the tool achieves a precision of 94.90%. The errors that were found are due to words that are missing from the exceptions list and also to compound words, which the current algorithm does not yet handle properly.

For more details, see (Branco and Silva, 2005b).

3.6. Verbal Featurizer and Lemmatizer

This tool assigns to each verbal token its lemma (the infinitive form) and an inflection tag indicating its values for mood, tense, person and number. Note that verbs are handled separately from nominal tokens as, in Portuguese, verbal inflection is a much more complex problem than nominal inflection.

Currently, the tool does not attempt to perform disambiguation. Hence, if a verbal type can have several possible lemma-feature pairs, they are all assigned to its occurrence. For example:

`diria` (*Eng.: would say*)
- `dizer, Cond-1s`
- `dizer, Cond-3s`
- `diriar, PresInd-3s`
- `diriar, ImperAfirm-2s`

The first two lemma-feature pairs correspond to the first and third persons of the conditional of the verb

3Note that, in this example, the arrow denotes the propagation of feature values. Additionally, POS tags were omitted for the sake of clarity.

`dizer` (*Eng.: to say*). The last two pairs correspond to a neologism, i.e. the infinitive form `diriar` is not a known verb[4] but, if conjugated with the listed features, it would produce `diria`.

Evaluated over a corpus of 12 000 fully conjugated verbs (which results in ca. 800 000 verb forms), this fully-fledged lemmatizer without disambiguation achieves 100% precision but only at 50% recall, as half of the verb forms receive more than one lemma-feature pair.

On the top of this exhaustive lemmatizer, we developed a tool for lemmatization with disambiguation, which scores 96.51% accuracy (Branco et al., 2005c).

3.7. Named-Entity Recognizer

This tool delimits and classifies various multi-word expressions with a XML-like markup.

In its current version, the tool identifies numbers, dates, addresses and measure expressions. The identified expressions are then classified and assigned a representation in a canonical format. For example, both `05-10-2005` and `5 de Outubro de 2005` are classified as being a date, and receive the same canonical representation, with explicit fields for the values of the year, month and day. In this regard, the assignment of a canonical representation can be seen as a shallow information extraction procedure. For example:

```
<EN Type='Date' Year='05' Month='10' Day='05'>
05-10-2005
</EN>
```

and

```
<EN Type='Date' Year='05' Month='10' Day='05'>
5 de Outubro de 2005
</TIMEX>
```

Every named-entity that the tool currently recognizes has some sort of internal structure, hence we opted for a rule-based, pattern matching algorithm. However, many other entities have a more free-form structure. Ongoing work will extend this tool with a stochastic approach which, trained over the IOB tagging scheme mentioned previously, will allow it to also recognize expressions that refer to persons, institutions and locations.

3.8. Other tools

Some other tools have also been developed that, though not strictly part of the pipeline (as they do not add any layer of annotation to the corpus), are still worth mentioning.

A verbal conjugator tool has been developed. It was used in the development of the exhaustive verbal lemmatization tool mentioned above in order to help ensuring its full accuracy. However, the verbal conjugator can also function as a standalone tool, taking an infinitive form and generating all of its inflected forms. Its most important feature is that this generation is exhaustive, i.e. it includes full pronominal conjugation, compound tenses, regular and irregular forms for past participles, inflected past participles, negative imperative forms and courtesy forms for second person.

A tool for producing inflected nominal forms has also been developed. This tool takes a word and a set of feature values and generates the corresponding inflected form.

4. Final Remarks

This paper presented the linguistic resources that were developed in the scope of a reserach initiative at the University of Lisbon.

The resources comprise a 1 million token corpus and a wide range of shallow processing tools. Both these resources have been developed in tandem: Initial versions of the tools were used to provide an initial, automatic tagging of the corpus that was then hand checked by trained linguists. The accurately tagged corpus was then used to evaluate the tools and to iteratively help to get improved versions of them.

The corpus is constituted by several types of text. Of special importance is the fact that approximately one third of the corpus is composed by transcribed spoken materials. The corpus also contained several levels of linguistic annotation, such as POS tags and lemmas.

The shallow processing software tools that have been developed work in a modular fashion, with a specialized tool for each level of annotation present in the corpus.

Some of these tools can currently be tested in on-line services:

The pipeline of tools, up to the POS tagging step, can be seen at `http://lxsuite.di.fc.ul.pt`.

The verbal featurizer and lemmatizer is available as a service at `http://lxlemmatizer.di.fc.ul.pt`.

The verbal conjugator is also available as an on-line service at `http://lxconjugator.di.fc.ul.pt`.

Future work will focus both on the corpus and on the tools. In particular, the annotation of the corpus will evolve to a more structured XML-like markup[5] while the current tools will be improved (e.g. better exceptions lists) and new tools will be added (e.g. a noun chunker and a stochastic named-entity recognizer) to the pipeline.

5. References

Bacelar do Nascimento, Maria Fernanda, M. L. Garcia Marques and M. L. Segura da Cruz (1987). *Português Fundamental, vol. II - Métodos e Documentos, tomo 1 - Inquérito de Frequência*. Lisboa: INIC, CLUL.

Bacelar do Nascimento, Maria Fernanda, José Bettencourt Gonçalves, Rita Veloso, Sandra Antunes, Florbela Barreto and Raquel Amaro (2005). 5. The Portuguese corpus. In Emanuela Cresti and Massimo Moneglia (eds.) *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins. pp. 163–207.

Branco, António and João Ricardo Silva (2003). Contractions: Breaking the Tokenization-Tagging

4The tool checks the lemmas against a list of approximately 16,500 infinitive forms of known verbs. Lemmas that are not found in this list are considered to be neologisms. Note that, by using this list as a filter it is possible for the tool to generate only known forms.

5This permits richer annotation schemes while actually making it easier to parse the token structure and access that annotation.

Circularity. In LNAI 2721, Berlin, Spinger, ISSN 0302-9743, pp.167–170.

Branco, António and João Ricardo Silva (2004). Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In *Proceedings of the 4th Language Resources and Evaluation Conference* (LREC), pp. 507–510.

Branco, António and João Ricardo Silva (2005a). Dedicated Nominal Featurization of Portuguese. In *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese* (PROPOR).

Branco, António and João Ricardo Silva (2005b). Nominal Lemmatization with Minimal Word List. University of Lisbon, ms.

Branco, António, Filipe Nunes and João Ricardo Silva (2005c). Verb Analysis in an Inflective Language: Simpler is better, submitted.

Brants, Thorsten (2000). TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing and the 1st North American Chapter of the ACL*. pp. 224–231.

Chinchor, Nancy (1997). MUC-7 Named-Entity Task Definition (version 3.5). At:
www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html

Marrafa, Palmira, José Bettencourt Gonçalves and Amália Mendes (1999). A Sintaxe do LE-PAROLE. In Palmira Marrafa and Maria Antónia Mota (eds.) *Linguística Computacional. Investigação Fundamental e Aplicações*. Lisboa: APL/Colibri, pp. 191–205.

Ramshaw, Lance and Mitchell Marcus (1995). Text Chunking Using Transformation-Based Learning. In ACL 3rd Workshop on Very Large Corpora, pp. 82–94.

TagShare (2006). Manual de Etiquetação e Convenções. University of Lisbon, ms.