Linguistic Resources and Software for Shallow Processing

Florbela Barreto,[†] *António Branco,*[‡] *Eduardo Ferreira,*[‡] *Amália Mendes,*[†] *Maria Fernanda Bacelar do Nascimento,*[†] *Filipe Nunes*[‡] *and João Ricardo Silva*[‡]

University of Lisbon

[†]Center of Linguistics, Corpus Linguistics Group

{florbela.barreto,amendes,fbacelar.nascimento}@clul.ul.pt

[‡]Department of Informatics, NLX Group

{ahb,eferreira,fnunes,jsilva}@di.fc.ul.pt

Abstract

This paper presents linguistic resources and software composed by a hand-tagged corpus with 1 million tokens and several shallow processing annotation tools.

1 Introduction

The main goal of this paper is to present the linguistic resources and software developed in a research initiative undertaken by the Department of Informatics, NLX-Natural Language Group (coord.) and the Center of Linguistics, Corpus Linguistics Group, both of the University of Lisbon, under research grants of the Portuguese Ministry of Science.

The main linguistic resource developed is a corpus with 1 million tokens, accurately hand-annotated with several layers of linguistic information. Additionally, various word lists were collected. The implemented software is composed by a set of several shallow processing annotation tools.

In Section 2, we give an overview of the corpus and the tools. In Section 3, each level of linguistic annotation is described, together with the tool that is associated with that level. Finally, in Section 4, we present some final remarks as well as possible improvements for future work.

2 Corpus and Tools

The TagShare corpus is a new corpus of Portuguese with 1 million tokens that is linguistically interpreted with high quality, accurately hand checked information relevant for linguistic research, in general, and for developing and evaluating shallow processing tools, in particular.

Against the backdrop of the existing language resources for Portuguese, another unique feature of this corpus worth underlining is that a considerable part (ca. 1/3) of it consists of transcribed spoken materials.

The spoken subset of the corpus was compiled from repositories gathered by the Center of Linguistics before the start of the current research initiative, in the scope of national and international projects, namely the C-ORAL-ROM corpus (Bacelar do

Nascimento *et al.*, 2005) and the Português Fundamental corpus (Bacelar do Nascimento *et al.*, 1987).

This spoken subcorpus includes materials from several registers (ranging from formal to informal) and several communicative situations (e.g. phone calls, media broadcasts, conversations, monologues, formal exposition, etc.).

The majority of the corpus, however, is composed of written texts. This also covers several genres: newspaper, books, magazines, journals and miscellaneous types, such as proceedings, dissertations and pamphlets. Parts of these materials were previously gathered also by the Center of Linguistics in other projects, namely the PAROLE project (Marrafa *et al.*, 1999).

A detailed breakdown of the corpus composition can be seen in Table 1, below:

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

Table 1: Composition of the TagShare corpus

Another result of the work in the TagShare corpus is a set of annotation tools, the LX-Suite. The tools in this suite follow a shallow processing rationale, where each tool is highly specialized and tackles a circumscribed NLP task. Additionally, shallow processing tools only make use of very local information (a single token or a limited window of context).

The tools work in a pipeline scheme, where each tool adds new annotation to the input text by building upon and adding linguistic information to the annotation that is already present. This information is typically added in the form of tags that are appended to the tokens.

The suite presented in this paper is formed by a sentence chunker, a tokenizer, a POS tagger, featurizers (one for nominal inflection and another for verbal inflection), lemmatizers (again, one for the lemmatization of nominal tokens and another for the lemmatization of verbal tokens), a verbal conjugator and a named-entity recognizer.

3 Linguistic Annotation

There are several layers of linguistic annotation present in the TagShare Corpus, ranging from the marking of sentence boundaries to the identification and classification of named-entities. Naturally, these annotation layers are of different nature and complexity, as are the tools that produce them.

Each of these layers and tools will be discussed in greater detail in the following sections.

3.1 Sentence Chunking

This task is frequently overlooked in the literature, but its importance should not be understated: The sentence is the maximum syntactic element in natural languages. Accordingly, one must know where each sentence begins and ends to be able to analyze it syntactically. However, from a purely computational point of view, unprocessed text is nothing more than a string of characters. Thus, it is necessary to have some sort of annotation scheme to mark sentence boundaries.

In the TagShare corpus, the sentence and paragraph boundaries are marked with XLM-like tags. More specifically, '`<s>...</s>`' delimits sentences and '`<p>...</p>`' marks paragraphs. The tool that adds this annotation is the sentence chunker.

In Portuguese, as well as in other languages with similar orthographic conventions, the ending of a sentence is signaled by specific punctuation symbols, the *terminator symbols*, such as `.` (period), `?` (question mark), `!` (exclamation mark) or `...` (ellipsis). This fact, when coupled with the convention stating that sentences begin with a capital letter, greatly eases the detection of sentence boundaries for most cases. For instance:

```
<s>Isto é uma frase.</s><s>Isto é outra frase.</s>
```

There are, however, harder cases which are also handled by this tool.

Dialog. Each language typically has very specific orthographic conventions pertaining the representation of dialog. Portuguese, in particular, uses paragraph boundaries to mark turn-taking, i.e. the change of speaker. The first sentence in character's turn is marked by a — (dash) symbol while the following sentences are separated as by the usual conventions:

```
<p><s>— Como estás?<s></p>
<p><s>— Estou bem, obrigado.<s></s>E tu?</s></p>
```

The narrator's aside is marked by the same dash symbol that marks the first sentence in a turn, but it does not initiate a new sentence and does not begin with a capital letter. It also influences the marking of the speaker's sentence: If that sentence was to be terminated by a period, that symbol is omitted. Any other terminator symbol is maintained for intonational purposes:

```
<p><s>— Bom dia — disse ele.</s></p>
<p><s>— Olá! — exclamou ela.</s></p>
```

Any sentence following a narrator's aside, even if within the same speaker's turn, must be marked by a dash. It is worth noting that the narrator's aside might even interrupt the speaker's utterance. In that case, the speaker's sentence continues unaltered after the narrator's aside is finished:

```
<p><s>— Não — respondeu.</s><s>— Eu não.</s></p>
<p><s>— Eu cá — disse — também.</s></p>
```

Abbreviations. Abbreviated forms pose a problem because, when followed by a word that begins with a capital letter, they originate an ambiguous situation where the chunker must decide if the period symbol marks the end of the sentence or is part of an abbreviation.

```
<s>Ocorreu um acidente na Av. Lusíada.</s>
```

Other cases. There are several other, rarer cases that, for completeness sake, are also handled. For instance, the occurrence of multiple terminator symbols, embedded quotations, etc.

The sentence chunker was implemented as a finite state automaton. Certain sequences in the input trigger state changes and emit boundary markings. As a simplified example, a terminator symbol followed by a capital letter would emit a sentence boundary mark.

This sentence chunker was evaluated over 12,000 sentences from the TagShare corpus, accurately hand-tagged for sentence and paragraph boundaries. The tool presents a recall of 99.94% and a precision of 99.93%.

More details, including a comparison with a statistical sentence chunker, can be found in (Branco and Silva, 2004).

3.2 Tokenization

Tokenization is the task of segmenting the text into lexically relevant elements, the tokens. For most cases, raw text is already segmented, as the orthographic conventions for most languages state that the whitespace character is used to separate words. However, by those very same conventions, not all tokens occur separately in raw text, thus the need for a tokenization step.

In the TagShare corpus, we maintain the whitespace character as the token separator. This eases the tokenization task for most cases. For instance:[1]

```
um exemplo simples • |um|exemplo|simples|
```

Most of other cases that the tokenizer handles are also simple: Punctuation symbols are separated from word tokens to which they are attached, contracted forms are expanded and clitic pronouns (when occurring in enclisis or mesoclisis position) are detached from the corresponding verb:

[1] In these examples we will use the | (vertical bar) symbol, instead of the whitespace, to mark token boundaries more clearly.

```
um, dois e três • |um|,|dois|e|três|
da • |de|a|
viu-o • |viu|-o|
```

At this point, it is worth noting that all tools try to ensure that the linguistic information in the text grows monotonically. That is to say, information is added to the text, but never removed. Although it might not be obvious, tokenization loses information. This can be seen, for instance, in the previous example, when tokenizing the word “da”, which is a contracted form: After tokenization, there is no way of knowing if the sequence of tokens “de a” originally occurred as two separate tokens or as a contracted form.

To prevent the loss of this information, the first token of an expanded contracted form is flagged with the '+' (plus) symbol. For the same reason, punctuation tokens are marked with the '*' (star) symbol to indicate adjoining spaces while the clitic removal procedure signals possible vocalic alterations of the verb with a '#' and leaves a '-CL-' mark at the original mesoclisis position:

```
da • |de+|a|
3.200 • |3|.|200|
um, dois e três • |um|,*|dois|e|três|
afirmá-lo-ia • |afirmá#-CL-ia|-lo|
```

The main difficulty in the cases seen so far is not in the complexity of the cases themselves but in finding such rare cases. The actual non-trivial aspect regarding tokenization of Portuguese text involves the tokenization of ambiguous strings that, depending on their morphosyntactic category, should or should not be considered a contracted form. For example, the word “deste” can be tokenized as “|deste|” when occurring as a Verb (*Eng.: [you] gave*) or as “|de+|este|” when occurring as the contraction (*Eng.: of this*) of the Preposition “de” and the Demonstrative “este”.

The problem raised by ambiguous strings is not a minor issue that can be overlooked. Such strings occur rather frequently, amounting to 2% of the corpus we used, and an error at such an early stage of processing will have a considerably negative influence on the subsequent processing steps, such as POS tagging.

To tackle the tokenization of ambiguous strings, one must first realize that they pose a circularity problem: To correctly tokenize an ambiguous string, one must know its POS tag, but POS tags can only be assigned after a tokenization step.

To break the circularity and solve this problem, a two-stage tokenization method is used, where the ambiguous strings are not immediately tokenized. Instead, the decision if left for the POS tagger. To allow for this, the tagger must be trained on a version of the corpus where the ambiguous strings are not tokenized. These strings, when occurring as a contracted form, are tagged with a portmanteau tag that combines the tags of both elements in the contraction. For instance, a 'PREP+DEM' tag would be assigned to a contraction of a Preposition and a Demonstrative. After the POS tagging step, the second stage of the tokenizer looks specifically for occurrences of these combined tags and splits the corresponding tokens:

```
deste/V • |deste/V|
deste/PREP+DEM • |de+/PREP|este/DEM|
```

This two-stage approach allowed us to correctly tokenize 99.4% of the ambiguous strings in the corpus.

More details about the two-stage method for resolving ambiguous strings and about the tokenization procedure in general, can be found in (Branco and Silva, 2003).

3.3 POS Tagging

Part-of-speech tagging is one of the most well studied tasks of linguistic annotation. This task requires designing a tagset and some procedure for assigning POS tags to tokens. As there are already several trainable, high-quality, language independent taggers available, our effort was directed towards developing a suitable tagset.

Naturally, the tagset includes the major POS tags, such as Definite Article (DA), Common Noun (CN), Adjective (ADJ) and Verb (V):

```
os/DA lobos/CN perseguiram/V a/DA presa/CN
```

However, when extending the tagset beyond these basic tags, one must bear in mind that the driving motivation behind this tagset is not only to allow linguistic analysis but also to be usable for training automatic POS taggers. These two requirements are in opposition, as the tagset must be rich enough to provide a high level of linguistic discrimination but, at the same time, have few tags to avoid data-sparseness problems.

Nonetheless, some guidelines can be followed that allow the tagset to balance these two requirements.

- Being that POS tags encode differences in syntactic distribution, different tags that are not justifiable by a difference in distribution are not included. For instance, tags indicating the degree of an Adjective.
- Tags that convey a difference in distribution but that can be inferred from the form of the token are not included. This is the case, for instance, of tags indicating inflectional features or the polarity of an adverb.
- Tags used to indicate differences in the constituency status of the phrase containing the relevant token are not included. For example, the tag for “indefinite pronoun”.

The guidelines that were followed do more than just excluding tags. They also lead us to consider tags that would normally not be included in a tagset.

For instance, although Gerund, Past Participle and Infinitive are all verbal forms, each has a particular distribution because they are predicators of subordinate clauses with a specific distribution. For that reason, these verbal forms have their own tags, and a different tag is provided for when verbal forms occur in an auxiliary function. Similarly, the Past Participle is split into those that occur in compound tenses (PPT) and those that occur with adjectival force (PPA):

```
tenho/VAUX estado/PPT
fui/V casado/PPA
cantando/GER
terem/INFAUX estado/PPT
tendo/GERAUX estado/PPT
```

Additionally, as a portion of the TagShare corpus consists of transcribed speech, some speech specific tags where included in the tagset, such as discourse markers (DM), extra-linguistic elements (EL), paralinguistic elements (PL) and fragments (FRAG):

```
olha/DM, pois/DM
hhh/EL
o/DA piano/CN faz/V pim/PL pim/PL
ga/FRAG ga/FRAG gato/CN
```

Finally, multi-word units (MWU), as the name implies, should be tagged as a single syntactic unit. Being that these sequences have already been tokenized as separate words, the tagset accounts for this by using indexed tags that bind several tokens under a same tag: Each POS tag in a MWU is prefixed with 'L' and extended with a number indicating its position inside the multi-word expression:

```
pois/LADV1 então/LADV2
a/LPREP1 respeito/LPREP2 de/LPREP3
sempre/LCJ1 que/LCJ2
é/LDM1 assim/LDM2
```

The final tagset has ca. 60 base tags. In the actual TagShare corpus ca. 100 different tags are found due to the indexed tags that are used for MWU.[2]

The POS tagger was developed with the TnT software (Brants, 2000), which uses Hidden Markov Models. The tool was trained over 90% of a small, accurately hand tagged corpus with approximately 260,000 tokens. Accuracy of 97.22% was obtained with one run test over a held out evaluation corpus with the 10% of the corpus not seen during training. This result is in line with state-of-the-art POS taggers for other languages like English or German, and probably the best for Portuguese (Branco and Silva, 2004).

3.4 Nominal Featurization

The assignment of inflection features is typically done during the POS tagging stage by using a tagset where the normal POS tags have been extended with inflection information. For instance, instead of having a single Adjective tag, the tagset would have four tags for Adjective, one for each possible combination of Gender and Number. However, extending the tagset in this way leads to worse tagging results due to the data-sparseness problem: For the same amount of training data, as the tagset

[2] For a complete overview of the tagset and a detailed account of the annotation decisions and procedures, see TagShare (2006).

increases, there will be more parameters that are harder to estimate due the lower availability of significant data.

Having this problem in mind, we explored what could be gained by having a standalone task for assigning inflection to nominal tokens.

Nominal featurization is thus defined as the procedure that assigns an inflection tag (Gender and Number) to nominal tokens. There are several categories that can bear inflection tags, such as Determiners or Demonstratives, but the main difficulty is in handling the opens categories (Adjectives, Common Nouns and Past Participles) as the inflection tags for words from the other nominal categories can be exhaustively listed.

Note that, when applicable, this procedure also assigns degree information (Superlative, Diminutive or Comparative):[3]

```
exemplos/CN • exemplos/CN#mp
altíssimas/ADJ • altíssimas/ADJ#fp-sup
casinha/CN • casinha/CN#fs-dim
```

To implement the nominal featurizer we followed several approaches: a rule-based approach, a statistical approach and a hybrid of the two.

In this article we report in greater detail on the rule-based approach as it best highlights the difficulties posed by this task.

Rule-based featurizer. To tackle nominal featurization through a rule-based approach we build upon the morphological regularities present in the language and collect a list of word terminations and their default feature values, i.e. a list of rules to assign a feature value to words based on their termination. Naturally, these rules have to be supplemented with a list of exceptions. For example:

Default rule: words ending in `-ção` are feminine singular
Exceptions: `cação`, `coração`, ... are masculine singular

Nevertheless, using rules and exceptions is not enough to ensure that every token receives an inflection tag. This is due mainly to the so-called *invariant* words, which are lexically ambiguous with respect to some feature value. For example, the word "`pianista`" (*Eng.: pianist*) can be masculine or feminine. By resorting only to the procedure described above, words such as this would always be tagged with underspecified inflection tags:

```
pianista/CN#?s
```

The basic algorithm is thus extended with a procedure that builds upon the fact that there is Gender and Number agreement within an NP.

For this procedure, one collects all words from the closed classes that have inflection features (Demonstratives, Determiners, Quantifiers, etc.) together with their

[3] The inflection feature values are `m`:masculine, `f`:feminine, `s`:singular and `p`:plural. For degree, they are `dim`:diminutive, `sup`:superlative and `comp`:comparative. These tags are appended to the POS tags, separated by a '#'.

respective inflection tags. When tagging a text, the inflection tags assigned to words from closed classes are propagated to the words from open classes that follow them. These, in turn, can propagate those values to other words.

For instance, the features for invariant word “pianista” can be easily determined if that word is preceded by a Definite Article. The feature values it received can then be again propagated to the invariant Adjective “ilustre” (*Eng.: illustrious*) that follows the Common Noun.

o/DA#ms pianista/CN#ms ilustre/ADJ#ms
vs.
a/DA#fs pianista/CN#fs ilustre/ADJ#fs

To prevent errors in tagging when using this mechanism, one must ensure that the propagated values remain within the boundaries of the NP. For this purpose, the featurizer uses a set of POS patterns that, when found in the text, block propagation from occurring. For instance, a typical case is that of a PP inside an NP context:

faca/CN#fs de/PREP aço/CN#ms azul/ADJ#fs
Eng. blue (steel knife)
or
faca/CN#fs de/PREP aço/CN#ms azul/ADJ#ms
Eng.: (blue steel) knife

In this example, “azul” (an Adjective, invariant with respect to Gender) might agree with “faca” or with “aço”. To prevent an erroneous tagging, propagation of feature values from “aço” to “azul” must be blocked. As “azul” is an invariant word, it will be tagged with an underspecified Gender value.

Besides the blocking of propagation, there are other situations that might lead to token being assigned an underspecified feature value. This happens in the so-called bare NPs, which do not have a specifier preceding, but also in non-bare NPs, provided that the specifier is itself an invariant word. Both these cases can be seen in the following example:

Ele/PRS#ms detesta/V pianistas/CN#?p
Eng.: He hates pianists
and
Cada/QNT#?s pianista/CN#?s
Eng.: Each pianist

One could attempt to resolve these underspecified tags by resorting to some heuristic, but for this rule-based tool we followed the rationale that it is preferable to abstain from tagging than it is to tag wrongly. This ensures a higher precision at the expense of not having full recall.

The two other approaches to featurization, however, always assign a fully specified inflection tag to every token.

Statistical featurizer. The statistical approach uses the TnT software (Brants, 2000) to train an HMM featurizer. For this approach, the tagger is trained over a corpus where the hidden states are the concatenation of word forms and POS tags and the inflection tags are the emitted symbols. Non-nominal tokens are tagged with a "`null`" inflection tag.

When tagging, the HMM featurizer runs over an input that has already been POS tagged and assigns inflection tags. For instance, taking one of the previous examples:[4]

Input:	`oDA pianistaCN ilustreADJ`
Output:	`oDA/ms pianistaCN/ms ilustreADJ/ms`

Hybrid featurizer. The hybrid approach proceeds as by the rule-based approach but falls back to the tag assigned by the statistical featurizer whenever the rule-based featurizer abstains from assigning a fully specified tag,

To evaluate these different featurizers, they were run over an accurately POS tagged corpus. In this way, there are no POS error to negatively influence the outcome of the featurizer. Additionally, we report only on the evaluation results for the featurization of Adjectives and Common Nouns. This is done for the sake of comparability, as the inventory and definition of the other inflected categories is variable among languages, and even open to dispute within a single language.

All three featurizers were run over a list of 8,750 Adjectives and Common Nouns from the vocabulary of the TagShare corpus.

The rule-based featurizer required a lexicon of words from closed classes and their respective inflection tags with 1,000 entries. The rule list was built with the help of a reverse dictionary. It has approximately 200 termination-inflection pairs. For these rules, we had to collect 9,500 exceptions, giving an average of 47.5 exceptions for each rule.

The rule-based featurizer achieves 95.05% recall, leaving only ca. 5% of the tokens with underspecified feature tags. As for precision, it scores 99.05%.

The statistical featurizer has 100% of recall but at the expense of a lower precision, achieving only 97.58% in this regard.

The hybrid featurizer combines the strengths of both approaches, achieving full recall while still maintaining a high precision, scoring 98.38%.

For a much more extensive analysis, see (Branco and Silva, 2005a).

3.5 Nominal Lemmatization

The next annotation layer in the TagShare corpus is the one concerning the lemma of nominal tokens (Adjectives, Common Nouns and Past Participles). The lemma corresponds to the entry that would be found in a dictionary for that word. Typically, this is the masculine singular form.

[4] In this example the inflection tags are shown separated from the token by a '`/`' instead of a '`#`' only to better highlight the fact that they are being assigned by a HMM tagger.

We followed a rule-based approach to this task. As with featurization, collecting an exhaustive list of every word and its lemma is not viable. We thus follow the same idea used for the featurizer and build upon morphological regularities in words.

The basic rationale is to gather a set of transformation rules that, depending on the termination of a word, replace that termination by another. For instance, a rule stating that a "-ta" termination should be replaced by a "-to" termination would correctly assign a lemma to the following tokens:

alta (*Eng.: [feminine] tall*) • alto (*Eng.: [masculine] tall*)
and
gata (*Eng.: [feminine] cat*) • gato (*Eng.: [masculine] cat*)

Naturally, there will be exceptions to these rules. For example, the feminine Common Noun "porta" (*Eng.: door*) falls under the above rule but its lemma is "porta" instead of "porto". These exceptions were collected by resorting to a machine-readable dictionary (MRD): As dictionaries only list lemmas and never the inflected forms, any word with the designed termination that is found in the dictionary must necessarily be an exception.

Even if this exceptions list grows to a large size, one must bear in mind that the number of word covered by a transformation rule is much larger than the number of exceptions one must collect for that rule. Additionally, any new words that eventually enter the lexicon will typically fall under the general rule instead of being exceptions.

It is worth noting that there are some cases where the lemma the tool should assign does not depend solely on the word form.

The most important of those cases is that of lemmas that depend on the sense of the word on that particular occurrence. For instance, the word "copas" may refer to the *hearts* suit of playing cards, in which case its lemma is "copas", or it might be the plural form of "copa" (*Eng.: cupboard*), in which case it should be lemmatized into the singular form "copa".

At this stage of processing it is not possible to resolve the ambiguity of these sense-dependent words, as that would require some sort of word sense disambiguation procedure. Following the rationale that it is preferable to abstain from tagging than it is to tag wrongly, the lemmatizer assigns all possible lemmas to such words. Consequently, these cases pose an effective upper-bound to the recall of any shallow nominal lemmatization process.

To implement the lemmatizer a list of 126 transformation rules was needed. The list of exceptions to these rules amounts to 9,614 entries.

The lemmatizer was evaluated over a list of 10,581 Adjectives and Common Nouns from the vocabulary of the TagShare corpus, where it achieved 99.82% recall and 94.90% precision.

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

3.6 Verbal Featurization and Lemmatization

The TagShare corpus also includes annotation for inflection tags and lemmas of verbal tokens. We chose to handle these separately from the nominal tokens as the task is intrinsically harder due to verbal inflection being more complex than nominal inflection.

In fact, while lemma ambiguity is rare for nominal tokens, most verbal forms will have several lemma-feature pairs. Following the same rationale as its nominal counterpart tools, all possible results are assigned to the token being processed. For instance, the form "`diria`" receives the following lemmas and features:

```
diria
• dizer,Conditional-1s
• dizer,Conditional-3s
• diriar,PresentIndicative-3s
• diriar,ImperativeAfirmative-2s
```

The first two possibilities are inflections of the verb "`dizer`" (*Eng.: to say*) while the following two pairs show a neologism, i.e. "`diriar`" is an infinitive form that, under the specified inflection, would inflect to "`diria`". Note that "`diriar`" is not an attested verb.[5] If the user so wishes, only attested forms are generated.

Tested over a list of ca. 80,000 verbal forms, this tool achieves 50% recall, as half of the verbal tokens are ambiguous and consequently receive more than one lemma-feature pair. However, those that do receive a single lemma-feature pair are always correct, ensuring 100% precision.

3.7 Verbal Conjugator

Strictly speaking, the verbal conjugator does not add any annotation layer to the TagShare corpus. Within the pipeline of the LX-Suite, this tool is only used by the verbal lemmatizer/featurizer to filter extraneous lemma-feature pairs.

However, the conjugator can also function as a standalone tool. In this mode, the verbal conjugator takes a verb in the infinitive form and generates all of its inflected forms. It is particularly important to note its exhaustiveness, in that it handles a much wider range of inflection situations, namely: full pronominal conjugation, compound tenses, regular and irregular forms for past participles, inflected past participles, negative imperative forms and courtesy forms for second person.

3.8 Named-Entity Recognition

The TagShare corpus also includes an annotation layer that delimits and classifies named-entities.

Every token in the corpus is marked with one of three tags in accordance with the MUC guidelines (Chinchor, 1997):

[5] As by a list of 11,640 infinitive forms of known verbs.

- '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 'B' tags. The tags account for various entity types: person (PER), organization (ORG), location (LOC), work (WRK), event (EVT) and others (MSC).

For example:[6]

```
a/O Associação/B-ORG Portuguesa/I-ORG de/I-ORG
Linguística/I-ORG
o/O Produto/B-MSC Interno/I-MSC Bruto/I-MSC de/O
França/B-LOC
```

This "OBI" tagging scheme delimits and classifies the named-entities in the text while allowing the training of a statistical named-entity recognizer.

In addition to this tagging scheme, which is best suited for a statistical approach to named-entity recognition, a rule-based tool has also been implemented.

This other tool uses XML-like markup tags to identify and classify named-entity expressions but it has the added functionality of assigning a normalized representation to each entity. This can be seen as a shallow information extraction step.

In the following example, the result of identifying and classifying "3.200" and "três mil e duzentos" is shown:

```
<entity type='number' norm='3200'>
3/DGT ./PNT 200/DGT
</entity>
```

and

```
<entity type='number' norm='3200'>
três/CARD mil/CARD e/CJ duzentos/CARD
</entity>
```

Both strings, though very different, are recognized as being a named entity that refers to a number with a normalized representation of '3200'.

4 Final Remarks

In this paper we presented a set of linguistic resources and software tools for the shallow processing of Portuguese.

The main linguistic resource is a 1 million token corpus, accurately hand-tagged with a variety of linguistic data, namely: Every sentence is delimited and every token is

[6] In these examples, all other annotation (POS, inflection, etc.) was removed for the sake of clarity.

circumscribed by blanks. Each token is associated to its morphosyntactic category, by means of POS tags. Each token is also tagged with its inflectional morphology, meaning every inflected token is associated with the corresponding lemma, and 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 from a nominal class. The latter includes also information about their degree, namely superlative for Adjectives, and diminutive for both Adjectives and Nouns. Finally, named-entities are delimited and classified by using the usual "OBI" scheme.

The software consists of shallow processing tools that produce the different layers of annotation mentioned.

Some of these tools can be tested through on-line demos: The on-line demo of the LX-Suite pipeline located at `http://lxsuite.di.fc.ul.pt` currently shows only the processing steps up to, and including, POS tagging. The verbal featurizer/lemmatizer can be tested at `http://lxlemmatizer.di.fc.ul.pt`. Finally, the verbal conjugator can be used as a standalone service/tool by visiting `http://lxconjugator.di.fc.ul.pt`.

Future work will focus on extending the suite with new tools such as a nominal inflector which, given a lemma and a feature value, returns the inflected form of that lemma; and a named-entity recognizer using a statistical approach.

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 Circularity. 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*.

Branco, António and João Ricardo Silva (2005a) Dedicated Nominal Featurization in Portuguese. University of Lisbon, ms.

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

Brants, Thorsten (2000) TnT – A statistical part-of-speech tagger. In *Proceedings of the 3rd Applied Natural Language Processing Conference and the 1st North American Chapter of the Association for Computational Linguistics.* 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.

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