LX-Abbreviations

1. BASIC INFORMATION

1.1 *Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc.)*

LX-Abbreviations resource is a collection of abbreviations of different types from European Portuguese composed by 208 words. Each abbreviation is annotated with grammatical categories, gender and number. Finally, abbreviations (see Branco & Silva, 2003) are grouped into types, as shown below:

LX-Abbreviations		
Types	Foreign abbreviations	4
	Nouns	5
	Units	3
	Possessives	1
	Personals	40
	Week days	7
	Months	12
	Social titles	125
	Parts for addresses	8
	Noun Phrases	3
Total	**10**	**208**

This resource was collected in the context of TagShare – Tagging and Shallow Tools and Resources project[1] with the following main goals: developing of a set of linguistic resources and software component tools to support the computational processing of Portuguese.

1.2 *Representation of the lexicon (flat files, database, markup)*

The corpus is represented in .txt format.

1.3 *Character encoding*

The characters are in UTF8 code.

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

[1] It can be visited at http://tagshare.di.fc.ul.pt/.

Name: António Branco

Address: Departamento de Informática NLX − Grupo de Fala e Linguagem Natural, Faculdade de Ciências da Universidade de Lisboa, Edifício C6, Campo Grande 1749-016 Lisboa

Position: Assistant Professor

Affiliation: Faculty of Sciences, University of Lisbon

Telephone: +351 217 500 087

Fax: +351 217 500 084

E-mail: antonio.branco@di.fc.ul.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be available on the META-SHARE platform.

2.3 Copyright statement and information on IPR

This resource is a free license-based for research and for commercial purposes, with attribution and no redistribution allowed. It will be available on the META-SHARE platform.

3. TECHNICAL INFORMATION

3.1 Directories and files

The archive that can be uploaded on the META-SHARE is a .zip file with two files: one .xml and one .xsd, which contains the .xml specification file.

3.2 Data structure of an entry

In the text file, the data is organized by types of abbreviations and each one of them is subdivided into entries with tags: grammatical categories, and grammatical features (gender and number), as exemplified below, when “WD” stands for “Week Days”, and “fs” for “female singular”; and the correspondent list of abbreviations:

```
<entry>
          <tag>_WD#fs</tag>
          <list>
               <abbrev>seg.</abbrev>
               <abbrev>qua.</abbrev>
               <abbrev>qui.</abbrev>
               <abbrev>sex.</abbrev>
               <abbrev>ter.</abbrev>
          </list>
     </entry>
     <entry>
          <tag>_WD#ms</tag>
          <list>
               <abbrev>sáb.</abbrev>
               <abbrev>dom.</abbrev>
          </list>
     </entry>
```

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

The corpus is composed by 208 words with 3.9 KB compressed (27 KB uncompressed) for disk storage.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the LX-Abbreviations is European Portuguese.

4.2 Entry Type

For this information, please see item 3.2.

4.3 Attributes and their values

There are three values for gender − <m> for male, <f> for female, and <g> for male or female − and other three for number − <s> for singular, <p> for plural, and <n> for singular or plural.

Taking as an example the entry exposed at Section 3.2 <_WD#fs>, the first value <WD> is the grammatical category tag (WD: Week Days) followed <#> by the tags for gender <m> and number <f>.

4.4 Coverage of the lexicon

The LX-Abbreviations lexicon covers the general language.

4.5 Intended application of the lexicon

LX-Abbreviations has been used as part of LX-Tokenizer in all NLP applications developed at NLX-Group, as a base list with string types considered hard cases for tokenization of Portuguese texts, involving the ambivalence between the end of a sentence and the end of an abbreviation (see Branco and Silva, 2003).

4.6 POS assignment

Each type of abbreviation was mannually annotated with proper grammatical category tag, according to the POS-Tagger used at NLX-Group (see Silva, 2007).

4.7 Reliability (automatically/manually constructed)

Mannually constructed (open list), under the standard abbreviations considered in grammars and spelling handbooks for Portuguese.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Nascimento, Filipe Nunes e João Silva, 2006, "Open Resources and Tools for the Shallow Processing of Portuguese: The TagShare Project", Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006).

Silva, João, 2007. Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization. MSc thesis, University of Lisbon. Published as Technical Report DI-FCUL-TR-07-16.