LX-Stopwords

1. BASIC INFORMATION

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

LX-Stopwords resource is a manual list of words from Portuguese composed by 2631 words of 51 types. The words are grouped in three big classes, arranged according to their morpho-syntactic category and inflectional feature value (closed classes, open classes, and multi-word units). This list was created as a support resource to develop CRIVO/*EtiFac* tool (see Branco & Silva, 2001), a tool for the semiautomatic annotation of corpora. With this in mind, the list seeks to be an as exhaustive as possible repository of all word forms that belong to closed classes, items typically with high frequency and fixity.

Taking into account the ambiguity between words of different categories, which means that some words from closed classes (1866 words) can be part of others categories, two classes were added to the list: open classes (592 words) and multi-word units (173 words), including only the words already contained in closed classes.

This wordlist was collected in the context of NeXing – Natural Negation Modeling and Processing[1] project whose the main goal was to contribute for improving the automated mapping between (orthographic) form and (linguistic) meaning, on the one hand, and between (linguistic) meaning and knowledge (representation), on the other hand, in what concerns natural language negation.

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

The corpus is represented in .txt format.

1.2 *Character encoding*

The characters are in UTF-8 code.

2. ADMINISTRATIVE INFORMATION

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

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

[1] It can be visited at http://www.di.fc.ul.pt/~ahb/nexing.htm.

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 purposes and free license-based 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 .txt file, the data are organized by classes of words according to their morpho-syntactic category which are sub-specified by inflectional feature value. As shown in the examples below, each class of words is divided in sub-classes taking into account the grammatical category introduced by the symbol <_> (cf. example A.), followed, when applicable, by <#> features values (gender <f/m/g(both)>, number <s/p/n(both)> (cf. example B.), and person <1/2/3> (cf. example C.):

```
A.

<entries>
            <sub-class>_PREP</sub-class>
                  <list>
                        <stopword>juntamente com</stopword>
                        <stopword>por causa de</stopword>
                        <stopword>até a</stopword>
                        <stopword>mediante</stopword>
                        <stopword>como</stopword>
                        <stopword>enquanto</stopword>
                        <stopword>segundo</stopword>
                        <stopword>quando de</stopword>
                        <stopword>a</stopword>
B.

<entries>
            <sub-class>_WD#fs</sub-class>
                  <list>
                        <stopword>segunda</stopword>
                        <stopword>segunda-feira</stopword>
                        <stopword>terça</stopword>
                        <stopword>terça-feira</stopword>
                        <stopword>quarta</stopword>
                        <stopword>quarta-feira</stopword>
                        <stopword>quinta</stopword>
                        <stopword>quinta-feira</stopword>
                        <stopword>sexta</stopword>
                        <stopword>sexta-feira</stopword>
                  </list>
</entries>
C.

<entries>
            <sub-class>_PRS#gs1</sub-class>
                  <list>
                        <stopword>eu</stopword>
                        <stopword>mim</stopword>
                  </list>
</entries>
```

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

The corpus is composed by 2631 words with 14.3 KB compressed (137.6 KB uncompressed) for disk storage.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of the list is European Portuguese.

4.2 Entry Type

For this information, see, please, item 3.2.

4.3 Attributes and their values

For the sake of completeness, the list of tags below contains all the tags that may occur in this lexicon, including open class tags.

Tagset	
Tag	**Description**
_DA	Definite Article
_UM	Occurrences of "um" or "uma"
_IA	Indefinite Articles (except "um" and "uma", see _UM)
_QNT	Quantifiers
_IND	Indefinites

_DEM	Demonstrative
_POSS	Possessive
_PRS	Personals
_CL	Clitics
_INT	Interrogative
_REL	Relatives
_EXC	Exclamatives
_CJ	Conjunctions
_PREP	Prepositions
_CARD	Cardinals (except "um" and "uma", see _UM)
_MGT	Magnitude classes
_ORD	Ordinals
_DFR	Denominators of fractions
_WD	Week Days
_MTH	Months
_ADV	Adverbs
_UNIT	Measurement Units (when in abbreviated form)
_EOE	End of Enumeration
_STT	Social Title
_EMP	Emphasis
_EL	Extra-linguistic
_DM	Discourse marker
_PL	Para-linguistic
_FRG	Fragment
_ITJ	Interjections
_CN	Common noun
_ADJ	Adjective
_VAUX	Auxiliar Verb "ter" and "haver" preceding _PPT in compound tenses
_INFAUX	Auxiliar Verb (Infinitive)
_GERAUX	Auxiliar Verb (Gerund)
_V	Verb (other than PPA, PPT, INF or GER)
_PPT	Past Participle preceded by aux. verb "ter" or "haver" in compound tenses
_PPA	Other Past Participles
_GER	Gerund
_INF	Infinitive
_NP	Noun Phrase
_PP	Prepositional Phrase
_PNM	Part of Name

_PADR	Part of Address
_LTR	Letters
_DGT	Digits
_DGTR	Roman numerals
_PNT	Punctuation
_SYB	Symbol
_EADR	Electronic Address
_TERMN	Terminations (for optional plu./fem./etc")

4.4 Coverage of the lexicon

The LX-Stopwords list works on the general language.

4.5 Intended application of the lexicon

The wordlist can be used in linguistic research and also in some NLP aplications.

4.6 POS assignment

All words were manually grouped and tagged (morpho-syntactic tagging) according to POS-Tagger tagset (see Silva, 2007).

4.7 Reliability (automatically/manually constructed)

The wordlist and the annotation were mannually constructed once the main and final goal was to construct a completely and accurately tagged resource.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Branco, António and João Silva, 2001. EtiFac: A Facilitating Tool for Manual Tagging. In *Actas do XVII Encontro Anual da Associação Portuguesa de Linguística (APL'02)*, pp. 81-90.

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.