EUROPARL Corpus
Parallel Corpora: Portuguese-English

1 BASIC INFORMATION

1.1 Corpus composition

The EUROPARL Corpus (subpart Portuguese-English of the parallel corpora), available at http://www.statmt.org/europarl/, was extracted from the proceedings of the European Parliament (Koehn, 2005). It contains transcriptions of sessions dating back from 1996 to 2011, in a total of approximately 58,324,562 tokens words of European Portuguese (L1) and 49,216,896 tokens of English (translation).

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

The file formats of this corpus are txt, and a three-column file with one token per line, followed by pos and lemma.

1.3 Character encoding

The characters have been encoded in UTF-8.

2 ADMINISTRATIVE INFORMATION

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

Name: Dr. Amália Mendes
Address: Complexo Interdisciplinar da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal
Affiliation: Centro de Linguística da Universidade de Lisboa
Position: Researcher
Telephone: +351 21 790 47 00
Fax: + 351 21 796 56 22
e-mail: amalia.mendes@clul.ul.pt

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

The resource will be available on the MetaShare platform.

2.3 Copyright statement and information on IPR

The resource is free license-based for research purposes and free license-based for commercial purposes. It is planned to be distributed under a Creative Commons BY-SA license.

3 TECHNICAL INFORMATION

3.1 Directories and files

The EUROPARL Corpus is composed by one text file for the English corpus and two files for the Portuguese version: a text file and an annotated file.

3.2 Data structure of an entry

The text version contains plain text and no further annotation. The Portuguese annotated file is a four-column file with one token per line, followed by PoS tag and lemma.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The corpus contains 58,324,562 tokens of European Portuguese and 49,216,896 tokens of English. It needs about 320 MB and 289 MB for disk storage for the text files in Portuguese and English, respectively, and about 883 MB for the annotated file.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

The corpus is a bilingual parallel corpus (Portuguese-English). The annotated corpus is monolingual (European Portuguese).

4.2 The natural language(s) of the corpus

The language of the corpus is European Portuguese and English.

4. 3 Domain(s)/register(s) of the corpus

This is a political corpus, composed by transcriptions of the Portuguese sessions in the European parliament and their translation into English.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The corpus is automatically annotated with PoS information at token level, including punctuation. It is also lemmatized.

4.4.2 Tags (if POS/MSD/TIME/discourse/etc –tagged or parsed),

The corpus was automatically PoS-tagged with MBT tagger (http://ilk.uvt.nl/mbt/), and lemmatized with MBLEM (http://ilk.uvt.nl/mbma/), following the annotation scheme of the Corpus of Reference of Contemporary Portuguese (Généreux et al., 2012).

4.4.3Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

n.a

4.4.4 Attributes and their values (if annotated)

The following tags were applied in the POS-tagging:

POS codification	
Tag	**Category**
ADJ	Adjectives
ADV	Adverbs
CARD	Cardinals
CJ	Conjunctions
CL	Clitics
CN	Common Nouns
DA	Definite Articles
DEM	Demonstratives
DFR	Denominators of Fractions
DGTR	Roman Numerals
DGT	Digits
DM	Discourse Marker
EADR	Electronic Addresses
EOE	End of Enumeration
EXC	Exclamatives

GER	Gerunds
GERAUX	Gerunds as auxiliary verbs
IA	Indefinite Articles
IND	Indefinites
INF	Infinitive
INFAUX	Infinitive auxiliary verb
INT	Interrogatives
ITJ	Interjection
LTR	Letters
LADV1…LADVn	Latin Multi-Word Adverbs
MGT	Magnitude Classes
MTH	Months
NP	Noun Phrases
ORD	Ordinals
PADR	Part of Address
PNM	Part of Name
PNT	Punctuation Marks
POSS	Possessives
PPA	Past Participles not in compound tenses
PP	Prepositional Phrases
PPT	Past Participle in compound tenses
PREP	Prepositions
PRS	Personals
QNT	Quantifiers

REL	Relatives
STT	Social Titles
SYB	Symbols
TERMN	Optional Terminations
UM	"um" or "uma"
UNIT	Measurement units in abbreviated form
VAUX	Finite "ter" or "haver" in compound tenses
V	Verbs (other than PPA, PPT, INF or GER)
WD	Week Days
LADV1…LADVn	Multi-Word Adverbs
Contracted forms	**Combinations of :**
CL+CL	Two clitics
PREP+ADV	Preposition and Adverb
PREP+DA	Preposition and Definite Articles
PREP+DEM	Preposition and Demonstratives
PREP+IND	Preposition and Indefinite
PREP+INT	Preposition and Interrogative
PREP+PRS	Preposition and Personal pronoun
PREP+QNT	Preposition and Quantifier
PREP+REL	Preposition and Relative
PREP+UM	Preposition and "um" or "uma"

4.5 Intended application of the corpus

The corpus can be used in linguistic research and for improving and developing Natural Language Processing tools and applications, namely training data for statistical machine translation, word sense disambiguation, anaphora resolution, information extraction, etc.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

The POS and lemma annotation was done automatically.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Généreux, M., I. Hendrickx and A. Mendes (2012), "A Large Portuguese Corpus On-Line: Cleaning and Preprocessing", in Caseli, H. et al. (eds.), *Computational Processing of the Portuguese Language. Proceedings of the 10th International Conference PROPOR1012*, Berlin, Heidelberg: Springer-Verlag, pp. 113-120.

Koehn, P. (2005), "EUROPARL: A Parallel Corpus for Statistical Machine Translation", in *Proceedings of the Tenth Machine Translation Summit*, Phuket, Thailand, pp. 79-86.