LT Corpus

1 BASIC INFORMATION

1.1 Corpus composition

The LT Corpus (Literary Corpus) contains approximately 1,781,083 running words of European and Brazilian Portuguese. It includes 70 copyright-free classics (61 Portugal and 9 from Brazil) published before 1940 (Annex 1, below, lists the works that constitute the corpus).

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

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

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 is available through ELRA on DVD.

2.3 Copyright statement and information on IPR

ELRA License.

3 TECHNICAL INFORMATION

3.1 Directories and files

The LT Corpus is composed by a text file (corpus) and an annotated file.

3.2 Data structure of an entry

The txt version has one sentence per line, an identification number for each text and no further annotation. The cqpweb file has one token per line, followed by PoS tag and lemma, and it is annotated for NP chunks and sentences.

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

This corpus contains 1,781,083 tokens (89,679 sentences and 218,432 noun phrases) and needs about 9 MB for disk storage for the text file and about 34 MB for the cqpweb file.

4 CONTENT INFORMATION

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

This corpus is a monolingual annotated corpus.

4.2 The natural language(s) of the corpus

The language of this corpus is European and Brazilian Portuguese.

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

This corpus is totally composed by literary register.

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)

This corpus is PoS-annotated at token level, including punctuation. Noun phrases were recognized and annotated with specific tags.

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). YamCha software (http://chasen.org/~taku/software/yamcha/) was used to recognize chunks that consist of noun phrases and identifies the elements that are in the beginning, in the middle and in the end of a noun phrase.

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"

The following tags were applied in the NP chunker:

Position	**Description**
B-NP	Beginning
I-NP	Inside
E-NP	End
O	Outside

4.5 Intended application of the corpus

This corpus can be used in linguistic research and for improving and developing Natural Language Processing tools and applications.

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

The POS and NP chunks annotation was done automatically.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Bacelar do Nascimento, F., L. A. Pereira and J. Saramago (2000), "Portuguese *Corpora* at CLUL", in *Second International Conference on Language Resources and Evaluation - Proceedings,* Volume III, Athens, pp. 1603-1607.

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 PROPOR2012*, Berlin, Heidelberg: Springer-Verlag, pp. 113-120.

ANNEX 1 – Corpus Constitution (list of works)

ALENCAR, José de (1865) *Iracema: Lenda do Ceará*, Brasil. L0476

ALENCAR, José de (1874) *Ubirajara: Lenda Tupi*, Empreza Literaria Universal, Brasil. L0592

ALENCAR, José de (1880) *O Guarany*, Livraria Garnier Collecção dos Autores Celebres da Litteratura Brasileira, Rio de Janeiro, Brasil. L0092

ALMEIDA, Fialho de (1882) *A Cidade do Vício*, 8ª edição, Clássica Lisboa. L0345

ALMEIDA, Fialho de (1889) *Os Gatos*, Ulisseia. L0460

ALMEIDA, Fialho de (1903) *À Esquina*, 7ª edição, Clássica, Lisboa (1960). L0462

ALMEIDA, Fortunato de (1894) *O Infante de Sagres*, 1ª edição, Livraria Portuense, Porto. L0088

ALMEIDA, Manuel António de (1852) *Memórias de um Sargento de Milícias*, Brasil. L0639

ASSIS, Machado de (1881) *Memorias Posthumas de Braz Cubas*, 1ª edição, Typographia Nacional, Rio de Janeiro, Brasil. L0292

ASSIS, Machado de (1891) *Quincas Borba*, Bertrand Obras-primas da Língua Portuguesa, Amadora, Brasil. L0418

ASSIS, Machado de (1900) *Dom Casmurro*, Livraria Chardron de Lello & Irmão - Editores Biblioteca Iniciação Literária, Brasil. L0635

BARBOSA, Ruy (1933) *Discursos e Conferências*, Companhia Portuguesa Editora, Porto, Brasil. L0248

BRANCO, Camilo Castelo (1854) *A Filha do Arcediago*, 9ª edição, Parceria A. M. Pereira Obras de Camilo Castelo Branco, Lisboa (1971). L0296

BRANCO, Camilo Castelo (1856) *A Neta do Arcediago*, 9ª edição, Parceria A. M. Pereira, Obras de Camilo Castelo Branco, Lisboa (1973). L0224

BRANCO, Camilo Castelo (1858) *Anos de Prosa*, 4ª edição, Parceria A. M. Pereira, Obras de Camilo Castelo Branco, Lisboa (1973). L0082

BRANCO, Camilo Castelo (1858) *O Que Fazem Mulheres*, 8ª edição, Parceria A. M. Pereira, Obras de Camilo Castelo Branco, Lisboa (1967). L0225

BRANCO, Camilo Castelo (1862) *As Três Irmãs*, Parceria A. M. Pereira Obras de Camilo Castelo Branco, Lisboa (1974). L0226

BRANCO, Camilo Castelo (1862) *Memórias do Cárcere*, Portugal. L0822

BRANCO, Camilo Castelo (1863) *Estrelas Propícias*, 6ª edição, Parceria A. M. Pereira, Obras de Camilo Castelo Branco, Lisboa (1965). L0114

BRANCO, Camilo Castelo (1865) *O Esqueleto*, 10ª edição, Parceria A. M. Pereira, Obras de Camilo Castelo Branco, Lisboa (1969). L0083

BRANCO, Camilo Castelo (1866) *O Santo da Montanha*, 6ª edição, Parceria A. M. Pereira, Obras de Camilo Castelo Branco, Lisboa (1972). L0509

BRANCO, Camilo Castelo (1868) *Mistérios de Fafe*, 8ª edição, Parceria A. M. Pereira, Obras de Camilo Castelo Branco, Lisboa (1969). L0227

BRANCO, Camilo Castelo (1869) *Os Brilhantes do Brasileiro*, 9ª edição, Parceria A. M. Pereira, Obras de Camilo Castelo Branco, Lisboa (1972). L0115

BRANCO, Camilo Castelo (1870) *A Mulher Fatal*, 10ª edição, Parceria A. M. Pereira, Lisboa (1968). L0728

BRANCO, Camilo Castelo (1875) *Novelas do Minho*, Ed. Crít. Mateus, M. H. M. Centro de Estudos Filológicos, Biblioteca de Clássicos Portugueses, Lisboa (1961). L0289

BRANCO, Camilo Castelo (1888) *A Doida do Candal*, 11ª edição, Parceria A. M. Pereira, Obras de Camilo Castelo Branco, Lisboa (1971). L0297

BRANCO, Camilo Castelo (1908) *O Condemnado / Como os Anjos se Vingam / Entre a Flauta e a Viola*, Parceria A. M. Pereira, Lisboa. L0081

BRANDÃO, Raúl (1923) *Os Pescadores*, 1ª edição, Livrarias Aillaud e Bertrand, Lisboa. L0009

BRANDÃO, Raúl (1926) *As Ilhas Desconhecidas: Notas e Paisagens*, Portugal. L0622

CAMACHO, Brito (s/d) *Contos Ligeiros*, 1ª edição, Guimarães, Lisboa. L0514

COELHO, Henrique Trindade (1902) *In Illo Tempore*, 5ª edição, Portugália, Lisboa. L0279

CORVO, João de Andrade (1903) *Um Anno na Corte*, 1ª edição, Lisboa. L0773

CUNHA, Euclides da (1909) *À Margem da História*, 5ª edição, Lello & Irmão, Porto (1941), Brasil. L0380

DINIS, Júlio (1867) *As Pupilas do Senhor Reitor*, Livraria Simões Lopes, Porto (1948). L0523

DINIS, Júlio (1868) *A Morgadinha dos Canaviais*, Porto (1935). L0016

DINIS, Júlio (1871) *Os Fidalgos da Casa Mourisca: Crónica da Aldeia*, Figueirinhas A Nossa Colecção, Porto (1953). L0432

GAMA, Arnaldo (1872) *O Balio de Leça*, 3ª edição, Livraria Simões Lopes, Colecção do Romance Português, Porto (1949). L0025

GARRETT, Almeida (1810) *Viagens na Minha Terra*, Livraria Tavares Martins, Porto (1946). L0003

GIL, Augusto (1909) *Luar de Janeiro*, Portugal. L0621

GRAVE, João (1930) *A Dor e a Ternura*, Portugal. L0626

HERCULANO, Alexandre (1847) *Eurico o Presbítero*, 36ª edição (ed. Crít), Vitorino Nemésio, Bertrand, Lisboa (1944). L0209

HERCULANO, Alexandre (1848) *O Monge de Cistér ou a Epocha de D. João I*, 17ª edição, Bertrand, Lisboa (1936). L0506

HERCULANO, Alexandre (1851) *Lendas e Narrativas Tomo I*, 18ª edição, Bertrand, Lisboa. L0215

JUNQUEIRO, Guerra (1885) *A Velhice do Padre Eterno,* Lello & Irmão, Porto (1967). L0229

PESSANHA, Camilo (1920) *Clepsidra*, Portugal. L0623

PESSOA, Fernando (1942) *Poesias*, Portugal. L0575

PESSOA, Fernando (1944) *Poemas de Álvaro de Campos*, Imprensa Nacional - Casa da Moeda, Lisboa (1992). L0078

PESSOA, Fernando (1945) *Odes de Ricardo Reis IV*, Ática, Lisboa (1946). L0516

PESSOA, Fernando (1994) *Mensagem*, Portugal. L0576

PIMENTEL, Alberto (1873) *O Anel Misterioso*, Figueirinhas, Porto (1945). L0211

QUEIROZ, Eça (1901) *A Cidade e as Serras*, 18ª edição, Lello & Irmão, Porto (1941). L0017

QUEIROZ, Eça de (1878) *O Primo Basílio: Episódio Doméstico*, Livros do Brasil, Obras de Eça de Queirós, Lisboa. L0443

QUEIROZ, Eça de (1887) *A Relíquia*, Livros do Brasil, Obras de Eça de Queirós, Lisboa. L0298

QUEIROZ, Eça de (1888) *Os Maias*, Livros do Brasil, Obras de Eça de Queirós, Lisboa. L0379

QUEIROZ, Eça de (1902) *Contos VIII*, Lello & Irmão, Obras de Eça de Queirós/edição do centenário, Porto. (1947) L0119

QUEIROZ, Eça de (1912) *Últimas Páginas*, s.e., Porto. L0299

QUEIROZ, Eça de (1925) *A Capital*, 10ª edição, Lello & Irmão, Porto (1978). L0508

QUEIROZ, Eça de (1925) *Alves & C.ª e Outras Ficções*, Livros do Brasil, Obras de Eça de Queirós, Lisboa. L0346

QUEIROZ, Eça de (1925) *Correspondência I*, 1ª edição completa, Biblioteca de Autores Portugueses, (1925; 1940; 1945; 1949; 1961 (publicações parciais)). L0290

QUEIROZ, Eça de (1925) *Correspondência II*, 1ª edição completa, Biblioteca de Autores Portugueses (1925; 1940; 1945; 1949; 1961 (publicações parciais)). L0288

QUENTAL, Antero de (1871) *Causas da Decadência dos Povos Peninsulares*, Portugal. L0823

QUENTAL, Antero de (1886) *Sonetos*, Sá da Costa, Lisboa. L0080

SABUGOSA, Conde de (1923) Bagos de História, 1ª edição, Portugália, Lisboa. L0125

SABUGOSA, Conde de (1923) Bôbos na Corte, 1ª edição, Portugália, Lisboa. L0124

SÁ-CARNEIRO, Mário de (1912) *Princípio – Diários*, Portugal. L0730

SÁ-CARNEIRO, Mário de (1912) *Princípio – Loucura*, Portugal. L0729

SÁ-CARNEIRO, Mário de (1912) *Princípio – O Incesto* Porto (1985). L0731

SÁ-CARNEIRO, Mário de (1914) *A Confissão de Lúcio,* 6ª edição, Ática, Obras Completas de Mário de Sá-Carneiro, Lisboa (1982). L0707

SÁ-CARNEIRO, Mário de (1946) *Poesias*, Portugal. L0580

TEIXEIRA-GOMES, M. (1899) *Inventário de Junho*, 3ª edição, Seara Nova, Lisboa (1933). L0348