ELRA-Distribution Agency (ELDA)
55-57 rue Brillat Savarin,
F-75013 PARIS, FRANCE
Tel. +33 1 43 13 33 33
Fax: +33 1 43 13 33 30
Email: choukri@elda.org
Web : http://www.elra.info or http://www.elda.org

The CINTIL Corpus – International Corpus of Portuguese

Tagset information

Tagset

The final tagset includes the following types of tags:

- **Major POS**: definite article (/DA), common noun (/CN), Verb (/V), etc. Ex.:
- **Distinguished verb forms**: infinitive (/INF), gerund (/GER), past participle in compound tenses (/PPT), and other participles (/PPA).
- **Auxiliary verbs**: auxiliary verb (/VAUX), infinitive auxiliary verb (/INFAUX), etc.
- **Speech specific elements**: discourse marker (/DM), extra-linguistic elements (/EL), fragment (/FRAG), etc. Ex.: hhhm/EL
- **MWU**: adverbials (/LADV*n*), prepositions (/LPREP*n*), etc.: each POS tag prefixed with 'L' is extended with information regarding the position (*n*) of the corresponding token inside the multi-word expression. Ex.:
  a/LPREP1 respeito/LPREP2 de/LPREP3
- **Inflectional feature values**: gender — feminine (f), masculine (m) or underspecified (g) —, number — singular (s), plural (p) or neutral (n) —, person — first (1), second (2) or third (3) —, superlative (-sup), mood — indicative (i), subjunctive (c) —, etc. Exs.:
  interessantíssimas/ADJ#fp-sup
  falou/V#ppi-3s
- **Lemmas:** the lemmas for tokens from nominal and verbal open classes. Exs.:
  interessantíssimas/INTERESSANTE/ADJ#fp-sup
  falou/FALAR/V#ppi-3s
- **Components of NE expressions**: denominators of fractions (/DFR), Part of Address (/PADR), Social Title (/STT), etc.: these tags identify major components of expressions for Named Entities.
- **NE expressions**: inside the NE expression (\I), outside (\O) and beginning (\B): these tags mark the boundaries of Named Entity expressions. Exs.:
  ...o/DA#ms/O Ministério/PNM/B de/PREP/I a/DA#fs/I Ciência/PNM/I decidiu/DECIDIR/V#ppi-3s/O...

Part-of-speech tags

Tag	Category	Examples
ADJ	Adjectives	bom, brilhante, eficaz, …
ADV	Adverbs	hoje, já, sim, felizmente, …
CARD	Cardinals	zero, dez, cem, mil, …
CJ	Conjunctions	e, ou, tal como, …
CL	Clitics	o, lhe, se, …
CN	Common Nouns	computador, cidade, ideia, …
DA	Definite Articles	o, os, …
DEM	Demonstratives	este, esses, aquele, …
DFR	Denominators of Fractions	meio, terço, décimo, %, …

DGTR	Roman Numerals	VI, LX, MMIII, MCMXCIX, …
DGT	Digits	0, 1, 42, 12345, 67890, …
DM	Discourse Marker	olá, …
EADR	Electronic Addresses	http://www.di.fc.ul.pt, …
EOE	End of Enumeration	etc
EXC	Exclamation	ah, ei, …
GER	Gerunds	sendo, afirmando, vivendo, …
GERAUX	Gerund "ter"/"haver" in compound tenses	tendo, havendo
IA	Indefinite Articles	uns, umas, …
IND	Indefinites	tudo, alguém, ninguém, …
INF	Infinitive	ser, afirmar, viver, …
INFAUX	Infinitive "ter"/"haver" in compound tenses	ter, haver, …
INT	Interrogatives	quem, como, quando, …
ITJ	Interjection	bolas, caramba, …
LTR	Letters	a, b, c, …
MGT	Magnitude Classes	unidade, dezena, dúzia, resma, …
MTH	Months	Janeiro, Dezembro, …
NP	Noun Phrases	idem, …
ORD	Ordinals	primeiro, centésimo, penúltimo, …
PADR	Part of Address	Rua, av., rot., …
PNM	Part of Name	Lisboa, António, João, …
PNT	Punctuation Marks	., ?, (, …
POSS	Possessives	meu, teu, seu, …
PPA	Past Participles not in compound tenses	sido, afirmados, vivida, …
PP	Prepositional Phrases	algures, …
PPT	Past Participle in compound tenses	sido, afirmado, vivido, …
PREP	Prepositions	de, para, em redor de, …
PRS	Personals	eu, tu, ele, …
QNT	Quantifiers	todos, muitos, nenhum, …
REL	Relatives	que, cujo, tal que, …
STT	Social Titles	Presidente, drª., prof., …
SYB	Symbols	@, #, &, …
TERMN	Optional Terminations	(s), (as), …
UM	"um" or "uma"	um, uma
UNIT	Abbreviated Measurement Unit	kg., km., …
VAUX	Finite "ter" or "haver" in compound tenses	temos, haveriam, …
V	Verbs (other than PPA, PPT, INF or GER)	falou, falaria, …
WD	Week Days	segunda, terça-feira, sábado, …
Tags for multi-word expressions		
LADV1…LADVn	Multi-Word Adverbs	de facto, em suma, um pouco, …
LCJ1…LCJn	Multi-Word Conjunctions	assim como, já que, …
LDEM1…LDEMn	Multi-Word Demonstratives	o mesmo, …
LDFR1…LDFRn	Multi-Word Denominators of Fractions	por cento
LDM1…LDMn	Multi-Word Discourse Markers	pois não, até logo, …
LITJ1…LITJn	Multi-Word Interjections	meu Deus
LPRS1…LPRSn	Multi-Word Personals	a gente, si mesmo, V. Exa., …
LPREP1…LPREPn	Multi-Word Prepositions	através de, a partir de, …
LQD1…LQDn	Multi-Word Quantifiers	uns quantos, …
LREL1…LRELn	Multi-Word Relatives	tal como, …
Tags specific to the spoken corpus		
EMP	Emphasis	
EL	Extra-linguistic	
DM	Discourse Marker	
PL	Para-linguistic	
FRG	Fragment	

Inflection tags

Tag	Description
Tags for nominal categories	
m	Masculine
f	Feminine
s	Singular
p	Plural
dim	Diminutive
sup	Superlative
comp	Comparative
Tags for verbs	
1	First Person
2	Second Person
3	Third Person
pi	Presente do Indicativo
ppi	Pretérito Perfeito do Indicativo
ii	Pretérito Imperfeito do Indicativo
mpi	Pretérito Mais que Perfeito do Indicativo
fi	Futuro do Indicativo
c	Condicional
pc	Presente do Conjuntivo
ic	Pretérito Imperfeito do Conjuntivo
fc	Futuro do Conjuntivo
imp	Imperativo

Named entity tags

Tag	Description	Type		Example
B	Beginning	PER	person	...o[O] *João[B-PER] Silva[I-PER]* disse[O]...
I	Inside	ORG	organization	...a[O] *Universidade[B-ORG] de[I-ORG] Lisboa[I-ORG]* comprou[O]...
		LOC	location	...de[O] *Londres[B-LOC]* a[O] *Paris[B-LOC]*...
		WRK	work	...a[O] *Mona[B-WRK] Lisa[I-WRK]* está[O]...
		MSC	other cases	...o[O] *RMS[B-MSC] Titanic[I-MSC]* afundou[O]...
O	Outside			

The annotation manual will be made available together with the corpus.

Package:

```
./CompanionTools.txt
./Corpus/FullVersion/MultipleFiles/CINTIL-1-NEWS.txt
./Corpus/FullVersion/MultipleFiles/CINTIL-2-OTHER.txt
./Corpus/FullVersion/MultipleFiles/CINTIL-SPOKEN.txt
./Corpus/FullVersion/MultipleFiles/FICTION-1-Clube.txt
./Corpus/FullVersion/MultipleFiles/FICTION-2-Detective.txt
./Corpus/FullVersion/MultipleFiles/FICTION-3-Lucio.txt
./Corpus/FullVersion/MultipleFiles/FICTION-4-Fradique.txt
./Corpus/FullVersion/MultipleFiles/FICTION-5-Carcere.txt
./Corpus/FullVersion/MultipleFiles/FICTION-6-Viagens.txt
./Corpus/FullVersion/MultipleFiles/FICTION-7-Eurico.txt
./Corpus/FullVersion/MultipleFiles/checksum.md5
./Corpus/FullVersion/TwoFiles/CINTIL-SPOKEN.txt
./Corpus/FullVersion/TwoFiles/CINTIL-WRITTEN.txt
./Corpus/FullVersion/TwoFiles/checksum.md5
./Corpus/VersionWoutIOB/MultipleFiles/CINTIL-1-NEWS.txt
./Corpus/VersionWoutIOB/MultipleFiles/CINTIL-2-OTHER.txt
./Corpus/VersionWoutIOB/MultipleFiles/CINTIL-SPOKEN.txt
./Corpus/VersionWoutIOB/MultipleFiles/FICTION-1-Clube.txt
```

```
./Corpus/VersionWoutIOB/MultipleFiles/FICTION-2-Detective.txt
./Corpus/VersionWoutIOB/MultipleFiles/FICTION-3-Lucio.txt
./Corpus/VersionWoutIOB/MultipleFiles/FICTION-4-Fradique.txt
./Corpus/VersionWoutIOB/MultipleFiles/FICTION-5-Carcere.txt
./Corpus/VersionWoutIOB/MultipleFiles/FICTION-6-Viagens.txt
./Corpus/VersionWoutIOB/MultipleFiles/FICTION-7-Eurico.txt
./Corpus/VersionWoutIOB/MultipleFiles/checksum.md5
./Corpus/VersionWoutIOB/TwoFiles/CINTIL-SPOKEN.txt
./Corpus/VersionWoutIOB/TwoFiles/CINTIL-WRITTEN.txt
./Corpus/VersionWoutIOB/TwoFiles/checksum.md5
./Corpus/cintil-spoken.dtd
./Corpus/cintil-written.dtd
./Description.txt
./ExcerptIndex.txt
./README
./CINTILManualAnotacaoV6.2_en.doc
```