CINTIL

Corpus Internacional do Português

Annotation Manual

Version 6.0

24/06/2005

Florbela Barreto, António Horta Branco, Amália Mendes,
Fernanda Bacelar Nascimento e João Silva.

Universidade de Lisboa

1 File format

Character set ISO-8859-1
DOS line breaks
Format .txt

2 Tokenization

Sentences and paragraphs

Sentences are divided by a line separator (= each sentence in one line).
Paragraphs are divided by two line separators (= one line between paragraphs).

Lexemes

Tokens are separated by a blank space:

```
um exemplo > um exemplo
```

Contractions are expanded, and the first token is concatenated with the symbol ('_'):

```
'da' -> 'de_ a'
consigo -> com_ si
pela ->  por_ a
```

Punctuation and symbols are marked with blank spaces:
(blank space on the left: '*'; blank space on the right: '*/')

```
5.3 > 5 . 3
1. 2 > 1 .*/ 2
8 . 6 > 8 \*.*/ 6
```

Clitics are separated from the verbal form, but keeping the hyphen. Vocalic change is marked with '#', and with '-CL-' in mesoclitic position:

```
deu-se-lho -> deu -se -lho
vê-las -> vê# -las
afirmá-lo-ia -> afirmá#-CL-ia -lo
```

Preposition 'de' is separated from the verb 'haver' (keeping the hyphen)

```
há-de -> há -de
```

Alternative word endings are separated from the word by a blank space

```
Caro(a) amigo(a) > Caro (a) amigo (a)
```

3 Prosodic annotation

The spoken subcorpus includes 2 symbols for prosodic annotation, which follow the CHAT/CHILDES annotation model.

The symbol “/”, preceded and followed by blank space, marks a non-terminal prosodic break.

The symbol “//”, preceded and followed by blank space, marks a terminal prosodic break (end of utterance).

The symbol “?” marks a terminal prosodic break in interrogative utterances.

4 Part-of-speech (POS) annotation

Symbol '/' to the right of the lemma, or immediately to the right of the token when there is no lemma. The POS tag follows '/':

```
com/PREP
viu/VER/V
```

In multiword expressions, the tag starts with the letter 'L'. The tag is repeated for each token of the expression and is followed by a number indicating the relative position of the token in the expression:

```
de_LCJ1 maneira_LCJ2 a_LCJ3 que_LCJ4
```

Other tokens only receive one POS tag.

Tagset

Tags are acronyms of the English name of the POS categories.

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 篬 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	Exclamatives	que, quanto, ...
GER	Gerunds	sendo, afirmando, vivendo, …
GERAUX	Gerunds as auxiliary verbs	tendo, havendo
IA	Indefinite Articles	uns, umas, …
IND	Indefinites	tudo, alguém 魬 ninguém 魬 …
INF	Infinitive	ser, afirmar, viver, …
INFAUX	Infinitive auxiliary verb	ter, havermos, ...
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 䢵 …
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, dr., prof., …
SYB	Symbols	@, #, &, …
TERMN	Optional Terminations	(s), (as), …
UM	"um" or "uma"	um, uma
UNIT	Measurement units in abbreviated form	Kg, h, seg, Hz, Mbytes,...
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, …
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 逢 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, …
Specific of transcriptions		
EL	Extra-linguistic	
EMP	Emphasis	
FRG	Fragment	
PL	Para-linguistic	

Past Participle

/PPT in compound tenses, with auxiliary verbs 'ter' and 'haver'.
/PPA in other contexts.

More detailed distinctions are also established at the lemmatization level (See below).

Occurrences of 'um' and 'uma'

Annotated with tag /UM.

Que

Occurrences of *que*: tagged with /REL in relatives, /INT in interrogatives, /EXC in exclamatives, and /CJ in the remaining cases, i.e. adverbials, clefts, embedded, comparatives and consecutives.

Exclamatives

/EXC for cases of pronouns starting exclamation sentences:

```
Que fadiga!
Quantas etiquetas ainda para atribuir!
```

Auxiliary Verbs

/VAUX Auxiliary verbs in compound tenses (occurrences of 'ter' or 'haver' followed by past participle).

Proper names

/PNM in anthroponyms, toponyms, titles of artistic works (literary works, songs, paintings, etc.), institutions, addresses, acronyms, siglas.

In cases of multiword proper names, only the words from open classes are tagged with /PNM:

```
Prof./STT Borges/PNM de/PREP Castro/PNM
Ministério/PNM de_/PREP a/DA Educação/PNM
Avenida/PADR de_/PREP a/DA Liberdade/PNM
```

5 Featurizer

Nominal features

Symbol '#' following the POS category, and features following '#':

```
gatos/GATO/CN#mp
```

Morphological gender and number (semantic ones are ignored).

Nominal foreign words are annotated with features.

Masculine: `m`; feminine: `f`.

Singular: `s`; plural: `p`.

First Person: `1`; second: `2`; and third: `3`:

```
ela/PRS#fs3
```

Diminutives in the sequences *-inho*, *-zinho*, *-ito* e *-zito* are tagged with `-dim`.

```
mesinha/MESA/CN#fs-dim
```

Superlatives (regular in *-íssimo* or irregular) are tagged with `-sup`.

```
normalíssimo/NORMAL/ADJ#ms-sup
o/ART#ms maior/GRANDE/ADJ#ms-sup
```

Comparatives (irregular) are tagged with `-comp`.

```
é/SER/V#pi-3s maior/GRANDE/ADJ#ms-comp
```

Open classes with nominal features of gender and number:

/CN	: Common noun
/ADJ	: Adjective
/PPA	: Other Past Participles

Open classes with nominal features of number and person:

/VAUX	: Auxiliary Verbs
/V	: Verbs (other than PPA, PPT, INF or GER)
/INF	: Infinitive

Closed classes with nominal features of gender, number and person:

/PRS	: Personals
/CL	: Clitics

Closed classes with nominal features of gender and number:

/DA : Definite Article
/UM : occurrences of "um" or "uma"
/IA : Indefinite Articles (except "um" and "uma", vd. /UM)
/QNT : Quantifiers
/IND : Indefinites
/DEM : Demonstrative
/POSS : Possessive
/INT : Interrogative (except *que, quem, quê* and *quão*)
/REL : Relatives (except *que, quem,*e *quê*)
/EXC : Exclamatives (except *que* and *quê*)
/CARD : Cardinals (except "um" and "uma", vd. /UM)
/MGT : Magnitude classes
/ORD : Ordinals
/DFR : Denominators of fractions (except symbol %)
/WD : Week Days
/MTH : Months
/UNIT : Measurement units (in abbreviated form)
/STT : Social Title
/LTR : Letter

Verbal Features

Symbol '#' following the POS category.
Features of tense and mood after '#', followed by '-'.
Features of person and number after '-'.

```
andarias/ANDAR/V#c-2s
```

Open classes with verbal features:
/VAUX : Auxiliary Verbs
/V : Verb (other than PPA, PPT, INF or GER)

Tempo/Modo	Etiqueta
Presente do Indicativo	`pi`
Pretérito Perfeito do Indicativo	`ppi`
Pretérito Imperfeito do Indicativo	`ii`
Pretérito Mais que Perfeito do Indicativo	`mpi`
Futuro do Indicativo	`fi`
Condicional	`c`
Presente do Conjuntivo	`pc`
Pretérito Imperfeito do Conjuntivo	`ic`
Futuro do Conjuntivo	`fc`
Imperativo	`imp`

Infinitives

/INF#ninf	non inflected
/INF#...	inflected
/INF#ndef	undetermined

6 Lemmatization

Symbol '/' following the token.
Lemma between '/' and '/'.
Lemma in capital letters:

```
gatos/GATO
```

Only one lemma for each token, except for tokens tagged as /PPA.
Tokens tagged as /PPA: infinitive form, masculine singular form:

```
cavada/CAVAR,CAVADO/PPA
```

Categories whose elements are lemmatized:
/CN, /ADJ, /V, /VAUX, /GER, /INF, /PPT and /PPA

Nominal lemmas

Open classes with nominal lemmas:
/CN, /ADJ and /PPA:

General case
The lemma is the masculine singular form, if it exists.
If not, it is the masculine (plural) form, if it exists.
If not, the feminine singular form, if it exists.
If not, the form itself.

Words with prefixes
They keep the prefix in the lemma.

Words with sufixes
In certain cases, the lemma is reduced to the radical: diminutives -*inho*. -*zinho*, -*ito*, and -*zito*; superlatives, either regular (ending in -*íssimo*) as irregular, and comparatives (irregular).

"Irregular" feminine forms
The lemma is the irregular form (e.g. actriz, etc)

Multiple orthographic forms
The lemma is the one occurring in the Rebelo's Vocabulary, if the word has several orthographic forms.
If it does not exist in Rebelo, lemmatization follows the rule defined in Rebelo.
If not, the lemma is the most frequent orthographic form.

Abreviations

The lemma of abreviations (of the categories /CN, /ADJ and /PPA) is not abbreviated.

Foreign words

The lemma of foreign words is the occurring form itself.

Verbal lemmas

Open classes with verbal lemmas:
/V, /VAUX, /GER, /INF, /PPT and /PPA.

General case

The lemma is the non inflected infinitive form.

Words with prefixes

They keep the prefix in the lemma.

Multiple orthographic forms

The lemma is the one occurring in the Rebelo's Vocabulary, if the word has several orthographic forms.
If it does not exist in Rebelo, lemmatization follows the rule defined in Rebelo.
If not, the lemma is the most frequent orthographic form.

Pronouns

Lemma has the same features of person, number and gender as the occurring form (there is no lemmatization for first person).

Remaining cases

No lemma is attributed to the remaining cases.

(for technical reasons, in the online concordancer, the lemma field of the remaining classes is filled with the form itself)

7 Named entities

Tags

The following tags are attributed:

B	to the beginning of the expression
I	for the other tokens of the expression
O	for tokens who do not belong to named entities

Tags B and I are further detailed as

-PER	when it is the name of a person
-ORG	of an organization
-LOC	of a place
-WRK	of a work (books, movies, paintings, etc)
-MSC	remaining cases

Format

'[' after the feature tags, if they occur
Tag after '['
']' after the tag

```
encontrei/ENCONTRAR/V#ppi-1s[O] o/DA#ms[O]
Pres./PRESIDENTE/STT#ms[O] Jorge/PNM[B-PER]
Sampaio/PNM[I-PER]
```

Criteria

Those of the manual of named entities:
http://www.itl.nist.gov/iaui/894.02/related/projects/muc/proceedings/ne/task.html