CINTIL
Corpus Internacional do Português

Convenções de Etiquetação
Versão 6.0
24/06/2005

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

Universidade de Lisboa

1 Formato de ficheiros

Character set ISO-8859-1
Quebras de linha DOS
Formato .txt

2 Segmentação

De frases e parágrafos

Frases separadas por quebra de linha (= cada frase numa linha).
Parágrafos separados por duas quebras de linha (= uma linha entre parágrafos).

De lexemas

Tokens separados por espaço em branco:

```
um exemplo > um exemplo
```

Contracções são expandidas, ficando uma marca ('_') no primeiro token da expansão:

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

Marcação do espaçamento em torno da pontuação e de símbolos:
(marca de espaço à esquerda: '*'; marca de espaço à direita: '*/')

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

Clíticos separados do verbo (mantêm o hífen). Marcação de alternância vocálica com '#', com '-CL-' no posição de mesóclise:

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

Separação da preposição 'de' do verbo 'haver' (mantêm o hífen)

```
há-de -> há -de
```

Separação de finais especiais que marcam terminações alternativas

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

3 Anotação prosódica

O subcorpus oral inclui 2 símbolos para a anotação prosódica, seguindo o modelo de anotação dos projectos CHAT/CHILDES.

O símbolo "/", precedido e seguido de espaço, assinala uma quebra prosódica não final.

O símbolo "//", precedido e seguido de espaço, assinala uma quebra prosódica final.

O símbolo "?" assinala uma quebra prosódica final em contextos interrogativos.

4 Etiquetação morfo-sintáctica

Símbolo '/' à direita do lema, ou imediatamente à direita do token quando não há lema. Etiqueta à direita de '/':

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

Em locuções (lexemas multi-palavra), etiqueta prefixada de 'L'. Em cada token da locução, a etiqueta repete-se seguida de numeração:

```
de_LCJ1 maneira_LCJ2 a_LCJ3 que_LCJ4
```

Outros tokens recebem uma única etiqueta categorial.

Conjunto de etiquetas

Etiquetas são acrónimos das designações em inglês das categorias.

Etiqueta	Categoria	Exemplos
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 篕 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 蘩 …
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 蘲 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	

Particípios

/PPT em tempos compostos, com auxiliares 'ter' e 'haver'.
/PPA nas restantes ocorrências.
Distinções mais finas são também feitas ao nível da lematização (vd. abaixo).

Ocorrências de 'um' e 'uma'

Etiquetados com /UM.

Que

Ocorrências de *que*: Etiquetadas com /REL nas relativas, /INT nas interrogativas, /EXC nas exclamativas, e /CJ nos restantes casos, i.e. adverbiais, clivadas, encaixadas, comparativas e consecutivas

Exclamativos

/EXC para pronomes que encabeçam exclamativas:

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

Verbos auxiliares

/VAUX Verbos auxiliares em tempos compostos (ocorrências de 'ter' ou 'haver' a preceder particípio passado).

Nomes próprios

/PNM em antropónimos, topónimos, títulos de obras (obras literárias, canções, pinturas, etc.), instituições, endereços, acrónimos, siglas.

Para nomes próprios multi-palavra, /PNM apenas em palavras de classes abertas:

```
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 Traçamento

Traços nominais

Símbolo '#' à direita de categoria, com traços à direita de '#':
`gatos/GATO/CN#mp`

Género e número morfológicos (semânticos são ignorados).

Estrangeirismos nominais são traçados.

Masculino: `m`; feminino: `f`.

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

Primeira pessoa: `1`; segunda: `2`; e terceira: `3`:
`ela/PRS#fs3`

Diminutivos em *-inho*, *-zinho*, *-ito* e *-zito* etiquetados com `-dim`.
`mesinha/MESA/CN#fs-dim`

Superlativos (regulares em *-íssimo* ou irregulares) etiquetados com `-sup`.
`normalíssimo/NORMAL/ADJ#ms-sup`
`o/ART#ms maior/GRANDE/ADJ#ms-sup`

Comparativos (irregulares) etiquetados com `-comp`.
`é/SER/V#pi-3s maior/GRANDE/ADJ#ms-comp`

Classes abertas com traços nominais de género e número:
/CN : Common noun
/ADJ : Adjective
/PPA : Other Past Participles

Classes abertas com traços nominais de número e pessoa:
/VAUX : Auxiliar Verbs
/V : Verbs (other than PPA, PPT, INF or GER)
/INF : Infinitive

Classes fechadas com traços nominais de género, número e pessoa:
/PRS : Personals
/CL : Clitics

Classes fechadas com traços nominais de género e número:

/DA : Definite Article
/UM : ocorrências de "um" ou "uma"
/IA : Indefinite Articles (excepto "um" e "uma", vd. /UM)
/QNT : Quantifiers
/IND : Indefinites
/DEM : Demonstrative
/POSS : Possessive
/INT : Interrogative (excepto *que, quem, quê* e *quão*)
/REL : Relatives (excepto *que, quem,*e *quê*)
/EXC : Exclamatives (excepto *que* e *quê*)
/CARD : Cardinals (excepto "um" e "uma", vd. /UM)
/MGT : Magnitude classes
/ORD : Ordinals
/DFR : Denominators of fractions (excepto símbolo %)
/WD : Week Days
/MTH : Months
/UNIT : Measurement units (em forma abreviada)
/STT : Social Title
/LTR : Letter

Traços verbais

Símbolo '#' à direita de categoria.
Traços de tempo e modo à direita de '#', seguidos de '-'.
Traços de pessoa e número a seguir a '-'.

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

Classes abertas com traços verbais:
/VAUX : Auxiliar 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

Infinitivos

/INF#ninf	não flexionado
/INF#...	flexionado
/INF#ndef	casos indeterminados

6 Lematização

Símbolo '/' à direita do token.
Lema entre '/' e '/'.
Lema em maiúsculas:

```
gatos/GATO
```

Apenas um lema para cada token, excepto para tokens /PPA.
Token PPA: forma no infinitivo, forma no masculino singular:

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

Classes cujos elementos recebem lema:
/CN, /ADJ, /V, /VAUX, /GER, /INF, /PPT e /PPA

Lemas nominais

Classes abertas com lemas nominais:
/CN, /ADJ e /PPA:

Caso geral

Lema é a forma masculina singular, se existir.
Se não, a forma masculina (plural), se existir
Se não, a forma feminina singular, se existir
Se não, a própria forma.

Palavras prefixadas

Mantêm o prefixo no lema.

Palavras sufixadas

Redução para radical apenas nos diminutivos *-inho*. *-zinho*, *-ito*, e *-zito*; e nos superlativos, tanto regulares (em *-íssimo*) como irregulares, e nos comparativos (irregulares).

Femininos "irregulares"

Mantém-se a forma irregular como lema (e.g. actriz, etc)

Múltiplas grafias

O lema no vocabulário do Rebelo, se a ocorrência admitir múltiplas grafias.
Se não exisitir no Rebelo, segue-se a regra definida pelo Rebelo.
Se não, opta-se pelo lema da grafia mais frequente.

Abreviaturas
Lema de uma abreviatura (das classes /CN, /ADJ e /PPA) não é abreviado.

Estrangeirismos
Lema de um estrangeirismo é a ocorrência.

Lemas verbais

Classes abertas com lemas verbais:
/V, /VAUX, /GER, /INF, /PPT e /PPA.

Caso geral
Lema é forma do infinitivo não flexionado.

Palavras prefixadas
Mantém o prefixo no lema.

Mtiplas grafias
Lema no vocabulário do Rebelo, se a ocorrência admitir múltiplas grafias.
Se não existir no Rebelo, segue-se a regra definida pelo Rebelo.
Se não, opta-se pelo lema da grafia mais frequente.

Pronomes

Lema na pessoa, número e género da ocorrência (não há lematização para primeira pessoa).

Casos restantes

Não se atribui lemas aos casos restantes.

(por razões de implementação, no concordanciador online, relativamente às restantes classes, o campo do lema é preenchido com a própria forma)

7 Nomes próprios multi-palavra

Etiquetas

Atribui-se etiqueta:

B	para início da expressão
I	para restantes tokens da expressão
O	para tokens que não pertencem a expressões

Etiquetas B e I:são continuadas por

-PER	caso se trate de designação de pessoa
-ORG	de organização
-LOC	de local
-WRK	de obra (livros, filmes, quadros, etc)
-MSC	restantes casos

Formato

'[' após traços de flexão se existirem
etiqueta após '['
']' após etiqueta

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

Critérios

Os do manual das MUC:
http://www.itl.nist.gov/iaui/894.02/related/projects/muc/proceedings/ne/task.html