CINTIL Corpus

1 BASIC INFORMATION

1.1 Corpus composition

CINTIL-Corpus Internacional do Português is a linguistically interpreted corpus of Portuguese. At present it is composed of 1 Million annotated tokens, verified by human expert annotators. The annotation comprises information on part-of-speech, open classes lemma and inflection, multi-word expressions pertaining to the class of adverbs and to the closed POS classes, and multi-word proper names (for named entity recognition). The corpus has been developed at the University of Lisbon by the NLX group at the Faculty of Sciences and the Anagrama group at the Cenro de Linguística da Universidade de Lisboa.

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

The corpus consists of 2 files with linguistic annotation (pos, inflection, lemma, named entities). The annotation has been manually revised.

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: Prof. António Branco
Address: Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Alameda da Universidade
1649-003 Lisboa - Portugal
Affiliation: Faculdade de Ciências da Universidade de Lisboa
Position: Professor
Telephone: +351 21 7500606
Fax: + 351 21
e-mail: Antonio.Branco@di.fc.ul.pt

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 as a set of 2 files annotated with linguistic information. The annotation manual is made available with the corpus.

2.3 Copyright statement and information on IPR

ELRA licence.

3 TECHNICAL INFORMATION

3.1 Directories and files

CINTIL is divided into 2 text files, one for written texts and another for transcriptions of spoken register.

3.2 Data structure of an entry

Each file is structured in one header per text, with sentences, paragraphs and excerpts mark-up, in XML format. The text and its annotation is presented in a 4 column format: token, lemma (for lexical categories), pos tag + inflection tag, named entities tag.

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

The corpus contains 1 million tokens and needs about 20MB for disk storage.

4 CONTENT INFORMATION

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

This corpus is a monolingual, annotated corpus, covering both written and spoken registers.

4.2 The natural language(s) of the corpus

The language of the corpus is European Portuguese.

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

The corpus contains a written (composed by 689 124 tokens) and a spoken subpart (composed by 502 622 tokens). The written subpart includes fiction, newspapers and technical texts, from 1990 till 2003, plus 3 fiction texts from the 19th century. The spoken subpart are transcriptions of recordings of both formal and informal registers, from 1970 to 2002. These spoken transcriptions cover very diversified situations: conversations, dialogues, phone conversations, radio and television programs, homilies, among others.

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)

Each written text is described in a header in xml format with information on the type of data, the author's name, title, place and date of edition. The header of each transcription refers to the type of data, title, description of participants, date and place of recording, situation, register, length, number of tokens, acoustic quality, transcriber and reviser's name. The corpus contains markup in xml at the following levels: paragraph, sentence (equivalent to speech turn in the spoken subpart) and excerpt.

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

The corpus was automatically tagged with LX-Tagger (Barreto et al.,2006). The POS named entities annotation were verified by two human annotators.

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

Does not apply.

4.4.4 Attributes and their values (if annotated)

Part-of-speech tags:

Tag	Category	Examples
ADJ	Adjectives	bom, brilhante, eficaz, …
ADV	Adverbs	hoje, já, sim, felizmente, algo…
CARD	Cardinals	zero, dez, cem, mil, …
CJ	Conjunctions	e, ou, mas, porque, pois…
CL	Clitics	o, lhe, se, …
CN	Common Nouns	computador, cidade, ideia, …
DA	Definite Articles	o, os, a, as.

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	Pronto, enfim, pá…
EADR	Electronic Addresses	http://www.di.fc.ul.pt, …
EL	Extra-linguistic	Riso, tosse: ah ah ah, ih ih ih, …
EMP	Emphasis	cá, lá, aí, …
EOE	End of Enumeration	etc.
EXC	Exclamatives	que, quanto, ...
GER	Gerunds	sendo, afirmando, vivendo, …
VAUXGER	Gerunds as auxiliary verbs	tendo, havendo
IA	Indefinite Articles	uns, umas, …
IND	Indefinites	tudo, alguém, ninguém, …
INF	Infinitive	ser, afirmar, viver, …
VAUXINF	Infinitive auxiliary verb	ter, havermos, ...
INT	Interrogatives	quem, como, quando, …
ITJ	Interjection	bolas, caramba, olá, fogo, alto, …
LTR	Letters	a, b, c, …
MGT	Magnitude Classes	unidade, dezena, dúzia, resma, …

MTH	Months	Janeiro, Dezembro, …
ORD	Ordinals	primeiro, centésimo, penúltimo, …
PADR	Part of Address	Rua, av., rot., …
PL	Para-linguistic	Onomatopeias: có-rócó-có, lá lá lá, pum pum pum, …
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, …
PPT	Past Participle in compound tenses	sido, afirmado, vivido, …
PREP	Prepositions	de, para, desde, em, …
PRS	Personals	eu, tu, ele, …
QNT	Quantifiers	todos, muitos, nenhum, …
REL	Relatives	que, cujo, quem, …
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	quando muito, de todo, por aí adiante, além disso, aos poucos, às tantas, por certo, além do mais, …
LCJ1…LCJn	Multi-Word Conjunctions	assim como, já que, apesar disso, tanto assim que, além de que, daí que, de cada vez que, tanto mais quanto, tanto que, por muito que, …
LDEM1…LDEMn	Multi-Word Demonstratives	o mesmo, o próprio, …
LDFR1…LDFRn	Multi-Word Denominators of Fractions	por cento
LDM1…LDMn	Multi-Word Discourse Markers	pois não□ pois claro, sim senhor, não sei quê, não sei quantos, quer dizer, a bem dizer, ora bem, tal e coisa, ou coisa assim, …
LITJ1…LITJn	Multi-Word Interjections	meu Deus, ora essa, com a breca, qual quê, até logo, alto lá, ó diabo, ….
LPRS1…LPRSn	Multi-Word Personals	a gente, si mesmo, V. Exa., …
LPREP1…LPREPn	Multi-Word Prepositions	para além de, a partir de, apesar de, ao pé de, por trás de, …
LQD1…LQDn	Multi-Word Quantifiers	uns quantos, …
LREL1…LRELn	Multi-Word Relatives	tal como, …
Specific of transcriptions		
FRAG	Fragment	&discut, &assi, &doen, …

Inflection tags:

Tag	Description
Tags for nominal features	
m	Masculine
f	Feminine
s	Singular
p	Plural
dim	Diminutive
sup	Superlative
comp	Comparative
1	First Person
2	Second Person
3	Third Person
Tags for verb features	
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
impaf	Imperativo (affirmative verb forms of the second person)
impneg	Imperativo (negative verb forms of the second person)
imp	Imperativo (subjunctive verb forms that have a imperative value)
Tags for infinitive verbs	
INF#ninf	non-inflected infinitives

INF#... (where '…' should be replaced by the person and number of each verb form)	Inflected infinitives
INF#ndef	Undetermined infinitives

Named-entity tags

Position	description	Semantic type	description	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			

4.5 Intended application of the corpus

The corpus can be used in linguistic research and for improving and developing numerous kinds of Natural Language Processing tools and applications.

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

The POS-tagging was done automatically in a first phase and then manually revised by two annotators.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Barreto, F., A. Branco, E. Ferreira, A. Mendes, M. F. Nascimento, F. Nunes, and J. Silva (2006), "Open Resources and Tools for the Shallow Processing of Portuguese". In 5th International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy.

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Fernanda Bacelar Nascimento, Filipe Nunes and João Silva, 2006, "Linguistic Resources and Software for Shallow Processing", In *Actas do XXI Encontro Anual da Associação Portuguesa de Linguística*, Lisbon, Portugal.

Branco, António and João Silva, 2006, "LX-Suite: Shallow Processing Tools for Portuguese", *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (EACL2006), Trento, Italy, pp.179-182.