CINTIL-PropBank

I. Basic Information

1.1. Corpus information

The CINTIL-PropBank (Branco *et al.*, 2012) is a set of sentences annotated with their constituency structure and semantic role tags, composed of 10,632 sentences and 132,260 tokens taken from different sources and domains: news (9,454 sentences; 123,524 tokens), and novels (399 sentences; 3,082 tokens) (cf. 3.2). In addition, there are 779 sentences (5,654 tokens) used for regression testing of the computational grammar that supported the annotation of the corpus (cf. Section 4.6).
For the creation of this PropBank we adopted a semi-automatic analysis with a double-blind annotation followed by adjudication. The resulting dataset contains three information levels: phrase constituency, grammatical functions, and phrase semantic roles.
The main motivation behind the creation of this resource was to build a high quality data set with semantic information that could support the development of automatic semantic role labelers for Portuguese.
The development of this resource started under the project SemanticShare – Resources and Tools for Semantic Processing (see more at: http://nlx.di.fc.ul.pt/projects.html), whose main goal was to generate a deep linguistic annotated corpus of Portuguese, with manually verified grammatical representations.

The following table displays a breakdown of the CINTIL-PropBank corpus:

CINTIL-PropBank				
Sub-corpus	**id**	**Sentences**	**Tokens**	**Domain**
Sentences for regression testing	aTSTS	779	5,654	Test
CINTIL-International Corpus of Portuguese[1]	bCINT	1,219	13,516	News
	cCINT	399	3,082	Novels
CETEMPúblico	eCTMP	8,134	108,996	News
Penn TreeBank (translation)	dPENN	101	1,012	News
Total		10,632	132,260	

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

The corpus is a single file in a XML format.

1.*3. Character encoding*

The characters are in UTF8 code.

1 CINTIL-International Corpus of Portuguese was the first corpus, but only partly, to being used to integrate our corpus of syntactic trees of constituencies and, thus, give it the name.

II. Administrative Information

2.1. Contact person

Name: António Branco
Address: Departamento de Informática NLX - Grupo de Fala e Linguagem Natural, Faculdade de Ciências da Universidade de Lisboa, Edifício C6, Campo Grande 1749-016 Lisboa
Position: Assistant professor
Affiliation: Faculty of Sciences, University of Lisbon
Telephone: +351 217 500 087
Fax: +351 217 500 084
E-mail: antonio.branco@di.fc.ul.pt

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

This resource is available through META-SHARE.

2.3. Copyright statement and information on IPR

This resource is available for both research and commercial purposes, with attribution, and no redistribution nor derivatives allowed.

III. Technical Information

3.*1. Directories and files*

The archive that can be uploaded on the META-SHARE is a ZIP file with two files: one XML and one XSD, which contains the XML specification file.

3.2. Data structure of an entry

For the XML file with the set of sentences, the data is organized with one sentence per entry (`<sentence>`). Each entry contains the sentence identifier (`<id>`), formed by the sub-corpus id (cf. the Table in Section 1.1) concatenated with a sentence number; the sentence in raw text (`<raw>`); and the CoNLL2005 (specification available at: http://www.lsi.upc.edu/%7Esrlconll/spec.html) or parenthesis format tree (`<conll>`), as shown in the example below:

```
<sentence>
  <id>a11</id>
  <tokenized>A criança obedece apenas a_ a mãe .</tokenized>
  <conll>
    A           *    DA          (NP*          (S(S*         -     -                *
    criança     *    CN             *)             *         -     -          (SJ-ARG1*
    obedece     *    V           (VP*              *         -     obedecer          *)))
    apenas      *    ADV            *              *         -     -             (M-M*
    a_          *    PREP           *              *         -     -          (IO-ARG2*
    a           *    DA          (NP*              *         -     -                *
    mãe         *    CN             *))            *)        -     -                *
    .           *    PNT            *              *)        -     -                *
  </conll>
</sentence>
```

3.3. *Corpus size (nmb. of tokens, NB occupied in disk)*

The corpus is composed by 10,632 sentences, which take up 1.2 MB compressed (13.8 MB uncompressed) for disk storage.

IV. Content Information

4.*1. Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*

This is a monolingual and a semi-automatic annotated corpus.

4.*2. The natural language(s) of the corpus*

The language of the corpus is Portuguese with pre-spelling reform of 1990[2].

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

Concerning the register of the texts represented in the corpus, it comprises excerpts of news articles from daily and general newspapers (9,454 sentences; 123,524 tokens), literary language from excerpts of novels (399 sentences; 3,082 tokens), and, additionally, test sentences used for regression testing of the grammar used in the annotation process (779 sentences; 5,654 tokens).

4.4. Annotation in the corpus (if an annotated corpus)

4.4.1. Types of annotation (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The annotation of the corpus consists of three levels of linguistic information: phrase constituency, grammatical functions, and phrase semantic roles.

4.4.2. Tags (if POS/WSD/TIME/discourse/etc – tagged or parsed)

For details on the linguistic options underlying the analyses, see (Branco *et al.*, 2011). In this Section we present the tag sets used at the various levels of annotation:

Phrasal and part-of-speech tags

Tag	Meaning
A	Adjective
AP	Adjective Phrase
ADV	Adverb
ADVP	Adverb Phrase
C	Complementizer
CP	Complementizer Phrase
CARD	Cardinal
CONJ	Conjuction
CONJP	Conjuction Phrase

2 This means that the orthography rules used are those that are described by the Orthography Reform of 1945. The orthographic agreement of 1990 was adopted just in may of 2009 and is being implemented until 2012.

D	Determiner
DEM	Demonstrative
N	Noun
NP	Noun Phrase
P	Preposition
PP	Preposition Phrase
POSS	Possessive
QNT	Predeterminer
S	Sentence
V	Verb
VP	Verb Phrase

Grammatical Function Tagset

Tag	Meaning
C	Complement
DO	Direct Object
IO	Indirect Object
M	Modifier
N	Relationship between words and named entities
OBL	Oblique Complement
PRD	Predicate
SJ	Subject
SP	Specifier

Semantic Role Tagset

Tag	Meaning
ARG1	First Argument
ARG2	Second Argument
ARG3	Third Argument
ARG11	Argument 1 of subordinating predicator and Argument 1 in the subordinate clause (semantic function of Subjects of so called Subject Control predicators)
ARG21	Argument 2 of subordinating predicator and Argument 1 in the subordinate clause (semantic function of Subjects of so called Direct Object Control predicators)
ARG1cp	Argument 1 in complex predicate constructions
ARG2cp	Argument 2 in complex predicate constructions
ARG2ac	Argument 2 of anticausative readings
ADV	Adverbial
CAU	Cause
DIR	Direction

EXT	Extension
LOC	Localization
MNR	Mode
PNC	Objective
POV	Viewpoint
TMP	Time
PRED	Secondary predication
NULL	Null

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

Not applicable.

4.4.4. Attributes and their values (if annotated)

Not applicable.

4.5. Intended application of the corpus

The corpus can be used in linguistic research and in the development of dependency parsers and semantic role labeling tools.

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

In order to achieve a gold-standard corpus with high accuracy, the CINTIL-PropBank is created by a two-phase process, where an automatic annotation is then manually revised by language experts with post-graduate degrees in Linguistics. More specifically, in the first stage, a deep computational grammar (see Branco and Costa, 2008) is used to generate all the possible parses for a given sentence (the parse forest). This is followed by a manual disambiguation stage where the correct parse is chosen from among those in the parse forest. This second stage follows a double-blind annotation method, where two annotators work independently and, for those cases where their decisions differ, a third annotator (the adjudicator) is brought in to make the final decision. For this corpus, the level of inter-annotator agreement (ITA) is 0.83 in terms of the specific inter-annotator metric developed for this kind of corpora and annotation (Castro, 2011).

The automatic annotation assigns only argumental semantic roles, leaving modifiers with a underspecified ‘M’. These tags are manually specified, again following the same annotation method as before (double-blind annotation with adjudication). For this task, the ITA is 0.76.

V. Relevant References and Other Information

Branco, A. Carvalheiro, C., Pereira, S., Avelãs, M., Pinto, C., Silveira, S., Costa, F., Silva, J., Castro, S., and Graça, J., 2012, “A PropBank for Portuguese: the CINTIL-PropBank”. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).*

Branco, A., Silva, J., Costa, F., and Castro, S., 2011, “CINTIL-TreeBank Handbook: Design options for the representation of syntactic constituency”. Technical Report 2011;02, University of Lisbon, Department of Informatics.

Branco, A. and Costa, F., 2008, “A computational grammar for deep linguistic processing of portuguese: LXGram, version A.4.1”. Technical Report DI-FCUL-TR-08-17. University of Lisbon, Department of Informatics, 2008.

Castro, Sérgio, 2011, *Developing Reliability Metrics and Validation Tools for datasets with deep linguistic Information*, MA Dissertation, Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática.