CINTIL-DeepBank 1.3

I. Basic Information

1.1. Corpus information

CINTIL-DeepBank (Branco *et al.*, 2010) is a corpus of Portuguese texts annotated with deep grammatical information. This document refers to version 1.3 of the corpus, delivered in September of 2015, which adds over 2,000 annotated sentences to the previous version from March 2015. The current version is composed by 17,030 sentences (166,933 tokens) taken from two different sources and domains: news (15,851 sentences; 159,525 tokens), novels (399 sentences; 2,547 tokens). In addition, there are 780 sentences (4,861 tokens) that are used for regression testing of the computational grammar that supported the annotation of the corpus (cf. section 4.6.).

CINTIL-DeepBank includes several levels of information for each sentence, including its derivation tree originated during parsing, its syntactic constituency tree, different renderings of MRS based representations of its meaning (Copestake, 2006), and its fully-fledged grammatical representation in AVM format. This is the result of a semi-automatic annotation process by means of automatic analysis by the grammar followed by a double-blind annotation followed by adjudication (see (Branco and Costa, 2008), for a full description of the process).

The main motivation behind the creation of this resource was to build a high quality data set with rich grammatical information that could support the development of a large set of high level language resources and processing tools for Portuguese.

The development of this resource started under the project SemanticShare – Resources and Tools for Semantic Processing (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, was continued in the project METANET4U-Enhancing the Linguistic Infrastructure of Europe, and in the project QTLeap-Quality Translation by Deep Language Engineering Approaches.

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

CINTIL-DeepBank 1.2				
Sub-corpus	**id**	**Sentences**	**Tokens**	**Domain**
Sentences for regression testing	aTSTS	780	4,861	Test
CINTIL-International Corpus of Portuguese	bCINT	1,220	11,618	News
	cCINT	399	2,547	Novels
Penn TreeBank (translation)	dPENN	105	942	News
CETEMPúblico	eCTMP	14,526	146,965	News
Total		17,030	166,933	

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

The corpus is stored in an archive composed by 3,409 folders. Each folder contains several files, one per sentence. These are plain text files, compressed with gzip.

1.3. Character encoding

The files are encoded in UTF-8.

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: Associate 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 required, and no redistribution nor derivatives allowed. It is available through META-SHARE.

III. Technical Information

3.1. Directories and files

The archive that can be downloaded on the META-SHARE site is a gzip file with 3,409 folders. Each file contains one gzip file per sentence.

3.2. Data structure of an entry

There is a file for each sentence. The file starts with a line at the top with the sentence id (between square brackets), followed by the sentence between quote marks in raw text. Under this there are a variety of analysis of the sentence, separated by a blank line, as illustrated by the example below:

```
[11] (1 of 1) {1} `a criança obedece apenas a a mãe.' []
```

Derivation:

```
(469 ROOT 4.95827e+17 0 7
 (468 SUBJECT-HEAD 3.84742e+17 0 7
  (461 FUNCTOR-HEAD-HCOMPS-SCOPAL -2.2851e+16 0 2
   (65 SG-NOMINAL 4.1552e+15 0 1
    (63 FEM-NOMINAL 4.1552e+15 0 1
     (8 O_DEFINITE-ARTICLE 2.0776e+15 0 1 ("a" 0 1))))
   (145 SG-NOMINAL 0 1 2
    (140 FEM-NOMINAL 0 1 2 (15 CRIANÇA 0 1 2 ("criança" 1 2)))))
  (358 HEAD-COMP_NOTCLITIC 2.34842e+17 2 7
   (98 3SG-VERB 0 2 3
    (97 PRES-IND-VERB 0 2 3 (16 OBEDECER 0 2 3 ("obedece" 2 3))))
   (357 FUNCTOR-HEAD-HCOMPS-SCOPAL 4.5623e+16 3 7
    (17 APENAS_NP-ADJUNCT 3.9016e+15 3 4 ("apenas" 3 4))
    (356 HEAD-COMP_NOTCLITIC 3.76979e+16 4 7
     (29 A_NONPREDICATIONAL-NP_OR_VP-PREPOSITION 1.03477e+16 4 5 ("a" 4 5))
     (355 FUNCTOR-HEAD-HCOMPS-SCOPAL 6.65476e+15 5 7
      (66 SG-NOMINAL 4.1552e+15 5 6
       (64 FEM-NOMINAL 4.1552e+15 5 6
        (38 O_DEFINITE-ARTICLE 2.0776e+15 5 6 ("a" 5 6))))
      (47 SG-NOMINAL 2.95058e+16 6 7
       (46 FEM-NOMINAL 2.95058e+16 6 7
        (45 MÃE_1_NOUN 0 6 7 ("mãe." 6 7))))))))))
```

Syntactic constituency tree:

```
(CP
 (S (NP-SJ-ARG1 (ART-SP (ART-SP (ART-SP (a)))) (N (N (N (criança)))))
  (VP (V (V (V (obedece))))
   (PP-IO-ARG2 (ADV-M-M (apenas))
    (PP (P (a)) (NP-C (ART-SP (ART-SP (ART-SP (a)))) (N (N (N (mãe.))))))))))
```

AVM: Due to its large size, this representation is left out of this document. You may find it in the sample document that is provided in the META-SHARE site.

MRS:

```
 [ LTOP: h1
   INDEX: e2 [ e ELLIPTICAL-PUNCT: BOOL SF: PROPOSITION-OR-QUESTION E.TENSE:
PRESENTE E.ASPECT.PERF: - E.MOOD: INDICATIVO ]
   RELS: <
          [ _o_q_rel
            LBL: h3
            ARG0: x6 [ x GENDER: FEMININE NUMBER: SINGULAR PERSON: 3RD ]
            RSTR: h4 [ h SCOPE: NARROW ]
            BODY: h5 [ h SCOPE: NARROW ] ]
          [ "_criança_n_rel"
            LBL: h7
            ARG0: x6 ]
          [ "_obedecer_v_-a-_rel"
            LBL: h8
            ARG0: e2
            ARG1: x6
            ARG2: x9 [ x PERSON: 3RD NUMBER: SINGULAR GENDER: FEMININE ] ]
          [ "_apenas_q_rel"
            LBL: h10 [ h SCOPE: SCOPE ]
            ARG0: e12
            ARG1: h11 [ h SCOPE: SCOPE ] ]
          [ _o_q_rel
            LBL: h11
            ARG0: x9
            RSTR: h13 [ h SCOPE: NARROW ]
            BODY: h14 [ h SCOPE: NARROW ] ]
          [ "_mãe_n_1-de-_rel"
            LBL: h15
            ARG0: x9
            ARG1: y16 ] >
   HCONS: < h1 qeq h8 h4 qeq h7 h13 qeq h15 > ]
```

Indexed MRS:

```
<h1,e2:BOOL:PROPOSITION-OR-QUESTION:PRESENTE:-:INDICATIVO,
{h3:_o_q(x6:FEMININE:SINGULAR:3RD, h4:NARROW, h5:NARROW),
h7:_criança_n(x6),
h8:_obedecer_v_-a-(e2, x6, x9:3RD:SINGULAR:FEMININE),
h10:_apenas_q(:SCOPEe12, h11:SCOPE),
h11:_o_q(x9, h13:NARROW, h14:NARROW),
h15:_mãe_n_1-de-(x9, y16)},
{h1 qeq h8,
h4 qeq h7,
h13 qeq h15}>
```

Prolog MRS:

```
psoa(h1,e2,[rel('_o_q',h3,
[attrval('ARG0',x6),attrval('RSTR',h4),attrval('BODY',h5)]),rel('_criança_n',h7,
[attrval('ARG0',x6)]),rel('_obedecer_v_-a-',h8,
[attrval('ARG0',e2),attrval('ARG1',x6),attrval('ARG2',x9)]),rel('_apenas_q',h10,
[attrval('ARG0',e12),attrval('ARG1',h11)]),rel('_o_q',h11,
[attrval('ARG0',x9),attrval('RSTR',h13),attrval('BODY',h14)]),rel('_mãe_n_1-
de-',h15,
[attrval('ARG0',x9),attrval('ARG1',y16)])],hcons([qeq(h1,h8),qeq(h4,h7),qeq(h13,
h15)]))
```

RMRS (Robust MRS):

```
 h1
 _o_q(h3,x6:)
 _criança_n(h7,x6:)
 _obedecer_v_-a-(h8,e2:)
 _apenas_q(h10,e12:)
 _o_q(h11,x9:)
 _mãe_n_1-de-(h15,x9:)
 RSTR(h3,h4:)
 BODY(h3,h5:)
 ARG1(h8,x6:)
 ARG2(h8,x9:)
 ARG1(h10,h11:)
 RSTR(h11,h13:)
 BODY(h11,h14:)
 ARG1(h15,u16:)
 qeq(h1:,h8)
 qeq(h4:NARROW:,h7)
 qeq(h13:NARROW:,h15)
```

XML MRS:

```
<rmrs cfrom='-1' cto='-1'a criança obedece apenas a a mãe.'11 @ 0 @ '>
<label vid='1'/>
<ep cfrom='-1' cto='-1'><realpred lemma='o' pos='q'/><label vid='3'/><var
sort='x' vid='6'/></ep>
<ep cfrom='-1' cto='-1'><realpred lemma='criança' pos='n'/><label vid='7'/><var
sort='x' vid='6'/></ep>
<ep cfrom='-1' cto='-1'><realpred lemma='obedecer' pos='v' sense='-a-'/><label
vid='8'/><var sort='e' vid='2'/></ep>
<ep cfrom='-1' cto='-1'><realpred lemma='apenas' pos='q'/><label vid='10'/><var
sort='e' vid='12'/></ep>
<ep cfrom='-1' cto='-1'><realpred lemma='o' pos='q'/><label vid='11'/><var
sort='x' vid='9'/></ep>
<ep cfrom='-1' cto='-1'><realpred lemma='mãe' pos='n' sense='1-de-'/><label
vid='15'/><var sort='x' vid='9'/></ep>
<rarg><rargname>RSTR</rargname><label vid='3'/><var sort='h' vid='4'/></rarg>
<rarg><rargname>BODY</rargname><label vid='3'/><var sort='h' vid='5'/></rarg>
<rarg><rargname>ARG1</rargname><label vid='8'/><var sort='x' vid='6'/></rarg>
<rarg><rargname>ARG2</rargname><label vid='8'/><var sort='x' vid='9'/></rarg>
<rarg><rargname>ARG1</rargname><label vid='10'/><var sort='h' vid='11'/></rarg>
<rarg><rargname>RSTR</rargname><label vid='11'/><var sort='h' vid='13'/></rarg>
```

```
<rarg><rargname>BODY</rargname><label vid='11'/><var sort='h' vid='14'/></rarg>
<rarg><rargname>ARG1</rargname><label vid='15'/><var sort='u' vid='16'/></rarg>
<hcons hreln='qeq'><hi><var sort='h' vid='1'/></hi><lo><label
vid='8'/></lo></hcons>
<hcons hreln='qeq'><hi><var sort='h' vid='4' SCOPE='NARROW'/></hi><lo><label
vid='7'/></lo></hcons>
<hcons hreln='qeq'><hi><var sort='h' vid='13' SCOPE='NARROW'/></hi><lo><label
vid='15'/></lo></hcons>
</rmrs>
```

Elementary dependencies:

```
{e2:
 x6:_o_q[]
 e2:_obedecer_v_-a-[ARG1 x6:_criança_n, ARG2 x9:_mãe_n_1-de-]
 e12:_apenas_q[ARG1 x9:_o_q]
 x9:_o_q[]
}
```

Discriminants:

```
{
  _o_q ARG0 _criança_n
  _obedecer_v_-a- ARG1 _criança_n
  _obedecer_v_-a- ARG2 _mãe_n_1-de-
  _apenas_q ARG1 _o_q
  _o_q ARG0 _mãe_n_1-de-
  _criança_n GENDER feminine
  _criança_n NUMBER singular
  _criança_n PERSON 3rd
  _obedecer_v_-a- ELLIPTICAL-PUNCT bool
  _obedecer_v_-a- SF proposition-or-question
  _obedecer_v_-a- E.TENSE presente
  _obedecer_v_-a- E.ASPECT.PERF -
  _obedecer_v_-a- E.MOOD indicativo
  _mãe_n_1-de- PERSON 3rd
  _mãe_n_1-de- NUMBER singular
  _mãe_n_1-de- GENDER feminine
  _apenas_q _criança_n
  _apenas_q _mãe_n_1-de-
  _apenas_q _o_q
  _apenas_q _obedecer_v_-a-
  _criança_n _mãe_n_1-de-
  _criança_n _o_q
  _criança_n _obedecer_v_-a-
  _mãe_n_1-de- _o_q
  _mãe_n_1-de- _obedecer_v_-a-
  _o_q _obedecer_v_-a-
}
```

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

The corpus is composed by 17,030 sentences with 473 MB compressed (5,322MB uncompressed).

IV. Content Information

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

This is a monolingual annotated corpus.

4.2. The natural language(s) of the corpus

The language of the corpus is Portuguese in the orthographic norm pre-dating the orthographic norm of 1990[1].

4.3. Domain(s)/register(s) of the corpus

The corpus comprises excerpt from news from daily and general newspapers (15,851 sentences), literary language from novels (399 sentences) and additional, 780 sentences from test set (cf. section 1.1.).

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)

Deep grammatical representations.

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

Not applicable.

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 and testing of language processing tools.

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

CINTIL-DeepBank is developed along a semi-automatic process, where an automatic annotation output by the grammar is manually revised by language experts with post-graduate degrees in Linguistics. In the first stage, a deep computational grammar (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 is performed along the double-blind annotation method followed by adjudication: two annotators work independently and, for those cases where their decisions differ, a third annotator (the adjudicator) makes 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).

V. Relevant References and Other Information

Branco, António, Francisco Costa, João Silva, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto, and João Graça, 2010. “Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: The CINTIL DeepGramBank”. In Proceedings of the Seventh International Conference on Language Resources and evaluation (LREC'10) May 19-21, Valetta, Malta pp. 1810-1815.

Branco, António and Francisco Costa, 2008, “A computational grammar for deep linguistic

[1] 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 in May of 2009.

processing of portuguese: LXGram”. In Technical Reports Series. University of Lisbon, Department of Informatics, 2008.

Castro, Sérgio, 2011, Developing Reliability Metrics and Validation Tools for datasets with deep linguistic Information, MA Dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.

Copestake, Ann, 2006, “Minimal Recursion Semantics: An Introduction”. In Research on Language and Computation, 3.4, pp. 281-332.