LogicalFormBankPT

I. Basic Information

1.1. Corpus information

The LogicalFormBankPT (Branco, 2009, and Branco et al., 2011) is a corpus of semantic dependencies of translated texts composed of 3,406 sentences and 44,598 tokens taken from the Wall Street Journal.

The LogicalFormBankPT is composed of MRS representations of each sentence's semantic relations resulting from a previous semi-automatic analysis with a double-blind annotation followed by adjudication (see Branco and Costa, 2008, with a full description of the process). The resulting dataset contains one information level: semantic relations.

The main motivation behind the creation of this resource was to build a high quality data set with syntactic information that could support the development of a large set of automatic resources and tools for Portuguese for NLP studies.

The development of this resource started under the METANET4U project (at: http://metanet4u.eu/) whose main goal is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.

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

The corpus is an archive composed of 913 folders. Each one of them contains a variable number of .gz files in a text format. Each file represents one sentence.

1.*3. Character encoding*

The characters are in UTF8 code.

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. It will be available on the META-SHARE.

III. Technical Information

3.*1. Directories and files*

The archive that can be uploaded on the Meta-Share is a .zip file with 913 folders and each folder has one file per sentence.

3.2. Data structure of an entry

Each text file with the one single sentence. The entry contains one line at the top with the sentence id (between square brackets), followed by the sentence between plicae) in raw text. Under of which are the six representations of MRS structure (see Copestake, 2006), divided by a white line, as shown in the example below:

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

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',h1
0,[attrval('ARG0',e12),attrval('ARG1',h11)]),rel('_o_q',h11,[attrval('ARG0',x9),att
rval('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(h
13,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[]
}
```

3.3. *Corpus size (nmb. of tokens, NB occupied in disk)*
The corpus is composed by 3,406 sentences with 89.5 *MB* compressed (92.6 *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[1].

4.*3. Domain(s)/register(s) of the corpus*
The corpus is exclusively composed of news.

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)
Sentences with semantic relations.

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)

---

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 just in may of 2009 and is being implemented until 2012.

It does not apply.

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, on the other hand, to development of components to add semantics to parsers.

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

Since the LogicalFormBankPT was an alternate representation of TreeBankPT, the reliability is similar. In order to achieve a gold-standard corpus with high accuracy, the TreeBankPT 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 0.868 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, A., 2009, "LogicalFormBanks, the Next Generation of Semantically Annotated Corpora: key issues in construction methodology", In *Proceedings of International Joint Conference on Intelligent Information Systems*, Kraków, Poland, June, pp. 15-18.

Branco, A. and Costa, F., 2008, “A computational grammar for deep linguistic 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, A., 2006, “Minimal Recursion Semantics: An Introduction”. In *Research on Language and Computation*, 3.4, pp. 281-332.