DepBankPT

I. Basic Information

1.1. Corpus information

The DepBankPT (Branco *et al.*, 2011a) is a corpus of grammatical dependencies of the translated news composed of 3,406 sentences and 44,598 tokens taken from the Wall Street Journal.
The DepBankPT is aligned to a constituency bank, the TreeBankPT (see Branco *et al.*, 2011b). The key bridging elements are the grammatical function tags decoring the nodes, in the treebank, and the arcs, in the dependencybank (see `http://lxcenter.di.fc.ul.pt/services/en/LXServicesSearcher.html`). This means that the DepBankPT was extended from the PropBankPT so that besides the tags for the different dependency relations, the arcs are further decorated with tags indicating the semantic relation at stake.
The main motivation behind the creation of this resource was to build a high quality data set with dependency 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 a single file in a xml format.

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 one .xml and one .xsd files.

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. Each entry contains the sentence id (concatenated with sub-corpus/sentence number), tokenized sentence and CoNLL format, as shown in the example below:

```
<sentence>
<id>aTSTS-001/11</id>
<raw>A criança obedece apenas à mãe.</raw>
<conll>
1	A	_	DA	DA	fs	2	SP	2	SP
2	criança	CRIANÇA	CN	CN	fs	3	SJ	3	SJ
3	obedece	OBEDECER	V	V	pi-3s	0	ROOT	0	ROOT
4	apenas	_	ADV	ADV	_	5	M	5	M
5	a_	_	PREP	PREP	_	3	IO	3	IO
6	a	_	DA	DA	fs	7	SP	7	SP
7	mãe	MÃE	CN	CN	fs	5	C	5	C
8	.	_	PNT	PNT	_	3	PUNCT	3	PUNCT
</conll>
</sentence>
```

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

The corpus is composed by 3,406 sentences with 495.5 MB compressed (2.3 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)

Trees with grammatical dependencies.

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)

It does not apply.

4.4.4. Attributes and their values (if annotated)

This is the tag set used:

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
D	Determiner
DEM	Demonstrative
N	Noun
NP	Noun Phrase

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.

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

4.5. Intended application of the corpus

The corpus can be used in linguistic research and, on the other hand, to development of semantic role labelers.

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

Once DepBankPT was directly extracted from the PropBankPT, the reliability is similar. In order to achieve a gold-standard corpus with high accuracy, the PropBankPT 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., Castro, S., Silva, J., and Costa, F. (2011a). “CINTIL-DepBank Handbook: Design options for the representation of grammatical dependencies”. In *Technical Reports Series*, University of Lisbon, Department of Informatics.

Branco, A., Silva, J., Costa, F., and Castro, S. (2011b). “CINTIL-TreeBank Handbook: Design options for the representation of syntactic constituency”. In *Technical Reports Series*, University of Lisbon, Department of Informatics.