

TreeBankPT

I. Basic Information

1.1. Corpus information

The TreeBankPT (Branco et al., 2011) is a corpus of syntactic constituency trees of the translated news composed of 3,406 sentences and 44,598 tokens taken from the Wall Street Journal.

For the creation of this TreeBank we adopted a semi-automatic analysis with a double-blind annotation followed by adjudication. The resulting dataset contains one information level: phrase constituency.

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 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 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. Each entry contains the sentence id (concatenated with sub-corpus/sentence number), sentence in raw text, and s-expressions or parenthesis format tree, as shown in the example below:

```
<sentence>
  <id>aTSTS-001/11</id>
  <raw>A criança obedece apenas à mãe.</raw>
  <tree>(S (S (NP (ART A) (N criança)) (VP (V obedece) (PP (ADV
apenas) (PP (P a_) (NP (ART a) (N mãe)))))) (PNT .))</tree>
</sentence>
```

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

The corpus is composed by 3,406 sentences with 256.1 KB compressed (1.2 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 1990¹.

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 syntactic constituency.

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.

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.

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	Conjunction
CONJP	Conjunction Phrase
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

4.5. Intended application of the corpus

The corpus can be used in linguistic research and, on the other hand, to development of constituency parsers.

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

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

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, Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática.