CINTIL-QATreeBank

I. Basic Information

1.1. Corpus information

CINTIL-QATreebank is a treebank composed of Portuguese sentences that can be used to support the development of Question Answering systems. This Treebank includes 111 declarative sentences from the pre-existing CINTIL-Treebank (see Branco et al. 2011) whose syntactic structure was manually transformed into their non-declarative counterpart: interrogative and imperative clauses.

The non-declarative sentences are annotated with several layers of linguistic information, namely (i) trees with information on constituency and grammatical function; (ii) sentence type; (iii) interrogative pronoun; (iv) question type; and (v) semantic type of expected answer. Moreover, these non-declarative sentences are paired with their declarative counterparts and associated with the expected answer snippets.

CINTIL-QATreebank composition according to the number of clause and question types:

Clause Types	Question Types	Answer Types	Qty.
Total Interrogatives	Affirmative	Yes/No	8
	Negative	Yes/No	4
	Alternative	Yes/No	1
Partial Interrogatives	Que_N (Which)	Factual	17
	Quem (Who)	Factual	8
	Onde (Where)	Factual	6
	Qual (Which)	Factual	7
	Quanto (How much/many)	Factual	9
	Quando (When)	Factual	18
	Como (How)	Factual	9
	O_que (What)	Definition	11
	Porque (Why)	WhyQuestion	11
Imperatives		Factual	2
Total			111

CINTIL-QATreebank composition of the corpus in terms of the semantic types of answers expected for the sentences contained in it:

Semantic Answer Types	Qty.
Personal	8
Temporal	18
Number	9
Location	8
Organization	7
Explanation	11
Yes/No	13

Miscellaneous	37

1.2. Representation of the corpora (flat files, database, markup)
The corpus is a single file in XML format.

1.*3. Character encoding*
The characters are in UTF8 encoding.

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
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 licensed for research purposes only, with no redistribution, nor derivatives allowed.

III. Technical Information

3.*1. Directories and files*
The archive that can be downloaded on the Meta-Share platform is a .zip file with one single file.

3.2. Data structure of an entry

The format of a CINTIL-QATreebank uses XML to store different levels of linguistic annotation, for example, information on constituency and grammatical function and any extra information that the corpus has. The adopted XML format was XCES2. The goal is to provide a fully-specified web-based format that enables maximal inter-operability not only among annotations of the same phenomena, but across annotation types.

The XCES file has several levels of annotation and extra information. We present below the levels of annotation:
• Id: internal identifier of CINTIL-QATreebank.
• Source: identify the source of information, in this case CINTIL-Treebank.
• SourceId: identifier in the source of information.
• OriginalSentence: Treebank original sentence in CINTIL- Treebank.
• OriginalTree: original constituency tree in the CINTIL-Treebank.
• Sentence: sentence QATreebank.modified for CINTIL-QATreebank.
• Tree: constituency tree of interrogative sentence.

• InterrogativePronoum: interrogative pronoun of the sentence, if the sentence does not have a pronoun is marked as none.
• QuestionType: question type can assume the values: Factual, Definition, WhyQuestion and Yes/No.
• AnswerType: answer type, can assume the values: Person, Date, Number, Localization, Organization, Explanation, Yes/No and Miscellaneous.
• SentenceType: sentence type can assume the values: Partial, Total Affirmative, Total Negative or Imperative.
• Restriction: restriction can assume the values: period or event. If the sentence does not have a restriction is marked as none.
• Answer: the answer of the interrogative sentence.
• Variant: language variant can assume the values: PE (European Portuguese), PB (Brazilian Portuguese) or PO (Portuguese – both variants).

```
<?xml version="1.0" encoding="UTF-8"?>
<cesAna>
<struct type="sentence">
    <feat value="S1" name="Id"/>
    <feat value="CINTIL" name="Source"/>
    <feat value="b104" name="SourceId"/>
    <feat value=" Washington acompanhou os movimentos de Saddam desde a primeira hora ." name="OriginalSentence"/>
    <feat value="[S [S [NP-SJ-ARG1 [N Washington]] [VP [VP [V acompanhou] [NP-DO-ARG2 [ART-SP os] [N' [N movimentos] [PP-OBL-ARG1
        [P de] [NP-C [N Saddam]]]]]] [PP-M-TMP [P desde] [NP-C [ART-SP a] [N' [ORD-M-PRED primeira] [N hora]]]]]] [PNT .]]" name="OriginalTree"/>
    <feat value="Quem acompanhou os movimentos de Saddam desde a primeira hora ." name="Sentence"/>
    <feat value="[CP [CP [NP-SJ-ARG1_1 [INT Quem]] [S [NP-SJ-ARG1_1 *GAP*] [VP [VP [V acompanhou] [NP-DO-ARG2 [ART-SP os] [N' [N movimentos] [PP-OBL-ARG1 [P de]
        [NP-C [N Saddam]]]]]] [PP-M-TMP [P desde] [NP-C [ART-SP a] [N' [ORD-M-PRED primeira] [N hora]]]]]]] [PNT ?]]" name="Tree"/>
    <feat value="QUEM" name="InterrogativePronoun"/>
    <feat value="FACTUAL" name="QuestionType"/>
    <feat value="PERSON" name="AnswerType"/>
    <feat value="PARTIAL" name="SentenceType"/>
    <feat value="PERIOD" name="Restriction"/>
    <feat value="Washington" name="Answer"/>
    <feat value="PO" name="Variant"/>
</struct>
</cesAna>
```

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

The corpus is composed by 111 sentences requiring 12 KB compressed (124 KB uncompressed) for disk storage.

IV. Content Information

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

This is a monolingual and 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*

Since this Treebank is based on the CINTIL-TreeBank, the text registers represented into the corpus comprises news from daily and general newspapers, literary language from novels, and, additionally, sentences from test set.

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

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.

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

Corpus composed of sentences from newspaper texts annotated with their syntactic constituency trees, further enriched with information on grammatical functions and semantic role labels (Branco et al., 2010; Gonçalves and Branco, 2009).

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

It does not apply.

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)

Please, see section 1.1.

4.5. Intended application of the corpus

CINTIL-QATreebank can be used in language science and technology general research, but it was created particularly for the development of automatic Question Answering systems.

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

CINTIL-QATreebank was built from CINTIL-TreeBank that was 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).

This treebank presented here was obtained by manually transforming the syntactic representation of declarative sentences in an existing treebank into their interrogative and imperative counterparts by a human linguistic expert.

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, 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, LREC2010 - The 7th international conference on

Language Resources and Evaluation, La Valleta, Malta, May 19-21, 2010.

Branco, António and Francisco Costa, 2008, "LXGram in the Shared Task "Comparing Semantic Representations" of STEP2008", In Johan Bos and Rodolfo Delmonte (eds.), Semantics in Text Processing, London, College Publications, Research in Computational Semantics Series, Vol. 1, pp.299-314.

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.

Gonçalves, Patrícia, Rita Santos and António Branco (2012). “Treebanking by Sentence and Tree Transformation: Building a Treebank to support Question Answering in Portuguese”. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Gonçalves, Patrícia, António Branco, 2009, "CINTIL-Treebank Searcher", In Proceedings of the I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages, Porto Salvo, September 3-4, 2009.