Treebanking by Sentence and Tree Transformation: Building a Treebank to Support Question Answering in Portuguese

Patrícia Nunes Gonçalves, Rita Santos, António Branco

University of Lisbon
Edifício C6, Departamento de Informática
Faculdade de Ciências, Universidade de Lisboa
Campo Grande, 1749-016 Lisboa
{patricia.nunes, rita.santos, antonio.branco}@di.fc.ul.pt

Abstract

This paper presents CINTIL-QATreebank, a treebank composed of Portuguese sentences that can be used to support the development of Question Answering systems. To create this treebank, we use declarative sentences from the pre-existing CINTIL-Treebank and manually transform their syntactic structure into a non-declarative sentence. Our corpus includes two clause types: interrogative and imperative clauses. CINTIL-QATreebank can be used in language science and techology general research, but it was developed particularly for the development of automatic Question Answering systems. 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.

Keywords: treebank, question answering, tree transformation

1. Introduction

A treebank is a corpus of texts in which every sentence has been annotated with syntactic structure. This resource is important for the development of advanced tools and applications in the NLP area.

This paper presents CINTIL-QATreebank, a treebank composed of Portuguese sentences that can be used as input to Question Answering (QA) systems and by their respective answers. Though this corpus is open to be used and explored for general purpose language science and technology, it´s development was motivated by the intention to build a dataset that supports the development and evaluation of processing tools (e.g. sentence parsers) which are important for the deployment of automatic QA systems.

Question Answering is the task of answering a query formulated in natural language. From a query and a set of documents, a QA system extracts and provides an answer (Paşca, 2003). Though interrogative sentences are the main object of study in the area of QA, we also include imperative clauses in our treebank since they tend to be commonly used in QA systems as well.

Most of the research in this area has been focused on English. For Portuguese, the field of QA is expanding and the systems that have been developed still have much room for improvement. To the best of our knowledge, there is no resource like CINTIL-QATreebank for Portuguese, which combines syntactically annotated sentences with a variety of information that is specific for QA tasks.

2. Corpus construction and annotation

CINTIL-QATreebank is built over CINTIL-Treebank, a 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).

The CINTIL-Treebank annotation is performed by experts in Linguistics according to the mainstream method of annotation that is deemed to ensure a more reliable outcome: double-blind annotation followed by adjudication.

The annotation work is supported, and its quality and consistency is ensured, by resorting to a computational grammar. Each sentence is automatically analyzed by LX-Gram (Branco and Costa, 2008), an advanced grammar for the deep linguistic processing of Portuguese. Once a parse forest is obtained for a given sentence, independent annotators choose the analysis that each consider to be correct. In case of divergence between annotators, an adjudicator makes the final decision.

Constituency trees hold the usual relations between syntactic constituents by following a basic X-bar scheme. The complete guidelines for CINTIL-Treebank annotation can be found in (Branco et al., 2011).

Sentence selection: The first step is the selection of those declarative sentences that will to be transformed into their interrogative/imperative counterparts. This selection is guided by the following criteria:

- Sentence length: Short sentences were discarded. This is an attempt to remove trivial sentences from the treebank, leaving in only those sentences that display interesting syntactic structure.
- Avoid repetition: In order to obtain a balanced representation of the various types of question, we did not convert several sentences that have syntactic structures which are already well represented in the treebank.

As a result, we obtained 85 declarative sentences that carry through to the tree transformation step.

Tree transformation: The second step is the manual transformation of the declarative sentences into their non-declarative counterparts. The same design options for the syntactic representation of the sentences adopted for CINTIL-Treebank were thus also adopted here for the new CINTIL-QATreebank. To account for both European and Brazilian variants of Portuguese, in cases where different syntactic options are possible, the non-declarative sentences generated are marked with respect to which variant they conform to.

In many cases, a single declarative sentence can lead to several interrogative sentences. This happens due to two main reasons: (i) by differences in the language variants for the same type of non-declarative; but also (ii) by the fact that it is possible to apply more than one transformation to a declarative in order to obtain a non-declarative, for example:

Original sentence in CINTIL-Treebank:

Washington acompanhou os movimentos de Saddam desde a primeira hora.
"Washington followed the movements of Saddam since the beginning."

New sentences in CINTIL-QATreebank:

QUEM acompanhou os movimentos de Saddam desde a primeira hora?
"Who followed the movements of Saddam since the beginning ?"
QUANDO é que Washington acompanhou os movimentos de Saddam?
"When did Washington follow the movements of Saddam?"
Washington acompanhou os movimentos de Saddam, QUANDO?
"Washington followed the movements of Saddam, when?"

The various transformations applied to the declarative sentences are the following:

A. By just adding an interrogative pronoun to the declarative sentence and a question mark:

   Original sentence:
   Negociações falham em Jerusalém.
   "Negotiations fail in Jerusalem."
   New sentence:
   QUE negociações falham em Jerusalém?
   "*WHICH* negotiations fail in Jerusalem?"

   Note that the interrogative pronoun does not necessarily need to be at the beginning of the sentence:
   Original sentence:
   O conselho emitiu 17 pareceres.
   "The board has issued 17 opinions."
   New sentence:
   O conselho emitiu QUANTOS pareceres?
   "The board has issued *HOW MANY* opinions?"

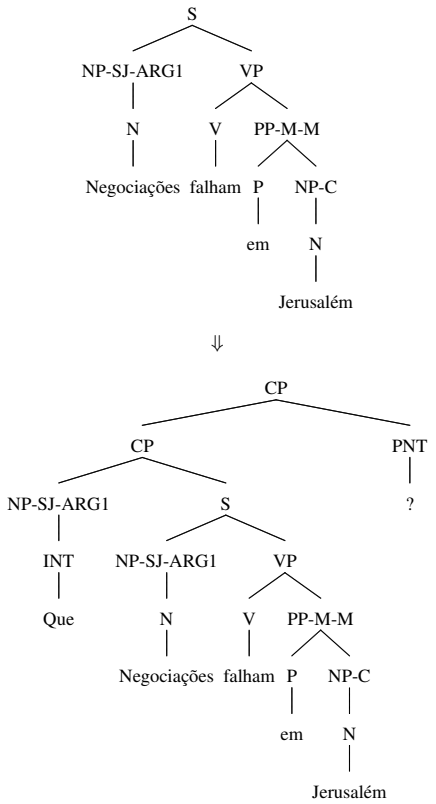

Figure 1: Transformation (A)

B. By adding the interrogative pronoun together with *é que* (by adding only the interrogative pronoun we get a question in the Brazilian Portuguese variant).
   New sentences:
   COMO é que as negociações falham em Jerusalém?
   "*HOW* do negotiations fail in Jerusalem?"
   PORQUE é que as negociações falham em Jerusalém?
   "*WHY* do negotiations fail in Jerusalem?"

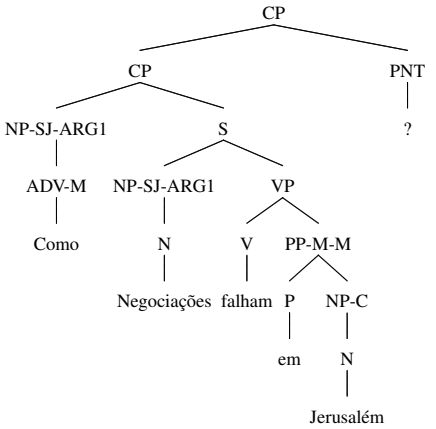

Figure 2: Transformation (B)

C. By replacing a phrase with an interrogative pronoun, which creates "echo-interrogatives". Note that "echo-interrogatives" are different from regular interrogatives by their declarative syntax and rising intonation: the speaker echoes an attributed utterance and expresses a questioning attitude to some aspect of

its form or its content. The function of an "echo-interrogative" is thus rather exclamatory or requiring (in that it demands a repetition or clarification) than interrogative:

Original sentence:
Fábrica de tintas centenária ardeu em Vila Nova de Gaia.
"Centennial ink factory burned in Vila Nova de Gaia."
New sentence:
Fábrica de tintas centenária ardeu ONDE?
"Centennial ink factory burned *WHERE*?"

But also by replacing a phrase with an interrogative pronoun and displacing it to the left periphery of the sentence:
New sentence:
ONDE ardeu a fábrica de tintas centenária?
"*WHERE* did the centennial ink factory burn?"

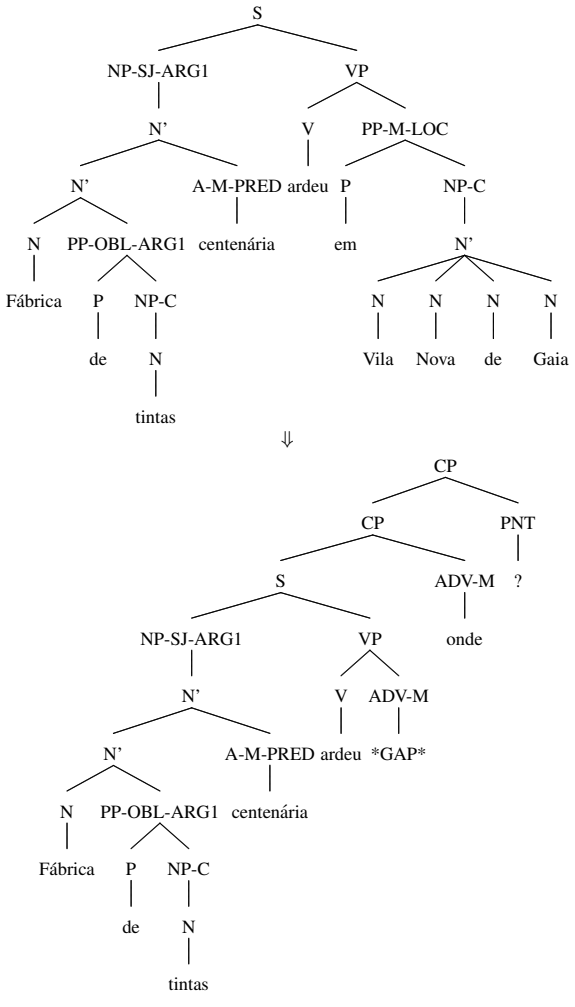

Figure 3: Transformation (C)

D. By just changing the final punctuation mark of the sentence to a question mark, generating, as a result, total interrogatives, which require an affirmative or negative answer:
Original sentence:
Gulbenkian não suspendeu financiamento.
"Gulbenkian did not suspend financing."
New sentence:
Gulbenkian não suspendeu financiamento?
"Did not Gulbenkian suspend financing?"

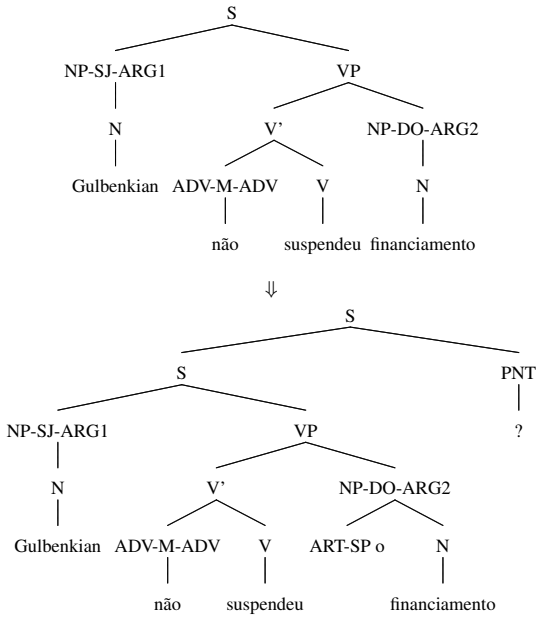

Figure 4: Transformation (D)

E. By adding a command verb in the imperative mood, and changing the clause type, from declarative to imperative:
Original sentence:
Este investimento é superior ao efectuado pela APDL nos últimos vinte anos.
"This investment is higher than the one made by APDL in the last twenty years."
New sentence:
INDIQUE um investimento superior ao efectuado pela APDL nos últimos vinte anos.
"*NAME* one investment higher than the one made by APDL in the last twenty years."

Note that all transformations were performed manually by experts in Linguistics, holding a BA degree.
The sentences were manipulated in their labeled bracket notation. To visualize the outcome of this transformation in terms of diagrammatic display as trees, and also to confirm and validate it, we used the PhpSyntaxTree tool[1]. This is a web application that creates syntax tree graphs from phrases entered manually in labeled bracket notation.

QA-specific annotation: The third and last step in the creation of CINTIL-QATreebank is where the QA-specific information is added to each parse tree. This information consists of:

- the sentence type: indicates whether the sentence is a partial interrogative, a total interrogative, or a question in the form of an imperative clause.
- the interrogative pronoun: information, which only applies to partial interrogatives, indicates what is the main interrogative pronoun of the sentence.

[1] `http://ironcreek.net/phpsyntaxtree/`

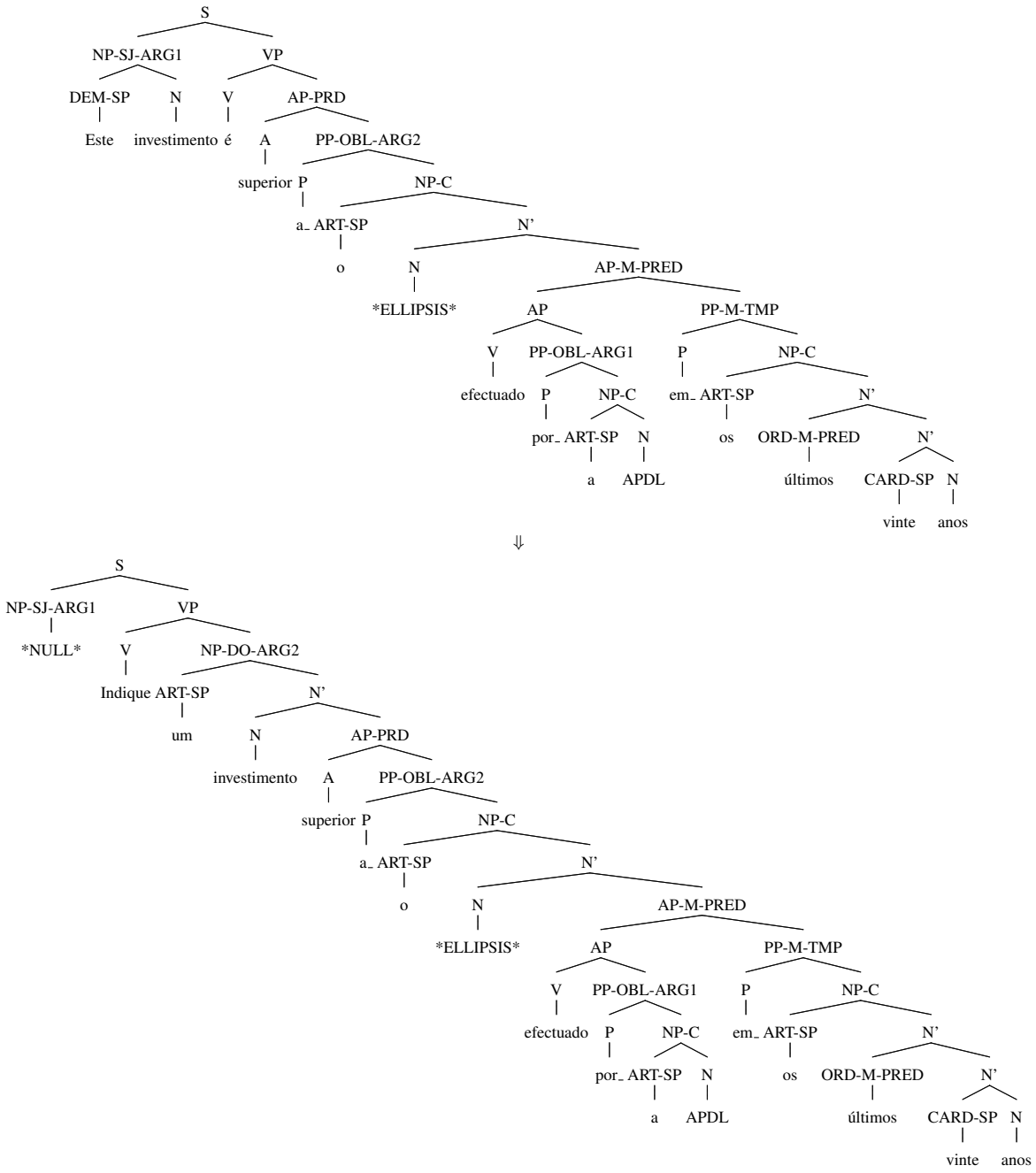

Figure 5: Transformation (E)

- the question type: factual, definition, yes/no, why-question
- the expected answer type: states what is the semantic type we expect for the answer to the question.

Additional details on this QA-specific information are given in the Sections 3 and 4.

3. Treebank composition: grammatical types of questions

Our corpus includes two clause types: on the one hand interrogative clauses, used in asking questions, and on the other hand imperative clauses, used in issuing orders or directives.

Grammatically, interrogative clauses can be of two distinct types: total interrogatives, which typically elicit a confirmation or a denial, and partial interrogatives, which typically elicit open-ended responses. The latter rely on the introduction of some interrogative pronoun or *wh*-word. Total interrogatives can be subdivided into three types:

- affirmative: *Negociações falham em Jersusalém?* "Do negotiations fail in Jerusalem?"
- negative: *A Itália não se apurou para os quartos-finais do Europeu?* "Didn't Italy make it to the EURO quarterfinals?"
- alternative: *A prova pode ser feita com uma declaração da embaixada respectiva ou um atestado*

de residência? "Can the exam be done with a declaration form from his embassy or with a residence certificate?".

Partial interrogatives, in turn, may be more complex. Just as a first example, consider the pronoun *Que* ("*What*/*Which*"). It differs from the rest of the wh- interrogative pronouns in as much as it may give raise to two different types of questions: *Que* ("*What*") and *Que_N* ("*Which*").
Finally, despite not being questions per se, imperative clauses are commonly used in QA systems. The most recurrent verbs used are the following: "mention", "say", "quote", "indicate", "name", "argue", "discuss", "demonstrate", "determine", "specify", "explain", "consider", "disclose", "report", "check", among others.
Table 1 shows the corpus 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

Table 1: Distribution of the grammatical types of sentences in the corpus.

4. Annotation with semantic types of expected answers

On a par with constituency trees, our corpus includes further relevant information for developing the area of QA systems.
Questions of a given grammatical type may elicit answers of a range of different semantic types. It is important to annotate the sentences in the corpus with such information. When the expected answer is factual, the questions are annotated with a label indicating the semantic type of the expected answer.
The answer types are organized along the following semantic categories:

- Person: *QUEM revelou a notícia num debate realizado em Alijó?* "*WHO* exposed the news in a debate which took place in Alijó?";
- Time: *QUANDO é que Monge toma posse?* "*WHEN* does Monge come into office?";
- Number: *QUANTOS anos é que a PSP festeja?* "*HOW MANY* years is PSP celebrating?";
- Location: *ONDE será realizada a eliminatória entre franceses e italianos?* "*WHERE* will the final round between French and Italian be?";
- Organization: *QUE banco reduz taxa de intervenção?* "*WHICH* bank is reducing the intervention fee?";
- Explanation: *PORQUE é que Itamar critica o Governo brasileiro?* "*WHY* is Itamar criticizing the Brazilian Government?";
- Yes/No: *A moeda é o espelho da economia?* "Is currency the mirror of the economy?";
- Miscellaneous (on every occasion none of the above types fit in): *COMO a Praça Luís Camões será embelezada?* "*HOW* will Luís Camões Square be adorned?".

Table 2 provides an overview of the composition of the corpus in terms of the semantic types of answers expected for the sentences contained in it:

Semantic Answer types	Qty
Person	8
Temporal	18
Number	9
Location	8
Organization	7
Explanation	11
Yes/No	13
Miscellaneous	37

Table 2: Distribution of the types of answers for the sentences in the corpus.

5. CINTIL-QATreebank format

The format of a CINTIL-QATreebank uses XML (eXtensible Markup Language). XML is used 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 XCES[2]. 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 aim to provide an environment in which annotations can be easily defined and validated. In summary, the XCES serves as an interface between different types of annotations. Figure 6 shows an excerpt of CINTIL-QATreebank XML format. The XCES file has several levels of annotation and extra information. We present below the levels of annotation:

- struct type: information about sentence type: partial, total or imperative.
- id: internal identifier of CINTIL-QATreebank.

[2] http://www.xces.org/

- source: identify the source of information, in this case CINTIL-Treebank.
- sourceId: identifier in the source of information.
- originalSentence: original sentence in CINTIL-Treebank.
- originalTree: original constituency tree in the CINTIL-Treebank.
- sentence: sentence 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 e Yes/No.
- answerType: answer type, can assume the values: Person, Date, Number, Localization, Organization, Explanation, Yes/No e Miscellaneous.
- answer: the answer of the interrogative sentence.
- variant: language variant (European and Brazilian)

```
<cesAna>
<struct type="partialInterrogative">
<feat name="id" value="s1"/>
<feat name="source" value="CINTIL-Treebank"/>
<feat name="sourceId" value="b104"/>
<feat name="originalSentence" value="Washington acompanhou os movimentos
          de Saddam desde a primeira hora."/>
<feat name="originalTree" 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 "/>
<feat name="sentence" value="Quem acompanhou os movimentos de Saddam
          desde a primeira hora?"/>
<feat name="tree" 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 ?]]"/>
<feat name="interrogativePronoum" value="Quem"/>
<feat name="questionType" value="Factual"/>
<feat name="answerType" value="Person"/>
<feat name="answer" value="Washington"/>
<feat name="variant" value="Brazilian"/>
</struct>
</cesAna>
```

Figure 6: An excerpt of the CINTIL-QATreebank format

6. CINTIL-QATreebank in use

In order to test an application of CINTIL-QATreebank, we performed an experiment in which a statistical parser was trained over the treebank. We chose the Stanford Parser (Klein and Manning, 2003), which we had previously used to create LX-Parser, a parser for Portuguese that achieved state-of-the-art performance scores (Silva et al., 2010). The goal of this experiment it to assess to what extent the performance of the parser over interrogative sentences improves when interrogative sentences are added to the training data.

For this experiment, we set aside 10% of the sentences in CINTIL-QATreebank for testing, and added the remaining sentences to CINTIL-Treebank (version 3 with 5422 declarative sentences) in order to form the training corpus.

The results obtained in this experiment were unsatisfactory, reaching a F-measure score of 66% for bracketing correctness (with and without the interrogative sentences in the training data), although some improvement was achieved in tagging accuracy.

An analysis of the results shows that adding CINTIL-QATreebank to training data does not have a measurable impact on the performance of a statistical parser. The reason for this is twofold. On the one hand, when building CINTIL-QATreebank, we tried to create a linguistically interesting dataset (e.g. long sentences with complex structures), which raise data-sparseness issues; one the other hand, the dataset is still rather small, having only 111 sentences, which further intensifies those issues.

7. Concluding remarks

In this paper we present CINTIL-QATreebank, a treebank composed of sentences from Portuguese that can be used to support the development of Question Answering systems.

This treebank was obtained by manually transforming the syntactic representation of declarative sentences in an existing treebank into their interrogative and imperative counterparts. We described the methodology used in these transformations as well as the composition of the corpus, in terms of domain, size and distribution of sentences. We also describe the several layers of annotation in CINTIL-QATreebank, which combine syntactically annotated sentences with a variety of information that is specific to QA tasks. The annotation contains: trees with information on constituency and grammatical function, sentence type, interrogative pronoun, question type, and semantic type of expected answer.

On a par with the construction of the treebank, we performed an experiment where a statistical parser was trained. The results obtained were unsatisfactory even though some improvement was achieved in terms of tagging accuracy. This experiment will be repetead with a larger version of the treebank that is under construction.

We present the CINTIL-QATreebank XML format. This format is portable, being easily handled by both humans and automatic processing, and allows inter-operability among different levels of information. This format is also easily extensible, allowing for new layers of linguistic annotation that may be incorporated in the future.

The description provided in this paper corresponds to the current status of the corpus. While the goals and the methodological approach are defined and the technical foundations are set up, the development work will now continue and the corpus will keep on being extended by the inclusion of more sentences. The final goal is to use the whole of CINTIL-Treebank to build a new version of CINTIL-QATreebank. We estimate that the final version may have around 5,000 sentences.

In addition, we are developing a tool to automatically transform the trees of declarative sentences. In the future we will train the statistical parser with the complete CINTIL-QATreebank. This parser will be used to annotate other interrogative sentences which may be used as a method for automatic, or semi-automatic, extension of the treebank.

8. References

António Branco and Francisco Costa. 2008. Lxgram in the shared task "comparing semantic representations" of step 2008. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 299–310, Stroudsburg, PA, USA. Association for Computational Linguistics.

António Branco, 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 Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

António Branco, João Silva, Francisco Costa, and Sérgio Castro. 2011. Cintil treebank handbook: Design options for the representation of syntactic constituency. TR 2011-02, University of Lisbon, Faculty of Sciences, Department of Informatics, Lisbon.

Patricia Gonçalves and António Branco, 2009. *Buscador Online do CINTIL-Treebank*. XXV Encontro Nacional da Associação Portuguesa de Linguística (APL), Lisbon, Portugal.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Marius Paşca. 2003. *Open-Domain Question Answering from Large Text Collections*. CSLI Studies in Computational Linguistics. CSLI.

João Silva, António Branco, and Patricia Gonçalves. 2010. Top-performing robust constituency parsing of portuguese: Freely available in as many ways as you can get it. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).