Timber! Issues in treebank building and use

Diana Santos[1]

[1] Linguateca, SINTEF Telecom & Informatics, Pb 1124 Blindern,
0314 Oslo, Norway
Diana.Santos@sintef.no

Abstract. We discuss several treebank conceptions in the literature and show that their requirements may be incompatible, describing then the options taken in the construction of a Portuguese treebank, in what concerns human vs. automatic intervention. Use cases are then listed in connection with a Web search tool (*Águia*), whose philosophy and implementation is presented.

1 Introduction

Treebank building has become fashionable lately with the number of treebank projects growing exponentially. However, there are quite different ways to conceive both the end result and the way to go about achieving it.

As far as treebank purpose is concerned, one can identify at least the following different views (an example of each is provided with no claim for exhaustiveness):

1. a treebank as a resource for the building of automatic processing tools [1]

2. a treebank is an evaluation resource to compare the performance of different parsers [2]

3. a treebank is a linguistic resource to fix and display the syntactic analysis of complex text (and can consequently be used for teaching purposes) [3]

4. a treebank is a proof of the qualities of a given theory[1]

Even though most papers on treebanks so far declare that they expect it to be used for (almost) all these purposes, a closer analysis show that the requirements to achieve these different goals are incompatible or, at least, difficult to harmonize.

For example, if you want to train computer programs on the treebank, you'd better only revise and clean information about which there is some understanding on how to program or achieve. In other words, information added by a human drawn from sources such as world knowledge or cognitive processing difficulties, as well as the result of complex inferences based on a distant context are not, in general, reproducible automatically and are therefore of no interest for goal 1.

In fact, desirable features for a treebank type 1 are: consistency, few information pieces and enough occurrences of each feature (so that systems have enough examples from which to learn).

[1] This is rarely stated but it often constitutes an additional motive to engage in treebank building.

On the other hand, if one wants to create a gold standard for ensuing evaluation endeavours, it is possible that one chooses not to annotate, or not to decide in cases where consensus was not reached. The result may not be consistent or complete, but it is empirically adequate.

If one wants to use a treebank for linguistic investigation, one would value most of all the information that only linguists could add, and actually almost "despise" the sort of low-level information that satisfaction of goal 1 would require (like correct morphological information). Consistency would be a platonic goal, but naturalness of the annotation and relevance to linguistic concerns would be features of such a treebank type 3.

Finally, a treebank type 4 should maximize diversity (although keeping consistency) in order to prove the expressiveness of the theory and therefore would again fail to be useful for goal 1.

Our treebank project, *Floresta Sintá(c)tica* [4], aimed (eventually) at building a type 3 treebank, given that we had an underlying symbolic parser which provided a lot of information and it was unrealistic to expect that a parser could be trained to learn it all. Reducing it would be a bold decision, which was not taken.

2 Annotation schemes

Wilson et al. [5] describe a set of desiderata for an annotation scheme where they emphasize that it should reflect distinctions a human could be expected to reliably annotate ("naturalness"). It is easy to find huge numbers of information tags that are not easy to annotate reliably (even though they may be used liberally by parsers); it is also the case that many of the easy to annotate categories for humans are, so far, never even attempted automatically.

2.1 Can our treebank type 3 be turned into an evaluation treebank (type 2)?

How to create a treebank that allows one to actually evaluate different parsers without forcing the linguistic view of the present treebank? Although we, as creators, might wish that it took the same role as the Penn Treebank [6] for English, used as a de facto standard, we are fully convinced of the need and advantage of cooperatively agreeing on a standard.

We believe that the present treebank can be used for experimentation and evaluation, and to make problems and disagreements explicit, but that one should try to build from scratch (or from a much stricter set of rules and using as point of departure the present treebank) a real evaluation resource that allows one to test given aspects of syntactic parsers for Portuguese, probably following Gaizauskas et al.'s proposal [7] for creating evaluation resources quickly, and using some manual analysis as in [8].

We are, in any case, convinced that it is totally unrealistic to expect that one can list parsers' outputs and try to harmonize or agree on the meaning of the different labels. This was already an enormous task for a field as (comparatively) simple as

Portuguese morphological analysis, for which an unexpected high degree of disagreement has been reported [9,10]. It is also enough to browse several different Portuguese grammar books to see that they verse about different subjects. Incidentally, it is also quite rare that they define their primitives.

2.1.1 Decisions as to the process

Let us give a concrete example of one of the many things that are far from trivial: The underlying parser – thoroughly described in [11] – assigns the two following syntactic categories to noun phrases attached to noun phrases: N<PRED (predicative adject) and APP ((adnominal) apposition), exemplified respectively as *Jerônimo, **um grande cacique**, temia ninguém* and *O grande cacique**, Jerônimo**, conhecia o seu país como mais ninguém*. The definitions in the treebank documentation follow:

APP: The prototypical apposition is a name or definite np, identifying the np-head it postmodifies: "Jerónimo, o grande cacique" or "o seu advogado, Marco da Silva".

N<PRED: The prototypical postnominal predicative is an adjective, attributive participle or indefinite np, predicating something about the np-head it postmodifies, typically with the semantic relation of 'IS' (=): "Jorge Gomes, funcionário" or "Jorge Gomes, contente com a vida".

It has proved, no matter the many heuristics or rules of thumb proposed[2], an extremely difficult decision to be done in practice, when one leaves the idealized landscape and comes to real utterances. Time and again there was uncertainty about which classification to assign. Examples are:

*No final do jogo, adeptos do Sporting lançam garras e pedras para a tribuna de honra, onde estavam Manuela Ferreira Leite, **ministra da Educação**, e Vítor Vasques, **presidente da FPF**.*

*Na mesma zona em que foi encontrado o templo, **a Alcáçova**, a caminho das Portas do Sol, foram ainda descobertas cisternas romanas que estão também a ser objecto de escavações e estudos arqueológicos.*

Several solutions about how to proceed concerning the assignment of these labels have been proposed, each of them showing, in fact, different conceptions of what a treebank should be for.

1. mark/revise the clear cases and leave the parser's output when no clear opinion
2. create a new non-committal label (let us call it here **npstack**) and
   a. transform all cases of either label into it, or
   b. use it only for the unclear cases

Even though no final decision was (so far) taken, this micro-controversy allows us to illustrate the consequences of each option with respect to the treebank goals mentioned in the beginning of the present paper: The first option was aimed at improving the parser, so that it agreed with human reasoning when humans had something to say. The result would probably not be consistent, and definitely not reflecting human performance, but was obviously ideal for parser improvement.

The second one was, on the contrary, aimed at describing human interpretation (and not a parser's). Option a) had the goal of making the task of building (and consequently revising) the treebank simpler, taking implicitly the view that this is probably not a human task – when we see two NPs following each other, it is not relevant to understand whether the second is APP or N<PRED.

[2] Such as: when an abbreviation follows what it is an abbreviation for, tag it @APP: *Partido da Terra **(PT)***; APP implies an identity relation, while N<PRED adds information, ...

Option b) aimed at pushing the limits of what is encoded in the treebank, to all decisions a human being could possibly make. So, if in some cases a person can reliably do some distinction, encode it, leaving the rest also encoded as “not possible to decide”, paving the way for a more thorough overview of what can be relevant in interpreting Portuguese text. The lurking assumption here is that the categories N<PRED and APP do have some relevance for Portuguese grammar, assumption supported by its being used by the parser and mentioned in several traditional grammars.

2.1.2 Decisions as to the encoding

In [12] it was argued that evaluation of parsed corpora has to take into account at which level a given phenomenon was (or not) represented. In particular, it was probably irrelevant to assign right or wrong to PoS classification of *clara* in *clara e sucintamente* [13], provided this is recognized as an adverbial phrase.

Also, it is even probably wrong to assign the gender and number features masculine plural to *surpresa* in *presentes surpresa* although it behaves rather like a common masculine plural adjective like *caros* (cf. *Estes presentes foram surpresa*, *os presentes surpresa estão no canto, acho estes presentes muito pouco surpresa!*) Similarly, when we have a fixed expression like *pele vermelha*, as in *o chefe pele vermelha bocejou*, we can assign to it, in addition to internal features, external features. These two sets of features may or may not agree. So, just as the question of whether *que* is a subject or an object has to be stated relative to the clause in consideration (relative clause or main clause), the question whether *pele* is a noun or an adjective depends on which context: in the noun phrase above or below?[3] What is relevant is that *pele* is a feminine noun for the lower NP (and so requires the adjective *vermelho* to agree) but behaves as a masculine adjectival phrase (or adjective), in the sentence above.

Conflicts may arise when a given lexical item is subject to conflicting requirements due to the different roles it plays, and this may actually even bring changes to the whole language system. For example, if *onde* vs. *aonde* should be selected according to whether the verb describes a movement to some place or not, the two sentences *Vi aonde ele foi* (‘I saw to-where he went’) and *Fui onde ele se escondeu* (‘I went where he hid himself’) are both suboptimal since the two verbs (*ver* and *ir* in the first sentence, *ir* and *esconder* in the second) have different features, and therefore requirements.

Contrary to a common view that parsed corpora should use the same information as lexicons, we believe that the interest of annotating corpora is precisely to investigate how language works and find out what <u>cannot</u> be predicted from the lexicon, as in *surpresa* above.

[3] In the present discussion, we are assuming a dependential framework where features are assigned to words (and functional roles are assigned to head words). The need for upwards and downwards marking remains in a more populated phrase structure formalism, we would just have to say "the clause headed by *que*," or "the phrase headed by *pele*".

3 Águia

Let us present a Web query tool that has been designed with two considerations in mind: 1. to furnish a higher level query language (in the sense of being as much as possible separate from the encoding realities and actual treebank syntax); 2. to be based on a powerful general purpose corpus system (the IMS CWB) instead of writing from scratch a particular treebank specific query system.

This tool is available on the Web (http://www.linguateca.pt/Floresta/) together with a guided tour that tries to give a feeling of the sort of possible queries – as high level as possible.

Águia's more radical (or unusual) feature is that its output is simple text, although the whole treebank is publicly available in its two internal coding formats, and therefore users can, if they want, see and use the tree structure at will. The basis for this feature is that we believe that a treebank user is not (or should not be) primarily concerned with trees, but with the information conveyed by these trees, in order to get at text, to get at language (which comes in the format of words in the written medium).

In addition, we are not yet sure about which are the most interesting questions users really want to ask a treebank. Therefore, we implemented also an open window where people can input questions in natural language and we help them to formulate their questions, with the proviso they are answerable by the actual treebank.

3.1 Kinds of queries

We can distinguish the following kinds of primary uses for a query tool for people (not for programs): The user wants quantitative information about the treebank, such as: What kind of clauses are most frequent? What kind of syntactic objects (phrases) have the function "question", and in what relative weight? What is the most frequent verb in each kind of clause? What is the most common function of a finite clause? In how many cases do adverbs occur in relative clauses?

The user wants to inspect some combinations or categories a little better – because s/he suspects they are wrongly assigned, or because they contradict her/his own beliefs about the language. Some (random) examples: How often can crosscategorial conjunctions be found? Are there subject complements with relative clauses?

The user may simply want to look for specific examples of special cases, related to his or her field of interest: Find clauses including an adverb as immediate constituent; find noun phrases including relative clauses in which the pronoun has the subject (or object, or dative) role; find finite clauses starting with the verb, etc.

The user may also want to look at the underlying generative grammar, according to the examples atested in the treebank: what is the generative grammar of a noun phrase? What is the generation grammar of a particular function?

Or the user may be more interested in the lexicon, and want to determine the grammatical properties of a lexical item: what is the valency grammar of a particular lexical item (verb, preposition)? Given a particular class of adverbs, in which patterns do they occur? When a given lexical item occurs as premodifier of a phrase, which functions does this phrase typically show?

Above, we showed a variety of different questions which could be answered by a single query with Águia. There is obviously no limit to the complexity of the interaction an experienced user may have with the treebank! We list here other questions that include more than one query but should not be too complicated to answer: What is the deepest embedding? (Find finite clauses under finite clauses.) How many prepositional phrases are not directly attached to the preceding phrase? How many noun phrases exhibit a potential attachment ambiguity?

Still, other metalinguistic questions, at the moment not catered for by Águia, but encoded in the treebank, can be asnwered: Which sentences were considered ambiguous in the treebank? Which utterances required world knowledge for disambiguation? (see examples in [14]). Which clauses involve ellipsis or required insertion of additional material in order to be parsed and represented by the human team?

3.2 Use of IMS CWB

The use of the underlying IMS CWB [15-17] is an obviously sound engineering decision, since it offers a well developed and tested set of capabilities, a powerful query language and several utilities. In addition, we believe that there should be, at least from a user point of view, a smooth transition between POS-tagged and annotated corpora, and the fact that the codification of the latter may pose complex problems to the language engineer should be transparent to the user.

The way we used the IMS CWB was straightforward but somehow imaginative: we created several different physical corpora from the manually edited output, that code the treebank in different ways. Depending on the query, the right corpus is used. This is, however, perfectly transparent for the user, who can only distinguish between the manually revised part (*Bosque*, the treebank proper) and the larger automatically produced part (*Floresta Virgem*, “the treebank to be”).

For example, we present an extract of one of the corpora in figure 1, having words as terminals and phrases as structural attributes, and therefore appropriate to look for words inside phrases, while the corpus of figure 2 has phrases as terminals and words as attributes.

```
<u C22-2>
<s>
<fcl0 STA>
“       “       pont    0       1
<fcl1 ADVL>
Se      se      SUB:conj-s      0       2
<vp2 P>
for     ser     AUX:v-fin       FUT_3S_SUBJ     3
firmado firmar  MV:v-pcp        M_S     3
</vp2>
</fcl1>
,       ,       pont    0       1
ninguém ninguém SUBJ:pron-indp  M_S     1
ficará  ficar   P:v-fin FUT_3S_IND      1
<ap1 SC>
mais    mais    >A:adv  <quant> 2
contente        contente        H:adj   M_S     2
<acl2 KOMP<>
do_que  do_que  COM:conj-s      0       3
nós     nós     SUBJ:pron-pers  M/F_1P_NOM/PIV  3
```

```
   </acl2>
   </ap1>
   .        .         pont     0        1
   </fcl0>
```

Figure 1: One of the views of the treebank encoded in the IMS-CWB

```
   <u C22-2>
   vp    P     'v-fin v-pcp '  'AUX MV '      "for firmado "  2
   fcl   ADVL  'conj-s vp2 '   'SUB P '    "Se for firmado "       3
   acl   KOMP< 'conj-s pron-pers '   'COM SUBJ ' "do_que nós "   2
   ap    SC    'adv adj acl2 ' '>A H KOMP<' "mais contente do_que nós" 4
   fcl  STA   'fcl1 pron-indp v-fin ap1 '  'ADVL SUBJ P SC '  "" Se for
firmado , ninguém ficará mais contente do_que nós . " 12
```

Figure 2: Another view of the treebank encoded in the IMS-CWB

While it is outside the scope of the present paper to dwell on technicalities, this small section should be read as a plea for using already existing powerful tools for dealing with large amounts of linguistically analysed text, instead of reinventing the wheel and create new treebank search tools from scratch, as was e.g. done in the TIGER project [18].

We conclude the present paper asking everyone interested in Portuguese syntax to look at *Floresta Sintá(c)tica* and try out *Águia* for the questions they are more interested in, so that we can have a representative idea of the shortcomings and the main user needs, and may be able to develop a tool that can be generally used, also later on, for different treebanks for Portuguese (and even other languages, if the concept turns out to be pertinent).

References

1. Marcus, Mitchell, Kim, Grace, Marcinkiewicz, Mary Ann, MacIntyre, Robert, Bies, Ann, Ferguson, Mark, Katz, Karen, Schasberger, Britta: The Penn treebank: Annotating predicate argument structure. In: Proceedings of the 1994 Human Language Technology Workshop (ARPA) (1994) 110—115.
2. Xia, Fei, Palmer, Martha, Xue, Nianwen, Okurowski, Mary Ellen, Kovarik, John, Chiou, Fu-dong, Huang, Shizhe, Kroch, Tony, Marcus, Mitch: Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In: Gavriladou, M. et al. (eds.), Proceedings of LREC 2000 (2000) 3—10.
3. Skut, Wojciech, Brants, Thorsten, Krenn, Brigitte, Uszkoreit, Hans: A Linguistically Interpreted Corpus of German Newspaper Text. In: Rubio, A. et al. (eds.), Proceedings of LREC 1998 (1998) 705—711
4. Afonso, Susana, Bick, Eckhard, Haber, Renato, Santos, Diana: "Floresta sintá(c)tica": a treebank for Portuguese. In: Rodríguez, M.G., Araujo, C.P.S. (eds.): *Proceedings of LREC 2002* (2002), 1698—1703
5. Wilson, G., Mani, I., Sundheim, B., Ferro, L.: A multilingual approach to annotating and extracting temporal information. In: Proceedings of the Worskhop for Temporal and Spatial Information Processing (Toulouse, July 7th 2001) (2001) 81—87

6. Marcus, Mitchell P., Santorini, Beatrice, Marcinkiewicz, Mary Ann: Building a large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, **19** (1993) 313—330
7. Gaizauskas, R., Hepple, M., Huyck, C. Modifying Existing Annotated Corpora for General Comparative Evaluation of Parsing. In: Workshop on Evaluation of Parsing Systems, at the LREC'98 (1998)
8. Carroll, John, Minnen, Guido, Briscoe, Ted: Corpus annotation for Parser Evaluation. In: Uszkoreit, H. et al (eds.): Proceedings of LINC-99: Linguistically Interpreted Corpora, EACL (Bergen, 12 June 1999) (1999) 35—41
9. Santos, Diana, Rocha, Paulo: AvalON: uma iniciativa de avaliação conjunta para o português. In: Actas do XVIII Encontro da Associação Portuguesa de Linguística (Porto, 2-4 de Outubro de 2002) (2003)
10. Santos, Diana, Costa, Luís, Rocha, Paulo: Cooperatively evaluating Portuguese morphology. In: this volume (2003)
11. Bick, Eckhard: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)
12. Santos, Diana, Gasperin, Caroline: Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation. In Rodríguez, M.G.; Araujo, C.P.S. (eds.): Proceedings of LREC 2002 (2002) 597—604
13. Afonso, Susana: Clara e sucintamente: Um estudo em corpus sobre a coordenação de advérbios em –mente. In: *Actas do XVIII Encontro da Associação Portuguesa de Linguística* (Porto, 2-4 de Outubro de 2002) (2003)
14. Afonso, Susana, Bick, Eckhard, Haber, Renato, Santos, Diana: Floresta sintá(c)tica: um treebank para o português. In: Gonçalves, Anabela, Correia, Clara Nunes (eds.): *Actas do XVII Encontro da Associação Portuguesa de Linguística* (Lisboa, 2-4 de Outubro de 2001) (2002) 533—545
15. Christ, Oliver: A modular and flexible architecture for an integrated corpus query system. In: *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research* (1994) 23—32
16. Evert, Stefan: CQP Query Language Tutorial. IMS Stuttgart, 13 Out 2001
17. Evert, Stefan; Kermes, Hannah: Annotation, storage, and retrieval of mildly recursive structures. In: Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003) (2003)
18. König, Esther, Lezius, Wolfgang: A description language for syntactically annotated corpora. In: *Proceedings of COLING 2000* (2000) 1056—1060