DeepBankPT and companion Portuguese treebanks in a multilingual collection of treebanks aligned with the Penn Treebank

António Branco, Catarina Carvalheiro, Francisco Costa, Sérgio Castro, João Ricardo Silva, Cláudia Martins and Joana Ramos
Universidade de Lisboa,
Departamento de Informática, Faculdade de Ciências

Abstract: We present a new collection of treebanks for the Portuguese language, comprising five datasets that cover major types of grammatically annotated corpora: TreeBankPT, PropBankPT, DependencyBankPT, LogicalFormBankPT and DeepBankPT. This collection is the Portuguese part of a broader multilingual collection of aligned treebanks that are developed for different languages, including English, under the same methodological principles and guidelines, and whose raw text versions are translations of the Penn Treebank, a de facto standard dataset for research on language technology.

Keywords: TreeBank, PropBank, DependencyBank, LogicalFormBank, DeepBank, aligned treebanks, multilingual datasets, Portuguese.

1 Introduction

In this paper we introduce a new collection of state of the art treebanks for the Portuguese language. This collection comprises five datasets that cover major types of grammatically annotated corpora:

- TreeBankPT encodes the syntactic constituency of sentences;
- PropBankPT expands the information in the dataset above with semantic roles;
- DependencyBankPT records the information on the grammatical dependencies holding among the expressions in the sentences;
- LogicalFormBankPT associates each sentence to the representation of its semantics in a logical formalism;
- DeepBankPT associates each sentence in the dataset with its fully fledged grammatical representation, thus including also the dimensions recorded in the other four treebanks above into a single integrated representation.

The treebanks in this collection were developed under the advanced design options of dynamic treebanks (Oepen *et al*., 2002), thus being the second collection developed under such conditions for the Portuguese language, together with the pioneer CINTIL collection (Branco *et al*., 2010). The present collection however goes a crucial step further in terms of covering an important gap that existed for Portuguese: these treebanks are parallel and sentence aligned to treebanks from English and other

languages, thus being suitable to support the development of a range of multilingual applications, including machine translation.

The above listed corpora for Portuguese are part of a broader multilingual collection of treebanks that are built not only over texts that are translationally equivalent among the different languages, but also under the same methodological principles and guidelines. What is more, these texts are translations from an English corpus that is a de facto standard dataset, over which most progresses on parsing have been obtained in the last decades, namely the WSJ corpus, upon which the Penn Treebank was established. That is, on the one hand, the WSJ corpus was re-annotated to develop a new dynamic English treebank, DeepBankENG (Flickinger *et al.*, 2012a); on the other hand, the WSJ corpus was translated into other languages, including Portuguese, and the corresponding collection of treebanks, ParDeepBank, has been developed, whose datasets *ipso facto* became aligned with each other and with the new English treebank.

In the next Section 2, the methodological conditions for the development of dynamic DeepBanks are introduced. The following Section 3 addresses how DeepBanks can support the establishment of a high quality collection of aligned treebanks for a given language. In Section 4, we present the ParDeepBank corpora, a multilingual collection of DeepBanks, which DeepBankPT is a component of, and in Section 5, the Portuguese DeepBankPT and companion treebanks are presented in detail. This paper closes with Section 6, with concluding remarks.

2 Advanced treebanks

Grammatically interpreted corpora have played a major role in the progress of language technology. Such accurately annotated data sets have been of fundamental importance for the development of language processing tools and solutions with increasingly improved performance and depth of analysis. These tools range from part-of-speech taggers to semantic role labelers, and include named entity recognizers, dependency parsers or lemmatizers, among many others.

The increased sophistication and depth of analysis of these tools has required corpora with increasingly more sophisticated linguistic information. As a consequence, this has lead to increasingly more demanding conditions on the annotation process, not only in terms of the linguistic expertise required from the human annotators, but also in terms of the organization and management of the annotation process (Branco, 2009).

In general, as the information and categories being associated with linguistic items grow in complexity, the concerns about the reliability of the data set increase. Categories with more complex structure to be handled by the annotators typically imply more chances for some parts of them to be incorrectly chosen, and thus more chances for the annotated dataset to be flawed. They also bring more chances for the categories assigned by different annotators to a markable be divergent, and thus for a lower inter-annotator agreement (Artstein and Poesio, 2008).

An extreme case both of the complexity of the information to be assigned and the need and importance of supportive tools can be found in DeepBanks. These are

corpora whose sentences are annotated with fully-fledged deep grammatical representations encompassing all different levels of grammatical dimensions for each sentence (from morphological analysis to meaning representation) (Cotton and Bird, 2002; Open *et al*., 2002, Böhmová *et al*., 2003, Rosén *et al*., 2005).

The complexity of the category to be assigned is such that, in practical terms, it is out of the range of human annotators ability to be able to compose it, even in a piecemeal fashion. In this case, the annotation process has to resort to an annotation tool, a computational grammar, which proposes a number of viable parses out of which the annotator eventually selects the one to be assigned via the selection of parse discriminants that progressively reduce the parse forest and thus the annotation space (Dipper, 2000; Oepen, 1999; Rosén *et al*., 2009; Rosén *et al*., 2012).

3 The collection of treebanks extracted from a DeepBank

As it occurs many times, sophistication comes at a cost, but extra cost may bring extra benefits. That happens also in the case of DeepBanks. The construction of a DeepBank is incomparably much more demanding, in resources and organization effort, than for instance a much simpler POS annotated corpus or even a constituency treebank. Among many other things, it requires a deep processing grammar, whose development is in itself an long term endeavor of non trivial prosecution (e.g. Copestake and Flickinger, 2000; Branco and Costa, 2010). But once the development of an annotated corpus of this highly advanced type is set in motion, one is opening the way to the construction of a resource that brings a range of unique advantages.

First, since deep processing grammars are developed under a thorough grammatical framework (e.g. HPSG), in DeepBanks, sentences are annotated with information that not only is linguistically principled, but that it is also consistent across the sentences in the corpus.[1]

Second, one can extract different "vistas" from a DeepBank: for instance, an extracted data set with sentences annotated with their syntactic constituency trees (a TreeBank); or another one with sentences annotated with those trees decorated with semantic roles (a PropBank), etc. Thus, when one builds a DeepBank, it is as if in practical terms, one is getting several corpora for the cost of one (Silva *et al*., 2012).

Third, while they capture different grammatical dimensions of their sentences, these corpora are fully aligned among themselves as they are built on the same set of raw sentences. Thus, a DeepBank allows for a collection of monolingual corpora that are aligned among each other and that encode different grammatical dimensions.

Against this background, an important line of progression is thus to have corpora that are aligned and that represent not only different grammatical dimensions, but also consistently represent such dimensions across different languages. These certainly are assets of utmost importance to support the training and development of multilingual and machine translation solutions of increased quality.

[1] As way of example, in the Peen Treebank, detecting inconsistencies became a topic of research in itself: see for instance Dickinson and Meurers, 2005.

4 A multilingual collection of aligned DeepBanks

This requires the development of multilingual aligned DeepBanks. This in turn presupposes some non-trivial conditions, among which the most demanding one is perhaps that there exist deep processing grammars for the different languages at stake. To facilitate that the alignment of the multiple dimensions may carry over to the alignment across languages, these grammars are expected to be developed under some similar guidelines, principles or grammatical framework.

An initiative to develop multilingual aligned DeepBanks is under way in the scope of the DELPH-IN consortium (`www.delph-in.net`). This endeavor relies on grammars developed by members of the consortium for different languages. And the aligned DeepBanks are obtained by annotating with those grammars raw texts in different languages, which are aligned among each other (Flickinger *et al*., 2012).

In order to maximize the potential of the possible research produced over these DeepBanks to be comparable to other results reported in the literature, the raw text in English is the one from Penn TreeBank (Marcus *et al*., 1993*)*. In the other languages, the aligned texts result from the translation of that English text into those languages.

The initial languages for which there are aligned DeepBanks being developed under this arrangement are Bulgarian, English, Portuguese and Spanish. More are being prepared to join. The Portuguese version of the raw text entering this multilingual collection, that is the translation of the WSJ corpus, contains over 40,000 sentences. It is the result of the translation of the text in the Penn Treebank by a paid professional translator, a translation that was subsequently submitted to a double checking by two reviewers.

5 DeepBankPT and the *BankPT collection of treebanks

In its current first released version, the DeepBank for Portuguese (DeepBankPT) comprises those sentences that have been annotated so far, out of the ca. 40,000 sentences available to be treebanked. These amount to 3,406 sentences, containing 44,598 tokens. The development of the DeepBankPT dataset resorted to the deep linguistic processing grammar for Portuguese LXGram (Branco and Costa, 2010). The dynamic treebanking was supported by the annotation environment `[incr tsdb()]` (Oepen, 1999).

The grammar produces the admissible grammatical representations, a so called parse forest, for each input sentence to be annotated. The annotation workbench permits the annotators to select one of the parse trees and annotate the sentence with it. The selection of the parse tree is performed by the annotators by setting up the appropriate option ("yes" vs. "no") of the set of so called binary discriminants. These discriminants are associated to the rules of the grammar that were applied and thus supported the different trees in the parse forest.

The DeepBankPT was developed following the widely acknowledged annotation methodology that ensures the best reliability of the dataset produced, namely with a double-blind annotation followed by adjudication. Each sentence was annotated by a pair of expert annotators, graduated in linguistics, working independently of each

other. The adjudication was performed by another expert researcher, with a post-graduation degree in computational linguistics. 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).

Besides this core data set, the collection of *BankPT treebanks includes four other data sets, namely TreeBankPT, PropBankPT, DependencyBankPT and LogicalFormBankPT. These treebanks are extracted from the DeepBankPT, following the procedures described in (Silva and Branco, 2012). They contain parts of the fully-fledged grammatical information contained in the DeepBankPT, displayed along widely acknowledged formats.

In the TreeBankPT, the sentences are associated to their syntactic constituency representations, along the lines of the Penn Treebank (Marcus *et al*., 1993). The linguistic options adopted for this annotated corpora follow the options that were assumed for the CINTIL TreeBank and are described in detail in (Branco *et al*., 2011a). The PropBankPT is an extension of the TreeBankPT where the constituency trees get decorated with semantic roles, along the lines of (Palmer *et al*., 2005). The specific tag set adopted is identical to the one adopted for the CINTIL PropBank, described in (Branco *et al*., 2012).

The DependencyBankPT, in turn, stores the sentences annotated with the representation of the grammatical dependencies among their component words. This dependency bank follows the design options adopted for the CINTIL DependencyBank, presented in (Branco *et al*., 2011b). Finally, in the LogicalFormBankPT, the sentences are associated with their semantic representation encoded with semantic description formalism MRS (Copestake *et al*., 2005).

Each one of the treebanks in the *BankPT collection is distributed free of charge for research purposes through the META-SHARE platform (`www.meta-share.eu`).

7 Concluding remarks

In this paper, we described the new collection of advanced treebanks *BankPT for the Portuguese language, developed around the core treebank DeepBankPT. The key innovative aspect of these corpora relies on the fact that they are the first parallel corpora for Portuguese that are aligned with corpora for several other languages, including English, thus opening the way for advanced research involving Portuguese in multilingual applications, including machine translation.

Acknowledgments

This research was supported by the grant of the European Commission FP7-ICT-2013-10-610516, to the project QTLeap, and by the grant of FCT-Fundação para a Ciência e Tecnologia, PTDC/EEI-SII/1940/2012, to the project DP4LT.

References

1. Artstein, Ron and Massimo Poesio, 2008, "Inter-Coder Agreement for Computational Linguistics", *Computational Linguistics*, 34 (4).

2. Böhmová, Alena, Jan Hajič, Eva Hajičová and Barbora Hladká, 2003, "The Prague Dependency Treebank ", in Anne Abeillé (ed.), *Treebanks*, Kluwer.
3. Castro, Sérgio, 2011, *Developing Reliability Metrics and Validation Tools for Datasets with Deep Linguistic Information*, MA Dissertaion, Universty of Lisbon.
4. Copestake, Ann, Dan Flickinger, Carl Pollard, Ivan A. Sag, 2005, "Minimal Recursion Semantics: An Introduction", *Journal of Research on Language and Computation* 3(4).
5. Copestake, Ann and Dan Flickinger, 2000, "An open-source grammar development environment and broad-coverage English grammar using HPSG". *LREC'2000*.
6. Cotton, Scott and Stephen Bird, 2002, "An Integrated Framework for Treebanks and Multilayer Annotations", in *Proceedings of LREC2002*.
7. Branco, António, Catarina Carvalheiro, Sílvia Pereira, Mariana Avelãs, Clara Pinto, Sara Silveira, Francisco Costa, João Silva, Sérgio Castro, João Graça 2012, "A PropBank for Portuguese: the CINTIL-PropBank", In *Proceedings of LREC2012*.
8. Branco, António, João Silva, Francisco Costa and Sérgio Castro, 2011, CINTIL TreeBank Handbook: Design options for the representation of syntactic constituency. Department of Informatics, University of Lisbon, Technical Reports nb. di-fcul-tp-11-02.
9. Branco António, Sérgio Castro, João Silva, Francisco Costa, 2011, *CINTIL DepBank Handbook: Design options for the representation of grammatical dependencies*. Department of Informatics, University of Lisbon, Technical Reports nb. di-fcul-tr-11-03.
10. 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 ", In *Proceedings of LREC2010*.
11. Branco, António and Francisco Costa, 2010, "A Deep Linguistic Processing Grammar for Portuguese", In *Lecture Notes in Artificial Intelligence*, 6001, pp.86-89, Berlin: Springer.
12. Branco, António, 2009, "LogicalFormBanks, the Next Generation of Semantically Annotated Corpora: key issues in construction methodology", M. Klopotek, A. Przepiorkowski, S. Wierzchón, K. Trojanowski, eds., *Recent Advances in Intelligent Information Systems*, Academic Publishing House EXIT, Warsaw.
13. Dickinson, Markus and Detmar Meurers, 2005, "Detecting Annotation Errors in Spoken Language Corpora". *Proceedings of the Special session on treebanks for spoken language and discourse at the 15th Nordic Conference of Computational Linguistics*.
14. Dipper, Stefanie, (2000), "Grammar-based Corpus Annotation", in P*roceedings of Workshop on Linguistically Interpreted Corpora*.
15. Flickinger, Dan, Valia Kordoni, Yi Zhang, António Branco, Kiril Simov, Petya Osenova, Catarina Carvalheiro, Francisco Costa and Sérgio Castro, 2012, "ParDeepBank: Multiple Parallel Deep Treebanking, Proceedings", *Proceedings of TLT2012*.
16. Flickinger, Dan, Valia Kordoni, and Yi Zhang, 2012, "DeepBank: A Dynamically Annotated Treebank of the Wall Street Journal, Proceedings", *Proceedings of TLT2012*.
17. Marcus, Mitchell, Beatrice Santorini, Mary Ann Marcinkiewicz, 1993, "Building a large annotated corpus of English: The Penn Treebank", *Computational Linguistics*, 19(2).
18. Oepen, Stephan, Dan Flickinger, Kristina Toutanova, Christopher D. Manning, and Thorsten Brants (2002), "The LinGO Redwoods Treebank: Motivation and Preliminary Applications", in *Proceedings of COLING 2002*.
19. Oepen, Stephan, 1999, [incr tsdb()] — Competence and Performance Laboratory. User Manual, Technical Report, Computational Linguistics, Saarland University, Germany.
20. Palmer, Martha, Paul Kingsbury and Daniel Gildea, 2005, "The Proposition Bank: An Annotated Corpus of Semantic Roles", *Computational Linguistics*, 31.
21. Rosén, Victoria, Paul Meurer, Gyri Smørdal Losnegaard, Gunn Inger Lyse, and Koenraad De Smedt, Martha Thunes, and Helge Dyvik, 2012, "An integrated web-based treebank annotation system", *Proceedings of TLT2012*.

22. Rosén, Victoria, Paul Meurer, and Koenraad de Smedt, 2009, "LFG Parsebanker: A Toolkit for Building and Searching a Treebank as a Parsed Corpus". In Frank Van Eynde, Anette Frank, Gertjan van Noord, and Koenraad De Smedt, editors, *Proceedings of TLT7*
23. Rosén, Victoria, Paul Meurer and Koenraad de Smedt, 2005, "Constructing a Parsed Corpus with a Large LFG Grammar", In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG'05 Conference*, CSLI Publications.
24. Silva, João, and António Branco 2012, "Deep, consistent and also useful: Extracting vistas from deep corpora for shallower tasks", In *Proceedings of the Workshop on Advanced Treebanking*, in *Proceedings of LREC2012*.