Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: the CINTIL DeepGramBank

António Branco, Francisco Costa, João Silva, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto, João Graça
University of Lisbon
Edifício C6, Departamento de Informática
Faculdade de Ciências, Universidade de Lisboa
Campo Grande, 1749-016, Portugal
antonio.branco@di.fc.ul.pt, fcosta@di.fc.ul.pt, jsilva@di.fc.ul.pt, sara.silveira@di.fc.ul.pt,
sergio.castro@di.fc.ul.pt, mariana.avelas@di.fc.ul.pt, clara.pinto@di.fc.ul.pt

Abstract

Corpora of sentences annotated with grammatical information have been deployed by extending the basic lexical and morphological data with increasingly complex information, such as phrase constituency, syntactic functions, semantic roles, etc. As these corpora grow in size and the linguistic information to be encoded reaches higher levels of sophistication, the utilization of annotation tools and, above all, supporting computational grammars appear no longer as a matter of convenience but of necessity.

In this paper, we report on the design features, the development conditions and the methodological options of a deep linguistic databank, the CINTIL DeepGramBank. In this corpus, sentences are annotated with fully fledged linguistically informed grammatical representations that are produced by a deep linguistic processing grammar, thus consistently integrating morphological, syntactic and semantic information.

We also report on how such corpus permits to straightforwardly obtain a whole range of past generation annotated corpora (POS, NER and morphology), current generation treebanks (constituency treebanks, dependency banks, propbanks) and next generation databanks (logical form banks) simply by means of a very residual selection/extraction effort to get the appropriate "views" exposing the relevant layers of information.

1. Introduction

An important methodological breakthrough took place in Language Technology with the advent of statistical approaches, which need large data sets for the estimation of relevant stochastic parameters as well as for the evaluation of the corresponding tools. These data sets have steadily grown not only in terms of their size but also in terms of the complexity of the linguistic information they store, as the application of stochastic techniques has moved from relatively shallow (e.g. POS tagging) to more deep processing tasks (e.g. semantic role labeling). Hence, development activities on annotated corpora have been deployed around extending lexical and morphological information with information concerning phrase constituency (aka TreeBanks (Marcus *et al.*, 1993)), with syntactic functions (aka DependencyBanks (Böhmova *et al.*, 2003)), with phrase-level semantic roles (aka PropBanks (Palmer *et al.*, 2005)), etc.

Progressing along this trend, next generations of annotated corpora will expand these annotations with semantic information of different sorts beyond the phrase level, e.g. by including sentence-level representations of meaning (logical forms). Accordingly, the linguistic information to be encoded will reach new levels of sophistication where the utilization of annotation tools and supporting computational grammars (Dipper, 2000, Oepen *et al.*, 2002) will appear no longer as a matter of convenience but of necessity.

To proceed towards the construction of such sophisticated annotated corpora, the annotation tools to be used will not be able to do without auxiliary grammars and lexicons for deep linguistic processing (Bos and Delmonte, 2008) in order:

- to obtain deep, accurate grammatical representations to serve as annotation materials, which can be selected from parse results but cannot be massively and accurately constructed or corrected by hand;
- to bring into the annotation process the benefits of principled linguistic theorizing of a deeper level than the shallow ones that have been put to use in the construction of previous generations of TreeBanks and PropBanks;
- to ensure the correct alignment and integration of annotations pertaining to the different linguistic dimensions and layers (morphology, syntax, semantics, etc.)

The usage of deep linguistic grammars in the construction process of annotated treebanks is becoming an essential move in supporting the progress in the area of Language Resources.

As a side effect, it permits to obtain very important payoffs: as the deep linguistic representation of a sentence may encode as much grammatical information as it is viable to associate to a sentence, by constructing a deep linguistic databank one is producing in tandem, and within the same amount of effort, a POS-tagged corpus, a constituency TreeBank, a DependencyBank, a PropBank, or even a LogicalFormBank.

We have developed a corpus annotated with deep linguistic representations for Portuguese along these design options, the CINTIL DeepGramBank.

Building on the experience gathered, our goal is twofold in this paper. On the one hand we aim at delivering a first report on this new corpus, its composition, size, construction, distribution, etc. On the other hand, we aim also at addressing generic issues worth being clarified concerning the development of this type of corpora for their construction to be practically viable.

In the remainder of this paper, we will focus mostly on the latter aspects (for an account of the first, see the companion metadata to this paper filled in for LREC2010 LRT Map).

2. One linguistically interpreted corpus, many vistas

A number of annotated corpora have been developed that store the parses of deep linguistic grammars (e.g. Simov *et al.*, 2002, Rosén *et al.*, 2005). An issue of concern frequently voiced is that such databanks may end up being too theory-centered, and that the relevant linguistic information is stored in a too theory-specific format. Figure 1 displays the fully-fledged deep representation of a sentence in our DeepGramBank, obtained with the LXGram grammar (Branco and Costa, 2008), which helps to easily visualize the source of this type of concerns and understand them.

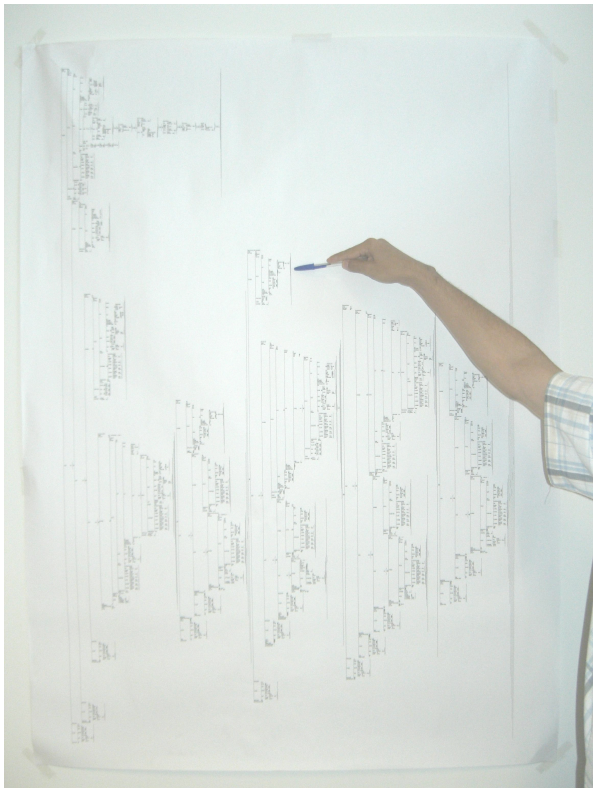


Figure 1: A 1 m x 1.4 m printout in font size 4 of the fully fledged HPSG representation, in the AVM format of (Pollard and Sag, 1994), for the 6 word sentence *Todos os computadores têm um disco* ("Every computer has a disk"). The arm and the pen are included to help appreciate the relative proportion of the whole representation.

We took these criticisms seriously into account, and developed a collection of tools to extract different vistas out of our kernel DeepGram databank. While sharing the same underlying sentences, each vista corresponds thus to a different annotated corpora. Henceforth, a CINTIL Treebank, a CINTIL DependencyBank, a CINTIL Propbank and a CINTIL LogicalFormBank are immediately available. They follow linguistic options that comply with current best practice, are encoded in de facto standards for data formats (viz. Penn Treebank format, etc.), are more manageable and less theory-specific than the kernel databank, and are distributed along with the DeepGramBank.

In the following sections we will be reporting on this collection of extraction/conversion tools and on the various databank vistas they produce.

2.1 CINTIL TreeBank

In the kernel DeepGramBank, punctuation symbols are not detached from words. Hence, one of the main non-trivial tasks that must be performed when extracting the TreeBank vista is to align leaves in the trees with their corresponding tokens in the annotated sentence. After a punctuation symbol is detached from a word, it is inserted into a new node and this node is moved into its appropriate position in the tree. This requires detecting if the punctuation is part of a coordination structure, an adjunct, etc. and act accordingly.

Another step in the extraction of the phrase constituency view consisted in the normalizing of the extracted tree to a basic X-bar representation. As the grammatical representations in the DeepGramBank are produced by a grammar that includes not only syntactic rules but also rules for morphological analysis, these representations may have several unary branches successively expanding each other. The extraction tool took care of collapsing such possible unary nodes.

This tool that extracts CINTIL TreeBank can produce several variants of this vista. These variants are controlled by toggling a set of options on or off. The main toggles are described below:

- Besides the constituency tag, tree nodes in the kernel databank bear information on grammatical function and semantic role. However, if one so wishes, the node tags in the resulting trees can be stripped down in order to include only constituency information.
- Null subjects and ellipsis can be represented in specific categories labelling the nodes or represented through the addition of empty nodes.
- A multi-word expression can be kept as a single string under a single pre-terminal node or expanded into several nodes. In the latter case, it is expanded into a set of sister nodes, one for each token in the multi-word, all bearing the same category as the multi-word.
- Morphological information contained in the annotated sentence—such as POS tag, lemma and inflection features—can be left out of the result or appended to the leaves of the resulting tree.

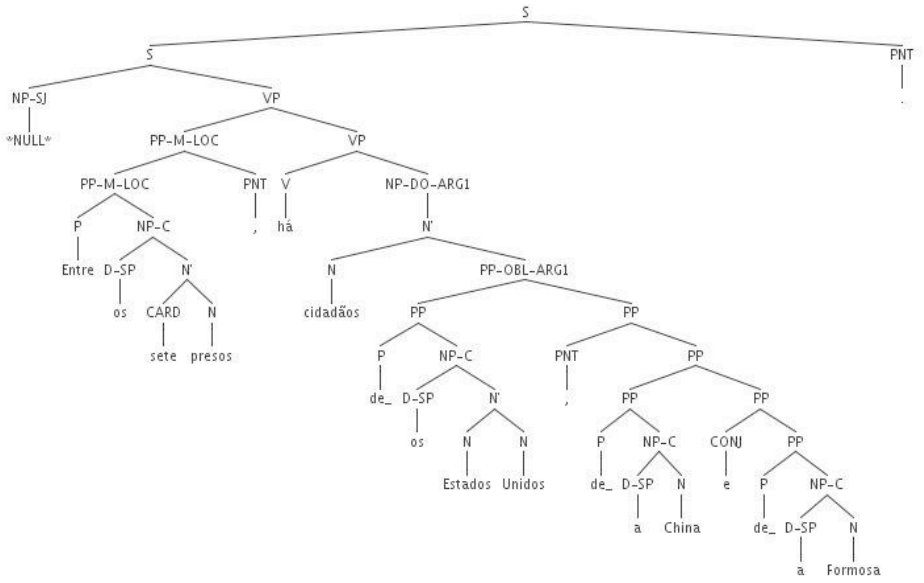


Figure 2: CINTIL TreeBank vista for the sentence
Entre os sete presos, há cidadãos dos Estados Unidos, da China e da Formosa
("Between the seven prisoners, there are citizens of the United States, China and Taiwan")

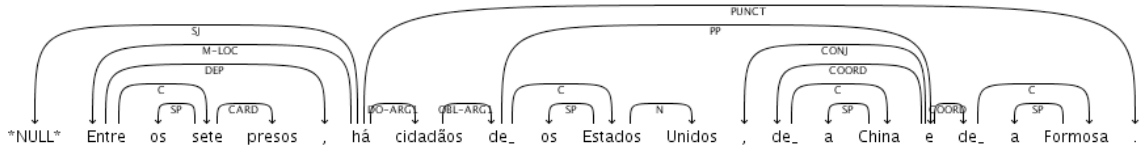


Figure 3: CINTIL DependencyBank vista

This tool makes use of the Tregex/Tsurgeon Java library,[1] created by the Stanford NLP Group, for all the required tree manipulations.

An online search service, running over this vista, can be found at `http://cintiltreebank.di.fc.ul.pt`. This service uses the Tregex syntax to search CINTIL TreeBank and display images of matching parses.

An example parse tree can be seen in Figure 2. It shows a tree where information on grammatical function and semantic role is kept; empty nodes have been explicitly added; multi-word expressions have been expanded (viz. *Estados Unidos* as two nodes); and morphological information from the annotated sentence has been left out.

Where appropriate, long-distance syntactic relations are represented by means of the widespread slashed labels convention in use in mono-level syntactic frameworks.

2.2 CINTIL DependencyBank

The tool that extracts the DependencyBank begins by taking as input the constituency vista—as produced by the tool described in the previous section—in a variant where tree nodes include information on grammatical function. A simple procedure can extract the dependencies from this annotated X-bar structure of the constituency tree. Given any two sister constituents, is be marked with a grammatical function tag while the other will not. The first constituent is then output as being dependent, under that grammatical function, on the second constituent.

CINTIL DependencyBank adheres to the commonly used CoNLL format. This is a tabular format, where each entry corresponds to a word in the sentence, and includes fields for the word's head, POS, lemma, among other data.

For coordination constructions, the first constituent in the coordination is taken as the head. That first constituent then dominates the final conjunction in the construction. This final conjunction then dominates every other constituent in the coordination construction with a COORD relation, and any comma in the coordination with a CONJ relation.

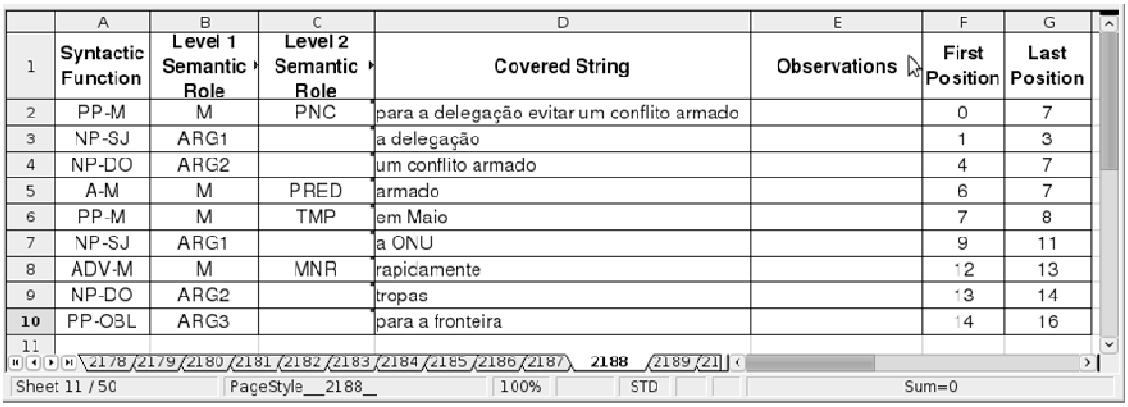
Similarly to what happens with CINTIL TreeBank, this vista can have variants. For instance, if one wishes that the dependency representation include empty categories, and since this tool runs over the TreeBank vista, it suffices to toggle on that option when generating the constituency vista.

The dependency representation corresponding to the constituency tree shown above can be seen in Figure 3.

2.3 CINTIL PropBank

Propbanks are treebanks whose trees have their constituents labelled with semantic role tags. In other words, propbanks are annotated corpora that result from the extension of the annotation associated to the sentences

[1]Tregex: `http://nlp.stanford.edu/software/tregex.shtml`



	A	B	C	D	E	F	G
1	Syntactic Function	Level 1 Semantic Role	Level 2 Semantic Role	Covered String	Observations	First Position	Last Position
2	PP-M	M	PNC	para a delegação evitar um conflito armado		0	7
3	NP-SJ	ARG1		a delegação		1	3
4	NP-DO	ARG2		um conflito armado		4	7
5	A-M	M	PRED	armado		6	7
6	PP-M	M	TMP	em Maio		7	8
7	NP-SJ	ARG1		a ONU		9	11
8	ADV-M	M	MNR	rapidamente		12	13
9	NP-DO	ARG2		tropas		13	14
10	PP-OBL	ARG3		para a fronteira		14	16
11							

Figure 4: Annotation interface for specifying semantic roles for the sentence
Para a delegação evitar um conflito armado em Maio a ONU enviou rapidamente tropas para a fronteira
("In order for the delegation to avoid an armed conflict in May the UN rapidly sent troops to the border")

in treebanks by means of an extra set of tags for semantic roles. Accordingly, propbanking can be seen as an annotation task for the semantic categorization of phrases, where the key issue involves semantic ambiguity resolution.

Some of the semantic role labels that are used in PropBank can be obtained from features that describe the semantics of the sentence in the kernel DeepGramBank, namely those used to tag the subject and the complements of predicators, ARG1 to ARGn. The grammar was expanded in order to add to the output tree these tags.

For the remaining semantic role labels, a completion step follows that consists in the manual specification of the occurrences of the portmanteau tag M in terms of one of the semantic roles available for modifiers in the tagset, LOC, TMP, MNR, etc. This manual annotation is supported by two tools: a converter from trees into an annotation format compatible with the annotation interface, and a reverser tool for the symmetric operation.

The annotation interface is based on a basic yet very efficient and powerful enough technology in view of the manual task it is aimed at supporting. A set of sentences to be annotated is presented in a spreadsheet file with each sentence in a different sheet.

These spreadsheets are created by the converter tool that takes as input an exported version of the treebanked sentences. For each suite of treebanked sentences, a spreadsheet is created with as many sheets as sentences in that suite. If a given sentence happens not to have received a parse, its sheet only contains its identification number and that sentence.

If in turn the sentence received a parse in the treebank, its tree is processed and, for each node with a syntactic function that ends label, a new line in the sheet is printed. Each sentence with a parse will then be annotated by a human annotator that specifies the semantic role tags of the modifiers.

A screenshot of this annotation interface can be seen in Figure 4, showing the result of manual specification in column C of four constituents that bear the portmanteau tag M in column B.

When the manual propbanking is finalized, the sentences—now extended with the newly assigned tags for the semantic roles of modifiers—are reverted to the original tree representation. This operation is ensured by a reverting tool that parses the data in the sheets of the spreadsheet and recombines the new information added by the human annotator with the original information about the parse tree of the sentence.[2]

2.4 CINTIL LogicalFormBank

The kernel DeepGramBank associates fully-fledged grammatical representations to sentences, including the formal representation of their meaning. Minimal Recursion Semantics (Copestake *et al.*, 2005) is used for this representation of meaning.

An MRS representation is a description of a set of possible logic formulas that differ only in the relative scope of the relations present in these formulas. In other words, it supports scope underspecification.

Semantic representations provide an additional level of abstraction, as they completely abstract from word order and language specific grammatical restrictions.

For instance, the fact that the Portuguese verb *gostar* selects for a PP complement and its English counterpart *like* is a transitive verb is not visible in the semantics, since in both cases they correspond to an equivalent binary predicate.

Additionally, the MRS format of semantic representation that is employed is well defined in the sense that it is known how to map between MRS representations and formulas of second order logic, for which there is a set-theoretic interpretation.

3. A dynamic data set

Another important issue is that the sentences that can be included in the annotated corpus are limited to those that the grammar is able to parse. This apparent difficulty deserves to be seriously taken also into account.

[2] For a more detailed account of this annotation environment and process, see (Branco *et al.*, 2009).

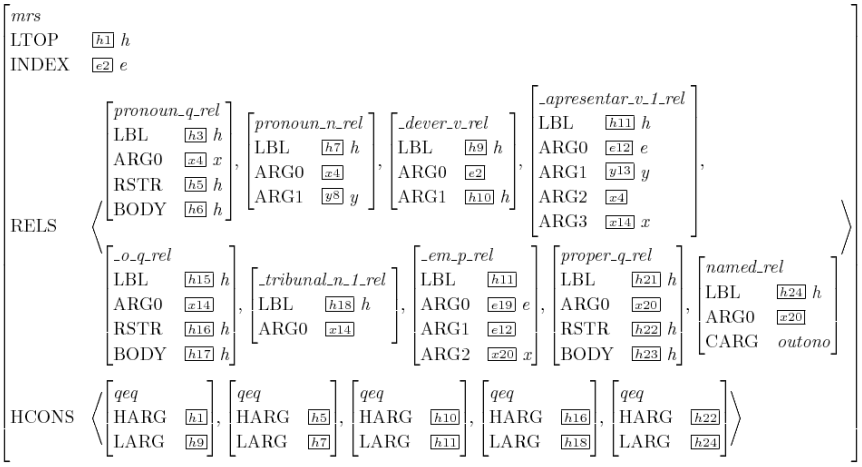


Figure 5: A MRS representation

In this concern, the important aspect to acknowledge is that, from this level of annotation complexity onwards, one has to live with dynamic databanks, as proposed in (Oepen *et al.*, 2002). Given that any computational grammar evolves and gets perfected from version to version, the corresponding annotated corpus will get perfected and enlarged as well.

The key contribution here was to design an annotation methodology and workflow that takes this as a central tenet. Crucially, it is important to avoid the re-annotation by the human annotators of sentences that are already annotated in the previous version of the DeepGramBank and whose parse was not altered by the new version of the grammar. It is also important that a given sentence already treebanked and whose representation is just enlarged by the new version of the grammar need not to go through the whole process of reinspection of its parse forest, sufficing to accept the proposed extension.

This annotation methodology is technically supported by the annotation environment `[incrs tsdb()]` (Oepen, 1999) that we resorted to and, in a nutshell, can be described as a "circular" enhancement of the common TAVA (train, annotate, validate, adjudicate) approach. For a thorough discussion of this enhancement and its implications see (Branco, 2009).

4. Data reliability

Finally, a third important issue relates to the reliability of the data produced. The construction of linguistically interpreted data with increasingly sophisticated annotation have raised concerns regarding the increased level of complexity of the annotation decisions that the annotators are asked to perform and the increased level of subjectivity possibly underlying such decisions. These concerns are motivated by a possible tradeoff that is reasonable to expect, between the higher sophistication of the judgments and the lower reliability of the annotated data produced (Artstein and Poesio, 2008).

The issue here is to find a sensible inter-annotator agreement coefficient that captures the level of coincidence in the decisions made during the annotation process of the DeepGramBank by different annotators, which consists in picking a parse out of a parse forest for each sentence. The few works that sought to take this into account (Brants, 2000; Civit *et al.*, 2003; Fujita *et al.*, 2006), resorted to the Parseval metric to compare the outcome of two different human annotators. This has been felt, however, not to be fully appropriate: what is being compared is the final representations picked up by the annotators, not the decisions they opted for during the parse selection process.

In this respect, we explored the log file with the annotators' decisions provided by the annotation environment `[incrs tsdb()]` (Oepen, 1999) that we resorted to. This tool permits to arrive at the parse with which to annotate a sentence by progressively narrowing down the parse forest by means of basic discriminants between parses. Each such discriminant supports a binary decision by the annotator, allowing him to keep (or reject) those parses for which a certain rule was used (or not used). This permitted to design a more sensible agreement coefficient and empirically verify that the annotation of corpora with deep grammars can be done at a reliable level of confidence, with an inter-annotator agreement score above the 0.8 threshold. A detailed report on this issue will appear in (Branco *et al.*, forth).

5. References

Artstein and Poesio, 2008, "Inter-Coder Agreement for Computational Linguistics*". Computational Linguistics* 34: 555–596.

Civit, Martí, Navarro, Bufí, Ferrández, Marcos, 2003, "Issues on the Syntactic Annotation of Cast3LB", *4th Workshop on Linguistically Interpreted Corpora.* Budapest, EACL.

Böhmova, Hajic, Hajicova and Hladka, 2003, "The Prague Dependency Treebank: Three-Level annotation scenario", in Abeillé (ed.), Treebanks, Kluwer, 103–128.

Bos and Delmonte (eds.), 2008, *Semantics in Text Processing*, College Publications.

Branco, 2009, "LogicalFormBanks, the Next Generation of Semantically Annotated Corpora: key issues in

construction methodology", In *Proceedings of International Joint Conference on Intelligent Information Systems*, Kraków, Poland, June 15–18, 2009.

Branco, Silveira, Castro, Avelãs, Pinto and Costa, 2009, "Dynamic Propbanking with Deep Linguistic Grammars", In *Proceedings of TLT009 - The 8th International Workshop on Treebanks and Linguistic Theories*, Milan, December 4–5, 2009.

Branco and Costa, 2008, "LXGram in the Shared Task "Comparing Semantic Representations" of STEP2008", In Bos and Delmonte (eds.), 299–314.

Branco and Castro, forth., "Inter-Annotator Agreement for Deep Linguistic Annotation".

Brants, 2000, "Inter-Annotator Agreement for a German Newspaper Corpus", LREC 2000.

Copestake, Flickinger, Sag and Pollard, 2005, "Minimal Recursion Semantics: An Introduction", Journal of Research on Language and Computation, 3, 2–3.

Dipper, 2000, "Grammar-based Corpus Annotation". *Workshop. on Linguistically Interpreted Corpora,* 56–64.

Fujita, Bond and Nakaiwa, 2006, "An Implemented Description of Japanese: The Lexeed dictionary and the Hinoki treebank". COLING/ACL2006.

Marcus, Santorini and Marcinkiewicz, 1993, "Building a Large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics*, 19: 313–330.

Oepen, 1999, [incr tsdb()] - Competence and Performance Laboratory, Saarland University.

Oepen, Toutanova, Shieber, Manning, Flickinger and Brants, 2002, "The LinGO Redwoods treebank. Motivation and preliminary applications". COLING2002.

Pollard and Sag, 1994, *Head-driven Phrase Structure Grammar*, CSLI.

Palmer, Gildea and Kingsbury, 2005, "The Proposition Bank: An Annotated Corpus of Semantic Roles", *Computational Linguistics*, 31: 71–106.

Simov, Osenova, Kolkovska, Balabanova, Doikoff, Ivanova, Simov and Kouylekov, 2002, "Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank". LREC 2002, 1729–1736.

Rosén, Meurer, De Smedt, 2005, "Constructing a Parsed Corpus with a Large LFG Grammar". In Butt and Tracy (eds.), *Proceedings of the LFG'05 Conference*, CSLI Publications, 371–387.