**L**inguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, January 2012

# Treebank Annotation in the Light of the Meaning-Text Theory

**Simon Mille**
**Leo Wanner**
**Alicia Burga**

Published by CSLI Publications

**LiLT** volume 7, issue 16 January 2012

# Treebank Annotation in the Light of the Meaning-Text Theory

Simon Mille, *Pompeu Fabra University,* Leo Wanner, *Catalan Institution for Research and Advanced Studies (ICREA), Pompeu Fabra University,* Alicia Burga, *Pompeu Fabra University*

## Abstract

A treebank may contain the annotation of different phenomena such as word order, morphological features, syntactic and semantic relations, etc., which are rather different in their nature. Quite often, the annotation of these phenomena is combined in a single structure, which leads to low-quality training results and is verifiably deficient from a theoretical (linguistic) perspective. We argue that the annotation of corpora requires a well-defined linguistic model which supports multi-level annotation, with one type of phenomenon per level. Our experience with dependency treebanks created or adjusted for surface-oriented natural language generation and based on the Meaning-Text Theory, a multi-level linguistic model, supports this argumentation.

**LiLT** Volume 7, Issue 16, January 2012.
*Treebank Annotation in the Light of the Meaning-Text Theory.*

## 1 Introduction

Over the last years, an increasing number of treebanks became available for training statistical Natural Language Processing (NLP) applications. A treebank may contain the annotation of word order, syntactic dependencies, morphological features, semantic relations, etc.—phenomena that are rather different in their nature. However, quite often, the annotations are agglomerated in a single structure, with no clear frontier between unrelated phenomena. Such a structure is verifiably deficient from the theoretical (linguistic) point of view. It also reduces the quality of the annotated resources, which in its turn hampers the quality of the applications trained on them. Annotation of corpora is of higher quality when a well-defined linguistic model which supports multi-level annotation, with one type of phenomenon per level, is followed. Our experience with dependency treebanks created or adjusted for surface-oriented natural language generation (NLG) buttresses this argumentation.

Dependency treebanks are increasingly popular in NLP applications; see, e.g., Penn TreeBank 3 for English (Marcus et al., 1993), Prague Dependency Treebank 2.0 for Czech (Hajič et al., 2006), Talbanken05 for Swedish (Nilsson et al., 2005), and SynTagRus for Russian (Apresjan et al., 2006). Only a few of them actually use separate levels of annotation, among them PDT 2.0 and the Italian Syntactic-Semantic Treebank (Montemagni et al., 2003). This is due to the fact that most dependency treebanks were meant to be used for syntactic parsing, for which only syntactic and linear order annotations are necessary. It is only very recently that there has been an increasing need for dealing with deeper levels of representation, for instance, in experiments on automatic semantic role labeling (Surdeanu et al., 2008). To respond to this need, the initially purely syntactic corpora were enriched with partial semantic annotation, but without prior discussion on what kind of deep annotation would be appropriate and which phenomena each level of representation should deal with. As a result, semantically enhanced annotations such as Penn Treebank/PropBank (Palmer et al., 2005)/NomBank (Meyers et al., 2004) prove insufficient, for instance, for NLG (Belz et al., 2011). In what follows, we analyze the problems encountered in the common annotation schemata such as the one used in PropBank and show then that the model of the Meaning-Text Theory (MTT) (Mel'čuk, 1988) supports the creation of linguistically sound annotation schemata which serve not only the purpose of analysis but also that of NLG.

## 2 Some problems in common annotation schemata: The case of the Penn Treebank/ PropBank

Currently, several corpora (AnCora, Tiger/Salsa, Chinese Treebank/ PropBank, etc.) are annotated following the annotation schema in the Penn Treebank (PTB)/PropBank (PB) corpus, which serves as the reference corpus regarding size and consistency of annotation. Figure 1 (see next page) illustrates the PTB/PB annotation in the popular CoNLL format. The first column is the position of the units in the sentence, the second holds the superficial form of each unit, the third its lemmatized form, the fourth indicates its POS, the fifth the position of its syntactic governor, the sixth the label of the edge from its governor. In the seventh column, we find the semantic annotation, starting with the semantic status of the unit—semantic predicate ("Y") or not ("_")—, and then, in the eighth column, its disambiguated meaning. The remaining columns, in this case columns nine to thirteen, contain the predicates of the sentence (five predicates ⇒ five columns) in the order they appear. For instance, *companies* is Argument 0 of the second (*gain.02*), the third (*knowledge.01*) and the fourth (*sale.01*) predicates.

Let us point out, in what follows, what we believe to be the main problems of corpora that follow the PTB/PB schema from the linguistic point of view.

### 2.1 *Confusion between levels of representation*

**Edge labels:** The edge labels mix semantics and syntax at the syntactic level and at the semantic level, which has consequences for the clarity and transparence of each tagset.

Some syntactic edge labels in PTB/PB encode semantic information. Thus, the preposition *through* (line 13 in Figure 1) is annotated as MNR of its governor *gain*, i.e., as a circumstantial carrying the meaning of manner. Further tags of this kind are, for instance, TeMPoral, LOCation, and PuRPose. All of these circumstantials behave in English syntactically in the same way; hence, their syntactic annotation should be identical. As a consequence, the tags do not reflect the level of idiosyncrasy of the syntactic analysis. Consider, e.g., the case of the NMOD relation, which links a noun to any modifier, be it a determiner, an adjective, a numeral, a relative or a PP. For example, a numeral can combine with a determiner, but it is impossible to combine two determiners. Syntactic tags should reflect this kind of difference instead of using different relations to annotate constructions with the same syntactic properties (e.g., circumstantials or appositions), based on their divergent meanings. By doing so, PTB/PB fails to offer a clear and

motivated point of view on English syntax.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | He | he | PRP | 5 | SBJ | _ | _ | A0 | _ | _ | _ | _ |
| 2 | and | and | CC | 1 | COORD | _ | _ | _ | _ | _ | _ | _ |
| 3 | Mr. | mr. | NNP | 4 | TITLE | _ | _ | _ | _ | _ | _ | _ |
| 4 | Bologna | bologna | NNP | 2 | CONJ | _ | _ | _ | _ | _ | _ | _ |
| 5 | emphasized | emphasize | VBD | 0 | ROOT | Y | emphasize.01 | _ | _ | _ | _ | _ |
| 6 | that | that | IN | 5 | OBJ | _ | _ | A1 | _ | _ | _ | _ |
| 7 | both | both | DT | 8 | NMOD | _ | _ | _ | _ | _ | _ | _ |
| 8 | companies | company | NNS | 9 | SBJ | _ | _ | _ | A0 | A0 | A0 | _ |
| 9 | would | would | MD | 6 | SUB | _ | _ | _ | AM-MOD | _ | _ | _ |
| 10 | gain | gain | VB | 9 | VC | Y | gain.02 | _ | _ | _ | _ | _ |
| 11 | technological | technological | JJ | 12 | NMOD | _ | _ | _ | _ | _ | _ | _ |
| 12 | knowledge | knowledge | NN | 10 | OBJ | Y | knowledge.01 | _ | A1 | _ | _ | _ |
| 13 | through | through | IN | 10 | MNR | _ | _ | _ | AM-MNR | _ | _ | _ |
| 14 | the | the | DT | 15 | NMOD | _ | _ | _ | _ | _ | _ | _ |
| 15 | sale | sale | NN | 13 | PMOD | Y | sale.01 | _ | _ | _ | _ | _ |
| 16 | of | of | IN | 15 | NMOD | _ | _ | _ | _ | _ | A1 | _ |
| 17 | Gen | gen | NNP | 19 | NAME | _ | _ | _ | _ | _ | _ | _ |
| 18 | - | - | HYPH | 19 | NAME | _ | _ | _ | _ | _ | _ | _ |
| 19 | Probe | probe | NNP | 16 | PMOD | _ | _ | _ | _ | _ | _ | A1 |
| 20 | , | , | , | 19 | P | _ | _ | _ | _ | _ | _ | _ |
| 21 | which | which | WDT | 22 | SBJ | _ | _ | _ | _ | _ | _ | R-A1 |
| 22 | will | will | MD | 19 | NMOD | _ | _ | _ | _ | _ | _ | AM-MOD |
| 23 | expand | expand | VB | 22 | VC | Y | expand.01 | _ | _ | _ | _ | _ |
| 24 | significantly | significantly | RB | 23 | ADV | _ | _ | _ | _ | _ | _ | AM-MNR |
| 25 | . | . | . | 5 | P | _ | _ | _ | _ | _ | _ | _ |

FIGURE 1: PTB/PB annotation of the sentence *"He and Mr. Bologna emphasized that both companies would gain technological knowledge through the sale of Gen-Probe, which will expand significantly [...]."*

In PTB/PB, there is semantics in syntax, but there is also syntax in the semantic annotation in that some edge labels clearly encode syntactic information. For instance, a relation such as AM-MNR in line 24 implies that the adverb *significantly* is a "modifying argument" of the predicate *expand.01*, ignoring the fact that such an adverb is itself a semantic predicate which takes as argument its syntactic governor. Along the same lines, in line 21, the R-A1 relation indicates that the semantic argument is a "relative" argument, in the sense that the relative pronoun is co-first argument of *expand.01*, whereas *expand.01* has only one first argument at the semantic level. AM-... or R-... edges actually reflect the syntactic structure of the sentence, not its semantic structure.

Another confusion induced by the semantic edge label nomenclature is the unjustified distinction between internal and external argument labels, a syntactic notion derived from the Government and Binding framework. According to the PropBank annotation guidelines, "A0 arguments are the arguments which cause the action denoted by the verb, either agentively or not, as well as those which are traditionally classified as experiencers, i.e. the arguments of stative verbs such as *love*, *hate*, *fear*. A1 arguments, on the other hand, are those that change due to external causation, as well as other types of patient-like arguments." (Babko-Malaya, 2005). Thus, *Gen-Probe* is A1 of *expand.01* because it is the entity which "changes due to external causation". As a consequence, A1 sometimes stands for the first argument of a predicate,

but sometimes it is used to annotate a second argument of a predicate (e.g., *knowledge* in line 12). For the sake of consistent and transparent predicate-argument structure, the distinction between A0 and A1 should be abandoned.

**Nodes:** The semantic annotation in PB contains not only semantic but also syntactic nodes. Thus, as already mentioned above, relative pronouns are annotated at the semantic level in spite of being purely syntactic elements, as are all pronouns (they have no own meaning since their antecedent carries it). Similarly, syntactically governed prepositions or conjunctions such as *that* and *of*, in lines 6 and 16 of Figure 1, receive a semantic arc, whereas the actual semantic arguments are *gain.02* and *Gen-Probe*, respectively.[1]

### 2.2 *Incompleteness of annotations*

At the syntactic and semantic levels, the annotation of PTB/PB is furthermore incomplete. This is partly due to the confusions mentioned in the previous subsection, but also due to the adopted annotation policies. For instance, the semantic annotation does not form a connected structure because only nominal and verbal predicates are annotated for the moment. This is a problem from the perspective of NLG since the algorithms generating from semantic representations must be able to navigate through an entire structure, which is impossible if some nodes are disconnected. However, this choice is understandable since the other semantic predicates (adjectives, adverbs, numbers, etc.) can be identified in the syntactic structure together with their arguments, which generally are their syntactic governors. That is, if a connected semantic structure is required, it can be obtained in an extra mapping step. However, the problem is that this is not trivial, nor will the final structure be flawless (Wanner et al., submitted).

From the perspective of NLG, the PTB/PB annotation also lacks two important types of data: information and coreferential structures. Information structure features such as theme/rheme, perspective, emphasis, given/new, etc. (see Mel'čuk, 2001) are crucial for NLG since they directly influence the syntactic organization of sentences. They can be derived only partially from the syntactic annotation (Wanner et al., submitted)—which is why they should be explicitly annotated

[1] One interesting example from the Spanish corpus AnCora shows a hybrid annotation of morphology and syntax: a combination such as *comerlo* lit. 'eat.it', a very productive construction in Spanish, appears as one single node in the syntactic representation, while it should be split into two functional nodes, the verb and the clitic object pronoun. PTB/PB actually split those morphological groupings: *don't=do+n't*.

on the semantic layer.

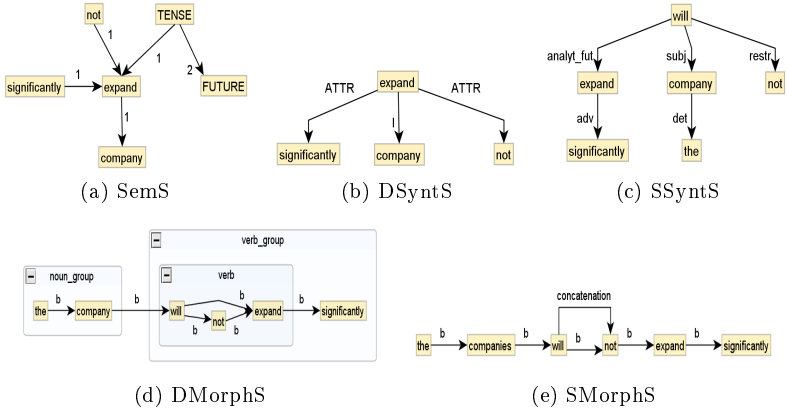
Coreferential structure controls pronominalization at the syntactic level. In PTB/PB, it is only introduced for relative pronouns; cf. *which* in Figure 1. For instance, in the sentence *The Japanese government has stated that it wants 10% to 11% of its gross national product to come from biotechnology products*, the two pronouns *it* and *its* are annotated as arguments of *wants* and *product* respectively. In both cases, the argument should be *the Japanese government*, but due to the introduction of syntactic nodes at the semantic level this is not how it is done. A coreference structure which not only links a pronoun with its antecedent but also nouns that co-refer would allow for retrieving this information.

## 3 The MTT model detailed

The MTT model supports fine-grained annotation at the three main levels of the linguistic description of written language: semantics, syntax and morphology, while facilitating a coherent transition between them via intermediate levels of deep-syntax and deep-morphology.[2] In total, thus five strata are foreseen. At each stratum, a clearly defined type of linguistic phenomena is described in terms of distinct dependency structures. **Semantic Structures** (SemSs) are predicate-argument structures in which the relations between predicates and their arguments are numbered in accordance with the order of the arguments. **Deep-syntactic structures** (DSyntSs) are actually the closest to the PropBank annotation: they are dependency trees, with the nodes labelled by meaningful ("deep") lexical units (LUs) and the edges by actant relations I, II, III, ..., VI (in accordance with the syntactic valency pattern of the governing LU) or one of the following three circumstantial relations: ATTR(ibute), COORD(ination), APPOS(ition). **Surface-Syntactic Structures** (SSyntSs) are dependency trees in which the nodes are labelled by open or closed class lexemes and the edges by grammatical function relations of the type 'subject', 'oblique_object', 'adverbial', 'modifier', etc. **Deep-Morphological Structures** (DMorphSs) are chains of lexemes in their base form (with inflectional and POS features being associated to them in terms of attribute-feature pairs) between which the precedence relation 'b(efore)' is defined and which are grouped in terms of constituents. **Surface-Morphological Structures** (SMorphSs) are chains of inflected word forms, i.e., sentences as they appear in the corpus, except that orthographic contractions still did not take place.

[2]Such a coherent smooth transition is especially relevant to NLG.

For illustration, consider in Figure 2 the representation of the sentence *The companies won't expand significantly* for each MTT-level.[3]




(a) SemS (b) DSyntS (c) SSyntS

(d) DMorphS (e) SMorphS

FIGURE 2: The variety of linguistic structures in a MTT-model

## 4 Meaning-Text Theory and multi-level corpus annotation

As became clear above, MTT offers a linguistically justified formal description of each layer of representation, each of them annotated following strict and independent principles. The rich stratification facilitates a clear separation of different types of linguistic phenomena and thus a straightforward handling for various NLP-applications. However, this is not to say that the MTT annotation is the only possible one. For instance, the *t-layer* in the Prague Dependency Treebank corresponds, roughly, to MTT's DSyntS+SemS; its *a-layer*, to MTT's SSyntS+DMorphS; and its *m-layer*, to MTT's SMorphS. Another possible theory candidate for multilevel annotation is Lexical Functional Grammar (LFG). However, LFG's two main structures— f- and c-structure— are complementary and of the same abstraction, while in MTT all levels differ with respect to their abstraction of the linguistic description. This MTT differentiation can be an advantage from the viewpoint of NLG. In any case, equivalent annotations from other theoretical frameworks can be easily derived from MTT representations, which is why we believe that MTT has considerable advantages. But this is just half of the story. The MTT model is a transductive model (Kahane, 2003). This means that it also provides the instruments

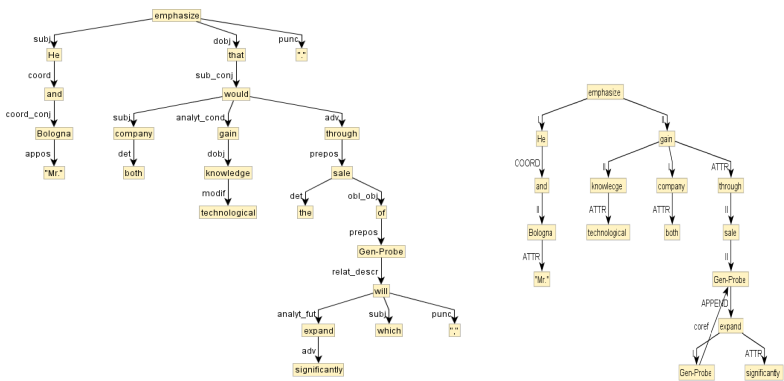
[3]For better readability, we show the structures as trees rather than in a one-word-per-line format.

for mapping the representation at a given level to representations at the adjacent levels, which has two interesting consequences as far as corpus annotation is concerned:

- Annotating two adjacent strata makes the automatic derivation of a broad-coverage mapping grammar for generation or analysis between these two strata possible; such mapping grammars are an essential component of MTT-based text generation, parsing, paraphrasing, and machine translation systems.
- Starting from a given stratum and a manually created mapping grammar (whose coverage does not need to be broad at first), the annotations at the adjacent strata can be easily obtained and then be used to derive the annotations at the next strata. That is, with an SSyntS-treebank, it is straightforward to derive parallel DSyntS- and SemS-treebanks using an adequate tool, namely a graph transducer such as MATE (Bohnet et al., 2000); see Mille et al. (2009).

This last point is particularly relevant, given that corpus annotation is an extremely demanding task; the process of annotation can be reduced to a minimal manual revision of automatically created structures as those in Figure 3. In the following, we illustrate this with the sentence shown in Figure 1 above.

In Figure 3a, the syntactic relations indicate exclusively syntactic properties: the edge *adv*, for instance, does not give any information concerning the meaning of the adverbial group since the meaning of 'manner' is encoded in the adverb itself (*significantly*) or the preposition (*through*).




(a) SSyntS (b) DSyntS

FIGURE 3: SSyntS and automatically derived DSynt annotation (same sentence as in Fig.1)

The syntactic relation *det* is differentiated from the relation *modif*, since both do not have the same combinatorial (see Section 2.1) or topological properties: no dependent of the noun can appear before the determiner, but this is possible for a modifier. The relation *oblique_object* indicates the presence of a governed preposition.

All relations have a direct correlation with Deep-Syntactic relations (Figure 3b): the *adverbials* and *modifiers* always correspond to ATTRibutes, the *subject* of an active verb is always its first actant, the *direct object* is the second, an *oblique object* is, most often, the second actant as well, etc.

Figure 4 shows a sample mapping rule from the mapping grammar. It maps the edge *obl_obj* to *II* while removing the governed preposition (*?Xl*):

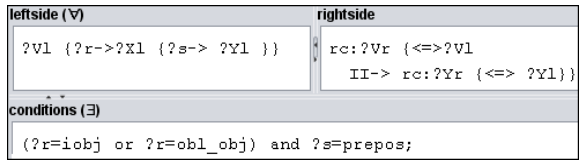



FIGURE 4: Sample mapping rule for graph transducer

The DSyntS in Figure 3b has required no manual modification after its automatic derivation, although this is not always the case.

Now that all grammatical units have been removed from the structure, the mapping to a pure predicate-argument structure is much easier, and another mapping grammar can transform (3b) into a SemS; see Figure 5.

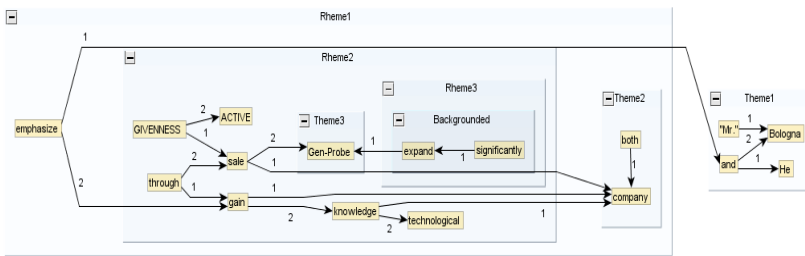



FIGURE 5: A derived semantic annotation (manually revised)

Unlike the semantic annotation of PTB/PB, the semantic structure in MTT has transparent semantic frames in the sense that no difference is made between external or internal arguments.

The structures at all levels have an information structure superimposed on them. Consider, for illustration, the information structure at the semantic level in Figure 5. It indicates what the sentence is about (Theme) and what is said about that (Rheme). This information dimension is recursive since a Theme or a Rheme can contain another Theme-Rheme opposition, which corresponds to an embedded sentence at the syntactic level. A perspective can be associated to a subgraph: 'significantly –1→ expand' is backgrounded, which will trigger the realization of a descriptive relative clause in the final sentence. The definite determiner *the*, which appears in the SSyntS as a node label and in the DSyntS as an attribute-value pair on the node of the noun *sale*, signals, according to Gundel's (1988) hierarchy of Givenness, that *sale* is "activated in the memory of both the Speaker and the Addressee". In Figure 5, this is expressed by a GIVENNESS predicate and its second argument ACTIVE. The case of Givenness illustrates well the fact that the meaning-oriented nature of SemS enables semantic inferences that syntactic structures do not directly allow and highlights the importance that stratification can have in a linguistic framework.

For the semantic annotation, most relations can be derived in a straightforward way: Roman numerals map to Arabic numerals, and ATTR, APPEND and COORD edges are inverted and relabelled with '1'. However, there is more manual workload needed because not all semantic links are expressed in DSyntS (e.g., 'knowledge –1→company'), and the information structure cannot always be deduced from the DSyntS and SSyntS.

## 5 Conclusions and future work

As Hajičová and Sgall (2006), we believe that corpus annotation is useful not only from the point of view of resource creation. The annotation of a corpus within the framework of a linguistic theory represents a large-scale test of this theory. Together with the development of statistical NLP tools, this is even truer, since the applicability of the theory can actually be witnessed. We are convinced that a theoretical framework such as MTT will prove very efficient when it comes to the development of those tools, largely due to its stratification and formal basis. The results of the 2011 Generation Shared task, in which the best scoring system uses an MTT-stratified model (Bohnet et al., 2011) confirms this assumption. To progress in this direction, we are currently working on the annotation of corpora for several languages (including English, Finnish and Spanish) with MTT structures.

## Acknowledgments

We would like to thank the three anonymous reviewers for their insightful comments. This work was partially supported by the European Commission under the contract numbers FP7-ICT-248594 and FP7-SME-286639 and by the Spanish Ministry of Science and Innovation under the contract number FFI2008-06479-C02-02.

## References

Apresjan, Juri, Igor Boguslavsky, Boris Iomdin, Leonid Iomdin, Andrei Sannikov, and Victor Sizov. 2006. A syntactically and semantically tagged corpus of Russian: State of the art and prospects. In *Proceedings of LREC 2006*, pages 1378–1381.

Babko-Malaya, Olga. 2005. *Propbank Annotation Guidelines*.

Belz, Anja, Mike White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 217–226.

Bohnet, Bernd, Andreas Langjahr, and Leo Wanner. 2000. A development environment for an MTT-based sentence generator. In *Proceedings of the First International Conference on Natural Language Generation*, pages 260–263.

Bohnet, Bernd, Simon Mille, Benoît, and Leo Wanner. 2011. StuMaBa: From deep representation to surface. In *Proceedings of ENLG 2011, Surface-Generation Shared Task*. Nancy, France.

Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, and Zdeněk Žabokrtský. 2006. Prague Dependency Treebank 2.0.

Hajičová, Eva and Petr Sgall. 2006. Corpus annotation as a test of a linguistic theory. In *Proceedings of LREC 2006*, pages 879–884.

Kahane, Sylvain. 2003. The Meaning-Text Theory. In *Dependency and Valency. Handbooks of Linguistics and Communication Sciences*, vol. 1-2. De Gruyter.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2):313–330.

Mel'čuk, Igor. 1988. *Dependency Syntax: Theory and Practice*. Albany: State University of New York Press.

Mel'čuk, Igor. 2001. *Communicative Organization in Natural Language: The Semantic-Communicative Structure of Sentences*. Philadelphia: John Benjamins.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank

Project: An interim report. In *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*.

Mille, Simon, Leo Wanner, Vanesa Vidal, and Alicia Burga. 2009. Towards a rich dependency annotation of Spanish corpora. In *Proceedings of SEPLN 2009*. San Sebastian, Spain.

Montemagni, S., F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, and R. Delmonte. 2003. Building the Italian syntactic-semantic treebank. In A. Abeillé, ed., *Building and Using Syntactically Annotated Corpora*, pages 189–210.

Nilsson, Jens, Johan Hall, and Joakim Nivre. 2005. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. In *Proceedings of NODALIDA 2005*, pages 119–132.

Palmer, Martha, Paul Kingsbury, and Daniel Gildea. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31.

Surdeanu, Mihai, Richard Johansson, Adam Meyers, Lluis Márquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of CoNLL 2008*, pages 159–177.

Wanner, Leo, Simon Mille, and Bernd Bohnet. submitted. Do we need new semantic corpus annotation policies for deep statistical generation? .