Annotation and Disambiguation of English Compound Units in the English DeepBank

Valia Kordoni

Department of English and American Studies
Humboldt-Universität zu Berlin
E-mail: evangelia.kordoni@anglistik.hu-berlin.de

Abstract

The aim of this paper is twofold. We focus, on the one hand, on the task of dynamically annotating English compound nouns, and on the other hand we propose disambiguation methods and techniques which facilitate the annotation task. These annotations are very rich linguistically, since apart from syntax they also incorporate semantics, which does not only ensure that the treebank is guaranteed to be a truly sharable, re-usable and multi-functional linguistic resource, but also calls for the necessity of a better disambiguation of the internal (syntactic) structure of larger units of words, such as compound units, since this has an impact on the representation of their meaning, which is of utmost interest if the linguistic annotation of a given corpus is to be further understood as the practice of adding interpretative linguistic information of the highest quality in order to give "added value" to the corpus.

1 Introduction

Disambiguating compounds is a challenging task for several reasons. The first challenge lies in the fact that the formation of compounds is highly productive. This is not only true for English, but for most languages in which compounds are found. Secondly, both the annotation and the disambiguation of compounds is particularly tricky in English, for there are no syntactic and hardly any morphological cues indicating the relation between the nouns: as has very often to date been proposed in the relevant literature, the nouns are connected by an implicit semantic relation. Being a true Natural Language Processing task, the third difficulty in compound noun annotation and disambiguation lies in ambiguity. One could say that compound nouns are double ambiguous: a compound may have more than one possible implicit relation. Therefore, the interpretation of the compound may also depend on context and pragmatic factors. The last main challenge lies in the fact that, even though finite sets of possible relations have been proposed (by among

others [12], [25]), there is no agreement on the number and nature of semantic relations that may be found in compounds. Since [6], it is generally assumed that theoretically, the number of possible semantic relations is infinite.

The annotation and disambiguation of the English compound units we focus on here are part of the English DeepBank, an on-going project whose aim is to produce rich syntactic and semantic annotations for the 25 Wall Street Journal (WSJ) sections included in the Penn Treebank (PTB: [14]). The annotations are for the most part produced by manual disambiguation of parses licensed by the English Resource Grammar (ERG: [7]), which is a hand-written, broad-coverage grammar for English in the framework of Head-driven Phrase Structure Grammar (HPSG: [18]).

The aim of the DeepBank project [8], has been to overcome some of the limitations and shortcomings which are inherent in manual corpus annotation efforts, such as the German Negra/Tiger Treebank ([1]), the Prague Dependency Treebank ([9]), and the TüBa-D/Z.[1] All of these have stimulated research in various sub-fields of computational linguistics where corpus-based empirical methods are used, but at a high cost of development and with limits on the level of detail in the syntactic and semantic annotations that can be consistently sustained. The central difference in the DeepBank approach is to adopt the *dynamic* treebanking methodology of Redwoods [17], which uses a grammar to produce full candidate analyses, and has human annotators disambiguate to identify and record the correct analyses, with the disambiguation choices recorded at the granularity of constituent words and phrases. This localized disambiguation enables the treebank annotations to be repeatedly refined by making corrections and improvements to the grammar, with the changes then projected throughout the treebank by reparsing the corpus and re-applying the disambiguation choices, with a relatively small number of new disambiguation choices left for manual disambiguation.

2 Annotation of Compound Units

The annotation of DeepBank as a whole, and thus also of the compounds in this text collection, is organised into iterations of parsing, treebanking, error analysis, and gramma/treebank update cycles.

2.1 Parsing

Each section of the WSJ corpus is first parsed with the PET unification-based parser [3] using the ERG, with lexical entries for unknown words added on the fly based on a conventional part-of-speech tagger, TNT [2]. Analyses are ranked using a maximum-entropy model built using the `TADM` [13] package, originally trained on out-of-domain treebanked data, and later improved in accuracy for this task by including a portion of the DeepBank itself for training data. A maximum of 500

[1] http://www.sfs.nphil.uni-tuebingen.de/en_tuebadz.shtml

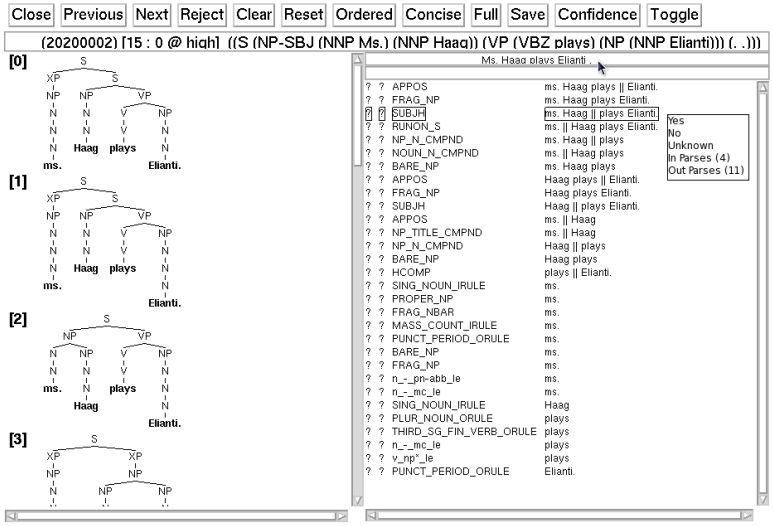


Figure 1: Treebanking Interface with an example sentence, candidate readings, discriminants and the MRS. The top row of the interface is occupied by a list of functional buttons, followed by a line indicating the sentence ID, number of remaining readings, number of eliminated readings, annotator confidence level, and the original PTB bracket annotation. The left part displays the candidate readings, and their corresponding IDs (ranked by the disambiguation model). The right part lists all the discriminants among the remaining readings. The lower part shows the MRS of one candicate reading.

highest-ranking analyses are recorded for each sentence, with this limit motivated both by practical constraints on data storage costs for each parse forest and by the processing capacity of the `[incr tsdb()]` treebanking tool [16]. The existing parse-ranking model has proven to be accurate enough to ensure that the desired analysis is almost always in these top 500 readings if it is licensed by the grammar at all. For each analysis in each parse forest, we record the exact derivation tree, which identifies the specific lexical entries and the lexical and syntactic rules applied to license that analysis, comprising a complete 'recipe' sufficient to reconstruct the full feature structure given the relevant version of the grammar. This approach enables relatively efficient storage of each parse forest without any loss of detail.

2.2 Treebanking

For each sentence of the corpus, the parsing results are then manually disambiguated by the human annotators, using the `[incr tsdb()]` treebanking tool which presents the annotator with a set of binary decisions, called *discriminants*, on the inclusion or exclusion of candidate lexical or phrasal elements for the desired analysis. (cf., Figure 1).

This discriminant-based approach of [5] enables rapid reduction of the parse

forest to either the single desired analysis, or to rejection of the whole forest for sentences where the grammar has failed to propose a viable analysis.[2] On average, given n candidate trees, $\log_2 n$ decisions are needed in order to fully disambiguate the parse forest for a sentence. Given that we set a limit of 500 candidate readings per sentence, full disambiguation of a newly parsed sentence averages no more than 9 decisions, which enables a careful annotator to sustain a treebanking rate of 30 to 50 sentences per hour on the first pass through the corpus.

2.3 Error analysis

During the course of this annotation effort, several annotators have been trained and assigned to carry out the initial treebanking of portions of the WSJ corpus, with most sections singly annotated. On successive passes through the treebank, two types of errors are identified and dealt with: mistakes or inconsistencies of annotation, and shortcomings of the grammar such that the desired analysis for a given sentence was not yet available in the parse forest. Errors in annotation include mistakes in constituent boundaries, in lexical choice such as verb valency or even basic part of speech, and in phrasal structures such as the level of attachment of modifiers or the grouping of conjuncts in a coordinated phrase. Our calculation of the inter-annotator agreeemnt using the Cohen's KAPPA[4] on the constituents of the derivation trees after the initial round of treebanking shows a moderate agreement level at $\kappa = 0.6$. Such disagreeements are identified for correction both by systematic review of the recorded 'correct' trees section by section, and by searching through the treebank for specific identifiers of constructions or lexical entries known to be relatively rare in the WSJ, such as the rules admitting questions or imperative clauses.

Shortcomings of the grammar are identified by examining sentences for which annotators did not record a correct analysis, either because no analysis was assigned, or because all of the top 500 candidate analyses were flawed. Some of the sources of error emerge quickly from even cursory analysis, such as the initial absence of a correct treatment in the ERG for measure phrases used as verbal modifiers, which are frequent in the WSJ corpus, as in *the index rose 20 points* or *the market fell 14%*. Other types of errors required more detailed analysis, such as missing lexical entries for some nouns taking verbal complements, as in *the news that Smith was hired* or *the temptation to spend the money*. These fine-grained lexical entries are not correctly predicted on the fly using the part-of-speech tagger, and hence must be added to the 35,000-entry manually supplied lexicon in the ERG.

[2]For some sentences, an annotator may be unsure about the correctness of the best available analysis, in which case the analysis can still be recorded in the treebank, but with a lower 'confidence' score assigned, so the annotation can be reviewed in a later cycle of updates.

2.4 Grammar & Treebank Update

While grammar development proceeds independent of the initial treebank annotation process, we have periodically incorporated improvements to the grammar into the treebank annotation cycle. When a grammar update is incorporated, the treebank also gets updated accordingly by (i) parsing anew all of the sentences in the corpus using the new grammar; (ii) re-applying the recorded annotation decisions; and (iii) annotating those sentences which are not fully disambiguated after step ii, either because new ambiguity was introduced by the grammar changes, or because a sentence which previously failed to parse now does. The extra manual annotation effort in treebank update is relatively small when compared to the first round of annotation, typically requiring one or two additional decisions for some 5–10% of the previously recorded correct analyses, and new annotation for previously rejected items, which were another 15% of the total in the second round, and much less in successive rounds. Hence these later rounds of updating the treebank proceed more quickly than the initial round of annotation.

Correcting errors of both classes based on analysis of the first pass through DeepBank annotation has resulted in a significant improvement in coverage and accuracy for the ERG over the WSJ corpus. Raw coverage has risen by some 10% from the first pass and the 'survival' rate of successfully treebanked sentences has risen even more dramatically to more than 80% of all sentences in the first 16 sections of the WSJ that have now gone through two rounds of grammar/treebank updates.

3 Examples of Compound Units in DeepBank

Being a collection of financial articles, the WSJ may not represent the English language in its most typical daily usage, but it is not in short of interesting linguistic phenomena. Having an average sentence length of over 20 words, loaded with tons of jargons in the financial domain, the corpus puts many natural language processing components (POS taggers, chunkers, NE recognizers, parsers) to the ultimate test. On the other hand, rich phenomena included in the corpus make it also interesting to test deep linguistic processing techniques. One particularly frequent and puzzling phenomenon in the corpus is the vast amount of compound nouns whose syntactic and semantic analyses are potentially ambiguous. Being symbolic systems, deep grammars like the ERG will not always disambiguate all the possibilities. For example, for the compound *"luxury auto maker"*, the ERG will assign both left-branching and right branching analyses (as shown in Figure 2), using the very unrestricted compounding rule NOUN-N-CMPND.

In some cases such branching decisions seem arbitrary and are defensible either way, but there are instances where a distinction should be made clearly. Consider the following two sentences from the WSJ section 3 of the PTB:

- *A form of asbestos once used to make **Kent cigarette filters** has caused a*

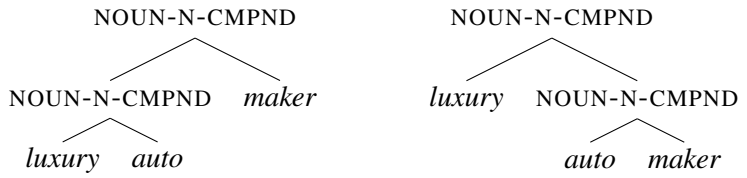


Figure 2: Two alternative analyses from the ERG

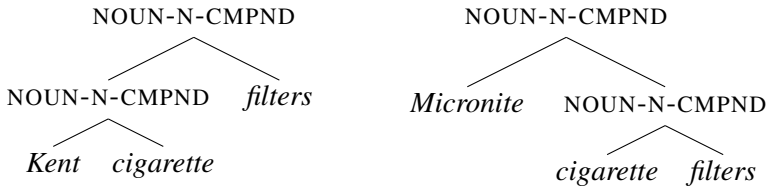


Figure 3: Similar noun compounds with different branching preferences

high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago, researchers reported.

- *Lorillard Inc., the unit of New York-based Loews Corp. that makes Kent cigarettes , stopped using crocidolite in its* ***Micronite cigarette filters*** *in 1956.*

In some cases, such branching preferences can be easily accounted for, if part of the compound is a multiword named entity, as in "Fortune 500 executives" and "auto maker Mazda Motor Corp.", where the words from the named entity should be grouped together.

More challenging cases come from the financial domain specific terminologies. While the majority of such terminologies conform to the largely right-branching structures of English, there are cases where left-branching structures may not be excluded in the analysis of the given compounds.

- *Nevertheless , said Brenda Malizia Negus, editor of Money Fund Report , yields " may blip up again before they blip down " because of recent rises in* ***short-term interest rates****.*
- *Newsweek said it will introduce the* ***Circulation Credit Plan*** *, which awards space credits to advertisers on "renewal advertising."*

While varying branching preference can hopefully be recovered partially by a statistical disambiguation model trained on the increasing number of manually disambiguated compounds in the treebanking project, there are also problems which need special treatment in the design of features for the disambiguation model. For instance, in a compound construction containing a deverbal noun, the predicate-argument relation from the deverbal noun to the other noun in the compound is

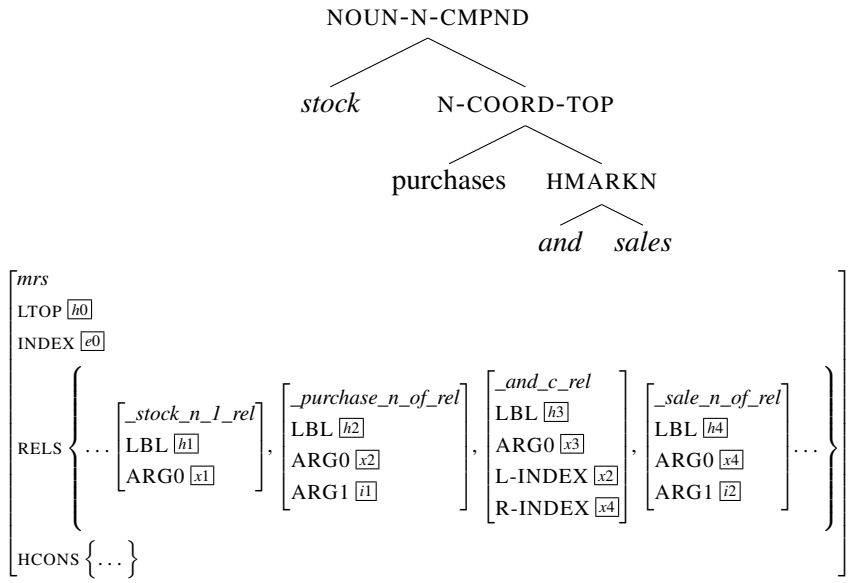


Figure 4: Missing semantic relation within a compound

left underspecified by the grammar, for the relation can be either an argument or a modifier. Consider the compound *"stock purchase and sales"*. A valid syntactic analysis (as shown in Figure 4) leaves an unbound semantic relation.

Ideally, in this example the semantic variables *i1* and *i2* should be both bound to *x1*. But resolving such an ambiguity within the grammar involves the risk of wrongly assigning the semantic roles in cases where, say, the first noun is serving as modifier instead of argument of the deverbal noun. The current disambiguation model does not recover such a kind of underspecified semantic information, as the model is trained exclusively on disambiguated treebanked data with underspecified semantics unchanged. Furthermore, such disambiguation requires a big number of bi-lexical preferences, in order, for instance, for the distinction between arguments and modifiers to be drawn clearly.

4 Disambiguation of Compound Nouns

Due to the lack of constraints on compound nouns in the ERG, the grammar tends to generate all possible internal structures to these NPs, leading to an combinatorial explosion to the number of candidate trees. Working with the DeepBank, we delay the decision on these internal structures of compounds until the other parts of the syntactic structures are disambiguated. Then the annotators go on to pick the preferred branching structures in line with the examples shown in the previous section.

The human annotators have been assisted with several disambiguation models

that help to rank the readings and treebanking decisions. The annotators have been warned to make use of this help with cautiousness. The inter-annotator agreements have been checked periodically to ensure the quality of the annotation.

In need to further facilitate and boost the performance of the parse disambiguation model used for the annotation of compounds in the DeepBank, we have also adopted the following strategies:

1. Almost in a preprocessing component manner, we have relied on the detection and evaluation methods for the automatic acquisition of Multiword Expressions (MWEs), thus also of compound nouns, for robust grammar engineering proposed in [24]. That is, we have first detected compound noun candidates for English on the basis of the distinct statistical properties of their component words, regardless of their type, comparing 3 statistical measures: mutual information (MI), χ^2 and permutation entropy (PE). Then we have validated the quality of such candidates against various corpora, investigating simultaneously the influence of the size and quality of different corpora, using the BNC and the Web search engines Google and Yahoo. At the end of this process, the eligible compound noun candidates have been introduced to the ERG-based parsing and treebanking procedure with the aim to also get validated by the human annotators before ultimately being used for the re-training of the parse disambiguation model.

2. That is, in a novel manner, we have incorporated the fine-grained treebanking decisions made by the human annotators as discriminative features for the automatic parse disambiguation of the compounds in the DeepBank. The advantage of such an approach and everyday treebanking practice is that use of human judgements is made. [11] show that annotators tend to start with the decisions with the most certainty, and delay the "hard" decisions as much as possible. As the decision process goes, many of the "hard" discriminants pertaining to compounds have received an inferred value from the certain decisions. This greedy approach has been shown to help guarantee high inter-annotator agreement. Concerning the statistical parse selection model, the discriminative nature of such treebanking decisions suggests that they are highly effective features, and if properly used, they contribute to an efficient disambiguation model.

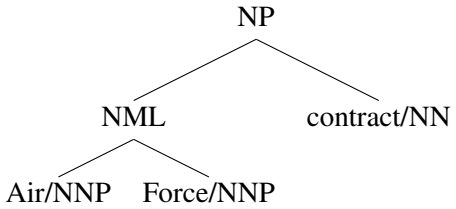
3. Use the annotated sections of the WSJ to retrain the parse disambiguation model and improve the syntactic bracketing prediction accuracy. The parse disambiguation model used here is that proposed in [20] and [19] which has been developed for use with so-called dynamic treebanking environments, like the Redwoods treebank [17]. In such a model, features such as local configurations (i.e., local sub-trees), grandparents, n-grams, etc., are extracted from all trees and used to build and (re-)train the model. Thus, as part of this procedure for our purposes, the eligible compound noun candidates have been introduced to the ERG-based parsing and treebanking procedure and

they have been validated through annotation by the human annotators before ultimately being used for the re-training of the inherent to the parser disambiguation model. The ultimate aim here has been to incorporate the fine-grained treebanking decisions made by the human annotators as discriminative features for the automatic parse disambiguation of the compounds in the DeepBank.

4. Use external large corpora to gather bi-lexical preference information as auxiliary features for the maximum-entropy based parse disambiguation model mentioned above. This is similar to the approach taken in [10] and [23], where pointwise mutual information association scores are used in order to measure the strength of selectional restrictions and their contribution to parse disambiguation. Because the association scores are estimated on the basis of a large corpus that is parsed by a parser and is aimed at getting improved through parse disambiguations, this technique may be described as a particular instance of self-training, which has been shown in the literature to serve as a successful variant of self-learning for parsing, as well. The idea that selection restrictions may be useful for parsing is not new. In our case at hand, i.e., the case of the disambiguation of compound nouns that we are interested in here, our approach and method is very much fine-tuned and targeted to the disambiguation of argument vs. modifier relations in the compound nouns.

5 Discussion and Outlook

In the work of [21], efforts to enrich the noun phrase annotations for the Penn Treebank have been reported. The extra binarization of the originally flat NP structures provides more information for the investigation of the internal structures of the compound nouns, although the enriched annotation adds very little information to the labels, and the semantic relations within the NPs are not explicitly revealed. More specifically, the work of [21] leaves the right-branching structures (which are the predominant cases for English) untouched, and just inserts labelled brackets around left-branching structures. Two types of new labels were assigned to these new internal nodes of the PTB NPs: NML or JJP, depending on whether each time the head of the NP is a noun or an adjective. Hence, in this analysis, for instance, the NP "Air Force contract" would receive the following structure:



As a consequence of such an annotation and treatment, *Air Force* as a unit is serving the function of the nominal modifier of *contract*.

Such enriched annotation enables one to investigate the bracketing preferences within the nominal phrases which was not available with the original PTB. By adapting the existing parsing models to use the enriched annotation, one can expect a fine-grained parsing result. Furthermore, this allows one to explore the treatment of NP in linguistically deep frameworks (see [22] for an example of such study in the framework of Combinatory Categorial Grammar (CCG)).

In DeepBank, the aim has always been to develop linguistic analyses independent from the PTB annotations. In the same spirit, we have decided not to incorporate the NP bracketing dataset from [21] directly during the annotation phase. On the other hand, as pointed out by the original PTB developers ([15]), asking annotators to directly annotate the internal structure of the base-NP significantly slows down the annotation process. We have made a similar observation in the DeepBank project. To help improve the annotation speed while maintaining quality, we have periodically updated the statistical models that re-rank the candidate trees and discriminants (binary decisions to made by human annotators) so that the manual decision making procedure has been made easier.

As an immediate next step in the research carried out for the dynamic annotation and disambiguation of English compound nouns in the DeepBank, we will compare the bracketing agreement with the NP dataset from [21].

References

[1] Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The tiger treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, pages 24–41, 2002.

[2] Thorsten Brants. TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP 2000)*, Seattle, USA, 2000.

[3] Ulrich Callmeier. Efficient parsing with large-scale unification grammars. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany, 2001.

[4] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[5] David Carter. The treebanker: a tool for supervised training of parsed corpora. In *Proceedings of the ACL Workshop on Computational Environments for Grammar Development and Linguistic Engineering*, pages 9–15, Madrid, Spain, 1997.

[6] Pamela Downing. On the creation and use of English compound nouns. *Language*, 53:4:810–842, 1977.

[7] Dan Flickinger. Accuracy vs. robustness in grammar engineering. In Emily M. Bender and Jennifer E. Arnold, editors, *Language from a Cognitive*

Perspective: Grammar, Usage, and Processing, pages 31–50. Stanford: CSLI Publications, 2011.

[8] Daniel Flickinger, Yi Zhang, and Valia Kordoni. Deepbank: A dynamically annotated treebank of the wall street journal. In *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories. International Workshop on Treebanks and Linguistic Theories (TLT-11), 11th, November 30 - December 1, Lisbon, Portugal*, pages 85–96. EdiÃğÃţes Colibri, Lisbon, 2012.

[9] Jan Hajič, Alena Böhmová, Eva Hajičová, and Barbora Vidová-Hladká. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 103–127. Amsterdam:Kluwer, 2000.

[10] Mark Johnson and Stefan Riezler. Exploiting auxiliary distributions in stochastic unification-based grammars. In *Proceedings of the 1st NAACL conference*, pages 154–161, Seattle, USA, 2000.

[11] Valia Kordoni and Yi Zhang. Annotating wall street journal texts using a hand-crafted deep linguistic grammar. In *Proceedings of The Third Linguistic Annotation Workshop (LAW III)*, Singapore, 2009.

[12] Judith Levi. *The Syntax and Semantics of Complex Nominals*. Academic Press, INC, New York, 1978.

[13] Robert Malouf, John Carroll, and Ann Copestake. Efficient feature structure operations without compilation. *Natural Language Engineering*, 6:29–46, 2000.

[14] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[15] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[16] Stephan Oepen. [incr tsdb()] — competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany, 2001.

[17] Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. The LinGO Redwoods treebank: motivation and preliminary applications. In *Proceedings of COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*, Taipei, Taiwan, 2002.

[18] Carl J. Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, USA, 1994.

[19] Kristina Toutanova, Christoper D. Manning, Dan Flickinger, and Stephan Oepen. Stochastic HPSG parse selection using the Redwoods corpus. *Journal of Research on Language and Computation*, 3(1):83–105, 2005.

[20] Kristina Toutanova, Christoper D. Manning, Stuart M. Shieber, Dan Flickinger, and Stephan Oepen. Parse ranking for a rich HPSG grammar. In *Proceedings of the 1st Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 253–263, Sozopol, Bulgaria, 2002.

[21] David Vadas and James Curran. Adding noun phrase structure to the penn treebank. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 240–247, Prague, Czech Republic, 2007.

[22] David Vadas and James R. Curran. Parsing noun phrase structure with CCG. In *Proceedings of ACL-08: HLT*, pages 335–343, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[23] Gertjan van Noord. Using self-trained bilexical preferences to improve disambiguation accuracy. In *Proceedings of the Tenth International Conference on Parsing Technologies*, pages 1–10, Prague, Czech Republic, 2007.

[24] Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 1034–1043, Prague, Czech, 2007.

[25] Beatrice Warren. *Semantic Patterns of Noun-Noun Compounds*. Acta Universitatis Gothoburgensis, Göteborg, 1978.