The Tibidabo Treebank

El treebank Tibidabo

Montserrat Marimon
Universitat de Barcelona
Gran Via de les Corts Catalanes 585, 08007-Barcelona
montserrat.marimon@ub.edu

Resumen: En este artículo presentamos el desarrollo de un nuevo recurso de código abierto para el español: el *treebank* Tibidabo. La anotación se está llevando a cabo de forma semi–automática en la que, en primer lugar, el corpus es analizado automáticamente con una gramática simbólica del español basada en HPSG e implementada en el sistema *Linguistic Knowledge Builder*, y, en segundo lugar, los resultados del proceso de análisis se desambiguan manualmente. La existencia del *treebank* Tibidabo nos permitirá futuros trabajos de investigación para el desarrollo y evaluación de una arquitectura híbrida que combine metodos simbólicos y estadísticos para el PLN, así como investigaciones orientadas a la hibridización de técnicas de bajo y alto nivel para el PLN.
Palabras clave: treebank, español, HPSG.

Abstract: This paper describes work in progress for the creation of a new open–source resource for Spanish: an HPSG–based treebank so–called Tibidabo. The annotation is performed semi–automatically. First, the corpus is automatically annotated by a symbolic HPSG–based grammar for Spanish implemented on the Linguistic Knowledge Builder system; then, the output is manually disambiguated. The existence of the Tibidabo treebank will facilitate research into the development and evaluation of a hybrid architecture combining symbolic and stochastic approaches to NLP, as well as investigations oriented to hybridization of shallow–deep techniques for NLP.
Keywords: Treebank, Spanish, HPSG.

1 Introduction

Linguistically interpreted natural language texts constitute a crucial resource both for theoretical linguistic investigations about language use, for instance, and for practical NLP purposes, such as the acquisition and evaluation of parsing systems. Thus, in recent years, there has been an increasing interest in the construction of treebanks and, nowadays, both theory–neutral and theory–grounded treebanks have been developed for a great variety of languages. Some of these treebanks are presented in (Hinrichs and Simov, 2004).

This paper describes ongoing work for the creation of a new resource for Spanish, an HPSG–based treebank so–called Tibidabo.

The annotation is carried out in two steps. First, the corpus is automatically annotated with the Spanish Resource Grammar, a multi–purpose large–coverage HPSG–based grammar for Spanish implemented on the Linguistic Knowledge Builder system. Then, ambiguous outputs are manually disambiguated.

The goal of the work we present is twofold: (i) to create the training data which will allow us to build up a parse selection model over the hand–built grammar following (Toutanova et al., 2005), and (ii) since language resource development and evaluation go hand in hand, to create the gold standard to evaluate the hybrid system. The Tibidabo treebank, in addition, will enable to evaluate foreseen investigations oriented to the hybridization of shallow–deep techniques for NLP aiming at efficient, robust, and accurate rule–based parsing.

The primary objective of our research work is to produce open–source reusable resources both for theoretical linguistic inves-

tigations and for application–oriented NLP. The Tibidabo treebank is open–source and, together with the Spanish Resource Grammar, is part of the DELPH–IN open–source repository of linguistic resources, which also includes HPSG–based grammars for a wide variety of languages, including English (Flickinger, 2002), German (Crysmann, 2005), Japanese (Siegel and Bender, 2002), French, Korean (Kim and Yangs, 2003), modern Greek (Kordoni and Neu, 2005), Norwegian (Hellan and Haugereid, 2005), and Portuguese (Branco and Costa, 2008), as well as HPSG–based treebanks, such as the LinGO Redwoods for English (Oepen et al., 2004) and Hinoki for Japanese (Hashimoto, Bond, and Siegel, 2007).[1]

The paper is organized as follows. First, in the following three sections, we present the main features of the system we use to annotate the corpus automatically (the architecture, the development environment and theoretical background, and the linguistic components and coverage) and the disambiguation process, and we show the representation of derivations produced by the grammar. Then, in section 5, we present the corpus which is the basis of the treebank and shows some figures about the treebank. Finally, in section 6, we conclude this paper with a summary and some directions for future work.

2 Treebank annotation

We have developed a hybrid architecture for the automatic processing of Spanish, shown in Figure 1, which integrates a shallow processing tool – FreeLing – into an HPSG–based grammar implemented on the LKB system – the Spanish Resource Grammar.

The advantage of our hybrid architecture is that it allows us to release the parser from certain tasks –namely, morphological analysis and recognition and classification of special text expressions that have been considered peripheral to the lexicon, e.g. numbers, dates, percentages, currencies, proper names, etc.– that may be robustly, efficiently, and reliably dealt with by shallow external components, thus making the whole system more adequate to deal with real world text.

[1]See http://www.delph-in.net/.

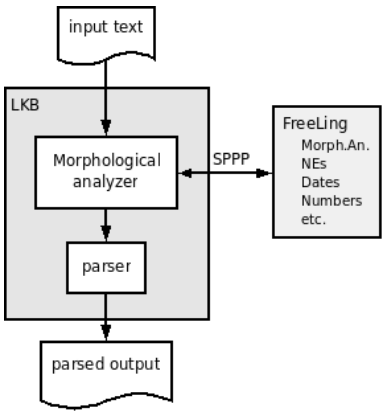

Figure 1: System architecture.

2.1 The FreeLing Tool

Before parsing input sentences with the LKB system, raw text is pre–processed by FreeLing, an open–source language analysis tool suite performing shallow processing functionalities ranging from text tokenization to dependency parsing (Atserias et al., 2006).[2]

Our system integrates the `morpho` processing module of FreeLing which receives a sentence and morphologically annotates each word. This module applies a cascade of specialized processors that includes:

- Punctuation symbol annotator.
- Multi–word recognizer.
- Numerical expression recognizer.
- Date/time expression recognizer.
- Ratio and percentage expression and monetary amount recognizer.
- Proper noun recognizer. A fast and simple pattern–matching module based on capitalization which yields an accuracy near 90%.[3]
- Dictionary look–up and affixes handler.
- Lexical probabilities annotator and unknown word handler.

Our system plugs the FreeLing tool into the system by means of the LKB Simple Pre-Processor Protocol (SPPP), which assumes

[2]The FreeLing toolkit may be downloaded from: http://www.lsi.upc.edu/~nlp/freeling.

[3]FreeLing also provides a NE recognizer based on the CoNLL–2002 shared task winning system (Carreras, Márquez, and Padró, 2002) with higher accuracy –over 92%– but which is rather slower.

that a preprocessor runs as an external process to the LKB system and communicates with its caller through its standard input and output channels.[4]

The SPPP, therefore, integrates into the parsing process the PoS tags and lemmata of each word in the sentence.

FreeLing and the Spanish Resource Grammar are two independently developed components and show some discrepancies in the tagset and the lexical categories. We also use this model to handle them.

2.2 The Spanish Resource Grammar

The Spanish Resource Grammar (SRG) is designed as multi–purpose (abstracted away from any particular application), and broad–coverage (aiming to cover not only all variations of the phenomena that have been implemented, but also the combinations of different phenomena).

2.2.1 Development environment and theoretical background

The grammar is implemented on the *Linguistic Knowledge Builder* (LKB) system – an interactive grammar development environment for typed feature structure grammars – (Copestake, 2002), based on the basic components of the LinGO Grammar Matrix, an open–source starter–kit for rapid development of precision broad–coverage grammars compatible with the LKB system (Bender and Flickinger, 2005).

The SRG is grounded in the theoretical framework of *Head–driven Phrase Structure Grammar* (HPSG) (Pollard and Sag, 1994), a constraint–based lexicalist approach to grammatical theory where all linguistic objects (i.e. words and phrases) are represented as typed feature structures, and uses *Minimal Recursion Semantics* (MRS) (Copestake et al., 2006) for the semantic representation. MRS is not a semantic theory in itself, but a kind of meta–level which has been defined for describing semantic structures. Using unification of typed features structures, MRS assigns a syntactically flat semantic representation to linguistic expressions.

2.2.2 Linguistic components and coverage

To parse a sentence, an LKB grammar requires three basic components: inflectional rules, a lexicon, and syntactic rules.

The inflectional rules. The inflectional rules in the LKB system perform the morphological analysis of the words in the input sentences.

Since we use an external morphological analyzer, the SRG does not need a morphology component, instead we use the inflectional rule component to propagate the morpho–syntactic information associated to full–forms, in the form of PoS tags, to the morpho–syntactic features of the lexical items. We have defined as many rules as tags we have in the FreeLing tagset.

The lexicon. The lexicon component in the LKB system contains the lexical entries of the grammar.

The SRG has a full coverage lexicon of closed word classes (pronouns, determiners, prepositions and conjunctions) and it contains about 50,000 lexical entries for open word classes (7,865 verbs, 28,025 nouns, 10,410 adjectives and 4,110 adverbs).[5] These lexical entries are organized into a multiple inheritance type hierarchy of about 500 leaf types which represent the type of words we have in our lexicon. The grammar also has 64 lexical rules to perform valence changing operations on lexical items (e.g. movement and removal of complements) which reduces the number of lexical entries to be manually encoded in the lexicon.

The syntactic rules. The syntactic rules in the LKB system are phrase structure rules that combine words and phrases into larger constituents and compositionally build up their semantic representation. The SRG has 191 phrase structure rules.

With these linguistic resources, the SRG handles the following range of linguistic phenomena: all types of subcategorization structures, surface word order variation and valence alternations, subordinate clauses, raising and control, determination, null–subjects and impersonal constructions, compound tenses, modification, passive constructions, comparatives and superlatives, cliticization, relative and interrogative clauses, sentential adjuncts, negation, noun ellipsis, and coordination, among others.

[4] See http://wiki.delph-in.net/moin/LkbSppp.

[5] The grammar also includes a set of generic lexical entry templates for open classes to deal with unknown words for virtually unlimited lexical coverage.

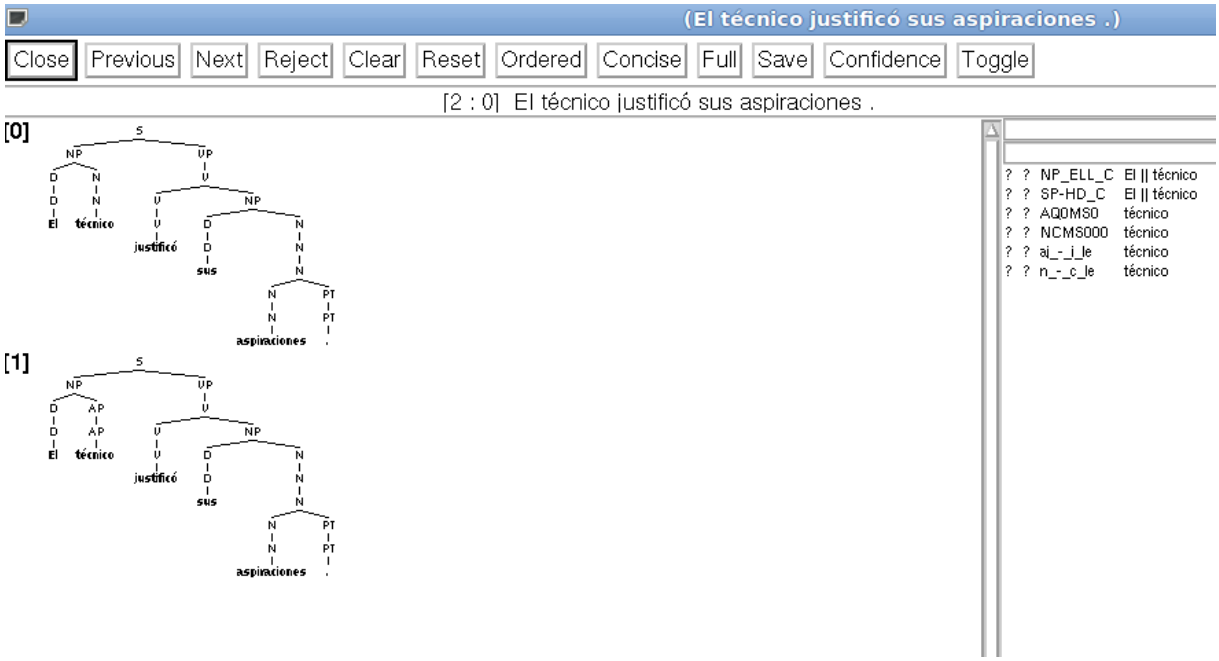

Figure 2: Screenshot of the annotation environment.

3 *Disambiguation*

Almost every sentence we parse is ambiguous. Thus, in order to construct the Tibidabo treebank, the outputs of the SRG have to be disambiguated by manually selecting the preferred analyses. This task is carried out with the `[incr tstb ()]` profiling environment.

Details of the `[incr tstb ()]` profiling environment can be found in (Oepen and Carroll, 2000). Basically, it includes: (1) a database that records the parsing results obtained from an LKB grammar, and (2) a tree comparison tool for the annotators to select the preferred analysis for each sentence, either by directly selecting it, as it is displayed as a labeled phrase structure tree, or, when dozens (or even hundreds) of analyses are displayed, to incrementally reduce the set of analyses either by selecting the (lexical or phrasal) local ambiguity that originates the multiple analyses, or by rejecting it.[6]

Figure 2 shows a screenshot of the annotation environment. The full set of analyses for the example *"el técnico justificó sus aspiraciones"* (the coach justified his goals)which appears on the top, are displayed on the left. Since *técnico* is both a noun and an adjective, the SRG produces two analyses: (0), where *técnico* is analysed as a noun and it is the head of the NP *el técnico*; and (1), where *técnico* is analysed as an adjective and it is attached to the article building an NP with an elliptical head. The window on the right shows the set of lexical and phrasal ambiguities that originates the multiple analyses for this sentence. That is, the two lexical entries –aj_-_i_le and n_-_c_le– and the two phrase structure rules that build up the NP node –SP-HD_C– and the elliptical NP node –NP_ELL_C–. AQ0MS00 and NCMS000 are the inflectional rules that propagate the PoS tags to the morpho–syntactic features of the lexical items. To disambiguate this sentence by selecting the first phrase structure tree, we can either select it directly, with the 'Save' option on the top of the environment, or we can select the rule or the lexical entry with which it is built up.[7]

All the decisions made by the annotators are recorded in the database of the `[incr tsdb()]` profiling environment so that they can be re–used to update the treebank semi–automalically with a revised version of the grammar, since only new ambiguities produced by the revised grammar need to be manually disambiguated.

[6] See (Oepen et al., 2004) for a detailed description of the tree comparison tool and some examples showing how the authors use it to construct the LinGO Redwoods treebank for English.

[7] Alternatively, we can reject either the second phrase structure tree, with the 'Reject' option on the top of the environment, or we can reject the lexical entry or the rule with which it is built up.

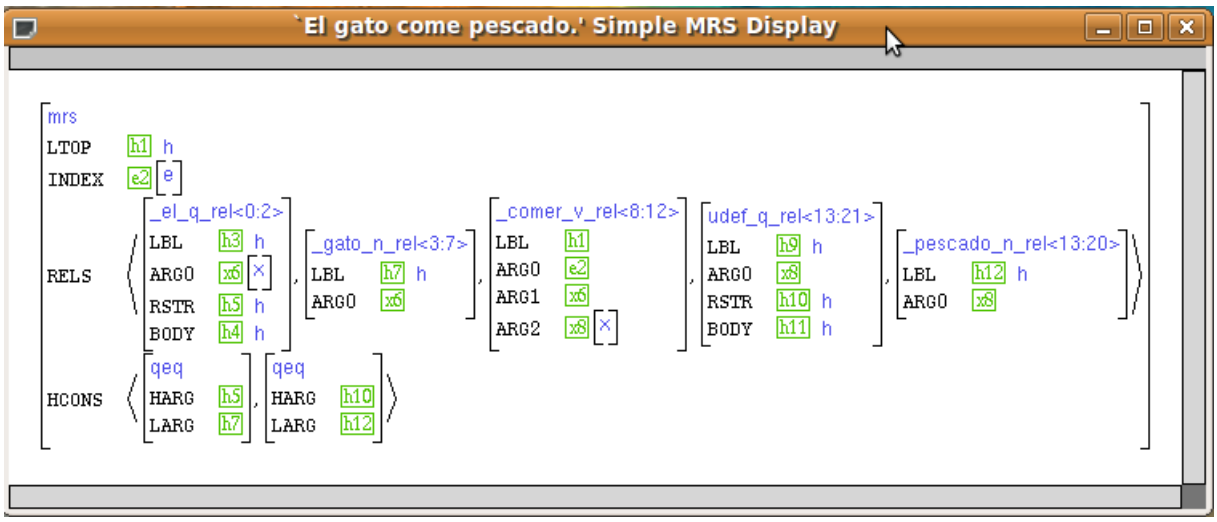

Figure 3: MRS semantic representation.

So far, analysis selection has been carried out by a single annotator, who has also been responsible for the development of the HPSG grammar. While disambiguating the outputs of the SRG, the annotator is writing down an annotation manual, or guidelines, which will be used by annotators that will join the project in the future. This manual include guidelines, for instance, (1) to select one lexical entry among the several entries we have defined to cope with all possible subcategorization frames of a given word, (2) to select the PP attachment, (3) to disambiguate the impersonal and passive constructions with "*se*", etc. Also, it is crucial to set a clear criteria to select one analysis when the grammar produces spurious ambiguity, i.e. when we get different structures showing the same semantic representation. That is the case, for example, of subordinating clauses in post–verbal position that the grammar attaches them both to the VP node and to the S node (i.e. before and after cancelling the subject). Another example is the rule removing optional complements which applies both before and after attaching post-verbal PP and/or adverbial modifiers.

4 *Representation of derivations*

The analyses that have been manually selected by the annotators are recorded in the `[incr tstb ()]` profiling environment database.

The annotation of each parsed sentence simultaneously represents two different descriptive levels: (i) a traditional phrase structure tree, and (ii) an MRS semantic representation. Thus, for an input sentence such as *"el gato come pescado."* (the cat eats fish.), the output of the SRG is as shown in Figure 3 and Figure 4.

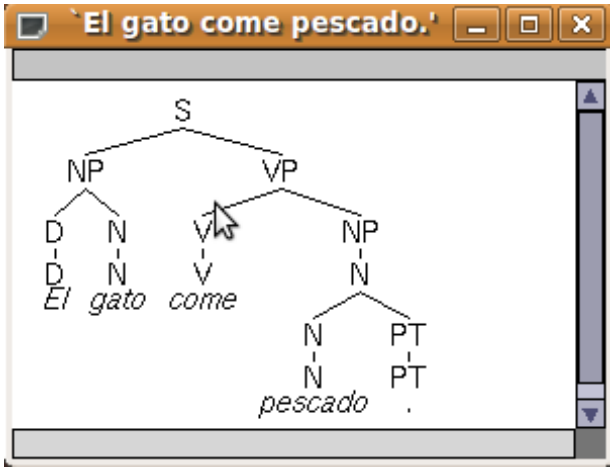

Figure 4: Phrase structure tree.

In the phrase structure tree each node is labeled with a set of atomic labels of the type 'S', 'VP', 'V', 'NP', etc.

The MRS semantic representation consists of: 1) a list of semantic relations (RELS), each with a "handle" (LBL) (used to express scope relations) and one or more roles (ARG0, ARG1,...). Relations are classified according to the number and type of arguments; 2) a set of handle constraints (HCONS), reflecting syntactic limitations on possible scope relations among the semantic relations, and 3) a group of distinguished semantic attributes of a linguistic sign. These attributes are: LTOP – the local top handle,

and INDEX – the salient nominal instance or event variable introduced by the lexical semantic head.

In addition, the database also records a derivational tree composed of the identifiers of the lexical entries and grammar rules which have been used to build up the tree and which may be used to reconstruct the full analysis.

5 The corpus

The basis of the Tibidabo treebank are newspaper text we borrow from the corpus AnCora, a corpus of about 500,000 words (17,364 sentences) (Taulé, Martí, and Recasens, 2008). Table 1 shows the number of sentences and ratio distributed along the sentence length.

sentence length	number of sentences	ratio
0-4	644	3.7%
5-9	1290	11.1%
10-14	1858	21.8%
15-19	2001	33.4%
20-24	2096	45.4%
25-29	1952	56.7%
30-34	1949	67.9%
35-39	1707	77.7%
40-44	1401	85.8%
45-49	1059	91.9%
50-54	615	95.4%
55-59	357	97.5%
60-64	206	98.7%
65-69	112	99.3%
70-74	52	99.6%
75-79	22	99.8%
80-84	11	99.8%
85-89	9	99.9%
90-94	4	99.9%
95-99	5	99.9%
100-104	2	99.9%
105-109	4	100%
110-114	4	100%
120-124	2	100%
125-129	1	100%
130-134	1	100%
125-129	1	100%

Table 1: The corpus AnCora.

Table 2 shows the ratio of sentences up to 40 words we have already processed, distributed along the sentence length. It is obvious that the longer the sentence are the more analyses they get, so sentences with have more that 40 words receive thundreds (even thousands) of analyses, this makes the disambiguation task difficult. Our objective is to disambiguate first the sentences which are up to 40 words and to use this treebank to create a parse selection model. The idea is to use this parse selection model to reduce the number of analyses produced by the grammar when parsing long sentences, which will then be disambiguated manually.

sentence length	processed
0-4	76%
5-9	85%
10-14	70%
15-19	60%
20-24	50%
25-29	40%
30-34	30%
35-39	20%

Table 2: The treebank Tibidabo.

6 Conclusions and future work

In this paper we have presented work in progress for the creation of a new open–source resource for Spanish: the Tibidabo HPSG–based treebank.

Besides improving the linguistic modules by extending the coverage of the grammar to deal with the phenomena which have not been implemented so far (e.g. VP ellipsis) and debugging the grammar to avoid errors and deficiencies to increase the treebank cases, future work includes the development of a parse selection model over the hand-built grammar. It is also foreseen investigations oriented to the hybridization of shallow–deep techniques for NLP aiming at efficient, robust, and accurate rule–based parsing.

Acknowledgments

This work was funded by the *Ramón y Cajal* program of the Spanish *Ministerio de Ciencia e Innovación.*

References

Atserias, J., B. Casas, E. Comelles, M. González, L. Padró, and M. Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of LREC'06*, Genoa, Italy.

Bender, E.M. and D. Flickinger. 2005. Rapid prototyping of scalable grammars: Towards modularity in extensions to a language-independent core. In *Proceedings of IJCNLP-05 (Posters/Demos)*, Jeju Island, Korea.

Branco, A. and F. Costa. 2008. *A Computational Grammar for Deep Linguistic Processing of Portuguese: LXGram, version A.4.1. Research Report TR-2008-17.* Universidade de Lisboa, Faculdade de CiÃªncias, Departamento de Informatica.

Carreras, X, L. Márquez, and L. Padró. 2002. Named entity extraction using adaboost. In *Proceedings of CoNLL Shared Task*, Taipei, Taiwan.

Copestake, A. 2002. *Implementing Typed Feature Structure Grammars.* CSLI Publications, CSLI lecture notes, number 110, Chicago.

Copestake, A., D. Flickinger, C.J. Pollard, and I. A. Sag. 2006. Minimal recursion semantics: an introduction. *Research on Language and Computation*, 3(4):281–332.

Crysmann, B. 2005. Syncretism in German: a unified approach to underspecification, indeterminacy, and likeness of case. In *Proceedings of the 12th International Conference on Head-driven Phrase Structure Grammar*, Lisbon, Portugal.

Flickinger, D. 2002. On building a more efficient grammar by exploiting types. In D. Flickinger, S. Oepen, J. Tsujii, and H. Uszkoreit, editors, *Collaborative Language Engineering.* Stanford: CSLI Publications, pages 1–17.

Hashimoto, C., F. Bond, and M. Siegel. 2007. Semi-automatic documentation of an implemented linguistic grammar augmented with a treebank. In *Language Resources and Evaluation. (Special issue on Asian language technology).*

Hellan, L. and P. Haugereid. 2005. Norsource - an excercise in the matrix grammar building design. In Emily M. Bender, Dan Flickinger, Frederik Fouvry, and Melanie Siegel, editors, *A Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI*, Vienna, Austria.

Hinrichs, E.W. and K. Simov. 2004. *Research on Language and Computation, 2 (4).*

Kim, J-B. and J. Yangs. 2003. Korean phrase structure grammar and its implementations into the lkb system. In *Proceedings of the 17th Pacific Asia Conference on Language, Information, and Computation.*

Kordoni, V. and J. Neu. 2005. Deep analysis of modern greek. In Jong-Hyeok Lee et al. Keh-Yih Su, Jun'ichi Tsujii, editor, *Lecture Notes in Computer Science, Vol 3248.* Springer-Verlag Berlin Heidelberg, pages 674 – 683.

Oepen, S. and J. Carroll. 2000. Performance profiling for parser engineering. In D. Flickinger, S. Oepen, J. Tsujii, and H. Uszkoreit, editors, *Natural Language Engineering (6)1 —Special Issue: Efficiency Processing with HPSG: Methods, Systems, Evaluation.* Cambridge University Press, pages 81–97.

Oepen, S., D. Flickinger, K. Toutanova, and C.D. Manning. 2004. Lingo redwoods. In E.W. Hinrichs and K. Simov, editors, *Research on Language and Computation, 2(4).*

Pollard, C.J. and I.A. Sag. 1994. *Head-driven Phrase Structure Grammar.* The University of Chicago Press and CSLI Publications, Chicago.

Siegel, M. and E.M. Bender. 2002. Efficient deep processing of japanese. In *3rd Workshop on Asian Language Resources and International Standardization, COLING-02*, Tapei, Taiwan.

Taulé, M., M.A. Martí, and M. Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC-2008*, Marrakech, Morocco.

Toutanova, K., C.D. Manning, D. Flickinger, and S Oepen. 2005. Stochastic hpsg parse disambiguation using the redwoods corpus. In *Journal of Logic and Computation.*