Towards Comparability of Linguistic Graph Banks for Semantic Parsing

Stephan Oepen[♣], Marco Kuhlmann[♠], Yusuke Miyao[♡], Daniel Zeman[♢], Silvie Cinková[♢], Dan Flickinger[◦], Jan Hajič[♢], Angelina Ivanova[♣], and Zdeňka Urešová[♢]

[♣] University of Oslo, Department of Informatics
[♠] Linköping University, Department of Computer and Information Science
[♡] National Institute of Informatics, Tokyo
[♢] Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
[◦] Stanford University, Center for the Study of Language and Information
sdp-organizers@delph-in.net

Abstract

We announce a new language resource for research on *semantic parsing*, a large, carefully curated collection of *semantic dependency graphs* representing multiple linguistic traditions. This resource is called SDP 2016 and provides an update and extension to previous versions used as Semantic Dependency Parsing target representations in the 2014 and 2015 Semantic Evaluation Exercises (SemEval). For a common core of English text, this third edition comprises semantic dependency graphs from four distinct frameworks, packaged in a unified abstract format and aligned at the sentence and token levels. SDP 2016 is the first general release of this resource and available for licensing from the Linguistic Data Consortium from May 2016. The data is accompanied by an open-source SDP utility toolkit and system results from previous contrastive parsing evaluations against these target representations.

Keywords: Semantic Dependency Parsing, Semantic Parsing, Semantic Dependencies, Meaning Representation

1. Background and Goals

Increased interest in natural language 'understanding' has brought into the focus of much current research a variety of techniques often described as 'semantic parsing'. In this work, we seek to contribute to improved comparability of representations and results in one sub-area of semantic parsing, viz. analysis of natural language strings into *bi-lexical semantic dependency graphs*; these are general representations of sentence meaning whose nodes correspond to surface lexical units and whose directed edges encode core predicate–argument relations.[1]

We have prepared for public release in May 2016 an assortment of semantic dependency graphs stemming from four distinct linguistic traditions, for substantial volumes of English, Chinese, and Czech running text, and providing both in-domain training and test, as well as out-of-domain test data. To eliminate minor divergences across resources, these graph banks are aligned at the sentence and token level, within each language, represented in a unified abstract graph model, and packaged in a common file format. Where applicable, the graphs are paired with original representations from which they were derived (e.g. underspecified logical-forms or tectogrammatical trees), and with corresponding 'companion' syntactic analyses from a broad variety of frameworks and sources, comprising gold-standard annotations as well as analyses delivered by state-of-the-art statistical parsers.

[1] Domain- and application-independence and lexicalization set these target representations apart from other strands of semantic parsing, into immediately actionable query languages in the tradition of Zelle & Mooney (1996), on the one hand, or into representations whose primitives need not be surface lexical units, on the other hand, as for example English Resource Semantics (Copestake & Flickinger, 2000; Flickinger et al., 2014), the Discourse Representation Structures of Bos (2008), or Abstract Meaning Representation (Banarescu et al., 2013).

Furthermore, the release package includes system submissions and scores from two parsing competitions against several of our target representations, viz. the Broad-Coverage Semantic Dependency Parsing (SDP) tasks at recent Semantic Evaluation Exercises (Oepen et al., 2014, 2015), together with Java and Python tools to read, manipulate, and score these graphs. We intend this overview paper to document relevant formal and (some of the) linguistic properties of these graph banks and to encourage broader use of this standardized collection for improved comparability and replicability; we refer to this new public resource as SDP 2016.

2. Varieties of Semantic Dependency Graphs

The earlier SDP tasks comprised three distinct target representations, dubbed DM, PAS, and PSD (see below for details). SDP 2016 derives an additional collection of semantic dependency graphs, which we term CCD, from CCGbank (Hockenmaier & Steedman, 2007). Dependencies of this type have been used as the target representations in some recent parsing work (Auli & Lopez, 2011; Du et al., 2015; Kuhlmann & Jonsson, 2015), but the exact procedure of extracting these graphs from CCGbank has yet to be standardized. The following paragraphs briefly summarize the linguistic genesis of each representation, with particular emphasis on CCD, because the other three have already been introduced by Oepen et al. (2014) and Miyao et al. (2014). With the exception of the DM graphs, all representations for English, to some degree, build on the venerable Penn Treebank (PTB; Marcus et al., 1993), though the connection is arguably more direct for CCD and PAS than for PSD (where substantial additional manual annotation was performed).

CCD: Combinatory Categorial Grammar Dependencies Hockenmaier & Steedman (2007) construct CCGbank from a combination of careful interpretation of the syntactic annotations in the PTB with additional, manually curated lexical and constructional knowledge. In CCGbank, the strings of

the PTB Wall Street Journal (WSJ) Corpus are annotated with pairs of (a) CCG syntactic derivations and (b) sets of semantic bi-lexical dependency triples. The latter "include most semantically relevant non-anaphoric local and long-range dependencies" and are suggested by the CCGbank creators as a proxy for predicate–argument structure. While these have mainly been used for contrastive parser evaluation (Clark & Curran, 2007; Fowler & Penn, 2010; inter alios), recent parsing work as mentioned above views each set of triples as a directed graph and parses directly into these target representations. Our CCD graphs combine the CCGbank dependency triples with information gleaned from the CCG syntactic derivations, notably the part of speech and lexical category associated with each token (interpreted as its argument *frame*), and the identity of the lexical head of the derivation, which becomes the semantic *top* node.

DM: DELPH-IN MRS Bi-Lexical Dependencies These semantic dependency graphs originate in a manual re-annotation, dubbed DeepBank[2], of Sections 00–21 of the WSJ Corpus with syntactico-semantic analyses from the LinGO English Resource Grammar (ERG; Flickinger, 2000; Flickinger et al., 2012). Native ERG semantics take the form of underspecified logical forms, which Oepen et al. (2002); Oepen & Lønning (2006); and Ivanova et al. (2012) map onto the DM bi-lexical semantic dependencies in a two-step conversion pipeline.[3] For this target representation, top nodes designate the highest-scoping (non-quantificational) predicate in the graph, e.g. the scopal adverb *almost* in Figure 1 below.

PAS: Enju Predicate–Argument Structures The Enju Treebank[4] is derived from automatic HPSG-style re-annotation of the PTB (Miyao, 2006). Our PAS graphs stem from the Enju Treebank, without contentful conversion, and from the application of the same basic techniques to the Penn Chinese Treebank (CTB; Xue et al., 2005). Top nodes in this representation denote semantic heads.

PSD: Prague Semantic Dependencies The Prague Czech-English Dependency Treebank (PCEDT; Hajič et al., 2012)[5] is a set of parallel dependency trees over the WSJ texts from the PTB, and their Czech translations. Our PSD dependencies have been extracted from the *tectogrammatical* annotation layer (so-called 't-trees').[6] Top nodes are derived from t-tree roots; i.e. they mostly correspond to matrix verbs; in case of clausal coordination, there can be multiple top nodes.

[2]See `http://www.delph-in.net/deepbank/`.

[3]The original logical forms and intermediate variable-free semantic networks (dubbed Elementary Dependency Structures by Oepen & Lønning, 2006) are included as background material in the SDP 2016 package.

[4]See `http://kmcs.nii.ac.jp/enju/`.

[5]See `http://ufal.mff.cuni.cz/pcedt2.0/`.

[6]Again, the original tectogrammatical trees from which the PSD bi-lexical semantic dependencies derive (through the elimination of empty 'generated' nodes and propagation of dependencies into paratactic constructions, as described by Miyao et al., 2014) are included in the released data.

3. A Contrastive Example

Figure 1 shows example target representations from the above four sources (not showing CCD frame identifiers, for layout reasons) for the WSJ sentence:

(1) A similar technique is almost impossible to apply to other crops, such as cotton, soybeans, and rice.

Semantically, *technique* arguably is dependent on the determiner (the quantificational locus), the comparative modifier *similar*, and the predicate *apply*. Conversely, the predicative copula, infinitival *to*, and the preposition marking the deep object of *apply* can be argued to not have a semantic contribution of their own. Thus, where common *syntactic* dependency representations take the form of fully connected trees, semantic dependency graphs are characterized by *node re-entrancies* and *partial connectivity*.

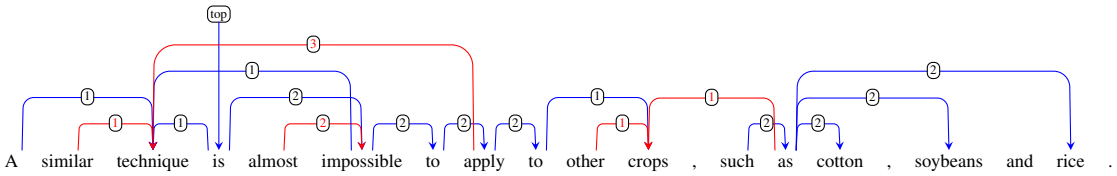
It is evident from this example alone that there are contentful differences between the four representations—as would be expected given stark differences in their linguistic pedigree. The different representations exhibit between twelve and twenty dependencies for (1), of which only five (highlighted in red in Figure 1) are shared across all representations, albeit at times with inverse directionality in PSD. Besides our technical goal of advancing semantic parsing research, we also hope that the parallel SDP 2016 collection will facilitate qualitative linguistic comparison of these representations (and possibly others).

For example, (1) invokes the so-called *tough* construction, where a restricted class of adjectives (*impossible* in our case) select for infinitival verb phrases containing an object gap and, thus, create a long-distance dependency (Rosenbaum, 1967; Nanni, 1980; inter alios). All four representations correctly capture this dependency (making *technique* a semantic argument of *apply*), but only DM and PSD further analyze *impossible* as an expletive predicate—recognizing the close paraphrase of (1) as *It is almost impossible to apply a similar technique* ... Therefore, in contrast to CCD and PAS, the latter two abstain from marking *technique* as a subject-like argument to the predicative adjective.

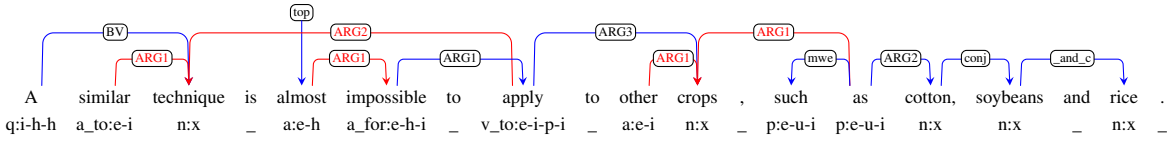
Similarly, DM and PSD pattern largely alike in considering the infinitival particle and argument-marking use of the preposition *to* as not meaning-bearing, but only DM further treats the predicative copula as vacuous, leading to a different choice of *top* node (see below). Finally, the representations also differ widely in their analysis of coordination, where CCD and PSD project the incoming argument dependency onto all conjuncts (and, for PSD, transitively also through the apposition established by *such as*), while DM and PAS apply different group forming mechanisms: the 'chaining' analysis of Mel'čuk (1988) and recursive binary nesting, respectively.

4. A Unified Graph Representation

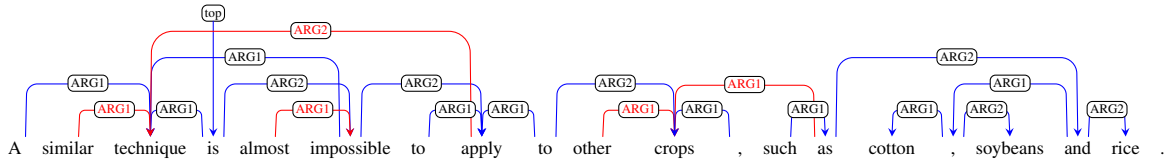
The SDP target representations can be uniformly characterized as node-ordered, labeled, directed graphs, i.e. as pairs $G = (V, E)$ where V is a set of *nodes*, $E \subseteq V \times V$ is a set of *edges*, and there is a strict total ordering on V (corresponding to the surface token sequence). Nodes in the SDP graphs can be labeled with up to five pieces of information: word *form*, optional *lemma*, *part of speech*, a Boolean flag



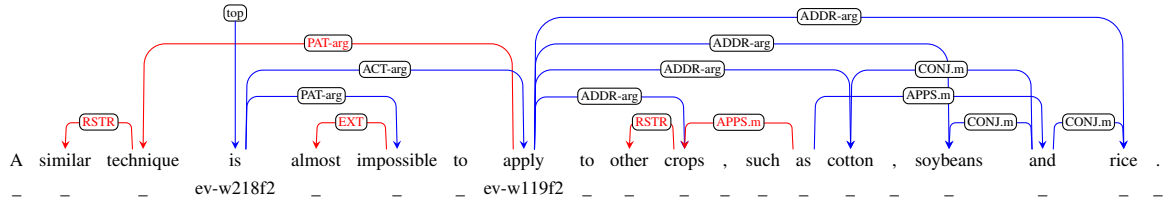
(a) CCGbank word–word dependencies (CCD; omitting lexical categories for space reasons).



(b) DELPH-IN Minimal Recursion Semantics–derived bi-lexical dependencies (DM).



(c) Enju Predicate–Argument Structures (PAS).



(d) Parts of the tectogrammatical layer of the Prague Czech-English Dependency Treebank (PSD).

Figure 1: Sample parallel dependency graphs for Example (1).

indicating whether the node represents a *top* predicate, and an optional *frame* (or sense) tag—for example the distinction between causative vs. inchoative predicates like *increase*. Edges are labeled with semantic relations that hold between their source and target nodes. In contrast to the unique root node in trees, graphs can have multiple (structural) roots, i.e. nodes with in-degree zero; in this sense, the majority of graphs are multi-rooted in all SDP graph banks. Thus, our format designates one (or more) *top* node(s) per graph to reflect notions like semantic headedness, top-level focus, or generally the most central semantic entities in the graph. All data in the SDP collection uses a column-based file format that conservatively extends the format of the Shared Task of the 2009 Conference on Natural Language Learning (CoNLL).

5. Facts and Figures

The bulk of the English SDP data is in-domain training and test data drawing on the WSJ Corpus; our collection only includes sentences for which gold-standard annotations are available in all representations and where sentence and token alignment was successful, in total some 80 % of the WSJ segment of the PTB. Table 1 shows sentence (graph) and token (node) counts for the various segments of the SDP 2016 release. For DM and PSD, additional English out-of-domain test data was constructed by fresh manual annotation of a balanced sample of twenty files from the Brown Corpus (Francis & Kučera, 1982), and for PAS parallel annotations were obtained by applying the Enju converter to the corresponding parts of the PTB. CCD and DM use theory-specific argument *frame* identifiers, whereas PSD draws on the (much larger inventory of) *sense* identifiers from the PCEDT valency lexicon.

	sentences	**tokens**
EN in-domain train	35,657	802,717
EN in-domain test	1,410	31,948
CS in-domain train	42,076	985,302
CS in-domain test	1,670	38,397
ZH in-domain train	31,113	649,036
ZH in-domain test	8,976	214,454
EN out-of-domain	1,849	31,583
CS out-of-domain	5226	214,454

Table 1: Sentence and token counts in SDP 2016.

The focus in much recent semantic dependency parsing work has been on English, but our SDP 2016 collection also includes some additional languages, albeit only for select target representations. Chinese in-domain training and test data for PAS is derived by conversion from Release 7.0 of the CTB. Czech in- and out-of-domain PSD graphs draw on the translations of the WSJ Corpus in PCEDT 2.0, and on the Prague Dependency Treebank 3.0 (PDT; Hajič, 1998). Again, see Table 1 for sentence and token counts.

To guide cross-framework comparison, Table 2 quantifies

		EN i-d				CS i-d	ZH i-d	EN o-o-d			CS o-o-d
		CCD	DM	PAS	PSD	PSD	PAS	DM	PAS	PSD	PSD
(1)	**# labels**	6	59	42	91	61	32	47	41	74	64
(2)	**# frames**	1263	297	–	5426	–	–	172	–	1208	–
(3)	**% nodes that are singletons**	12.10	22.97	4.38	35.76	28.91	0.11	25.40	5.84	39.11	29.04
(4)	**% graphs that are trees**	1.45	2.30	1.22	42.19	37.66	3.49	9.68	2.38	51.43	51.49
(5)	**edge density**	1.07	1.02	1.07	1.07	1.07	1.02	0.95	1.02	0.99	1.00
(6)	**% nodes with reentrancies**	28.09	27.44	29.36	11.42	11.80	24.96	26.14	29.36	11.46	11.44

Table 2: High-level graph statistics across target representations, languages, and domains.

a number of structural properties for each graph bank and segment. Again, there is great variation across the various representations. The granularity of dependency labels (1) is much greater in PSD than in CCD, for example; CCD labels merely are positional indices into argument slots of the lexical frame. A similar observation applies to the granularity of 'frames' (2), where CCD and DM encode more general 'linking patterns', i.e. unlexicalized mappings from syntactic to semantic arguments; PSD, on the other hand, employs actual sense identifiers and would, thus, show different values for distinct lexemes.[7] Conversely, PSD only annotates senses on verbal predicates, whereas CCD and DM provide frame identifiers for all semantically contentful nodes (as seen in Figure 1).

Oepen et al. (2014) call unconnected (with in- and out-degree zero) non-top nodes *singletons*, and by construction these nodes correspond to semantically vacuous lexical units in the SDP graphs. As observed in Figure 1 above, differences in the analysis of function words and punctuation marks largely account for the much smaller proportions of singletons in CCD and PAS (3); these trends are stable across languages and the in- vs. out-of-domain splits. Subsequent statistics in Table 2 discard singletons, and measures (5) and (6) quantify structural properties in each graph bank that transcend simple, rooted trees (4)—or, informally speaking, degrees of graph complexity: *edge density* (5) is the ratio of edges to nodes; nodes with reentrancies (6) are nodes with more than one incoming edge. Presumably owing to its roots in the PCEDT tectogrammatical trees, PSD stands out in this comparison with the highest proportion of actual trees and smallest percentage of reentrant nodes. In future work, we plan to further correlate quantitative observations and differences in linguistic design.

6. Outlook: Connecting the Dots

The SDP 2016 collection of graph banks becomes available through the Linguistic Data Consortium (LDC) in May 2016 with catalogue number LDC2016T10; please see `http://sdp.delph-in.net/` for further information. We envision that general availability of a standardized and comprehensive set of semantic dependency graphs and associated tools will stimulate more research in this sub-area of semantic parsing. To date, reported 'parsing success' measures in terms of dependency F_1 range between the high seventies for PSD (Martins & Almeida, 2014) and high eighties to low nineties for CCD, DM, and PAS (Du et al., 2015; Miyao et al., 2014). Such variation may in principle be owed to differences in the number and complexity of linguistic distinctions made, to homogeneity and consistency of training and test data, and of course to the cumulative effort that has gone into pushing the state of the art on individual target representations. A deeper understanding of these parameters, as well as of contentful vs. superficial linguistic differences across frameworks, will be a prerequisite to judging the relative suitability of different resources and approaches. At the same time, we expect that a more in-depth analysis of the algebraic properties of various subclasses of semantic dependency graphs will aid the design of specialized parsing algorithms and probability models.

While the SDP 2016 collection of target representations limits itself to bi-lexical semantic dependency graphs, i.e. graphs whose nodes correspond one-to-one with surface tokens, we plan to extend the quantitative and qualitative comparison across frameworks and representations to additional graph-structured meaning representations that transcend bi-lexical dependencies. Closely related but formally somewhat different (non-lexicalized and unordered) semantic graphs include, for example, some of the underlying representations from which the SDP graphs derive (Elementary Dependency Structures for DM; tectogrammatical trees for PSD), as well as semantic networks in the framework of Abstract Meaning Representation.

Acknowledgments

We are grateful to Željko Agić and Bernd Bohnet for consultation and assistance in preparing our companion parses, to the Linguistic Data Consortium (LDC) for support in distributing the SDP data, as well as to three anonymous reviewers for feedback on an earlier version of this manuscript. We warmly thank the SemEval 2014 and 2015 chairs, Preslav Nakov and Torsten Zesch, for always being role-model organizers, equipped with an outstanding balance of structure, flexibility, and community spirit. Data preparation was supported through the ABEL high-performance computing facilities at the University of Oslo, and we acknowledge the Scientific Computing staff at UiO, the Norwegian Metacenter for Computational Science, and the Norwegian taxpayers. Part of the work was supported by the grants 15-10472S, 15-20031S, and GP13-03351P of the Czech Science Foundation, and by the LINDAT/CLARIN projects LM2015071 and LM2010013 of the Ministry of Education, Youth, and Sports of the Czech Republic.

[7] To seek to relate these different approaches to the encoding of lexical valency, one can multiply out the DM frame identifiers with verb lemmata, which yields a count of some 4,600 distinct combinations, i.e. slightly less than the set of observed sense distinctions in PSD.

References

Auli, M., & Lopez, A. (2011). Training a log-linear parser with loss functions via softmax-margin. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (p. 333 – 343). Edinburgh, Scotland, UK.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... Schneider, N. (2013). Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (p. 178 – 186). Sofia, Bulgaria.

Bos, J. (2008). Wide-coverage semantic analysis with Boxer. In *Proceedings of the Semantics in Text Processing conference* (p. 277 – 286). Venice, Italy.

Clark, S., & Curran, J. R. (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, *33*(4), 493 – 552.

Copestake, A., & Flickinger, D. (2000). An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation.* Athens, Greece.

Du, Y., Sun, W., & Wan, X. (2015). A data-driven, factorization parser for CCG dependency structures. In *Proceedings of the 53rd Meeting of the Association for Computational Linguistics and of the 7th International Joint Conference on Natural Language Processing* (p. 1545 – 1555). Bejing, China.

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering*, *6 (1)*, 15 – 28.

Flickinger, D., Bender, E. M., & Oepen, S. (2014). Towards an encyclopedia of compositional semantics. Documenting the interface of the English Resource Grammar. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (p. 875 – 881). Reykjavik, Iceland.

Flickinger, D., Zhang, Y., & Kordoni, V. (2012). DeepBank. A dynamically annotated treebank of the Wall Street Journal. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories* (p. 85 – 96). Lisbon, Portugal: Edições Colibri.

Fowler, T. A. D., & Penn, G. (2010). Accurate context-free parsing with Combinatory Categorial Grammar. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics* (p. 335 – 344). Uppsala, Sweden.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage. Lexicon and grammar*. New York, USA: Houghton Mifflin.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., ... Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (p. 3153 – 3160). Istanbul, Turkey.

Hajič, J. (1998). Building a syntactically annotated corpus. The Prague Dependency Treebank. In *Issues of valency and meaning* (p. 106 – 132). Prague, Czech Republic: Karolinum.

Hockenmaier, J., & Steedman, M. (2007). CCGbank. A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, *33*, 355 – 396.

Ivanova, A., Oepen, S., Øvrelid, L., & Flickinger, D. (2012). Who did what to whom? A contrastive study of syntacto-semantic dependencies. In *Proceedings of the Sixth Linguistic Annotation Workshop* (p. 2 – 11). Jeju, Republic of Korea.

Kuhlmann, M., & Jonsson, P. (2015). Parsing to noncrossing dependency graphs. *Transactions of the Association for Computational Linguistics*, *3*, 559 – 570.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpora of English. The Penn Treebank. *Computational Linguistics*, *19*, 313 – 330.

Martins, T. A. F., & Almeida, C. M. S. (2014). Priberam. A turbo semantic parser with second order features. In *Proceedings of the 9th International Workshop on Semantic Evaluation* (p. 471 – 476). Dublin, Ireland.

Mel'čuk, I. (1988). *Dependency syntax. Theory and practice*. Albany, NY, USA: SUNY Press.

Miyao, Y. (2006). *From linguistic theory to syntactic analysis. Corpus-oriented grammar development and feature forest model*. Doctoral dissertation, University of Tokyo, Tokyo, Japan.

Miyao, Y., Oepen, S., & Zeman, D. (2014). In-House. An ensemble of pre-existing off-the-shelf parsers. In *Proceedings of the 9th International Workshop on Semantic Evaluation* (p. 63 – 72). Dublin, Ireland.

Nanni, D. (1980). On the surface syntax of constructions with *easy*-type adjectives. *Language*, *56*(3), 568 – 591.

Oepen, S., Flickinger, D., Toutanova, K., & Manning, C. D. (2002). Lingo Redwoods. A rich and dynamic treebank for HPSG. In *Proceedings of the 1st International Workshop on Treebanks and Linguistic Theories* (p. 139 – 149). Sozopol, Bulgaria.

Oepen, S., Kuhlmann, M., Miyao, Y., Zeman, D., Cinková, S., Flickinger, D., ... Urešová, Z. (2015). SemEval 2015 Task 18. Broad-coverage semantic dependency parsing. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (p. 915 – 926). Denver, CO, USA.

Oepen, S., Kuhlmann, M., Miyao, Y., Zeman, D., Flickinger, D., Hajič, J., ... Zhang, Y. (2014). SemEval 2014 Task 8. Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation* (p. 63 – 72). Dublin, Ireland.

Oepen, S., & Lønning, J. T. (2006). Discriminant-based MRS banking. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (p. 1250 – 1255). Genoa, Italy.

Rosenbaum, P. S. (1967). *The grammar of English predicate complement constructions*. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA.

Xue, N., Xia, F., Chiou, F.-D., & Palmer, M. (2005). The Penn Chinese TreeBank. Phrase structure annotation of a large corpus. *Natural Language Engineering*, *11*, 207 – 238.

Zelle, J. M., & Mooney, R. J. (1996). Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (p. 1050 – 1055). Portland, OR, USA.