*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 1902–1909
Marseille, 11–16 May 2020

# A Tale of Three Parsers[†]: Towards Diagnostic Evaluation for Meaning Representation Parsing

**Maja Buljan[♣], Joakim Nivre[♠], Stephan Oepen[♣], and Lilja Øvrelid[♣]**
[♣] University of Oslo, Department of Informatics
[♠] Uppsala University, Department of Linguistics and Philology
{majabu|oe|liljao}@ifi.uio.no, joakim.nivre@lingfil.uu.se

**Abstract**

We discuss methodological choices in contrastive and diagnostic evaluation in meaning representation parsing, i.e. mapping from natural language utterances to graph-based encodings of semantic structure. Drawing inspiration from earlier work in syntactic dependency parsing, we transfer and refine several quantitative diagnosis techniques for use in the context of the 2019 shared task on Meaning Representation Parsing (MRP). As in parsing proper, moving evaluation from simple rooted trees to general graphs brings along its own range of challenges. Specifically, we seek to begin to shed light on relative strenghts and weaknesses in different broad families of parsing techniques. In addition to these theoretical reflections, we conduct a pilot experiment on a selection of top-performing MRP systems and two of the five meaning representation frameworks in the shared task. Empirical results suggest that the proposed methodology can be meaningfully applied to parsing into graph-structured target representations, uncovering hitherto unknown properties of the different systems that can inform future development and cross-fertilization across approaches.

**Keywords:** Data-Driven Parsing, Sentence Semantics, Meaning Representation Parsing, Contrastive Evaluation, Diagnostics

## 1. Introduction

In no small part following on from decades of progress in (surface) syntactic parsing, there is now growing interest in parsing into 'deeper' and more abstract representations of sentence structure. In particular, encoding *sentence meaning* in the form of *labeled directed graphs* has been the focus of a series of annual parsing competitions since the first shared task on Semantic Dependency Parsing (SDP) at the 2014 Workshop on Semantic Evaluation (Oepen et al., 2014). Adopting the perspective of the most recent such shared task, at the 2019 Conference on Computational Natural Language Learning (CoNLL), we refer to this line of research as *meaning representation parsing* (Oepen et al., 2019).

While most representations of syntactic structure limit themselves to *rooted trees*, common frameworks for meaning representation assume general graphs. These structures make the parsing task much more complex—often moving from techniques with polynomial worst-case complexity to problems that are in principle NP-hard. Among other things, meaning representations transcend syntactic trees in allowing nodes with in-degree greater than one ('reentrancies'), multiple roots, and ignoring semantically 'vacuous' parts of the parser input. Besides greatly increased modeling and algorithmic complexity, meaning representation parsing also poses its own set of methodological challenges for parser evaluation and diagnostics.

The contrastive studies initiated by McDonald and Nivre (2007) and McDonald and Nivre (2011) have been influential in comparing the performance of two core types of approaches to syntactic dependency parsing, i.e. different families of parsing approaches. In this work, we investigate to what degree these techniques can be transferred to meaning representation parsing, and how they can be adapted and extended to reflect the formal and linguistic differences in the nature of target representations. We develop the blueprint of a general framework for quantitative diagnostic evaluation and experimentally seek to validate this proposal through a small-scale pilot study.

The remainder of the paper is structured as follows: In § 2., we present the relevant background, including previous studies in syntactic parsing that provide our point of departure and the 2019 shared task on meaning representation parsing. § 3. gives a review of established dimensions of contrastive diagnostic evaluation for syntactic dependency parsing, and discusses their transferability and adaptation to semantic graphs. In § 4., we present a small-scale empirical study, and apply two of the querying dimensions to a select trio of currently top-performing semantic parsers. Finally, § 5. concludes the paper and discusses avenues for future research.

## 2. Background

The following paragraphs establish relevant methodological and technological context for our work, out of necessity summarizing prior efforts in rather broad strokes.

**A Tale of Two Parsers** One inspiration for this study is the contrastive error analysis of *graph-based* vs. *transition-based* syntactic dependency parsers carried out by McDonald and Nivre (2007) and McDonald and Nivre (2011). Based on data from the CoNLL 2006 shared task on multilingual dependency parsing (Buchholz and Marsi, 2006), they analyzed the performance of the two parser types in relation to a number of structural factors, such as sentence length, dependency length, and tree depth, as well as linguistic categories, notably parts of speech and dependency types. The analysis showed that, although the best graph-based and transition-based syntactic dependency parsers at the time achieved very similar accuracy on average, they had quite distinctive error profiles. Moreover, these differences could be explained by inherent strengths and weaknesses of the two algorithmic approaches. Thus, for exam-

[†] We acknowledge and thank (Zhang and Clark, 2008) for inspiring our title.

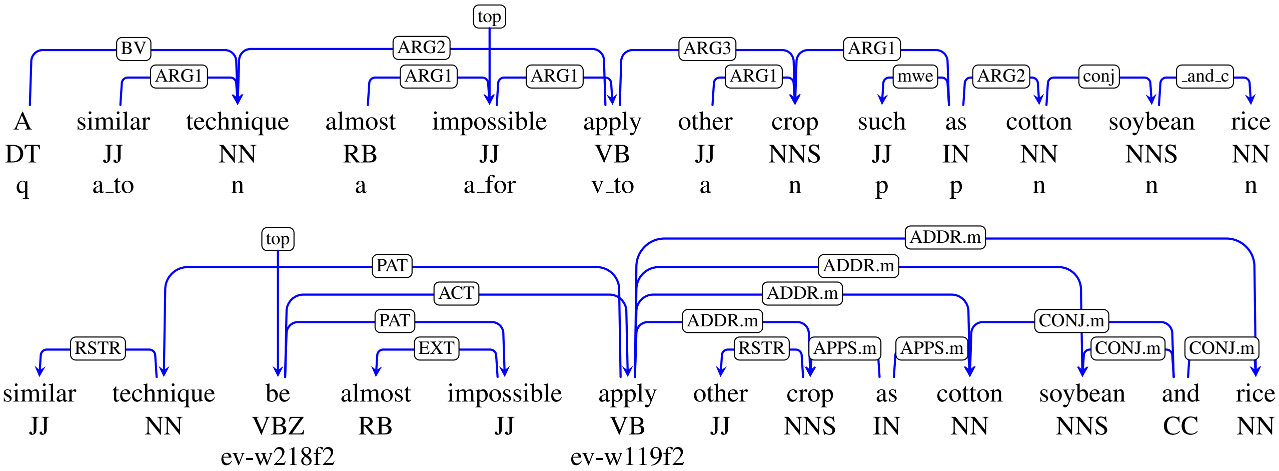

Figure 1: Sample bi-lexical semantic dependency graphs for the example sentence *A similar technique is almost impossible to apply to other crops such as cotton, soybeans, and rice.* The top graph shows DELPH-IN MRS Bi-Lexical Dependencies (DM), and the bottom one Prague Semantic Dependencies (PSD).

ple, transition-based parsers were more accurate on short dependencies thanks to a richer feature model, but degraded more because of error propagation in greedy decoding. Conversely, graph-based parsers showed a more graceful degradation thanks to global optimization and exact decoding, but had a disadvantage for local structures because of a more restricted feature model. More recently, Kulmizev et al. (2019) replicated this analysis for neural graph-based and transition-based parsers and showed that, although the distinct error profiles are still discernible, the differences are now much smaller and are further reduced by the use of deep contextualized word embeddings (Peters et al., 2018; Devlin et al., 2019).

**MRP 2019** The 2019 Shared Task at the Conference for Computational Language Learning (CoNLL) was devoted to Meaning Representation Parsing (MRP) across frameworks (Oepen et al., 2019). For the first time, this task combined *formally* and *linguistically* different approaches to meaning representation in graph form in a uniform training and evaluation setup. The training and evaluation data for the task comprised five distinct approaches—which all encode core predicate–argument structure, among other things—to the representation of sentence meaning in the form of directed graphs, packaged in a uniform abstract structure and serialization. This task design seeks to enable cross-framework comparison of different parsing approaches and to advance learning from complementary knowledge sources (e.g. via parameter sharing). The MRP 2019 competition received submissions from eighteen teams, and there will be a follow-up shared task, again hosted by CoNLL, in 2020.

Figure 1 shows two example graphs for one sentence from the venerable Wall Street Journal (WSJ) corpus in the two bi-lexical MRP frameworks (of five total), DELPH-IN MRS Bi-Lexical Dependencies (DM) of Oepen and Lønning (2006) and Ivanova et al. (2012), and Prague Semantic Dependencies (PSD) by Hajič et al. (2012) and Miyao et al. (2014). The DM and PSD frameworks are *bi-lexical* in the MRP collection, characterized by a one-to-one relation between graph nodes and surface tokens. But even within this limiting assumption, which makes these graphs formally somewhat similar to standard syntactic dependency trees, the examples in Figure 1 exhibit all the non-tree properties sketched in § 1. above (reentrancies, multiple roots, and semantically vacuous surface tokens). DM and PSD nodes are labeled with lemmas, parts of speech, and (for verbs only, in the PSD case) frame or sense identifiers; jointly, these properties define a semantic predicate. Edges represent semantic argument roles: DM mostly uses overtly order-coded labels, e.g. ARG1, ARG2, etc. Abstractly similar, PSD labels like ACT(or), PAT(ient), or ADDR(essee) indicate 'participant' positions in an underlying valency frame.

Regarding lexical anchoring, on the opposite end of the range of frameworks in the MRP 2019 shared task is Abstract Meaning Representation (AMR; Banarescu et al. (2013)), which by design does not spell out how nodes relate to sub-strings of the underlying parser input; Figure 2 shows the same example sentence in AMR. Without an explicit relation to the surface string, several of the 'querying'

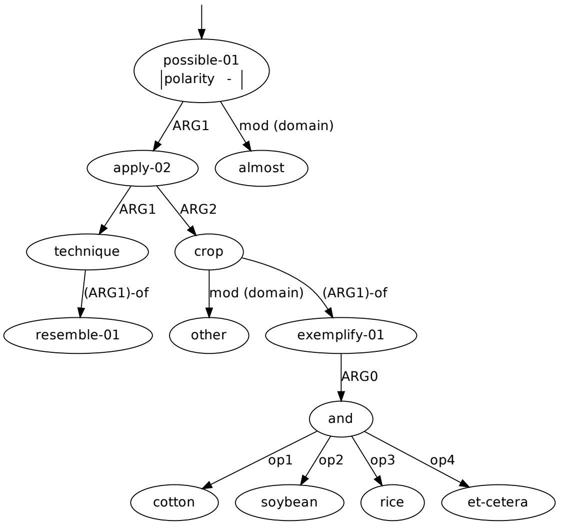

Figure 2: Sample unanchored Abstract Meaning Representation (AMR) graph for the same sentence as in Figure 1.

dimensions of McDonald and Nivre (2011) will need to either be derived or replaced by other structural properties, and for simplicity we focus subsequent discussion in §3. and §4. on the bi-lexical MRP frameworks. However, we will at times speculate about how relevant querying dimensions can be obtained for the more abstract frameworks, including AMR.

Evaluation in the MRP 2019 competition is in terms of $F_1$ scores at the level of individual graph elements, e.g. node labels, additional node-local properties, identification of the top node(s), and individual labeled edges. The latter two of these components of the MRP evaluation metric closely correspond to established evaluation practices in syntactic dependency parsing, essentially scoring isolated dependency edges. However, the sub-problem of node identification and labeling takes a much more prominent role in meaning representation parsing (even for the bi-lexical MRP graphs), and some of our reflections below explicitly seek to tease apart parser behavior on node-local vs. more structural predictions.

## 3. Methodological Reflections

When considering dimensions in which the accuracy of a meaning representation graph may be analysed, the queries can be separated into two broad categories: (1) structural factors, stemming from formal graph theory—the aspects of a tree or graph such as root node labels, and edge lengths; and (2) linguistic factors—related to underlying characteristics of the input strings, such as part-of-speech tags, and sentence length.

Drawing from the body of previous work on syntactic dependency tree analysis, we observe a number of dimensions with varying degrees of applicability to semantic dependency graphs. Furthermore, there are two marked differences between syntactic trees and meaning representation graphs that require additional attention. First, from a structural viewpoint, graph nodes allow for multiple incoming edges, as well as outgoing. Secondly, from a linguistically-informed viewpoint, meaning representation graphs also use the concept of node for an additional layer of information, with nodes having particular properties that differ by level of abstraction (in contrast to syntactic dependencies, where nodes are equivalent to tokens, and information on dependency relations is contained in labelled edges).

**Sentence Length** Universally, both syntactic and semantic parsers show lower accuracy for longer sentences. In the context of semantic dependencies, this is true regardless of the level of abstraction of a particular meaning representation. Longer sentences commonly contain complex syntactic constructions, which call for more parsing decisions to be made, and thus increase the chance of errors, as well as error propagation. In previous work, sentence length has been expressed in terms of the number of tokens. However, word count is less closely related to node count in meaning representation frameworks of higher levels of abstraction, where nodes may represent token substrings (e.g. affixes) or multiple tokens (e.g. multiword expressions), or there is no clear mapping at all between tokens and graph nodes. An alternative could be to measure sentence length at character level, but since the semantics of a single word does not necessarily depend on its length in characters, we propose calculating sentence length over nodes in the semantic dependency graph.

**Dependency Distance** Previous work on syntactic dependencies has shown that different parsing algorithms show variation in accuracy relative to dependency distance. As in the case of sentence length, long-range dependencies increase the chance of parsing errors. However, performance on lower- vs. shorter-range dependencies also depends on the particular parsing approach. As a side note, dependency distance is related to part of speech: shorter dependencies being typical, for example, of noun–adjective relations, while longer dependencies tend to be associated with predicates. In the case of more abstract meaning representations, it is not universally possible to define the connection between dependency distance, part of speech, and node anchoring, as discussed below.

**Distance to Root** In meaning representation graphs, an equivalent to the root node exists in the *top* node, but depending on the framework, multiple top nodes may exist (as, for example, in what is sometimes called run-on sentences, i.e. the juxtaposition of two independent clauses by e.g. a punctuation mark, such as a semicolon). Additionally, there is no guarantee of a directed path from the top node to any other node (as there would be in a directed tree). Several possible solutions exist in defining this dimension for semantic graphs. To solve the problem of guaranteed paths, one can generalise to undirected edges. In the case of multiple top nodes, the distance-to-root measure may be defined as the minimal distance to the closest top node. However, a search for this value raises the classic issue of efficient graph traversal.

**Types of Dependency Relations** This dimension, justified in the case of syntactic dependency analysis, does not translate well to semantic dependencies. Instead of the linguistically-motivated relation labels common to syntactic dependency parses (e.g. indirect object, attribute), semantic graph frameworks label edges using formal notions of structure (e.g. argument $n$, type-modifier). Furthermore, these formal labels are not necessarily coherent and comparable across frameworks—for example, ARG1 cannot be assumed to universally represent a subject-relation. A possible alternative to this dimension is quantification: consider parsing accuracy depending on the number of argument slots taken by a head.

**Parts of Speech** Evaluating, for example, the parsing accuracy for all (edges involving) nodes corresponding to nominal tokens is transferable from syntactic to semantic parsing only in the case of bi-lexical meaning representations, as discussed earlier, while the more abstract representations do not necessarily link their nodes to surface tokens and their morpho-syntactic properties. In other words, in a representation like AMR, there is no explicit distinction of, for example, ‘nominal’ vs. ‘non-nominal’ semantic predicates, even though there are of course, systematic underlying regularities.

**Node-Related Dimensions** One of the evaluation perspectives introduced in the MRP 2019 shared task is the

| | MRP Score | | | Ranking | | |
|---|---|---|---|---|---|---|
| | P | R | F | Overall | PSD | DM |
| Saarland | .83 | .80 | .819 | 4 | 1 | 4 |
| SJTU-NICT | .87 | .83 | .853 | 2 | 3 | 1 |
| HIT-SCIR | .87 | .85 | .862 | 1 | 4 | 2 |

Table 1: System scores and rankings in MRP 2019.

distinction between different types of semantic 'information', specifically properties of the dependency graphs local to individual nodes vs. ones that involve relations between nodes (edges) or the graph structure at large (top nodes). Employing this distinction in error analysis might reveal useful insights into parsing systems—e.g. which types of nodes are harder for the models to predict. However, as previously discussed, analysis involving node anchoring (into the underlying string) is only applicable in the case of bi-lexical representations. A more universal approach to node-local querying dimensions would be to group nodes by their in- and out-degree, or, more generally, by the number of undirected edges assigned to a node.

## 4. Pilot Study

We perform a small-scale empirical study as a first step towards in-depth contrastive analysis of semantic dependency parsing systems. For our initial experiments, we choose to compare three parsing systems, parsing into the two bi-lexical MRP frameworks, and analyzing parser performance depending on two querying dimensions: *input complexity* (string length) and *edge distance* (between two dependent nodes).

### 4.1. Data and Scoring

We restrict our pilot study to the PSD and DM frameworks, because (unlike for some of the other MRP 2019 target frameworks) their evaluation data is publicly available (Oepen et al., 2016). All statistics in this section are against the standard 3,359-sentence PSD and DM test set, comprising gold-standard graphs drawn from the WSJ and Brown corpora. Overall and component-wise MRP evaluation scores were computed using an instrumented version of the official scorer, the `mtool` Swiss Army knife of meaning representation.[1]

### 4.2. Parsing Systems

Our choice of models for contrastive evaluation was motivated by the characterisation of systems into three broad families of approaches, as presented, amongst others, by Koller et al. (2019) and Oepen et al. (2019): transition-, factorisation-, and composition-based parsers. Of these, the first two abstractly parallel the two families represented in the studies by McDonald and Nivre (2011), whereas composition-based parsing approaches are not found in syntactic parsing. We consider participating systems in the MRP 2019 competition, and, within each family of approaches, choose the top-performing systems for the PSD and DM frameworks.[2]

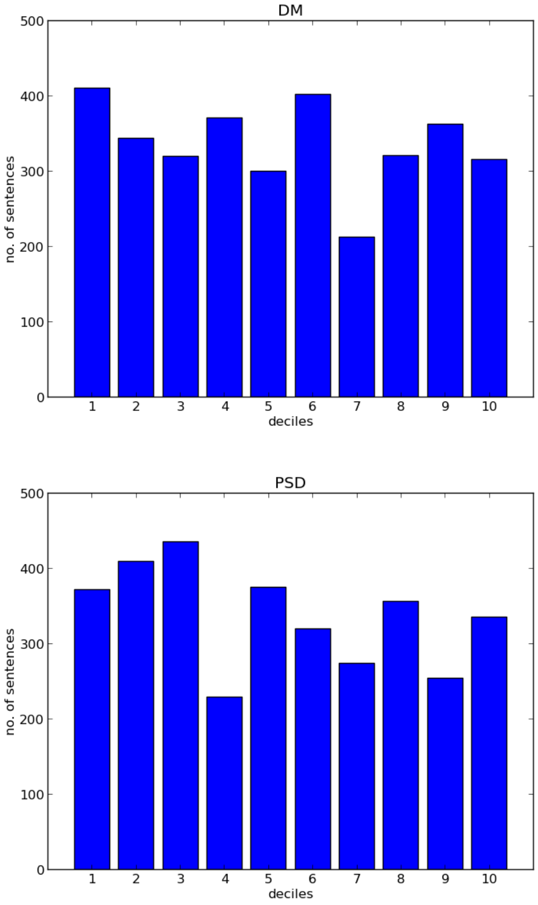

Figure 3: Distribution of sentences by length (node count), binned to ten aggregates.

Among the transition-based systems in MRP 2019, the best-performing parser is the HIT-SCIR parser (Che et al., 2019), which is also the top-performing overall parser; in the factorisation-based family, the SJTU-NICT system (Li et al., 2019) performs best on DM; and among the composition-based submissions, the Saarland system (Donatelli et al., 2019) obtains the best PSD results. Table 1 shows the absolute output quality (in terms of MRP precision, recall, and $F_1$) and the rankings of these systems on the PSD evaluation data, reproducing the official shared task results presented by Oepen et al. (2019).

The Saarland parser, an extension of Lindemann et al. (2019), uses a compositional approach, employing the Apply–Modify Algebra of Groschwitz et al. (2017) to build semantic graphs through highly constrained combinations of smaller graph fragments. A BiLSTM sequence labeling model is used for semantic tagging of word tokens, and the BiLSTM 'feature extractor' architecture of Kiperwasser and Goldberg (2016) is employed for predicting dependency trees, with input representations combining ELMo

[1] See `https://github.com/cfmrp/mtool` for details.

[2] We use framework-specific performance on PSD and DM, rather than the overall ranking across frameworks within the shared task, as the selection criterion, given that this pilot study is focused on comparing and analysing the results of parsing into these particular frameworks.

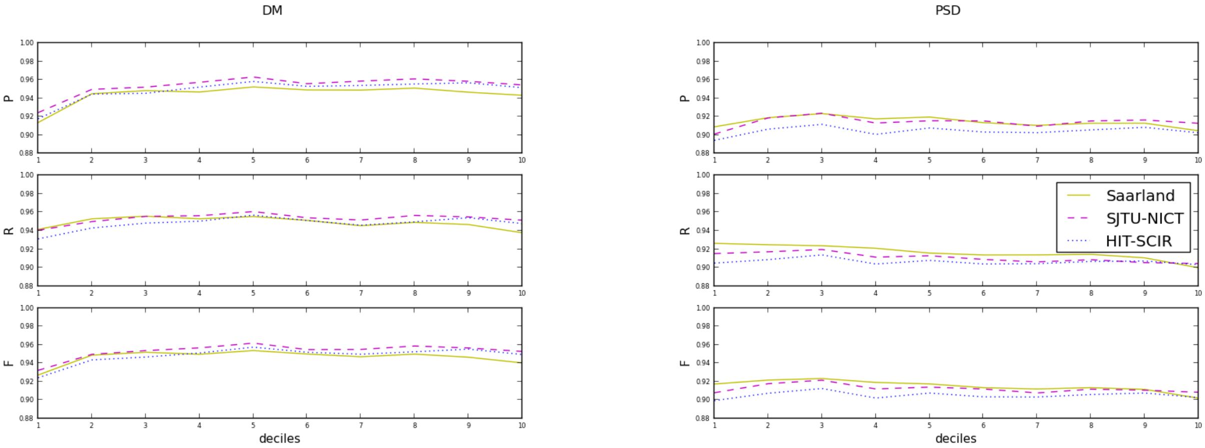

Figure 4: Overall MRP precision, recall, and $F_1$ by sentence length for DM (left) and PSD (right).

(Peters et al., 2018), and BERT (Devlin et al., 2019) contextualised word embeddings. Additionally, a decomposition step into subgraphs is necessary for training the model, which is handled using manually defined heuristics.

The SJTU-NICT parser is a factorisation-based (or 'graph-based', in the terminology of McDonald and Nivre (2007)) system, using a feed-forward network and a biaffine attention mechanism for edge and node property predictions on top of BERT embeddings. For the prediction of node-local properties, such as part-of-speech tags and frame labels for the PSD and DM frameworks, the parser also implements a multi-tasking objective.

Finally, the HIT-SCIR parser is an extended transition-based system designed to predict semantic graphs; it is the overall top-performing system in the shared task and the best transition-based parser for the PSD framework. The HIT-SCIR system is built upon the parser of (Wang et al., 2018), with the introduction of a stack LSTM architecture for batch training, and BERT contextualised word embeddings.

## 4.3. Preliminary Results

We conduct our study considering two querying dimensions of diagnostic evaluation—the length of the input sentence, and the distance between two dependent nodes.

**Sentence Length** Although the error analysis is focused on systems parsing into bilexical dependency graphs, which ensure an injective node-to-token mapping, this is not universal across meaning representations (see §2.). As proposed earlier, we choose to consider sentence length at the node level. Figure 3 shows the distribution of sentences by node length in decile bins.

Figure 4 plots the average P, R, and $F_1$ scores (i.e. the standard MRP metric) by sentence length at the node level. We find that the overall results for the two representations, DM and PSD, are fairly similar. DM results are in general somewhat higher, however, within the same range. We further observe only minor differences between the three parsers which exhibit very similar behaviours over the different sentence lengths. At this level of analysis, we do not observe the expected downward trend indicative of a drop in parsing accuracy for longer sentences. Rather, all three parsers seem relatively robust to sentence length, varying by less than 2 points over length bins.

Figure 5 plots the means of labeled $F_1$ for edges and $F_1$ for top nodes, comparable to labeled attachment score in syntactic dependency parsing evaluation. In capturing the structural properties when building an input sentence representation, there is a marked drop in accuracy for longer sentences, across all systems and for both representations. This is clear, despite an initial increase in performance which likely due to particularities of very short sentences (headings, fragments). We here observe clear differences between the different parsers. While the factorisation-based parser (SJTU-NICT) seems most resilient to the effects of longer inputs, the degradation is most prominent for the composition-based parser (Saarland). Overall the most successful parser on the PSD framework, the composition-based system, nevertheless suffers the most dramatic drop in results, and is universally the weakest-performing system when considering structural properties in isolation. Here there is also a clear difference between the two frameworks, where the Saarland parser exhibits a markedly more dramatic drop in results for DM as compared to PSD. Generally speaking, DM results are on average somewhat lower than the results for PSD, perhaps indicating that DM structural analysis is a harder task. This is possibly related to differences in formal graph properties between these two frameworks. Kuhlmann and Oepen (2016) present comparative statistics for all the MRP frameworks along several dimensions relating to nodes, treeness, and order. Their analysis shows that when it comes to the proportion of graphs that are rooted trees, there is a clear difference between the frameworks (2.31% vs. 42.24% for DM and PSD, respectively). DM furthermore exhibits a larger proportion of reentrant nodes (27.43% vs. 11.41%) as well as a much higher percentage of fragmented graphs (6.57% vs. 0.7% for DM and PSD, respectively).

These observations provide directions for future work in the cross-framework comparison of semantic parsers. More extensive testing is also needed to evaluate the role of training differences, i.e. the use of contextualised word embeddings (common across all three parsers) versus decomposition heuristics (specific to the composition-based system)

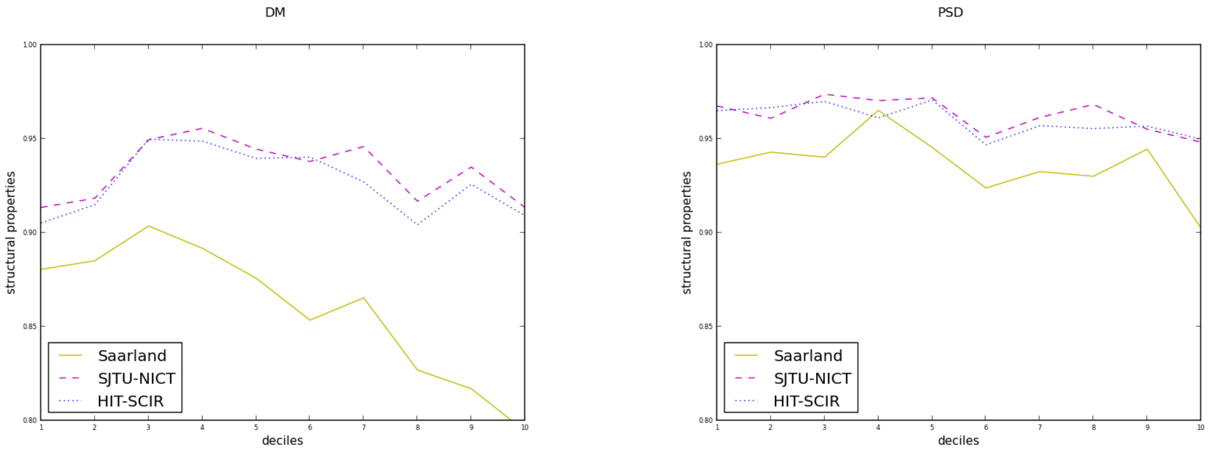

Figure 5: Structural $F_1$ (edges and top nodes) by sentence length for DM (left) and PSD (right).

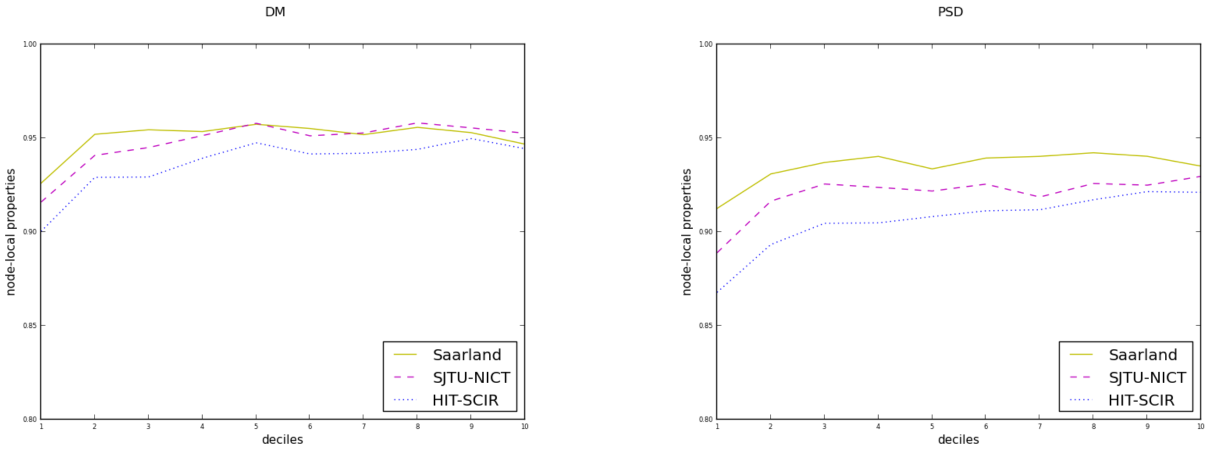

Figure 6: Node-local $F_1$ (labels and properties) by sentence length for DM (left) and PSD (right).

in fluctuation across sentence length.

So far, these findings are largely in line with those of previous studies, most notably (Kulmizev et al., 2019)—demonstrating that this dimension of analysis is indeed applicable to semantic dependency parsing. These preliminary observations potentially also point to a similar trend as seen with the introduction of neural networks to syntactic parsing: narrowing of the margin of difference in model performances.

By contrast, Figure 6 plots system performance considering the mean of $F_1$ scores for node properties and token-to-node anchors—a measure of how accurately the parsers capture node-local information. As discussed in §2., this concept has no clear equivalent in syntactic dependency parsing. We here observe a similar trend for all three parsers across the two representations; the prediction of node-local properties does not seem to be notably affected by sentence length and is fairly stable over sentences of increasing length. It is also clear that the Saarland parser, which is the top-performing system for the PSD representation, outperforms the other parsers for the task of node-local property prediction.

**Dependency Distance** The PSD and DM frameworks retain the word order of the input sentence, so we calculate the dependency distance between tokens $w_i$ and $w_j$ as $|i - j|$. Figure 7 shows the distribution of dependencies by distance in the dataset; we observe that there is a clear dominance of shorter dependencies in the gold data.

Looking at the systems' performance in the dependency distance dimension, there is a dramatic drop in performance with an increase in distance, as shown in Figure 8 ($F_1$ for edges). Parallel to the results for sentence length, we also observe differences between the three parsers, and in particular for the performance of the composition-based Saarland parser on the DM data. This raises the question: in which aspects of building a representation graph does this system demonstrate its strengths? Given the earlier mentioned importance of node-local information—not a matter of consideration in error analysis for syntactic parses—possibly other querying dimensions would help build a clearer picture of where particular approaches excel or fail.

## 5. Conclusion

We have given a methodological overview of previous error analyses for syntactic dependency parsing, and discussed the merits of particular query dimensions, as well as their applicability to semantic graphs. Using the evaluation results of the 2019 shared task on meaning repre-

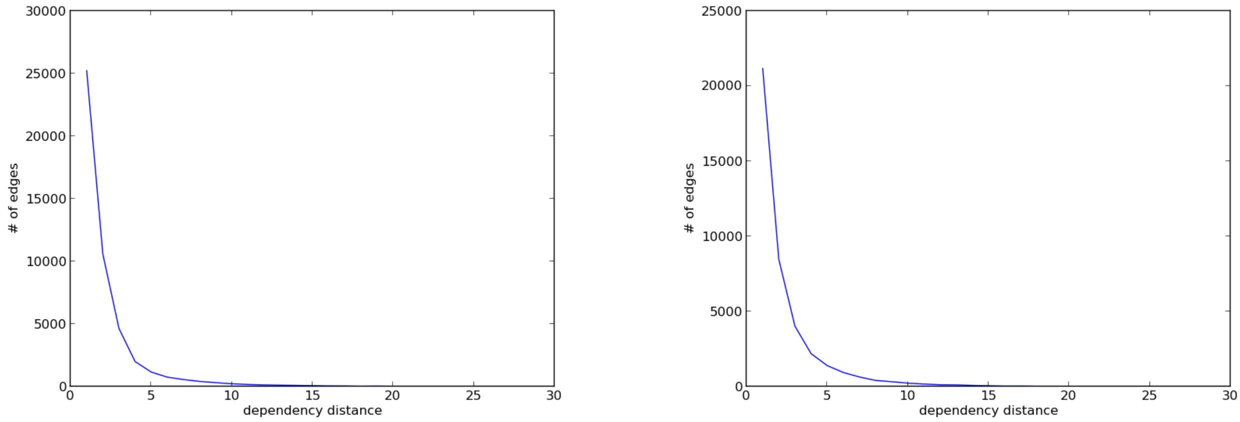

Figure 7: Distribution of edges by dependency distance.

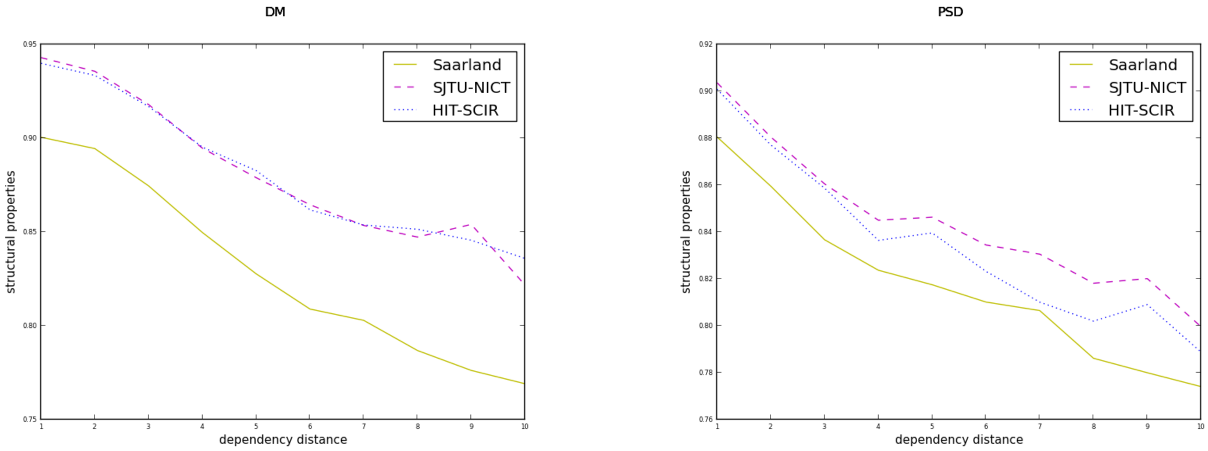

Figure 8: Structural $F_1$ for DM (left) and PSD (right).

sentation parsing (Oepen et al., 2019), we focused on two bi-lexical meaning representation frameworks, and three top-performing parsing systems. We conducted a small-scale empirical study, in line with similar work on syntactic parser analysis (McDonald and Nivre, 2011; Kulmizev et al., 2019), to verify our assumptions about query dimensions, necessary modifications, and possible additions.

The pilot study confirmed that there is merit to performing in-depth error analysis with the discussed query dimensions, but also served to highlight drawbacks of directly transferring the methodological approach. The experiments raised questions about modifying and extending the methodology—in particular, concerning the dichotomy between structural and node-local properties of meaning representation graphs.

Building upon this study, we intend to carry out a contrastive error analysis of semantic parsing systems through the discussed query dimensions. An additional challenge is to broaden the methodology to more abstract meaning representation frameworks.

## 6. Bibliographical References

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

Buchholz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Natural Language Learning*, pages 149 – 164, New York, NY, USA.

Che, W., Dou, L., Xu, Y., Wang, Y., Liu, Y., and Liu, T. (2019). HIT-SCIR at MRP 2019: A unified pipeline for meaning representation parsing via efficient training and effective encoding. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 76 – 85, Hong Kong, China.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA.

Donatelli, L., Fowlie, M., Groschwitz, J., Koller, A., Lindemann, M., Mina, M., and Weißenhorn, P. (2019). Saarland at MRP 2019: Compositional parsing across all graphbanks. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 66 – 75, Hong Kong, China.

Groschwitz, J., Fowlie, M., Johnson, M., and Koller, A. (2017). A constrained graph algebra for semantic parsing with AMRs. In *Proceedings of the 12th International Conference on Computational Semantics*, Montpellier, France.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3153 – 3160, Istanbul, Turkey.

Ivanova, A., Oepen, S., Øvrelid, L., and Flickinger, D. (2012). Who did what to whom? A contrastive study of syntacto-semantic dependencies. In *Proceedings of the 6th Linguistic Annotation Workshop*, pages 2 – 11, Jeju, Republic of Korea.

Kiperwasser, E. and Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Koller, A., Oepen, S., and Sun, W. (2019). Graph-based meaning representations. Design and processing. In *Proceedings of the 57th Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6 – 11, Florence, Italy.

Kuhlmann, M. and Oepen, S. (2016). Towards a catalogue of linguistic graph banks. *Computational Linguistics*, 42(4):819 – 827.

Kulmizev, A., de Lhoneux, M., Gontrum, J., Fano, E., and Nivre, J. (2019). Deep contextualized word embeddings in transition-based and graph-based dependency parsing: A tale of two parsers revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2755–2768, Hong Kong, China.

Li, Z., Zhao, H., Zhang, Z., Wang, R., Utiyama, M., and Sumita, E. (2019). SJTU–NICT at MRP 2019: Multi-task learning for end-to-end uniform semantic graph parsing. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 45 – 54, Hong Kong, China.

Lindemann, M., Groschwitz, J., and Koller, A. (2019). Compositional semantic parsing across graphbanks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4576 – 4585, Florence, Italy.

McDonald, R. and Nivre, J. (2007). Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Conference on Natural Language Learning*, Prague, Czech Republic.

McDonald, R. and Nivre, J. (2011). Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.

Miyao, Y., Oepen, S., and Zeman, D. (2014). In-House. An ensemble of pre-existing off-the-shelf parsers. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 63 – 72, Dublin, Ireland.

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

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

Oepen, S., Kuhlmann, M., Miyao, Y., Zeman, D., Cinková, S., Flickinger, D., Hajič, J., Ivanova, A., and Urešová, Z. (2016). Towards comparability of linguistic graph banks for semantic parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 3991 – 3995, Portorož, Slovenia.

Oepen, S., Abend, O., Hajič, J., Hershcovich, D., Kuhlmann, M., O'Gorman, T., Xue, N., Chun, J., Straka, M., and Urešová, Z. (2019). MRP 2019: Cross-framework Meaning Representation Parsing. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 1 – 27, Hong Kong, China.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, LA, USA.

Wang, Y., Che, W., Guo, J., and Liu, T. (2018). A neural transition-based approach for semantic dependency graph parsing. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhang, Y. and Clark, S. (2008). A tale of two parsers: Investigating and combining graph-based and transition-based dependency parsing using beam-search. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 562–571, Honolulu, HI, USA.