The long and the short of it: DRASTIC, a semantically annotated dataset containing sentences of more natural length

Dag T. T. Haug
University of Oslo, Norway
d.t.t.haug@ifikk.uio.no

Jamie Y. Findlay
University of Oslo, Norway
jamie.findlay@iln.uio.no

Ahmet Yıldırım
University of Oslo, Norway
ahmet.yildirim@iln.uio.no

Abstract

This paper presents a new dataset with Discourse Representation Structures (DRSs) annotated over naturally-occurring sentences. Importantly, these sentences are more varied in length and on average longer than those in the existing gold-standard DRS dataset, the Parallel Meaning Bank, and we show that they are therefore much harder for parsers. We argue, though, that this provides a more realistic assessment of the difficulties of DRS parsing.

1 Motivation

Corpora with deep, logic-based semantic annotations are quite rare because they are so hard to annotate. The arrival of the Groningen Meaning Bank (Bos et al., 2017) and the Parallel Meaning Bank (PMB; Abzianidze et al., 2017) changed this situation by offering full Discourse Representation Structures (DRSs; Kamp, 1981b) for substantial amounts of text in Dutch, English, German, and Italian. The current release, 4.0.0, contains more than 10,000 sentences in English and between 1,400 and 2,800 sentences in other languages. However, the dataset contains both bronze (automatic), silver (partial manual disambiguation), and gold (full manual disambiguation) data, and the gold sentences are consistently very short (mostly $<$10 words). Since the dev, test, and eval sets contain only gold data, this means that DRS parsers are tested only on very short sentences, yielding an overly optimistic assessment of results in this area.

In this paper, we improve on the situation by offering a gold standard dataset containing DRSs with a more realistic sentence length distribution. We call this dataset DRASTIC, for 'Discourse Representation Annotation with Sentence Texts of Increased Complexity'.[1] An additional strength of DRASTIC is that the texts it contains – three contiguous documents plus a selection of medium-length sentences – are from the GUM corpus (Zeldes, 2017), allowing users to explore connections between the DRS annotation and the rich annotation available in GUM: beside morphosyntactic annotation following the Universal Dependencies (UD) scheme (de Marneffe et al., 2021), this also includes entity recognition, coreference, discourse structure and more.[2] The current size of our dataset is small, at 157 sentences with full manual disambiguation, but around 1,000 more sentences have received a first manual annotation by student annotators and will subsequently be integrated into the dataset.

DRS parsing gets harder as sentences grow longer (cf. van Noord et al., 2020, 4594f.). This is natural, but some peculiarities of the PMB annotation are especially hard to capture, and contribute only little extra information. Cases in point are recursive presuppositions, strict separation of different presuppositions of a single sentence, and the use of discourse relations with relatively bland content such as CONTINUATION. As the sentence grows in length, these result in a complex network of embedded DRSs. In such cases, parser output that is (more or less) logically equivalent to the gold can still get a low score. To avoid this, we simplify the annotation of such structures (see Section 2.2). Since our corpus is small and does not include training data, we provide a script that flattens PMB-style annotations to our format. This can be used to flatten PMB data before training a parser, or alternatively to flatten the output of a parser trained on the PMB.

The structure of the paper is as follows. In Section 2, we introduce Discourse Representation Theory (DRT), as well as the PMB annotation and our simplifications of it. In Section 3, we describe our corpus, and Section 4 studies the effects of sen-

[1] The dataset and accompanying scripts are available here: https://github.com/Universal-NLU/DRASTIC.

[2] The list at https://gucorpling.org/gum/annotations.html (accessed 31 May 2023) provides the full set of annotation layers.

Proceedings of the 4th International Workshop on Designing Meaning Representations, pages 89–98
June, 2023.

tence length on DRS parsing and offers baseline modelling results on our data.

2 The format: Discourse Representation Structures

In Discourse Representation Theory (Kamp, 1981b; Kamp and Reyle, 1993; Kamp et al., 2011; Kamp and Reyle, 2019) the meaning of a sentence is analysed as its contribution to the existing semantic representation of the discourse context, called a Discourse Representation Structure (DRS). This means that DRT belongs to the family of theories called *dynamic semantics*, although DRT treats only the process of interpretation as dynamic, not the notion of meaning itself.[3]

DRSs are traditionally represented as boxes divided into two: a universe of discourse at the top, containing a number of *discourse referents*, which can then be referred to by the set of *conditions* in the lower part of the box. DRS conditions are by and large simply formulae of some predicate logic, but can also contain complex conditions relating multiple DRSs via logical operators like negation, implication, and disjunction, or modal operators like possibility and necessity. By way of illustration, Figure 1 gives a DRS for the sentence *Jadzia thought that Miles or Julian had been hurt*. DRT is compatible with many different specific theoretical approaches to semantics; in Figure 1, as in our corpus, we use a Neo-Davidsonian event semantics where events and states (collectively called *eventualities*) are treated as first-class entities in the ontology, and semantic dependents are related to their heads via thematic role predicates such as Agent, Patient, etc. (on event semantics see e.g. Davidson, 1967; Parsons, 1990). A basic representation of tense is also given, by including the relation Time between an eventuality and its time, and relating that time to the constant 'now' (referring to the time of utterance) or to other times.

Aside from the rich body of theoretical work in DRT exploring various knotty semantic phenomena such as anaphora (Kamp, 1981b; Haug, 2014), tense (Kamp, 1981a), rhetorical structure (Lascarides and Asher, 1993; Asher and Lascarides, 2003), propositional attitudes (Asher, 1986; Kamp, 1990), and others, one other good reason for using DRSs as our semantic representations is the existence of the Parallel Meaning Bank (PMB; Abzianidze et al., 2017), a multilingual corpus of DRS-annotated texts in English, Dutch, Italian, and German, to which we aim to contribute.

2.1 DRT in the PMB

The PMB makes a number of specific choices with regard to its DRS representations, which we endeavour to follow. Firstly, it represents DRSs not as graphical boxes, but as machine-readable text files, in a clausal format (van Noord et al., 2018a). An example PMB-style DRS and its corresponding translation into the clausal format is shown in Figure 2. Each clause begins with the label of a DRS (a 'box', hence the `b`), indicating where the condition is introduced. It then contains one of three types of condition: (1) a unary or binary predicate name, followed by its argument(s), as in `b1 scowl.v.01 e1`; (2) the explicit introduction of a discourse referent, as in `b1 REF e1`; or (3) a relation between DRSs, as in `b2 PRESUPPOSITION b1`, which states that the contents of DRS `b2` is a presupposition of `b1`.[4] Finally, the clause contains information about which word it originates from, and gives the character offsets of that word in square brackets.

As indicated in Figure 2, the PMB also represents presupposition, following the approach of Projective DRT (Venhuizen, 2015; Venhuizen et al., 2018). In the graphical representation, we indicate presupposed material with a prefixed asterisk, '`*`', since we flatten any embedded presupposition structure so that we just have a single box containing all presupposed material for the sentence (see Section 2.2 for more on our simplifications of the clausal format). On the clausal side, `b2 PRESUPPOSITION b1` means that DRS `b2` is a presupposition of DRS `b1`. A full list of PMB relations, including temporal relations (such as `TPR`, temporal precedence, used in Figure 2) is available on the PMB website.[5].

The PMB representations do not include any indication of number (singular vs. plural, etc.), nor of aspect, but they do contain detailed lexical semantic information, because each lexical concept, i.e. unary predicate, is identified with a WordNet synset (Fellbaum, 1998) indicating which

[3] Muskens (1994, 1996) provides a compositional interpretation of DRT using the lambda calculus, which also treats meaning itself as dynamic, thus uniting two divergent approaches within the dynamic semantics family.

[4] Other relations between DRSs used in the PMB follow the rhetorical relations of Segmented DRT (SDRT; see e.g. Asher and Lascarides 2003), but we do not use these in the DRASTIC corpus – see Section 2.2.

[5] https://pmb.let.rug.nl/drs.php (accessed 31 May 2023).

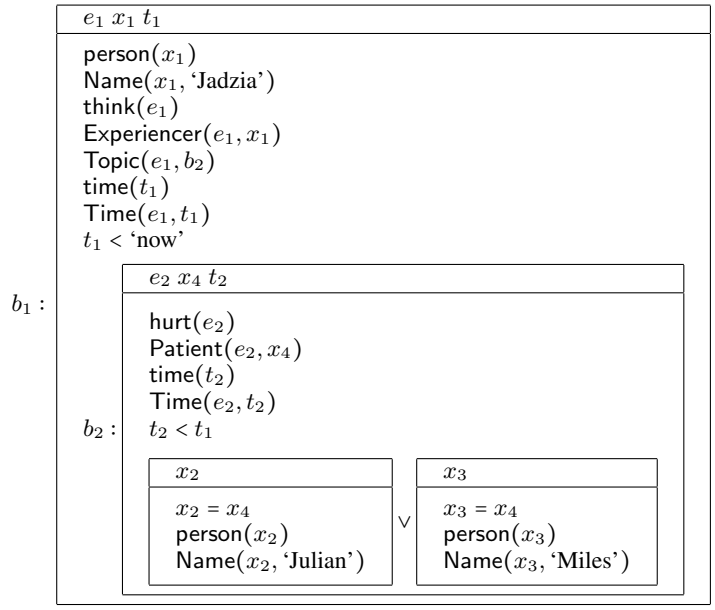
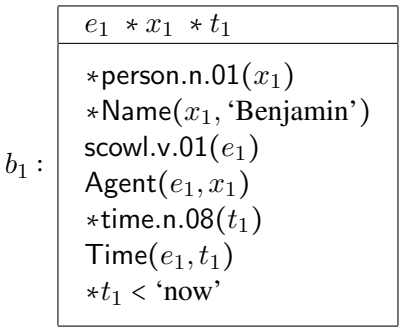


Figure 1: DRS for *Jadzia thought that Miles or Julian had been hurt*



(a) Graphical representation

```
b2 PRESUPPOSITION b1
b2 REF x1               % Benjamin [0...8]
b2 person.n.01 x1       % Benjamin [0...8]
b2 Name x1 "Benjamin"   % Benjamin [0...8]
b1 REF e1               % scowled [9...16]
b1 scowl.v.01 e1        % scowled [9...16]
b1 Agent e1 x1          % scowled [9...16]
b2 REF t1               % scowled [9...16]
b2 time.n.08 t1         % scowled [9...16]
b1 Time e1 t1           % scowled [9...16]
b2 TPR t1 "now"         % scowled [9...16]
```

(b) Clausal notation

Figure 2: Graphical vs. clause-based representation of a PMB-style DRS for the sentence *Benjamin scowled*

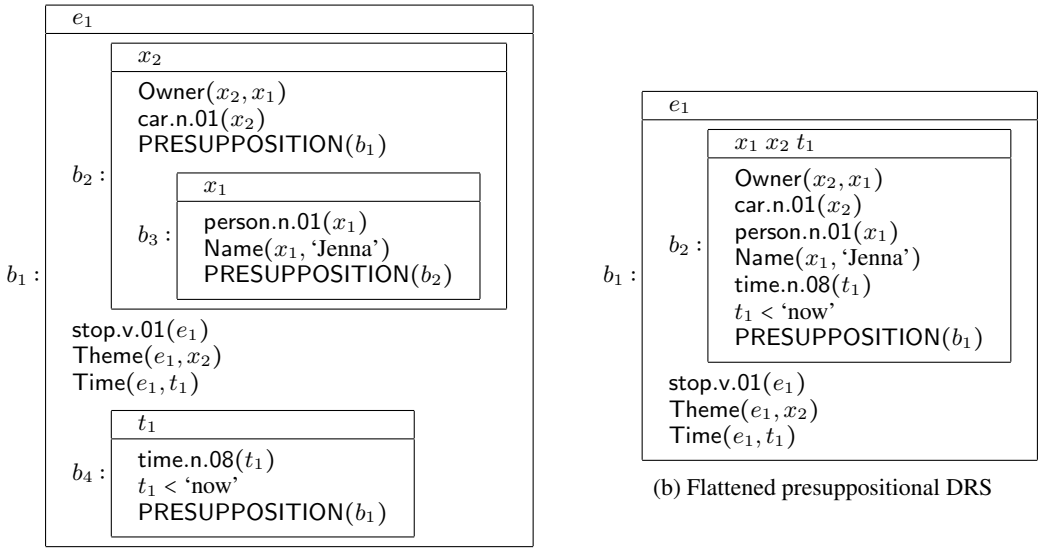
particular word sense is implicated. That is, the clause `b2 person.n.01 x1` indicates that the discourse referent `x1` falls under the first nominal sense of the lexeme PERSON listed in WordNet, i.e. a human being, as opposed to a body or the grammatical category (senses 2 and 3).

2.2 Simplifications

In general, the DRASTIC corpus follows the PMB annotation style, to allow the transfer of tools and techniques developed for the PMB, and in particular to provide test data involving longer sentences for the evaluation of parsers trained on the PMB. However, there are two areas in which we have chosen to simplify the PMB scheme in DRASTIC.

Firstly, we flatten DRSs by removing extraneous presuppositional sub-DRSs. To see what this means, consider the sentence *Jenna's car stopped*. Here we have (at least) three distinct existential presuppositions: the possessive construction presupposes the existence of Jenna's car; the proper noun *Jenna* itself introduces a presupposition that someone called 'Jenna' exists; and the past tense presupposes the existence of some time before the present. In PMB, this would result in three separate presuppositional DRSs, with two related directly to the main, outer DRS, and one related indirectly, via another of the presuppositional DRSs. This is shown in Figure 3a. In our own representations, all presuppositional material that originates in a given DRS is collapsed into a single sub-DRS, as shown in Figure 3b.

Since presuppositional material is ultimately not interpreted where it originates, but at the level to which it projects (on presupposition projection in DRT, see Venhuizen et al. 2018), this move is harmless with respect to the content of the DRSs in question. We lose track of which presuppositions



(a) Unflattened presuppositional DRS

(b) Flattened presuppositional DRS

Figure 3: Unflattened vs. flattened DRS for *Jenna's car stopped*

originated together, but this is not essential for interpretation. Moreover, when it comes to evaluating DRS parsing, we avoid many cases where logically equivalent DRSs are identified as distinct, owing to inconsequential differences in presupposition structure, which will then inappropriately suppress performance scores for DRS parsers.

This move also has the major advantage of making the representations easier for contemporary general-purpose neural networks to learn in the first place. As van Noord et al. (2018b, 619) observe, "DRSs are recursive structures and thus form a challenge for sequence-to-sequence models because they need to generate a well-formed structure and not something that looks like one but is not interpretable". By collapsing largely extraneous structure, we reduce one major source of difficulty for sequence-to-sequence models in producing DRSs.

The second simplification that we make is to eliminate rhetorical/discourse relations from our representations. This is more destructive than our first change since some such relations are genuinely informative (e.g. EXPLANATION). However, by far the most common relation in the PMB is CONTINUATION, the semantics of which reduces to conjunction, meaning that nothing is lost by eliminating it. Annotation of such rhetorical relations is also rather more subjective than other aspects of semantic annotation, which can inevitably lead to inconsistencies within or between annotators. Finally, removing these relations once again results in flatter DRSs, and so also serves to aid machine learning of DRS parsing.

However, since our corpus is too small to train a parser on our simplified format, model training must still rely on the PMB training set. Since most sentences there are very short, the structures that we simplify are unlikely to arise in large numbers; nevertheless, to make sure that the annotations are compatible, we provide a script that flattens PMB-style annotations as described above. This can be used to flatten the PMB data before training (to train a parser directly on this simplified format) or to flatten the output of a parser trained on the PMB directly. In Section 4, we report the results of some experiments using this second approach.

3 The corpus

3.1 The texts

The DRASTIC corpus consists of four sub-corpora: three entire documents from the biographical section of GUM, and one selection of shorter sentences drawn from different sub-parts of the GUM corpus.

The three biographical texts are Wikipedia articles relating to Czech composer Antonín Dvořák (`GUM_bio_dvorak`), YouTuber Jenna Marbles (`GUM_bio_marbles`), and translation theorist Eugene Nida (`GUM_bio_nida`), while the short texts corpus contains sentences 6–19 words long from 6 academic articles included in the 216 texts of the GUM corpus (the specific texts are shown in Table 1). Table 2 gives details about the size of the sub-corpora. 'Tokens' in this table refers to orthographic words separated by whitespace or hyphens,

GUM_academic_art
GUM_academic_census
GUM_academic_eegimaa
GUM_academic_enjambment
GUM_academic_epistemic
GUM_academic_games

Table 1: GUM texts from which the short-texts corpus draws

Sub-corpus	Sentences	Tokens	UD tokens
dvorak	28	668	678
marbles	43	842	926
nida	46	878	917
short-texts	40	512	539
TOTAL	157	2900	3060

Table 2: Size breakdown of the DRASTIC corpus

and to some punctuation characters (., ,, !, ?, ;). The UD tokenisation used in the GUM CoNLL-U files is more morphosyntactically motivated (e.g. possessive *'s* is separated from its host), and as such gives a larger number.

The major contribution of our corpus is that the sentence length distribution is more evenly spread and has a far wider range than that of the PMB data (especially the test set). For instance, the median sentence length in our corpus is 17, compared to 8 in the PMB data overall, and 6 in the PMB test set. The full distributions are shown in Figure 4, while Table 3 gives some further descriptive statistics about sentence length across the (sub-)corpora.

(Sub-)corpus	Median	Mean	St.dev.
dvorak	23	23.9	9.68
marbles	17	19.6	12.4
nida	18	19.1	11.1
short-texts	13	12.8	4.29
DRASTIC (all)	17	18.5	10.6
PMB (all)	8	10.0	9.53
PMB (test only)	6	6.60	2.08

Table 3: Sentence length across (sub-)corpora

Although it only has a modest number of sentences, the DRASTIC corpus nevertheless also manages to exemplify a range of complex linguistic phenomena, including negation, modal expressions, meta-linguistic usage, appositions, relative clauses, complement clauses, and a variety of other kinds of multi-clausal structure.

3.2 The annotation procedure

In the first instance, our annotation procedure follows that of the PMB as described in Abzianidze et al. (2017).[6] Our texts were uploaded to the PMB, where they were automatically analysed on several layers: tokenisation, CCG parsing, semantic tagging and WordNet sense selection. With this information, the Boxer system (Bos, 2008, 2015) then automatically produces a DRS representation for the sentence. All layers were subject to manual correction by trained annotators, and annotations were harmonised through weekly meetings and subsequent retagging of texts. This was done for around 1,000 sentences. For the 157 sentences released in the current version of the corpus, all sentences were also checked by the authors of this paper, and this process will continue.

The PMB interface imposes compositionality, in the sense that the final representation cannot be edited; only the representation of the tokens can be changed, and Boxer will then assemble a new representation of the sentence. While this is theoretically desirable, it can be practically limiting. Consider the sentence *She paid $800 rent by working various jobs, like bartending, working at a tanning salon, blogging, and go-go dancing at nightclubs.* Because this is a sequence of coordinated gerund VPs, Boxer produces disjoint DRSs connected by the discourse relation CONTINUATION. By manual intervention, we can instead make sure that the gerunds are coordinated to form one complex event, which bears an Instrument role to the matrix event.

To deal with such complicated cases, we therefore exported our data from the PMB and manually corrected remaining errors. Because we wanted to factor out anaphora resolution as a separate task, we exported the data with no anaphora resolved. However, all cases of sentence-internal anaphoric reference were noted, and we distribute the data in two versions: with and without anaphoric resolution. The former is the standard of the PMB and was used in our subsequent experiments. Unfortunately, it is not easy to represent cross-sentential anaphoric references when each DRS represents one sentence only; for that the DRSs must be merged or connected by discourse relations. This is a task for full-fledged discourse parsing, which we do not

[6] We are grateful to the PMB team, in particular Johan Bos and Rik van Noord, for helping us with both technical and linguistic issues in using the PMB interface.

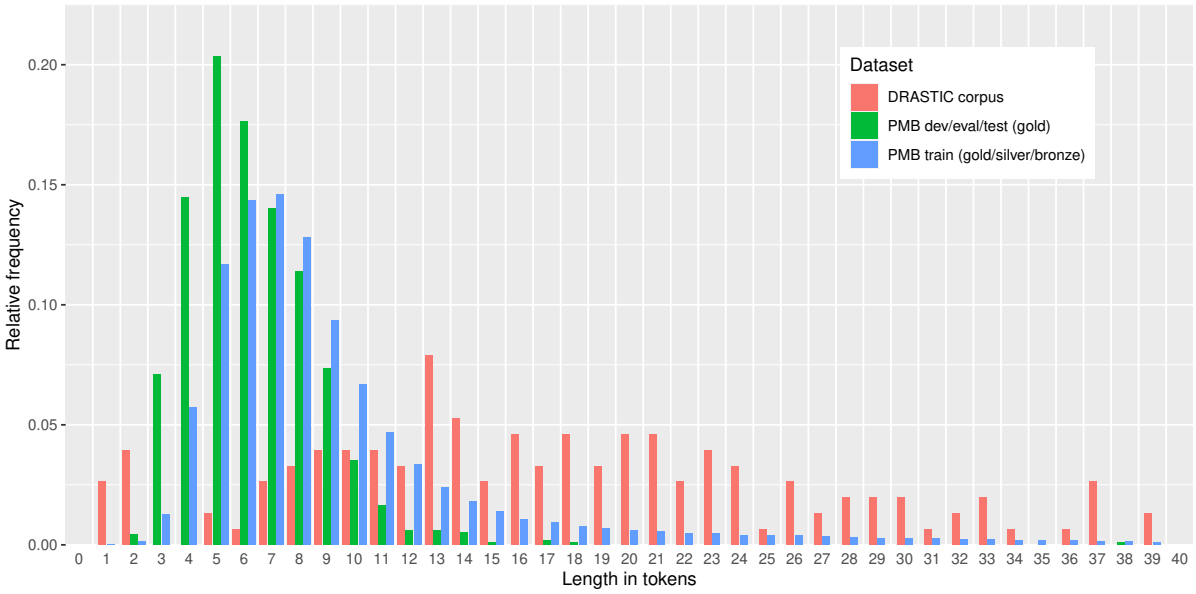


Figure 4: Length distribution in PMB 4.0.0 datasets compared to our corpus

attempt here. However, it is worth noting that this is an area where the additional annotation layers of the GUM corpus will be particularly useful. In this instance, the discourse annotation layer, based on Rhetorical Structure Theory (Mann and Thompson, 1988; Taboada and Mann, 2006), may aid in, for example, reconstructing the SDRS rhetorical/discourse relations which DRASTIC omits.

3.3 The format

Each of the two versions of the corpus, with and without anaphoric resolution, is provided as a set of files, one for each sentence, named after their GUM sent_id. Each file contains the raw text of the sentence and its clausal DRS annotation. We make the connection from the DRS annotation to both the original text and the GUM UD format explicit by decorating clauses in our data not only with character offsets, as shown in Figure 2, but also with UD token offsets, taken from the CoNLL-U files. This indicates which word(s) the clause in question originates from.

4 Modelling results

4.1 State of the art DRS parsing

Work on DRS parsing has recently involved applying deep neural networks. The majority of the work in this area (van Noord et al., 2018b; van Noord, 2019; Evang, 2019) has used sequence-to-sequence (seq2seq) LSTMs (Hochreiter and Schmidhuber, 1997). Table 4 presents recently reported performances of DRS parsing on the PMB datasets, along with the best results from our seq2seq experiments (Yıldırım and Haug, 2023), which, unlike previous work, also reports results on PMB 4.0.0.[7] We trained this state-of-the-art parser following the design principles used by van Noord et al. (2020), but instead of an LSTM we used transformer-based encoders and decoders. Here, we report the results obtained by using bert_base_cased as a frozen encoder along with a non-pretrained (randomly initialized) transformer as the decoder (12 layers, 12 attention heads per layer, using the Wordpiece tokenizer (Wu et al., 2016) used by the input (encoder) for the output as well).

The results in Table 4, with F1 scores in the high 80s/low 90s, clearly leave room for improvement, but do suggest that DRS parsing is a relatively straightforward task for current systems. The results are better, for example, than state-of-the-art parsing for Abstract Meaning Representation (AMR; Langkilde and Knight, 1998), which is in the low to mid 80s (Bai et al., 2022). This is surprising, because the expressive power of AMR is strictly less than that of DRT (Bos, 2016), and because the PMB DRSs capture many phenomena that AMR ignores, particularly involving scope.

However, there is reason to believe that DRS parsing as evaluated on the PMB test set understates the difficulty of the task. One issue that was

[7] Poelman et al. (2022) report performances of parsing Discourse Representation Graphs (DRG), a simpler form of DRSs, using PMB 4.0.0.

	PMB 2.2.0		PMB 3.0.0		PMB 4.0.0			
	dev	test	dev	test	dev	test	eval	DRASTIC
van Noord et al. (2020)	86.1	88.3	88.4	89.3	–	–	–	–
Liu et al. (2021)	–	88.7	–	–	–	–	–	–
Yıldırım and Haug (2023)	87.5	89.2	89.8	90.3	88.1	89.0	86.9	36.2

Table 4: Recently reported F1 scores for PMB 2.2.0, 3.0.0, and 4.0.0 datasets, and our result for DRASTIC

noticed by van Noord et al. (2020, 4594f.) is that, unsurprisingly, all models in their experiments performed worse as sentences got longer. In this context, the short length of the sentences in the PMB test set becomes especially noteworthy. The distribution of sentence lengths in the PMB was already shown in Figure 4. We see that it is very different between the training set and the dev/eval/test sets. As noted above, this is because the latter only include data that have been fully corrected manually – generally very short sentences – whereas the training set also contains data with no or only partial manual disambiguation, and those sentences are much longer. This mismatch is in itself a potential problem and may be the reason why several teams have fine-tuned their models on only the gold data of the training set, which has a similar length distribution to that of the test set.

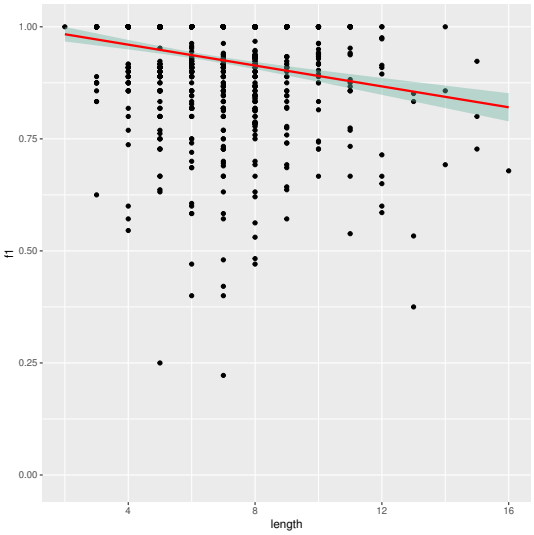


Figure 5: Performance vs. length in the PMB test set obtained by using the model reported under PMB 4.0.0 in Table 4 by Yıldırım and Haug (2023)

4.2 DRS parsing and sentence length

More worrying than the mismatch between training and evaluation is the overall short length of the sentences in the PMB dataset. We observe that sentences longer than 10 tokens are very rare. This is quite different from what one encounters in most genres of running text. Owing to the small range of sentence lengths in the PMB test set, the deleterious effect of increased length noted by van Noord et al. (2020) is only weakly felt there. The correlation between sentence length and F1 score in the PMB test set has a Pearson's r value of -0.21 ($p < 5 \times 10^{-10}$), a trend shown in Figure 5 (with regression line and 95% confidence intervals). In the DRASTIC data, with its more varied sentence lengths, the correlation with F1 scores is slightly more pronounced, as shown in Figure 6 (Pearson's $r = -0.29$, $p < 4 \times 10^{-4}$). Nevertheless, it is still fairly weak. Although longer sentences may confuse the transformer architecture by virtue of their length alone (because there was little or no data with the same positional encodings in the training phase), linguistic complexity (e.g. the presence of negation or other scopal operators, along with embedded structures) is another, semi-orthogonal source of difficulty, which will affect performance independently of length. Of course, the two are not entirely unrelated, since longer sentences also tend to be linguistically more complex (especially in terms of sentential embedding), exhibiting more structures that are rarely seen in the training data.

4.3 Performance on our dataset

Since our data structures are simplified ('flattened') compared to the PMB annotations, as described in Section 2.2, we transform the output of our parsers, which are trained on the original PMB data. This is done automatically in three steps:[8]

1. Removing discourse relations: Each clause of the form `x REL y`, where `REL` is one of `CONTINUATION`, `CONTRAST`, `ELABORATION` or `EXPLANATION`, is eliminated. All occurrences of the box variable `x` are replaced by `y` in all clauses.

[8]The script to perform this transformation has been made available along with the data at https://github.com/Universal-NLU/DRASTIC.

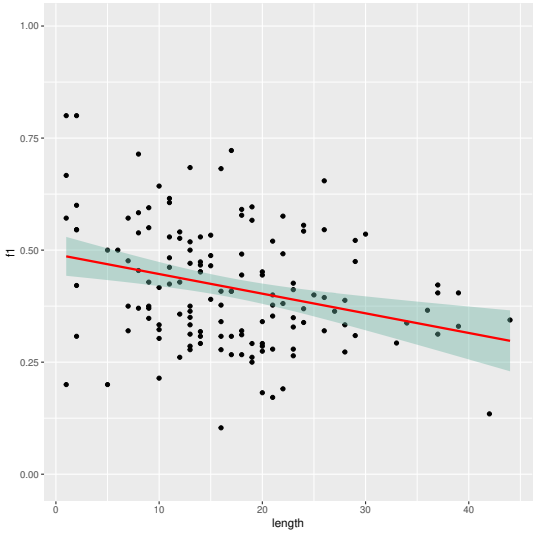


Figure 6: Performance vs. length in the DRASTIC corpus obtained by using the model reported under PMB 4.0.0 in Table 4 by Yıldırım and Haug (2023)

2. Flattening recursive presuppositions: for all occurrences of pairs of clauses of the form `x PRESUPPOSITION y`, `y PRESUPPOSITION z`, we remove the first clause and replace all occurrences of the box variable `x` by `y`.

3. Grouping presuppositions: for all occurrences of clause pairs `x PRESUPPOSITION y`, `z PRESUPPOSITION y`, we remove the first clause and replace all occurrences of the box variable `x` by `z`.

For the purposes of this paper, we perform these transformations on the output of the DRS parser before measuring performance on our dataset. This allows us to use the same model both on PMB data (with unflattened output) and on our data (with flattened output). As an alternative, it would be possible to train the model on flattened PMB output, so that the model will have seen the simplified structures directly during training; we leave this for future research.

We saw in Figure 6 the performance of our best model across sentences of different lengths in the DRASTIC corpus. Often for longer sentences the output of the model contains far fewer clauses than the gold data, suggesting an effect of length alone. But the model also performs much worse on DRASTIC than on the PMB in general, as witnessed by the low F1 score of 36.2 shown in Table 4.[9] Partly, this is because our dataset is more linguistically complex than the PMB. Sentences involving negation, for example, cause particular problems, and the negative meaning is often absent from the model output. Interaction between scopal elements such as negation and modality is also difficult: for the sentence *While the impact of a translation may be close to the original, there can be no identity in detail*, the model incorrectly stacks the possibility operators and flips the scope of negation and possibility, so that the meaning of the second clause becomes "it is possible that it is possible that there is no identity in detail", while in *This is, perhaps, not the best example of the technique* . . . , the negation disappears altogether.

Linguistic complexity cannot be the whole story, however. There are also unusual errors such as names that occur in our data but not in the PMB being incorrectly rendered in the parser output: e.g. the name "Marbles" becomes 'georgia strawberry', 'margau', 'margis', and 'name'. It is surprising to see such behaviour in a parser that performs so well on the PMB test set. This might indicate that the models overfit on peculiarities of the PMB.[10] A deeper investigation into what causes this drop in performance is clearly required – for example, one could replace names in the DRASTIC corpus with frequently-occuring names in the PMB to see if this improves performance. Whatever the exact origins of these deficiencies turn out to be, we believe our more varied data can contribute to more robust DRS parsers, especially as DRASTIC grows in size.

5 Summary

We have presented a new dataset, the DRASTIC corpus, which contains PMB-style DRSs annotated over sentences with more realistic lengths than the original PMB dataset, and which is, accordingly, much more of a challenge for state-of-the-art parsers. We hope that this will lead both to a more realistic assessment of the difficulty of DRS parsing and, in the longer term, to the development of more robust models.

[9] And this is true even when length is held constant: for example, in the PMB, the majority of sentences of length 8 are parsed to an F1 score of 0.75 or higher, whereas in our data, only one "sentence", of length 1, gets a score at this level.

[10] As an anecdotal example, we can mention that that 15-20% of the sentences across the PMB subsets contain the proper name *Tom*.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Nicholas Asher. 1986. Belief in Discourse Representation Theory. *Journal of Philosophical Logic*, 15:127–189.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Johan Bos. 2008. Wide-coverage semantic analysis with Boxer. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 277–286. College Publications.

Johan Bos. 2015. Open-domain semantic parsing with boxer. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 301–304, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Johan Bos. 2016. Squib: Expressive power of Abstract Meaning Representations. *Computational Linguistics*, 42(3):527–535.

Johan Bos, Valerio Basile, Kilian Evang, Noortje Venhuizen, and Johannes Bjerva. 2017. The groningen meaning bank. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, volume 2, pages 463–496. Springer.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The logic of decision and action*, pages 81–120. University of Pittsburgh Press, Pittsburgh, PA.

Kilian Evang. 2019. Transition-based DRS parsing using stack-LSTMs. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.

Dag T. T. Haug. 2014. Partial dynamic semantics for anaphora: compositionality without syntactic coindexation. *Journal of Semantics*, 31(4):457–511.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hans Kamp. 1981a. Événements, représentation discursive et référence temporelle. *Langages*, 64:39–64. [English version published in 2017 in *Semantics & Pragmatics* 10(2), available online here: https://doi.org/10.3765/sp.10.2].

Hans Kamp. 1981b. A theory of truth and semantic representation. In Jeroen Groenendijk, Theo M. B. Janssen, and Martin Stokhof, editors, *Formal methods in the study of language*, pages 277–322. Mathematical Centre Tracts, Amsterdam.

Hans Kamp. 1990. Prolegomena to a structural account of belief and other attitudes. In C. Anthony Anderson and Joseph Owens, editors, *Propositional attitudes – the role of content in logic, language, and mind*, pages 27–90. CSLI Publications, Stanford, CA.

Hans Kamp, Josef van Genabith, and Uwe Reyle. 2011. Discourse Representation Theory. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of philosophical logic*, 2nd edition, volume 15, pages 125–394. Springer, Berlin.

Hans Kamp and Uwe Reyle. 1993. *From discourse to logic*. Kluwer, Dordrecht.

Hans Kamp and Uwe Reyle. 2019. Discourse Representation Theory. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: theories*, pages 321–384. De Gruyter, Berlin.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 704–710, Montreal, Quebec, Canada. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations, and commonsense entailment. *Linguistics and Philosophy*, 16:437–449.

Jiangming Liu, Shay B. Cohen, Mirella Lapata, and Johan Bos. 2021. Universal discourse representation structure parsing. *Computational Linguistics*, 47(2):445–476.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Reinhard Muskens. 1994. A compositional Discourse Representation Theory. In Paul Dekker and Martin Stokhof, editors, *Proceedings of the 9th Amsterdam Colloquium*, pages 467–486. ILLC, Amsterdam.

Reinhard Muskens. 1996. Combining Montague semantics and discourse representations. *Linguistics and Philosophy*, 19:143–186.

Rik van Noord. 2019. Neural boxer at the IWCS shared task on DRS parsing. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden. Association for Computational Linguistics.

Rik van Noord, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. 2018a. Evaluating scoped meaning representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rik van Noord, Lasha Abzianidze, Antonio Toral, and Johan Bos. 2018b. Exploring neural methods for parsing discourse representation structures. *Transactions of the Association for Computational Linguistics*, 6:619–633.

Rik van Noord, Antonio Toral, and Johan Bos. 2020. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

Terence Parsons. 1990. *Events in the semantics of English: a study in subatomic semantics*. MIT Press, Cambridge, MA.

Wessel Poelman, Rik van Noord, and Johan Bos. 2022. Transparent semantic parsing with Universal Dependencies using graph transformations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4186–4192, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Maite Taboada and William C. Mann. 2006. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Noortje J. Venhuizen. 2015. *Projection in discourse: a data-driven formal semantic analysis*. Ph.D. thesis, University of Groningen.

Noortje J. Venhuizen, Johan Bos, Petra Hendriks, and Harm Brouwer. 2018. Discourse semantics with information structure. *Journal of Semantics*, 35(1):127–169.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint: 1609.08144.

Ahmet Yıldırım and Dag Trygve Truslew Haug. 2023. Experiments in training transformer sequence-to-sequence DRS parsers. In *Proceedings of the 15th International Conference on Computational Semantics (IWCS 2023)*, Nancy, France. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.