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

Dag T. T. Haug, Jamie Findlay and Ahmet Yıldırım

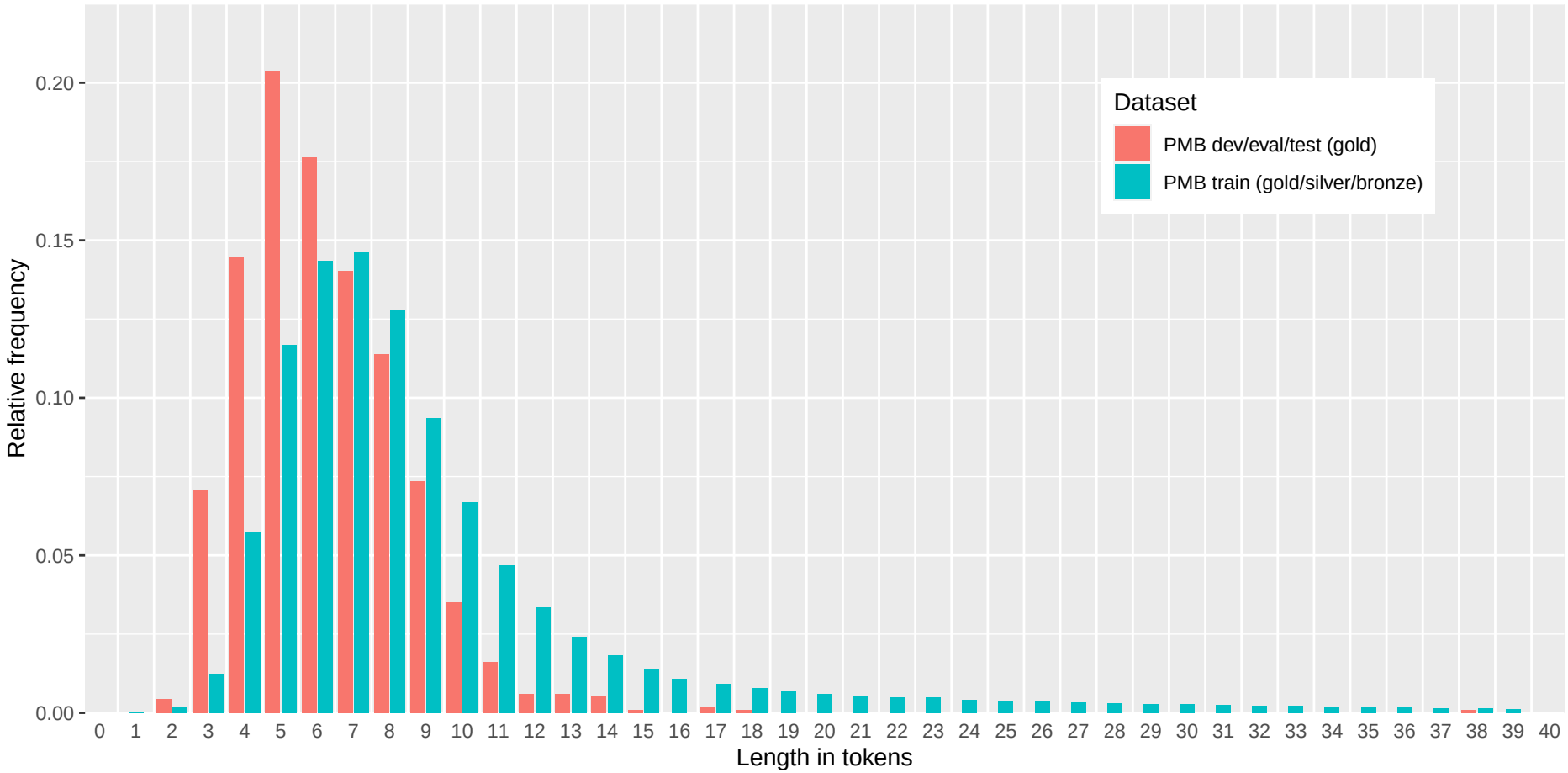
University of Oslo

20 June 2023
DMR 2023

The problem

- Corpora with deep, logic-based semantic annotations are quite rare because they are so hard to annotate.
- The **Parallel Meaning Bank (PMB)** is a major exception, containing texts annotated with **Discourse Representation Structures (DRSs)**.
  - >10,000 sentences in English
  - 1,400–2,800 sentences in each of Dutch, German, and Italian
- **However**:
  - data includes **gold**, **silver**, and **bronze** annotations
  - **gold** sentences are very short (mostly <10 words)
  - dev/test/eval sets contain only **gold**
  - **so DRS parsers are tested against only very short sentences!**

Sentence length in the PMB



The problem

- DRS parsing gets harder as sentence length increases.

  (van Noord et al. 2020b: 4594f.)

- Aside from greater string length, longer sentences are also more likely to introduce linguistic complexities:
  - embedding structures
  - coordination
  - interacting scopal elements

- Structural generalisations are also hard for seq2seq models.

  (Yao and Koller 2022; Donatelli and Koller 2023)

- **So testing against short sentences gives an overly optimistic account of parsers' performance.**

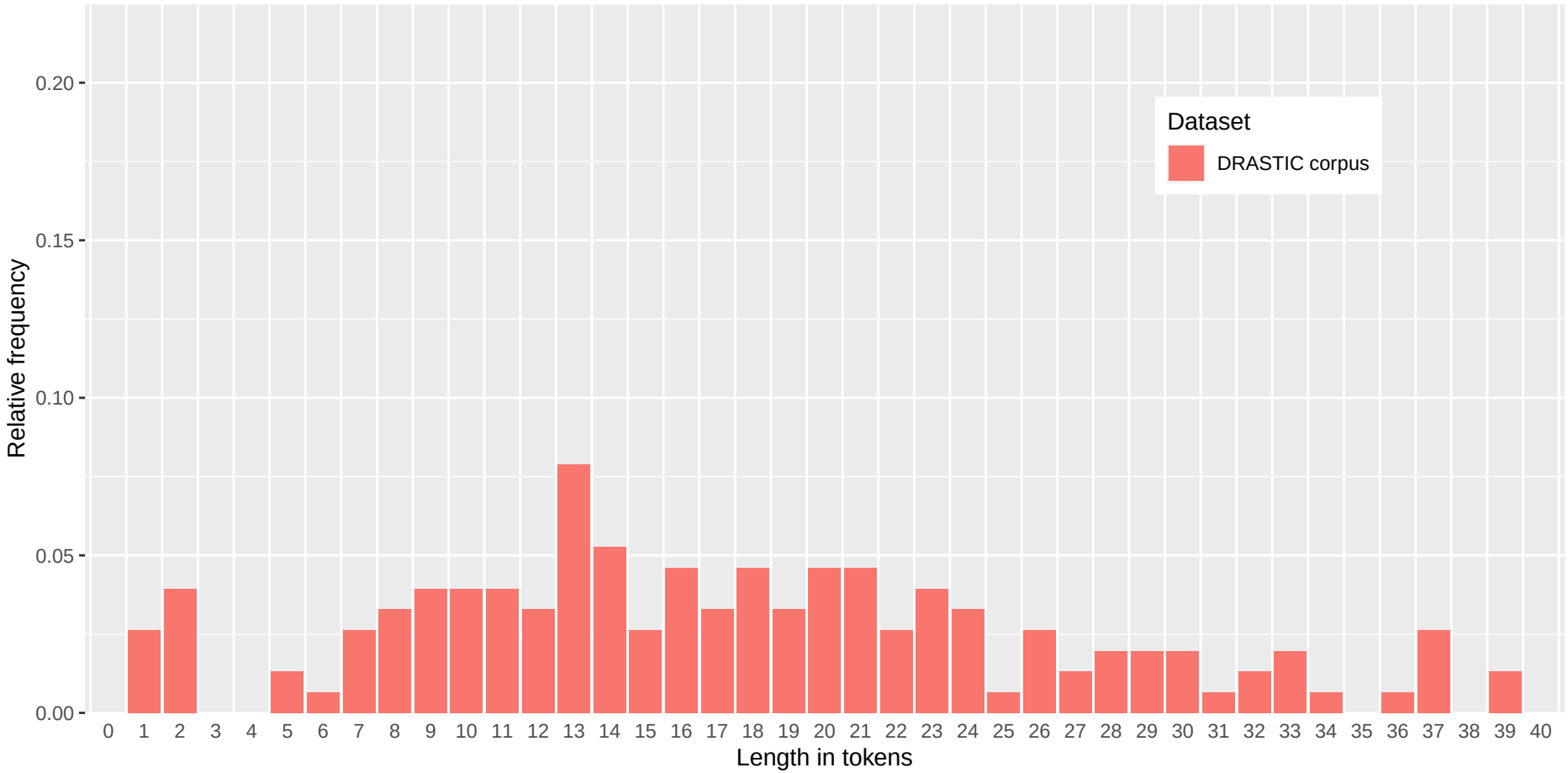
The solution

- DRASTIC: Discourse Representation Annotations with Sentence Texts of Increased Complexity.
- Texts drawn from the biographical and academic sections of the GUM corpus. (Zeldes 2017)

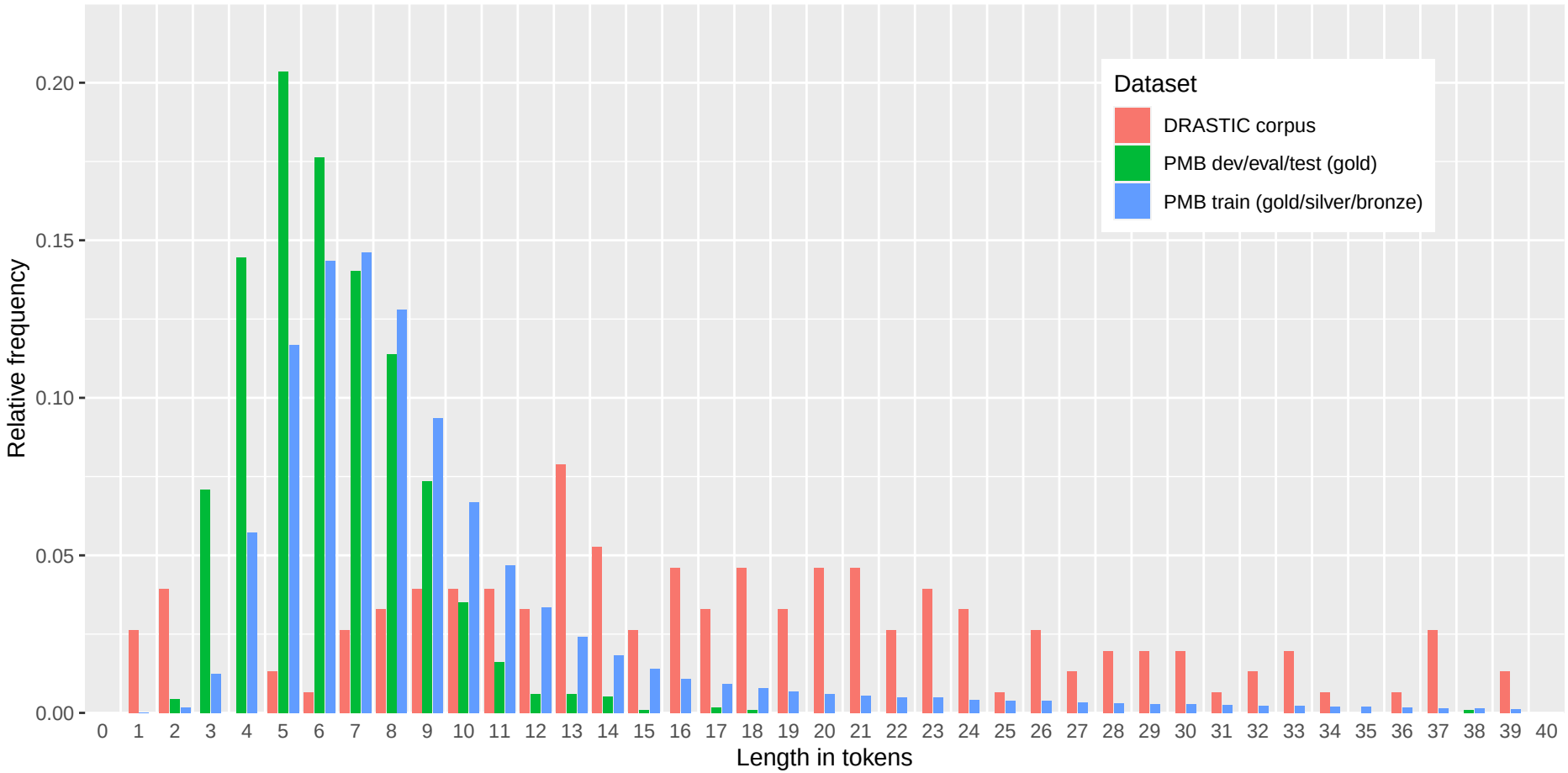
Sub-corpus	Sentences	Tokens	UD tokens
`dvorak`	28	668	678
`marbles`	43	842	926
`nida`	46	878	917
`short-texts`	40	512	539
TOTAL	157	2900	3060

Table: Size breakdown of the DRASTIC corpus

Sentence length in DRASTIC



Sentence lengths compared

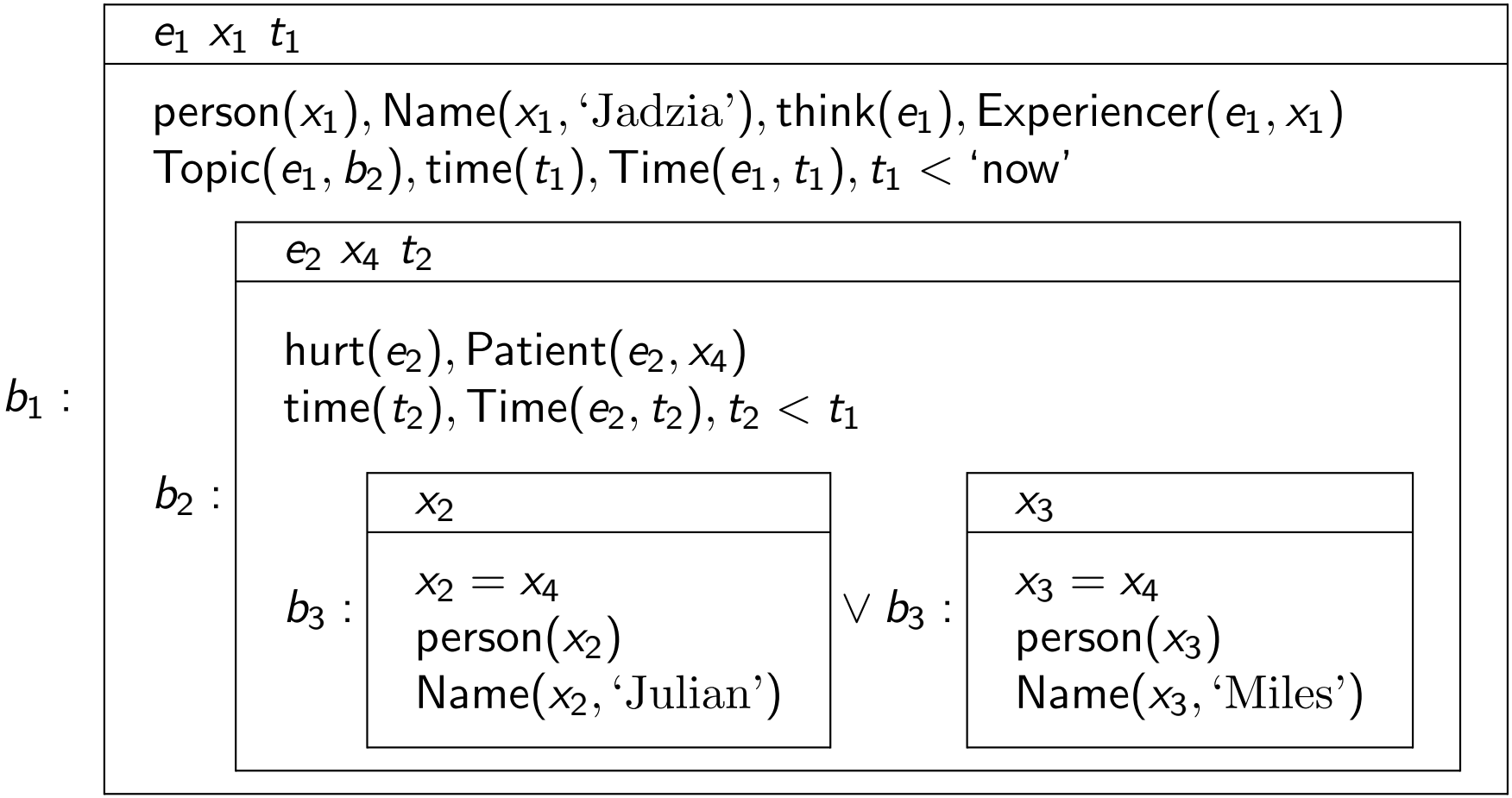


Sentence lengths compared

(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

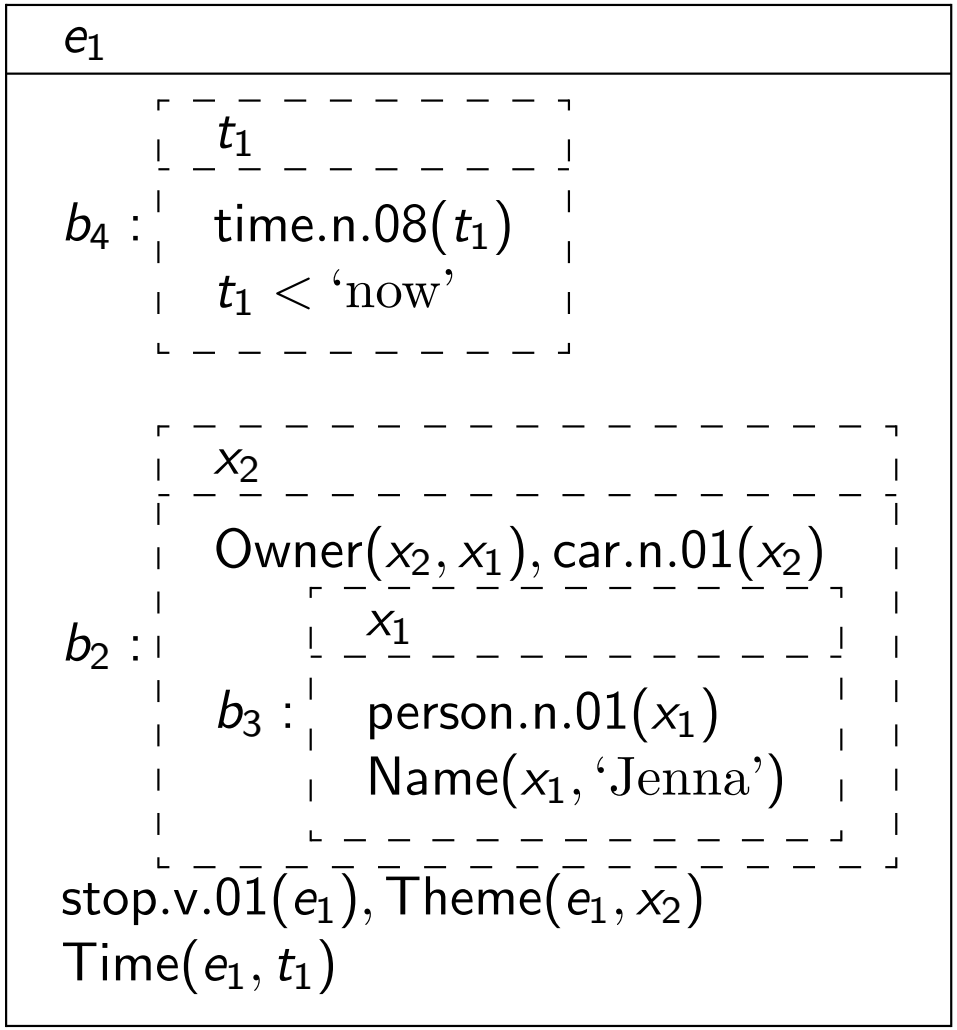
Table: Sentence length across (sub-)corpora

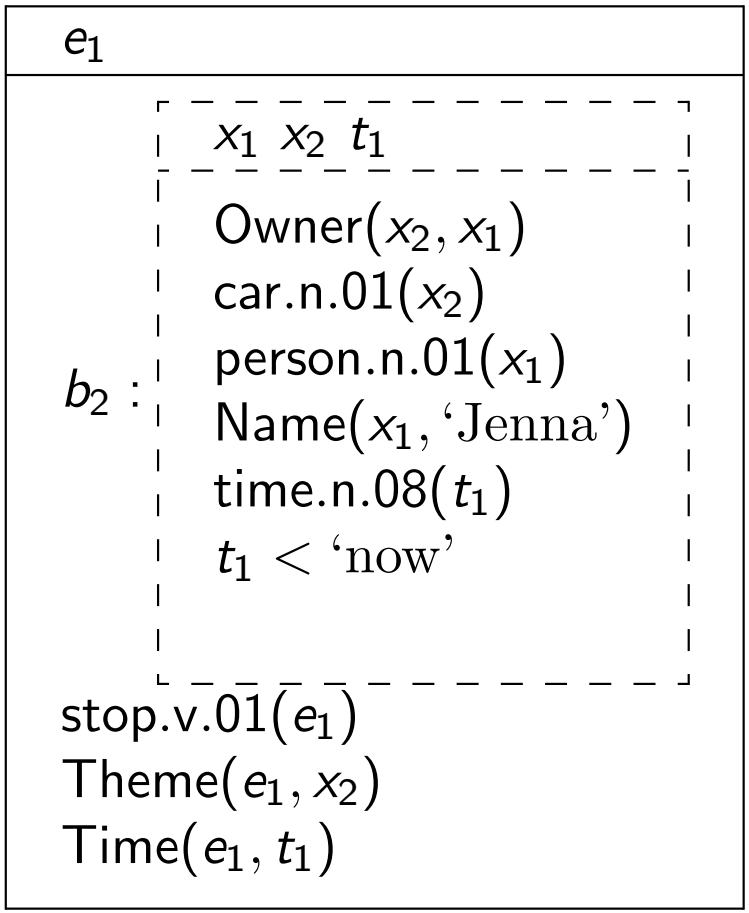
Discourse Representation Theory



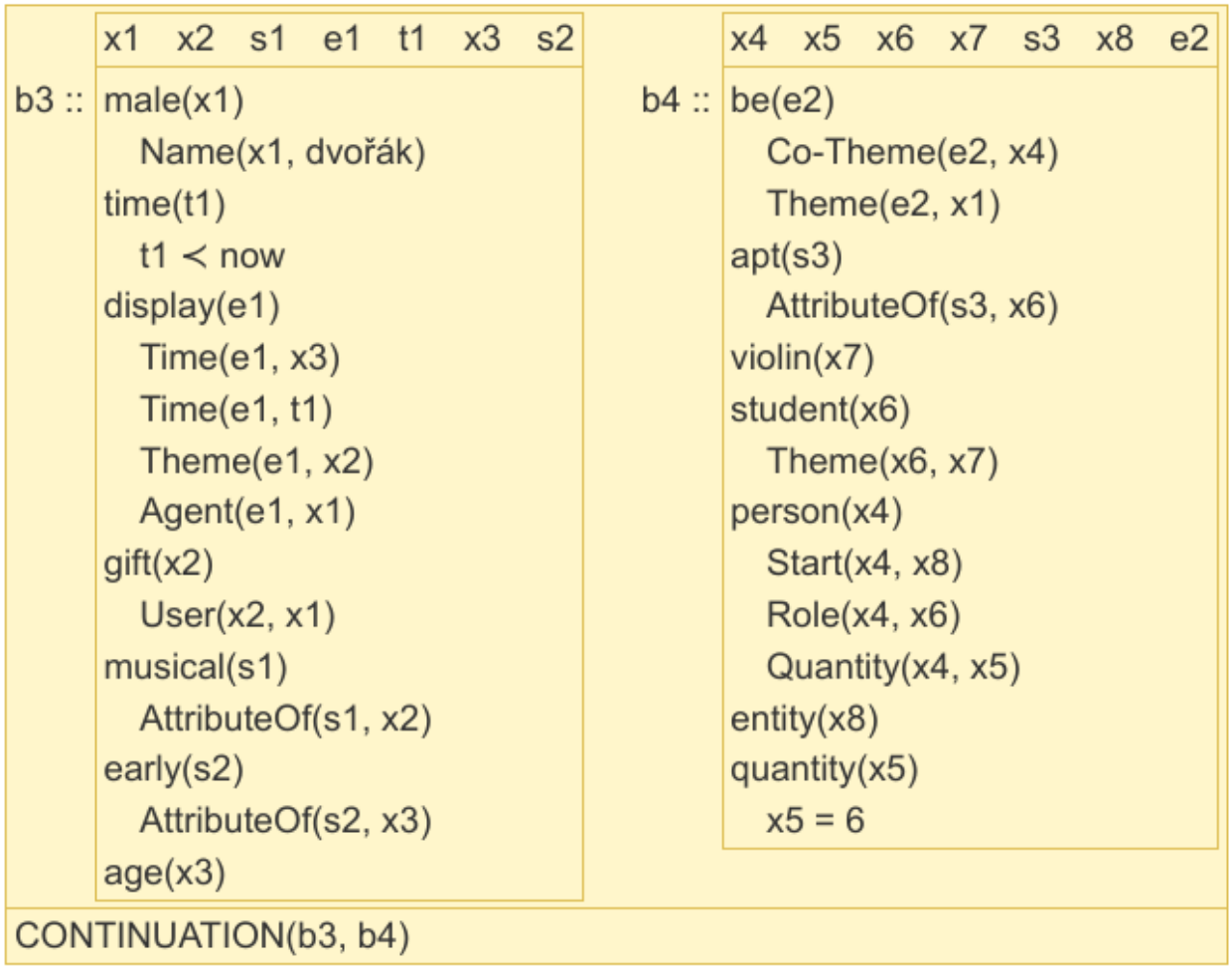
DRS for *Jadzia thought that Miles or Julian had been hurt*

Flattening presupposition structure





Flattening discourse structure



Dvořák displayed his musical gifts at an early age, being an apt violin student from age six.

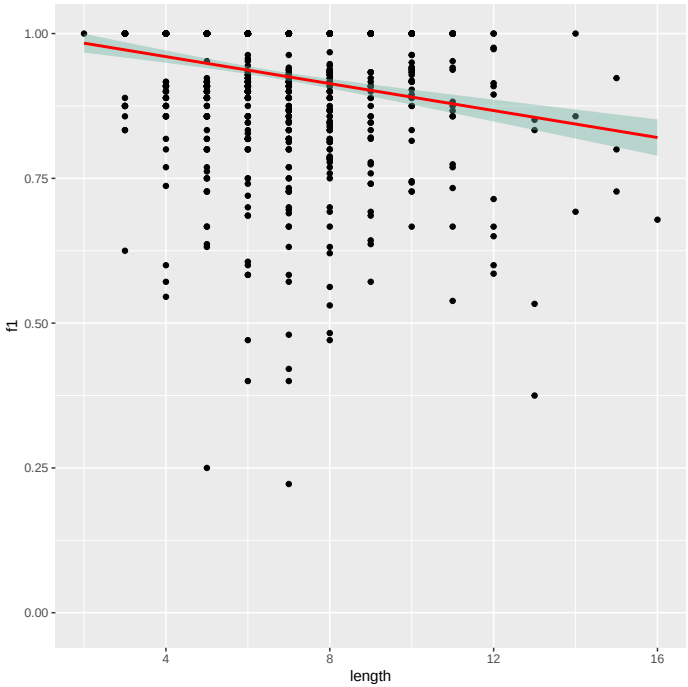
State of the art DRS parsing

	PMB 2.2.0		PMB 3.0.0		PMB 4.0.0			
	dev	**test**	**dev**	**test**	**dev**	**test**	**eval**	DRASTIC
van Noord et al. (2020a)	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

- Sequence-to-sequence networks, mostly LSTMs
- Haug and Yıldırım improve on the results by using transformers
- F1 scores in high 80s/low 90s
- Surprising because better than parsing to (less expressive) AMR
- However, sentence lengths in PMB may underestimate the difficulty of DRS parsing

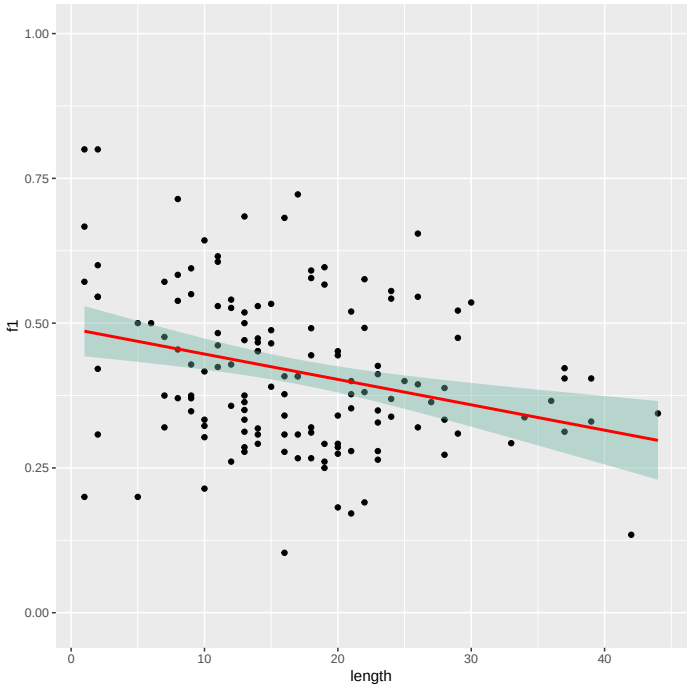
DRS parsing and sentence length

PMB



$r = -0.21$ ($p < 5 \times 10^{-10}$)

DRASTIC



$r = -0.29$ ($p < 4 \times 10^{-4}$)

- Pure length has a small effect
- Concomitant complexity is likely more important

Error analysis

- Caveat: we flattened the parser output rather than the training data (script included with the data)
- Allows reuse of the parser that we trained on the PMB data
- Error analysis is anecdotal due to data set size, but
  - negation is problematic and sometimes disappear
  - relative scope of negation and possibility is problematic
  - names that were not seen in training are problematic (15-20% of PMB sentences contain the name *Tom*)

▸ Examples

Summary and outlook

- PMB data has enabled neural DRS parsing but may yield an overly optimistic picture given the short sentence lengths
- DRASTIC provides sentences of more natural length, and tried to remove some of the ensuing complexity by flattening DRS structure
- Still much harder than PMB
- Currently very small at 157 sentences, but ca. 1000 more in the pipeline
- Will hopefully help improve DRS parsing by offering more varied data
- Also, opportunities to connect with the GUM annotation (discourse relations, coreference)

References I

Lucia Donatelli and Alexander Koller. Compositionality in computational linguistics. *Annual Review of Linguistics*, 9:463–481, 2023. DOI: 10.1146/annurev-linguistics-030521-044439.

Jiangming Liu, Shay B. Cohen, Mirella Lapata, and Johan Bos. Universal discourse representation structure parsing. *Computational Linguistics*, 47(2):445–476, 07 2021. ISSN 0891-2017. DOI: 10.1162/coli_a_00406. URL https://doi.org/10.1162/coli_a_00406.

Rik van Noord, Antonio Toral, and Johan Bos. 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, November 2020a. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.371. URL https://www.aclweb.org/anthology/2020.emnlp-main.371.

Rik van Noord, Antonio Toral, and Johan Bos. 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, nov 2020b. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.371.

References II

Yuekun Yao and Alexander Koller. Structural generalization is hard for sequence-to-sequence models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5048–5062, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.337.

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

Amir Zeldes. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017. DOI: 10.1007/s10579-016-9343-x.