# Is neural semantic parsing good at ellipsis resolution, or isn't it?

**Xiao Zhang**
Center for Language and Cognition
University of Groningen
xiao.zhang@rug.nl

**Johan Bos**
Center for Language and Cognition
University of Groningen
johan.bos@rug.nl

## Abstract

Neural semantic parsers have shown good overall performance for a variety of linguistic phenomena, reaching semantic matching scores of more than 90%. But how do such parsers perform on strongly context-sensitive phenomena, where large pieces of semantic information need to be duplicated to form a meaningful semantic representation? A case in point is English verb phrase ellipsis, a construct where entire verb phrases can be abbreviated by a single auxiliary verb. Are the otherwise known as powerful semantic parsers able to deal with ellipsis or aren't they? We constructed a corpus of 120 cases of ellipsis with their fully resolved meaning representation and used this as a challenge set for a large battery of neural semantic parsers. Although these parsers performed very well on the standard test set, they failed in the instances with ellipsis. Data augmentation helped improve the parsing results. The reason for the difficulty of parsing elided phrases is not that copying semantic material is hard, but that they usually occur in linguistically complicated contexts, causing most of the parsing errors.

## 1 Introduction

Semantic parsing is the task of providing a formal meaning representation for an input sentence of a natural language such as English, Dutch, or Italian. Semantic parsing is crucial for applications that require the precise translation of unstructured data (i.e., text and images) into structured data (e.g., databases and robot commands). Currently, the most promising approaches to semantic parsing are based on neural models (Bai et al., 2022; Wang et al., 2023; Zhang et al., 2024b, 2025) trained or fine-tuned on large semantically annotated corpora (Banarescu et al., 2013; Abzianidze et al., 2017), reaching high performance with F scores greater than 90%. Little is known about the ability of neural semantic parsers to cope with *ellipsis*, a linguistic construction in which elements are omitted and are supplied by the discourse context. In this paper, we will study how neural semantic parsers deal with Verb Phrase Ellipsis (VPE) in English. An example of a VPE is shown in (1) together with its fully expressed surface interpretation in (2).

(1) Ann likes grapes, and Bea does, too.
(2) Ann likes grapes, and Bea does **like grapes**, too.

As this very simple example already demonstrates, ellipsis interpretation is a challenging task, for the only way to recover the elided material is to consider the discourse context. The (computational) linguistics literature abounds with many more complicated examples of VPE, including sloppy-strict interpretation of pronouns appearing in the elided material, cascaded ellipsis, antecedent contained deletion, gapping, and embedded ellipsis (Dahl, 1973; Williams, 1977; Roberts, 1989; Dalrymple et al., 1991). Nevertheless, our aim is not to focus on these linguistically interesting examples carefully crafted by linguists, but rather to investigate how data-driven semantic parsers deal with instances of VPE found in corpora.

As far as we know, this is the first in-depth study of VPE interpretation in neural semantic parsing. Related, but taking a different perspective, is work by Hardt (2023), who found that large language models have difficulty processing ellipsis.

In Section 2 we give an overview of earlier computational approaches to VPE. In Section 3 we introduce the Parallel Meaning Bank (PMB) and a VPE challenge test set distilled from the PMB. In Section 4 we outline our approach to enhance the semantic parsing for VPE, while in Section 5 the parsing results are presented, showing that neural approaches face a difficult time in interpreting elliptical constructions, even with substantial fine-tuning, but not for the reasons we initially thought would cause the difficulty.



*Proceedings of the 16th International Conference on Computational Semantics*, pages 137–142

## 2 Background

VPE interpretation has drawn considerable attention in formal linguistics (Dahl, 1973; Sag, 1976; Klein, 1987; Dalrymple et al., 1991). These early approaches can be summarized as identifying an antecedent verb phrase in the context, providing a logical form while abstracting over the subject, and applying the result to the subject noun phrase of the elided verb phrase. Computational approaches were introduced later (Alshawi, 1992; Kehler, 1993; Bos, 1994; Crouch, 1995; Hardt, 1997), with the landmark paper by Dalrymple et al. (1991) introducing a set of benchmark VPE examples and a sophisticated algorithm based on higher-order unification to construct fully resolved meaning representations for elliptical phrases. These approaches, although computational of nature, still required external modules to identify the source verb phrase and the parallel elements between source and target phrase.

Data-driven approaches based on annotated corpora (Nielsen, 2005; Bos and Spenader, 2011; Bos, 2016) demonstrated the large gap between theoretical ideas and practical implementations (McShane and Babkin, 2016; Kenyon-Dean et al., 2020; Zhang et al., 2019), and were considered to be specific tasks rather than an integral part of wide-coverage semantic parsing. In this paper, we take a different computational perspective and depart with an overall well-performing general-purpose semantic parsing and investigate how well it succeeds on ellipsis data.

## 3 Data

**The Parallel Meaning Bank** The PMB Abzianidze et al., 2017 is a multilingual corpus enriched with semantic annotations, covering a wide range of linguistic phenomena. It contains a substantial set of parallel texts, each paired with a formal meaning representation known as a Discourse Representation Structure (DRS) based on Discourse Representation Theory (DRT, Kamp and Reyle, 1993). While DRSs are typically presented in a human-readable box format, a clause-based linear representation was introduced by van Noord et al. (2018) to enable their use in sequence-based models. More recently, Bos (2023) proposed Sequence Box Notation (SBN), a simplified, variable-free version of DRS aimed at further facilitating sequence processing. In this paper we use SBN as meaning representation format (see Figure 1).

Sentence with VPE:
*Life never ends but earthly life does.*

DRS:

```
life.n.11                        % Life never   [0-10]
          NEGATION <1            %
end.v.01  Theme -1 Time +1       % ends         [11-15]
time.n.08 EQU now                %
          CONTRAST <2            %
earthly.a.01 AttributeOf +1      % but earthly  [16-27]
life.n.11                        % life         [28-32]
end.v.01  Theme -1 Time +1       % does.        [33-38]
time.n.08 EQU now                %
```

DRG:

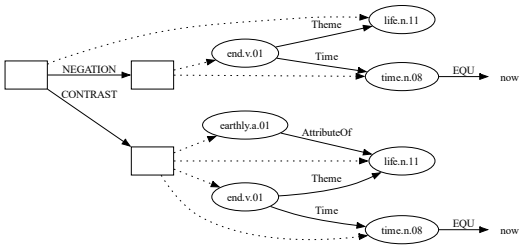


Figure 1: An example sentence with Verb Phrase Ellipsis and meaning representation in sequence notation and drawn as a directed acyclic graph.

**Annotated VPE Instances** As occurrences of VPE are relatively rare (Bos and Spenader, 2011), it is rather challenging to yield a reasonably sized corpus. A total of 120 cases were identified in the PMB and their corresponding meaning representations manually corrected. Slightly more than half of the cases (71) contained some kind of negation in the elided construction (e.g., ”and neither am I”, ”Greenland is not”, ”but she didn’t”). Half of the instances are accompanied by the auxiliary verb *to do*, a third by *to be*, and the remaining cases are formed by other auxiliary verbs, the infinitival particle *to* or instances of gapping. An annotated example taken from the corpus is shown in Figure 1, where the elliptical phrase ”life does” is semantically interpreted as ”life does end”.

## 4 Experimental Setup

**Training Sets** For training our neural semantic parsers, we consider two settings: (1) the **Standard Training Set**, the default training data provided by PMB version 5.1.0, with all texts that are included in the VPE test set removed; and (2) the **Augmented Training Set**, an augmented dataset to enhance the model’s ability to handle verb phrase ellipsis. We construct the augmented dataset apply-

ing the following data augmentation strategies:

- We employ GPT-4 to generate 600 sentence pairs, each consisting of a sentence containing VPE and its corresponding resolved version (i.e., the full sentence with the elided verb phrase explicitly restored in the surface text).
- We use the state-of-the-art DRS parser from Zhang et al. (2024a) to generate DRSs for the resolved sentences. These DRSs are then paired with the original VPE sentences as their target semantic representations.
- We incorporate the generated VPE data into the standard training set in varying quantities (from 100 up to 600) to examine how the scale of augmentation affects model performance and to identify the point at which performance improvements begin to converge.

**Test sets** We evaluate the trained parsers on two test sets: the **Standard Test Set**, which serves as a general, broad-coverage set for comparison, and **VPE120**, a targeted test set focusing on VPE, as described in Section 3.

**Evaluation** We evaluate model performance using two metrics: **Smatch**[1] and **Ill-Formed Rate (IFR)**. Smatch (Cai and Knight, 2013; Opitz, 2023) measures the similarity between the predicted and reference semantic graphs by converting each graph into a set of triples and computing the optimal variable mapping via a hill-climbing algorithm. Precision (P), recall (R), and F1 score are calculated as follows:

$$\mathrm{P} = \frac{m}{p}, \quad \mathrm{R} = \frac{m}{g}, \quad \mathrm{F1} = \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}, \qquad (1)$$

where $m$ denotes the number of matching triples, $p$ is the number of predicted triples, and $g$ is the number of gold-standard triples.

To assess the structural validity of generated graphs, we additionally report the **Ill-Formed Rate (IFR)**. A graph is considered ill-formed if it exhibits structural defects such as cyclic dependencies, isolated nodes, or dangling edges referencing non-existent elements. Graphs identified as ill-formed are assigned a Smatch score and F1 score of zero, thereby contributing to a quantitative measure of structural failure.

[1]We adopt the Smatch++ implementation (Opitz, 2023), which uses Integer Linear Programming (ILP) instead of the standard hill-climbing approach.

**Models** We evaluated three encoder–decoder models–mBART (Liu et al., 2020), mT5 (Xue et al., 2021), and ByT5 (Xue et al., 2022), as well as four decoder-only models: Qwen2.5-7B (Yang et al., 2024), Ministral-8B, LLaMA3.1-8B (Grattafiori et al., 2024), and Gemma2-9B (Team et al., 2024).

## 5 Results and Analysis

The performance of models on both test sets is presented in Table 1. Overall, sentences containing VPE instances pose significantly greater challenges for semantic parsing, as evidenced by substantially lower Smatch scores and elevated ill-formed rates (IFR). We analyze these results in detail below.

Table 1: Smatch and IFR performance on the Standard Test Set and VPE120 for models trained with the Standard Training Set, Aug300, and Aug600.

| Model | Train set | Standard Test | | VPE120 | |
|---|---|---|---|---|---|
| | | Smatch | IFR | Smatch | IFR |
| mBart-Large | Standard | 83.50 | 6.95 | 70.90 | 33.17 |
| | Aug300 | 85.40 | 7.00 | 77.90 | 27.33 |
| | Aug600 | 85.00 | 6.60 | 78.10 | 24.83 |
| mT5-Large | Standard | 82.61 | 11.20 | 70.38 | 29.83 |
| | Aug300 | 84.50 | 9.80 | 75.20 | 24.83 |
| | Aug600 | 84.00 | 9.20 | 75.50 | 24.00 |
| ByT5-Large | Standard | 91.40 | 8.73 | 66.22 | 27.33 |
| | Aug300 | 92.50 | 7.50 | 73.00 | 22.33 |
| | Aug600 | 92.90 | 7.00 | 72.50 | 22.33 |
| Qwen2.5-7B | Standard | 94.19 | 5.34 | 77.09 | 17.33 |
| | Aug300 | 94.35 | 5.17 | 85.31 | 9.83 |
| | Aug600 | 95.50 | 5.09 | 84.64 | 12.33 |
| Ministral-8B | Standard | 95.45 | 4.67 | 82.77 | 13.17 |
| | Aug300 | 95.50 | 4.25 | 89.00 | 6.50 |
| | Aug600 | 95.42 | 4.59 | 90.61 | 6.50 |
| LLaMA3.1-8B | Standard | 95.56 | 4.51 | 83.11 | 12.33 |
| | Aug300 | 95.32 | 5.18 | 88.89 | 12.33 |
| | Aug600 | 95.44 | 4.76 | 89.21 | 8.17 |
| Gemma2-9B | Standard | 96.31 | 4.59 | 78.09 | 17.33 |
| | Aug300 | 96.46 | 4.42 | 88.52 | 7.33 |
| | Aug600 | 96.59 | 4.09 | 89.20 | 8.17 |

**Performances on Standard Test** Decoder-only architectures consistently outperform encoder–decoder models on the Standard Test set. Gemma2-9B achieves the highest performance with a Smatch score of 96.59 following augmentation (compared to 96.31 on the standard training set). Other decoder-only models demonstrate similarly strong performance: LLaMA3.1-8B (95.44), Ministral-8B (95.50), and Qwen2.5-7B (95.50) all maintain scores above 95. In contrast, encoder–decoder architectures (mBART-Large, mT5-Large, and ByT5-

Large) achieve lower performance, with *standard* scores ranging from 82.61 to 91.40. This performance disparity likely stems from both architectural differences and parameter scale advantages, where larger decoder-only models may benefit from more stable fine-tuning dynamics and in-context learning capabilities.

**Performances on VPE120** All models exhibit substantially degraded performance on VPE120 relative to the Standard Test, confirming the inherent difficulty of parsing elliptical constructions semantically. When trained solely on the standard dataset, models achieve VPE120 Smatch scores between 66.22 and 83.11, accompanied by markedly increased IFR (e.g., 33.17% for mBART-Large and 29.83% for mT5-Large), indicating frequent generation of malformed outputs.

VPE-specific data augmentation yields substantial improvements across all architectures. Ministral-8B achieves the highest score of 90.61 with Aug600, closely followed by LLaMA3.1-8B (89.21) and Gemma2-9B (89.20). These top-performing models also demonstrate the most significant IFR reductions (e.g., Ministral-8B: 13.17% $\rightarrow$ 6.50%). Encoder–decoder models also benefit from augmentation: mBART-Large improves from 70.90 (standard) to 78.10 (Aug600), while mT5-Large advances from 70.38 to 75.50. Notably, ByT5-Large shows improvement from 66.22 to 73.00 with Aug300.

These findings demonstrate that VPE-specific data augmentation effectively narrows the performance gap between the Standard and VPE test sets, particularly for larger decoder-only models. The convergence of performance scores beyond Aug300 (see Figure 2) suggests diminishing gains from additional augmentation data, indicating that current models may be approaching their capacity limits for ellipsis resolution. This shows the need for more advanced architectures or specialized training strategies to further improve performance on complex elliptical phenomena.

**Qualitative Analysis** The previous section showed that sentences with ellipsis are a lot harder to parse for the neural semantic models. But why is this the case? Is this because they are a bad at copying semantic information, or is it something else? In order to answer we examined the output of the best performing model and manually inspected the results.

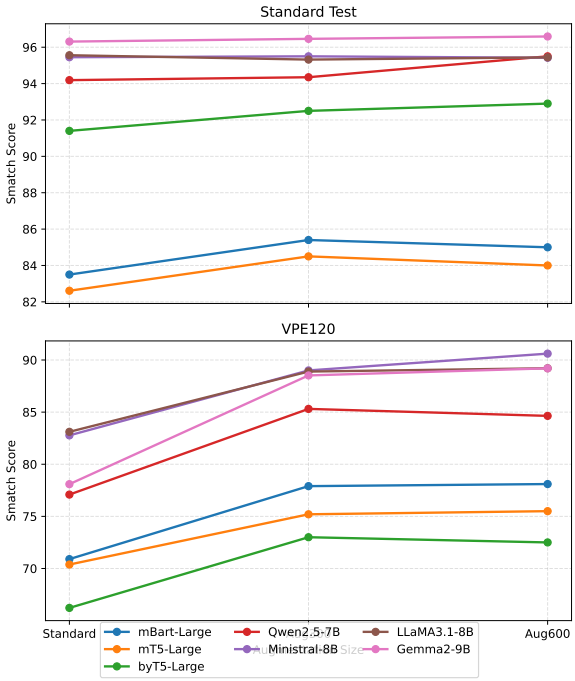


Figure 2: Model performance on VPE120 with increasing augmentation sizes (100 to 600).

Surprisingly, what we thought would be hard for the models, copying semantic material from the source to the target, was not hard at all. Only in three of the 120 cases did this not happen. Actually, what contributed to the low score was the wrong choice of discourse relation (22% of overall errors), the wrong attachment of a discourse relation (20%), incorrect scope order between tense and negation (16%), incorrect choice of word sense (16%), incorrect choice of thematic role (10%), incorrect choice of concept (10%), and incorrectly resolved anaphora (4%).

One reason why selecting the correct VP antecedent might have to do with the amount of ambiguity, or lack thereof. For instance, in the VPE example in Figure 1 there is only one potential verb phrase that could serve as antecedent for the elliptical phrase. Closer inspection of the dataset reveals that most (81%) of the texts with VPE are relatively short and provide only one verb phrase that could act as antecedent; only 23 examples provide two or more potential verb phrase antecedents, as in (3).

(3) Ann hoped to succeed, but she didn't.

Here there are two verb phrases in the context: *hope to succeed* and *succeed*. For most of these cases picking the most recent verb phrase usually yields the correct interpretation.

## 6 Conclusion

Although open-domain semantic parsing achieves good overall performance on the standard test sets, its shortcomings arise at the surface when looking at more complex linguistic phenomena. We demonstrated this by looking specifically at how neural parsing models deal with cases of English VP Ellipsis. Although we observed a drop in performance, the reason for the drop was not the context-sensitive nature of ellipsis, but rather the fact that elliptical phenomena are often surrounded by complex phenomena such as tense, negation, and discourse structure, causing parsing errors. So, is neural semantic parsing good at ellipsis resolution? Yes, it is!


### Acknowledgements

We would like to thank the three anonymous reviewers for their comments. Reviewer 1 pointed out that the developed datasets will be valuable for the community, and indeed the VPE dataset will be made public via the Parallel Meaning Bank data releases. Reviewer 2 wondered how many different VP targets there are for each case of ellipsis, as the complexity for ellipsis resolution "depends greatly on the number of available VP targets for each VPE". We added a discussion in this topic in Section 5. Reviewer 3 noted that the question in the paper's title is not actually answered. This was a correct observation. But now we did explicitly in Section 6. Special thanks go to Juri Opitz, who pointed out that using a hill-climber for evaluation (as we did in the submitted version of this paper) is not optimal. So we recalculated our scores using ILP instead.


## 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*, pages 242–247, Valencia, Spain.

Hiyan Alshawi, editor. 1992. *The Core Language Engine*. The MIT Press, Cambridge, Massachusetts.

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.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.

Johan Bos. 1994. Presupposition & vp ellipsis. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling 1994)*, pages 1184–1190, Kyoto, Japan.

Johan Bos. 2016. Verb phrase ellipsis and sloppy identity: a corpus-based investigation. In Martijn Wieling, Martin Kroon, Gertjan Van Noord, and Gosse Bouma, editors, *From Semantics to Dialectometry*, volume 32 of *Tributes*, pages 57–64. College Publications.

Johan Bos. 2023. The sequence notation: Catching complex meanings in simple graphs. In *Proceedings of the 15th International Conference on Computational Semantics (IWCS 2023)*, pages 1–14, Nancy, France.

Johan Bos and Jennifer Spenader. 2011. An annotated corpus for the analysis of vp ellipsis. *Language Resources and Evaluation*, 45(4):463–494.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Richard Crouch. 1995. Ellipsis and Quantification: A Substitutional Approach. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–236, Dublin, Ireland.

Östen Dahl. 1973. On so-called'sloppy identity'. *Synthese*, pages 81–112.

Mary Dalrymple, Stuart M. Shieber, and Fernando C.N. Pereira. 1991. Ellipsis and Higher-Order Unification. *Linguistics and Philosophy*, 14:399–452.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daniel Hardt. 1997. An Empirical Approach to VP Ellipsis. *Computational Linguistics*, 23(4):525–541.

Daniel Hardt. 2023. Ellipsis-dependent reasoning: A new challenge for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages

39–47, United States. Association for Computational Linguistics. The 61st Annual Meeting of the Association for Computational Linguistics ; Conference date: 09-07-2023 Through 14-07-2023.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht.

Andrew Kehler. 1993. A discourse copying algorithm for ellipsis and anaphora resolution. In *Proceedings of EACL*, pages 203–212.

Kian Kenyon-Dean, Edward Newell, and Jackie Chi Kit Cheung. 2020. Deconstructing word embedding algorithms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8479–8484, Online. Association for Computational Linguistics.

Ewan Klein. 1987. VP Ellipsis in DR Theory. In Jeroen Groenendijk et al., editors, *Studies in Discourse Representation Theory and the Theory of Generalised Quantifiers*, volume 8, pages 161–187. FLORIS, Dordrecht.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Marjorie McShane and Petr Babkin. 2016. Detection and resolution of verb phrase ellipsis. *Linguistic Issues in Language Technology*, 13.

Leif Arda Nielsen. 2005. *A corpus-based study of Verb Phrase Ellipsis Identification and Resolution*. Ph.D. thesis, King's College London.

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

Juri Opitz. 2023. SMATCH++: Standardized and extended evaluation of semantic graphs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1595–1607, Dubrovnik, Croatia.

Craige Roberts. 1989. Modal subordination and pronominal anaphora in discourse. *Linguistics and Philosophy*, 12(6):683–721.

Ivan Sag. 1976. *Deletion and Logical Form*. Ph.D. thesis, MIT.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Chunliu Wang, Huiyuan Lai, Malvina Nissim, and Johan Bos. 2023. Pre-trained language-meaning models for multilingual parsing and generation. In *Findings of the Association for Computational Linguistics*, page 5586–5600. Association for Computational Linguistics (ACL).

Edwin Williams. 1977. Discourse and logical form. *Linguistic Inquiry*, 8(1):101–139.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv e-prints*, pages arXiv–2412.

Wei-Nan Zhang, Yue Zhang, Yuanxing Liu, Donglin Di, and Ting Liu. 2019. A neural network approach to verb phrase ellipsis resolution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7468–7475.

Xiao Zhang, Gosse Bouma, and Johan Bos. 2025. Neural semantic parsing with extremely rich symbolic meaning representations. *Computational Linguistics*, 51(1):235–274.

Xiao Zhang, Qianru Meng, and Johan Bos. 2024a. Retrieval-augmented semantic parsing: Using large language models to improve generalization. *arXiv preprint arXiv:2412.10207*.

Xiao Zhang, Chunliu Wang, Rik van Noord, and Johan Bos. 2024b. Gaining more insight into neural semantic parsing with challenging benchmarks. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 162–175, Torino, Italia. ELRA and ICCL.