The Hoosier Ellipsis Corpus (HELC): Documenting Linguistic Dark Matter

Damir Cavar Ludovic Mompelat Muhammad S. Abdo NLP-Lab

Indiana University at Bloomington

Ellipsis Constructions

- Omission of words in sentences that are usually obligatory in a given syntactic context
- Example: noun phrase (NP) or Forward Conjunct Reduction (FCR), as in example (1)

(1) a. My sister lives in Utrecht and ___ works in Amsterdam.
b. My sister lives in Utrecht and ***she***/***my sister*** works in Amsterdam.

- *gapping* in (2a) where the verb complex *is reading* is elided
- VP-ellipsis in (2b) where the entire predicate or Verb Phrase (VP) is elided

(2) a. Peter is reading a book and Mary ___ a newspaper.
b. She will hi-five Daniel, but I won't ___

- Context-dependent forms of ellipsis in responses to questions as in (3), the words *each candidate will talk* are elided:

(3) a. Will each candidate talk about taxes?
b. No, ___ about foreign policy.

- Lexical mismatches of elided word forms as in (4a)
- In highly inflecting languages like Hindi or Croatian (4b) elided words do not have to be homophonous

(4) a. John **reads** a book, but Paul and Mary (**read**) a newspaper.
b. Ivan **je čitao** knjigu a Marija i Petar (**su čitali**) novine. (Croatian)
I. be read book but M. and P. be read newspaper

- Elided elements scattered over multiple positions in example (5) where the words *will*, *greet*, and *first* are elided

(5) Will Jimmy greet Jill first, or ___ Jill ___ Jimmy ___ ?

- ellipsis constructions are very common and often accompanied by specific semantic effects (Testa et al., 2023; Hardt, 2023)
  - various quantifier scope effects
  - semantic issues involve so-called *zeugma* (Sennet, 2016) as in example (6)

(6) a. John stole a book and Peter stole kisses from Mary.
b. John stole a book and Peter ___ kisses from Mary.

HELC Data

- HELC is constructed as a pair of sentences with optional context.
- The sentence pairs are separated by 4 dashes.
- The first line contains a sentence with ellipses.
- The second line contains the same sentence with the elided words spelled out.

Sample entry in the corpus:

```
Wird sie kommen oder ___ er gehen?
----
Wird sie kommen oder wird er gehen?
# TR eng: Will she come or will he go?
# added by: John Smith
# source: Wolfgang Klein (1981)
# Some Rules of Regular ...
```

- The canonical position of the elided word(s) is indicated by 3 underscores.
- Complex ellipsis constructions may have several elided positions.

Coverage

Languages: Arabic, Mandarin Chinese, Croatian, English, German, Gujarati, Hindi, Japanese, Kumaoni, Korean, Navajo, Norwegian, Polish, Russian, Spanish, Swedish, Telugu, Ukrainian

In preparation: Bengali, Bosnian, Bulgarian, Hebrew, Kanada, Serbian, Slovak, Slovenian, Tamil

Availability:

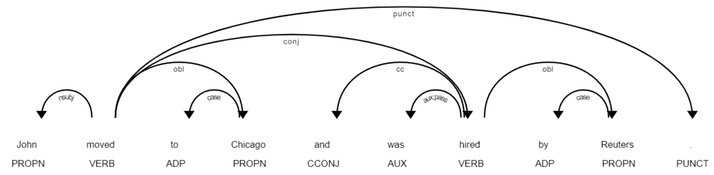
- Data website: `https://nlp-lab.org/ellipsis/`
- GitHub repositories: `https://github.com/dcavar/hoosierellipsiscorpus`

IU NLP-Lab Team and Contributors:

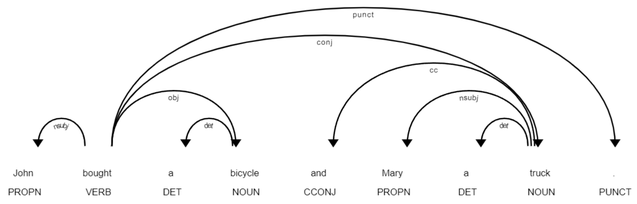
Dr. Damir Cavar, Muhammad S. Abdo, Andrew Davis, Dhananjay Srivastava, Billy Dickson, Vance Holthenrichs, Soyoung Kim, Dr. Zoran Tiganj, Khai Anthony Willard, Calvin Josenhans, Yuchen Yang, John MacIntosh Phillips, Luis Abrego, Ian Devine, Anshul Kumar Mangalapalli, Tanmayi Balla, Koushik Reddy Parukola, Dr. Ludovic Mompelat

NLP Challenges

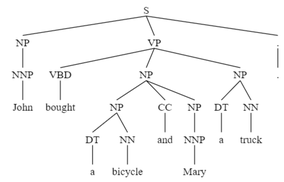
- Common State-of-the-Art NLP-pipelines fail, as in the following Stanza Dependency Trees: The syntactic subject in the second conjunct is not identified



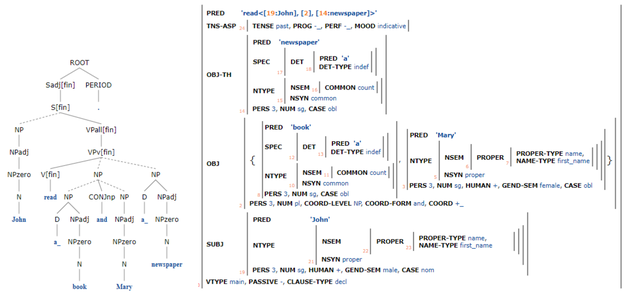
- Coordination and ellipsis with Stanza: Useless Dependency parse tree



- Constituent parsing with Stanza: no improvement – the common tendency is to analyze every coordination as local NP-coordination



- Lexical-functional Grammar using Xerox Linguistic Environment (XLE) and the English grammar:



- All NLP-pipelines fail with most constructions containing:
  - ellipsis
  - syntactic discontinuities
  - long-distance dependencies

  independent of underlying syntactic theory or ML model!

NLP Pipelines Tested

- Benepar Kitaev and Klein (2018); Kitaev et al. (2019)
- spaCy 3.x Honnibal and Johnson (2015)
- Stanford Stanza Qi et al. (2020)
- Stanford CoreNLP Manning et al. (2014)
- Xerox Linguistic Environment (XLE) Crouch et al. (2011)
- Quantum NLP pipelines, e.g., Lambeq Kartsaklis et al. (2021)
- LLMs: GPT-4

Testing Ellipsis in Different Models

- Baseline: Linear Regression
- Neural classifier using BERT
- SOTA LLMs: GPT-4, Claude 3, etc.
- LLMs tested using linguistic bias prompt and 0-shot or few-shot with 5 or more examples

Test 1: Binary Classification

- Does the sentence contain ellipses? Yes/No
- Test data: mix of distractor and target sentences (language dependent: e.g., English 575 target and 658 distractor sentences; Arabic 375 target and 500 distractor sentences)
- ten-fold randomized rotation for experiments

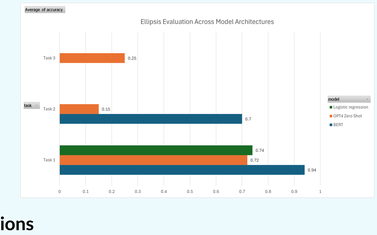
Test 2: Ellipsis Location

- Identify the location of the ellipses.
- Neural classifier using BERT
- SOTA LLMs: GPT-4, Claude 3, etc.

Test 3: Missing Words

- Identify the elided words.
- Only SOTA LLMs: GPT-4, Claude 3, etc.

Task 1:



model	accuracy
LR	0.74
BERT	**0.94**
GPT-4 zero-shot	0.72

Conclusions

- Logistic Regression outperforms GPT-4 zero-shot on Task 1
- BERT model outperforms GPT-4 zero-shot on Task 2
- GPT-4 on Task 3 only 25% accuracy with zero-shot

References

Richard Crouch, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell, III, and Paula Newman. *XLE Documentation*. Xerox Palo Alto Research Center, Palo Alto, CA, 2011. URL `https://ling.sprachwiss.uni-konstanz.de/pages/xle/doc/xle_toc.html`.

Daniel Hardt. 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: Short Papers)*, pages 39–47, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.4. URL `https://aclanthology.org/2023.acl-short.4`.

Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1162. URL `https://aclanthology.org/D15-1162`.

Dimitri Kartsaklis, Ian Fan, Richie Yeung, Anna Pearson, Robin Lorenz, Alexis Toumi, Giovanni de Felice, Konstantinos Meichanetzidis, Stephen Clark, and Bob Coecke. lambeq: An efficient high-level python library for quantum nlp, 2021.

Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1249. URL `https://aclanthology.org/P18-1249`.

Nikita Kitaev, Steven Cao, and Dan Klein. Multilingual constituency parsing with self-attention and pre-training. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1340. URL `https://aclanthology.org/P19-1340`.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. URL `https://nlp.stanford.edu/pubs/qi2020stanza.pdf`.

Adam Sennet. Polysemy. In S. Goldberg, editor, *Oxford Handbooks Online: Philosophy*. Oxford University Press, 2016.

Davide Testa, Emmanuele Chersoni, and Alessandro Lenci. We understand elliptical sentences, and language models should too: A new dataset for studying ellipsis and its interaction with thematic fit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, pages 3340–3353. Association for Computational Linguistics, July 2023.

The NLP-Lab (`https://nlp-lab.org/`)