The Hoosier Ellipsis Corpus: Building a Corpus of Ellipsis for Arabic Natural Language Processing

Muhammad S. Abdo   Damir Cavar   NLP-Lab

Indiana University at Bloomington

What is Ellipsis?

- Winkler (2011) defines ellipsis as the omission of words in a sentence that are understood from the context. It involves leaving out elements that are not necessary for the listener or reader to comprehend the meaning, thus making communication more concise.
- **Verb Phrase (VP) Ellipsis**: Cannon (2023) argues that VP always targets an entire VP, which usually occurs where two clauses are coordinated, and an equivalent VP exists in the other clause.

  أنا سأسافر اليوم وهو ___ غدًا

'I'm traveling today, and he ~~is traveling~~ tomorrow.'

- **Noun Phrase (NP) Ellipsis**: According to Saab (2018), NP Ellipsis occurs within the noun phrase or when the whole noun phrase is elided.

  شاهدت الجزء الثاني من الفيلم قبل أن أشاهد ___ الأول

'I saw the second part of the movie before I saw the first ~~part~~.'

- **Gapping**: According to Park (2019) Carnie (2021) and Mansour (2007), for gapping to occur in Arabic, three conditions must be met: 1) There must be surrounding lexical material on both sides of the elided verb in the second conjunct. 2) Constituents, after the verb in the second conjunct and in the first conjunct, must be syntactically and semantically parallel, and 3) At least two remnants must be left behind.

  أنا دخلت كلية الآداب وهو ___ هندسة

'I joined the school of Arts and he ~~joined~~ engineering'

- **Stripping**: In Stripping, an entire clause is omitted except for one constituent, i.e., the remnant, as mentioned in Algryani (2019).

  الطلاب يريدون عطلات أطول ونحن كذلك ___

'Students want longer vacations, and we ~~want longer vacations~~, too.'

- **Sluicing**: a form of anaphora that omits part of a sentence, linking back to a prior sentence for context. The remaining wh-phrase functions as a complete wh-question, as mentioned in Carnie (2021) and Merchant (2006).

  نعم السودان يمكنه أن يحقق طفرات ولكن كيف؟ ___

'Yes! Sudan can make some breakthroughs, but how ~~can Sudan make these breakthroughs~~ ?'

- **Fragment Answer**: According to Algryani (2017), these are brief responses to questions comprising non-sentential expressions. While they lack complete sentence structure, they convey the same meaning as full sentences.

  ما هي آخر أعمال نجيب محفوظ؟ الأحلام الأخيرة ___

'What was Naguib Mahfouz's last work? *Last Dreams* ~~was Naguib Mahfouz's last work~~'

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, and the canonical position of the elided word or words is indicated by 3 underscores.

Sample entry in the corpus:

تركز عليها أغلب الدول والشركات كذلك ___
____
تركز عليها أغلب الدول والشركات كذلك تركز عليها
# TR eng: Most countries focus on them, and so do companies.
# added by: Muhammad S. Abdo
# source: ArabicTenTen (2024)

Coverage

Languages: Currently, the corpus comprises ellipsis constructions from Modern Standard Arabic only, but we plan to expand the corpus to include data from various Arabic dialects.

Availability:

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

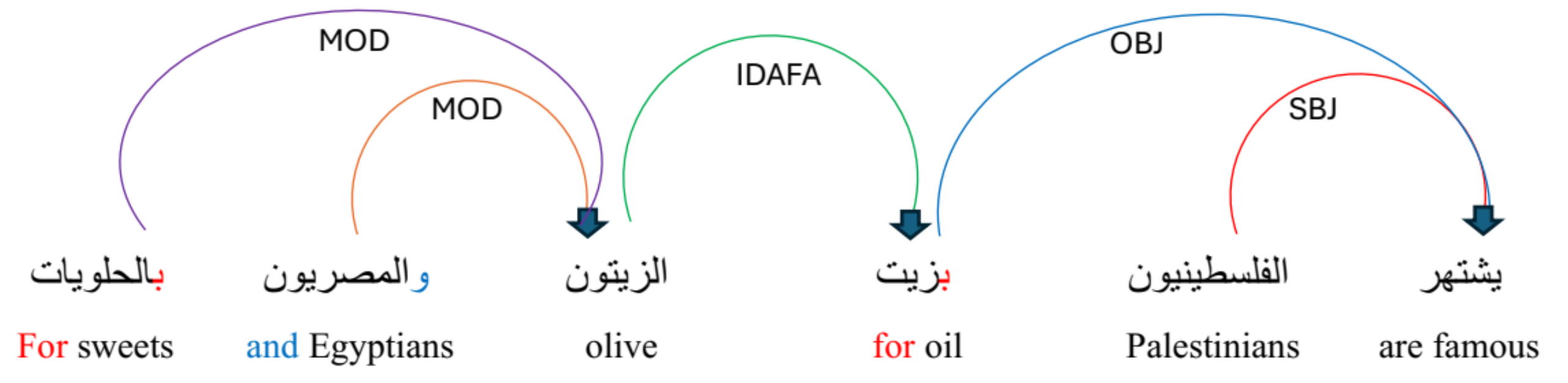
IU NLP-Lab Team and Contributors:

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

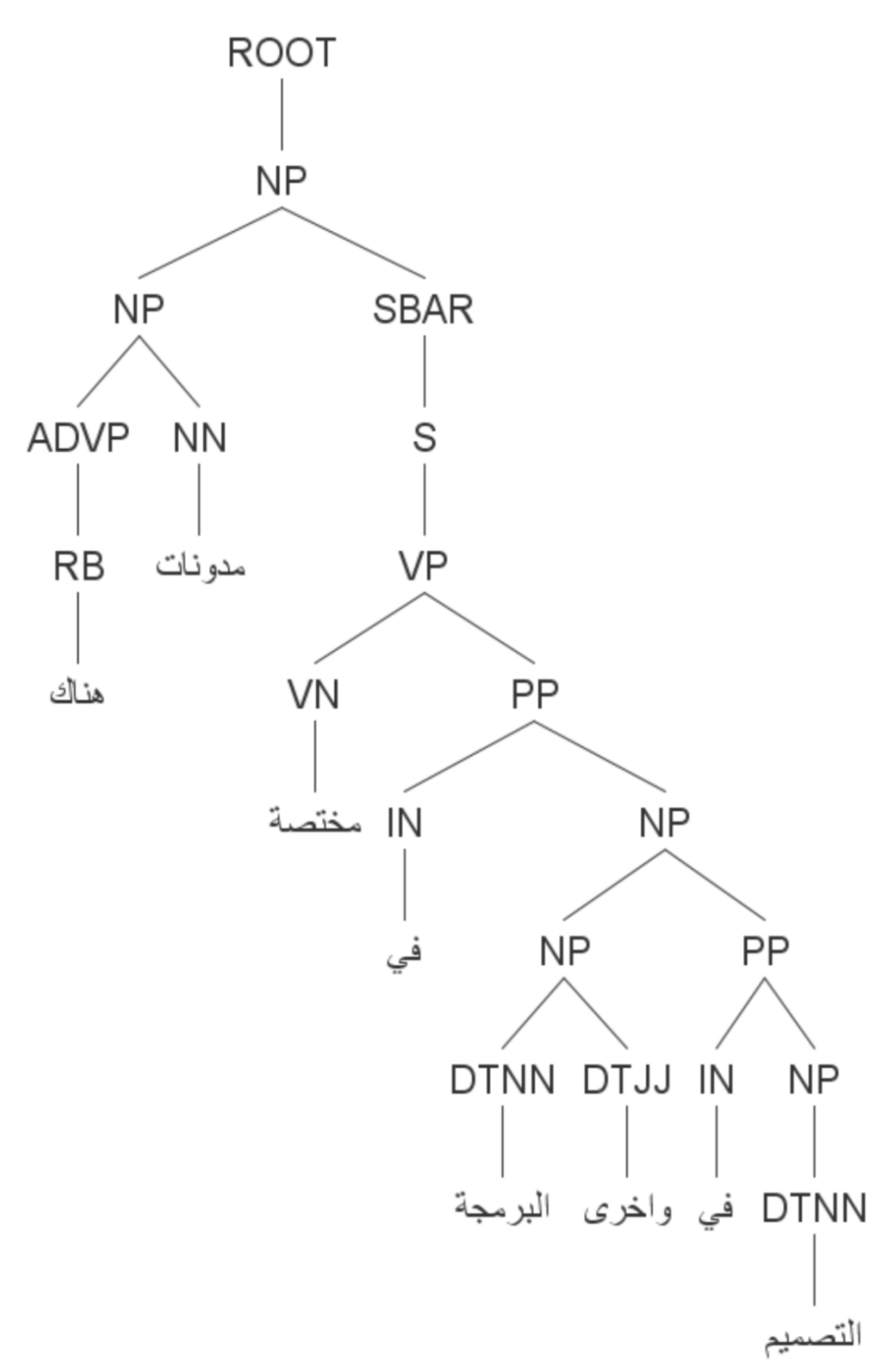
- Common State-of-the-Art NLP-pipelines fail, as in the following Dependency and Constituency Trees.

FARASA Arabic Dependency Parser (Zhang et al., 2015)



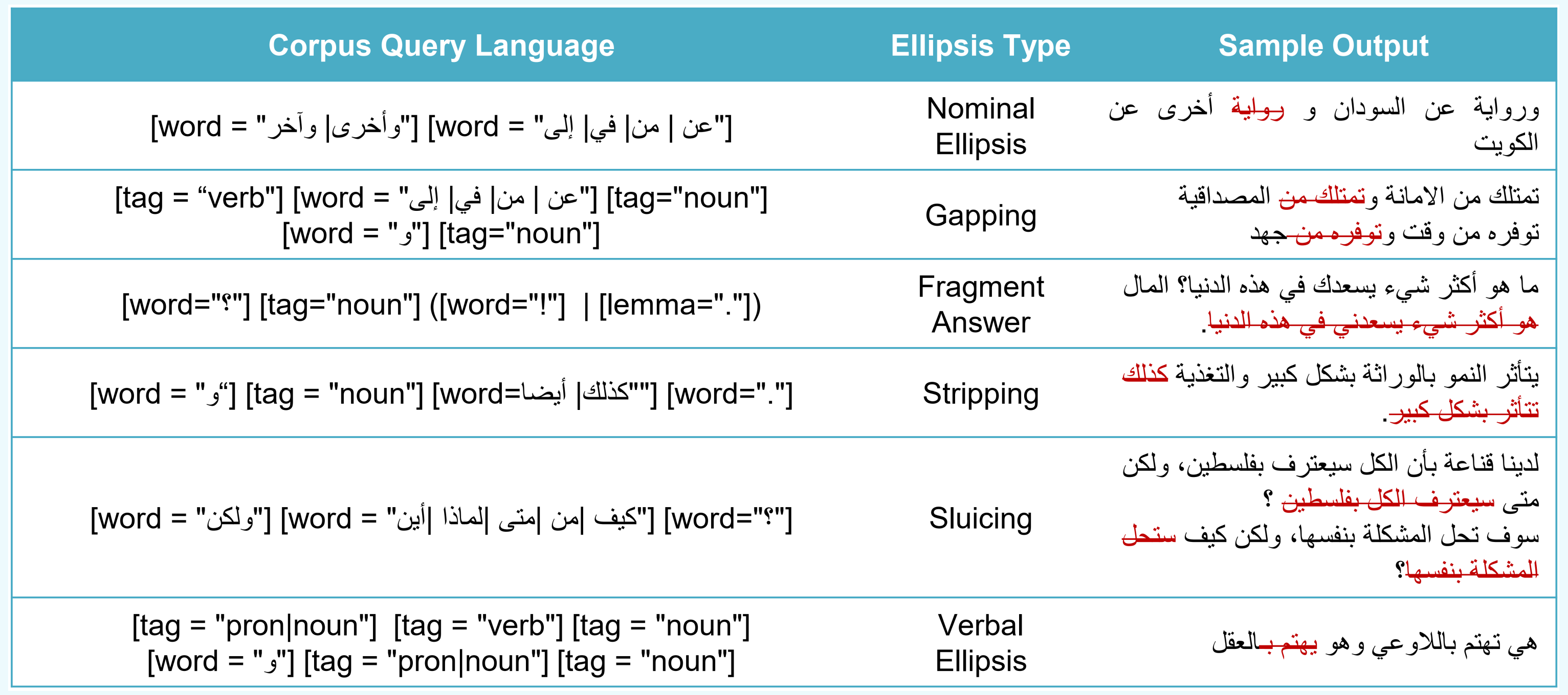
Palestinians are famous for their olive oil and Egyptians for their sweets.

Stanford Parser (Green and Manning, 2010)



Data Collection Using Corpus Query Language

- **ArabicTenTen**: 7b words covering topics from Arts, travel, and sports to religion, finance, and history. **Corpus Query Language (CQL):** is a query language used in Sketch Engine to search for complex grammatical or lexical patterns. The top 1000 query results, if any, are currently being labeled to indicate whether they contained ellipsis constructions or not, in order to evaluate the effectiveness of the query in extracting sentences with ellipsis. The provided sample includes a selection of queries. The sample queries reviewed demonstrated that, while initial annotations indicated an 86% success rate in eliciting ellipsis constructions from the responses across most ellipsis types, the queries designed specifically for verbal ellipsis only succeeded in generating ellipsis in about 5 % of the cases, which is not surprising given that Al-Khawalda (2002) and Hawkins (2012) confirmed that VP ellipsis in Arabic is either not possible or very rare.



Corpus Query Language	Ellipsis Type	Sample Output
[word = "وأخرى| وآخر"] [word = "عن | من| في| إلى"]	Nominal Ellipsis	ورواية عن السودان و ~~رواية~~ أخرى عن الكويت
[tag = "verb"] [word = "عن | من| في| إلى"] [tag="noun"] [word = "و"] [tag="noun"]	Gapping	تمتلك من الامانة و~~تمتلك من~~ المصداقية توفره من وقت و~~توفره من~~ جهد
[word="؟"] [tag="noun"] ([word="!"] | [lemma="."])	Fragment Answer	ما هو أكثر شيء يسعدك في هذه الدنيا؟ المال ~~هو أكثر شيء يسعدني في هذه الدنيا~~.
[word = "و“] [tag = "noun"] [word="كذلك| أيضا""] [word="."]	Stripping	يتأثر النمو بالوراثة بشكل كبير والتغذية كذلك ~~تتأثر بشكل كبير~~.
[word = "ولكن"] [word = "كيف| من| متى| لماذا| أين"] [word="؟"]	Sluicing	لدينا قناعة بأن الكل سيعترف بفلسطين، ولكن متى ~~سيعترف الكل بفلسطين~~ ؟ سوف تحل المشكلة بنفسها، ولكن كيف ~~ستحل المشكلة بنفسها~~؟
[tag = "pron|noun"] [tag = "verb"] [tag = "noun"] [word = "و"] [tag = "pron|noun"] [tag = "noun"]	Verbal Ellipsis	هي تهتم باللاوعي وهو ~~يهتم ب~~العقل

Figure 1. Extracting Ellipsis Constructions using CQL queries

Testing Ellipsis in Different Models

- Baseline: Logistic Regression
- SOTA LLMs: GPT-4.
- 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 with 375 target and 500 distractor sentences.
- ten-fold randomized rotation for experiments

Test 2: Ellipsis Location

- Identify the location of the ellipses.
- Currently experimenting with BERT classifiers, GPT-4, and Claude.

Test 3: Missing Words

- Identify the elided words.
- Only SOTA LLMs: GPT-4.

Results:

- **Task 1:** Baseline Logistic Regression 0.83.
- Zero-shot experiments with GPT-4 resulted in an accuracy of 0.87.
- **Task 3: GPT-4** Accuracy 0.80.

Conclusions

- Zero-shot GPT-4 experiments were found to outperform Logistic Regression on Task 1, which is not the case for other languages, e.g., English, where GPT-4 underperformed in comparison with Logistic Regression.
- Also, for the third task, in comparison with English, GPT-4 achieved a relatively high accuracy score, with 0.80 for Arabic and only 0.25 for English.

References

Mohammad Al-Khawalda. Ellipsis in arabic and english. *International Journal of Arabic-English Studies*, 3(1):183–199, 2002.

Ali Algryani. Ellipsis in arabic fragment answers. *Order and structure in syntax II*, page 319, 2017.

Ali Algryani. The syntax of sluicing in omani arabic. *International Journal of English Linguistics*, 9(6):337–346, 2019.

Cutler Cannon. A theoretical account of whale song syntax: A new perspective for understanding human language structure. *Proceedings of the Linguistic Society of America*, 8(1):5571–5571, 2023.

Andrew Carnie. *Syntax: A generative introduction*. John Wiley & Sons, 2021.

Roger Hawkins. Knowledge of english verb phrase ellipsis by speakers of arabic and chinese. *Linguistic Approaches to Bilingualism*, 2(4):404–438, 2012.

Mohamed Abdelmageed Mansour. Semantic constraints on licensing vp-ellipsis and vp-gapping in arabic. *Bulletin of the Faculty of Arts, Assiut University*, 2007.

Jason Merchant. Sluicing. *The Blackwell companion to syntax*, pages 271–291, 2006.

Sang-Hee Park. *Gapping: A constraint-based syntax-semantics interface*. PhD thesis, State University of New York at Buffalo, 2019.

Andrés Saab. 526 Nominal Ellipsis. In *The Oxford Handbook of Ellipsis*. Oxford University Press, 12 2018. ISBN 9780198712398. doi: 10.1093/oxfordhb/9780198712398.013.26. URL `https://doi.org/10.1093/oxfordhb/9780198712398.013.26`.

Susanne Winkler. *Ellipsis and focus in generative grammar*, volume 81. Walter de Gruyter, 2011.

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

https://nlp-lab.org   Midwest Speech and Language Days 2024, Ann Arbor — University of Michigan   dcavar@iu.edu