On Ellipsis in Slavic: The Ellipsis Corpus and Natural Language Processing Results

Damir Cavar & Van Holthenrichs (+ the NLP-Lab Team)

Indiana University at Bloomington

Formal Approaches to Slavic Linguistics 33, Halifax, Canada

May 2024

Team

- NLP-Lab at Indiana University Bloomington (https://nlp-lab.org/)

  Coauthors: Billy Dickson, Zoran Tiganj

  Lab team: Muhammad S. Abdo, Andrew Davis, Dhananjay Srivastava, Soyoung Kim, Khai Anthony Willard, Calvin Josenhans, Yuchen Yang, John MacIntosh Phillips, Ludovic Mompelat, Luis Abrego, Ian Devine, Anshul Kumar Mangalapalli, Tanmayi Balla, Koushik Reddy Parukola

  Slides and publications, and links to more material available at:

  https://nlp-lab.org/

Agenda

- Ellipsis Constructions and Syntax
- The Hoosier Ellipsis Corpus
- Slavic Sub-Corpora
- Machine Learning Experiments

SECTION 1

Introduction and Motivation

Ellipsis Constructions

- Common phenomena like gapping, sluicing, forward or backward conjunction reduction
  - Lexical elements are elided under certain conditions (e.g., syntactic, discourse)
  - Native speakers have no cognitive issues processing and understanding ellipsis constructions
  - Ellipsis constructions are very frequent in common genres

- Examples...

Ellipsis Constructions

Sluicing:

- *Moja sestra živi u Utrechtu ali ne znam gdje ___*
  → *Moja sestra živi u Utrechtu ali ne znam gdje (moja sestra u Utrechtu živi)*

Gapping

- Иван и Андрей смотрели новости, а Ольга ___ фильм.
  → Иван и Андрей смотрели новости, а Ольга смотрела фильм.

- *Czy Marek zobaczył Annę pierwszy, czy Anna ___ Marka ___ ?*
  → *Czy Marek zobaczył Annę pierwszy, czy Anna zobaczyła Marka pierwsza ?*

Ellipsis Constructions

- **Discourse Licensed Ellipsis:**
- A: *Tko želi sresti koga?*
- B: *Suzana ___ Petra.*

  → *Suzana želi sresti Petra.*

- **Semantic Issues:**

- *Marek pojechał do Warszawy i ___ został aresztowany w Poznaniu.*

- *Peter stole a book and John ___ kisses from Mary.* (zeugma, Sennet 2016)

Ellipsis Constructions

- Publicly available datasets:
  - Sluicing corpus for English
  - VP-ellipsis corpus for English
  - ELLies corpus for English
- Small datasets
- Limited to English and a few common languages
- Limited to specific ellipsis phenomena (gapping, sluicing, or VP-ellipsis)

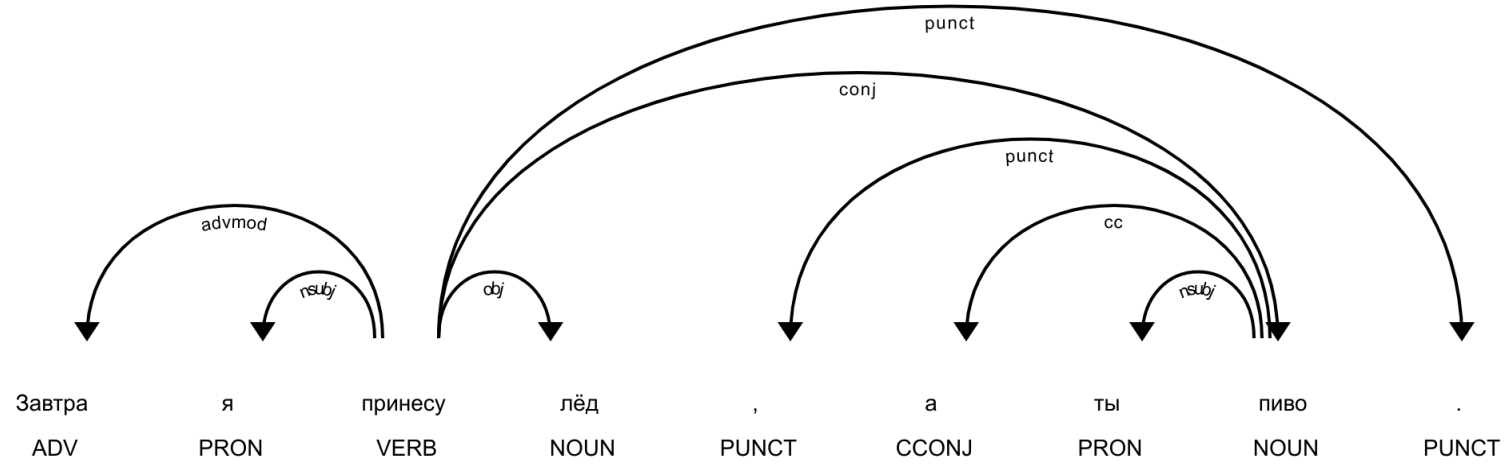
Ellipsis Constructions

- Lack of a cross-linguistic typological overview of ellipsis types
- Explanatory theoretical analysis of ellipsis constructions
- Frameworks like Dependency Grammar or Lexical-functional Grammar do not provide descriptive or explanatory means
    - even Generative frameworks like Minimalist Program do not general explanations

Ellipsis Constructions: Classic NLP

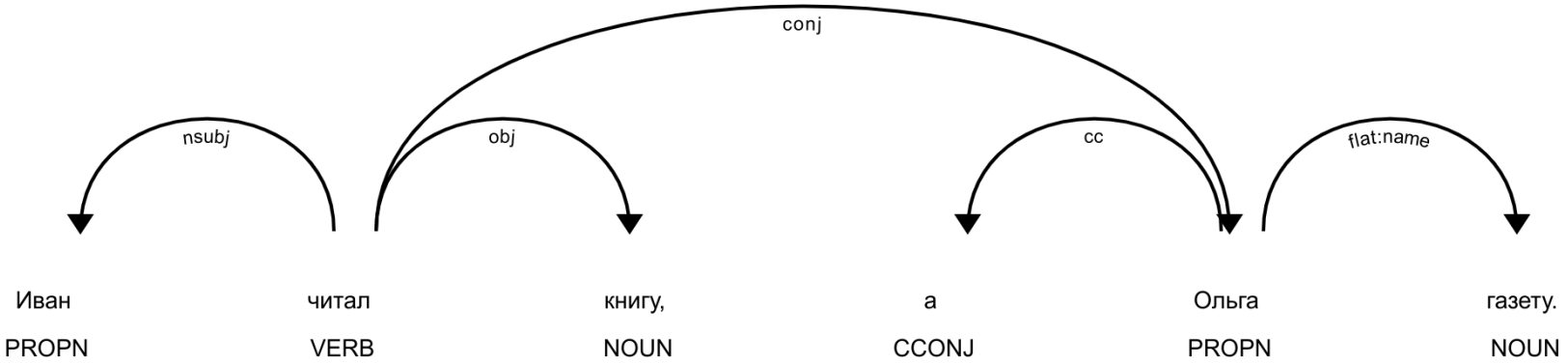
- Current State of the Art (SOTA) Natural Language Processing-pipelines and parsers perform poorly (or not at all)
- Tested SOTA parsers:
  - Stanford CoreNLP
  - Stanford Stanza (V 1.8.2) (Dependency & Constituent Parser)
  - Berkley Neural Parser (benepar) (V 0.2.0)
  - SpaCy 3.7
  - XLE (Web-XLE, Lexical-functional Grammar Parser)
- All parsers fail with Ellipsis (and other constructions) → not useful for downstream NLP tasks (e.g., relation extraction)

Dependency Parsers



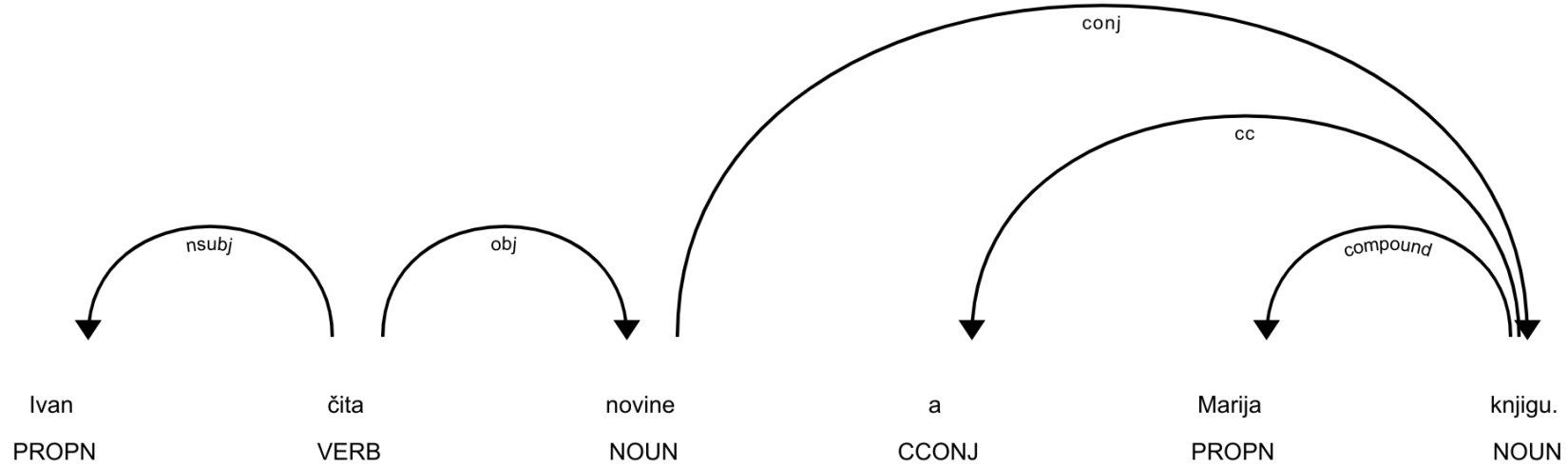
Stanza 1.8.2
Resulting assumption:
I will bring ice and beer (you?); coordination of "bring" and "beer"

Dependency Parsers



Stanza 1.8.2
Resulting assumption:
Coordination of "read" and "Olga" - compound direct object newspaper

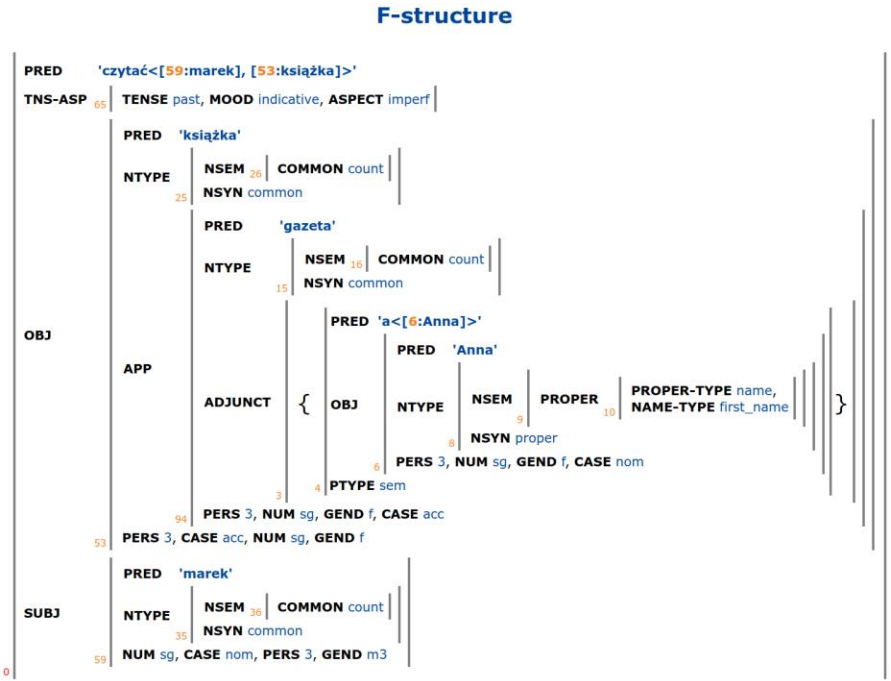
Dependency Parsers

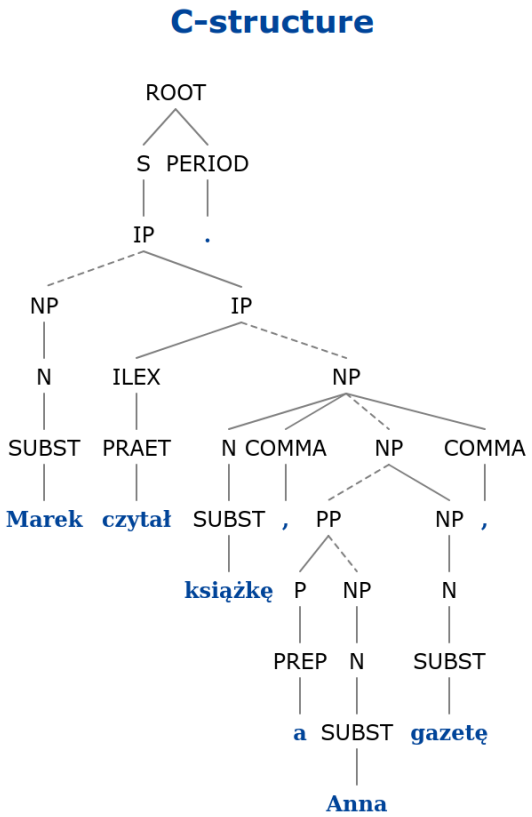


Spacy 3.7
Result: Coordination of "newspaper" and "book"

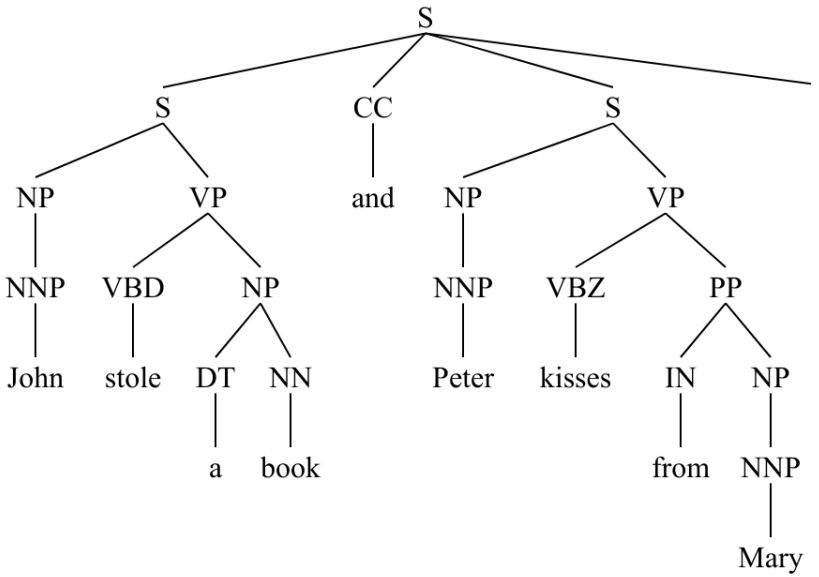
LFG Parser

XLE Web
Polish (POLFIE)





Constituent Parsers



Berkley Neural Parser

Head Noun of the object (kisses) is assumed to be the predicate head of the second conjunct.

Computational Methods and Experiments

- **Cloze test:**
  1. Used in Machine Learning – Masked Word Prediction in BERT (LM)
     - *The house ___ I was born.* (a. *where* , b. *which*)
- **LLMs:**
  1. Next word prediction as in GPT and other Large Language Models (LLMs)
- **Tasks:**
  1. Classification of sentences / utterances: Does it contain ellipsis or not?
  2. Detection of locus of ellipsis: indicate the space
  3. Guess of the missing words: fill in the missing words

Experiments

- 18 Languages with varying numbers of examples.
  - Largest: Russian, Polish, Ukrainian
  - Included: Croatian, Navajo, Gujarati, Chinese, Arabic, English, Spanish...
- Picked:
  - Up to 500 target sentences
  - 1000 or more distractors
  - For task 2 & 3: only examples with ellipsis are used.
- Algorithms:
  - Logistic Regression
  - BERT/RoBERTa-based Deep Learning model
  - GPT-4 Large Language Model (ChatGPT), Claude 3, Falcon2, Llama2, etc.
    - Using API = no memory related to testing N examples

SECTION 2

Data Collection

Building the Corpus

Sources:

- Literature (peer-reviewed articles and scientific books, literature)
- Corpora (valid and checked)
  - Croatian Language Corpus (Literature)
  - National Corpus of Polish

Polish	139
Russian	202
Ukrainian	158

Arabic (375), English (267), German (79), Gujarati (9), Hindi (127), Japanese (105), Korean (40), Kumaoni (85), Mandarin Chinese (40), Navajo (9), Norwegian (55), Spanish (171), Swedish (20), Telugu (20), Croatian, Bosnian, Serbian, Slovenian, Slovak...

Data Structure

```
A Nina ___ na pianinie.
----
A Nina gra na pianinie.
B: Kasia gra na klarnecie.
A: Marek śpiewa.
# source: Marjorie J. McShan (2000)
# TR eng: Nina plays piano.
```

Polish THEC gapping example with additional information.

Data Structure

The THEC data format allows for:

- adding syntactic tree annotations (bracketed notation, triples for dependencies, c- and f-structure)
- Easy processing with our own Python tools, simple scripts
- Extensible to allow for other linguistic annotations

- Simplifies evaluation using classical NLP-technologies and AIs/Large Language Models

Corpus Access

- In the next days: See NLP-Lab page
  - https://nlp-lab.org/ellipsis/
- Link to GitHub, allowing for collaboration and contribution.
  - https://github.com/dcavar/hoosierellipsiscorpus
  - Contribution and feedback welcome!

SECTION 3

Experiments

Goals

NLP Problems:

- Failing parsers and lack of downstream processing ⟶ semantic and pragmatic analysis
- Sentences with ellipsis undone parse mostly correctly

⟶ Undo ellipsis

- Assumption: ellipsis constructions do differ from unelided constructions, but syntactically this makes no difference, if ellipsis information is not lost and feeds the semantic mapping process

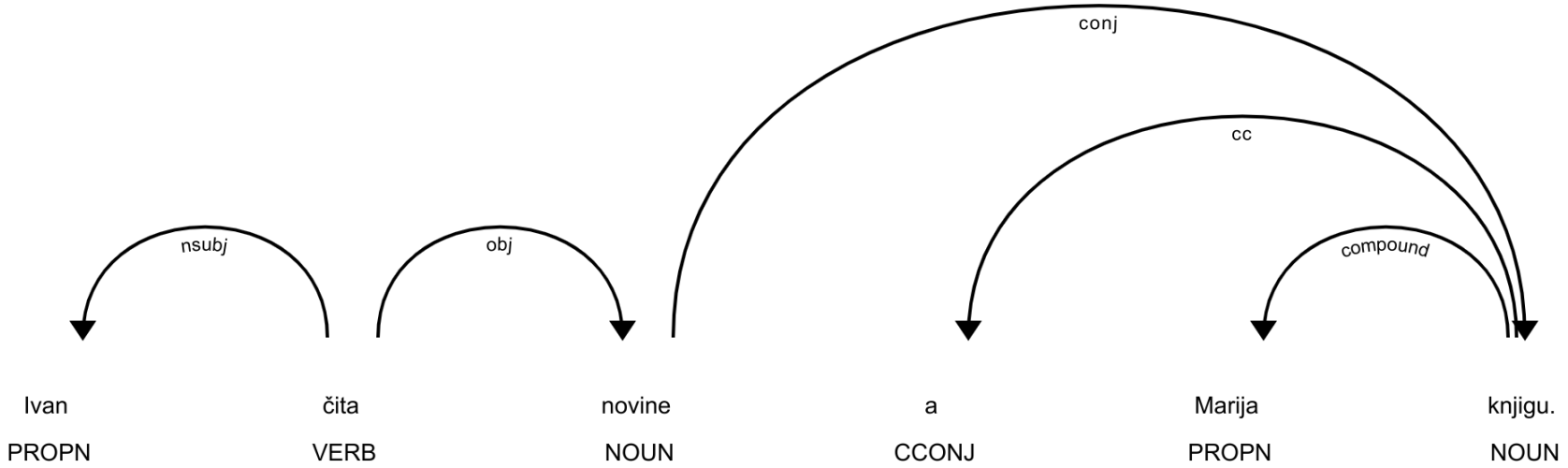
Example of ellipsis affecting quantifier scope or interpretation:

Nobody was smoking and ___ drinking.

Nobody was smoking and nobody was drinking.

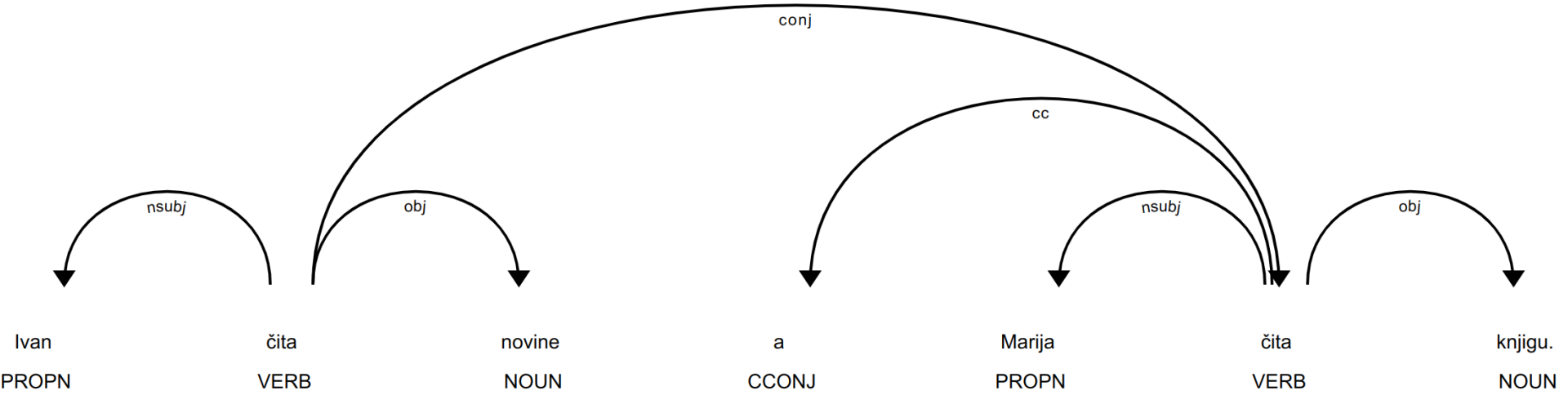
Example

Croatian: spaCy 3.7 Large Croatian Model



Example

Croatian: spaCy 3.7 Large Croatian Model



Experiments

Task 1: Does the sentence contain ellipsis?

Task 2: Where is the ellipsis? (growing complexity: one, two, three slots)

Task 3: What are the missing words / phrases?

Baseline: Logistic Regression using 8 basic features

- the number of nouns/NPs, subject dependency labels, object dependency labels, conjunctions, verbs, auxiliaries
- a boolean whether an interrogative pronoun is sentence/clause-final position

Experiments

English in comparison:

- Task 1:
  - Logistic Regression (baseline): accuracy 72%
  - BERT-based Transformer: accuracy 94%
  - GPT-3.5: accuracy: 35%
  - GPT-4: accuracy: 60%

BERT/Transformer > Logistic Regression > GPT-4

Experiments

- **For Arabic:**
  - We utilized GPT-4 (no other LLM was capable of processing Arabic)
    - Missing useful BERT-type LM for Arabic, we need to train one
  - Task 1: 0-shot classification
    - Baseline: Logistic Regression **83%**
    - GPT-4: Precision 0.56, Recall 0.18
  - Task 3: 0-shot word filling (single word task)
    - Accuracy ~80%

Experiments

	Russian	Polish	Ukrainian
LogReg	0.63	0.76	0.76
Task 1			
GPT-4 0-shot	0.74	0.73	0.7
GPT-4 few-shot	0.77	0.84	0.8
Task 2			
GPT-4 0-shot	0.28	0.44	0.35
GPT-4 few-shot	0.28	0.36	0.5
Task 3			
GPT-4 0-shot	0.2	0.26	0.24
GPT-4 few-shot	0.32	0.29	0.4

Conclusion

- Problems with "invisible words" in all parsers and LLMs
  - Parsers perform without a problem with "ellipsis undone"
- The problem is:
  - Theoretical – Dependency Grammar, Lexical-functional Grammar, etc.
  - Data-based – missing corpora with annotated ellipsis constructions
  - Computational – LLMs predict next words, and not next missing words (while BERT is trained on masked words)

Conclusion

- Goal:
  - We are developing tools to detect and undo ellipsis and to use classical NLP for representation learning/generation = syntactic trees, to build semantic representations in a subsequent step
  - Work on alternative computational grammar formalisms to handle ellipsis (as well as other construction types: islands, LDDs, discontinuities)

Thank you ~~for your attention~~!