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

Van Holthenrichs, Damir Cavar, Zoran Tiganj, Billy Dickson

March 22, 2024

Ellipsis constructions are common in all languages. In formal text types like legal documents and medical or financial reports, different types of ellipsis constructions are highly frequent. Often, complex sentences in such genres are long and contain more than one case of ellipsis in subordinate structures. Despite their ubiquity, Natural Language Processing (NLP) technologies and syntactic parsers fail to analyze constructions like the Polish example in (1).

(1) Piotr lubi waniliowe lody a Janek i Basia ____ czekoladowe ____ . (Pol)
Piotr likes vanilla ice.cream and Janek and Basia ____ chocolate ____ .
'Piotr likes vanilla ice cream and Janek and Basia (<u>like</u>) chocolate (<u>ice cream</u>).'

The underscore in (1) indicates the position of elided elements. In this particular construction, the words *lubi* and *lody* are elided in two independent positions.

Our motivations for studying qualitative and quantitative properties of ellipsis in different Slavic languages are the following:

1. Study cross-linguistically the effects of ellipsis on semantic interpretation, as well as the prosodic properties and morpho-syntactic peculiarities of ellipsis, and to our knowledge, no comparable corpus for Slavic languages exists.

2. Provide computational means to generate semantic representations of sentences, in particular, those containing ellipsis.

Solving goal 2 requires us to use a parser that is capable of generating an adequate syntactic representation that allows us to map it to appropriate semantic representations in a generative framework.

Unfortunately, current state-of-the-art (SOTA) Natural Language Processing (NLP) pipelines for Slavic languages (in particular for Croatian, Czech, Polish, Russian, Slovak, and Ukrainian) fail to provide adequate analyses for ellipsis constructions. Independent of the syntactic framework and the type of syntactic parser that we evaluated, the resulting syntactic representations generated by these parsers were inadequate and of limited use for our goals. Our experiments included SOTA parsers based on the Dependency Grammar framework, simple Constituent parsers, and advanced Lexical-Functional Grammar parsers. For all common ellipsis construction types, the parsers fail to provide adequate parse trees.

We defined core tasks to engineer computational models to detect and parse ellipsis constructions. The computational tasks are the following:

- Detect whether a sentence contains ellipsis.
- Identify the positions of elided elements in sentences with ellipsis.
- Reconstruct the elided words in the correct positions.

For these tasks, we collected corpora from multiple Slavic languages. In this article, we present the corpus data and format that we designed for the core tasks. We demonstrate the experiments with SOTA technologies like the Stanza and spaCy Dependency Parsers trained on the Universal Dependencies treebank, the Berkeley Neural Parser, the Xerox Linguistic Environment (XLE) using Polish grammar, and various other NLP technologies. We demonstrate that Large Language Models (LLMs) like GPT-4 perform worse than our own trained models on the third task.