Krystyna Bojałkowska
Nicolaus Copernicus University in Toruń
Marcin Woliński
Institute of Computer Science Polish Academy of Sciences

On correlation of inflectional and syntactic description of the Polish language in treebanks based on the National Corpus of Polish

Key words: corpus, inflection, syntax, treebank, multi-word expression

The paper concerns the representation of syntactic relations in treebanks based on the National Corpus of Polish (www.nkjp.pl).

The National Corpus of Polish (henceforth: NKJP), and especially its manually annotated one-million-word subcorpus, has become the basis for various syntactic studies in the form of treebanks. Descriptions in various formalisms can be mentioned here: Polish Dependency Bank (Wróblewska 2014), reporting dependency relationships between words; Składnica (Woliński et al. 2011, Woliński 2019), describing the constituent structure; a treebank comprising LFG structures (Patejuk and Przepiórkowski 2018). The treebanks mentioned were created or verified manually, but parsebanks created using automatic tools also get created.

Since Polish is an inflectional language, the description of the syntactic structure is strictly dependent on the inflectional characteristics of individual word forms (which involves assigning them to the grammatical class of lexemes and identifying the values of inflectional categories in the case of inflected lexemes). In this paper, we will focus on the problems that arise at the junction of these two descriptions. The NKJP enforces a rule that inflectional forms cannot contain spaces. As a result, it is necessary to reflect on all types of multi-word expressions appearing in the text, having various levels of lexicalization, independence of meaning, regularity of creating analogous constructions or, conversely, atypicality.

Some of such units can be treated as syntactically compositional, for example, multi-word inflectional forms of verbs (complex future tense, the subjunctive mood and complex forms of the imperative mood). Some of them sometimes create discontinuous structures, cf. *Student **będzie** jutro **zdawał** egzamin* ('A student **will** tomorrow **take** an exam'). Another construction on the border between inflection and syntax are complex adjective forms of the *biało-czerwony*

('white-red') type (compositional and productive). However, there are also syntactically non-compositional units, such as particles comprising more than one word form (e.g. *co najmniej* 'at least'), conjunctions (e.g. *o ile* 'as long as'), prepositions (e.g. *wraz z* 'along with') or adverbs (e.g. *na długo* 'for a long time'). In their case, the subject of discussion may be what description their components should receive in inflectional marking.

Moreover, problems related to the syntactic representation of structures containing words whose assignment to a specific class of lexemes is difficult will also be presented. Examples include *niż* 'than', *jako* 'as', *zamiast* 'instead', which are described in studies of the Polish language as prepositions or conjunctions – depending on the context and properties of use.

In the talk, we will present problematic constructions and show structures we have selected for them in the *Składnica* treebank. However, we hope that the analysis can be useful for other treebanks.

Literature

Kieraś Witold, Woliński Marcin (2017), *Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego,* "Język Polski" XCVII, p. 75-83.

Patejuk Agnieszka, Przepiórkowski Adam (2018), *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically Informed Treebanks of Polish,* Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Przepiórkowski Adam et al. (2012): *Narodowy Korpus Języka Polskiego,* eds. A. Przepiórkowski, M. Bańko, R. Górski, B. Lewandowska-Tomaszczyk, Warsaw.

SGJP: *Słownik gramatyczny języka polskiego* (2020), Marcin Woliński, Zygmunt Saloni at al., ed. 4., https://sgjp.pl.

Woliński Marcin, Głowińska Katarzyna, and Świdziński Marek (2011), *A preliminary version of Składnica—a treebank of Polish*. In: *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, ed. Z. Vetulani, Poznań, p. 299–303.

Woliński Marcin (2019), *Automatyczna analiza składnikowa języka polskiego*, Warsaw.

Wróblewska Alina (2014), *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*, Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.