

Grammatical profiling of Czech nouns: what do cases tell us about nouns' meanings

Michal Láznička & Vojtěch Janda
Charles University, Prague
michal.laznicka@ff.cuni.cz

Keywords: animacy; case semantics; Czech nominal morphology; cluster analysis; grammatical profiles

Nouns in Czech are inflected for case and number and are categorized by gender and inflectional class. While some cases have limited and well-defined functions (e.g. the nominative as a subject case), others are more complex (e.g. the instrumental expresses commitatives, instruments, locations, and agents in passive constructions). In this paper, we analyze the relationship of case and meaning from a novel perspective of grammatical profiling. Grammatical profiles, relative frequency distributions of all inflected forms of lemmas in a corpus, were introduced by Janda & Lyashevskaya (2011) as a subtype of the behavioral profile method (e.g. Gries & Divjak 2009) and provide an opportunity to study variation in form and its relation to meaning on a relatively large scale.

By employing the grammatical profile method, we show to what extent patterns of use of grammatical features are interpretable through word meaning and how stable this relationship is across different corpora. We extracted grammatical profiles of all nouns in four different corpora of Czech (written and spoken Czech, movie subtitles, translations of the proceedings of the European Parliament). After filtering out low frequency lemmas ($ipm < 7$), we analyzed a set of 821 noun lemmas that appear in all four corpora. The distributions in each corpus were analyzed using hierarchical clustering to identify lemmas with similar profiles. Obtained clusters were studied with respect to the relative frequency of individual cases and the meanings of cluster members.

The results show that there are potentially meaningful groupings in the data. We identified four clusters that are, with one exception, found in all four corpora. The clusters suggest that the analyzed distributional patterns may be interpreted as an interplay between word meaning and more abstract syntactic patterns. For instance, one of the clusters is characterized by a high proportion of nominative uses and contains almost exclusively human referents, confirming the expectation that animacy will be a strong predictor of noun behavior. A different cluster has a high proportion of genitives and contains nouns that express quantity (e.g. units of measure), a consequence of the fact that most Czech numerals govern the genitive case (e.g. *pět litrů* 'five liter.pl.gen'). The results also show similarities and differences between corpora, such as the differences in relative frequencies of the nominative, genitive, and instrumental case in the "human" cluster. The results of the analysis thus show that it is possible to "go from form to meaning via statistics" (Kuznetsova 2015) and that the distribution of relatively abstract grammatical categories which oftentimes have multiple functions is to an extent driven by meaning, as much in potentially abstract written texts as in spontaneous spoken data.

References

- Gries, Stefan Th. & Dagmar Divjak (2009). Behavioral profiles: a corpus-based approach towards cognitive semantic analysis. In Evans & Purcel (eds.), *New Directions in Cognitive Linguistics*. John Benjamins, 57-75.
- Janda, Laura A. & Olga Lyashevskaya (2011). Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian. *Cognitive Linguistics* 22 (4): 719-763.
- Kuznetsova, Julia (2015). *Linguistic Profiles*. De Gruyter Mouton.