Grammar and Corpora 2020
Second Submission

Same question, different annotation depths: early Slavic dative absolutes in deeply versus shallowly annotated treebanks

This paper exploits data from the TOROT Treebank (Eckhoff & Berdicevskis 2015) to discuss the extent to which corpora with different annotation depths can contribute to the investigation of a syntactic phenomenon in early Slavic. The dative absolute (DA), a type of participial adjunct, is used as a case study. The widespread intuition in the literature is that its usage is better understood at the discourse-structural level, where it seems to have a framing, backgrounding function (Worth 1994; Collins 2004, 2011, Sakharova 2010). The goal is to address this intuition systematically by combining treebank data with formal frameworks for discourse representation, particularly Segmented Discourse Representation Theory (SDRT; Asher & Lascarides 2003).

Drawing from similar approaches to Ancient and New Testament Greek participles (i.e. Bary & Haug 2011; Haug 2012), the results strongly indicate that DAs are identifiable as 'frames' or 'stage-setters' from a discourse-structural perspective, regardless of their syntactic configuration (e.g. subject co-reference vs. switch-reference with the matrix clause; juxtaposition vs. overt subordination or coordination). This is established in the paper on the basis of indirect evidence emerged from lexical, morphosyntactic and information-structural annotation on Old Church Slavonic (OCS) texts, as well as their aligned Greek parallels, that is, thanks to corpus data containing the deepest annotation available in the treebank. Later Church Slavonic and early Slavic original texts are instead only partially represented in the corpus, and have overall much shallower annotation. This is at best limited to morphology (automatically performed, but often in need of post-correction), and fragmentary dependency annotation. A comparison between translated and original Slavic texts, as well as between early and later Slavic sources, can however be crucial for a proper understanding of early Slavic constructions which do not stand out as either slavish Greek calques (e.g. *eže*-nominalised infinitive, cf. MacRobert 1986) or genuinely Slavic phenomena (e.g. *possessive constructions*, cf. Eckhoff 2018), as is the case with participle clauses. Since speedy deep annotation is unfeasible through manual tagging, an alternative strategy has been employed to tackle our specific case study using treebanks with very different annotation depths: all potential prototypical configurations of the dative absolute are isolated on the basis of the frequencies calculated on OCS across several variables (e.g. position in the sentence; lemma-tense correlations; aspectual shifts; information status) in order to minimise the amount of targeted annotation needed for newly added texts. This is achieved by extracting highly predictable configurations first, allowing for closer inspection only of less prototypical uses.

Throughout the paper the usefulness and limitations of exploiting treebanks with different annotation depths are noted. In particular, shallowly annotated corpora proved useful in the possibility to extract occurrences of the relevant constructions by means of morphological pre-processing alone, which is shown to be relatively inexpensive from the computational perspective, thanks to the recent advances in automatic morphological analysis of pre-modern Slavic (Scherrer et al. 2018; Scherrer & Rabus 2019). In our case study, texts containing only morphological annotation where also strategically annotated with dependency annotation, which not only demostrated to be a useful means of corroborating patterns emerged from deeply annotated treebanks, but also revealed that DAs may be employed exclusively as specialized topic-shifters, a usage which did not immediately emerged from the distant-reading approach employed on deeply annotated treebanks. One of the main advantages of having syntactic-dependency annotation in addition to morphological analysis is instead the opportunity to compare potentially competing constructions by looking for grammatical functions rather than inflection. In our case study, the properties of subjects gave crucial insights into the different functions of dative absolutes and finite temporal subordinates (*egda*-clauses), the former having information-structurally more prominent subjects than the latter. Particularly for texts which contain an overall limited number of event participants, information-structural annotation thus seems to be particularly useful to systematically assess anaphoric phenomena in large stretches of discourse, rather than relying on a case-by-case approach when checking the relation between referents across sentences.

Overall, this study confirms that even only on the basis of translations - which is necessarily the case for the earliest stages of Slavic - and of datasets of limited size, historical corpora can be used to test hypotheses and provide valuable insights on the syntax of the relevant language. This is particularly encouraging for the study of early Slavic syntax: deeply annotated OCS treebanks can be exploited to formulate informed predictions on a given construction in previously unannotated texts, and, ideally, as a relatively solid guideline to analyse its behaviour in later Slavic texts, before giving in to a fully corpus-driven approach.

References

Asher, Nicholas & Lascarides, Alex. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

Bary, Corien & Dag Haug. 2011. Temporal anaphora across and inside sentences: The function of participles. *Semantics and Pragmatics* 4. 1-56.

Collins, Daniel E. 2004. Distance, subjecthood, and the early Slavic dative absolute. *Scando-Slavica* 50. 165–181.

Collins, Daniel E. 2011. The pragmatics of "unruly" dative absolutes in Early Slavic. In Eirik Welo (ed.), *Indo-European syntax and pragmatics: contrastive approaches*, Oslo Studies in Language 3(3). 103–130.

Eckhoff, Hanne M. & Aleksandrs Berdicevskis. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta* 14–15. 9-25.

Eckhoff, Hanne M. 2018. Quantifying syntactic influence: Word order, possession and definiteness in Old Church Slavonic and Greek. *Diachronic Slavonic Syntax: The Interplay between Internal Development, Language Contact and Metalinguistic Factors*. Berlin/Boston: Mouton De Gruyter. 29-62.

Haug, Dag. 2012. Open verb-based adjuncts in New Testament Greek and the Latin of the Vulgate. In Cathrine Fabricius-Hansen & Dag Haug (eds.), *Big events and small clauses*, 287-321. Berlin: Mouton de Gruyter.

MacRobert, Mary. 1986. Foreign, naturalized and native syntax in Old Church Slavonic. *Transactions of the Philological Society* 84(1). 142-66

Scherrer, Yves & Achim Rabus. 2019. Neural morphosyntactic tagging for Rusyn. *Natural Language Engineering* 25(5). 633–650.

Scherrer, Yves, Achim Rabus & Susanne Mocken. 2018. *New developments in tagging pre-modern orthodox Slavic texts*. *Scripta & E-Scripta* 18. 9–33.

Sakharova, Anna. 2010. K voprosu o diskursivnych funkcijach pričastnych konstrukcij v russkoj letopisi. *Russian Linguistics* 34. 87–11

Worth, Dean S. 1994. The dative absolute in the Primary Chronicle: Some observations. *Harvard Ukrainian Studies* 18. 29–46.