Keeping linguistic variation across time and space: tagging of a diachronic and diatopic corpus of Balkan Slavic

Teodora Vuković, Ivan Šimko
Slavisches Seminar, University of Zurich

A major challenge for annotating historical corpora are morphosyntactic categories that underwent changes in both form and function. This includes considerable synchronic variation in both respects, such that processing tools for contemporary standard languages cannot be applied. Preserving information on variation is needed in order to approach questions such as internal change, dialectal distribution and areal developments. In order to accommodate for the encountered variation, we utilize mixed approach of morphological and syntactic tagging matching forms with their functions.

One prime example is the expression of case from the oldest attested documents (Old Church Slavonic) up to the contemporary standard languages (Bulgarian, Macedonian, Serbian) and non-standardised varieties (Torlak). Diachronic change concerns the replacement of nominal inflectional marking and the emergence of morphological coding outside of nouns. See the marking of the third argument of '*give*' by means of nominal inflection in (1), dependence marking by means of *na* in (2), marking by both *na* and nominal inflection in (3) and additional preverbal pronominal clitic marking in (4).

(1)	*Dati*	*razumъ*	*[...]*	*ljudemъ.*
	give.INF	reason.M.SG.ACC		people.M.PL.DAT

'give reason to the people'

(OCS, Luke 1.77, Codex Marianus, TOROT, Eckhoff and Berdicevskis (2015))

(2)	*Toi*	*go*	*davaše*	*skrišom*	*na*	*syromasi.*
	he.M.SG.ACC	he.CL.ACC	give.3SG.IMPF	secretly	to	poor.M.PL

'He was giving secretly to the poor'

(Balkan Slavic 17th c., Tixonravovskij damaskin, 2.10b, authors' data)

(3)	*Dadem*	*si*	*na*	*unuku.*
	give.1SG.PRES	REFL.DAT.CL	to	grand-daughter.F.SG.ACC

'I give to my grand-daughter'

(contemp. Torlak, Gornja Sokolovica, authors' data)

(4) *Mu* *ja* *dadov* *knigata* *na*
DAT.M.3SG she.CL.ACC give.1SG.PST book.F.SG.DEF to
studentot.
student.M.SG.DEF

'I the book to the student.'

(contemp. Macedonian., Tomic (2006))

In order to empirically analyze diachronic change and areal variation in the morphosyntax of Balkan Slavic we are building a corpus of different diachronic and diatopic varieties of Bulgarian, Macedonian and Torlak. Since morphological realizations of categories under scrutiny are different in each stage, currently available automatic annotation fail in recognizing them. For this reason, we are creating custom taggers for the corpus, one for older stages and another one for contemporary data. We are adapting a widely accepted standard scheme for morphosyntactic tagging developed within the MULTEXT-East project, with the goal of making the resources internally and externally comparable.

In doing so, we are faced with one particular challenge. Obviously, one and the same grammatical category uses different morphological forms or a different set of morphological exponents in each variety. Morphological formulations can be linked to syntactic functions. We are adding syntactic annotation to help us with the task of tracking change and variation within individual grammatical units, modifying the scheme given by the Universal Dependencies project.

Combining two annotation methods, as proposed, we will bridge the problem of variation. Consequently, the method will be applicable not only on the diachronic but also on the diatopic dimension. In addition, it allows to trace longer term changes (OCS > contemporary), smaller-scale and more recent changes, and facilitate comparison across families, including potential areal developments.

References

Hanne Martine Eckhoff and Aleksandrs Berdicevskis. Linguistics vs. digital editions: The tromsø old russian and ocs treebank. pages 9–25, 2015.

Olga Mišeska Tomic. *Balkan Sprachbund Morpho-Syntactic Features.* Studies in Natural Language and Linguistic Theory. Springer Netherlands, 2006.