On the mutual benefit of matrėški, triangulation and token-based research:  
the eastern part of the Circum-Baltic area as a test case

Abstract:
My talk divides into two parts. In the first part I will survey cases of structural convergence in 
the eastern part of the Circum-Baltic area, with a focus on the Slavic-Baltic contact zone (for 
which cf. Wiemer 2013; Wiemer et al. 2014). The surveyed cases will be evaluated in terms of 
the matrėški-approach, which assesses the areal spread of some feature, or a feature aggregate, 
on the background of immediately surrounding geographic circles. To this an account of 
genealogic affiliation and possible more universal tendencies should be added. This approach 
has been designed to help single out unique convergence in smaller areas and determine whether 
smaller areas are not just parts of larger clines (Wiemer, forthcoming). In particular, a key 
question is whether single or aggregate features in the western part of East Slavic are better 
captured as (parts of) areal continua or rather as clusters that stand out on some larger areal 
background. Put otherwise: with respect to which features can certain groups of varieties more 
adequately be grouped into subareas (clusters)? And with respect to which features do smooth 
transitions obtain (compare Heeringa/Nerbonne’s 2001 investigation on Dutch dialects)?

I will also address problematic issues of this approach, which concentrates on diatopic 
variation, because an adequate understanding of variation also has to account (i) for diastratic 
dimensions, (ii) for the range of admissible lexical input and (iii) for the productivity of the 
chosen features. The latter can roughly be captured by type/token ratios. However, so far both 
dialect geography and typology have avoided token-based investigations, one of the reasons 
being the lack of suitable corpora.

This brings us to the second part of my talk, which will be devoted to an argument in favour 
of corpus-based, or even corpus-driven, research into variation in non-standard varieties. I will 
discuss the pluses and minuses of such an approach, mainly by using examples from extant 
research into East Slavic and Baltic, part of which has been carried out in the TriMCo-project 
(www.trimco.uni-mainz.de). This part of my talk is closely connected to Kirill Kozhanov’s talk 
at the same conference.

Apart from the debate about pros and cons of corpus-based or -driven research there is a 
more general problem, which hitherto seems to have been overlooked (or neglected) in corpus-
building. Namely, dialectal (or simply spoken) data are often ambiguous, or vague, in that the 
category membership of particular tokens often cannot be clearly determined. (A very similar 
problem occurs with diachronic corpora.) I argue that, for an adequate treatment of primary 
data and natural variation, one should avoid the preconceived projection from notions and 
categorial distinctions based on standard languages. Instead, discourse tokens which are 
ambiguous or vague should be treated in their own right in bottom-up analyses. They should as 
such remain retrievable, and annotation should suit this requirement.

References
Heeringa, Wilbert & John Nerbonne (2001): Dialect areas and dialect continua. Language Variation and 
Change 13, 375-400.
Wiemer, Björn (2013): Zur arealen Stufung im baltisch-slavischen Kontaktgebiet (und dabei 
auftretenden methodischen Desideraten). In: Kempgen, Sebastian, Monika Wingender, Norbert