Rusyn as a language between state borders – a statistical approach to variation (for small sample sizes).

This paper aims to a) report on current developments in the research of Carpatho-Rusyn varieties, b) address the issues that arise when applying state-of-the-art statistics and variational linguistics methods to small data samples.

With the help of computational methods, we endeavor to investigate dynamic processes in the Slavic minority language Rusyn, which is spoken predominantly in South-West Ukraine, East Slovakia and South-East Poland. The setting of Rusyn language is divided by several political and linguistic borders as it not only marks a point between East and West Slavic, but also spreads across many national borders as well as the outer border of the EU. Nowadays, the speakers of Rusyn live to a greater extent in a dynamic environment and under constant and evident pressure by their respective roofing state languages Ukrainian, Polish and Slovak. In this fashion, new divergences within the old Rusyn dialect continuum due to contact with the majority language, that is, so-called border effects, are to be expected (Wieling et al. 2018, Rabus, 2015; Woolhiser, 2005).

Not only in order to trace these divergences but also to make Rusyn vernacular more accessible for further empirical research, an online corpus of Spoken Rusyn has been created. The transcribed and annotated collection of spoken Rusyn allows us to draw conclusions about the impact of various sociolinguistic factors on spoken Rusyn. For specific cases of variation, sociolinguistic factors, such as age, gender, living / birth place etc. are contrasted to the impact of the factor "variety" (e.g. Ukrainian, Slovak, Polish), to confirm or refute the above mentioned hypothesis of border effects.

A case of variation within spoken Rusyn we want to discuss is the threefold variation of the verbs мати$_{3Ps.Sg.Pres.}$(*ма*, *має*, *мат(ь)*) and знати$_{3Ps.Sg.Pres.}$(*зна*, *знає*, *знат(ь)*) within three varieties of Rusyn. We assume that the usage of the several verb forms could be mainly influenced by one sociolinguistic factor, which is the respective roofing languages. The distribution of the forms is very heterogeneous in-between the varieties, but also the frequency in-between speakers of one and the same variety is varying greatly in size. This leads to an unbalanced data set with many outliers. On top of that, due to the time intensive transcription process of spoken language data, the sample size is rather small. Thus, in preparation for statistical analysis we want to avoid the cleaning of the data (to leave out speakers with an especially high or low number of utterance) at any cost but instead use statistical methods in order to validate the conclusions we draw, based on our tests. For statistical analysis the software R (R Core Team 2019) is used.

We want to draw attention to the pitfalls of statistical analysis of categorical data and how we can limit the danger of drawing wrong conclusions (type I and type II errors). Doing quantitative sociolinguistic research on spoken language, data of the categorical type is more the rule than the exception. Nevertheless, the methodology dealing with categorical data is rather not common in Slavistics. We want to encourage the use of the modern methods and propose to report bootstrapped statistics (Davison, A. C. & Hinkley 1997) and reliable confidence intervals instead of falsely positive results or disclaiming statistical approaches. Ultimately we use multinomial logistic regressions (with the R-Package *nnet* (Venables & Ripley 2002) to infer about the effects of sociolinguistic factors. We can validate the result of the regression 1. by bootstrapping and 2. cross-validation of the model predictions. In this way we can prove our hypothesis, without losing a great amount of our rather small data set due to cleaning of outliers or splitting for training/test purposes. State of the art methods like multinomial regressions combined with bootstrapping and cross-validation allow us to conclude, that the "variety" of Rusyn is by far the most important factor determining the outcome variable

compared to other sociolinguistic factors, without solely relying on a small data set or purely qualitative analysis.

Keywords: Corpus linguistics; dialectology; Rusyn dialect; R; categorical data; multinomial regression; bootstrapping; cross-validation.

Random examples of variation of the verbs мати$_{3Ps.Sg.Pres.}$ (*ма, має, мат(ь)*) and знати$_{3Ps.Sg.Pres.}$ (*зна, знає, знат(ь)*)

HM1950/SLO: 157337- 162002: «Але інакше зо женов, фурт по руснацькы бісїдує, ай дївча *знає* по руснацькы.»

JB1958/SLO 695517- 700422: «Планы, я сі думам, же каждый, каждый *мать* планы лен залежіть як...»

VSh1965/TRA: 488583- 494430: «Мама ще жыє, має вісімдесят єден рік, отиць має дівяносто чотыри, тоже ще жыє.»

References:

Бандрівський, Д., Л. Григорук, Ф. Жилко, Й. Закревка, А. Залеський, Т. Назарова, М. Онишкевич, П. Приступа. (1988). Атлас українької мови: Волинь, Наддністрянщина, Закарпаття і суміжні землі. Том другий. Київ.

Chomiak, M., H. Fontański (2004): Грамматика лемковского языка = Gramatyka języka łemkowskiego. Warszawa.

Davison, A. C. & Hinkley, D. V. (1997) Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge

Латта, Василь (1991): Атлас українських говорів східної словаччини: Пряшев: Словацьке педагогічне видавництво.

Rabus, A. & A. Šymon (2015): На новых путях ісслідованя русинськых діалекту. Корпус розговорного русинського языка. In: Копорова, Кветослава (Hrsg.): *Русиньскый літературный язык на Словакії. 20 років кодіфікації*. Prešov, 40–54.

Rabus, A. (2019): Sprachwissen versus Sprachgebrauch im gesprochenen Karpatorussinischen. *Zeitschrift für Slavische Philologie*, 75(2), 347–370

Wieling M., Valls E., Baayen R.H., Nerbonne J. (2018) Border Effects Among Catalan Dialects. In: Speelman D., Heylen K., Geeraerts D. (eds) Mixed-Effects Regression Models in Linguistics. Quantitative Methods in the Humanities and Social Sciences. Springer, Cham

Woolhiser, C. (2005). Political borders and dialect divergence/convergence in Europe. In Peter Auer, Frans Hinskens, and Paul Kerswill, editors, Dialect change, 236–262. Cambridge Univ. Press, Cambridge

Software:

R: R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Url: https://www.R-project.org/.

Package “nnet”: Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer.