A tool for measuring spatial saturation in large-scale dialectological survey data

Curdin Derungs
University of Zurich, Department of Geography
Adrian Leeman
Lancaster University
Luca Scherrer
University of Zurich

Present-day technology allows for large-scale crowdsourcing of dialect data of unseen sample sizes (e.g. Goncalvez & Sanchez 2014, Vaux 2016). In 2015, Leemann et al. (2015) released 'Grüezi, Moin, Servus', a web-app that asks users 24 questions about their language use, in particular regional lexical variation. More than 800,000 users have participated, resulting in response data from 18,000 localities. Given this extraordinarily rich and dense spatial distribution of lexical variation, we raise the following questions:

> *To what degree is the data saturated? Is more data needed to cover gaps of information or is the majority of information redundant?*

These seemingly trivial questions are difficult to tackle given that we lack a sound definition of saturation. Apart from cursory discussions by Goebl (e.g. 2006) we are not aware of a methodological framework that accounts for this type of saturation . In Leemann's survey, for instance, each question on average has 10 answers (regional lexical variants). In theory, this results in 10^{24} combinations to answer the survey, assuming all answers are independent. In reality, however, answers are dependent and only 650,000 of all 800,000 responses are unique, with some occurring >100 times. This dependency is an advantage, as it allows to define reasonable sample sizes – yet, it introduces statistical complexity. Further, it would be worthwhile exploring saturation for global as well as local patterns.

In this contribution, we introduce an open-source R tool (*LingSat*) which enables in-depth analysis of saturation patterns in dialectological survey data. Based on several dialect surveys we demonstrate how a thorough understanding of saturation not only fosters i) a better understanding of how many participants should be incorporated in a linguistic experiment, but also ii) which local information gaps should be covered with additional responses.