Quantitative analysis of language contacts within the Finnic group

Finnic languages have been vigorously studied since the beginning of the 20th century with traditional linguistic methods (e.g. Setälä 1926, Sammallahti 1977, Itkonen 1984, Kallio 2015). Nevertheless, there is as yet no uniform view on the history of the Finnic-speaking area and how the Finnic languages have taken shape (Kallio 2015). The challenges in detecting the past of the Finnic group are largely due to the contacts within the group. I study here the language contacts within the Finnic group with quantitative methods adopted from biology and with datasets including both standard and non-standard varieties. I focus specifically on the role of the Gulf of Finland as Finnic languages encircle it and it has been proposed to act as both an inhibitor and a promoter of contacts (Leskinen 1999).

I study contacts within the Finnic group by calculating “language flow” between languages in a comparable way as gene flow is calculated between biological populations. I calculate “language flow” with BAPS (Corander et al. 2003), which is a model-based clustering method used in population genetics to cluster genetic data and to define the direction and strength of gene flow between populations. BAPS has been used earlier to study language flow between the dialects of Finnish (Santaharju et al. ms). I will also illustrate the connections between languages with a NeighborNet network (Huson & Bryant 2006) in a comparable way as was done for the Uralic languages in Lehtinen et al. (2014). In this study, more Finnic languages will be included in the analysis to obtain a more complete picture of the connections within Finnic. To study the role of the Gulf of Finland, I will calculate linguistic and geographical distances and correlate these to investigate the importance of geographical distance in linguistic similarity. Depending on whether I find connections between linguistic populations on different sides of the Gulf of Finland it will be possible to make inferences about its role in contacts.

I study this question with two kinds of linguistic datasets. The first is Atlas Linguarum Fennicarum (ALFE; Tuomi 2004, 2007, 2010), which describes lexical, morphological and phonological variation in the Finnic languages and in their dialects. The atlas covers the whole Finnic area including also a variety of non-standard varieties. ALFE has been compiled as a multinational effort since the 1980s and it was recently published in digital format (in 2013-2014) and is openly accessible through the AVAA portal of the Institute for the Languages of Finland. The second dataset, UraLex, is a database of standard lexical variation among the Uralic languages. This data consists of basic vocabulary (Swadesh lists; Swadesh 1952, 1955, items from the Leipzig-Jakarta list; Tadmor 2009) in 25 Uralic languages coded according to their shared origins (i.e. cognate relationship).
There is eight Finnic languages in the dataset, which I will use in my analyses. UraLex data has been collected by the BEDLAN team and will be published as part of the Lexibank database of Max Planck Institute for the Science of Human History.

According to the networks made with the standard varieties of the UraLex data, the clearest division was between the groups of Finnish, Karelian, Veps, Ingrian and Votic, Estonian, Võru, Livonian. Of these languages Votic and Estonian showed more connections with the northern group than Võru and Livonian. Thus, the preliminary results suggest that contacts are most intense between the neighboring languages circulating the Gulf of Finland. However, this view will be further adjusted when including the non-standard varieties in the analyses.

**Bibliography**


