New wave of quantitative dialect studies

The linguistic variation seen between languages now stems from earlier language internal variation. Language internal, i.e. dialectal, variation has been studied with traditional linguistic methods since 19th century, and since the 1950´s there have been attempts to quantify the spatial patterns of linguistic variation in a more objective way with quantitative methods (Chambers & Trudgill, 1998). However, the step where methods developed for biological research are applied to dialect data has not yet been taken even though these methods have been applied to study between-languages variation through suggested analogies between biological and linguistic evolution (Croft, 2000; Pagel, 2009).

In the same way as in biology within-species genetic variation may be structured as populations, analogously language internal linguistic variation is structured as dialects. Thus, population genetic methods could be suggested to be suited for analyzing dialect data. Some of these methods have already been applied to language data (Reesink, Singer, & Dunn, 2009; Bowern, 2012) but whether the applicability extends also to dialect data remains unsolved.

Here we study language internal variation and linguistic population structure with old Finnish dialect data and reflect the dialectal variation also to variation in extralinguistic variables. We apply population genetic methods to dialect data to get a better understanding about forces that have shaped and maintained the language internal variation of Finnish language through times and which possibly could play a role in shaping the pattern of variation also in other languages. Thus, the advantages of population genetic methods arise for example from extending the study of language evolution beyond linguistic material.

We used data from Finnish dialect Atlas, which includes 213 map sheets of phonological, morphological and lexical features collected from Finnish speaking municipalities in the area of Finland in the 1920-1930 (Kettunen, 1940). We applied population genetic STRUCTURE (Pritchard, Stephens, & Donnelly, 2000) software to objectively cluster the data to dialects. We show how STRUCTURE turns the linguistic variation seen in the map sheets into frequency data per municipality, which allows us to see how strongly each municipality belongs to each dialect area and separate the municipalities to focal and transitional dialect areas. We compared the clusters created with STRUCTURE to clusters obtained with K-medoids and to the traditional view of Finnish dialects and got strong support for usability of STRUCTURE to dialect data. We further calculated various metrics from the dialect data (e.g. heterozygosity and Shannon-Wiener diversity index). Also linguistic distances (Φ_{PT} values) were calculated and compared to averages of extralinguistic (geographical, environmental and cultural) variables for each dialect area with Mantel, Partial Mantel and MRM analyses to see the relative contributions of each of these variable groups to linguistic variation.

We found that language internal variation is structured equally by geographical distance and by environmental and cultural variables, which we now consider to be candidates for causing and maintaining variation within a language. Further, due to the highly comparable results of our and traditional dialect divisions the applicability of population genetic framework to dialect studies is suggested.

References:

Bowern, C. (2012). The riddle of Tasmanian languages. *Proceedings of the Royal Society B, 279, 4590-4595.*

Chambers, J. K., & Trudgill, P. (1998). *Dialectology*.2nd edition. Cambridge: Cambridge University Press.

Croft, W. (2000). *Explaining language change: an evolutionary approach*. Harlow: Pearson Education Limited.

Kettunen, L. (1940). *Suomen murteet. 3 A, Murrekartasto*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Pagel, M. (2009). Human language as a culturally transmitted replicator. *Nature Reviews Genetics 10, 405-415.*

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics, 155, 945-959.*

Reesink, G., Singer, R., & Dunn, M. (2009). Explaining the Linguistic Diversity of Sahul Using Population Models. *Plos Biology, 7(11), e1000241.*