Spatio-social separation and linguistic complexity in Daghestan

It has been suggested that linguistic separation of speech communities may influence the complexity of their languages (Trudgill 2011, Nichols 2020). However, investigating the supposed correlation is hindered by methodological issues. Previous studies (Baechler 2017) present metrics dependent both on the type of landscape data available and the quality of the linguistic description. The degree of complexity is also influenced by inheritance. One way to overcome the issues is, respectively, (i) using absolute values as a proxy for the degree of isolation, (ii) using metrics based on corpora or dictionaries rather than on grammatical descriptions, and (iii) comparing genealogically related languages. In this study, several statistical experiments are carried out to probe for a correlation between the degree of separation of the languages of the East Caucasian family and their relative morphological and phonological complexity.

As a proxy for spatial separation, we used (the medians of) the altitudes of the villages where each of the languages is spoken: assumedly, the higher in the mountains, the more separated the language. To estimate morphological complexity, we analyzed distributions of the *type-token ratio* (Kettunen 2014) in random samples from non-annotated narrative corpora and translation of the Gospel of Luke. No correlation between morphological complexity and altitude has been found (Figure 1, Figure 2).

We then applied the same methodology to measuring phonological complexity. The *type-phoneme ratio* was obtained from the comparative dictionary (Kibrik, Kodzasov 1990). As Figure 3 with languages ordered by the median shows, the genealogical factor may be significant: Lezgic gravitate towards the right (though with a strong branch-internal variation), Tsezic towards the left, and Andic are intermediate (as confirmed by pairwise Mann-Whitney tests applied to medians). In other words, Lezgic may be more complex and Tsezic less complex not because of their relative attitude but due to the difference in inherited complexity.

To avoid this problem, we checked correlations between phonological complexity and altitude for the languages within Lezgic, Andic and Tsezic branches. As can be seen from Figures 4-6, the results of the experiment are different for all three branches. Lezgic languages seem to conform to the theoretical expectations; after excluding Budugh and Kryz, which are heavily influenced by Azerbaijani, the correlation between altitude and complexity becomes significant (p-value 0.006). However, Andic languages show no correlation between the variables, while for the Tsezic languages, one even sees a negative trend (though no significant correlation has been found).

We conclude that the hypothesis on the relation between complexity of the language and the degree of separation of its speakers, suggested in numerous case studies and in cross-linguistic comparison, is not supported on the local level, at least if one uses altitude as a proxy for separation and relies on corpus or lexicon frequency metrics instead of structural complexity (as in Nichols 2020). If such correlation exists, it may be weaker than the signal from inherited complexity.

Baechler, R. (2017). Complexity, Isolation, and Language Change. Zeitschrift für Dialektologie und Linguistik, 84(2-3), 178-201.

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages?. Journal of Quantitative Linguistics, 21(3), 223-245.

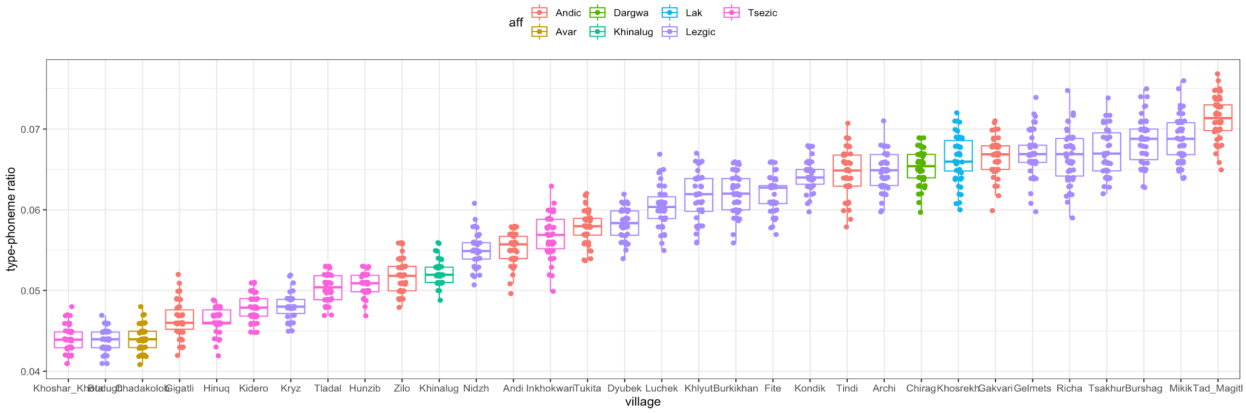
Kibrik, Aleksandr & Sandro Kodzasov. (1990). Sopostavitel′noe izučenie dagestanskix jazykov. Imja. Fonetika. Moscow: MGU.

Nichols, J. (2020). Canonical complexity. In P. Arkadiev & F. Gardani (Eds.), The complexities of morphology (pp. 163–192). Oxford University Press.

Trudgill, Peter. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press.

Figures 1-2.

Figure 3. Type-phoneme ratio in the EC branches



Figures 4-6.