Measuring cross-linguistic distances

Workshop of the 59th Annual Meeting of the Societas Linguistica Europaea

(26 – 29 August 2026, Osnabrück, Germany)

Ian Joo

Otaru University of Commerce

Keywords: Quantitative typology; Linguistic distance; Comparative linguistics; Geolinguistics; Dialectometry

To what degree a given variety of human language is similar to or different from another variety has been a topic of great interest, not only to linguists but also to the general public. With the advent of big data and computational tools in language science, we can now address this perennial question in ways that were not possible or realistic before. In this proposed workshop, we will venture into the many dimensions of **cross-linguistic distance**, the quantifiable degree of dissimilarity between multiple lects ("language" or "dialect"), including:

- **Geographical distance.** Different varieties of the human language are used by human groups residing in different geographical spaces. As such, despite the intangible nature of language, most lects are definable as belonging to a specific spatial range, at least within a given time frame. It is thus possible to measure the geographical distance between two lects, not only for its own sake but also to compare it to other types of distances.

  The most simplistic way of measuring a cross-linguistic geographical distance is to define a given lect as a dot on a surface, i.e. in two-dimensional coordinate values, and measure the linear distance between two dots. While convenient, this method has the obvious limitations of (i) limiting the dynamic space of a given lect into a unifocal location; and (ii) ignoring all the geographical barriers between two spaces that make human transport more difficult and non-linear, such as elevated lands, bodies of water, and political borders. To overcome this simplicity, more sophisticated measures of geographical distance are needed, taking into consideration not only the physical distances between two ends but also the human capacity and willingness to travel those distances.

- **Genealogical distance.** By using the traditional historical comparative method, the lects of a given language family can be hierarchically classified in terms of when each lect has diverged from which other lect. This hierarchy, often visualized as a tree, make it possible the measure how genealogically close a given lect is to its intrafamilial lect, analogous to measuring how distant a relative is from a human individual.

Such genealogical distance can be quantified either as a count or continuous variable, depending on whether the estimated time of familial divergence is in relative or absolute scale. If the temporal data of the divergences are only relatively ordered, i.e. if we only know that lect A has split off from lect B before lect C has from lect B but not when in history these two splits have happened, we can only quantify the genealogical distances in terms of the number of divergences, or "layers". If we do know when they happened, however, with the aid of non-linguistic data from genetics, archaeology, and written history, we are able to quantify the genealogical distance as a continuous variable, weighing each divergence differently in terms of how far back in time it has happened. Such temporal data, however, are often unreliable, controversial, or unavailable.

- **Typological distance.** Human groups using different lects can come into contact and develop conduct-induced similarities in their lects. Such convergence can occur in all linguistic domains, namely phonology, syntax, and semantics.

  - **Phonological distance**. Lects can be compared in terms of which forms they employ to represent concepts. As defined by Joo and Hsu (accepted), a phonological distance can be **intersequential**, i.e. between two phonological sequences, such as between English word *man* /mæn/ and its German cognate *Mann* /man/ (Heeringa 2004; Do and Lai 2021, e.g.). It can also be **interstructural**, i.e. between two phonological structures, such as between the English phonemic inventory and the German phonemic inventory (Avram 1964; Eden 2018; Joo and Hsu accepted, e.g.).
  - **Syntactic distance**. The increasing number of cross-linguistic databases encoding morphosyntactic parameters, such as the World Atlas of Language Structures (WALS, Dryer and Haspelmath 2013) and Grambank (Skirgård et al. 2023), allow us to quantify the degree of cross-linguistic grammatical similarity. The parametrical values may be categorical (e.g. the six possible values of *What is this lect's basic order of subject, object, and verb?*) or binary (e.g. the logical values of *Is this lect verb-final?*).

    While the parametrical format is a convenient way to retrieve the values from descriptive sources and encode them in a comparable manner, it has some non-negligible limitations. First, the choice of the parameters is inevitably biased by what aspects of grammar the designer of the database deems important. The parameters may also bear different weights, some being more typologically significant than others, which is difficult to quantify and code into the database. Moreover, the parameters are also almost always non-independent, such as the Greenbergian universal of the correlation between word order and adpositions, which must always be controlled for when using the parametrical values for distance measurement. Innovative alternatives are in need in order to overcome and complement such limitations.
  - **Semantic distance.** Semantic properties, due to their qualitative and sociocultural nature, are arguably more difficult for quantitative comparison than syntax and phonology. Recent advances in computational semantics are helping us to overcome this difficulty. Natural Language Processing techniques like Word2vec enable the quantification of the semantic values of lexical items based on collocational data retrieved from large corpora. Colexification databases, such as the Database of Cross-Linguistic Colexifications (CLICS, Rzymski et al. 2020), also allow us to

measure the distance between two lects in terms of how many colexification patterns they share. As computational semantics is a rapidly growing field, many possibilities await us to measure semantic distances in various innovative ways.

The presentations of this workshop will provide new methodologies in measuring any form of cross-linguistic distances. The distance may be at any scale, between genealogically distinct lects ("languages") or closely related ones ("dialects"), and in any domain, such as phonology, syntax, lexicon, genealogy, geography, chronology, or demographics. By shedding light on the various possible ways of measuring different dimensions of cross-linguistic distances, our workshop will inspire novel ideas for the everlasting quest of measuring one form of language to another.

References

Avram, Andrei (1964). "Sur la typologie phonologique quantitative [On the quantiative phonological typology]". In: *Revue roumaine de Linguistique* IX, pp. 131–134.

Do, Youngah and Ryan Ka Yau Lai (2021). "Accounting for lexical tones when modeling phonological distance". In: *Language* 97.1, e39–e67.

Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online (v2020.4).* Zenodo. URL: doi.org/10.5281/zenodo.13950591.

Eden, S. Elizabeth (2018). "Measuring phonological distance between languages". PhD thesis. University College London.

Heeringa, Wilbert Jan (2004). "Measuring dialect pronunciation differences using Levenshtein distance". PhD thesis. University of Groningen.

Joo, Ian and Yu-Yin Hsu (accepted). "Phonological distances between Eurasian lects measured via Phonotacticon 1.0 reveal areal patterns". In: *Linguistics*.

Rzymski, Christoph et al. (2020). "The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies". In: *Scientific Data* 7.1. DOI: 10.1038/s41597-019-0341-x.

Skirgård, Hedvig et al. (2023). "Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss". In: *Science Advances* 9.16, eadg6175. DOI: 10.1126/sciadv.adg6175.