Margaret J. Blake & Harald Hammarström

Swadesh, meet Levenshtein: A Quantitative Method from Dialectology Applied to Phonetic Glottochronological Data

Glottochronology and lexicostatistics, despite their flaws, do provide useful information about language relationships and lexical change. However, it is crucial that lexical change not be viewed as equivalent to language change, and that other complementary tools and techniques be developed to give a better picture of overall language change, which also includes morphosyntactic change, phonetic/phonological change, etc. Additionally, the assumption that language change occurs at a constant rate for all languages is one of the major flaws in Swadesh's original approach to glottochronology; simple inspection of, for example, modern Icelandic and modern Danish against Old Norse suggests quite the opposite (see Trudgill 2007 for exploration of the causes of this difference). To that end, the authors propose a novel technique for quantifying phonetic change, using the Levenshtein string distance algorithm to calculate degree of phonetic change through time upon a cognate subset of the Swadesh-200 list for the insular and peninsular Scandinavian languages, as well as their parent language, Old Norse.

The Levenshtein distance algorithm calculates the difference between two strings by calculating the minimum number of additions, deletions, and substitutions necessary to transform one string into another, normalized by dividing by the length of the longer string. In linguistic applications, the algorithm is typically applied to phonetic transcriptions of words rather than the strings themselves. In addition, modifications have been made to the Levenshtein algorithm to make it more reflective of language change, such as calculating only vowel-to-vowel and consonant-to-consonant changes (Heeringa & Gooskens 2003), as well as the more conservative Almeida-Braun variation, which considers a greater number of phonetic features and incorporates each set of relevant features as an axis in a multidimensional distance calculation (Heeringa & Braun 2003). Levenshtein distance has been shown to correspond significantly with perceptual evaluations of dialect distance (Gooskens & Heeringa 2004), and thereby may yield useful results for other areas of language study where difference and/or change are involved.

Although there are numerous criticisms of the validity of the Swadesh list, it does have many strengths to offer (see e.g. Renfrew et al., eds. 2000 for discussion of the relevant issues): it is pre-existing, readily available for many languages, and contains enough cognates for closely-related languages to provide a statistically significant data set (Kessler 2001). Criticisms have also been leveled against the Levenshtein algorithm (Heggarty 2006), but both methods can yield meaningful results when applied in a conservative, linguistically-informed manner (ibid., Heggarty 2000). To that end, the Scandinavian languages provide an ideal data set, as they are long-documented, well-studied, and have a long history of the written word from which to draw data (as well as extensive reconstruction of former incarnations of their phonetic systems). The accuracy of our analysis will be judged against non-glottochronological measures, such as socio-historical data (König and van der Auwera 1994), data concerning intermediate forms between the parent and modern varieties, and mutual intelligibility judgments (Delsing & Åkesson 2005).

References

Delsing, L.-O. & Åkesson, K.L. (2005) Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska. Nordic Ministerråd, Copenhagen.

Gooskens, C. & Heeringa, W. (2004) Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. Language Variation and Change, **16**, 189-207.

Heeringa, W. & Braun, A. (2003) The Use of the Almeida-Braun System in the Measurement of Dutch Dialect Distances. Computers and the Humanities, **37**, 257-271.

Heeringa, W. & Gooskens, C. (2003) Norwegian Dialects Examined Perceptually and Acoustically. Computers and the Humanities, **37**, 293-315.

Heggarty, P. (2000) Quantifying change over time in phonetics. In Time Depth in Historical Linguistics, (Renfrew, C., McMahon, A. & Trask, L., eds.) The McDonald Institute for Archaeological Research, Cambridge, pp. 531-562.

Heggarty, P. (2006) Interdisciplinary Indiscipline? Can Phylogenetic Methods Meaningfully Be Applied to Language Data – and to Dating Language? In Phylogenetic Methods and the Prehistory of Languages, (Forster, P. & Renfrew, C., eds.) McDonald Institute for Archaeological Research, Cambridge, pp. 183-194.

Kessler, B. (2001) The Significance of Word Lists. Center for the Study of Language and Information, Stanford, CA.

König, E. & van der Auwera, J. (Eds.) (1994) *The Germanic Languages*. Routledge, London.

Renfrew, C., McMahon, A. & Trask, L. (Eds.) (2000) Time Depth in Historical Linguistics The McDonald Institute for Archaeological Research, Cambridge.

Trudgill, P. (2007) Sociolinguistic dialect typology: contact and isolation in Nordic dialects. In Nordisk dialektologi og sociolingvistik, (Arboe, T., ed.) Peter Skautrup Centret for Jysk Dialekforskning, Århus, 33-53.