# DATABASES FOR THE NEW HISTORICAL LINGUISTICS

## AN EXPERT-LED DATABASE OF COGNACY IN BASIC VOCABULARY ACROSS INDO-EUROPEAN

Cormac Anderson & Paul Heggarty — DLCE, Max Planck Institute for the Science of Human History, Jena

cormacanderson@gmail.com — Paul.Heggarty@gmail.com

A 'new historical linguistics' seems to be coming of age. Traditional comparative/historical language data are now being reanalysed using a set of highly sophisticated computational analysis tools. Most were originally devised for the biological sciences, but languages too 'descend with modification' from a common ancestor. So where appropriately adapted, these techniques are breathing new life into the long-standing debates with which linguistics first began, not least the Indo-European question. Bayesian phylogenetics, in particular, has catapulted the origins and divergence histories of major language families back into leading journals such as *Language* (Chang *et al.* 2015) and even *Science* (Gray *et al.* 2009 and, controversially, Bouckaert *et al.* 2012).

Linguistics has not kept pace, however, in devising accessible and reliable language database resources, (re)structured and compiled to new policies to ensure that we make the most out of the new quantitative toolkit — and avoid certain potential pitfalls. Many Swadesh lists, for example, allow multiple (near-)synonymous lexemes for a single meaning. The linguistic justification is ill thought out, for what is only a tiny *sample* data-set in any case, and where the real imperative is cross-linguistic consistency. Under normal Bayesian phylogenetic models, excessive synonymy in practice only introduces new distortions, far worse than those they were intended to remedy.

This paper reports on a project — www.cobl.info — to develop a new model database structure for encoding Cognacy relationships in Basic Lexicon across any given language family, designed to maximise utility for qualitative as well as quantitative research purposes.

We devise a new reference list of 200 comparison meanings, combined out of the Swadesh 100, Swadesh 200 and Leipzig-Jakarta 100 lists (Tadmor 2009), but freed of those meanings found most open to serious coding inconsistencies, especially in languages of radically different structural types or spoken in different cultures and contexts.

Our first implementation is to Indo-European. Our new database follows in part the relational structure of the existing IELex database used by many recent publications. The actual data in IELex, however, originated mostly in Dyen, Kruskal & Black (1992), long identified by many linguists as highly unreliable and seriously inconsistent, as well as incomplete and insufficient for many other desired research applications. The IELex website (ielex.mpi.nl) also lacked much-needed functionality, especially for efficient data entry and consistency.

Corresponding lexeme lists for each language are compiled by a consortium of specialist authors by language or sub-branch of Indo-European, then cross-checked and reconciled with a second expert opinion. All experts work to a new and very explicit set of lexeme selection guidelines and precisely (re)defined target meanings. Entries are given in native orthography, Roman transliteration where necessary, phonemic and IPA phonetic (major allophone) transcriptions, all searchable and linked to published sources. All lexemes entered are assigned into cognate sets, explicitly defined by a shared root (or 'loanword event', where applicable). Ultimately these will be broken down further into cognate subsets, which share their root but differ in other morphemes (e.g. presence/absence of s-mobile, derivation from different case or tense forms, etc.), to allow more precise analyses, both quantitative and qualitative.

## References

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S.J., Alekseyenko, A.V., Drummond, A.J., Gray, R.D., Suchard, M.A., & Atkinson, Q.D. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097): p.957–960. http://dx.doi.org/10.1126/science.1219669

Chang, W., Cathcart, C., Hall, D., & Garrett, A. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1): p.194–244. http://dx.doi.org/10.1353/lan.2015.0005

Dyen, I., Kruskal, J.B., & Black, P. 1992. *An Indoeuropean Classification: A Lexicostatistical Experiment*. Philadelphia: American Philosophical Society. www.wordgumbo.com/ie/cmp/iedata.txt

Gray, R.D., Drummond, A.J., & Greenhill, S.J. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913): p.479. http://dx.doi.org/10.1126/science.1166858

Tadmor, U. 2009. Loanwords in the world's languages: findings and results. In M. Haspelmath & U. Tadmor (eds) *Loanwords in the World's Languages: A Comparative Handbook*, 55–75. Berlin: de Gruyter. www.degruyter.com/viewbooktoc/product/41172