ON DIGITALISATION OF THE COMPARATIVE METHOD OF RECONSTRUCTION IN THE INDO-EUROPEAN FAMILY OF LANGUAGES IN PIE LEXICON

Mans Hulden, Fedu Kotiranta, Jouna Pyysalo, and Aleksi Sahala

Proto-Indo-European Lexicon (PIE Lexicon) is the generative etymological dictionary of Indo-European languages. The reconstruction is obtained by means of the comparative method, whose output equals the Indo-European (IE) data. The IE sound laws leading from PIE to IE can be coded using finite-state transducers (FST). The foma finite-state compiler by Mans Hulden (2009) has been chosen for this purpose. At this point PIE Lexicon generates data of some 120 IE languages with an accuracy rate of over 99% and is therefore the first dictionary in the world capable of generating (predicting) its data entries.

1 **The IE language family, PIE, and the generation of the data**

The IE language family, with three billion native speakers, is the largest in the world, comprising some six hundred languages including one hundred and fifty archaic, partly extinct ones, the oldest of which were attested more than three thousand years ago. The comparative method of reconstruction is the gold standard in the postulation of the proto-language, PIE. The method is based on comparison of IE morphemes and the postulation of PIE is determined by the measurable features of the data. By coding the sound laws, the words of the IE languages can be generated with sound law (foma) scripts. The most ancient IE sound laws, critically chosen and revised in Pyysalo 2013, have been formulated in foma at http://pielexicon.hum.helsinki.fi.

2 **On the coding of sound laws and generation of data**

The digitalization of the IE sound law system starts with coding the individual sound laws, which are chronologically arranged as foma scripts. The foma equivalents of the IE sound laws are expressed together with their environments, i.e. they assume the format A -> B || C _ D (‘A changes into B in environment C_D’). Each rule is tested by the compiler and placed in the foma (sound law) scripts. Each IE language is equipped with a foma script. All archaic sound laws (i.e. ones applying to at least two subgroups) have already been coded. After the rules have been arranged into scripts, their internal and external consistency is tested.

By now PIE Lexicon consists of some 120 foma scripts and generates 50000 phonemes, about 200 of which are erroneous, i.e. the general accuracy rate is over 99%. Since the choice of the material is random, the theory presented in Pyysalo 2013 is valid, sound and complete. The remaining errors represent open research problems. These can divided into:
(a) The PIE accent/tone problem, not treated in Pyysalo 2013, accounts for almost half of the errors, making it the fundamental remaining problem of IE linguistics.
(b) A dozen minor sound law problems related to individual subgroups, languages and/or dialects are also open or only partially solved.

3 **Concluding observations, remarks, and an outline of the project’s future**

The high success rate of PIE Lexicon in the automatic generation of the IE data shows that managing historical sound laws by applying finite-state transducers provides a rigorous formal calculus for mapping cognates from PIE to the daughter languages. The success rate in the generation of the IE data is explained by the following factors:

(a) In the comparative method the input (the PIE reconstruction) is not hypothetical, but a sum of the measurable features of the data and its comparison, the logical equivalent of the data.
(b) All consistent sound law proposals, perfected if necessary, of two centuries of research in IE linguistics have been chosen for PIE Lexicon.

The sound laws of the most archaic languages have now been coded, and the next, more abstract phase, coding the IE language family tree on the basis of the common sound laws, has begun. For this purpose a foma rule bank, consisting of some 800 rules, has been coded. This allows us to place the rules in a template in which identical rules are placed on the same row. This first sound law-based IE language family tree will be published once ready.

Once the main features of the IE language family tree have been coded, the preconditions for the digitalization of the decision method of IE etymology have been created. This feature, originally outlined by August Schleicher (1852: iv-v), has a counterpart in language technology: The (foma) scripts can be run in the reverse direction ("apply up") to generate all possible PIE prototypes of the IE words. This requires the addition of tailored, language-family-specific phonological constraints, after which an intersection function seeking identical PIE prototypes between the disjunctions can be coded. This allows us to test every etymology ever proposed and to mechanically identify all potential etymologies.

Taken together the coding of the IE sound law system, the IE language family tree, and the decision method of IE etymology mean that the critical components of the comparative method of reconstruction itself have been digitized. Once achieved, PIE Lexicon will be able to manage comparative IE linguistics digitally for the first time in history. Thus IE linguistics will be at the frontline of digital humanities, equipped with a next-generation theory embedded in the methodic framework of natural sciences.

References

Kenneth E. Beesley & Lauri Karttunen. 2003. Finite State Morphology. Studies in computational
linguistics 3. Center for the Study of Language and Information, Stanford.
N. E. Collinge. 1985. The Laws of Indo-European. Benjamins, Amsterdam.
N. E. Collinge. 1995. Further Laws of Indo-European. In: On Languages and Language: The Presidental Adresses of the 1991 Meeting of the Societas Linguistica Europaea. ed, Werner Winter. B. Trends in Linguistics. Studies and Monographs, 78. Mouton, Berlin: 27-52.
N. E. Collinge. 1999. The Laws of Indo-European: The State of Art. Journal of Indo-European Studies, 27:355-377.
Mans Hulden. 2009. Finite-State Machine Construction Methods and Algorithms for Phonology and Morphology, PhD Thesis. University of Arizona.
Jouna Pyysalo. 2013. System PIE: The Primary Phoneme Inventory and Sound Law System for Proto-Indo-European. Publications of the Institute for Asian and African Studies 15. Unigrafia Oy, Helsinki.
August Schleicher. 1852. Die Formenlehre der kirchenslavischen Sprache, erklärend und vergleichend dargestellt. H. B. König, Bonn.