Proto-Indo-European Lexicon: The Generative Etymological Dictionary of Indo-European Languages

Jouna Pyysalo
Department of Modern Languages
University of Helsinki
jouna.pyysalo@helsinki.fi

Abstract

Proto-Indo-European Lexicon (PIE Lexicon) is the generative etymological dictionary of Indo-European languages. The reconstruction of Proto-Indo-European (PIE) is obtained by applying the comparative method, the output of which equals the Indo-European (IE) data. Due to this the Indo-European sound laws leading from PIE to IE, revised in Pyysalo 2013, can be coded using Finite-State Transducers (FST). For this purpose the *foma* finite-state compiler by Mans Hulden (2009) has been chosen in PIE Lexicon. At this point PIE Lexicon generates data of some 120 Indo-European languages with an accuracy rate of over 99% and is therefore the first dictionary in the world capable of generating (predicting) its data entries by means of digitized sound laws.

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

The Indo-European language family, with some three billion native speakers, is the largest in the world. The family comprises 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 IE language family is divided into a dozen subgroups (e.g. Germanic, Italic, and Anatolian), each of which was given a specific character by the common vocabulary of the subgroup and the set of common sound changes distinguishing it from the other subgroups.

The comparative method of reconstruction is the gold standard in the postulation of the proto-language, PIE. The method is based on comparison of originally identical IE morphemes. The postulation of the proto-language is exclusively determined by the measurable features of the data. Consequently the reconstruction of a data segment is the equivalent of the data set it was inferred from. By coding the sound laws of each Indo-European language the words of the languages can be generated by sound law *(foma)* scripts detailing the changes applying to the languages.

In order to generate the data the most ancient Indo-European sound laws,[1] critically chosen and revised in Pyysalo 2013, have been formulated in the *foma* finite-state compiler and arranged chronologically.[2] By now also later sound laws have been added and the scripts have been implemented in PIE Lexicon at http://pielexicon.hum.helsinki.fi, generating the data in a manner explicated in this paper.

2 On the coding of sound laws and generation of data in PIE Lexicon

The digitization of the IE sound law system starts with the coding of the individual sound laws as the *foma* rules. These are then arranged in a chronological order into *foma* scripts of the languages.

2.1 Coding of the IE sound laws

The IE sound laws are of the implicational form PIE $*x \rightarrow \mathrm{IE}_z\ y$ ('the PIE sound $*x$ turns into sound y in the Indo-European language z').

1 For the core of the traditional Indo-European sound law system, see Collinge 1985, 1995, and 1999.

2 For a state-of-the-art formulation of the finite-state technology, see Beesley & Karttunen 2003.

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 259–262,
Gothenburg, Sweden, 23-24 May 2017.

In *foma* the equivalents of the IE sound laws are expressed together with the environments in which they are operational. In practice the *foma* rules assume the format *A -> B || C _ D* ('*A* changes into *B* in environment *C_D*'). Corresponding to the principle of regularity of sound changes in comparative Indo-European linguistics, the *replacement* of *A* with *B* is *obligatory* in *foma*, i.e. it always occurs in all instances of the environment *C_D*.

For example, the revised Brugmann's law adds the glottal *H* to the environment: PIE **oHCV* → PIIr. *āCV* (Pyysalo 2013: 121-125). This is written in *foma* as:

o -> ā || _ Glottal Consonant Vowel ;

Each *foma* rule coded is tested by the compiler and then placed in the *foma* (sound law) scripts.

2.2 Coding of the IE sound law *foma* scripts

After the digitization of the IE sound laws, each IE language is equipped with a *foma script* consisting of chronologically ordered *foma* rules. The data of PIE Lexicon has required all major archaic sound laws (i.e. ones applying to at least two subgroups) to be coded. The *foma* scripts of some 120 languages or dialects have already been implemented on the PIE Lexicon site, and are available in the control bar at the bottom of the site. E.g. the Hittite sound law script can be opened by first clicking *Select rule set*, then *Hitt.*, and finally *Show rules*, opening the file: http://pielexicon.hum.helsinki.fi/?showrule=24.

Once the *foma* rules have been arranged into scripts, the consistency of the rules is tested both internally (with regard to the other rules of the script) and externally (with regard to the portion of the data the script is capable of generating).

2.3 On the generation of the IE data in PIE Lexicon

By March 2017 PIE Lexicon consists of some 120 *foma* scripts and generates some 35000 phonemes, some 175 of which are erroneous, i.e. the general accuracy rate is 99,5%.

Since the choice of the material is random, the digitized glottal fricative theory (GFT) of Pyysalo 2013, now tested and digitally published in PIE Lexicon, is valid, i.e. sound and complete.

The remaining errors – shown in red in PIE Lexicon – represent open research problems, for which *foma* rules cannot be specified, because the sound laws remain unknown. The errors can divided into two subsets:

(a) The PIE accent/tone problem, not treated in Pyysalo 2013, accounts for almost half of the errors, making it the fundamental problem of Indo-European linguistics at this point.

(b) A dozen minor sound law problems related to individual subgroups, languages and/or dialects are also open or only partially solved. In order to solve these problems the project has opened a journal, *Proto-Indo-European Linguistics*, at http://pielinguistics.org.

2.4 On the explicit *foma* proof chains

PIE Lexicon has added a special feature to the basic version of *foma*, which makes the proof chains of the generation of the data explicit for the PIE Lexicon users and editors. Consequently the entire generation of data can be immediately verified. Clicking the reconstruction on the left side of the individual IE data entries reveals the respective *foma* proof chain, and all *foma* chains can be opened simultaneously by clicking the *Chains* button at the bottom of the site.

A proof chain explicitly detailing all sound laws applied and their mutual order in the generation is thus attached to every IE form. This makes the proofs fully explicit and to a degree confirms the PIE reconstructions serving as the starting point of the generation of the data.

3 Concluding observations, remarks, and an outline of the project's future development

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 and automatically evaluating the consistency of such a system.

3.1 On the background of the successful generation of IE data

The success rate achieved in the generation of the Indo-European data is also 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, hence the logical equivalent of the data. This in turn results in the respective success in its generation.

(b) The GFT comprises all correct sound law proposals, perfected if necessary, of two centuries of research in IE linguistics. Consequently, the sound law system stands on the shoulders of giants.

3.2 The coding of the IE language family tree

By the spring 2017 the coding of the ancient sound laws of the most archaic languages has been completed. Only late, unproblematic sound laws need be added when appearing in the new data to be published in PIE Lexicon.

Consequently the next, more abstract phase, the coding of the IE language family tree on the basis of the common sound laws, has already begun. For this purpose a *foma* rule bank, consisting of some 800 rules, has been coded. Instead of full sound law scripts the operating system uses the names of the rules which call the rules from the bank. This allows us to place the rules in an excel template in which identical rules are placed on the same row.

By means of this procedure the first IE language family tree based purely on the common sound laws is already being coded and will be published in PIE Lexicon once ready.

3.3 The coding of the decision method of Indo-European etymology

Once the main features of the IE language family tree have been coded, the preconditions for the digitization of the decision method of IE etymology have been created. This feature, originally outlined by August Schleicher (1852: iv-v) [3] has a counterpart in language technology. Once the sound law (*foma*) scripts are ready it is possible to run them in reverse direction, i.e. starting from the bottom ("apply up"). The technology already exists and once implemented it will generate all possible PIE prototypes of an Indo-European word. The coding only requires the addition of tailored, language-family specific phonological constraints in order to eliminate potential infinite chains caused by historically lost phonemes.

Once the disjunctions of possible PIE prototypes of all IE words have been digitally generated it is not complicated to code an intersection function that seeks identical PIE prototypes between the disjunctions and proposes an etymology when there is an intersection of the disjunctions of two languages. If the identity is semantically feasible, the computer has found a PIE etymology.

After the coding of this feature it is in principle possible to test every etymology proposed during the history of Indo-European linguistics and to mechanically identify all potential etymologies, which in turn may reveal identities not noticed by the scholars, and thus revitalize the research.

3.4 Conclusion: Coding of the comparative method of reconstruction in Indo-European linguistics

Taken together the coding of the IE sound law system (§3.1), the IE language family tree (§3.2), and the decision method of IE etymology (§3.3) mean that the critical components of the comparative method of reconstruction itself have been digitized.

Once achieved, Operating System (OS) PIE Lexicon will be able to manage the comparative IE linguistics digitally for the first time in history. Thus in the 21st century, Indo-European linguistics will be in the frontline of digital humanities, equipped with a next-generation theory embedded in the methodic framework of natural sciences.

Credits

This paper summarizes a joint collaboration of Dr. Jouna Pyysalo (IE languages and their reconstruction), Prof. Mans Hulden (language technology and *foma*) and Mr. Fedu Kotiranta (web design and node.js), the coding of PIE Lexicon that started in the autumn 2014.

References

Kenneth E. Beesley & Lauri Karttunen. 2003. *Finite State Morphology*. Studies in computational

3 For a precise formulation of the decision method with a data example, see Pyysalo 2013: 475-476.

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*. BTrends 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.