Verifying the Consistency of the Digitized Indo-European Sound Law System Generating the Data of the 120 Most Archaic Languages from Proto-Indo-European

Jouna Pyysalo, Aleksi Sahala, and Mans Hulden

ABSTRACT: Using state-of-the-art finite-state technology (FST) we automatically generate data of the some 120 most archaic Indo-European (IE) languages from reconstructed Proto-Indo-European (PIE) by means of digitized sound laws.
The accuracy rate of the automatic generation of the data exceeds 99%, which also applies in the generation of new data that were not observed when the rules representing the sound laws were originally compiled.
After testing and verifying the consistency of the sound law system with regard to the IE data and the PIE reconstruction, we report the following results:
a) The consistency of the digitized sound law system generating the data of the 120 most archaic Indo-European languages from Proto-Indo-European is verifiable.
b) The primary objective of Indo-European linguistics, a reconstruction theory of PIE in essence equivalent to the IE data (except for a limited set of open research problems), has been provably achieved.
The results are fully explicit, repeatable, and verifiable.

1. On the digitalization of Indo-European sound laws with finite-state technology (FST)

1.1 Sir William JONES' (1788) groundbreaking announcement of a genetic relationship between European languages including Greek and Latin, Sanskrit, and other language groups and the existence of a common ancestor of these Indo-European (IE) languages marks the birth of modern comparative linguistics. Soon after, the pioneers Rasmus RASK, Franz BOPP, and others confirmed the existence of systematic correspondences between the 'letters' (phonemes) of the IE languages, thus verifying Sir William's initial assessment: the Indo-European languages are indeed genetically related and descended from a common source, now known as Proto-Indo-European (PIE).

1.2 By the 1860s, August SCHLEICHER had presented the basic pillars of the comparative method of reconstruction:

(a) The phonemes of languages change regularly (without exception). The principle of *regularity of sound change* won the day through August LESKIEN, whose formulation *Ausnahmslosigkeit der Lautgesetze* became the slogan of the famous Neogrammarian movement in the 1870s.

(b) The sound changes are described and the proto-language reconstructed by means of the *comparative method*. The method identifies and compares cognates reflecting the same underlying forms, reconstructs (postulates) their proto-phonemes, and accounts for the changed (or lost) phonemes by means of sound laws describing their historical developments.

1.3 After SCHLEICHER had introduced the concept of reconstruction, its justification, the *principle of postulation*, was explicitly formulated by August FICK (1870-1871) in his motto:

Durch zweier Zeugen Mund wird alle Wahrheit kund. (FICK's rule)

Accordingly, as later on explained by Holger PEDERSEN, every feature postulated into the proto-language must be independently confirmed by at least two pieces of data belonging to different subgroups.[1] Furthermore, a correctly postulated reconstruction is exclusively based on measurable features of the data and the comparative inferences are justified by the principle of postulation. In an ideal case the PIE reconstruction and the IE data are logically equivalent, i.e. the PIE reconstruction is implied by the IE data and the IE data result from the PIE reconstruction. In other words, if a sound law system generates the entire (inherited) data without errors, it is valid, i.e. complete and sound.[2]

1.4 The comparative method is an empirical science and was therefore quickly digitalized as language technology emerged. By now the state-of-the-art *finite-state compilers* (or automata) contain tools to express, test, and evaluate intuitive (non-formal) sound law systems presented by linguists.[3] A finite-state compiler produces transducers (finite-state machines) from the digitized rules, formally expressing the respective sound laws.[4] The transducers are subsequently tested to confirm the internal and external consistency of the rules representing the sound laws. Finite-state transducers have long since been recognized as being suitable computational models for treating sequential derivations of phonological change. The fact that phonological rules of the type introduced in *The Sound Pattern of English* (CHOMSKY and HALLE 1968) could be expressed as finite-state transducers was first noted by C. Douglas JOHNSON (1972). The same observation was later elaborated on and refined (KAPLAN and KAY 1981) to produce a complete

[1] See PEDERSEN (1962:274): "If a word [or an object of any level] is found in the two branches, then it was also to be found in the original language which divided into these branches." For a more general detailed discussion, see PYYSALO (2013 §1.5.5).

[2] The comparative method is thus a decision method in the Hilbertian sense, i.e. an algorithm distinguishing between right and wrong for each step of inference.

[3] For the state-of-the-art of finite-state technology, see BEESLEY & KARTTUNEN 2003.

[4] A finite-state compiler is a *Turing machine* based on a digitized *predicate calculus*, operated by the code reader (compiler) of a computer.

computational model for sound changes. The possibilities to use finite-state calculus to explicitly describe sound change through the combination of finite-state transducers have since then been greatly expanded (see KARTTUNEN 1995, KEMPE and KARTTUNEN 1996, YLI-JYRÄ 2008, HULDEN 2009b, 2009c), and several essentially equivalent finite-state compilers are currently available. In the research reported in this paper we use Mans HULDEN's *foma* (2009a), an open-source compiler particularly suited for modeling alternation and derivation processes in phonology, while noting that naturally many other devices, equally neutral, can be used instead.

1.5 As the target for the testing we have chosen Jouna PYYSALO's (2013) recent, revised reconstruction of Proto-Indo-European. The stated goal of this *glottal fricative theory* (GFT), based upon Oswald SZEMERÉNYI's (1967, 1970) monolaryngealism, is to present a valid synthesis of the research in IE linguistics by picking among all IE sound laws proposed during the history of the field those which lead to a single consistent system of sound laws generating the data.[5] The GFT is currently our best chance of success in the automatic generation of the IE data due to its comprehensive approach and because the digitalized version of the theory has already been successfully implemented in PIE Lexicon at http://pielexicon.hum.helsinki.fi with representative data.[6]

1.6 All correspondences of the research data contain at least one Hittite, Palaic, Cuneiform Luwian, or Hieroglyphic Luwian form in order to maximally contribute to the solution of the etymology of the Old Anatolian languages, which are currently in a critical position for IE linguistics.[7] As the data corresponding to the Old Anatolian words may, however, belong to any IE language, the material is in essence arbitrary. With regard to the research data it should be also noted that:

(a) All twelve established sub-branches of the Indo-European language family (viz. Albanian, Anatolian, Armenian, Baltic, Celtic, Germanic, Greek, Indo-Aryan, Iranian, Italic, Slavic, and Tocharian) and all most archaic languages and/or dialects are included in the data.

(b) All major Indo-European sound laws (i.e. ones applying to more than one subgroup) are included, and later sound laws have also been added as required by the published research data.

[5] For the classical IE sound law system, see COLLINGE 1985, 1995, and 1999.

[6] PIE Lexicon is a digital online application of the GFT into which *foma*, the sound law scripts and the IE data have been implemented for immediate verification of the results.

[7] In 1917 the Czech scholar Bedřich HROZNÝ proved that Hittite belonged to the Indo-European language family. Hittite, as well as the other Old Anatolian languages, preserves a (segmental) 'laryngeal' Hitt. ḫ, which has been lost in all other Indo-European languages.

(c) Although the data are not complete, new material is constantly being added and has not brought forth significant problems, which suggests that they are representative. Hence the sound law system is not susceptible to significant changes in the future except for the addition of the later sound laws expanding the coverage of the system itself.

2. The Indo-European sound laws and their digitalization

2.1 *Coding of individual* IE *sound laws in* foma

The Indo-European sound changes are implications according to which the PIE sound **x* turns into the Indo-European sound **y* in a certain environment. Accordingly the respective *foma* rules assume the format

A -> *B* || *C* _ *D* (‘*A* changes into *B* in environment *C_D*’)[8]

For the sake of illustration, the (unconditioned) change of PIE *o into a vowel IE /a/, common for several subgroups (e.g. Anatolian, Germanic, and Indo-Iranian), is coded in *foma* as follows:

define Ro›a o -> a ; # PIE *o → a | Change of *o into a |[9]

After this phase a finite-state transducer, a device for modeling regular relations between sequences of symbols, is formulated for the written *foma* rule. The transducer reads the input strings, matches these against the input symbols on the transitions, exports the corresponding output strings, and immediately tests and evaluates their correctness in terms of internal consistency.

2.2 *Coding of the sound laws of the* IE *languages as* foma *scripts*

After the digitalization of the individual sound laws, each IE language is equipped with a foma *script* consisting of all of its sound laws as far as they are present in the data. In this phase, which requires correct chronological ordering of the rules, we have coded *foma* scripts for some 120 of the most archaic IE languages and dialects.[10] The current number of actual *foma* rules in the scripts alternates to a degree. For instance Palaic, an extremely conservative Old Anatolian language, currently has 60

[8] The replacement of *A* with *B* is obligatory in *foma*, i.e. it always occurs in all instances of the environment *C_D* (where C and D can be empty in the case of an unconditioned change).

[9] With regard to the formulation, note that we use a version of *foma* that is modified so that in addition to the sound laws in *foma* (e.g. define Ro›a o -> a ;) the same rules are expressed in the conventional notation of IE linguistics (PIE *o → a) and once more in English (‘Change of *o into a’) in order to make the code understandable also for non-experts.

[10] The *foma* (sound law) scripts, as far as they have been coded, are available in the control bar at the bottom of all PIE Lexicon data pages. Thus e.g. the Hittite sound law script can be obtained by first clicking *Select rule set*, then *Hitt.*, and finally *Show rules*, opening the file: http://pielexicon.hum.helsinki.fi/?showrule=24.

rules, whereas Vedic Sanskrit, which has undergone substantially more sound changes, requires 140 rules.[11]

2.3 *Describing the Indo-European language family on the basis of the* foma *rules*

The *foma* scripts define a third level of the IE sound law system, viz. the Indo-European language family itself, through the common sound laws of the languages, especially important ones such as the *centum-satem* split, and the subgroups defined by these. In order to describe the IE family from a bird's-eye view we have already compiled a *rule bank* consisting of all the *foma* rules coded so far, the total figure currently comprising some 800 distinct rules. In the currently ongoing coding phase, the results of which will be published later on, we will illustrate the mutual relations of the IE languages with a self-organizing map, and most likely also with a parallel digital mapping of the language family tree in a traditional sense.

3. On the testing and verifying of the sound laws and the generation of the IE data

3.1 *Testing and evaluating the individual* foma *rules*

We have tested the bulk of the some 800 individual *foma* rules attested in the data for internal consistency (i.e. against the logical syntax of *foma*) by modeling them as transducers and confirmed their consistency. There are no problems or inconsistencies involved.

3.2 *Testing and evaluating the* foma *scripts*

Subsequently we have tested the *foma* scripts with regard to their consistency (i.e. the absence of contradictions in the sets of rules) and external performance (i.e. their capability to generate the data of the IE languages). The scripts do not contain mutually contradicting rules, and there are no problems involved in these.

At this point (early February 2018) PIE Lexicon contains some 10,000 Indo-European words with five phonemes on average. In terms of consistency with the data, of about 50,000 automatically generated phonemes of the data some 250 are erroneous, i.e. the accuracy rate of the *foma* scripts in the generation of the data phonemes exceeds 99%.

4. Results and concluding remarks

[11] Note that there are multiple other factors affecting the number of *foma* rules required by a language, such as the extent of the corpus, the level of generality of the coded rules, the number of words present in the data published so far etc.

4.1 We conclude that

(a) *The* foma *scripts digitally replicating the rules of the glottal fricative theory* (GFT) *are consistent and capable of generating the* IE *data* without errors except for a well-definable set of open research problems.[12] This result supports the view that the sound law system studied is indeed a close representation of the historical Indo-European sound laws and represents the split of PIE into the IE languages with sufficient accuracy to make it of considerable interest for the future development and improvement of IE linguistics.

(b) After two centuries of research the *primary objective of comparative Indo-European linguistics, a reconstruction of* PIE *essentially equivalent to the* IE *languages, has been preliminarily achieved.* As the *foma* rules already coded do not only apply to the data on the basis of which they were initially compiled, but also to the Indo-European data in general, it is possible to expand the published data to cover the main bulk of the vocabularies of the ancient languages without significant problems, which in turn will provide IE linguistics with a new digital discussion matrix.

4.2 The immediate repeatability and verification of these results has been facilitated by explicit *foma* chains, proving each reconstruction rule by rule.[13] As the *foma* scripts and the data are open source, the results are reproducible and verifiable in a transparent manner.[14] Finally, all errors in the generation of the data are automatically identified, marked in red and collected into a single PIE Lexicon page for future study of the open research problems they reflect.[15]

4.3 An independent confirmation for these results will be sought in the future from the hitherto unutilized potential of the *foma* scripts: If the IE *foma* scripts are consistent, as proven by the finite-state test, we can expect other methodologically solid applications such as a self-organizing map and/or a digitized language family tree to display the *foma* scripts in a consistent and informative manner reflecting the actual distinctions and features of the IE language family.

4.4 As specifically related to the future theme of this congress we would like to conclude with some general remarks on Indo-European linguistics, the comparative

[12] The remaining errors are generic, i.e. they represent well-defined classes of open research problems in IE linguistics that can be also be solved at least to a degree in the future. In addition to the PIE accent/tone problem concerning all Indo-European languages and only tentatively approached in PYYSALO 2013, there are currently a dozen minor research problems related to individual subgroups or languages.

[13] Clicking a PIE reconstruction (in blue) on the left side of the respective IE stem generates the form with a *foma* chain, explicitly stating all the rules applied and their mutual order.

[14] For the *foma* rules and the data package please contact jouna.pyysalo@helsinki.fi.

[15] Note that the red marking is placed in the phoneme of the attested form, despite being correct, in order to note the failure to generate that particular item. For the PIE Lexicon mismatch page, see http://pielexicon.hum.helsinki.fi/?alpha=ALL&view=mismatch.

method of reconstruction, finite-state methods, language technology, and artificial intelligence.

a) For Indo-European linguistics the successful digitalization of the core IE sound law system constitutes a merger of the traditions of the humanities and technology into a next-generation computational environment. In the future IE linguistics will be essentially practiced as a natural science as originally envisioned by August SCHLEICHER already in the 1850s. This conclusion may initially sound surprising, but if we consider Alan TURING's idea that language, as the manifest form of thinking, is logical, it is only natural that languages are studied in a fully digitalized, logical environment. This will, once achieved, allow the researchers of the field to concentrate on new, higher-level problems, such as the structure of proto-language and language itself as the vehicle of delivery of thinking.

b) The compatibility of FST and the comparative method demonstrate that FST provides a rigorous formal calculus for mapping cognates from proto-forms to daughter languages and evaluation of the consistency of the systems involved. The result applies not only to Indo-European but all language families. We welcome other similar implementations and hope to facilitate these efforts by making the PIE Lexicon platform open-source and available as a download that can be applied to any data in the future.

c) In addition to the finite-state methods, also infinite approaches, including especially machine learning (or artificial intelligence, AI), can of course be used in solving the problems of the sound law systems of languages. It is not excluded that AI could produce solutions to the open research problems left unsolved by the human-controlled finite-state methods. Since all errors appear in specific contexts, they constitute well-defined objects for artificial intelligence. Should machine learning provide the first ever non-man-made sound law, it would not only be a significant breakthrough as such, but could enable us to develop more efficient methods in language technology in fields facing similar problems.[16]

References

BEESLEY, Kenneth E. & KARTTUNEN, Lauri. 2003. Finite State Morphology: Xerox tools and techniques. (Studies in computational linguistics 3). Center for the Study of Language and Information, Stanford.

[16] Although we haven't implemented any AI applications as of yet, the preliminary estimate of the (Peer) Review 3 of this article is definitely positive: "It is likely that many of the *foma* rules can be learned in an automatic manner if appropriate training data is available. Machine learning could potentially also be used to reveal potentially other (latent) properties and relations in this type of data when it comes to data generation and cross-language comparison."

CHOMSKY, Noam & HALLE, M. 1968. The Sound Pattern of English. New York: Harper and Row.

COLLINGE, N. E. 1985. The Laws of Indo-European. Benjamins, Amsterdam.

----- 1995. Further Laws of Indo-European. In: On Languages and Language: The Presidential Addresses of the 1991 Meeting of the Societas Linguistica Europaea. ed, Werner Winter. Trends in Linguistics. Studies and Monographs, 78. Mouton, Berlin: 27-52.

---- 1999. The Laws of Indo-European: The State of Art. Journal of Indo-European Studies, 27:355-377.

FICK, August. 1870-71. Vergleichendes Wörterbuch der indogermanischen Sprachen. Göttingen: Vandenhoeck & Ruprecht.

HROZNÝ, Bedřich. 1917. Die Sprache der Hethiter, ihr Bau und ihre Zugehörigkeit zum indogermanischen Sprachstamm. (Boghazköi-Studien 1–2). Leipzig: Hinrichs.

HULDEN, Mans. 2009a. "Foma: a finite-state compiler and library". In: Proceedings of the EACL 2009 Demonstrations Session (Athens, Greece). Association for Computational Linguistics, pp. 29–32.

----- 2009b. Finite-State Machine Construction Methods and Algorithms for Phonology and Morphology. PhD Dissertation, University of Arizona.

----- 2009c. Regular expressions and predicate logic in finite-state language processing. In Proceedings of the 2009 conference on Finite-State Methods and Natural Language Processing: Post-proceedings of the 7th International Workshop FSMNLP, pp. 82-97.

JOHNSON, C. D. 1972. Formal Aspects of Phonological Description. The Hague: Mouton.

JONES, Sir William. 1788. Anniversary Discourse (February 2nd 1786). Asiatick Researches 1: 415–431.

KAPLAN, R. and KAY, M. 1981. Phonological rules and finite-state transducers. Paper presented to the Winter Meeting of the Linguistic Society of America, New York.

KARTTUNEN, Lauri. 1995. The replace operator. In: Proceedings of the 33rd annual meeting of the Association for Computational Linguistics (Cambridge, MA). Association for Computational Linguistics, pp. 16–23.

KEMPE, A., and KARTTUNEN, L. 1996. "Parallel replacement in finite state calculus". In: Proceedings of the 16th conference on Computational linguistics (Copenhagen, Denmark), Volume 2. Association for Computational Linguistics, pp. 622–627.

LESKIEN, August. 1876. Die Deklination im slavisch-litauischen und germanischen. Leipzig: Hirzel.

PEDERSEN, Holger. 1962. The discovery of language. Linguistic science in the nineteenth century. Trans. John W. Spargo. Bloomington (IN): Indiana University Press.

PYYSALO, Jouna. 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. https://helda.helsinki.fi/handle/10138/41760

SCHLEICHER, August. 1852. Die Formenlehre der kirchenslavischen Sprache, erklärend und vergleichend dargestellt. H. B. König, Bonn.

SZEMERÉNYI, Oswald. 1967. The new look of Indo-European. Reconstruction and typology. *Phonetica* 17, 67-99.

----- 1970. *Einführung in die vergleichende Sprachwissenschaft.* Darmstadt: Wissenschaftliche Buchgesellschaft.

YLI-JYRÄ, Anssi. 2008. Transducers from parallel replace rules and modes with generalized lenient composition. In: 6th International Workshop on Finite-State Methods and Natural Language Processing, FSMNLP 2007 (Helsinki). Revised Papers, pp. 197–212.