Proto-Indo-European Lexicon and the Next Generation of Smart Etymological Dictionaries: The Technical Issues of the Preparation

Jouna Pyysalo[1], Fedu Kotiranta[2], Aleksi Sahala[1], Mans Hulden[3]

[1] University of Helsinki, Faculty of Arts, PL 24, 00014 Helsingin yliopisto
[2] Independent
[3] University of Colorado Boulder, Department of linguistics, Hellems 290,
295 UCB Boulder, CO 80309

E-mail: jouna.pyysalo@helsinki.fi, fedu@mediamoguli.fi, aleksi.sahala@helsinki.fi,
mans.hulden@gmail.com

Abstract

Proto-Indo-European Lexicon (PIELex) is the generative etymological dictionary of Indo-European (IE) languages at http://pielexicon.hum.helsinki.fi. It is the first dictionary in the world capable of mechanically generating its data entries, i.e. the lexical stems of more than 120 of the most archaic IE languages. In addition, in order to solve the reverse process work has already begun on the problem of the mechanical generation of Proto-Indo-European (PIE) from the IE data,. The plan of the project as a whole is to run PIE Lexicon using an operating system (OS), a computer, under which the dictionary and its data are exclusively governed by smart features ranging from semantics to morphology, and the very root structure of Proto-Indo-European itself.

In principle PIE Lexicon is compatible with all digitized etymological dictionaries of IE languages, and as the operating system is scientifically neutral, material of any language or language family can be implemented onto the platform. By outlining the key features of the future coding plan we hope to offer ideas, assistance and support for other enterprises in the field of electronic lexicography.

Keywords: Indo-European linguistics; Proto-Indo-European; electronic lexicography; finite-state technology; historical linguistics

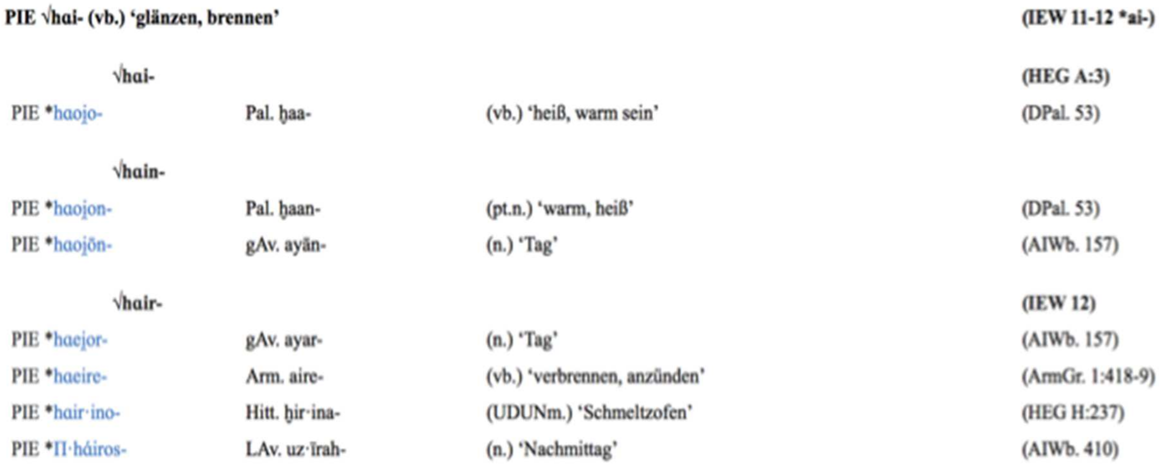
1. General introduction to PIE Lexicon

An etymological dictionary deals with at least two genetically related languages, and is therefore smart by default when compared to dictionaries of a single language. The Indo-European (IE) language family is one of the largest in the world, comprising some 400 languages. This naturally increases the complexity at the outset, as the preserved inherited data appear in mutually incompatible native writing systems. This problem is solved by means of the comparative method of reconstruction, a procedure that allows arranging etymologically related items into correspondence sets and projecting them back into the unitary phoneme system of a single language, Proto-Indo-European.

Furthermore, the IE languages are usually attested in several successive chronological phases. This entails additional complex requirements, the most important of them being that since the older the language is, the fewer changes it has undergone, it is necessary to start the reconstruction with the oldest form of every language in order to optimize the output. The full addition of the later layers becomes possible once these preconditions have been met.

Initially PIE Lexicon will be dealing with perhaps some 150-200 languages, mostly representing the oldest or a middle period in the written history of the languages, but also already including modern ones when the language is attested only in two periods such as, for instance, Lithuanian and Russian.

The etymological entries of PIE Lexicon, an example of which is shown in Figure 1, are of the following form:



PIE √hɑi- (vb.) 'glänzen, brennen'			**(IEW 11-12 *ai-)**
√hɑi-			**(HEG A:3)**
PIE *hɑojo-	Pal. ḫaa-	(vb.) 'heiß, warm sein'	(DPal. 53)
√hɑin-			
PIE *hɑojon-	Pal. ḫaan-	(pt.n.) 'warm, heiß'	(DPal. 53)
PIE *hɑojōn-	gAv. ayān-	(n.) 'Tag'	(AIWb. 157)
√hɑir-			**(IEW 12)**
PIE *haejor-	gAv. ayar-	(n.) 'Tag'	(AIWb. 157)
PIE *haeire-	Arm. aire-	(vb.) 'verbrennen, anzünden'	(ArmGr. 1:418-9)
PIE *hair·ino-	Hitt. ḫir·ina-	(UDUNm.) 'Schmeltzofen'	(HEG H:237)
PIE *II·háiros-	LAv. uz·īrah-	(n.) 'Nachmittag'	(AIWb. 410)

Figure 1: An etymological entry of PIE Lexicon

The topmost horizontal line (in bold) starting with PIE √hai- (vb.) 'glänzen, brennen' and ending with (IEW 11-12 *ai-) represents a Proto-Indo-European root with a reference to earlier research. The root and its extensions (PIE √hɑi-, √hɑin- √hɑir-) are morphologically arranged as nodes of the root.

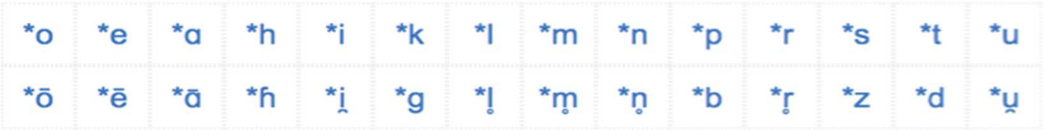
The PIE Lexicon data entries, consisting of a PIE reconstruction (e.g. PIE *hɑoi̯o-) and the respective IE stem, (e.g. Pal. haa-), the morphological classifier of the IE stem '(vb.)', translation ('heiß, warm sein'),[1] and the reference '(DPal. 43)' are arranged under the nodes from which they were originally derived.

[1] Note that in the initial version of PIE Lexicon the translations are those provided in the quoted source (usually a dictionary). In addition to this, future versions of PIE Lexicon will provide translations in several main languages, initially at least German and English.

2. Mechanical generation of the Indo-European data from PIE

In traditional (non-digital) etymological dictionaries the PIE reconstructions and the proto-phoneme system are not necessarily explicit. Furthermore, the sound laws leading from PIE to the IE languages are not always evident, and sometimes they are even inconsistent. In short, the entire traditional reconstruction is more or less intuitive, to a degree necessitating scholars to take leaps of faith instead of allowing them to rely on robust proofs by digitized sound laws.

In contrast to the traditional etymology, PIE Lexicon uses an explicitly defined PIE proto-phoneme inventory shown in Figure 2:



*o	*e	*ɑ	*h	*i	*k	*l	*m	*n	*p	*r	*s	*t	*u
*ō	*ē	*ɑ̄	*ɦ	*i̯	*g	*l̥	*m̥	*n̥	*b	*r̥	*z	*d	*u̯

Figure 2: The PIE phoneme inventory of PIE Lexicon

In the reconstruction these and only these phonemes are allowed, which blocks the use of ad hoc-phonemes.[2] The fact that the set is sufficient to reconstruct the IE forms proves the completeness of the PIE phoneme inventory.[3]

The most archaic IE sound laws, revised in Pyysalo (2013) have been digitized with the foma finite-state-compiler developed by Mans Hulden (2009).[4] In practice this means that the non-formal sound laws used by the rest of the field have been replaced with their foma counterparts, 800 unique sound laws having been coded at this point. For illustration's sake, the loss of PIE *ɦ/h as a segmental phoneme is coded with the following two rules:

define Rɦ›0 ɦ -> 0 || .#. | \Stop _ ;   define Rh›0 h -> 0 || .#. | \Stop _ ;

In order to facilitate the mechanical generation of the IE stems the individual sound laws coded in foma have been arranged in chronological order for each language, forming the sound law system of that language in digitized form. These sound law (foma) scripts can in turn be used to mechanically generate the actual forms of the language from their respective PIE reconstructions. The sound law scripts, as far as coded, can be found in the control bar at the bottom of the PIE Lexicon site. By

[2] For the revised PIE phoneme inventory used in PIE Lexicon, a further revision of Szemerényi (1967), see Pyysalo (2013).

[3] For the completeness (i.e. sufficiency in the generation of the IE data) of the phoneme inventory, see Pyysalo, Sahala and Hulden (2018).

[4] For the latest version of *foma*, see https://code.google.com/archive/p/foma/.

clicking ‘Select rule set’, choosing one (e.g. gAv.) and clicking ‘Show rules’, the respective sound law script is opened:

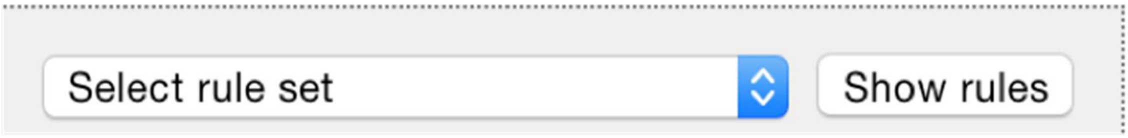
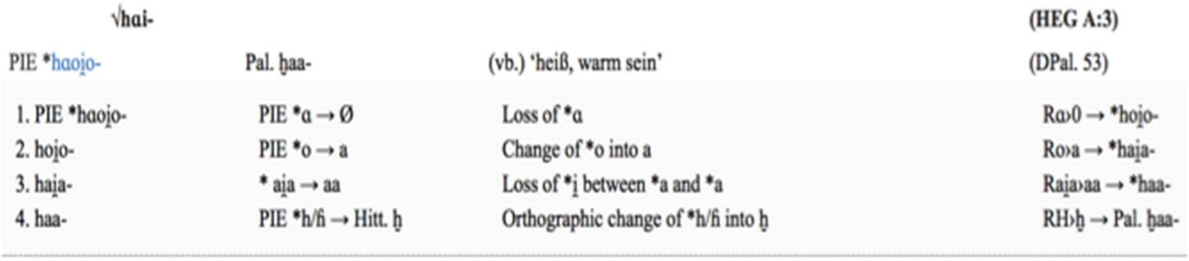


Figure 3: The control bar access to PIE Lexicon sound law scripts

By now some 120 of the most archaic IE languages have been provided with a sound law script in PIE Lexicon, and new scripts are constantly added as new languages emerge when new data is published. The sound laws provably form a consistent system and generate the IE data with an accuracy rate exceeding 99% (see Pyysalo, Hulden & Sahala 2018), strongly suggesting that the system is valid, i.e. sound and complete.

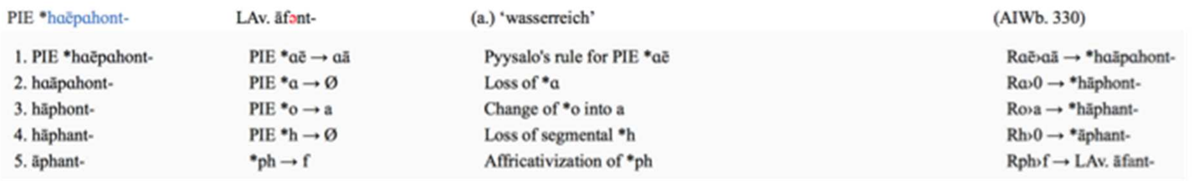
Due to the availability of the sound law scripts the PIE Lexicon operating system mechanically generates the IE stems (output) from their PIE reconstructions (input). PIE Lexicon editors, users, and visitors can explicitly verify the mechanical generation of the data by clicking a reconstruction (in blue). This is a command for the code reader to execute the foma script and create an explicit foma proof chain consisting of successive, explicitly stated sound laws leading from the PIE reconstruction to the respective IE stem, as shown in Figure 4:



√hɑi-			(HEG A:3)
PIE *hɑoj̯o-	Pal. ḫaa-	(vb.) ‘heiß, warm sein’	(DPal. 53)
1. PIE *hɑoj̯o-	PIE *ɑ → Ø	Loss of *ɑ	Rɑ›0 → *hoj̯o-
2. hoj̯o-	PIE *o → a	Change of *o into a	Ro›a → *haj̯a-
3. haj̯a-	* aj̯a → aa	Loss of *j̯ between *a and *a	Raj̯a›aa → *haa-
4. haa-	PIE *h/ɦ → Hitt. ḫ	Orthographic change of *h/ɦ into ḫ	RH›ḫ → Pal. ḫaa-

Figure 4: An example of a foma proof chain in PIE Lexicon

When the output form has been generated, an additional function of the operating system (OS) compares the output to the actual stem form, and if these match, the letters of the attested form are shown in black as in the previous screenshot. If, on the other hand, any phoneme is erroneously generated, the error is shown in red in the attested form (Figure 5).



PIE *hɑĕpɑhont-	LAv. āfənt-	(a.) ‘wasserreich’	(AIWb. 330)
1. PIE *hɑĕpɑhont-	PIE *ɑĕ → ɑā	Pyysalo's rule for PIE *ɑĕ	Rɑĕ›ɑā → *hɑāpɑhont-
2. hɑāpɑhont-	PIE *ɑ → Ø	Loss of *ɑ	Rɑ›0 → *hāphont-
3. hāphont-	PIE *o → a	Change of *o into a	Ro›a → *hāphant-
4. hāphant-	PIE *h → Ø	Loss of segmental *h	Rh›0 → *āphant-
5. āphant-	*ph → f	Affricativization of *ph	Rph›f → LAv. āfant-

Figure 5: An example of an error (in red) in foma proof chain

All errors have been collected on a separate 'mismatch' page at the address http://pielexicon.hum.helsinki.fi/?alpha=ALL&view=mismatch. Although about half of the currently listed errors are typos or result from a necessary rule have not yet been coded, there are some 200 errors forming a dozen (or so) open research (sound law) problems to be solved.

Finally, and as particularly relevant to lexicography, the capability of the operating system to generate the Indo-European languages from the PIE phoneme inventory reduces the some 150 IE languages which are to be treated into a single, uniform language to manage, an advantage readily understood by anyone familiar with the complexities of lexicography in an environment requiring the treatment of a relatively large set of languages.

3. The automatic generation of PIE on the basis of Indo-European data

The second most challenging problem of historical linguistics in language technology after the automatic generation of IE data from PIE discussed above involves the mechanical reconstruction of the proto-language (here: PIE) and the definition of etymologies based on the attested data (here: IE). With regard to this problem there are two main solutions available, the original (traditional) and the recently emerged digital one. These ultimately represent the same process, that of reversing the order of the historical sound changes that have taken place during the development of a language and, based upon this, engineering a decision method allowing for the identification of originally identical Indo-European forms and their etymologies.

The traditional decision method of Indo-European etymology was originally outlined by August Schleicher. In Schleicher's (1852b: iv-v) words, quoted here in Koerner's (1982: 24) translation:

> "When comparing the linguistic forms of two related languages, I firstly try to trace the forms to be compared back to their probable base forms, i.e., that structure [gestalt] which they must have [had], excepting phonetic laws [lautgesetze] which became effective at a later time, or at least I try to establish identical phonetic situations in historical terms for both of them."

In modern terminology the identification of a PIE prototype and its reconstruction is based on creating a disjunction of possible PIE prototypes of an Indo-European morpheme. This disjunction, in turn, is compared to the similar disjunctions of other Indo-European languages, and if a formal match that is also semantically acceptable is found between two disjunctions, then an etymology (and a reconstruction) has been found.

This procedure is a decision method in a mathematical sense, i.e. it leads to the solution if sufficient data have been preserved. For this reason the comparative method has proven its worth in allowing scholars to reconstruct the proto-forms of the discovered correspondences, simultaneously settling their etymologies.

The attempts to mechanize the reconstruction (here: PIE) have been unsuccessful up to this day, and have by now been largely abandoned and replaced by AI-based attempts to identify the processes involved (see Sims-Williams 2018). In the case of the Indo-European languages, the reason for the failure does not lie in the decision method or in its digitized formulation, the latter equally functional as the former, but in an imperfect set of sound laws leading from IE to PIE. If this (or any similar) set does not actually represent a consistent system of historical sound laws, then the system does not yield correct reconstructions, because the decision method essentially consists of reversing the sound laws, allowing the back-projection of the PIE prototypes mentioned by Schleicher. This can be seen from the digitized version of the method, consisting in essence of the following steps:

a) The order of the sound law (foma) scripts is reversed so that the first rules become the last ones and the last ones become the first.

b) In addition, the direction of the individual sound laws of the scripts, basically implications of the form ‘if X, then Y’, is also reversed, i.e. each rule X → Y is turned into Y → X.

This reversing of the sound law scripts makes it possible to generate digital counterparts of Schleicher’s disjunctions, except for the fact that the code reader lacks the common sense applied in the intuitive use of the method. Without this the code reader generates infinite chains of phonemes, especially lost ones. In order to eliminate the problems related to this it is necessary to add morphophonological constraints to the code that exclude impossible prototypes such as †hhhhhhhhhhhep-.

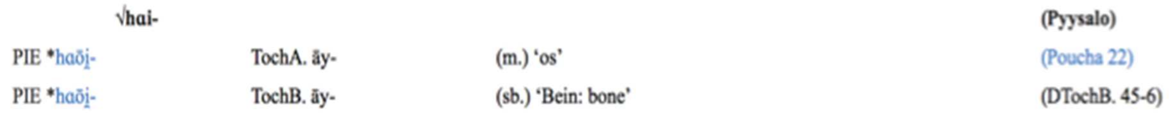
Once the morphophonological constraints have been added to the reversed sound law scripts, their output is in essence identical with the intuitively used decision method, i.e. the algorithm generates disjunctions of possible PIE prototypes for the IE forms used as input. At this point it is possible to code and implement the decision method function, basically an intersection seeking identities from the terms of each two PIE disjunctions. If a common denominator is identified by the function then a PIE reconstruction has been defined and an etymology has been found, if the equation satisfies the semantic criteria.

With the decision method function coded, also the intuitive comparison, done manually until now, has been explicated and may be used in automatically reconstructing PIE prototypes, testing the hitherto suggested etymologies as well as finding new ones, discovered by a computer for the first time in the history of the field.

4. On the digitalization of other key features of PIE Lexicon

The core idea of PIE Lexicon, illustrated above with mechanized generation of IE data and the PIE reconstructions, is to digitize every possible feature and aspect of the linguistic data. This will ultimately result in an etymological dictionary exclusively containing smart or digitized features. In order to illustrate this in further detail several other key features to be digitized will be outlined in this paragraph.

Initially the focus of PIE Lexicon is placed on etymology and therefore we do not aim at full coverage of the entire IE data like the dictionaries of individual IE languages. This partial display of the material is compensated for with active links attaching the IE data entries of PIE Lexicon to other electronic dictionaries available on the internet. This automatic linking has already begun in a manner illustrated by the screenshot below, where the blue in '(Poucha 22)' indicates an active link leading to the respective entry in another electronic dictionary:[5]



√hai-			(Pyysalo)
PIE *haŏi̯-	TochA. äy-	(m.) 'os'	(Poucha 22)
PIE *haŏi̯-	TochB. äy-	(sb.) 'Bein: bone'	(DTochB. 45-6)

Figure 6: An automatic external hyperlink in PIE Lexicon

This exploitation of language resources allows the PIE Lexicon users to verify the entries and, something of equal importance, reach comprehensive internal data and description of the IE entries.

Automated customization of the dictionary to the users' needs and characteristics is already provided in a preliminary form in the search function located in the control bar at the bottom of the site:

Figure 7: The PIE Lexicon search engine window

Initially the search function is referential, only allowing the user to search for a single, untagged item, but this function will be upgraded into a full-scope advanced search with any number of search variables of all categories to exactly define any data segments needed by scholars in their work.

[5] For the actual entry in CEToM, see https://www.univie.ac.at/tocharian/?āy.

The rightmost (optional) column is reserved for the attested forms of the IE stems and their grammatical analysis, as shown in Figure 8:

Hitt. ḫaa-	(vb.) 'vertrauen, jemandem etwas glauben'	(HHand. 34)	(Hitt. ḫa-a, ḫa-a-mi [1sg], ḫa-a-ši, ḫa-a-ir)

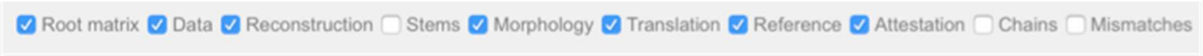
Figure 8: A PIE Lexicon data entry line including attested forms

Once the priority coding tasks have been established, a key NPL tool, the automatic grammatical analysis of the attested forms, will be implemented in this section. In addition, the attested forms and their exact locus, possibly in the context of the original text, will be added to each form, if not already present.

Until this point the data of the pilot versions of PIE Lexicon have been limited to correspondence sets containing at least one of the best preserved Old Anatolian languages: Hittite, Palaic, Cuneiform Luwian, or Hieroglyphic Luwian. These languages have uniquely preserved the PIE 'laryngeal' (i.e. glottal fricative PIE *h) as such, giving them priority in the reconstruction of PIE ever since their discovery. In the next coding phase, however, such limitations no longer apply, and inherited data of all languages will be used equally to compile the first complete initial PIE *u/u̯, comprising the main bulk of the entire most archaic data starting with this initial. As this data segment, the first of the total of eleven main entries,[6] will be about a thousand pages long, its publication will turn PIE Lexicon into a big data program proper and, equally importantly, the stable, largely permanent initial display of the data will allow scholars of IE linguistics as well as other fields to begin the study of the data in earnest.

As the entry PIE *u/u̯ is representative in terms of the preserved material, its publication will make possible especially the study of the morphology, the original structure, formation and the origin of Proto-Indo-European. This is facilitated by the fact that the reconstructions contain the information of the respective IE correspondence sets in compressed form, i.e. this single, unified language can be taken as the primary object of the study instead of the earlier material divided into some 150 distinct languages. This study has already been anticipated in the control bar at the bottom of the site (Figure 9).

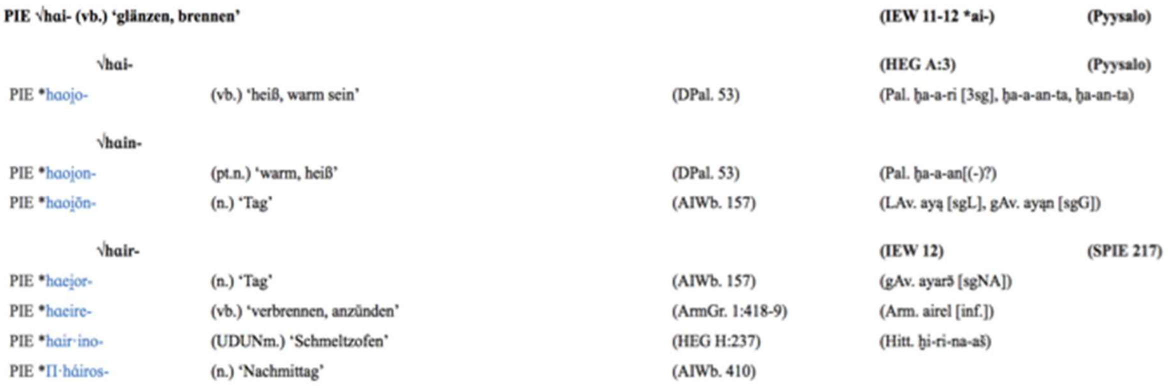
[6] The PIE phoneme inventory (see §2.2.1) comprises of fourteen items, each with two varieties in columns. Of these fourteen phonemes the three leftmost are vowels, which occur as independent roots in only a few cases, to be dealt with the introduction in a small separate work. Due to this the dictionary proper splits into eleven main entries corresponding to the remaining consonantal phonemes of the inventory.



☑ Root matrix ☑ Data ☑ Reconstruction □ Stems ☑ Morphology ☑ Translation ☑ Reference ☑ Attestation □ Chains □ Mismatches

Figure 9: The PIE Lexicon data selection control bar

When the ‘Stems’ option is deactivated in the manner shown in the screenshot above, the IE forms are not shown and the translations apply to the PIE reconstructions instead. A description of a single language, PIE, gives the following results:

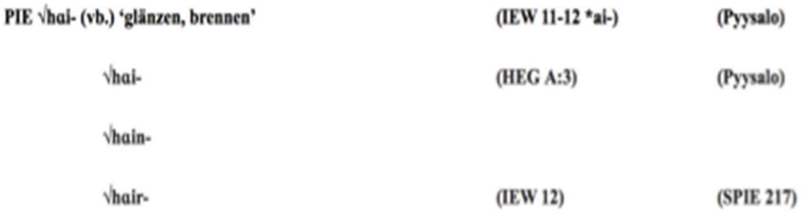


PIE √hai- (vb.) ‘glänzen, brennen’			**(IEW 11-12 *ai-)**	**(Pyysalo)**
√hai-			**(HEG A:3)**	**(Pyysalo)**
PIE *haojo-	(vb.) ‘heiß, warm sein’	(DPal. 53)	(Pal. ḫa-a-ri [3sg], ḫa-a-an-ta, ḫa-an-ta)	
√hain-				
PIE *haojon-	(pt.n.) ‘warm, heiß’	(DPal. 53)	(Pal. ḫa-a-an[(-)?)	
PIE *haojōn-	(n.) ‘Tag’	(AIWb. 157)	(LAv. ayą [sgL], gAv. ayąn [sgG])	
√hair-			**(IEW 12)**	**(SPIE 217)**
PIE *haejor-	(n.) ‘Tag’	(AIWb. 157)	(gAv. ayarə̄ [sgNA])	
PIE *haeire-	(vb.) ‘verbrennen, anzünden’	(ArmGr. 1:418-9)	(Arm. airel [inf.])	
PIE *hair·ino-	(UDUNm.) ‘Schmeltzofen’	(HEG H:237)	(Hitt. ḫi-ri-na-aš)	
PIE *Π·háiros-	(n.) ‘Nachmittag’	(AIWb. 410)		

Figure 10: The PIE Lexicon in the PIE mode without IE languages

With this simple device the IE data has turned into PIE data, and the further digitalization of these structures, including simplifications, enables us to digitally define and manage the entire word formation of Proto-Indo-European.

Similarly, by releasing all buttons in the control bar except ‘Root matrix’, the root structure of PIE becomes directly observable, as shown in Figure 11:



PIE √hai- (vb.) ‘glänzen, brennen’	**(IEW 11-12 *ai-)**	**(Pyysalo)**
√hai-	**(HEG A:3)**	**(Pyysalo)**
√hain-		
√hair-	**(IEW 12)**	**(SPIE 217)**

Figure 11: PIE Lexicon in PIE root and extension mode

As soon as the first representative data set becomes available, these and other similar devices will facilitate the study and mechanization of the proto-language PIE in an exact manner, similar to how the Indo-European languages themselves have already been mechanized in PIE Lexicon.

The complete data entries enable the coding and digital management of the semantics of Proto-Indo-European. This observation is based on the fact that every PIE morpheme is associated with the meaning conveyed by its IE counterpart, which associates the morpheme with a specific morphological category (e.g. verb or adjective). Under these circumstances it is possible to define the semantic fields of the PIE roots. Each of these contains a number of IE stems (e.g. verbs and nouns) with meanings, the combination of which defines the semantic field of the root in question. Once these meanings have been defined and coded for the individual PIE roots, it becomes possible to compare multiple PIE roots having similar semantic fields. This will provide a warning of potential errors in the classification of the data if a parallel for the meaning of a semantic field is absent in other roots with otherwise identical semantic fields. Reversely, forms that have hitherto failed to be connected to any root can be attached to one, if a semantic parallel is available in the semantic field of another, morphologically different root. As a whole this means that the relatively complex and abstract study of meaning in Proto-Indo-European can be established in a strictly scientific environment.

Initially PIE Lexicon uses IE stems, supported by some attested forms, as its data entries. Naturally this restriction is artificial, and PIE Lexicon can be expanded to contain all the attested data and the related scientific discussion so far. Achieving this is not problematic, because a separate article page can be simply opened for each data entry, allowing the editors and contributors to compile an article containing the full attested data, the related scientific discussion so far, and other relevant observations.

5. Summary

As a whole the underlying plan of PIE Lexicon is to digitize (and turn smart) all of its features, ranging from reconstruction to semantics and its data. In other words, the long-term aim is to critically summarize two centuries of Indo-European linguistics as a whole into a single file, ultimately containing every piece of data or material bearing relevance to it, and offer it to scholars and others interested. While this task is too ambitious to be achieved by a single person or even a single team, the PIE project is built upon the chassis of natural science and is thus open-ended. This allows new administrators and teams to take over the management and continuation of the project in future decades, possibly even centuries, during which corrections, improvements, supplementations, and extensions to the original can be executed when needed until all problems of the field, including the new high-level ones emerging during the process, have been solved.

As specifically related to the content, the project is initially designed to optimize the digital treatment, analysis and presentation of the primary material, the Indo-European languages themselves. However, as soon as the basic problems involved are satisfactorily managed, the aim is to increasingly shift the focus to the digitized study of Proto-Indo-European, the inductive equivalent of the Indo-European languages. This will take the field far beyond the scope of traditional Indo-European linguistics, resulting not only in the triumph of the electronic Neogrammarians mentioned by Sims-Williams (2018), but also of electronic lexicography as a whole in the 21st century.

In order to reach such ambitious goals the importance of electronic lexicography cannot be exaggerated: As an empirical science Indo-European linguistics is exclusively data-based. Accordingly, the more advanced and smarter electronic dictionaries of the field get, the more advantages result for science. In addition, the cooperation of electronic dictionaries will play a vital role in future science: Not only the active links, guiding the users to other sites and thus promoting these, but more abstract sharing of data, e.g. in the forms of etymologies, improves the content of electronic dictionaries..

6. References

Hulden, M. (2009). *Finite-State Machine Construction Methods and Algorithms for Phonology and Morphology.* PhD Dissertation, University of Arizona.

Koerner, K. (1982). The Schleicherian Paradigm in Linguistics. *General Linguistics* 22: 1-39.

Pyysalo, J. (2013). *System PIE: The Primary Phoneme Inventory and Sound Law System for Proto-Indo-European.* PhD Dissertation, University of Helsinki. Publications of the Institute for Asian and African Studies 15. Helsinki: Unigrafia Oy. https://helda.helsinki.fi/handle/10138/41760

Pyysalo, J., Sahala, A. & Hulden, M. (2018). Verifying the Consistency of the Digitized Indo-European Sound Law System Generating the Data of the 120 Most Archaic Languages from Proto-Indo-European. In Eetu Mäkelä, Mikko Tolonen, and Jouni Tuominen (eds) *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference Helsinki, Finland, March 7-9, 2018.* Helsinki, University of Helsinki. http://ceur-ws.org/Vol-2084/paper7.pdf

Schleicher, A. (1852b). *Die Formenlehre der kirchenslawischen Sprache, erklärend und vergleichend dargestellt.* Bonn: H.B. König.

Sims-Williams, P. (2018). Mechanising historical Phonology. *Transactions of the Philological Society* 116, pp. 555-573.

Szemerényi, O. (1967). The new look of Indo-European reconstruction and typology. *Phonetica* 17, pp. 65-99.