Towards a Framework of Etymological Relations

Nathanael E. Schweikhard and Johann-Mattis List
(Max Planck Institute for the Science of Human History, Jena)

Computational approaches to linguistics have made huge progress in the last decades. However, when reconstructing language history, they often work with limited models of language change, for example not including word formation (List 2016, 126-128), and sometimes not even regular sound change (e.g. Boc et al. 2010).

Etymological dictionaries on the other hand host a rich amount of linguistic data and are based on a vast range of considerations. Yet they present it in a traditional format which inhibits the fast retrieval of larger amounts of information from these dictionaries. Furthermore, due to the unstructured prose format in which at least parts of some etymological entries are written, their authors are not forced to always make clear on which grounds two given words were deemed to be or not to be cognate.

Therefore we propose a more exhaustive digital framework for etymological relations. This will allow for the representation of etymological relations in a machine-readable fashion, ready to be adopted for a variety of quantitative studies on language history and change.

Additionally it also serves as a way to test etymological reconstructions for consistency. By specifying in which way exactly the members of a cognate set are considered related, including both regular sound change and irregular processes like word formation, mistakes become more noticable and the number of assumptions becomes clear. Thereby also different proposals regarding the reconstruction of language history can be compared more transparently (Gray et al. 2007, 13).

In order to test and further enhance our framework, we have initiated a pilot project in which we use it to digitize the etymological relations of about 100 entries of *Nomina im indogermanischen Lexikon* (NIL, Wodtko et al. 2008). We further limit this endeavor to include only material from some of the attested languages (Ancient Greek, Latin, Old High German, Vedic) and only those etymological relations which were deemed certain by the dictionary's editors.

In figure 1 and 2 you see some attested forms and reconstructions from one of NIL's entries (and a reconstruction from Mallory and Adams 2006) and how this word family is handled by our framework. This format consists of two tables which were inspired by the CLDF-initiative (Forkel et al. 2018).

In the first table we annotate cognacy between morphemes, here based on morpheme borders in the reconstruced proto-language. Morphemes that differ between languages only by regular sound change are given the same ID in COGNATES, whereas in those cases where a non-concatenative morphological process like ablaut was involved, they receive the same ID only in the column ROOTS.

The second table explicitely notes the sound change and word formation processes in which words differ from each other. In the final version, the linguistic data will be presented in IPA, and we will specify the regular sound changes involved in a separate file. In my talk I will present first results of this project.

ID	LANGUAGE	CONCEPT	FORM	MORPHEMES	COGNATES	ROOTS
1	Old High German	eternity	ēwo	ēw o	1 2	1 2
2	Ancient Greek	life	aiōn	ai ōn	1 2	1 2
3	Vedic	life	ā́yu	ā́yu	3	1
4	Vedic	long-living	dīrghā́yu	dīrgh á ā́yu	4 5 3	3 4 1
5	Vedic	young	yúvan	yúv an	6 7	1 5
6	Latin	(deity name)	iūnō	iū n ō	6 8 2	1 5 2
7	Indo-European	life	*h₂aḭ̯-u-on-	h₂aḭ̯u on	3 2	1 2
8	Indo-European	life	*h₂oḭ̯-u-	h₂oḭ̯u	1	1
9	Indo-European	long-living	*dl̥h₁gʰ-ó-h₂oḭ̯-u-	dl̥h₁gʰ ó h₂oḭ̯u	4 5 1	3 4 1
10	Indo-European	young	*h₂i̯-u-h₃on-	h₂i̯u h₃on	6 7	1 5
11	Indo-European	the young one	*h₂i̯-u-h₃n-on-	h₂i̯u h₃n on	6 8 2	1 5 2

Figure 1: Annotating cognacy between morphemes.

Source	Source-ID	Target	Target-ID	Change
*h₂aḭ̯-u-on-	7	aiōn	2	sound change
*h₂oḭ̯-u-	8	*h₂aḭ̯-u-on-	7	e-grade, on-suffix
*h₂oḭ̯-u-	8	*dl̥h₁gʰ-ó-h₂oḭ̯-u-	9	compound with *dl̥h₁gʰ-ó-
*dl̥h₁gʰ-ó-h₂oḭ̯-u-	9	dīrghā́yu	4	sound change
...	...	...	...	...

Figure 2: Annotating etymological relations between full words.

Bibliography:

Boc, Alix, Anna Maria Di Sciullo, and Vladimir Makarenkov. 2010. “Classification of the Indo-European Languages Using a Phylogenetic Network Aproach.” In *Classification as a Tool for Research*, edited by Hermann Locarek-Junge and Claus Weihs, 647–55. Berlin; Heidelberg: Springer.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. “Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics.” *Nature Scientific Data* 180205 (5). https://doi.org/10.1038/sdata.2018.205.

Gray, Russell D., Simon J. Greenhill, and Malcolm D. Ross. 2007. “The Pleasures and Perils of Darwinzing Culture (with Phylogenies).” *Biological Theory* 2 (4): 360–75.

List, Johann-Mattis. 2016. “Beyond Cognacy: Historical Relations Between Words and Their Implication for Phylogenetic Reconstruction.” *Journal of Language Evolution* 1 (2): 119–36. https://doi.org/10.1093/jole/lzw006.

Mallory, James P., and Douglas Q. Adams. 2006. *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World*. Oxford; New York: Oxford University Press.

Wodtko, Dagmar, Britta Irslinger, and Carolin Schneider, eds. 2008. *Nomina im indogermanischen Lexikon*. Heidelberg: Winter.