Tiago Tresoldi (Max-Planck-Institut für Menschheitsgeschichte, Jena)

"CompIE Database of Comparative Indo-European Material"

Etymological studies are generally published in lexicographic prose, relying on bibliographies for cross-referencing (see, for examples, Mallory & Adams, 2006, and Beekes & van Beek, 2009). Authoritative sources tend to be available in unstructured formats only, with on-line resources in most cases either unstructured, commendable but limited in scope or intended for phylogenetics, criticized in their theoretical assumptions, or academically unsuitable. Extensive labor is necessary for aggregating information such as competing reconstructions, imposing limits on data sharing and comparison (Forkel et al., 2018; Rzymski, Tresoldi et al., 2020).

CompIE is a proposal for a database cross-referencing authoritative Indo-European comparative material. It orients towards inter-operable and open data by adopting tools from the CLDF Initiative (Forkel et al., 2018; Forkel et al., 2019) and the relational model, with data as tuples of relations defined in consistence with first-order predicate logic (Codd, 1970). Information is collected in textual long-table format, editable with common spreadsheet programs and suitable for documenting changes. Entries are cross-referenced with internal and external catalogs (Anderson et al., 2018; List et al., 2016; Hammarström et al., 2019; Rzymsky, Tresoldi et al., 2020), allowing composite queries and data validation. The emphasis on referencing and the mandatory bibliographic field orients the proposal as a tool for linking, and not replacing, authoritative data (Wilkinson et al., 2016).

The core of the database are collections of cognates linked through identifiers, listing sources where scholars can obtain additional, supporting, or opposing material:

ID	PIE-ID	Language	Form	Segments	Source
HT32	PIEyoke	Hittite	𒅗𒌅𒀸	j u k a n	DeVaan08
AT567	PIEyoke	Attic	ζυγόν	z d y g ó n	DeVaan08,Liddell40
PG9	PIEyoke	Proto_Germanic	*juką	j u k ã	Kroonen13
IL448	PIEyoke	Imperial_Latin	jugum	j u g u m	DeVaan08
PT24	PIEyoke	Portuguese	jugo	ʒ u ɣ u	Aurelio87

By adopting meaningless unique identifiers, the proposal takes an agnostic view in face of competing information and facilitates internal catalog referencing:

ID	JPM	Rix	Concept	Morphemes
PIEyoke	*yugóm	*Hyugóm	YOKE	BTjoin CCnonab

Internal reconstruction is accounted for by extending models for the annotation of word-formations, still with partial support for processes such as apophonies and reduplication:

ID	Pokorny	Watkins	Rix	Gloss	Semantic_Field
BTjoin CCnoab	*i̯eu̯-	*yeug- *-om	*Hyug- *-om	join non-ab. nom. sing.	Transport Grammar

The proposal extends sequence alignment, as used in phylogenetics for automatic cognate detection (List, 2014, Kilani, forth), to support non-phonological alignment, with algorithms for identifying patterns of correspondence that experts can correct. The segmented sequences can later be filtered with multi-tiered representations:

ID	Language	Root	Form	Alignment
IE345	PIE	IE-deh3	*dh₃m̥h₁nós	- - d h m h n ó s
AT140	Attic	AT-didoomi	δῐδόμενος	δ ῐ δ ό μ ε ν ο ς

Multi-sequence alignments of any kind of corresponding sequence are possible, including orthographies, reconstructions, and pronunciations:

ID	Language	Field	Alignment
PIEyoke	PIE	Reconstruction	- H y u g ó m
AT567	Proto_Germanic	Reconstruction	- - j u k ã -
HT32	Hittite	Orthography	- - 𒄿 𒌅 𒅗 - -
AT567	Attic	Pronunciation	z d y - g o n
IL448	Imperial_Latin	Pronunciation	- - j u g u m

along with partial cognacy alignment:

Language	Alignment			Morphemes
PIE	[dh é h1 s]	-	[t o s]	ROOT-4 SUFFIX-5
Imperial_Latin	[f eː - s]	-	[t u s]	ROOT-4 SUFFIX-5

Language	Alignment			Morphemes
Attic	[tʰ e - -]	[o]	[s]	ROOT-4 THEM_VOWEL SUFFIX-3

The long-table format facilitates the referenced and structured record of alternative information, such as competing etymologies, once more cross-linking etyma and glosses:

PIE	Concept	Etymon	Gloss	Source
*hₐr̥t-ḱo-	BEAR	*h₂rétĥk-	destroyer	MalloryAdams: 55
*hₐr̥t-ḱo-	BEAR	*h₂ŕ̥tĥko-	(sharp-)pawed	GJP

A proof-of-concept of the proposal has been developed internally, with automatic deployment to a website when changes are committed, and is being used for data exploration, notably for morphology. At the time of submission it counts circa 2,500 Proto-Indo-European entries, 10,000 reflexes in over 100 languages, along with dozens of morphological annotations and alignments. It will be made available before presentation as a public website offering downloadable data.

References

Anderson, C. et al. (2018). A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting* 4, pp. 21–53.

Beekes, R. P. & van Beek, L. (2009). *Etymological Dictionary of Greek*. Leiden/ Boston: Brill.

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM Classics* 13 (6): pp. 377–87.

Forkel, R. et al. (2018) Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Sci. Data* 5, pp. 1–10.

Forkel, R. & Bank, S. (2019). *CLLD: A toolkit for Cross-Linguistic Databases*. Version 5.0. Jena: Max Planck Institute for the Science of Human History.

Kroonen, G. (2013). *Etymological Dictionary of Proto-Germanic*. Leiden/Boston: Brill.

Hammarström, H., Haspelmath, M. & Forkel, R. (2019) Glottolog. Version 4.1.

Zenodo. (Available online at http://glottolog.org, Accessed on 2020-01-20.)

Kilani, M. (forth). FAAL: a Feature-based Aligning Algorithm. *Language Dynamics and Change*.

Liddell, H. G., Scott, R., Jones, H. S., & McKenzie, R. (1940). *A Greek-English lexicon*. Oxford: Clarendon Press.

List, J.-M. (2014). *Sequence comparison in historical linguistics.* Düsseldorf : Düsseldorf University Press.

List, J.-M., Cysouw, M. & Forkel, R. (2016= Concepticon. A resource for the linking of concept lists. In Calzolari, N. *et al*. (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2393–2400 (European Language Resources Association).

Mallory, J. P., Adams, D. Q. (2006). *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World*. Oxford: OUP Oxford.

Rzymski, C., Tresoldi, T. *et al.* (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Sci Data* 7, 13.

De Vaan, M. (2008). *Etymological Dictionary of Latin and the other Italic Languages*. Leiden/Boston: Brill.

Schweikhard N. E. (2018). “Enhancing morphological annotation for internal language comparison,” in *Computer-Assisted Language Comparison in Practice*, 10/10/2018, https://calc.hypotheses.org/570.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3.