Dialexification: A tool for studying cross-linguistic patterns of semantic change

Alexandre François[1] & Siva Kalyan[2]
[1]Lattice–CNRS; A.N.U. – alexandre.francois@ens.fr
[2] A.N.U.; Univ. of Queensland – siva.kalyan@anu.edu.au

Keywords: lexicon, etymology, semantic change, historical linguistics, lexical typology

In order to decide whether two words with different meanings are cognate, historical linguists must be able to assess the likelihood of the semantic changes that might link the two meanings historically. While the general cognitive mechanisms behind semantic change are well-understood (e.g. Sweetser 1990, Traugott & Dasher 2002), we still lack an empirical catalogue of attested semantic changes across the world's languages, which linguists could turn to for guidance when judging the cognacy of words.

One could propose to use synchronic COLEXIFICATION (François 2008, 2022) as a proxy for likelihood of semantic change. That is, if two senses A and B are close enough to be frequently "colexified" (expressed by a single polysemous word), we may expect that over time, a word with sense A is likely to acquire sense B, or vice versa. If so, a weighted colexification network of the sort provided by CLiCS (Rzymski et al. 2020) could serve as a preliminary catalogue of likely semantic changes. However, in practice, meanings that are related historically are not always attested as colexified pairs: e.g. the cognate pair {Latin *hortus* 'garden' – Greek χόρτος *khórtos* 'food'} points to a semantic link <garden>–<food> that is not attested, to our knowledge, as a synchronic colexification.

We address this issue by introducing the novel concept of "DIALEXIFICATION" (short for "diachronic colexification"). Two meanings are "dialexified" if they are attached to words from the same *cognate set* – that is, to descendants of the same etymon. For example, descendants of the PIE root *g^{h}erd^{h}- 'enclose' include such meanings as 'belt' (Old Norse *gjǫrð*), 'fence' (Albanian *gardh*), 'yard' (Old Norse *garðr*), 'garden' (German *Garten*), 'earth' (Scots *yird*), 'region' (Old English *ġeard*), 'estate' (Danish *gård*), 'castle' (Czech *hrad*), 'city' (Russian город *gorod*), 'house' (Romani *kher*), 'family' (Bengali ঘর *ghor*), and 'wife' (Sanskrit गृह *gṛhá*). By targeting cognate sets rather than lexemes, dialexification can capture a broader range of semantic connections than synchronic colexification alone.

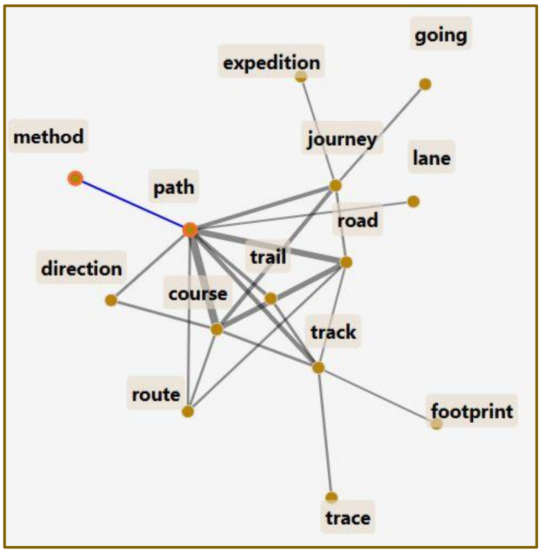
Crucially, certain dialexifications are attested repeatedly across the world's languages. For each pair of senses, the number of different etymons that dialexify them can be taken as a reliable indicator of their semantic proximity, and thus, of the likelihood that a word expressing one sense can eventually express the other. We report here on our efforts to build *EvoSem*, a cross-linguistic database of dialexifications, assembled from open-access online etymological resources. We have begun with the collaborative online dictionary Wiktionary (English version: https://en.wiktionary.org/), for three major language families; Fig. 1 shows the current state of our database. We also plan to include data from Austronesian, based on the *Austronesian Comparative Dictionary* (Blust & Trussel 2013) – as well as other language families for which online comparative dictionaries are available.

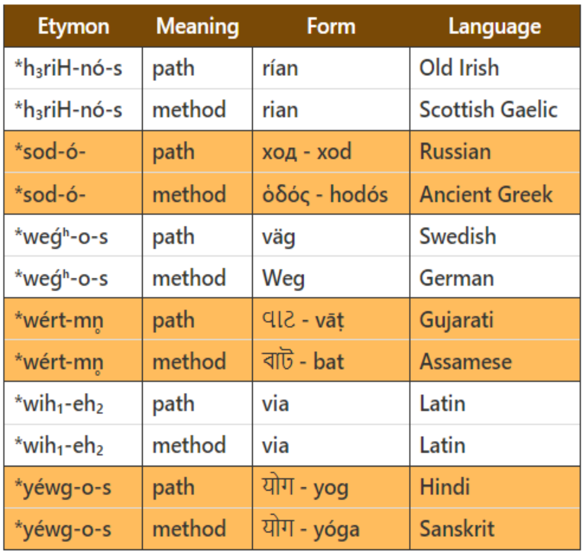
Based on this database, our interface (not yet public) can produce weighted dialexification graphs (Fig. 2a), where links are drawn between the most frequently dialexified pairs of meanings. The thickness of lines is proportional to how frequently each connection is attested (by different cognate sets), and thus how likely it is to constitute a pathway of semantic change. A table is produced dynamically to illustrate each case of dialexification, showing cognate forms and their shared etymon (Fig. 2b).

In sum, we hope to provide both a new conceptual tool (*dialexification*) and a growing database (*EvoSem*) to support empirically-grounded work in comparative linguistics.

	Indo-European	Semitic	Uralic
# source lemmas and roots in the highest proto-language	1,304	196	292
# reflexes in descendant languages	62,930	1,855	2,854
# languages covered, including intermediate proto-languages	650	139	122
# languages covered, excluding proto-languages	620	138	111
# distinct meanings covered	21,736	2,714	1,749

Fig. 1: Statistics of the EvoSem database (under construction) as of March 2023. Etymological data extracted from the Wiktionary collaborative lexical database.





Etymon	Meaning	Form	Language
*h$_3$riH-nó-s	path	rían	Old Irish
*h$_3$riH-nó-s	method	rian	Scottish Gaelic
*sod-ó-	path	ход - xod	Russian
*sod-ó-	method	ὁδός - hodós	Ancient Greek
*weǵ^h-o-s	path	väg	Swedish
*weǵ^h-o-s	method	Weg	German
*wért-mn̥	path	વાટ - vāṭ	Gujarati
*wért-mn̥	method	বাট - bat	Assamese
*wih$_1$-eh$_2$	path	via	Latin
*wih$_1$-eh$_2$	method	via	Latin
*yéwg-o-s	path	योग - yog	Hindi
*yéwg-o-s	method	योग - yóga	Sanskrit

Fig. 2: A weighted dialexification graph built around the notion PATH, *showing the number of cognate sets supporting various semantic links. In the current database, the dialexification* ‹PATH – METHOD› *is attested under six etymons – as displayed in the table (right).*

References

Blust, Robert & Stephen Trussel. 2013. The Austronesian Comparative Dictionary: A work in progress. *Oceanic Linguistics* 52(2), 493–523.

François, Alexandre. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Martine Vanhove (ed.), *From Polysemy to Semantic change: Towards a Typology of Lexical Semantic Associations* (Studies in Language Companion Series), vol. 106, 163–215. New York, Amsterdam: Benjamins.

—— 2022. Lexical tectonics: Mapping structural change in patterns of lexification. *Zeitschrift für Sprachwissenschaft* 41/1: 89–123. DOI:10.1515/zfs-2021-2041.

Rzymski, Christoph, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus Bodt, Abbie Hantgan, Gereon Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell Gray, Robert Forkel & Johann-Mattis List. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data* 7(1). 13.

Sweetser, Eve. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure* (Cambridge Studies in Linguistics 54). Cambridge: Cambridge University Press.

Traugott, Elizabeth & Richard Dasher. 2002. *Regularity in semantic change*. (Cambridge Studies in Linguistics 96). Cambridge: Cambridge University Press.