Matthew Scarborough
matthew.scarborough@cantab.net

SPLITTING HAIRS OVER THE LEXICON
ISSUES IN UTILISING LEXICAL DATA FOR INDO-EUROPEAN CLADISTIC RESEARCH

Matthew Scarborough
Max Planck Institute for the Science of Human History

§1. Introduction:

- 'Current' Issues: Current, perhaps, but certainly not new.
- Cognate coding for lexicostatistical studies may seem simple in principle.
- In practice is quite messy for many reasons.

§2. Context: The *Cognacy in Basic Lexicon: Indo-European* (CoBL-IE) Project

- Database of Cognacy Judgements Across Indo-European Languages, as successor to IELex Database used in BOUCKAERT et al. (2012), CHANG et al. (2015)[1]
- Basic cognacy policy: Tracing etymologies as far back as possible, ideally to an IE root etymology.
- Customised Jena 200 comparison list drawn from the IELex 207 + Leipzig-Jakarta lists[2]

§3. Methodological Problems of Deciding Between Etymologies (Cognacy Policy)

- "[I]f we are to maintain scientific rigor, we must reject etymologies that are attractive but flawed." (RINGE 2017:2; cf. similarly RINGE 1996:xvi-xvii)

§3.1. Issue 1: Hypercriticism over forms in closely related varieties

- In the absence of evidence to the contrary, assume common inheritance rather than borrowing or independent innovation (cf. HENNIG 1966:121).

§3.2. Issue 2: Uncertainties in cross-branch etymologies: some BAD examples

(1) Hitt. *idālu-* 'bad, evil, evilness' Luw. *āddu̯ā̯li-*, TochB *yolo* 'bad, evil' < ?**h₁ed-u̯ol-*.[3]

(2) Gk. κακός 'bad' < Proto-Greek **kako-*, Alb. *keq* 'bad' < Proto-Albanian **kaki̯a/ā-*[4]

- Methodologically best solution is to split uncertain cognate sets (cf. RINGE 2017:2 above).
- CoBL has developed a new 'proposed cognacy' system to index levels of scholarly disagreement over disputed etymologies.

[1] IELex: http://ielex.mpi.nl
[2] Cf. TADMOR (2009), HASPELMATH & TADMOR (2009).
[3] Cf. SCHINDLER (1975), RASMUSSEN (1984:144-145), PUHVEL (1984:487-493), KLOEKHORST (2008:420-422), ADAMS (2013:555-556).
[4] Cf. FRISK (1960–1972:758-759), CHANTRAINE (1968–1980:482), BEEKES (2010:619-620), HULD (1984:79-80), DEMIRAJ (1997:216-217), OREL (1998:175), SCHUMACHER & MATZINGER (2013:223, 239).

Matthew Scarborough
matthew.scarborough@cantab.net

§4. Practical Problems in Coding and Cognacy

§4.1. Indo-European Specific Problems

§4.1.1. Root Extensions and *s*-mobile: The same root etymology or different?

(3) 2. *(*s*)*ker-* 'scheren, kratzen, abschneiden': ON *skera*, OE *scriþ* (LIV² 556-557, IEW 938-40, but possibly extended *(*s*)*kerH-* LIV² 558, cf. Lith. *skìrti* 'trennen, teilen, unterscheiden', so KROONEN 2013:443-444).

(4) *(*s*)*kert-* '(zer)schneiden': YAv. *kərəṇtaiti*, OPers. **kart-*, Ved. *k̥rt-* (LIV² 559-560, IEW 941-2).

➢ Practice determining root etyma basically follows LIV², but perhaps requires justification.

§4.1.2. Taboo Deformation

(5) TONGUE from PIE **dn̥ĝʰu̯éh₂* : TochA *käntu*, TochB *kantwo*, Arm. *lezow*, Av. *hizuua*, Ved. *jihvā́*, OCS ѩзыкъ, Lith. *liežùvis*, OPr. *insuwis*, Goth. *tuggō*, Osc. **fangvam**, Lat. *lingua*, OIr. *tengae*.[5]

(6) ANT from PIE ?**mor̥u̯i-* : TochB *warme**, Gk. μύρμηξ, Arm. *mrǰiwn*, YAv. *maoiri-*, Ved. *vamrá-*, OCS мравии, ON *maurr*, Lat. *formīca*, OBret. *morion*.[6]

(7) LOUSE (IEW 692: **lū̆s* (gen. **lu̯u̯-ós*) 'Laus') → to be split into four different cognate sets:
   a. Germanic and Celtic ?**lū̆-* : ON *lús*, OE *lūs*, OS *lûs*; MW *lleuen*, OBret. *louenn*
   b. Ved. *yūka-* (> MIA, NIA forms, e.g. Pāli *ūkā*, Hindi *jūm̐*, Bengali *ukun*, Nepali *jumro*, etc., cf. EWAia II:415, TURNER 1962-1966:608).
   c. PSlav. **vŭš-* (SCr. *uš*, Maced. вошка, Russ. вошь, Ukr. воша, Pol. *wesz*, Cz. *veš*, Slk. *voš*, cf. DERKSEN 2008:532).
   d. Baltic forms: Lith. *utėlẽ*, Latv. *uts*, Latgal. *vuts* (to be connected with Slavic forms according to FRAENKEL 1962-1965:1173; against connection with Slavic forms cf. DERKSEN 2008:532).

§4.2. Cognacy Coding Problems that are (probably) Universal

§4.2.1. Meanings in High-frequency Grammatical Words

(8) BECAUSE in Slavic: OCS занѥ(же), понѥ(же); Bulg. защото; Maced. затоа што; Russ. потому что; Uk. тому що, Cz., Slk. *preto-že*

[5] Cf. MALLORY & ADAMS (2006:175), MARTIROSYAN (2010:307-308), ADAMS (2013:147), EWAia I:591-593, DE VAAN (2008:343), DERKSEN (2008:159), DERKSEN (2015:285), LEHMANN (1986:349), MATASOVIĆ (2009:368).
[6] Cf. MATASOVIĆ (2009:278), BEEKES (2010:982), EWAia II:507, ESSJa 19:241-249, DERKSEN (2008:326) MARTIROSYAN (2010:482-483), ADAMS (2013:630), DE VRIES (2000:380).

Matthew Scarborough
matthew.scarborough@cantab.net

(9) BECAUSE in Hellenic: AGk. διότι (= διά + ὅ + τι), ἕνεκα; SMG γιατί (για + τί), επειδή (ἐπεί + δή), SMG διότι; Pontic επειδήσκαι (επει + δη + (σ?) + και), Cappadocian ασο (ας + το), Cypriot επειδή (επει + δη), Tsakonian γιατὅί (= SMG γιατί), Italiot τι.

- Some grammatical words apparently more stable:

(10) **ne* 'not' : Hitt. *natta*, Alb. *nuk*, Av. *nōiṯ*, Ved. *ná*, OCS не, Lith. *ne-*, Goth. *ni*, Lat. *nōn*, OIr. *ní* (cf. *LIPP* Vol.2 530-549 s.v. 1. **né* 'nicht')

- But problems still arise: When does a cognate cease to be cognate?

(11) NOT in Hellenic:[7]
  a. Ancient varieties: Myc. *o-u-*, AGk. οὐ(κ), NT οὐ(κ)
  b. Modern Varieties: SMG δεν, Tsak. δε(ν), ο-, ου- (prefix), Capp. δεν (< Gk. οὐδέν = οὐ + δέ + ἕν 'and not one', cf. ANDRIOTIS et al. 1999, BABINOTIS 2010:338)

- Should one keep coding 'ghost morphemes' that have been lost? Arguably no complete lexical replacement has taken place.
- Similar problems for other grammatical words, pronouns, and demonstratives.
- Proposed solution: elimination of pronouns, demonstratives, and high-frequency grammatical words from comparison lists.

§4.2.2. Meanings Prone to Onomatopoeia ± Sound-Symbolism(?)

(12) SPIT from PIE **spṭ̂ieu̯H-* 'spucken, speien'? : Lat. *spuī* (pf.), Lith. *spiáuti*, OCS плю̑вати, Goth. *speiwan*, Gk. πτύω (LIV² 583-584, IEW 999-1000)

(13) CUT:
  a. PIE 2. **kers-* '(ab)schneiden' : Luw. *karš-*, TochA *kärṣ-*, TochB *kars-* (LIV² 355-356; MALZAHN 2010:582-583)
  b. PIE **kʷer-* '(ab)schneiden, schnitzen' : Hitt. *kuer-*/*kur-*, Luw. *k(u)u̯ar-*/*kūr-* Lyc. *xurzei-* (LIV² 391-392, KLOEKHORST 2008:486)
  c. PIE **sekH-* 'abtrennen' : Lat. *secāre*, Umbr. **prusekatu** (IIa 28, etc.), PSlav. **sěkti* > SCr. sjeći, Maced. сече (LIV² 524, DE VAAN 2008:550-551, DERKSEN 2008:446)
  d. PIE *(*s*)*ker-* 'scheren, kratzen, abschneiden' : ON *skera*, OE *scriþ* (LIV² 556-557; or to LIV² 558 *(*s*)*kerH-*, cf. Lith. *skìrti* 'trennen, teilen, unterscheiden', so KROONEN 2013:443-444)
  e. PIE *(*s*)*kert-* 'zerschneiden' : YAv. *kərəṇtaiti*, OPers. **kart-*, Ved. *kr̥t-* (LIV² 559-560)
  f. PIE ?*(*s*)*kelp-* : Marathi *kāpaṇe*, Sinhalese *kapanavā* < Skt. *kalpáyati* 'sets in order; trims, cuts' (TURNER 1962-1966:150; according to EWAia I:232-233 perhaps to be connected with Lat. *scalpere* 'to cut, scrape, scratch', Goth. *halba* 'half')

[7] I follow the reserved judgment of CLACKSON (1994:158), CLACKSON (2004/2005:155-156), and MARTIROSYAN (2010:531), who see Arm. *oč'* more likely as an inner-Armenian creation based on the simple pronoun *o-* (cf. *o-k'* and *o-mn* 'someone') + simple negative *č'* < **kʷid*. As such I regard Gk. οὐ(κ) to have no secure etymology outside of Greek and not cognate with PIE **né* (cf. COWGILL (1960) arguing from **ne h₂oi̯u* (*kʷid*)).

Matthew Scarborough
matthew.scarborough@cantab.net

g. PIE 1. *(*s*)*kep*- 'hacken, hauen' : Gk. κόπτω (LIV² 555, with Balto-Slavic, Albanian cognates. BEEKES 2010:748-749 prefers to see 'Pre-Greek')

(14) SCRATCH:[8]

a. PIE 2. *(*s*)*ker*- 'scheren, kratzen, abschneiden' : Arm. *kʻerem* (LIV² 556-557, MARTIROSYAN 2010:662-663)
b. PIE **kes*- 'ordnen' : OCS *česati*, Latv. *kasît* (LIV² 357, DERKSEN 2008:86, DERKSEN 2015:231)
c. PIE **kseu̯*- 'schaben, schliefen': Mod.Gk. ξύνω (< AGk. ξύω), Hindi *khuracanā* (cf. Skt. *kṣuráti*) (LIV² 372, BEEKES 2010:1039-1040, TURNER 1962-1966 3729, EWAia I:435-436.)
d. PIE **ksneu*- 'schärfen' : Ved. *kṣṇav*- (LIV² 373, EWAia I:441, cf. Lat. *novācula* 'Rasiermesser')
e. PIE **skab*ʰ- 'kratzen, schaben' : Lat. *scabere*, OS *skaƀan* (LIV² 549, DE VAAN 2008:541, KROONEN 2013:438, cf. Lith. *skõbti* 'to plane')
f. PIE *(*s*)*kreb*- 'schaben, kratzen' : OE *screpan*, MW *crauu* (LIV² 562, cf. Orel 2003:344)
g. PIE *(*s*)*kerp*- 'abschneiden, abrufen' : Lith. *krapštýti* (LIV² 559)
h. PIE *(*s*)*kep*- : Rusyn šǐpatı (according to our Slavicist Lechosław Jocz; I have been unable to confirm this root etymology.)
i. PGmc. **krat*- > Germ *kratzen*, etc. (EWA 5:762-764)

§5. Some Outstanding Anomalies

§5.1. Parallel Semantic Shift / Derivation

(15) CHILD derivations from PIE **d*ʰ*eh*$_1$(*i̯*)- '(Muttermilch) saugen' : Lyc. *tideimi*- 'child'; Sp. *hijo*, It. *figlio* 'child', etc. < Lat. *filius* 'son' (LIV² 138-139, KLOEKHORST 2008:875-877, NEUMANN 2007:359-360, DE VAAN 2008:219).

(16) GIVE based on PIE **b*ʰ*er*- 'tragen' : Khot. *heḍä*, Sogd. *ϑbr*- < PIr. **fră-bar*-; OIr. *do·beir* < PCl. **to-ber*- (BAILEY 1979:499, CHEUNG 2007:6-10, PEDERSEN 1913:469-471 MATASOVIĆ 2009:62).

§5.2. Semantic calquing

- In preliminary results Modern Greek Tsakonian (from West Greek) subgroups with other Modern Greek dialects descended from the Attic-Ionic *koiné*.

- I suspect semantic calquing / parallel semantic developments probably to blame here, but also good intermediate data for West Greek dialects is also lacking.

- Exactly what is happening with the data requires further investigation, but I suspect this is likely to be a situation (lack of good intermediate documentation + sociolinguistic situation favouring convergence in the lexicon) where this methodology does not work.

[8] Cf. stock sound effects for shovelling (https://www.youtube.com/watch?v=xOxBGcZROoM) and scraping: (https://www.youtube.com/watch?v=vPtSEPhBuWk)

Matthew Scarborough
matthew.scarborough@cantab.net

§6. Some Conclusions

1. Comparison meanings need to be carefully selected to minimise bad data in lexicostatistics:

   - **Generally:** Comparative wordlists of basic vocabulary need to be easy to elicit so that lexemes to ensure that comparable semantics are being compared across all languages.

   - **More Specifically:** Comparative wordlists have to consist of lexemes that are also easy to encode consistently.

2. Contingencies need to be designed into database structures in order to deal with issues of unclear or partial cognacy, with a principled decision-making process or cognacy coding policy in place in order to deal with disputed etymologies that do not lend themselves to binary true/false coding decisions.

References:

ADAMS, Douglas Q. 2013. *A Dictionary of Tocharian B* (*2nd ed.*). Amsterdam: Rodopi.

ANDRIOTIS, Nikolaos (ed.) 1999. *Λεξικό της κοινής νεοελληνικής* [Dictionary of Standard Modern Greek]. Θεσσαλονίκη: Ινστιτούτο Νεοελληνικών Σπουδών.

BABINOTIS, George. 2010. *Ετυμολογικό λεξικό της νέας ελληνικής γλώσσας* [Etymological Dictionary of the Modern Greek Language]. Αθήνα: Κέντρο λεξικογραφίας.

BAILEY, Harold W. 1979. *A Dictionary of Khotan Saka*. Cambridge: Cambridge University Press.

BEEKES, Robert S. P. 2010. *An Etymological Dictionary of Greek*. Leiden: Brill.

BOUCKAERT, Remco, Philippe LEMEY, Michael DUNN, Simon J. GREENHILL, Alexander V. ALEKSEYENKO, Alexei J. DRUMMOND, Russell D. GRAY, Marc A. SUCHARD & Quentin D. ATKINSON. 2012. "Mapping the origins and expansion of the Indo-European language family" *Science* 337 : 957-60.

CHANG, Will, Chundra CATHCART, David HALL & Andrew GARRETT. 2015. "Ancestry-constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis" *Language* 91 : 194-244.

CHANTRAINE, Pierre. 1968-1980. *Dictionnaire étymologique de la langue grecque*. Paris: Klinksieck.

CLACKSON, James. 1994. *The Linguistic Relationship Between Armenian and Greek*. Oxford: Blackwell.

CHEUNG, Johnny. 2007. *Etymological Dictionary of the Iranian Verb*. Leiden: Brill.

CLACKSON, James. 2004/2005. Review of Kortlandt, Frederik. 2003 *Armeniaca: Comparative Notes, with an Appendix on the Historical Phonology of Classical Armenian by Robert Beekes*. Ann Arbor: Caravan Books. *Annual of Armenian Linguistics* 24-25 : 153-58.

CoBL-IE: *Cognacy in Basic Lexicon – Indo-European*. Jena: Max Planck Institute for the Science of Human History. (Project co-ordinators: Paul HEGGARTY & Cormac ANDERSON. Database authored with the co-operation of many scholars. Forthcoming 2017 at http://www.cobl.info).

COWGILL, Warren. 1960. "Greek *ou* and Armenian *oč*'." *Language* 36 : 347-50.

DE VAAN, Michiel. 2008. *Etymological Dictionary of Latin and the other Italic Languages*. Leiden: Brill.

DE VRIES, Jan. 2000. *Altnordisches etymologisches Wörterbuch*. Leiden: Brill.

DEMIRAJ, Bardhyl. 1997. *Albanische Etymologien*. Amsterdam: Rodopi.

DERKSEN, Rick. 2008. *Etymological Dictionary of the Slavic Inherited Lexicon*. Leiden: Brill.

DERKSEN, Rick. 2015. *Etymological Dictionary of the Baltic Inherited Lexicon*. Leiden: Brill.

DYEN, Isidore, Joseph B. KRUSKAL & Paul BLACK. 1992. "An Indoeuropean Classification: A Lexicostatistical Experiment" *Transactions of the American Philosophical Society* 82 : iii-iv+1-132.

ESSJa 19 = TRUBAČEV, O. N. 1992. *Этимологический словарь славянских языков. Выпуск 19* (**męs⁽ⁱ⁾arъ — *morzakъ*) [Etymological Dictionary of the Slavic Languages. Volume 19]. Москва: Наука.

EWA 5 = Lühr, Rosemarie, Harald BICHLMEIER, Maria KOZIANKA & Roland SCHUHMANN. 2014. *Etymologisches Wörterbuch des Althochdeutschen* (*Bd. 5*). Göttingen: Vandenhoeck & Ruprecht.

EWAia = MAYRHOFER, Manfred. 1992-2001. *Etymologisches Wörterbuch des Altindoarischen* (3 vols.) Heidelberg: Carl

Matthew Scarborough
matthew.scarborough@cantab.net

Winter Universitätsverlag.

FRAENKEL, Ernst. 1962-1965. *Litauisches Etymologisches Wörterbuch*. Heidelberg: Carl Winter Universitätsverlag.

FRISK, Hjalmar. 1960-1972. *Griechisches Etymologisches Wörterbuch*. Heidelberg: Winter Verlag.

HASPELMATH, Martin & Tadmor, Uri (eds.) 2009. *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wold.clld.org, Accessed on 2017-08-30.)

HENNIG, Willi. 1966. *Phylogenetic Systematics*. Urbana: University of Illinois Press.

HULD, Martin. 1984. *Basic Albanian Etymologies*. Columbus, OH: Slavica Publishers.

IELex: *Indo-European lexical cognacy database* (IELex). Nijmegen: Max Planck Institute for Psycholinguistics. (Available online at http://ielex.mpi.nl, Accessed on 2017-08-30.)

IEW = Pokorny, Julius. 1959-1969. *Indogermanisches Etymologisches Wörterbuch* (*3 Vols.*). Bern und München: Francke Verlag.

KLOEKHORST, Alwin. 2008. *Etymological Dictionary of the Hittite Inherited Lexicon*. Leiden: Brill.

KROONEN, Guus. 2013. *Etymological Dictionary of Proto-Germanic*. Leiden: Brill.

LEHMANN, Winfred P. 1986. *Gothic Etymological Dictionary*. Leiden: Brill.

LIPP = DUNKEL, George. 2014. *Lexikon der indogermanischen Partikeln und Pronominalstämme. Band 1: Einleitung, Terminologie, Lautgesetze, Adverbialendungen, Nominalsuffixe, Anhänge und Indices, Band 2: Lexikon*. Heidelberg: Universitätsverlag Winter.

LIV² = RIX, Helmut, Martin Joachim KÜMMEL, Thomas ZEHNDER, Reiner LIPP & Brigette SCHIRMER. 2001. *Lexikon der indogermanischen Verben: Die Wurzeln und ihre Primärstammbildungen* (*2e Aufl.*). Wiesbaden: Reichert Verlag.

MALLORY, James Patrick & Douglas Q. ADAMS. 2006. *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World*. Oxford: Oxford University Press.

MALZAHN, Melanie. 2010. *The Tocharian Verbal System*. Leiden: Brill.

MARTIROSYAN, Hrach K. 2010. *Etymological Dictionary of the Armenian Inherited Lexicon*. Leiden: Brill.

MATASOVIĆ, Ranko. 2009. *Etymological Dictionary of Proto-Celtic*. Leiden: Brill.

NEUMANN, Günter. 2007. *Glossar des Lykischen* (*Überarbeitet und zum Druck gebracht von Johann Tischler*). Wiesbaden: Harrassowitz.

OREL, Vladimir. 1998. *Albanian Etymological Dictionary*. Leiden: Brill.

OREL, Vladimir. 2003. *A Handbook of Germanic Etymology*. Leiden: Brill.

PEDERSEN, Holger. 1913. *Vergleichende Grammatik der keltischen Sprachen. Zweiter Band: Bedeutungslehre* (*Wortlehre*). Vandenhoeck und Ruprecht: Göttingen.

PERELTSVAIG, Asya & Martin D. LEWIS. 2015. *The Indo-European Controversy: Facts and Fallacies in Historical Linguistics*. Cambridge: Cambridge University Press.

RASMUSSEN, Jens Elmegård. 1984. "Miscellaneous Morphological Problems in Indo-European Languages" *Arbejdspapirer udsendt af Institut for Lingvistik, Københavns Universitet* 4:135-49.

RINGE, Donald A. 1996. *On the Chronology of Sound Changes in Tocharian. Volume 1: From Proto-Indo-European to Proto-Tocharian*. New Haven: American Oriental Society.

RINGE, Donald A. 2017. *From Proto-Indo-European to Proto-Germanic* (*2nd ed.*). Oxford: Oxford University Press.

SCHINDLER, Jochem. 1975. "Armenisch *erkn*, griechisch ὀδύνη, irisch *idu*" *Zeitschrift für vergleichende Sprachforschung* 89 : 53-65.

SCHUMACHER, Stefan & Joachim MATZINGER. 2013. *Die Verben des Altalbanischen: Belegwörterbuch, Vorgeschichte, und Etymologie*. Wiesbaden: Harrassowitz Verlag.

TADMOR, Uri. 2009. "Loanwords in the World's Languages: Findings and Results" In: *Loanwords in the World's Languages: A Comparative Handbook*, ed. by M. Haspelmath & U. Tadmor, 55-75. Berlin: de Gruyter.

TURNER, Ralph Lilley, Sir. 1962-1966. *A Comparative Dictionary of Indo-Aryan Languages*. London: Oxford University Press.

WATKINS, Calvert. 1982. "Notes on the Formations of the Hittite Neuter" In: *Investigationes Philologicae et Comparativae. Gedenkschrift für Heinz Kronasser*, ed. by E. Neu, 250-62. Wiesbaden: Harrassowitz.