IRFD Project Description
Connecting the Dots: Reconfiguring the Indo-European family tree
Thomas Olander

1. Problem

At this very moment the prehistory of Eurasia is being rewritten. The staggering recent advances in ancient genetics and isotope analysis have led to a new view on prehistoric population movements and contacts, and to a reinterpretation of the archaeological record (Kristiansen et al. 2017). An indispensable component of population history is language – and in the case of western Eurasia, primarily the Indo-European (IE) languages. The new evidence from ancient genetics has made it very likely that the spread of IE languages was intimately connected with the spread of material culture and genes during the Late Neolithic and Bronze Age. The early spread and diversification of the IE language family is thus an essential component in this new understanding of our past.

All IE languages – e.g. English, Danish, Spanish, Russian, Persian and Hindi – descend from a common ancestor, Proto-Indo-European (PIE), not attested in writing but reconstructed on the basis of its daughter languages. Around 4000 BCE, more than two millennia before the first appearance of written attestations of any IE language, the proto-language began to dissolve into distinct languages. During the following millennia, the descendants of PIE continued to develop in different directions. When they eventually enter the historical stage one by one through written records, the IE languages fall into ten related, yet clearly distinct branches: Italic, Celtic, Germanic, Greek, Armenian, Albanian, Indo-Iranian, Balto-Slavic and the extinct branches Anatolian and Tocharian.

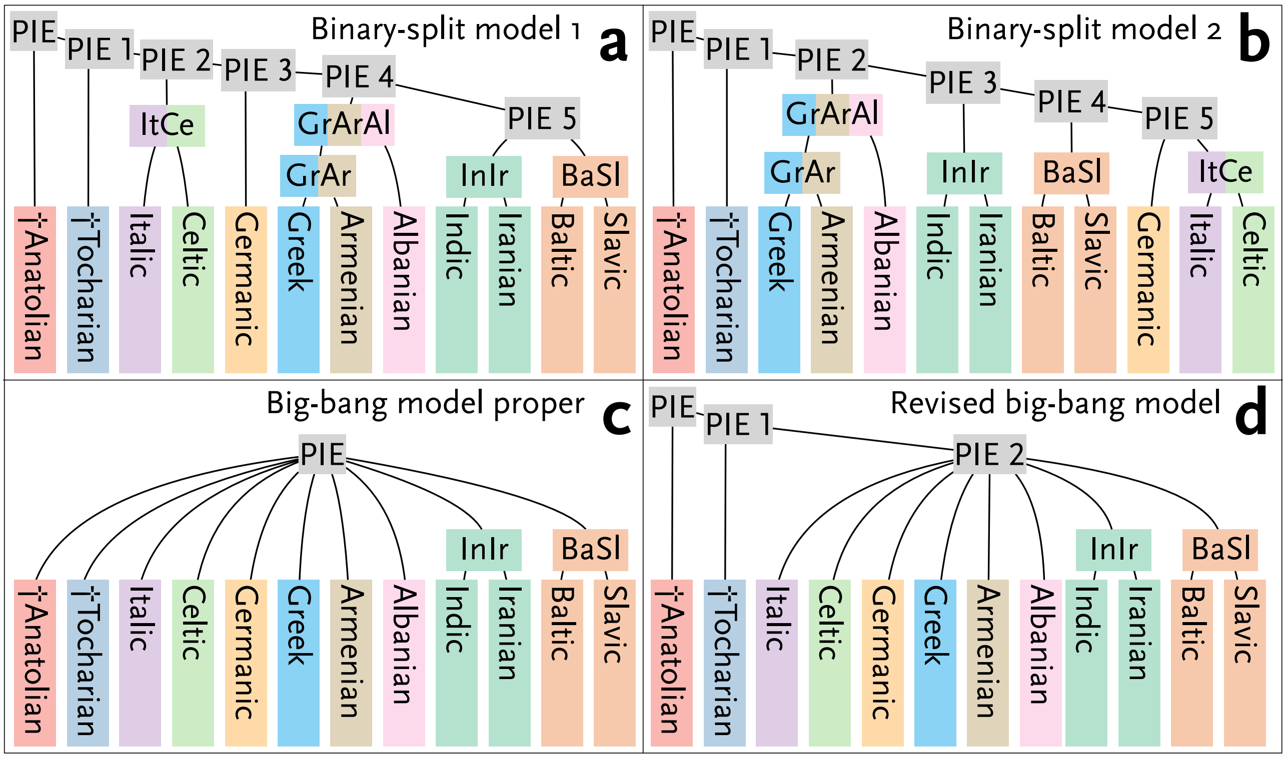

Fig. 1 Four models of the disintegration of IE. Model **a** is based on Nakhleh et al. 2005a; **b** is that of Chang et al. 2015; **c** and **d** are traditional IE models. † = extinct branch. P = “Proto-”.

The branching structure of the IE family tree is disputed: Did the branches split off from the main stem one at a time or in small groups, each time leaving a residual proto-language behind (Models **a** and **b** in fig. 1)? Or did they split off more or less simultaneously in a linguistic "big bang" (Models **c** and **d**)? If the branching-off was gradual, in which order did the branches split off (Model **a** vs. Model **b**)? Since the different models relate different accounts of the early spread of the IE languages and their speakers, how do the models square with the archaeological and genetic record?

The goal of the project is to determine how the early disintegration of the IE language family unfolded and how this matches the dispersals of the speakers. In order to achieve this goal, the project is broken down into three subprojects, each of which aims at providing a missing part in the puzzle of the early spread of IE languages (see **5**). Focusing on (1) a crucial problem in the early branching structure of the IE language family, (2) the methodological aspects and (3) the linguistics–archaeology interface, the subprojects benefit mutually from a continuous exchange of problems and results, enabling us to connect the dots that lead to the integration of language as a key component in the new picture of Eurasian prehistory.

2. Objectives and hypotheses

In the last decades some consensus has been reached that the first branch to split off from the rest was Anatolian, and the second one was Tocharian (Models **a**, **b** and **d**; see e.g. Clackson 2007: 13; Fortson 2015: 646; Olander forthc. a; forthc. b). The next question is thus which branch was the third one to split off: in some scenarios Italo-Celtic is favoured (Model **a**), in others it is Graeco-Armeno-Albanian (Model **b**; note that Albanian is included in this group only for convenience's sake; its position is unclear, see Nakhleh et al. 2005a: 396); still others stick to the more traditional view according to which the remaining branches split off from each other in a big-bang scenario (Model **d**). Objective 1 is to determine which branch was the third one to split off, a prerequisite for determining the structure of the remainder of the tree. The most common way for language families of a certain age to diverge is through binary (or ternary) splits (Nichols 1990: 489; 1997: 138; Fortson 2004/2010: 11); this favours Models **a** and **b** over **d** (and **c**). Moreover, Model **a** (based on Nakhleh et al. 2005a; see also Kortlandt 2016) is more solidly founded than Model **b** (from Chang et al. 2015: 200; cf. Bouckaert et al. 2012) since it is based on significant shared phonological and morphological innovations as well as shared lexical items between various IE languages (see **4**). Our Hypothesis 1 is that the third branch to split off was Italo-Celtic as in Model **a**, implying that there are linguistic innovations in PIE 3 in this model vis-à-vis PIE 2.

Why do researchers arrive at so widely differing family trees as those in fig. 1? The trees are the results of different interpretations of the material, but an even more significant factor is the methodology employed (with implications for the selection of material): Models **a**, **c** and **d** are based primarily on significant shared innovations, whereas Model **b** is based on a statistical analysis of the amount of shared lexical items between IE languages (see **4** for these methods). Objective 2 is to assess and refine the methodology for linguistic phylogenetics, a crucial device for the understanding of the IE family tree. Our Hypothesis 2 is that the most reliable method of linguistic phylogenetics is based on significant shared innovations.

As the structure of the IE family tree may reflect prehistoric migration events, it has recently become the object of interest from geneticists and archaeologists (e.g. Haak et al. 2015; Allentoft et al. 2015, from Eske Willerslev's Centre for GeoGenetics at UCPH; Anthony & Ringe 2015). One of the arguments against a prominent hypothesis which puts the IE homeland in Anatolia in present-day Turkey (Renfrew 1987; Bouckaert et al. 2012) is precisely the branching order: the common innovations of the non-Anatolian branches are hard to explain if these branches spread in different directions from an Anatolian homeland. Objective 3 is to relate the branching structure of the IE family tree to migration events. Our Hypothesis 3 is that the linguistic evidence for the early disintegration of IE will shed light on the partly conflicting and unresolved archaeological and genetic evidence for migrations in Late Neolithic and Bronze Age Eurasia.

3. Impact

With its focus on one of the biggest unsolved problems in the phylogenetics of IE, the "third split", *Connecting the Dots* aims to provide future research with the key to the understanding of the remainder of the family tree. By refining the methodological framework for linguistic phylogenetics, the project contributes to the general theory of language evolution. By statistically correlating linguistic and archaeological data (see **4**) it opens up new vistas of research in this field. By staging language as a key component in the new account of prehistoric Eurasia that is currently being worked out by scholars from other disciplines, it makes this new prehistory directly relevant to a broad audience, as half the world's population speaks an IE language.

4. Methods and concepts

In linguistic phylogenetics two fundamentally different approaches are commonly applied: one based on a qualitative assessment of identical non-trivial innovations in related languages, and another based on a quantitative, statistical analysis of the shared lexicon of the languages.

The first approach relies on the idea that if a significant shared innovation is found in two or more languages, it is most likely to have taken place once, in a common ancestor of these languages (see e.g. Hock 1986/1991: 556–567; Ringe & Eska 2013: 256–263). This methodology makes it possible to determine the degree of relationship among related languages, i.e. to establish subgroups within a language family. In all languages belonging to the Germanic branch, for example, a PIE *p* sound (preserved in Latin *pater, piscis*) has become *f,* as in English *father, fish* and German *Vater, Fisch* etc. Together with several other significant innovations found in all Germanic languages, this sound change is most likely to have occurred once, at a common pre-stage of all Germanic languages, indicating that these languages form a group that descends from a common ancestor, Proto-Germanic. The most reliable features for linguistic phylogenetics are innovations found at the phonological and, especially, morphological levels, whereas syntax and lexicon are less helpful since they are more liable to parallel development and borrowing (Clackson 1994: 6, 17–19; Ringe et al. 2002: 65; Nakhleh et al. 2005a: 395–6; 2005b: 172, 180; Ringe & Eska 2013: 256–263;

Barbançon et al. 2013: 149).

An alternative approach to linguistic phylogenetics is lexicostatistics, which evaluates the relative likelihood for subgroups in a language family on the basis of the amount of shared lexicon across the languages. Most lexicostatistical analyses are based on the “Swadesh list” of around 100 or 200 basic concepts (for IE see e.g. Dyen et al. 1992; Rexová et al. 2003). Extensions of lexicostatistical methods have attempted to use relative rates of lexical replacement in languages in order to estimate approximate dates for the splitting of branches (e.g. Gray & Atkinson 2003; Bouckaert et al. 2012; Chang et al. 2015; cf. also the discussions in Embleton 1986; McMahon & McMahon 2005; Nichols & Warnow 2008).

Both of these approaches evaluate quite different sources of linguistic data for determining linguistic subgroups, and as can be seen in the differences between Model **a** (based on analysis of significant innovations) and Model **b** (based on lexicostatistics), the amount of weight that one places on either methodology has a significant impact on the resulting structure of the IE family tree. Since shared items are a less reliable indicator for genealogical relationship than shared innovations (because shared items may simply be archaisms), we will develop statistical methods based on shared innovations rather than lexicon (as in Model **a**). In order to test the results of the project, they are compared to a model based exclusively on lexicostatistics (Model **b**) and to the big-bang models (Models **c** and **d**) that are current in IE linguistics.

The methodological challenges in correlating reconstructed proto-languages with the archaeological record are immense (see e.g. Dressler 1965; Mallory 1989; 1997). The conventional approach uses the principle of linguistic palaeontology: when a word can be reconstructed back to a proto-language, we may assume that the speakers of the proto-language knew the concept referred to by the word, allowing us to make inferences about the world they lived in (see e.g. Hock & Joseph 1996/2009: 477–509; Campbell 1998/2013: 406–446). This can be a useful device and, applied carefully, has produced key insights (Anthony 2007; Olander 2017; forthc. c). However, due to several well-known constraints the method is often of limited use in practice. *Connecting the Dots* therefore investigates novel, alternative methods of integrating linguistic and archaeological models of population history, including (a) more robust extensions of linguistic palaeontology, (b) the cross-validation of linguistic and cultural phylogenies (cf. Lipo et al. 2005, Mace et al. 2005) and (c) the use of cultural distance analysis (Nakoinz & Knitter 2018: 193–212) to test reconstructed language branching events archaeologically.

References

Allentoft, Morten E., et al. 2015. Population genomics of Bronze Age Eurasia. *Nature* 522(7555). 167–172.

Anthony, David W. 2007. *The horse, the wheel, and language: How Bronze-Age riders from the Eurasian steppes shaped the modern world*. Princeton (NJ) & Oxford: Princeton University Press.

Anthony, David W., & Donald A. Ringe. 2015. The Indo-European homeland from linguistic and archaeological perspectives. *Annual Review of Linguistics* 1. 199–219.

Barbançon, François, et al. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica* 30(2). 143–170.

Bouckaert, Remco, et al. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337. 957–960.

Campbell, Lyle. 1998. *Historical linguistics: An introduction*. Edinburgh: Edinburgh University Press. (3rd ed., 2013.)

Chang, Will, Chundra Cathcart, David Hall & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244.

Clackson, James. 1994. *The linguistic relationship between Armenian and Greek* (Publications of the Philological Society 30). Oxford & Cambridge, MA: Blackwell.

Clackson, James. 2007. *Indo-European linguistics: An introduction*. Cambridge: Cambridge University Press.

Dressler, Wolfgang U. 1965. Methodische Vorfragen bei der Bestimmung der "Urheimat". *Die Sprache* 11(1–2). 25–60, 217.

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.

Embleton, Sheila M. 1986. *Statistics in historical linguistics* (Quantitative Linguistics 30). Bochum: Brockmeyer.

Fortson, Benjamin W., IV. 2004. *Indo-European language and culture: An introduction* (Blackwell Textbooks in Linguistics 19). Malden, MA & Oxford: Blackwell. (Second edition, 2010.)

Fortson, Benjamin W., IV. 2015. Indo-European: Methods and problems. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics,* 645–656. Oxon & New York: Routledge.

Garrett, Andrew. 1999. A new model of Indo-European subgrouping and dispersal. In Steve S. Chang, Lily Liaw & Josef Ruppenhofer (eds.), *Proceedings of the Twenty-Fifth Annual Meeting of the Berkeley Linguistics Society, February 12–15,* 146–156. Berkeley: Berkeley Linguistics Society.

Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426. 435–439.

Haak, Wolfgang, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522(7555). 207–211.

Hock, Hans Henrich. 1986. *Principles of historical linguistics*. Berlin & New York: Mouton de Gruyter. (2. rev. and updated ed., 1991.)

Hock, Hans Henrich, & Brian D. Joseph. 1996. *Language history, language change, and language relationship: An introduction to historical and comparative linguistics* (Trends in Linguistics. Studies and Monographs 93). Berlin: Mouton. (Second revised edition, 2009.)

Hyllested, Adam, Benedicte Nielsen Whitehead, Thomas Olander & Birgit Anette Olsen (eds.). 2017. *Language and prehistory of the Indo-European peoples: A cross-disciplinary perspective* (Copenhagen Studies in Indo-European 7). Copenhagen: Museum Tusculanum.

Kim, Ronald I. 2018. Greco-Armenian: The persistence of a myth. *Indogermanische Forschungen* 123. 247–272.

Kortlandt, Frederik. 2016. Baltic, Slavic, Germanic. *Baltistica* 51(1). 81–86.

Kristiansen, Kristian. 2017. When language meets archaeology: From Proto-Indo-European to Proto-Germanic in northern Europe. In Bjarne Simmelkjær Sandgaard Hansen, Adam Hyllested, Anders Richardt Jørgensen, Guus Kroonen, Jenny Helena Larsson, Benedicte Nielsen Whitehead, Thomas Olander & Tobias Mosbæk Søborg (eds.), *Usque ad radices: Indo-European studies in honour of Birgit Anette Olsen* (Copenhagen Studies in Indo-European 8), 427–437. Copenhagen: Museum Tusculanum.

Lipo, C., M. Collard, M. O'Brien & S. J. Shennan. 2005. *Mapping our ancestors: Phylogenetic approaches in anthropology and prehistory.* London: Taylor and Francis.

Mace R., C. J. Holden & S. J. Shennan. 2005. *The evolution of cultural diversity: A phylogenetic approach.* Walnut Creek, CA: Left Coast.

Mallory, James P. 1989. *In search of the Indo-Europeans: Language, archaeology, and myth*. London: Thames & Hudson.

Mallory, James P. 1997. The homelands of the Indo-Europeans. In Roger Blench & Matthew Spriggs (eds.), *Archaeology and language.* Vol. 1. Theoretical and methodological orientations (One World Archaeology 27), 93–121. London & New York: Routledge.

McMahon, April M. S. & Robert McMahon. 2005. *Language classification by numbers*. Oxford & New York: Oxford University Press.

Nakhleh, Luay, Donald A. Ringe & Tandy Warnow. 2005a. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81. 382–420.

Nakhleh, Luay, Tandy Warnow, Donald A. Ringe & Steven N. Evans. 2005b. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. *Transactions of the Philological Society* 103(2). 171–192.

Nakoinz, O., & D. Knitter. 2018. *Modelling human behaviour in landscapes: Basic concepts and modelling elements.* Cham: Springer.

Nichols, Johanna. 1990. Linguistic diversity and the first settlement of the New World. *Language* 66(3). 475–521.

Nichols, Johanna. 1997. The epicentre of the Indo-European linguistic spread. In Roger Blench &

Matthew Spriggs (eds.), *Archaeology and language.* Vol. 1. Theoretical and methodological orientations (One World Archaeology 27), 122–148. London & New York: Routledge.

Nichols, Johanna, & Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2(5). 760–820.

Olander, Thomas. 2006. *Accentual mobility: The prehistory of the Balto-Slavic mobile accent paradigms.* PhD thesis, University of Copenhagen.

Olander, Thomas. 2009. *Balto-Slavic accentual mobility* (Trends in Linguistics. Studies and Monographs 199). Berlin & New York: Mouton de Gruyter.

Olander, Thomas. 2014. *Proto-Slavic inflectional morphology.* Habilitation thesis, Dept. of Nordic Studies and Linguistics, University of Copenhagen.

Olander, Thomas. 2015. *Proto-Slavic inflectional morphology: A comparative handbook* (Brill's Studies in Indo-European Languages and Linguistics 14). Leiden & Boston: Brill.

Olander, Thomas. 2017. Drinking beer, smoking tobacco and reconstructing prehistory. In Bjarne Simmelkjær Sandgaard Hansen et al. (eds.), *Usque ad radices: Indo-European studies in honour of Birgit Anette Olsen* (Copenhagen Studies in Indo-European 8), 605–618. Copenhagen: Museum Tusculanum.

Olander, Thomas. Forthc. a. Connecting the dots: The Indo-European family tree as a heuristic device. In David Goldstein, Stephanie Jamison & Brent Vine (eds.), *Proceedings of the 29th UCLA Indo-European Conference*, 181–202. Bremen: Hempen. [Accepted.]

Olander, Thomas. Forthc. b. The nomenclature of the Indo-European language family. [Submitted to *Indogermanische Forschungen*.]

Olander, Thomas. Forthc. c. The Indo-European homeland: Introducing the problem. In Olsen, Olander & Kristiansen forthc. [Submitted.]

Olander, Thomas (ed.). Forthc. d. *The Indo-European language family.* Cambridge University Press. [Book proposal accepted.]

Olsen, Birgit A., Thomas Olander & Kristian Kristiansen (eds.). Forthc. *Indo-European languages and society in prehistory.* [Book manuscript submitted to Oxbow Books.]

Renfrew, Colin. 1987. *Archaeology and language: The puzzle of Indo-European origins*. New York: Cambridge University Press.

Rexová, Kateřina, Daniel Frynta & Jan Zrzavý. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19. 120–127.

Ringe, Donald A. 1998. Some consequences of a new proposal for subgrouping the IE family. In Benjamin K. Bergen, Madelaine C. Plauché & Ashlee C. Bailey (eds.), *Proceedings of the Twenty-Fourth Annual Meeting of the Berkeley Linguistics Society, February 14–16, 1998. Special session on Indo-European subgrouping and internal relations,* 32–46. Berkeley: Berkeley Linguistics Society.

Ringe, Donald A., Tandy Warnow & Ann Taylor. 2002. Indo-European and computational linguistics. *Transactions of the Philological Society* 100(1). 59–129.

Ringe, Donald A., & Joseph F. Eska. 2013. *Historical linguistics: Toward a twenty-first century* *reintegration*. Cambridge etc.: Cambridge University Press.

Roe, Joe. 2013. Cultural phylogeny of kurgan burial mounds in the west Eurasian steppe. Unpublished MA dissertation, UCL Institute of Archaeology.

Serangeli, Matilde & Thomas Olander (eds.). Forthc. *Dispersals and diversification: Linguistic and archaeological perspectives on the early stages of Indo-European*. Brill. [Book proposal accepted.]

Weiss, Michael. 2009. *Outline of the historical and comparative grammar of Latin*. Ann Arbor, MI & New York: Beech Stave. (2nd, corrected printing, 2011.)