From phonology to phylogeny

This abstract is for a **talk** in the workshop **Recent advances in computational historical linguistics: New methods and results**.

Language	Lexeme	Class
Latin	*caput*	1
Spanish	*cabeza*	2
Portuguese	*cabeça*	2
Catalan	*cap*	2
Italian	*testa*	3
French	*tête*	3

Table 1: Lexical cognates

Introduction. Linguistic phylogenies are standardly inferred from cognate relationships among lexical items (e.g., [1], [2], [3], [4], [5], [6]), an example of which is presented in Table 1. Shared ancestry among lexemes for 'head' in Latin and Romance are encoded with the values 1, 2, and 3, which denote cognate classes. Despite the prevalence of this approach, it suffers from well-known flaws. First, it ignores segmental information. The words for 'stone' in Table 2 illustrate the problem. Since these words all descend from Latin *petra* 'stone', they belong to the same cognate class and as such are not phylogenetically informative. There is, however, phylogenetic signal in the forms of the words themselves. For instance, the shared development of /ð/ in Portuguese and Spanish provides evidence for Ibero-Romance. Second, the cognate class values 1, 2, and 3 in Table 1 are arbitrary and as such lack consistent reference across cognate sets ([7:602]). Estimated transition rates among cognate-class values are therefore not linguistically meaningful.

Language	Aligned cognate word forms					
Latin	p		e	t	r	a
Portuguese	p		ɛ	ð	ɾ	ɐ
Spanish	p	j	e	ð	ɾ	a
Catalan	p		e	d	ɾ	ə
French	p	j	ɛ		ʁ	
Italian	p	j	ɛ	t	r	a
Romanian	p	j	a	t	r	ə

Table 2: Romance 'stone'

Incorporating segmental information. There is a model capable of handling of segmental information, however—the TKF91 model ([8], [9]). The TKF91 model defines three possible events over aligned sequences of segments: insertions, deletions, and transitions. These are the very processes that give rise to the Romance word forms in Table 2, where we have segmental insertion (e.g., /-j-/ in Spanish); segmental deletion (e.g., loss of the coronal stop in French); and transitions between segments (e.g., /e/ > /ɛ/). Insertions and deletions are modeled as continuous-time birth-death processes while transitions can be modeled in a variety of ways (e.g., with JC69, F81, GTR, or Covarion). In this talk I present the first application of the TKF91 model to linguistic data.

Data and method. The phylogeny is inferred from aligned phonemic sequences of 2,628 cognate word forms from nine Romance languages and Latin. Concepts for the cognate sets are selected from the Swadesh 200-word list. The tree topology and transition rates were estimated in a Bayesian-MCMC framework ([10]). Although the model is provided with alignments, they are treated as a random variable, so posterior distributions are not conditioned on any particular alignment.

*Results and discussion.*Preliminary results demonstrate the utility of the TKF91 model for linguistic phylogenetics. In particular, posterior probabilities for widely recognized inner-Romance clades (e.g., Ibero-Romance, Gallo-Romance) are high. The TKF91 model is not only an important addition to the toolkit of historical linguists, but can also shed new light on theoretical questions of sound change. For instance, this method has contributions to make to the hypothesis of functional load theory ([11], [12]), since it is now possible to estimate meaningful rates of changes among segments. More broadly, the TKF91 model brings linguistic phylogenetics closer to the study of molecular phylogenetics, in as much as segmental sequences parallel those of nucleotides.

References

[1] Donald A. Ringe, Tandy Warnow, and Ann Taylor. Indo-European and computational cladistics. *Transactions of the Philological Society* 100.1 (Mar. 2002), 59–129. DOI: 10.1111/1467-968X.00091.

[2] Luay Nakhleh, Tandy Warnow, Donald A. Ringe, and Steven N. Evans. A comparison of phylogenetic linguistic reconstruction methods on an Indo-European dataset. *Transactions of the Philological Society* 103.2 (Aug. 2005), 171–192. DOI: 10.1111/j.1467-968X.2005.00149.x.

[3] Claire Bowern and Quentin D. Atkinson. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88.4 (Dec. 2012), 817–845. DOI: 10.1353/lan.2012.0081.

[4] Remco R. Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. Mapping the origins and expansion of the Indo-European language family. *Science* 337.6097 (Aug. 2012), 957–960. DOI: 10.1126/science.1219669.

[5] Simon J. Greenhill and Russell D. Gray. Basic vocabulary and Bayesian phylolinguistics. Issues of understanding and representation. *Diachronica* 29.4 (2012), 523–537. DOI: 10.1075/dia.29.4.05gre.

[6] Will Chang, Chundra Aroor Cathcart, David P. Hall, and Andrew J. Garrett. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91.1 (Mar. 2015), 194–244. DOI: 10.1353/lan.2015.0005.

[7] April M. Wright, Graeme T. Lloyd, and David M. Hillis. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology* 65 (2016), 602–611. DOI: 10.1093/sysbio/syv122.

[8] Jeffrey L. Thorne, Hirohisa Kishino, and Joseph Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* 33.2 (Aug. 1991), 114–124. DOI: 10.1007/BF02193625.

[9] Gerton A. Lunter, István Miklós, Yun S. Song, and Jotun Hein. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *Journal of Computational Biology* 10.6 (Dec. 2003), 869–889. DOI: 10.1089/106652703322756122.

[10] Simon J. Greenhill, Paul Heggarty, and Russell D. Gray. Bayesian phylolinguistics. *The handbook of historical linguistics*. Ed. by Brian D. Joseph, Richard D. Janda, and Barbara S. Vance. Vol. 2. Malden, MA: Wiley, 2021. DOI: 10.1002/9781118732168.ch11.

[11] André Martinet. *Économie des changements phonétiques. Traité de phonologie diachronique*. Berne: Francke, 1955.

[12] Charles F. Hockett. The quantification of functional load. *Word* 23.1–3 (1967), 300–320. DOI: 10.1080/00437956.1967.11435484.