Fossilized birth-death models and the diversification of Latin

Divergence-time estimation is one of the most contentious issues in Indo-European linguistics (Gray et al. 2003, Nicholls et al. 2008, Bouckaert et al. 2012, Chang et al. 2015, Heggarty 2021). Most of the debate has focused on the root age of the family, which is understandable given that it is a key component of the Steppe and Anatolian Hypotheses. At the same time, the focus on the root age has led to the neglect of the interior clades (although see recently Hartmann 2021 on Germanic for an exception). For instance, the estimated root ages in Chang et al. 2015 and Rama 2018 are consistent with the Steppe Hypothesis, but the estimated ages of some interior clades conflict with the empirical data. In Chang et al. 2015, the speciation of Latin begins around 1000 CE, which is too late (Heggarty 2021:381–382). In Rama 2018:199 the earliest estimate for the emergence of French is 1000 CE, but Old French is already attested in 842 CE.

To put Indo-European divergence-time estimates on a more secure footing, more work is required at the clade level. This talk addresses this need by investigating the diversification of Latin and focuses on the following two questions. First, when does Latin begin to diversify into what will become the Romance languages? This is a question with a long history of research from the perspective of both Latin and Romance (e.g., Hall 1974, Dardel 1985, Adams 2007). Two main hypotheses have emerged (Väänänen 1983:486–494). One maintains that Latin began to diversify in the early centuries CE; the other dates this process to 600–800 CE. Second, is Latin the direct ancestor of the Romance languages? It is widely believed that the most recent common ancestor of the Romance languages is not written Classical Latin but rather the colloquial language (i.e., Vulgar Latin). So the question then arises of what the phylogenetic relationship between the written language and Vulgar Latin is. Opinion is sharply divided on this point (e.g., Chang et al. 2015:206–207, Heggarty 2021:381).

Recent developments in birth-death models—in particular, among fossilized birth-death (FBD) models—offer a powerful new approach to these long-debated issues (Heath et al. 2014, Stadler et al. 2018). The crucial advantage of fossilized birth-death models is that they treat extinct languages such as Latin on a par with contemporary languages and also allow them to be sampled as ancestors. For the study of divergence times in Indo-European, where we have a rich stock of ancient languages, these models are of paramount importance. I investigate in particular the standard FBD model and the FBD range model.

Data and Methods. This study jointly infers a tree topology and divergence-time estimates from the A3 dataset of Chang et al. 2015, which includes fourteen Romance languages and Latin. This is a dataset of binary-state basic vocabulary characters, which ultimately descends from the IELEX database. It contains 197 meaning classes and loanwords were not culled. Model parameters including divergence times were estimated in a Bayesian framework. A variety of clock models (e.g., strict, uncorrelated exponential, uncorrelated lognormal), fossilized birth-death models (standard and range), and character models (Mk and F81 mixture) were examined. Model comparison was carried out with Bayes Factor analysis. Tests of model adequacy were carried out with posterior predictive simulation.

Results. Fossilized birth-death range models in general fit the data better than standard fossilized birth-death models. The results from the best model unequivocally support the early model of the diversification of Latin. Figure 1 presents the posterior distribution of the age of Proto-Romance

(ages are presented in thousands of years before the present, which is assumed to be the year 2000 CE; the maximum a posteriori estimate of 1.65 kya is highlighted in green). The probability that Latin started to diversify after 500 CE is less than ten percent. Second, the probability that Latin is a direct ancestor of the Romance languages is over eighty percent. In the maximum a posteriori (MAP) tree in Figure 2, Classical Latin is represented as a fossil along the lineage that will become the common ancestor of the Romance languages.

Conclusions. The results of this study support the following general conclusions. First, the fossilized birth-death range process is of greater utility for estimating the timing of linguistic events than has thus far been appreciated. Second, the methodological challenges highlighted by Heggarty (2021:382) can be mitigated with extensive tests of model comparison, model sensitivity, and model adequacy. Finally, the results of this study accord with the view of Adams (2007:725), who writes that "[w]e should get away from the idea that Latin was monolithic until a very late date, when some catastrophic event caused it to 'split up.'"

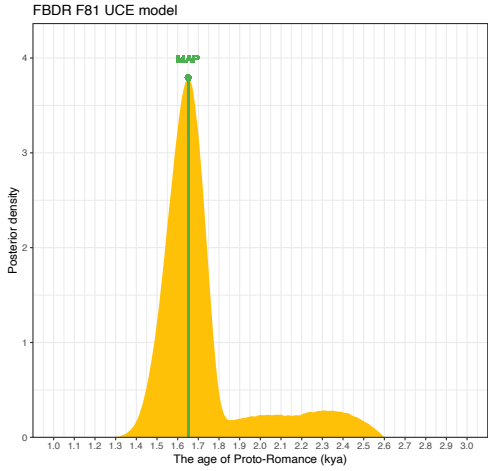

Figure 1: Posterior distribution of the age of Proto-Romance

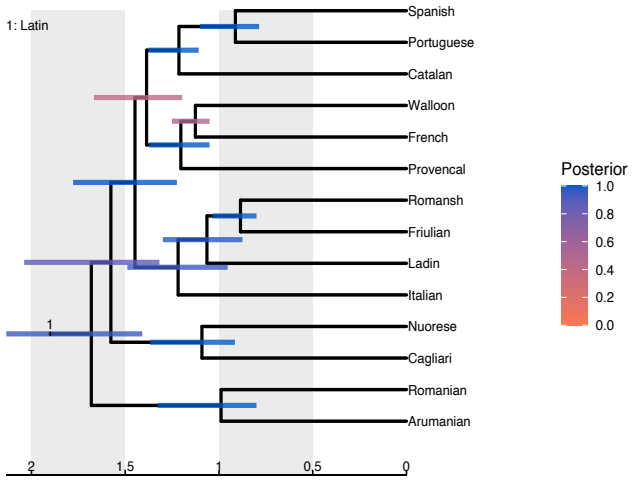

Figure 2: Maximum a posteriori (MAP) tree

Adams, James N. (2007). *The regional diversification of Latin 200 BC–AD 600*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511482977.

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

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

Dardel, Robert de (1985). Le sarde représente-t-il un état précoce du roman commun? *Revue de Linguistique romane* 49, 263–269.

Gray, Russell D. and Quentin D. Atkinson (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426, 435–439. DOI: 10.1038/nature02029.

Hall Jr., Robert A. (1974). *External history of the Romance languages*. New York: Elsevier.

Hartmann, Frederik (2021). Germanic phylogeny. A computational investigation using Bayesian inference and agent-based models. PhD thesis. Universität Konstanz.

Heath, Tracy A., John P. Huelsenbeck, and Tanja Stadler (2014). The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences of the United States of America* 111.29 (July 2014), E2957–E2966. DOI: 10.1073/pnas.1319091111.

Heggarty, Paul (2021). Cognacy databases and phylogenetic research on Indo-European. *Annual Review of Linguistics* 7, 371–394. DOI: 10.1146/annurev-linguistics-011619-030507.

Nicholls, Geoff K. and Russell D. Gray (2008). Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of The Royal Statistical Society, Series B* 70.3, 545–566. DOI: 10.1111/j.1467-9868.2007.00648.x.

Rama, Taraka (2018). Three tree priors and five datasets. A study of the effect of tree priors in Indo-European phylogenetics. *Language Dynamics and Change* 8.2, 182–218. DOI: 10.1163/2210583 2-00802005.

Stadler, Tanja, Alexandra Gavryushkina, Rachel C. M. Warnock, Alexei J. Drummond, and Tracy A. Heath (2018). The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. *Journal of Theoretical Biology* 447, 41–55. DOI: 10.1016/j.jtbi.2018.03.005.

Väänänen, Veikko (1983). Le problème de la diversification du latin. *Aufstieg und Niedergang der römischen Welt. Sprache und Literatur (Sprachen und Schriften)*. Ed. by Wolfgang Haase. Vol. 29. Berlin: de Gruyter, 480–505. DOI: 10.1515/9783110847024-009.