A vanishing, multiple-gain lexical trait model

Challenges and opportunities in lexical data and analysis

Will Chang
University of California, Berkeley

17–20 September, 2014
Workshop Towards a Global Language Phylogeny, Jena

SUMMARY. I motivate a novel lexical trait model, fit its parameters to Indo-European (IE) data, and estimate the number of related items in a language that have diverged from IE 10,000–50,000 years ago.

1. IE lexical traits are homoplastic

In IE datasets, a lexical trait is a root-meaning correspondence, e.g.:

> French *homme* 'adult male' and Modern Irish *duine* 'adult male' both derive from the IE root **d^{h}ǵ^{h}om-*. Thus French and Modern Irish both have the trait [**d^{h}ǵ^{h}om-*, 'adult male'].

More precisely, "has trait [X,Y]" means that root X is used in the most semantically general and stylistically neutral word for meaning Y.

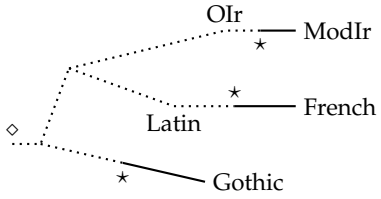
This trait is homoplastic, due to a recurrent semantic shift.

Latin *homō* 'person'	→	French *homme* 'adult male'
Old Irish *duine* 'person'	→	Modern Irish *duine* 'adult male'

Both shifts exemplify:

PRECURSOR TRAIT	→	COROLLARY TRAIT
[**d^{h}ǵ^{h}om-*, 'person']	→	[**d^{h}ǵ^{h}om-*, 'adult male']

The corollary trait is gained independently in Romance and Irish. This is NOT borrowing, since contact is not involved. Rather:



- ◇ [**d^{h}ǵ^{h}om-*, 'person'] is born in a common ancestor of Italic and Celtic (and Germanic too, cf. Gothic *guma* 'adult male').
- ⋆ It is replaced by [**d^{h}ǵ^{h}om-*, 'adult male'] three times.

Alternatively, if we posit that parallel gains in lexical traits are impossible without borrowing, the implications are awkward:

- In Proto-Italo-Celtic (MRCA of Latin and Irish) the root **d^{h}ǵ^{h}om-* was used for both 'person' and 'adult male'.
- In Proto-Italo-Celtic, the root **wiHrós* was also used for 'adult male'; Latin *vir* 'adult male' and Old Irish *fer* 'adult male' both reflect it.
- Classical Latin is not directly ancestral to French; while Classical Latin was attested, the contempory ancestor of French was not. (Same for Old and Modern Irish.)

2. *Homoplasy is common*

In IELEX, of the 789 traits attested in 14 Romance languages, 64 traits are missing in Latin but attested elsewhere in IE. (Items tagged as loans were first excluded.) Thus 8.1% of the Romance traits are homoplastic. For closer inspection, those in French are listed below. (French etymologies are from Gamillscheg, 1969).

MEANING	LATIN	FRENCH	ETYMOLOGY	
because	*quod*	*parce que*	*par* < L *per* 'through', *ce* < L *ecce* 'behold' + *hoc* 'that' (?), *que* < L *quid* 'what'	
here	*hīc*	*ici*	L *ecce* 'behold' + *hīc* 'here'	
this	*hic*	*ceci*	*ce* < L *ecce* 'behold' + *istum* 'that', *ci* < L *ecce* 'behold' + *hīc* 'here'	
dirty	*sordidus*	*sale*	MHG *sal*	'cloudy, dark'
stick	*baculum*	*bâton*	VL *bastum*	'staff'
white	*albus*	*blanc*	Frank **blank*	'white, shining'
big	*magnus*	*gros*	L *grossus*	'thick, coarse'
far	*procul*	*loin*	L *longe*	'long, far-off'
fear	*tīmere*	*craindre*	L *tremere*	'tremble'
man	*homō*	*homme*	L *homō*	'person'
rightside	*dexter*	*droit*	L *directus*	'straightened'
river	*flūmen*	*rivière*	L *rīpārius*	'of a riverbank'
sand	*arēna*	*sable*	L *sabulō*	'coarse sand'
short	*brevis*	*court*	L *curtus*	'shortened'
skin	*cutis*	*peau*	L *pellis*	'pelt, hide'
split	*scindere*	*fendre*	L *findere*	'cleave, break up'
thin	*tenuis*	*mince*	L *minūtia*	'trifle'
warm	*tepidus*	*chaud*	L *calidus*	'hot'

There are three kinds of etymologies.

- French function words ('because', 'here', 'this') are phonologically reduced phrases, whose elements may themselves be phonologically reduced phrases. The accretion of IE deictic morphemes results in homoplasy.
- Some French forms ('dirty', 'stick', 'white') are not found in Classical Latin, and are probably loans.
- The remainder exemplify semantic shifts that can plausibly recur. Such shifts are the primary mechanism of lexical replacement, and tend to be unidirectional.

Loanwords do not support my thesis that recurrent semantic shifts result in parallel gain. However, they do show how semantic shift is a mechanism of trait gain. They were probably not the basic words for 'dirty', 'stick', and 'white' until long after they were borrowed into Late Latin or Early French.

3. *Root-meaning traits are vanishing*

Semantic shift and morphological processes may cause a form to spread from meaning to meaning, but eventually a form vanishes completely. Traits associated with the form ought never to recur.

Restriction site characters and covarion characters are multiple-gain trait models in which a trait may always recur. They are awkward for modeling lexical traits.

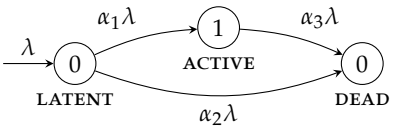
- They are not marginally invariant. As more languages are added to an analysis, the stationary frequency of trait presence (π_1) falls. A universal characterization of lexical trait behavior is impossible.
- In a recent analysis involving 94 IE languages, there were 5700 lexical traits; π_1 was estimated to be around one percent. Implication: no matter how distantly a language is related to IE, it is expected to contain 57 IE cognates in basic vocabulary!

In English, some forms hang on by a thread. PIE **medhu-* 'honey' is reflected only in *mead*. OE *wer* 'man' is reflected only in *werewolf*.

4. *Latent birth traits*

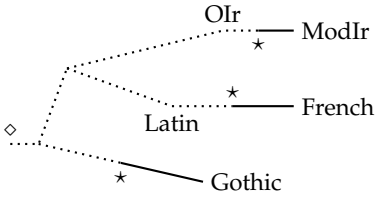
Let's craft a trait model that is MULTIPLE-GAIN but VANISHING by extending Nicholls and Gray's single-gain trait model (Nicholls & Gray, 2008).

- There is a global trait birth rate on the tree of λ.
- On birth, a trait is LATENT rather than ACTIVE.
- A latent trait becomes active at rate $\alpha_1\lambda$ and dies at rate $\alpha_2\lambda$; an active trait dies at rate $\alpha_3\lambda$.
- Active traits are present (1); latent and dead traits are absent (0).



How would the trait [**d*^{h}*ǵ*^{h}*om-*, 'adult male'] evolve?

◇ It is born in a common ancestor of Germanic, Italic, and Celtic.
⋆ It becomes active three times.



Interpretation: trait birth is when the precursor trait [**d*^{h}*ǵ*^{h}*om-*, 'person'] became active.

But why can't trait birth correspond to when the precursor trait's precursor became active? PIE **d*^{h}*ǵ*^{h}*om-* probably meant 'earth' (cf. Greek *khthōn* 'soil, earth') before additional morphology gave 'earthling' > 'person'. Since any language that has [**d*^{h}*ǵ*^{h}*om-*, 'earth'] has the potential to develop [**d*^{h}*ǵ*^{h}*om-*, 'male adult'], the trait birth for the latter should go back at least that far. Latency ought to be a graded thing, with multiple levels of actualization. A binary value for latency could be the Achilles heel of this trait model.

5. *Experiment*

Bayesian inference is used to fit model parameters to IE lexical data (39 languages, 143 meaning classes).

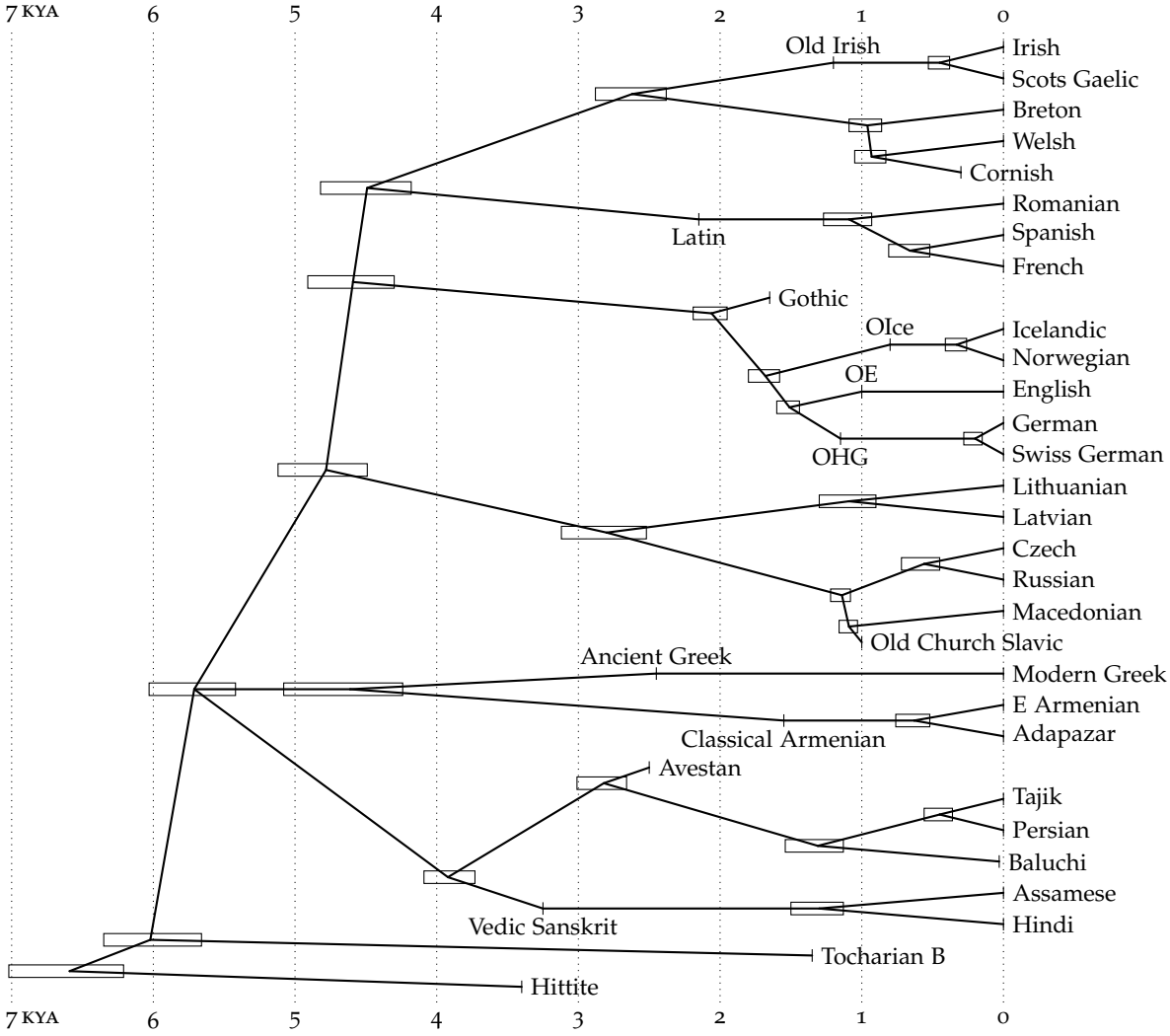
- To simplify inference, I fix the topology and use a strict clock.
- Eight languages are constrained to be directly ancestral to other languages. Attested languages serve as the only calibration points. Tagged loans are retained in the data.
- Traits are partitioned by meaning. Each meaning k has its own birth rate λ_k. Birth rates $\lambda_1, \ldots, \lambda_K$ are log-normal distributed. Transition rate parameters $\alpha_1, \alpha_2, \alpha_3$ are shared over all meanings.

I left out many languages from dialect continua, to minimize the effect of language contact.

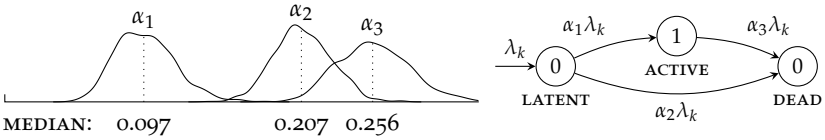
Ancestry constraints was the basis of recent work that found a later date for PIE (Chang et al., under review).

$\log \lambda_k \sim \mathcal{N}(\mu, \sigma^2)$, with μ taken uniformly from $(-\infty, \infty)$ and σ taken from $(0, \infty)$ with probability proportional to σ^{-1}.

The estimated chronology, with nodes at posterior medians; boxes show 5–95%ile ranges:



6. Transition rates, relative to birth rate



For the sake of concreteness, consider 'bird', which has a birth rate of once per millenium ($\lambda_k \approx 0.001$).

- On average a 'bird' trait remains latent for $1/[(\alpha_1 + \alpha_2)\lambda_k] \approx 3300$ years.
- A latent 'bird' trait goes active with probability $\alpha_1/(\alpha_1 + \alpha_2) \approx 32\%$. Otherwise it dies.
- The longevity of a 'bird' trait (mean duration active) is $1/(\alpha_3\lambda_k) \approx 3900$ years.

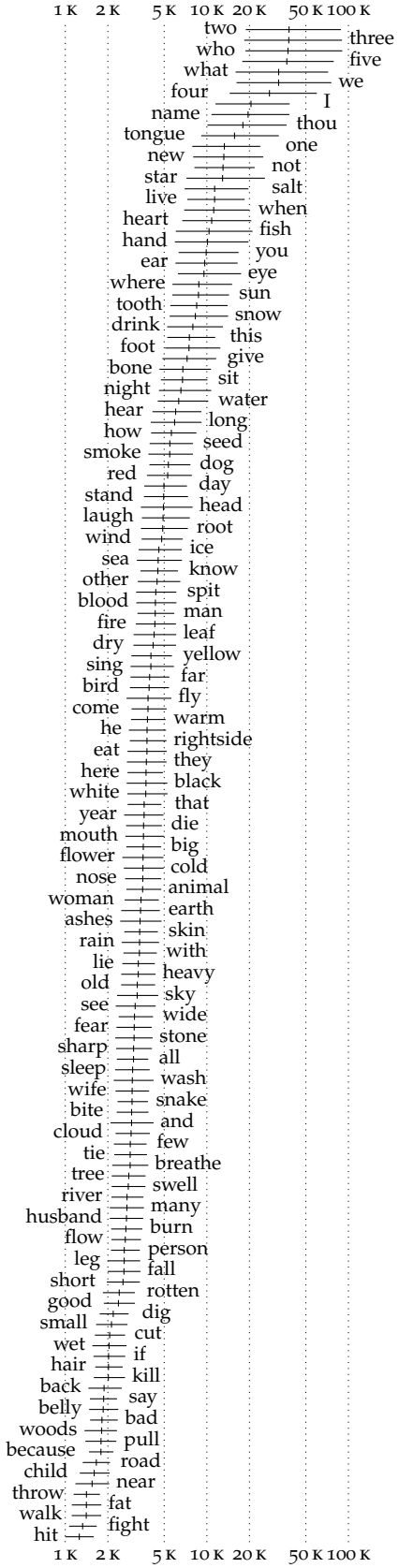
Some general implications:

- On average there are $\alpha_3/\alpha_1 \approx 2.6$ latent traits for every active one. That's the number of traits in 'adjacent' meanings that could supplant the active trait.
- The mean latent duration is close to the mean active duration, suggesting that a precursor trait and a corollary trait have similar longevity.

7. Longevity

Plotted on the right is the longevity $1/(\alpha_3\lambda_k)$ for each meaning class; shown are the median posterior and the 5–95%ile range.

- The longevity of the most stable meanings ('two', 'three', 'five', 'who') are very sensitive to the prior; these traits are invariant within IE, so the likelihood is relatively insensitive to a wide range of values for λ_k.
- Had I modeled rate variation with a gamma distribution (rather than a log-normal distribution) their longevities would double.
- Compared to the log-normal distribution, the gamma distribution gives more weight to values near zero when its shape parameter is around 2, as is the case when fitting this dataset.

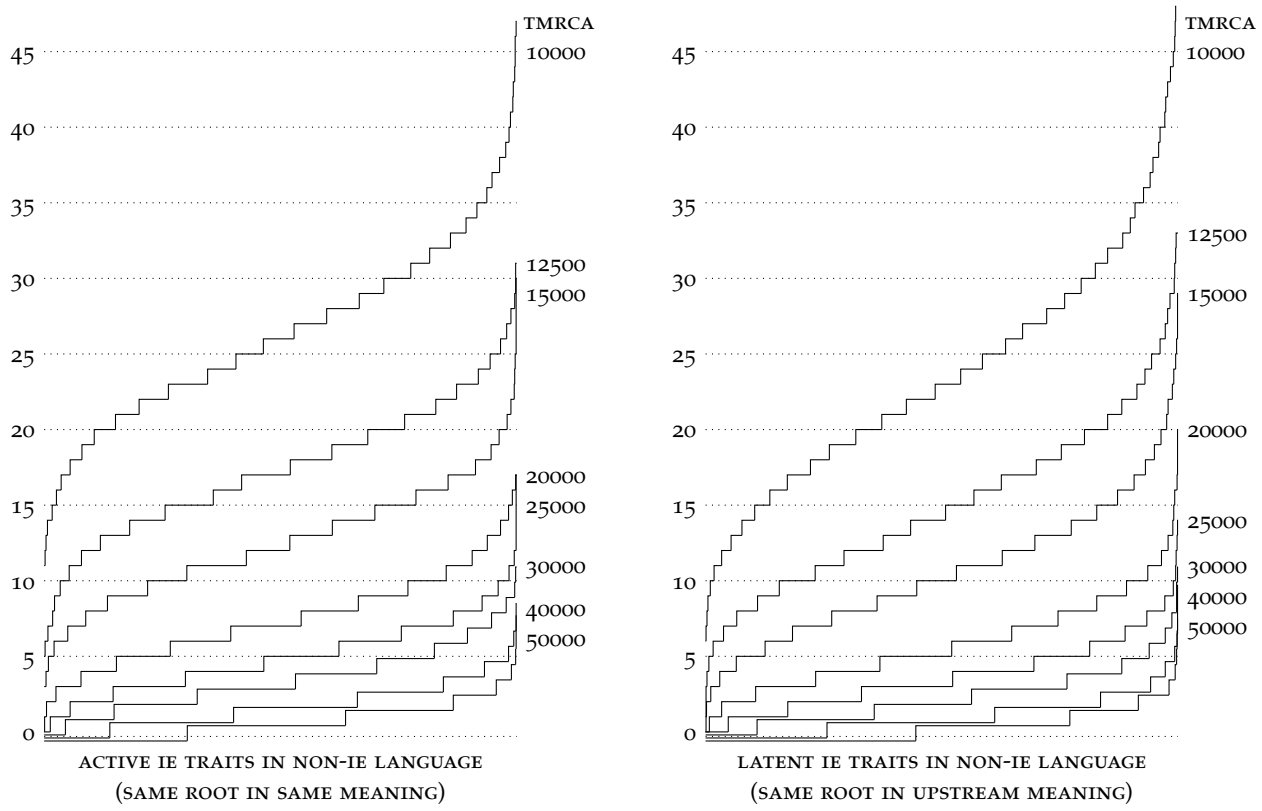


8. IE traits outside IE

If a non-IE language has a given TMRCA with IE:

- How many IE traits (in the basic vocabulary) would it attest?
- How many IE traits would be latent in the language?

These are random quantities. I simulate 1000 outcomes (drawing from their posterior predictive distributions) for various choices of TMRCA, and plot the outcomes.



The left pane gives results similar to Atkinson (2010). The right pane shows that roughly as many related forms are in upstream meanings, and presumably as many are in downstream meanings. In all, perhaps 30 IE forms could be found in a language that diverged 20,000 years ago. If only we could weave together:

- A model of semantics, to know where to look for related forms.
- A model of sound change, to know what related forms look like.
- A model of cognate judgment uncertainty that factors in the phonological and semantic plausibility of a match.
- A more refined model of lexical borrowing.

The fine print: since this analysis treats loans and non-loans equally, the count of true cognates may be inflated. Tagged loans comprised 2.1% of the forms in this analysis.

9. Model evaluation (added later)

As a rough indicator of how the latent-birth (LB) trait model performs relative to restriction site characters (RSC) or covarion (CV), I give the log probability of the data under each trait model. Since each has just a few parameters, the marginal likelihoods should be similar.

RSC	-13419 ± 10
CV	-13239 ± 11
LB	-9183 ± 10

The improvement from RSC to CV is about 1% in log probability, which is similar to the 0.5% gain reported in the supplement of Bouckaert et al (2013). An interpretation is that under CV, each trait is assigned a probability that is 9% higher, on average. (There are 2169 traits in these analyses.) Compared to CV, LB assigns each trait pattern a probability that is 6.5 times in size, on average.

References

Atkinson, Quentin D. 2010. The prospects for tracing deep language ancestry. *Journal of Anthropological Sciences* 88.231–233

Bouckaert, Remco; Philippe Lemey; Michael Dunn; Simon J. Greenhill; Alexander V. Alekseyenko; Alexei J. Drummond; Russell D. Gray; Marc A. Suchard; and Quentin D. Atkinson. 2013. Corrections and clarifications. *Science* 342.1446.

Chang, William; David Hall; Chundra Cathcart; and Andrew Garrett. Under review. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis.

Gamillscheg, Ernst. 1969. *Etymologisches wörterbuch der französischen sprache*. Carl Winter Universitätsverlag, Heidelberg.

Nicholls, Geoff K. and Russell D. Gray. Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society, Series B* 70.545–566