Phylogenetic Models of Language Change: Validating Interference and Quantifying Uncertainty

Robin J. Ryder

Centre de Recherche en Mathématiques de la Décision
Université Paris-Dauphine

4 October 2019

Statistics in historical corpus linguistics, Maynooth University

Based on work with Geoff Nicholls and with Guillaume Jacques, Laurent Sagart, Yunfan Lai, Valentin Thouzeau, Simon Greenhill and Mattis List

What to remember

For a statistical analysis to be trustworthy, it needs to include:

- A measure of uncertainty
- A validation of the inference procedure
- “All models are wrong, but some are useful.”

Introduction

A large number of recent papers describe computationally-intensive statistical methods for Historical Linguistics

- Increased computational power
- Advances in statistical methodology
- New datasets
- Complex linguistic questions which cannot be answered with traditional methods

Caveats

- I am not a linguist
- I am a statistician
- Some of these papers were not written by me; figures were created by the papers' authors
- I use the word "evolution" in a broad sense
- "All models are wrong, but some are useful"

Aims of this talk

- Review of several recent papers on statistical models for Historical Linguistics
- Walk through statistical methodology
- Statisticians *won't* replace linguists
- When done correctly, collaborations between statisticians and linguists can provide useful results

Advantages of statistical methods

- Analyse (very) large datasets
- Test multiple hypotheses
- Cross-validation
- Estimate uncertainty

Languages diversify

- Languages “evolve” similarly to biologically species
- Similarities between languages indicate they may be cousins
- Most standard model: tree

Questions of interest

- Which languages are related?
- Given a set of related languages, can we reconstruct their history and the age of the most recent common ancestor (MRCA)?
- What mechanisms drive language change?
- How do the various parts of language change? Vocabulary, syntax, phonetics...

Why be Bayesian?

In the settings described in this talk, it usually makes sense to use Bayesian inference, because:

- The models are complex
- Estimating uncertainty is paramount
- The data are not "big"
- Some prior information is available
- The output of one model is used as the input of another
- We are interested in complex functions of our parameters

Bayesian statistics

- Statistical inference deals with estimating an unknown parameter θ given some data D.
- In the Bayesian framework, the parameter θ is seen as inherently random: it has a distribution.
- Before I see any data, I have a *prior* distribution on $\pi(\theta)$, usually uninformative.
- Once I take the data into account (through the likelihood function L), I get a *posterior* distribution, which is hopefully more informative.

$$\pi(\theta|D) \propto \pi(\theta)L(\theta|D)$$

- Different people have different priors, hence different posteriors. But with enough data, the choice of prior matters little.
- We are allowed to make probability statements about θ, such as "there is a 95% probability that θ belongs to the interval $[78\ ;\ 119]$" (credible interval)

Bayes factors

Two models $\mathcal{M}_1$ and $\mathcal{M}_2$ can be compared using a Bayes factor. Compute the marginal likelihood:

$$m_1(D) = \int L_1(\theta_1; D)\pi_1(\theta_1)\, d\theta_1$$

and $m_2(D)$ similarly. Then

$$BF_{12}(D) = \frac{m_1(D)}{m_2(D)}$$

Usually interpreted on the log scale: if $\log BF > 2$, decisive evidence in favour of model 1; if $\log BF < -2$, decisive evidence in favour of model 2; between -2 and 2, weaker evidence.

- Includes a natural penalty of more complex models.
- Treats models symmetrically (no "null" hypothesis)
- Related to the BIC (Bayesian Information Criterion)
- Can be long and painful to compute

Advantages and drawbacks of Bayesian statistics

- More intuitive interpretation of the results
- Easier to think about uncertainty
- In a hierarchical setting, it becomes easier to take into account all the sources of variability
- Prior specification: need to check that changing your prior does not change your result
- Computationally intensive

Statistical method in a nutshell

1. Collect data
2. Design model
3. Perform inference (MCMC, ...)
7. Conclude

Statistical method in a nutshell

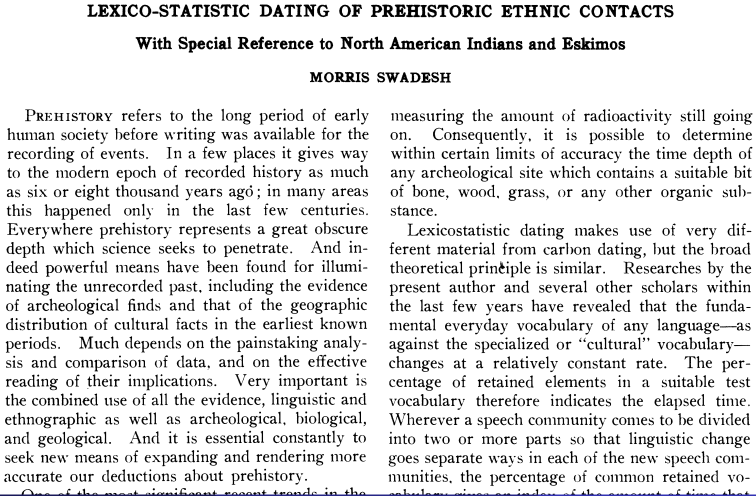
1. Collect data
2. Design model
3. Perform inference (MCMC, ...)
4. Check convergence
5. In-model validation (is our inference method able to answer questions from our model?)
6. Model mis-specification analysis (do we need a more complex model?)
7. Conclude

In general, it is more difficult to perform inference for a more complex model.

Outline

1. Swadesh: Glottochronology
2. Gray & Atkinson: Language phylogenies
3. Pagel et al.: Frequency of use
4. Sagart et al.: Sino-Tibetan phylogenies
5. Re-examining Bergsland and Vogt
6. Conclusions

Swadesh (1952)



LEXICO-STATISTIC DATING OF PREHISTORIC ETHNIC CONTACTS

With Special Reference to North American Indians and Eskimos

MORRIS SWADESH

PREHISTORY refers to the long period of early human society before writing was available for the recording of events. In a few places it gives way to the modern epoch of recorded history as much as six or eight thousand years agȯ; in many areas this happened only in the last few centuries. Everywhere prehistory represents a great obscure depth which science seeks to penetrate. And indeed powerful means have been found for illuminating the unrecorded past, including the evidence of archeological finds and that of the geographic distribution of cultural facts in the earliest known periods. Much depends on the painstaking analysis and comparison of data, and on the effective reading of their implications. Very important is the combined use of all the evidence, linguistic and ethnographic as well as archeological, biological, and geological. And it is essential constantly to seek new means of expanding and rendering more accurate our deductions about prehistory.

One of the most significant recent trends in the

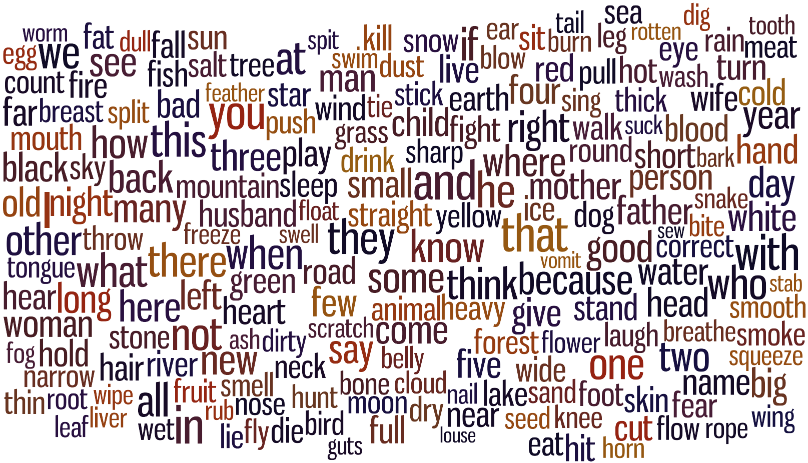
measuring the amount of radioactivity still going on. Consequently, it is possible to determine within certain limits of accuracy the time depth of any archeological site which contains a suitable bit of bone, wood, grass, or any other organic substance.

Lexicostatistic dating makes use of very different material from carbon dating, but the broad theoretical principle is similar. Researches by the present author and several other scholars within the last few years have revealed that the fundamental everyday vocabulary of any language—as against the specialized or "cultural" vocabulary—changes at a relatively constant rate. The percentage of retained elements in a suitable test vocabulary therefore indicates the elapsed time. Wherever a speech community comes to be divided into two or more parts so that linguistic change goes separate ways in each of the new speech communities, the percentage of common retained vocabulary gives an index of the amount of time that

First attempt: Swadesh (1952)

Aim: dating the MRCA (Most Recent Common Ancestor) of a pair of languages.
Data: "core vocabulary" (Swadesh lists). 215 or 100 words.

Core vocabulary



Assumptions

Swadesh assumed that core vocabulary evolves at a constant rate (through time, space and meanings). Given a pair of languages with percentage C of shared cognates, and a constant retention rate r, the age t of the MRCA is

$$t = \frac{\log C}{2 \log r}$$

The constant r was estimated using a pair of languages for which the age of the MRCA is known.

Issues with glottochronology

Many statistical shortcomings. Mainly:

1. Simplistic model
2. No evaluation of uncertainty of estimates
3. Only small amounts of data are used

Bergsland and Vogt (1962) debunked glottochronology, showing on 3 pairs of languages with known history that the assumption of constant rates does not hold.

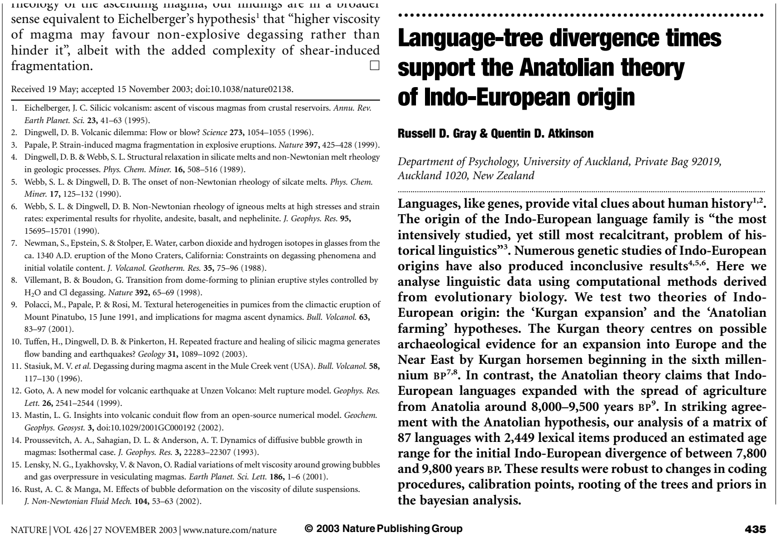
What has changed?

- More elaborate models + model misspecification analyses
- We can estimate the uncertainty (⇒easier to answer "I don't know")
- Large amounts of data

Outline

1. Swadesh: Glottochronology
2. Gray & Atkinson: Language phylogenies
3. Pagel et al.: Frequency of use
4. Sagart et al.: Sino-Tibetan phylogenies
5. Re-examining Bergsland and Vogt
6. Conclusions

Gray & Atkinson (2003)



rheology of the ascending magma, our findings are in a broader sense equivalent to Eichelberger's hypothesis[1] that "higher viscosity of magma may favour non-explosive degassing rather than hinder it", albeit with the added complexity of shear-induced fragmentation. □

Received 19 May; accepted 15 November 2003; doi:10.1038/nature02138.

1. Eichelberger, J. C. Silicic volcanism: ascent of viscous magmas from crustal reservoirs. *Annu. Rev. Earth Planet. Sci.* **23,** 41–63 (1995).
2. Dingwell, D. B. Volcanic dilemma: Flow or blow? *Science* **273,** 1054–1055 (1996).
3. Papale, P. Strain-induced magma fragmentation in explosive eruptions. *Nature* **397,** 425–428 (1999).
4. Dingwell, D. B. & Webb, S. L. Structural relaxation in silicate melts and non-Newtonian melt rheology in geologic processes. *Phys. Chem. Miner.* **16,** 508–516 (1989).
5. Webb, S. L. & Dingwell, D. B. The onset of non-Newtonian rheology of silcate melts. *Phys. Chem. Miner.* **17,** 125–132 (1990).
6. Webb, S. L. & Dingwell, D. B. Non-Newtonian rheology of igneous melts at high stresses and strain rates: experimental results for rhyolite, andesite, basalt, and nephelinite. *J. Geophys. Res.* **95,** 15695–15701 (1990).
7. Newman, S., Epstein, S. & Stolper, E. Water, carbon dioxide and hydrogen isotopes in glasses from the ca. 1340 A.D. eruption of the Mono Craters, California: Constraints on degassing phenomena and initial volatile content. *J. Volcanol. Geotherm. Res.* **35,** 75–96 (1988).
8. Villemant, B. & Boudon, G. Transition from dome-forming to plinian eruptive styles controlled by H_2O and Cl degassing. *Nature* **392,** 65–69 (1998).
9. Polacci, M., Papale, P. & Rosi, M. Textural heterogeneities in pumices from the climactic eruption of Mount Pinatubo, 15 June 1991, and implications for magma ascent dynamics. *Bull. Volcanol.* **63,** 83–97 (2001).
10. Tuffen, H., Dingwell, D. B. & Pinkerton, H. Repeated fracture and healing of silicic magma generates flow banding and earthquakes? *Geology* **31,** 1089–1092 (2003).
11. Stasiuk, M. V. *et al.* Degassing during magma ascent in the Mule Creek vent (USA). *Bull. Volcanol.* **58,** 117–130 (1996).
12. Goto, A. A new model for volcanic earthquake at Unzen Volcano: Melt rupture model. *Geophys. Res. Lett.* **26,** 2541–2544 (1999).
13. Mastin, L. G. Insights into volcanic conduit flow from an open-source numerical model. *Geochem. Geophys. Geosyst.* **3,** doi:10.1029/2001GC000192 (2002).
14. Proussevitch, A. A., Sahagian, D. L. & Anderson, A. T. Dynamics of diffusive bubble growth in magmas: Isothermal case. *J. Geophys. Res.* **3,** 22283–22307 (1993).
15. Lensky, N. G., Lyakhovsky, V. & Navon, O. Radial variations of melt viscosity around growing bubbles and gas overpressure in vesiculating magmas. *Earth Planet. Sci. Lett.* **186,** 1–6 (2001).
16. Rust, A. C. & Manga, M. Effects of bubble deformation on the viscosity of dilute suspensions. *J. Non-Newtonian Fluid Mech.* **104,** 53–63 (2002).

Language-tree divergence times support the Anatolian theory of Indo-European origin

Russell D. Gray & Quentin D. Atkinson

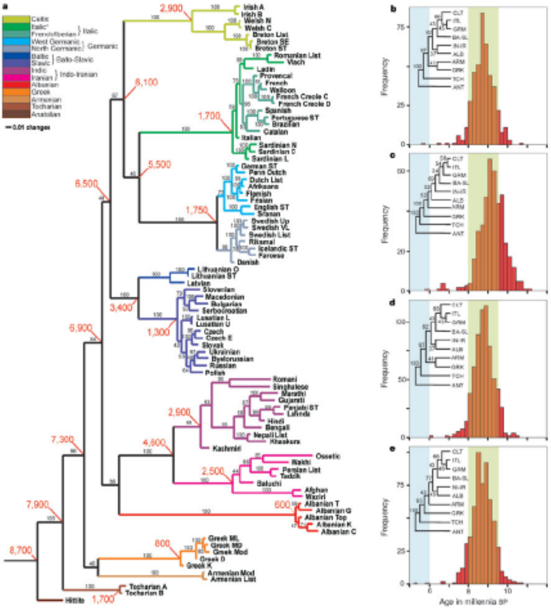
Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1020, New Zealand

Languages, like genes, provide vital clues about human history[1,2]. The origin of the Indo-European language family is "the most intensively studied, yet still most recalcitrant, problem of historical linguistics"[3]. Numerous genetic studies of Indo-European origins have also produced inconclusive results[4,5,6]. Here we analyse linguistic data using computational methods derived from evolutionary biology. We test two theories of Indo-European origin: the 'Kurgan expansion' and the 'Anatolian farming' hypotheses. The Kurgan theory centres on possible archaeological evidence for an expansion into Europe and the Near East by Kurgan horsemen beginning in the sixth millennium BP[7,8]. In contrast, the Anatolian theory claims that Indo-European languages expanded with the spread of agriculture from Anatolia around 8,000–9,500 years BP[9]. In striking agreement with the Anatolian hypothesis, our analysis of a matrix of 87 languages with 2,449 lexical items produced an estimated age range for the initial Indo-European divergence of between 7,800 and 9,800 years BP. These results were robust to changes in coding procedures, calibration points, rooting of the trees and priors in the bayesian analysis.

Swadesh lists, better analysed

- Use Swadesh lists for 87 Indo-European languages, and a phylogenetic model from Genetics
- Assume a tree-like model of evolution with constant rate of change
- Bayesian inference via MCMC (Markov Chain Monte Carlo)
- Reconstruct trees and dates
- Main parameter of interest: age of the root (Proto-Indo-European, PIE)

Lexical trees



Correcting the issues with glottochronology

Returning to the issues with Swadesh's glottochronology:

1. Simplistic model → Slightly better, but the model of evolution is rudimentary
2. No evaluation of uncertainty of estimates → Bayesian inference
3. Only small amounts of data are used → Large number of languages reduces variability of estimates

Bayesian inference for lexical trees

- The tree parameter is seen as random: it has a distribution
- Via MCMC, G & A get a sample of possible trees, with associated probabilities, rather than a single tree
- The uncertainty in trees is thus made explicit

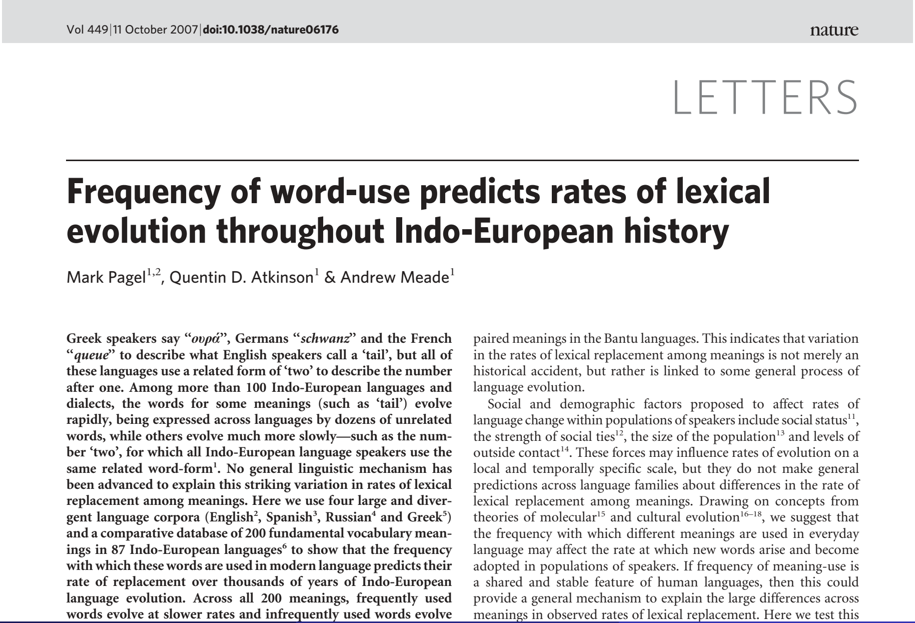
G & A: conclusions

- Age of PIE: 7800-9800 BP (Before Present)
- Large error bars, but this is a good thing
- Reconstruct many known features of the tree of Indo-European languages
- Little validation of the model, no model misspecification analysis
- These trees can also be used as a building block to answer other questions.

Outline

1. Swadesh: Glottochronology
2. Gray & Atkinson: Language phylogenies
3. Pagel et al.: Frequency of use
4. Sagart et al.: Sino-Tibetan phylogenies
5. Re-examining Bergsland and Vogt
6. Conclusions

Pagel et al. (2007)



Vol 449|11 October 2007|**doi:10.1038/nature06176** nature

LETTERS

Frequency of word-use predicts rates of lexical evolution throughout Indo-European history

Mark Pagel[1,2], Quentin D. Atkinson[1] & Andrew Meade[1]

Greek speakers say "*ουρά*", Germans "*schwanz*" and the French "*queue*" to describe what English speakers call a 'tail', but all of these languages use a related form of 'two' to describe the number after one. Among more than 100 Indo-European languages and dialects, the words for some meanings (such as 'tail') evolve rapidly, being expressed across languages by dozens of unrelated words, while others evolve much more slowly—such as the number 'two', for which all Indo-European language speakers use the same related word-form[1]. No general linguistic mechanism has been advanced to explain this striking variation in rates of lexical replacement among meanings. Here we use four large and divergent language corpora (English[2], Spanish[3], Russian[4] and Greek[5]) and a comparative database of 200 fundamental vocabulary meanings in 87 Indo-European languages[6] to show that the frequency with which these words are used in modern language predicts their rate of replacement over thousands of years of Indo-European language evolution. Across all 200 meanings, frequently used words evolve at slower rates and infrequently used words evolve

paired meanings in the Bantu languages. This indicates that variation in the rates of lexical replacement among meanings is not merely an historical accident, but rather is linked to some general process of language evolution.

Social and demographic factors proposed to affect rates of language change within populations of speakers include social status[11], the strength of social ties[12], the size of the population[13] and levels of outside contact[14]. These forces may influence rates of evolution on a local and temporally specific scale, but they do not make general predictions across language families about differences in the rate of lexical replacement among meanings. Drawing on concepts from theories of molecular[15] and cultural evolution[16–18], we suggest that the frequency with which different meanings are used in everyday language may affect the rate at which new words arise and become adopted in populations of speakers. If frequency of meaning-use is a shared and stable feature of human languages, then this could provide a general mechanism to explain the large differences across meanings in observed rates of lexical replacement. Here we test this

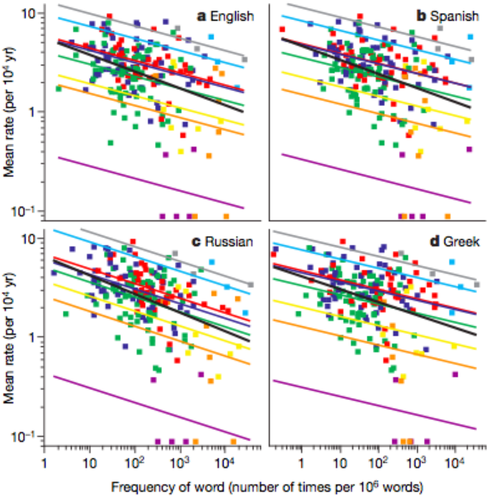
Question at hand

- Check link between frequency of use and rate of change for vocabulary.
- Hypothesis: when a meaning is used more often, the corresponding word has less chances of changing.
- Problem: since this rate is expected to be very slow, we need to look at the deep history. But then the evolutionary history is unknown.

Workaround

- Use Indo-European core vocabulary data, and frequencies from English, Greek, Russian and Spanish
- Get a sample from the distribution on trees and ancestral ages using G&A's method
- For each tree in the sample, estimate the rate of change for each meaning.
- Average across all trees.

Results



(The different colours correspond to different classes of words: numerals, body parts, adjectives...)

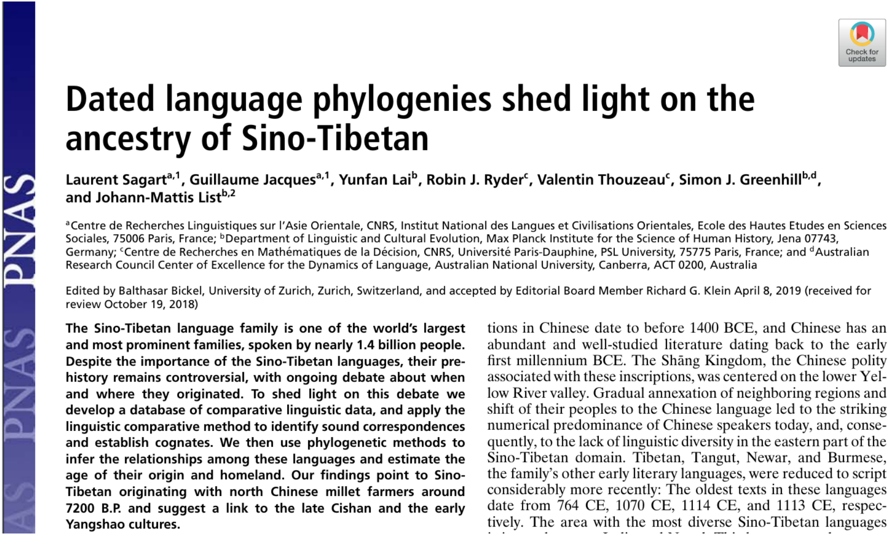
Comments

- There is significant (negative) correlation between frequency of use and rate of change.
- Even if there is high uncertainty in the phylogenies, we can still answer other questions (integrating out the tree)
- Similar results for Bantu (Pagel & Meade 2006)
- It would have been much harder to evaluate this hypothesis without the Bayesian paradigm.

Outline

1. Swadesh: Glottochronology
2. Gray & Atkinson: Language phylogenies
3. Pagel et al.: Frequency of use
4. Sagart et al.: Sino-Tibetan phylogenies
5. Re-examining Bergsland and Vogt
6. Conclusions

Sagart et al. (2019)



Dated language phylogenies shed light on the ancestry of Sino-Tibetan

Laurent Sagart[a,1], Guillaume Jacques[a,1], Yunfan Lai[b], Robin J. Ryder[c], Valentin Thouzeau[c], Simon J. Greenhill[b,d], and Johann-Mattis List[b,2]

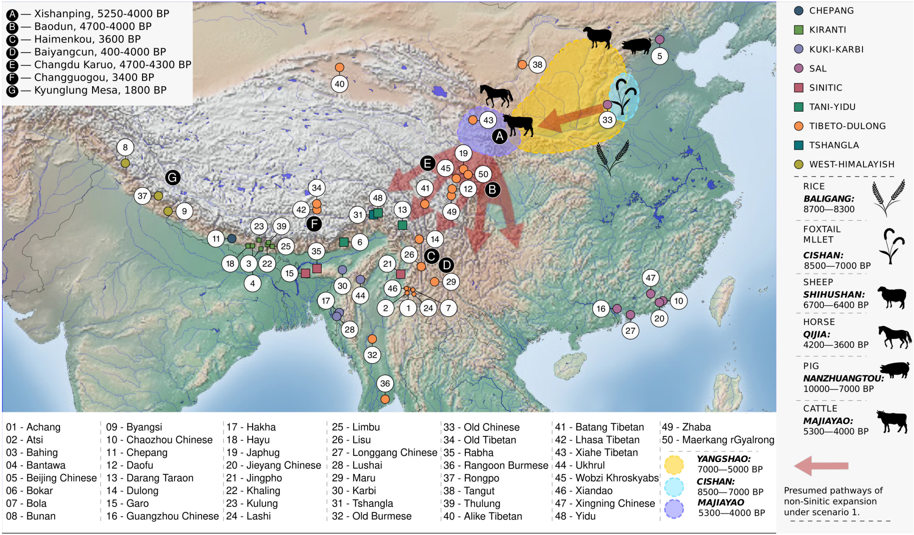
[a]Centre de Recherches Linguistiques sur l'Asie Orientale, CNRS, Institut National des Langues et Civilisations Orientales, Ecole des Hautes Etudes en Sciences Sociales, 75006 Paris, France; [b]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena 07743, Germany; [c]Centre de Recherches en Mathématiques de la Décision, CNRS, Université Paris-Dauphine, PSL University, 75775 Paris, France; and [d]Australian Research Council Center of Excellence for the Dynamics of Language, Australian National University, Canberra, ACT 0200, Australia

Edited by Balthasar Bickel, University of Zurich, Zurich, Switzerland, and accepted by Editorial Board Member Richard G. Klein April 8, 2019 (received for review October 19, 2018)

The Sino-Tibetan language family is one of the world's largest and most prominent families, spoken by nearly 1.4 billion people. Despite the importance of the Sino-Tibetan languages, their prehistory remains controversial, with ongoing debate about when and where they originated. To shed light on this debate we develop a database of comparative linguistic data, and apply the linguistic comparative method to identify sound correspondences and establish cognates. We then use phylogenetic methods to infer the relationships among these languages and estimate the age of their origin and homeland. Our findings point to Sino-Tibetan originating with north Chinese millet farmers around 7200 B.P. and suggest a link to the late Cishan and the early Yangshao cultures.

tions in Chinese date to before 1400 BCE, and Chinese has an abundant and well-studied literature dating back to the early first millennium BCE. The Shāng Kingdom, the Chinese polity associated with these inscriptions, was centered on the lower Yellow River valley. Gradual annexation of neighboring regions and shift of their peoples to the Chinese language led to the striking numerical predominance of Chinese speakers today, and, consequently, to the lack of linguistic diversity in the eastern part of the Sino-Tibetan domain. Tibetan, Tangut, Newar, and Burmese, the family's other early literary languages, were reduced to script considerably more recently: The oldest texts in these languages date from 764 CE, 1070 CE, 1114 CE, and 1113 CE, respectively. The area with the most diverse Sino-Tibetan languages

Sino-Tibetan languages



A — Xishanping, 5250-4000 BP
B — Baodun, 4700-4000 BP
C — Haimenkou, 3600 BP
D — Baiyangcun, 400-4000 BP
E — Changdu Karuo, 4700-4300 BP
F — Changguogou, 3400 BP
G — Kyunglung Mesa, 1800 BP

● CHEPANG
🟩 KIRANTI
● KUKI-KARBI
● SAL
🟥 SINITIC
🟩 TANI-YIDU
● TIBETO-DULONG
🟩 TSHANGLA
● WEST-HIMALAYISH

RICE
BALIGANG:
8700—8300

FOXTAIL
MLLET
CISHAN:
8500—7000 BP

SHEEP
SHIHUSHAN:
6700—6400 BP

HORSE
QIJIA:
4200—3600 BP

PIG
NANZHUANGTOU:
10000—7000 BP

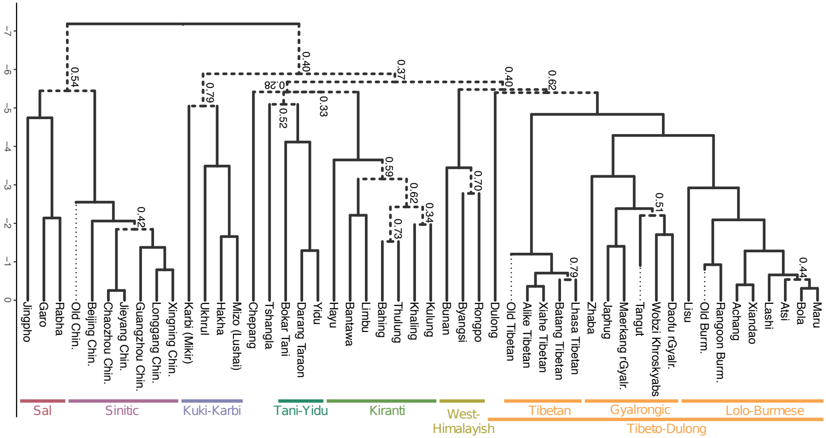
CATTLE
MAJIAYAO:
5300—4000 BP

01 - Achang
02 - Atsi
03 - Bahing
04 - Bantawa
05 - Beijing Chinese
06 - Bokar
07 - Bola
08 - Bunan
09 - Byangsi
10 - Chaozhou Chinese
11 - Chepang
12 - Daofu
13 - Darang Taraon
14 - Dulong
15 - Garo
16 - Guangzhou Chinese
17 - Hakha
18 - Hayu
19 - Japhug
20 - Jieyang Chinese
21 - Jingpho
22 - Khaling
23 - Kulung
24 - Lashi
25 - Limbu
26 - Lisu
27 - Longgang Chinese
28 - Lushai
29 - Maru
30 - Karbi
31 - Tshangla
32 - Old Burmese
33 - Old Chinese
34 - Old Tibetan
35 - Rabha
36 - Rangoon Burmese
37 - Rongpo
38 - Tangut
39 - Thulung
40 - Alike Tibetan
41 - Batang Tibetan
42 - Lhasa Tibetan
43 - Xiahe Tibetan
44 - Ukhrul
45 - Wobzi Khroskyabs
46 - Xiandao
47 - Xingning Chinese
48 - Yidu
49 - Zhaba
50 - Maerkang rGyalrong

YANGSHAO:
7000—5000 BP
CISHAN:
8500—7000 BP
MAJIAYAO
5300—4000 BP

Presumed pathways of
non-Sinitic expansion
under scenario 1.

Example of a (consensus) tree



Questions to answer

- Topology of the tree: subfamilies and their links
- Age of ancestor nodes
- Age of root

Swadesh lists

- 100 or 200 words, present in almost all languages: *bird, hand, to eat, red*...
- Cognacy judgments performed by experts
- "Obvious" borrowings removed

Data are transformed to a binary matrix. This observation process is the first aspect we need to model.

Binary data: *he dies, three, all* (data: Ringe et al. ’02)

	he dies	*three*	*all*
Old English	stierfþ	þrīe	ealle
Old High German	stirbit, touwit	drī	alle
Avestan	miriiete	þrāiiō	vispe
Old Church Slavonic	umĭretŭ	trĭje	vĭsi
Latin	moritur	trēs	omnēs
Oscan	?	trís	súllus

Binary data: *he dies, three, all* (data: Ringe et al. '02)

	he dies	*three*	*all*
Old English	stierfþ	þrīe	ealle
Old High German	stirbit, touwit	drī	alle
Avestan	miriiete	þrāiiō	vispe
Old Church Slavonic	umĭretŭ	trĭje	vĭsi
Latin	moritur	trēs	omnēs
Oscan	?	trís	súllus

Cognacy classes (traits) for the meaning *he dies*:

1. {stierfþ, stirbit}
2. {touwit}
3. {miriiete, umĭretŭ, moritur}

Binary data: *he dies, three, all* (data: Ringe et al. ’02)

	he dies	*three*	*all*
Old English	stierfþ	þrīe	ealle
Old High German	stirbit, touwit	drī	alle
Avestan	miriiete	þrāiiō	vispe
Old Church Slavonic	umĭretŭ	trĭje	vĭsi
Latin	moritur	trēs	omnēs
Oscan	?	trís	súllus

O. English	1	0	0
OH German	1	1	0
Avestan	0	0	1
OC Slavonic	0	0	1
Latin	0	0	1
Oscan	?	?	?

Cognacy classes (traits) for the meaning *he dies*:

1. {stierfþ, stirbit}
2. {touwit}
3. {miriiete, umĭretŭ, moritur}

Binary data: *he dies, three, all* (data: Ringe et al. ’02)

	he dies	*three*	*all*
Old English	stierfþ	þrīe	ealle
Old High German	stirbit, touwit	drī	alle
Avestan	miriiete	þrāiiō	vispe
Old Church Slavonic	umĭretŭ	trĭje	vĭsi
Latin	moritur	trēs	omnēs
Oscan	?	trís	súllus

O. English	1	0	0	1
OH German	1	1	0	1
Avestan	0	0	1	1
OC Slavonic	0	0	1	1
Latin	0	0	1	1
Oscan	?	?	?	1

Cognacy classes for the meaning *three*:

1. {þrīe, drī,þrāiiō, trĭje, trēs, trís}

Binary data: *he dies, three, all* (data: Ringe et al. ’02)

	he dies	*three*	*all*
Old English	stierfþ	þrīe	ealle
Old High German	stirbit, touwit	drī	alle
Avestan	miriiete	þrāiiō	vispe
Old Church Slavonic	umĭretŭ	trĭje	vĭsi
Latin	moritur	trēs	omnēs
Oscan	?	trís	súllus

O. English	1	0	0	1	1	0	0	0
OH German	1	1	0	1	1	0	0	0
Avestan	0	0	1	1	0	1	0	0
OC Slavonic	0	0	1	1	0	1	0	0
Latin	0	0	1	1	0	0	1	0
Oscan	?	?	?	1	0	0	0	1

Cognacy classes for *all*:

1. {ealle, alle}
2. {vispe, vĭsi}
3. {omnēs}
4. {súllus}

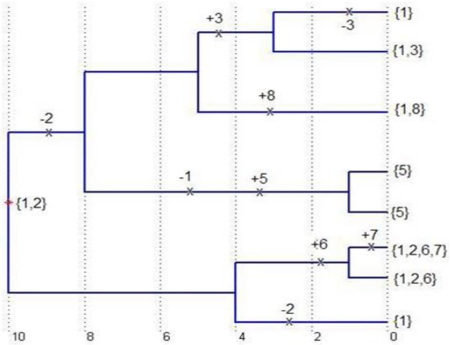
Final data

Old English	1	0	0	1	1	0	0	0
Old High German	1	1	0	1	1	0	0	0
Avestan	0	0	1	1	0	1	0	0
Old Church Slavonic	0	0	1	1	0	1	0	0
Latin	0	0	1	1	0	0	1	0
Oscan	?	?	?	1	0	0	0	1

Constraints

- Constraints on the tree topology
- Constraints on the age of some nodes or ancient languages
- These constraints are used to estimate the evolution rates and the age.
- Also provide one way of validating the model and inference procedure.

Model (1): birth-death process



- Traits (=cognacy classes) are born at rate λ.
- Traits die at rate μ.
- λ and μ are constant.

1	1	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0
3	1	0	0	0	0	0	0	1
4	0	0	0	0	1	0	0	0
5	0	0	0	0	1	0	0	0
6	1	1	0	0	0	1	1	0
7	1	1	0	0	0	1	0	0
8	1	0	0	0	0	0	0	0

Statistical method in a nutshell

1. Collect data
2. Design model
3. Perform inference (MCMC, ...)
4. Check convergence
5. In-model validation (is our inference method able to answer questions from our model?)
6. Model mis-specification analysis (do we need a more complex model?)
7. Conclude

In general, it is more difficult to perform inference for a more complex model.

Limitations of this model

Limitations of this model

1. Constant rates across time and space
2. No handling of missing data
3. No handling of borrowing
4. Treats all traits in the same fashion
5. Binary coding loses part of the structure
6. Assumes a tree structure
7. ...

Do any of these limitations introduce systematic bias?

Limitations of this model

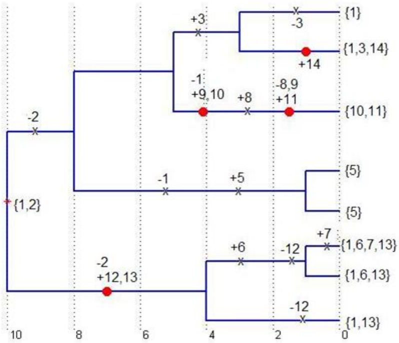
1. Constant rates across time and space
2. No handling of missing data
3. No handling of borrowing
4. Treats all traits in the same fashion
5. Binary coding loses part of the structure
6. Assumes a tree structure
7. ...

Do any of these limitations introduce systematic bias? (Answer: YES, some do.)
Check each misspecification in turn, and adapt the model if necessary.

How to check for misspecifications

1. Compute the Bayes factor to choose between two models $\mathcal{M}_1$ and $\mathcal{M}_2$ (gold standard, but often mathematically challenging and computationally demanding)
2. If a misspecification can be represented by a single estimable parameter θ, estimate it and check whether $\theta = 0$
3. Perform a simulation study: synthesize data from the complex model, infer parameters using the simple model, and check whether we are able to reconstruct the "truth"
4. If for some reason two reasonable models cannot be compared using the above: infer under both, and see where the output agrees.

Model (2): catastrophic rate heterogeneity



- Catastrophes occur at rate ρ
- At a catastrophe, each trait dies with probability κ and *Pbetoiss*(ν) traits are born.
- $\lambda/\mu = \nu/\kappa$: the number of traits is constant on average.

1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	1	1	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	1	1	0	0	0	0	0	1	0
7	1	0	0	0	0	1	0	0	0	0	0	0	1	0
8	1	0	0	0	0	0	0	0	0	0	0	0	1	0

Influence of catastrophes

Checks 2 and 4:

- Estimate the number of catastrophes: posterior distribution is between 0 and 2, with ρ close to 0.
- Infer with and without catastrophes: here we get essentially the same distribution for the parameters of interest (topology, ages)

Conclusion: we can ignore catastrophes. (As it turns out, we will need another kind of rate heterogeneity.)

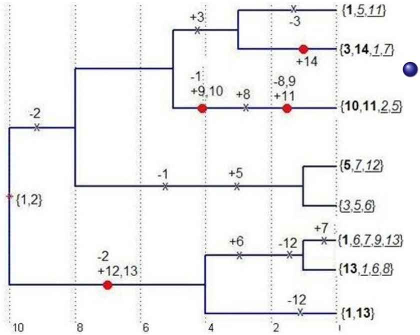
Influence of missing data

Check 3:

- If we simulate synthetic data with missing entries, replace those with 0s, and infer the parameter values, we get biased results

Hence, we need to model missing data

Model (3): missing data



- Observation process: each point goes missing with probability ξ_i
- Some traits are not observed and are thinned out of the data

1	1 0 0 0 ? 0 0 0 0 0 ? 0 0 0
2	? 0 1 0 0 0 ? 0 0 0 0 0 0 ?
3	0 ? 0 0 ? 0 0 0 0 1 1 0 0 0
4	0 0 0 0 ? 0 ? 0 0 0 0 ? 0 0
5	0 0 ? 0 1 ? 0 0 0 0 0 0 0 0
6	1 0 0 0 0 ? ? 0 ? 0 0 0 ? 0
7	? 0 0 0 0 ? 0 ? 0 0 0 0 1 0
8	1 0 0 0 0 0 0 0 0 0 0 0 1 0

Influence of missing data

Check 3: if we simulate synthetic data for which data go missing in blocks, then infer using our simple model of missing data, we get no bias.
Conclusion: this model of missing data is useful enough.

Inference

- BEAST and TraitLab software
- Bayesian inference
- Markov Chain Monte Carlo
- (Almost) uniform prior over the age of the root
- Extensive validation (in-model and out-model; real data and synthetic data)

Posterior distribution

$$
\begin{aligned}
& p(g,\mu,\lambda,\kappa,\rho,\xi|\mathbf{D}=D) \\
= \quad & \frac{1}{N!}\frac{\lambda^N}{\mu^N}\exp\left(-\frac{\lambda}{\mu}\sum_{\langle i,j\rangle\in E} P[\mathcal{E}_Z|Z=(t_i,i),g,\mu,\kappa,\xi](1-e^{-\mu(t_j-t_i+k_iT_C)})\right) \\
& \times\prod_{a=1}^{N}\left(\sum_{\langle i,j\rangle\in E_a}\sum_{\omega\in\Omega_a} P[M=\omega|Z=(t_i,i),g,\mu](1-e^{-\mu(t_j-t_i+k_iT_C)})\right) \\
& \times\frac{1}{\mu\lambda}p(\rho)f_G(g|T)\frac{e^{-\rho|g|}(\rho|g|)^{k_T}}{k_T!}\prod_{i=1}^{L}(1-\xi_i)^{Q_i}\xi_i^{N-Q_i}
\end{aligned}
$$

Likelihood calculation

$$\sum_{\omega\in\Omega_a^{(c)}} P[M=\omega|Z=(t_i,c),g,\mu] =$$

$$\begin{cases} \delta_{i,c}\times \sum_{\omega\in\Omega_a^{(c)}} P[M=\omega|Z=(t_c,c),g,\mu] & \text{if } Y(\Omega_a^{(c)})\geq 1 \\ (1-\delta_{i,c})+\delta_{i,c}v_c^{(0)} & \text{if } Y(\Omega_a^{(c)})Q(\Omega_a^{(c)})=0 \\ & \quad (\textit{i.e. } \Omega_a^{(c)}=\{\varnothing\}) \\ 1-\delta_{i,c}(1-\sum_{\omega\in\Omega_a^{(c)}} P[M=\omega|Z=(t_c,c),g,\mu]) & \text{if } Y(\Omega_a^{(c)})=0 \\ & \text{and } Q(\Omega_a^{(c)})\geq 1 \end{cases}$$

$$\sum_{\omega\in\Omega_a^{(c)}} P[M=\omega|Z=(t_c,c),g,\mu] = \begin{cases} 1 & \text{if } \Omega_a^{(c)}=\{\{c\},\varnothing\} \text{ or } \{\{c\}\} \\ & \quad (\textit{i.e. } D_{c,a}\in\{?,1\}) \\ 0 & \text{if } \Omega_a^{(c)}=\{\varnothing\} \ (\textit{i.e. } D_{c,a}=0) \end{cases}$$

Tests on synthetic data

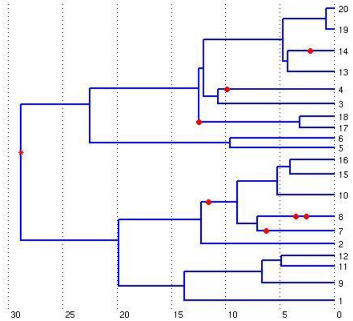


Figure: True tree, 40 words/language

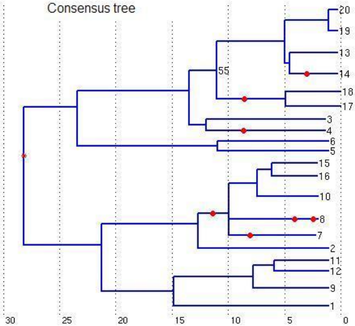


Figure: Consensus tree

With in-model synthetic data, the tree is well reconstructed.

Tests on synthetic data (2)

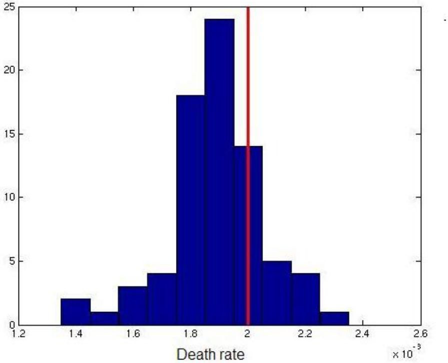


Figure: Death rate (μ)

(Not shown: other parameters are also well reconstructed.)

Influence of borrowing (1)

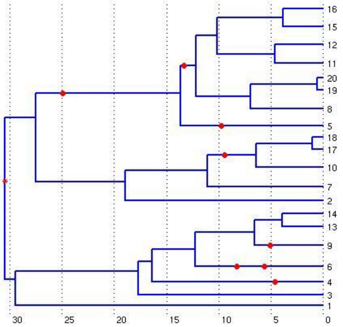


Figure: True tree, 40 words/language, 10% borrowing

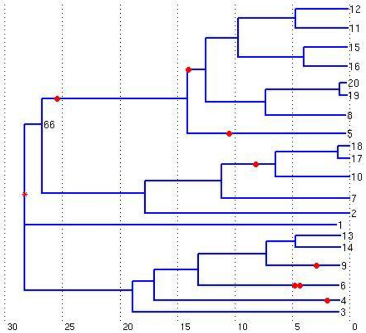


Figure: Consensus tree

With out-of-model synthetic data with low levels of borrowing, the tree is well reconstructed.

Influence of borrowing (2)

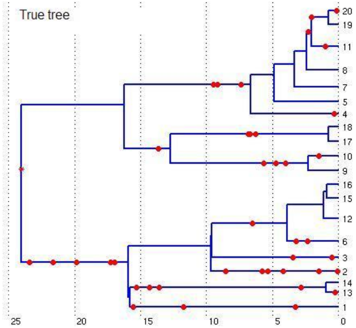


Figure: True tree, 40 words/language, 50% borrowing

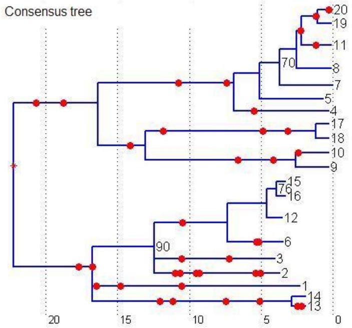


Figure: Consensus tree

Influence of borrowing (3)

- The topology is well reconstructed
- Dates are under-estimated if borrowing levels are high

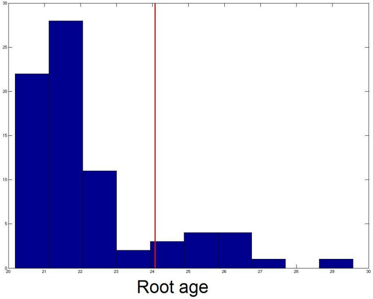


Figure: Root age

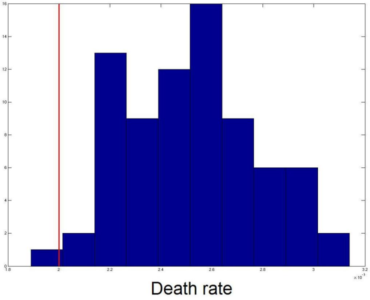


Figure: Death rate (μ)

Borrowing: what to do?

High levels of undetected borrowing (or other non-treeness) would introduce bias in our results.

Fortunately, Kelly & Nicholls (2017) provide a methodology to:

- Essentially infer a network superimposed on a tree
- Test for treeness (check 1)
- Estimate levels of borrowing (allowing check 2 and check 3)

Borrowing: what to do?

High levels of undetected borrowing (or other non-treeness) would introduce bias in our results.

Fortunately, Kelly & Nicholls (2017) provide a methodology to:

- Essentially infer a network superimposed on a tree
- Test for treeness (check 1)
- Estimate levels of borrowing (allowing check 2 and check 3)

Here:

- log Bayes factor is non-decisive at 1.8, indicating that including the network does not improve the model fit
- the level of borrowing is estimated at $\hat{\beta}/\mu = 0.104$, a level at which we have no systematic bias with synthetic data

Analysis restricted to 15 languages (chosen randomly across subfamilies) for computational reasons. Took 83 hours on 8 cores.

Mis-specifications

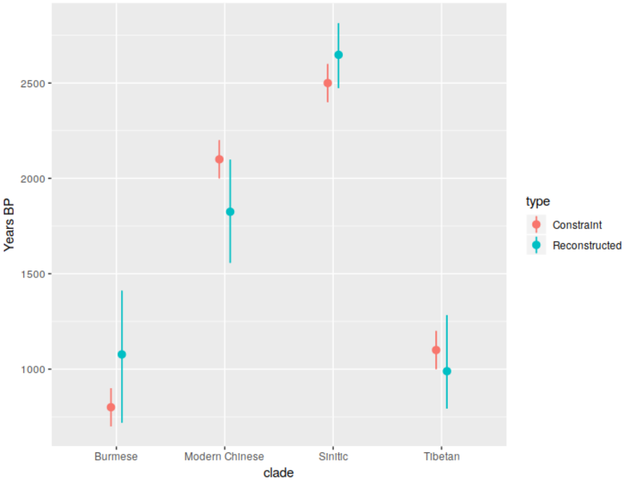
Heterogeneity between traits	Analyse subset of data+ simulated data
Heterogeneity in time/space (non catastrophic)	Infer from 3 distinct models, giving similar results
Borrowing	Bayes factor + Simulated data analysis + check level of borrowing
Data missing in blocks	Simulated data analysis
Non-empty meaning categories	Simulated data analysis
Heterogeneity across subfamilies	Analyse subset of the data
...	...

Cross-validation

Final check: we can take out the age constraints one by one, and check whether we are able to reconstruct them.

Cross-validation

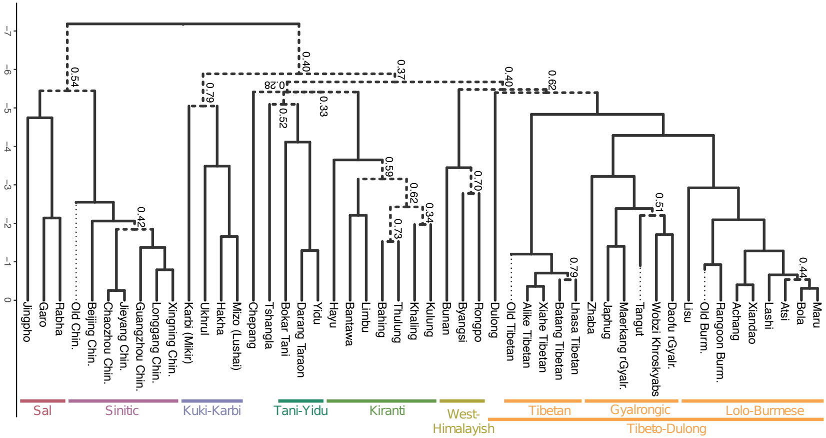
Final check: we can take out the age constraints one by one, and check whether we are able to reconstruct them.



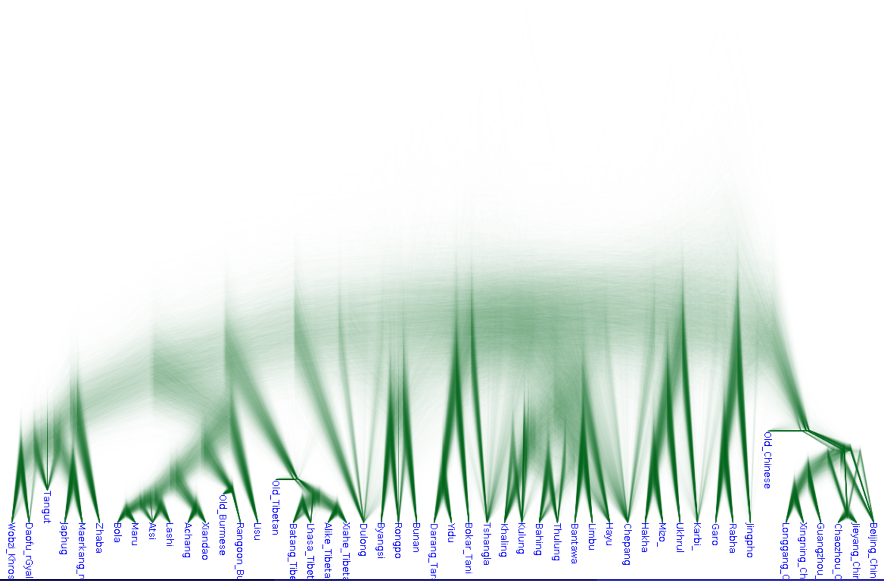
MCMC output

See animation.

Sino-Tibetan consensus tree



Densitree



Wobzi_Khros
Daofu_rGyal
Tangut
Japhug
Maerkang_r
Zhaba
Bola
Maru
Atsi
Lashi
Achang
Xiandao
Old_Burmese
Rangoon_Bu
Lisu
Old_Tibetan
Batang_Tibe
Lhasa_Tibet
Alike_Tibeta
Xiahe_Tibeta
Dulong
Byangsi
Rongpo
Bunan
Darang_Tar
Yidu
Bokar_Tani
Tshangla
Khaling
Kulung
Bahing
Thulung
Bantawa
Limbu
Hayu
Chepang
Hakha
Mizo_
Ukhrul
Karbi_
Garo
Rabha
Jingpho
Old_Chinese
Longgang_C
Xingning_Ch
Guangzhou_
Chaozhou_C
Jieyang_Chi
Beijing_Chin

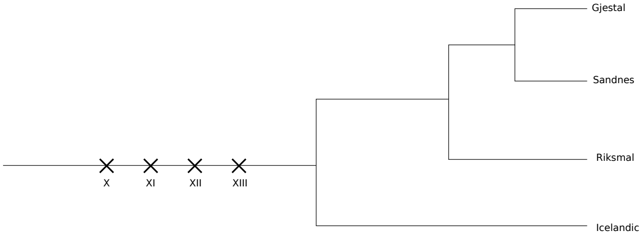
Outline

1. Swadesh: Glottochronology
2. Gray & Atkinson: Language phylogenies
3. Pagel et al.: Frequency of use
4. Sagart et al.: Sino-Tibetan phylogenies
5. Re-examining Bergsland and Vogt
6. Conclusions

Back to Bergsland and Vogt

- Norse family, 8 languages
- Selection bias
- B&V claim that the rate of change is significantly different for these data.
- B&V included words used only in literary Icelandic, which we exclude.
- We can handle polymorphism.
- Do not include rate heterogeneity (would be cheating!)

Known history



Tests

Two possible ways to test whether the same model parameters apply to this example and to Indo-European:

1. Assume parameters are the same as for the general Indo-European tree, and estimate ancestral ages.
2. Use Norse constraints to estimate parameters, and compare to parameter estimates from general Indo-European tree

Results

- If we use parameter values from another analysis, we can try to estimate the age of 13th century Norse.
- True constraint: 660–760 BP. Our HPD: 615 – 872 BP.
- If we analyse the Norse data on its own, we estimate parameters.
- Value of μ for Norse: $2.47 \pm 0.4 \cdot 10^{-4}$
- Value of μ for IE: $1.86 \pm 0.39 \cdot 10^{-4}$ (Dyen et al.), $2.37 \pm 0.21 \cdot 10^{-4}$ (Ringe et al.)

But...

- We can also try to estimate the age of Icelandic (which is 0 BP)
- Find 439–560 BP, far from the true value
- B&V were right: there was significantly less change on the branch leading to Icelandic than average
- However, we are still able to estimate internal node ages.

Georgian

- Second data set: Georgian and Mingrelian
- Age of ancestor: last millenium BC
- Code data given by B&V, discarding borrowed items
- Use rate estimate from analysis of Indo-European (Ringe et al. data)

Georgian

- Second data set: Georgian and Mingrelian
- Age of ancestor: last millenium BC
- Code data given by B&V, discarding borrowed items
- Use rate estimate from analysis of Indo-European (Ringe et al. data)
- 95% HPD: 2065 – 3170 BP

B&V: conclusions

- Third data set (Armenian) not clear enough to be recoded.
- There is variation in the number of changes on an edge.
- Nonetheless, we are still able to estimate ancestral language age.
- Variation in borrowing rates
- B& V: "we cannot estimate dates, and it follows that we cannot estimate the topology either".
- We can estimate dates, and even if we couldn't, we might still be able to estimate the topology.

Outline

1. Swadesh: Glottochronology

2. Gray & Atkinson: Language phylogenies

3. Pagel et al.: Frequency of use

4. Sagart et al.: Sino-Tibetan phylogenies

5. Re-examining Bergsland and Vogt

6. Conclusions

Overall conclusions

- When done right, statistical methods can provide new insight into linguistic history
- Importance of collaboration in building the model and in checking for mis-specification.
- Bayesian statistics play a big role, for estimating uncertainty, handling complex models and using analyses as building blocks
- Accept and embrace the uncertainty
- Major avenues for future research. Challenges in finding relevant data, building models, and statistical inference:
    - Models for morphosyntactical traits
    - Putting together lexical, phonemic and morphosyntactic traits
    - Incorporate geography
    - ...

References

- Swadesh, Morris. "Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos." Proceedings of the American philosophical society (1952): 452-463.
- Gray, Russell D., and Quentin D. Atkinson. "Language-tree divergence times support the Anatolian theory of Indo-European origin." Nature 426.6965 (2003): 435-439.
- Kelly, Luke J., and Geoff K. Nicholls. "Lateral transfer in stochastic Dollo models." The Annals of Applied Statistics 11.2 (2017): 1146-1168.
- Pagel, Mark, Quentin D. Atkinson, and Andrew Meade. "Frequency of word-use predicts rates of lexical evolution throughout Indo-European history." Nature 449.7163 (2007): 717-720.
- Ryder, Robin J., and Geoff K. Nicholls. "Missing data in a stochastic Dollo model for binary trait data, and its application to the dating of Proto-Indo-European." Journal of the Royal Statistical Society: Series C (Applied Statistics) 60.1 (2011): 71-92.
- Ryder, Robin J. "Phylogenetic Models of Language Diversification". DPhil Diss. University of Oxford, UK, 2010.
- Sagart, Laurent, et al. "Dated language phylogenies shed light on the ancestry of Sino-Tibetan." Proceedings of the National Academy of Sciences 116.21 (2019): 10317-10322.

Questions

otázky
spørgsmåler
pytania
preguntas
kláusimai
вопросы
întrebări
vragen
запитанні
domande
questões
vprašanja

kesses
cwestiwnau
preguntes
vrae
Fragen
quaestiones
questions
ερωτήσεις
spurningar
spørsmåler
frågor