Temporal bias:
A new bias for typologists to worry about

Steven Moran (University of Neuchatel)
& Eitan Grossman (Hebrew University of Jerusalem)

Usage-Based Linguistics 5
Tel Aviv University, 5-7 July 2021

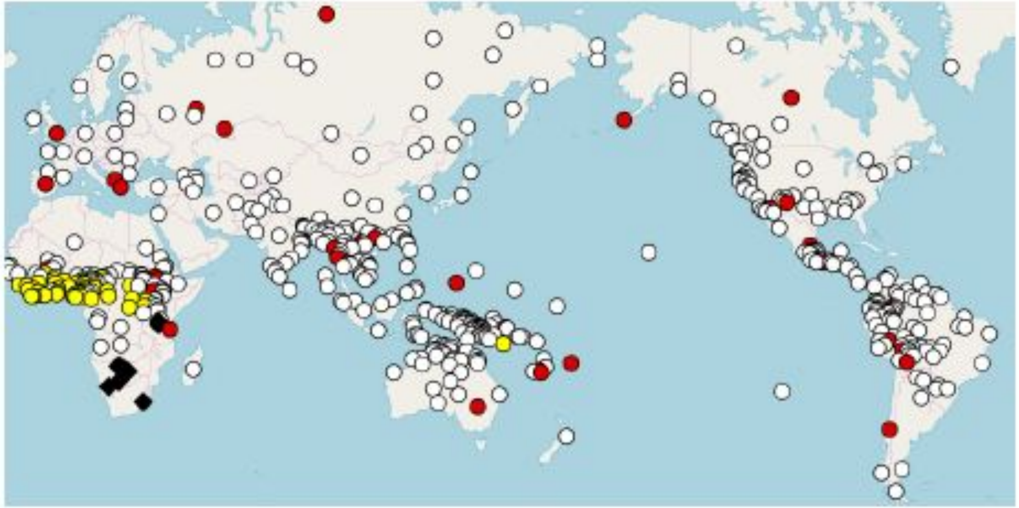
A point of departure

Present-day distributions of linguistic properties may conceal substantive evolutionary changes in human language.

These changes may have occurred even in the relatively recent past.

If this is the case, then linguists should be cautious about making inferences about Language that are read directly from present-day distributions.

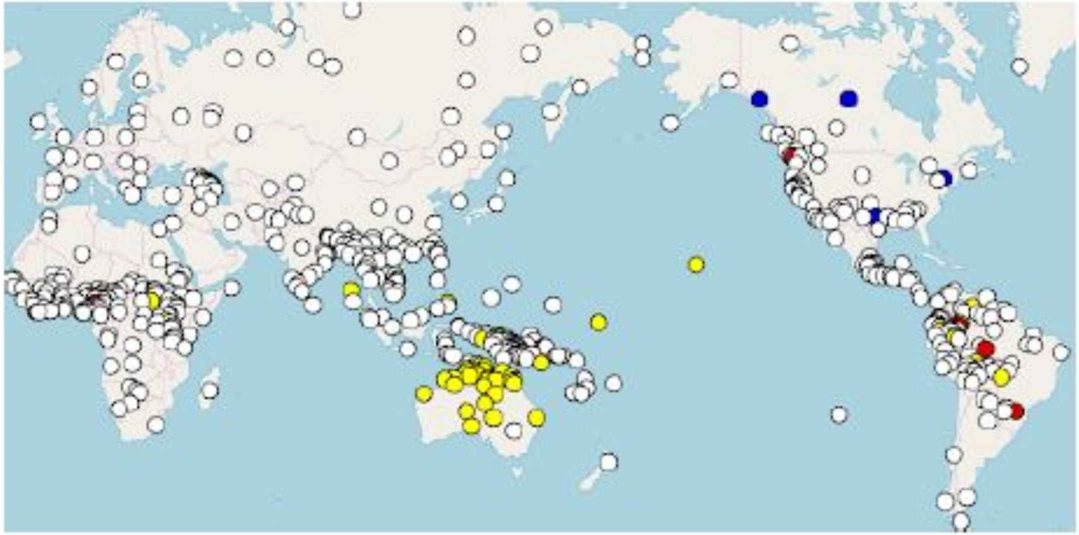
Crosslinguistic distributions are skewed



🔴 front rounded vowels 🟡 labiovelar plosives ◆ clicks

Maddieson 2013 (WALS Online)

Crosslinguistic distributions are skewed



● no bilabials ● no fricatives ● no nasals

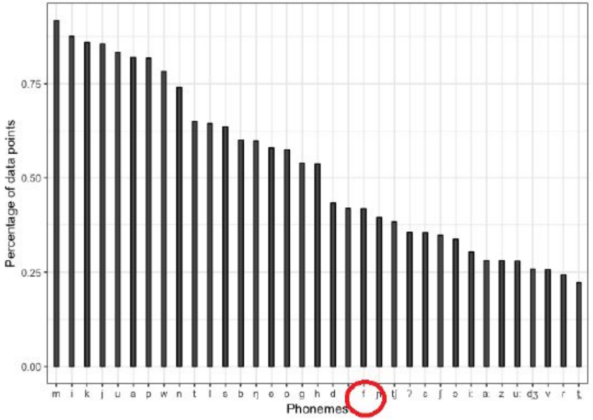
Maddieson 2013 (WALS Online)

Frequency distributions

Two labiodental fricatives, /f/ and /v/, are among the most common phonological segments in the world’s languages.

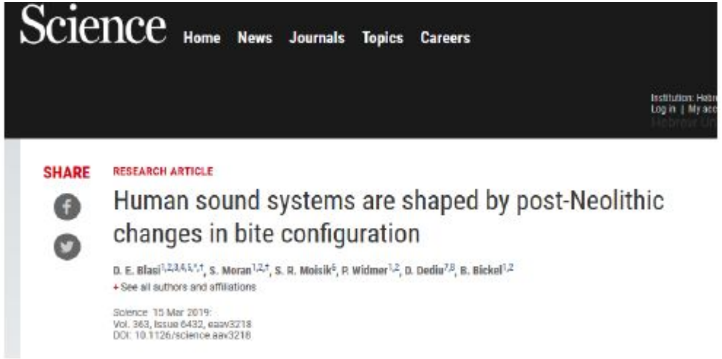
- /f/ is the 21st most frequent segment in the world’s languages (around 40%), /v/ is the 33rd (around 25%).
- /f/ is the 16th most frequent consonant segment in the world’s languages, /v/ is the 23rd.

So maybe they’re ‘unmarked’ in some way…



Recent substantive changes (Blasi et al. 2019)

- Labiodentals are late in human evolution.
- They postdate the advent and spread of agriculture several thousand years ago.
- Languages spoken by hunter-gatherer societies tend not to have labiodentals, in comparison with agriculturalist societies.
- Labiodentals are rarely reconstructed for proto-languages, particularly those at relatively great time-depths.



RESEARCH ARTICLE

Human sound systems are shaped by post-Neolithic changes in bite configuration

D. E. Blasi[1,2,3,4,5,*,†], S. Moran[1,2,†], S. R. Moisik[6], P. Widmer[1,2], D. Dediu[7,8], B. Bickel[1,2]

+ See all authors and affiliations

Science 15 Mar 2019:
Vol. 363, Issue 6432, eaav3218
DOI: 10.1126/science.aav3218

From distributions to ‘naturalness’?

It would be nice if linguists could read off the cognitive, communicative, or biological ‘goodness’ or ‘naturalness’ of a linguistic type from the empirical frequency of that type (Cysouw 2011).

The possibility to draw valid inferences of this sort depends, however, to a large extent on some version of the **Uniformitarian Principle**.

The Uniformitarian Principle

The Uniformitarian Principle has been interpreted in a number of ways in linguistics.

- ‘the forces which operated to produce the historical record are the same as those which can be seen operating today’. (Labov 1974/1978)
- ‘what we can reconstruct is … limited by our empirical knowledge of things that occur in present day languages’. (Lass 1978)
- ‘the general processes and principles which can be noticed in observable history are applicable in all stages of language history’. (Hock 1991)

For extensive and nuanced surveys of uniformitarianism, see Janda & Joseph (2003) and Walkden (2019).

The Implicit Uniformitarian Assumption

‘[H]uman languages have always been pretty much the same in terms of the typological distribution of the units that compose them’ (Newmeyer 2002: 300).

The Implicit Uniformitarian Assumption

Throughout the history of what linguists call ‘human language,’ cross-linguistic distributions of linguistic properties, whether simple or complex, have always been more or less the same.

In particular, linguistic properties that are currently rare have always been rare, and linguistic properties that are currently frequent have always been frequent.

That is, cross-linguistic distributions are time-independent (Moran et al. 2021).

Reasons for skepticism

1. Relatively small number of top-level families, many of which are isolates. (Hammarström 2016)
2. Cross-linguistic distributions have not yet reached a stationary distribution. (Maslova 2000, 2002)
3. Language contact scaling up to continent-sized areas. (Dryer 1989, Nichols 1992)
4. The number of present-day language families and areas does not provide enough independent data points to infer universal properties of language. (Piantadosi & Gibson 2014)

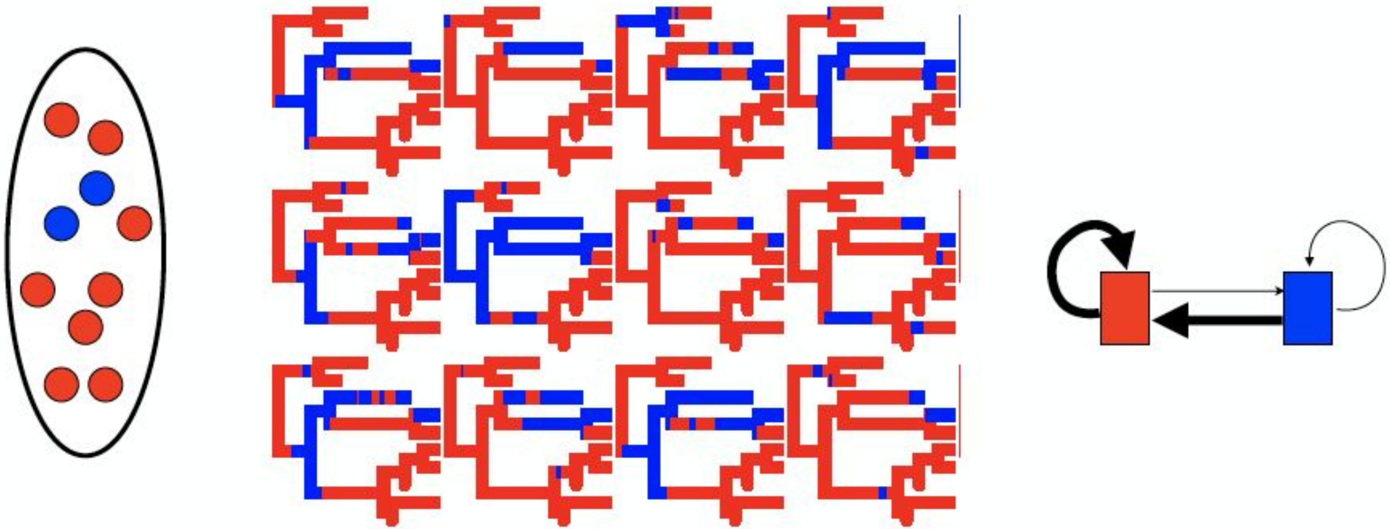
Present-day distributions are at least to some extent artefacts of inheritance and of language-external historical events, such as language spread and contact.

More reasons for skepticism

Non-linguistic factors may shape language structure in ways that influence cross-linguistic distributions.

1. Speech community size and other aspects of demographics (Trudgill 1989, Nettle 1996, Lupyan & Dale 2010, and many more)
2. Genes and aspects of speech-relevant anatomy (Dediu & Ladd 2007, Creanza et al. 2015, Dediu et al. 2017, Moisik & Dediu 2017))
3. Geography and other environmental factors (Everett 2013, Everett et al. 2015, Maddieson & Coupe 2016)
4. Technology, in particular food production technology (Blasi et al. 2019)
5. Aspects of culture, including sexual mores (Ember & Ember 2007)

From Balthasar Bickel’s talk



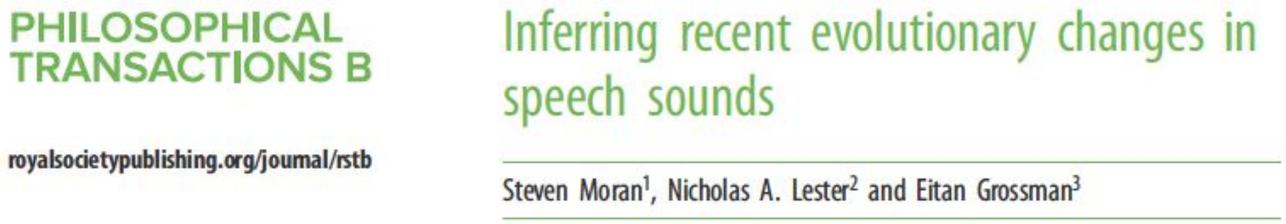
Dunn et al 2011, Widmer et al 2017

Balthasar Bickel’s talk

What we have is a degenerate sample of linguistic diversity, possibly reflecting post-Neolithic population history rather than Language.

- Forget post-hoc generalizations [based on present-day cross-linguistic distributions].
- Focus on mechanisms [of change and the causal factors/triggers that bias transition probabilities in one direction or another].

A story…



PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Inferring recent evolutionary changes in speech sounds

Steven Moran[1], Nicholas A. Lester[2] and Eitan Grossman[3]

We were interested in investigating the short-term evolution of phonological systems over the past few thousand years.

We wanted to do this based on empirical evidence -- that is, we wanted to compare the distribution of speech sounds in present-day and ancient or reconstructed languages.

We have the relevant databases: PHOIBLE (Moran & McCloy 2019), on the one hand, and BDPROTO (Moran et al. 2020), on the other.

PHOIBLE

A repository of cross-linguistic phonological inventory data (Moran 2012).

- Includes 3020 inventories that contain 3183 segment types found in 2186 distinct languages (Moran & McCloy 2019).
- Currently the most comprehensive cross-linguistic database on phonological inventories, which is openly available.

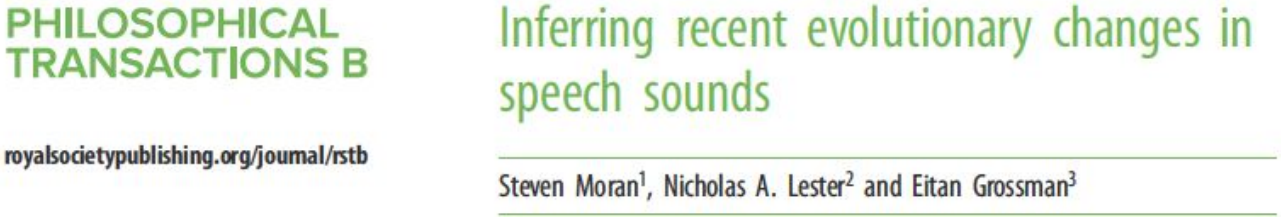
https://phoible.org/

BDPROTO

A database of 257 phonological inventories from ancient and reconstructed languages that were extracted from historical linguistic reconstructions and then interpreted by experts (Marsico et al. 2018, Moran et al. 2020).

https://github.com/bdproto/bdproto

A story…



PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Inferring recent evolutionary changes in speech sounds

Steven Moran[1], Nicholas A. Lester[2] and Eitan Grossman[3]

We also wanted to see to what extent recent language contact (borrowing events) have shaped present-day distributions, based on SEGBO.

But we soon realized that there is a data problem in comparing present-day and past distributions of linguistic properties (of any sort).

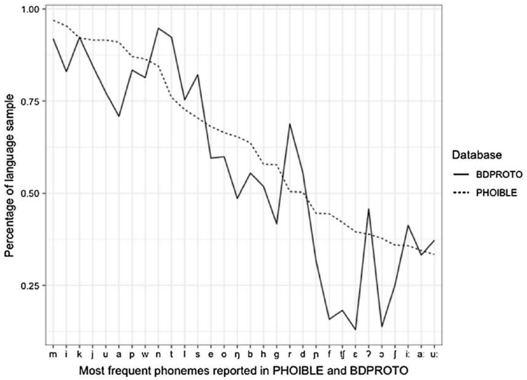


Fig. 3 The 30 most frequency phonemes reported in PHOIBLE compared with BDPROTO

Sampling in typology

The traditional goal of typology is to find **universals of language** (post hoc generalizations) and to explain them.

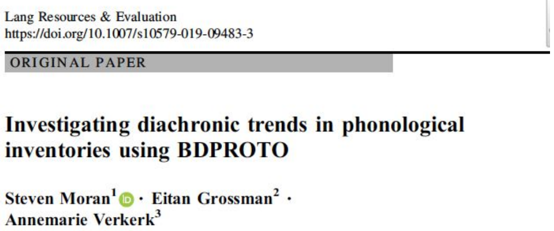
Since there are numerous ways in which languages can be non-independent data points, typologists have traditionally resorted to **sampling**.

In particular, typologists have tried to create maximally stratified language samples.

Biases in sampling

- Genealogical bias (Greenberg 1966, Bell 1978)
- Geographical bias (Greenberg 1966, Bell 1978)
- Typological bias (= dependencies between variables, Bell 1978)
- Cultural bias (Perkins 1992)
- Bibliographical bias (Bakker 2011)
- And a few more…

Temporal bias


Lang Resources & Evaluation
https://doi.org/10.1007/s10579-019-09483-3

ORIGINAL PAPER

Investigating diachronic trends in phonological inventories using BDPROTO

Steven Moran[1] · Eitan Grossman[2] · Annemarie Verkerk[3]

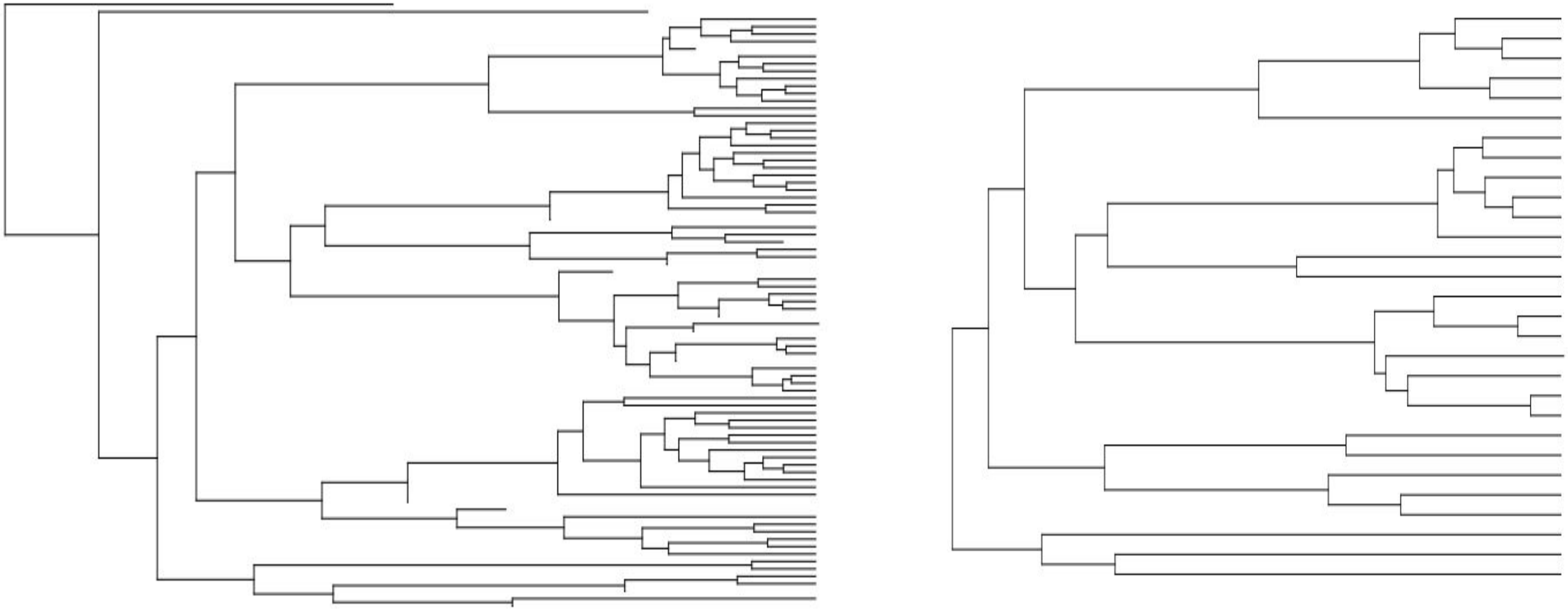
A set of problems that complicate the empirical study of cross-linguistic distributions in the past.

Basically stems from the fact that each language exists at a particular time.

Its main effect is to compound already-existing problems of data sparsity.

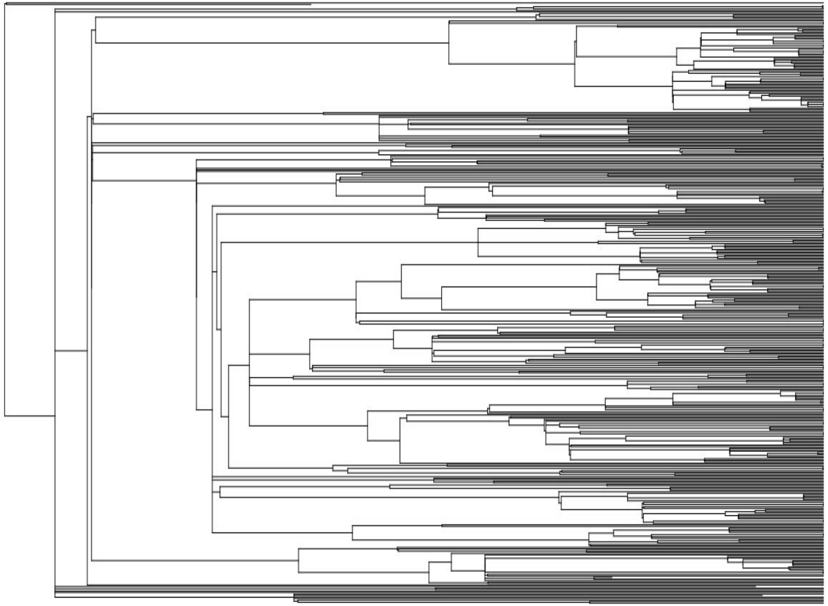
Data sparsity

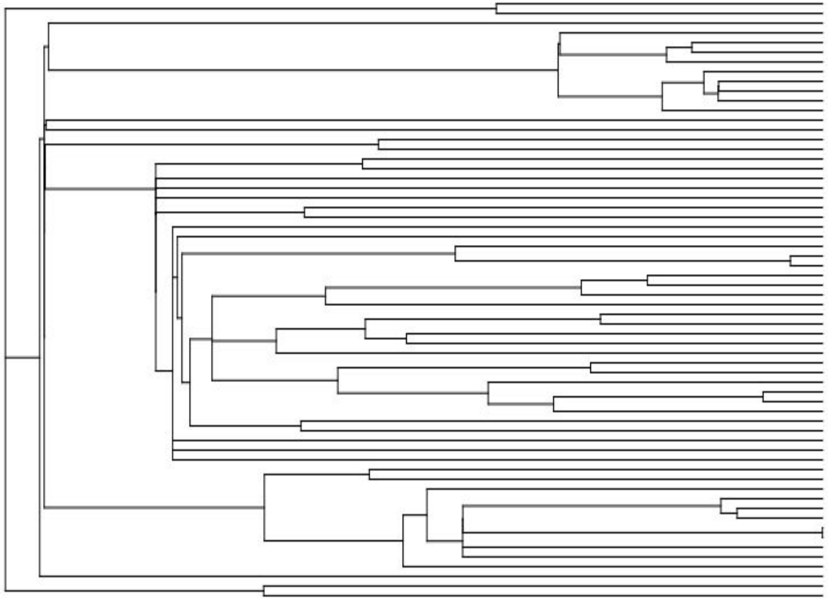
Indo-European full tree (left) vs pruned tree with existing data points (PHOIBLE)



Data sparsity

Austronesian full tree (left) vs pruned tree with existing data points (PHOIBLE)





Temporal bias (1): Chronological bias

Any database of past and reconstructed languages might include languages spoken at various time depths, like:

- Early Modern English (400-500 YBP)
- Classical Latin (1800-2100 YBP)
- Proto-Germanic (2500 YBP)
- Old Egyptian (4600 YBP)
- Proto-Dravidian (5000 YBP)

For each point in time, our data are increasingly sparse.

Temporal bias (2): Phylogenetic depth bias

Language families are heterogeneous in terms of their size and diversification.

For example, one stock may have many nodes (e.g., Indo- European), while another may have very few (e.g., Basque).

This means that the comparison of ancient and reconstructed languages may be like comparing the branches of an angel oak tree with those of a birch tree:

(a) An angel oak tree

(b) Some birch trees

Temporal bias

If we want to study distributions of linguistic properties in the past for particular periods, we would either need:

1. To eliminate nodes from the more diversified families or
2. To inflate the genealogical complexity of the less diversified families.

Both of these would exacerbate the already-acute problem of data sparsity when existing genealogical and geographical diversity is taken into account.

Mitigating or avoiding temporal bias

The answer isn’t found in better sampling techniques.

Focus on **mechanisms** - or, more broadly, **dynamics of change**.

Case study: evolutionary rates of C and V systems

- Languages dating back as far as 10kya are equally complex (Marsico 1999)
  - number of segments, consonant/vowel ratio, average number of consonants and vowels, and frequency hierarchy of the segments
- We expect reconstructed phonological systems are likely to show more complexity than their daughter languages, due to inherent biases of the comparative method
  - Fox (1995): some reconstructions of the Proto-Indo-European consonant system contain more consonants than any of the daughter languages
- However, modern languages tend to have slightly more consonants than their ancestors did in the past; does not apply to vowels
  - Consonants and vowels across proto-languages in BDPROTO are 18 and 8
  - Modern spoken languages have on average 22 consonants and 8 vowels (Maddieson 1984)
- Why is it that we observe more consonants in phonological inventories today than we see in reconstructed ancient languages of the past?

Case study: evolutionary rates of C and V systems

- Tested whether eight language families show greater rates of change in consonant inventory size as compared to vowel inventory size using phylogenetic comparative methods

  - Arawakan (Walker & Ribeiro, 2011)
  - Austronesian (Gray et al., 2009)
  - Bantu (Grollemund et al., 2015)
  - Dravidian (Kolipakam et al. 2018)
  - Indo-European (Bouckaert et al., 2012).
  - Pama-Nyungan (Bowern & Atkinson, 2012)
  - Tupi-Guarani (Michael et al., 2015)
  - Turkic (Hruschka et al., 2015)

Case study: evolutionary rates of C and V systems

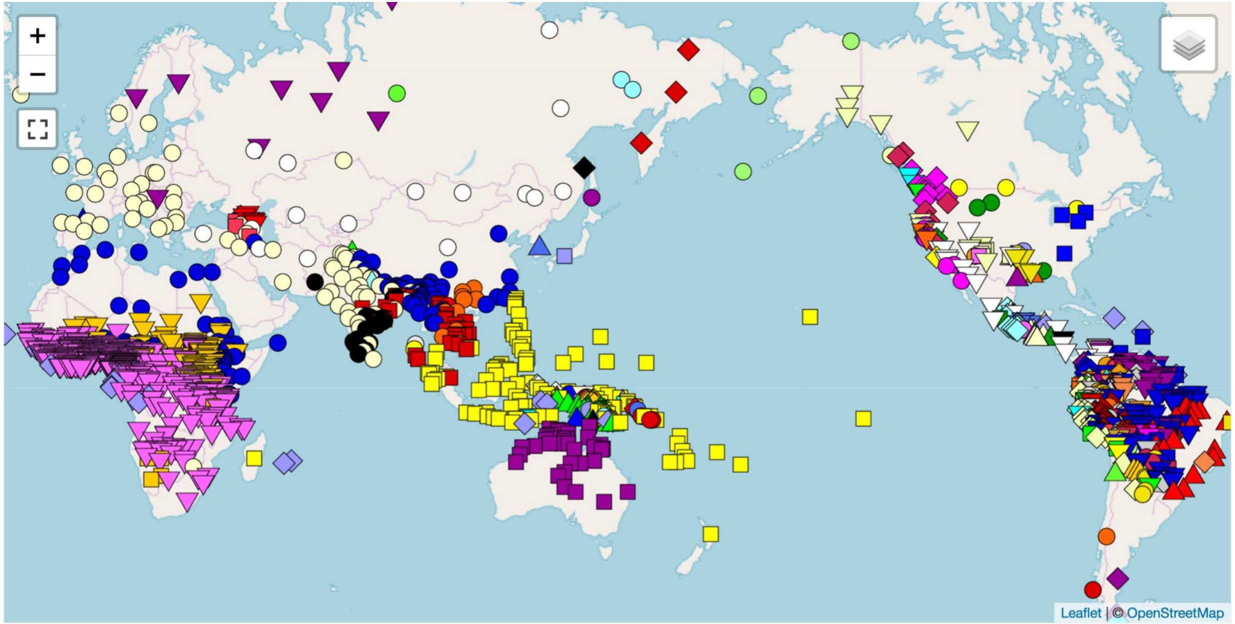
- Determined the number and composition of consonants and vowels in phonological inventories from PHOIBLE and BDPROTO
    - Each segment is encoded with a set of 37 distinctive phonetic features
- Map PHOIBLE data as traits to high-resolution phylogenetic trees
- Computational phylogenetic analysis
    - BayesTraits V3 (Meade and Pagel 2014)
    - generalized least squares approach to modeling the evolution of continuously varying traits (Pagel 1997, 1999) and multistate traits (Pagel et al. 2004)

Case study: evolutionary rates of C and V systems

1. Determined the number and composition of consonants and vowels in phonological inventories from PHOIBLE and BDPROTO

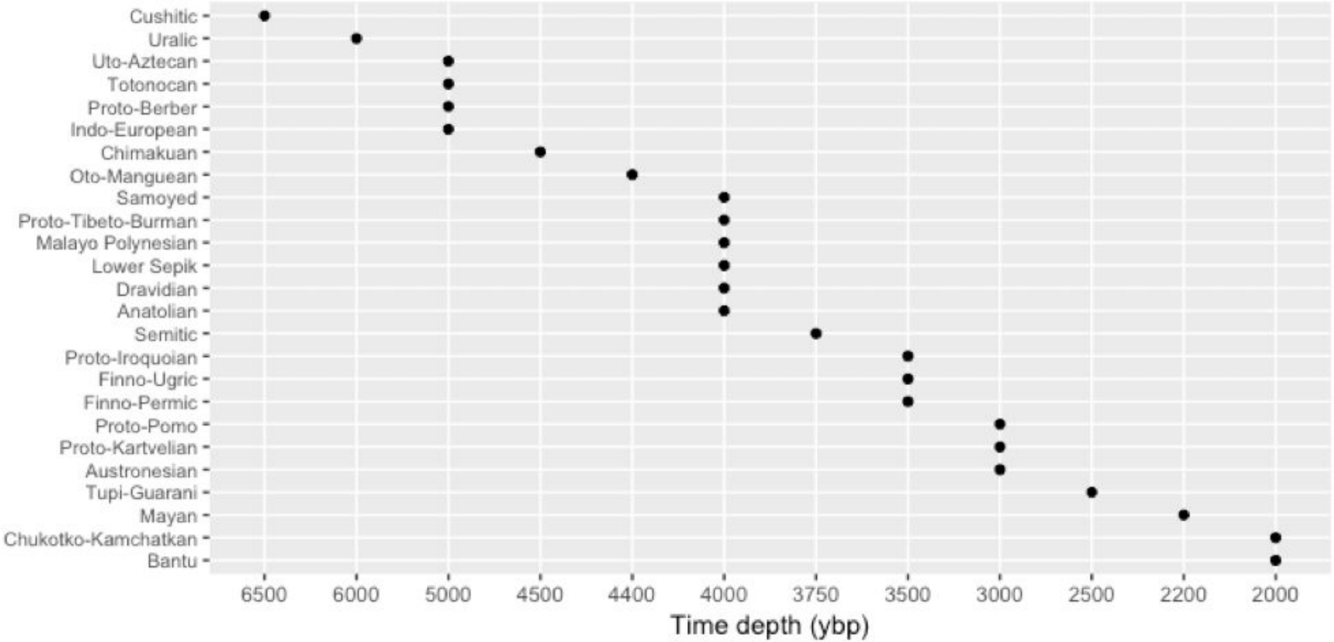
Case study: evolutionary rates of C and V systems

- PHOIBLE Online (Moran et al., 2014): repository of cross-linguistic phonological inventory data (n=1672)



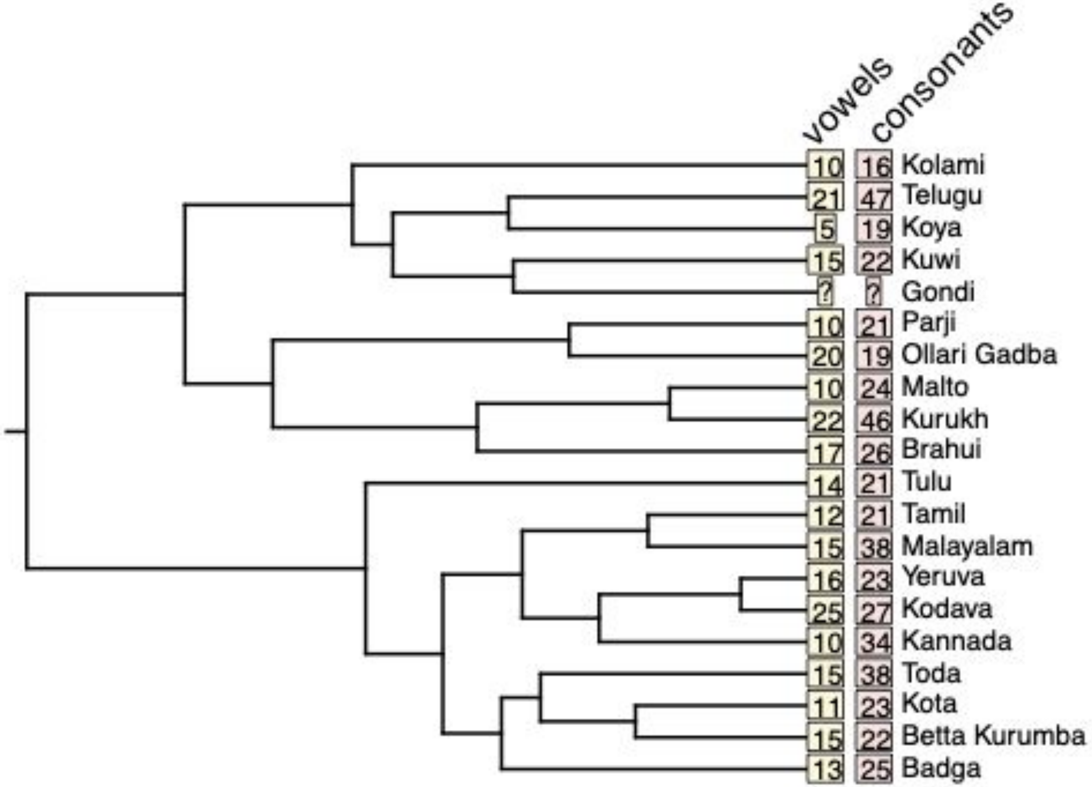
Case study: evolutionary rates of C and V systems

- BDPROTO (Marsico et al., 2018): ancient and reconstructed languages' phonological inventories (n=132)



Case study: evolutionary rates of C and V systems

- Dravidian



Case study: evolutionary rates of C and V systems

1. Determined the number and composition of consonants and vowels in phonological inventories from PHOIBLE and BDPROTO
2. Use standard computational phylogenetic analysis techniques (Pagel 1997, 1999, Pagel et al. 2004) with BayesTrait V3 (Meade and Pagel 2014)

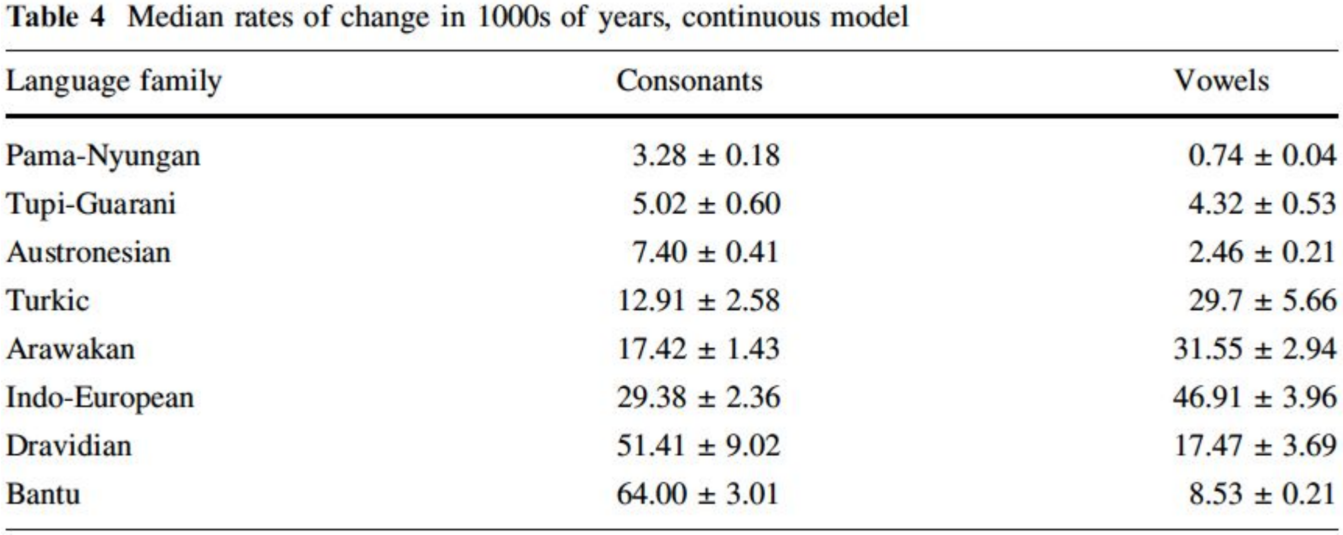
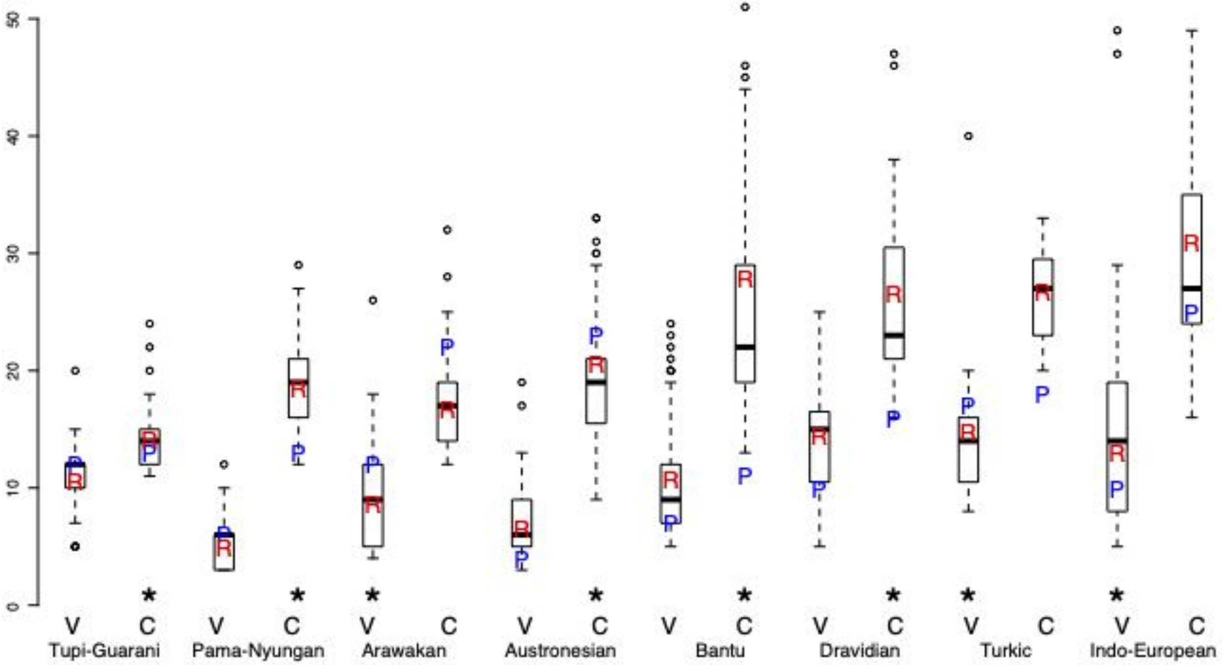


Table 4 Median rates of change in 1000s of years, continuous model

Language family	Consonants	Vowels
Pama-Nyungan	3.28 ± 0.18	0.74 ± 0.04
Tupi-Guarani	5.02 ± 0.60	4.32 ± 0.53
Austronesian	7.40 ± 0.41	2.46 ± 0.21
Turkic	12.91 ± 2.58	29.7 ± 5.66
Arawakan	17.42 ± 1.43	31.55 ± 2.94
Indo-European	29.38 ± 2.36	46.91 ± 3.96
Dravidian	51.41 ± 9.02	17.47 ± 3.69
Bantu	64.00 ± 3.01	8.53 ± 0.21

Case study: evolutionary rates of C and V systems

Ranges, reconstructions, and ancestral state estimations of vowel and consonant inventory size



Case study: evolutionary rates of C and V systems

- The evolution of phonological inventories does not follow a universal pathway.
- We find differential rates of change in consonant and vowel inventories across different language families.
- In contrast to the null hypothesis -- no relationship between family and rate of change.

Case study: evolutionary rates of C and V systems

- This finding is in line with previous work on rates of change and typological stability:
    - Some typological features are far more stable than others (Wichmann & Holman 2009, Dediu 2011, Greenhill et al., 2017)
    - Pathways leading to high diversity among related languages (implying high rates of change) have not been systematically researched
    - Phonological change may be dependent not only on language-internal processes but also on environmental factors (Everett et al., 2015, Everett 2017, Roberts 2018)

Case study: evolutionary rates of C and V systems

- We cannot assume uniform rates of change across all languages across all areas of grammar (cf. Greenhill et al., 2010, Greenhill et al., 2018, Verkerk, 2015) and further research on the relevant dynamics is in order.

Case study: evolutionary rates of C and V systems

- Family-specific rates of change.
- Each family changes in a particular direction, i.e., no unbiased families.

These findings can be interpreted in light of Bickel’s (2013) proposed Family Bias Method.

Case study: evolutionary rates of C and V systems

Bickel (2013): when there is a cross-family preference for change while individual families show particular preferences:

(i) the operation of universal pressures leading to preferred pathways or directions of change, and

(ii) the relative strength of these pressures.

In the present case, our findings provide preliminary evidence for both pressures that lead to an expansion of consonant inventories and pressures that lead to an expansion of vowel inventories.

These pressures must be relatively strong, since no unbiased families were found in our sample, but it is hard to say which pressures are stronger, because the respective proportions of vowel- increasing and consonant-increasing families in the sample are very close (5/8 vs. 3/8).

Case study: evolutionary rates of C and V systems

Such pressures might include:

- Differential functional load of vowels and consonants in maintaining lexical distinctions (possibly favoring higher rates of change in vowels).
- Differential borrowability of vowels and consonants (favoring higher rates of change in consonants).
- Number and type of dimensions through which vowels and consonant inventories expand (vowel inventories tend to grow more economically).
- Other language-external -- socio-cultural, environmental, biological -- factors.

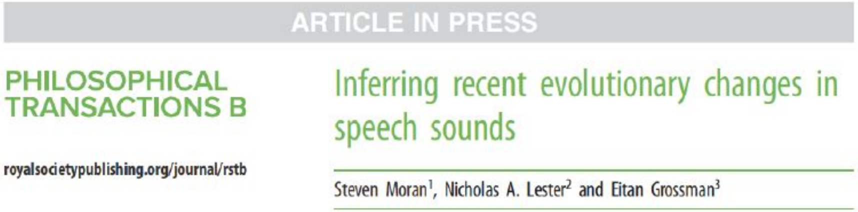
In case you were wondering…

Phonological inventories have indeed changed substantially over the past few millennia.

The segments whose distribution changed the most correlates strongly with the most frequently borrowed segments in SEGBO.

Points to the possible role of recent language spreads in shaping phonological inventories.

No support for the Implicit Uniformitarian Assumption.



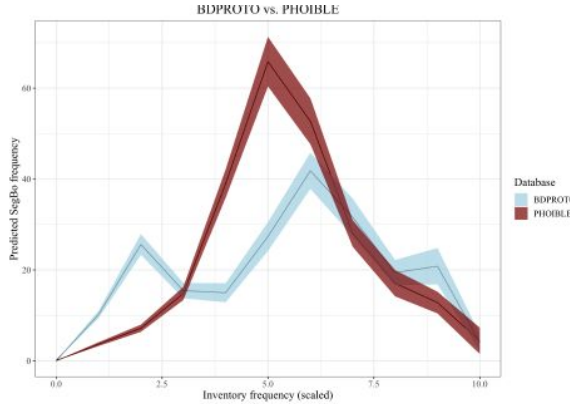
ARTICLE IN PRESS

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Inferring recent evolutionary changes in speech sounds

Steven Moran[1], Nicholas A. Lester[2] and Eitan Grossman[3]



Conclusions

We cannot simply assume that the distribution of linguistic properties in the past is identical to the present-day distributions.

That means we have to care about temporal bias when studying distributions in the past.

The way forward is dense (not stratified) samples, better statistical methods, and targeting mechanisms and causes of change.

Thank you!

steven.moran@unine.ch

eitan.grossman@mail.huji.ac.il

We can at this point, however, avoid genealogical and temporal biases by investigating specific language families, regardless of their age, by comparing certain properties of proto-languages directly with the currently available data of its daughter languages. Moreover, although we cannot give precise phonetic values for reconstructed sounds from thousands of years ago, we can be fairly certain that proto-languages derived through the historical-comparative method provide us with a fairly accurate number of contrastive sounds in the phonological inventories that the parent languages would have had. Hence, we can compare whether ancient and modern languages have shifted in their number of sounds, and in which direction, over the millennia.