Universals of phonological segment borrowing? Questions, evidence, methods, findings

Eitan Grossman
Hebrew University of Jerusalem

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM
الجامعة العبرية في اورشليم القدس

43 Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS)
February 23–26, 2021

Today's talk is based on joint work with...

Dmitry Nikolaev

(Stockholm University)

Steven Moran

(University of Neuchatel)

Elad Eisen

(Hebrew University)

A roadmap for today's talk

1. Cross-linguistic diversity in phonological segment inventories
2. Questioning the uniformitarian hypothesis
3. SegBo: A world survey of phonological segment borrowing
4. Targeted research questions
    - Are there universals of borrowing?
    - To what extent does borrowing show macro-areal patterning?
    - To what extent has borrowing shaped present-day distributions of phonological segments?
5. Conclusions

A few words on cross-linguistic diversity in phonological segment inventories

Some languages have small consonant inventories

	bilabial	alveolar	velar
voiceless	p	t	k
voiced	b	d	g

Central Rotokas
(Non-Austronesian Papuan, Bougainville)

Some language have big consonant inventories

Table 7: Consonant chart, using Nakagawa's (2006) orthography

Series		Extended Place of Articulation										
		Lb	Dt	Dt-Af	Dt-Af-Cl	Al-Cl	Al-Af (Lt)-Cl	Pl	Pl-Cl	Vl	Uv	Gl
Stop segments	plain	p	t	ts	|	!	‖	(c)	ǂ	k	q	ʔ
	voiced	b	d	dz	g|	g!	g‖	(ɟ)	gǂ	g	(ɢ ʔ)	
	voiceless ejective		t’	ts’	|’		‖’	(c’)	ǂ’	k’		
	voiceless aspirated		t^h	ts^h	$\|^h$	$!^h$	$\|\|^h$	(c^h)	$ǂ^h$	k^h		
Stop cluster	plain + x		tx	tsx	|x	!x	‖x	(cx)	ǂx			
	plain + q				|q		‖q	(qy)	ǂq			
	plain + ɢ				|ɢ		‖ɢ		ǂɢ			
	plain + ʔ				|ʔ	!ʔ	‖ʔ	(ʔy)	ǂʔ			
Nasal	voiced	m	n		ŋ|		ŋ‖	(ɲ)	ŋǂ	ŋ		
Pre-nasalised		mb	nd		ŋ|g	n!g	ŋ‖g	(ɲɟ)	ŋǂg	ŋg		
Fricative	voiceless		s							x		h
Tap or Flap			(ɾ)									
Glides		w						j				

(Abbreviations used: Lb = labial, Dt = dental, Dt-Af = dental affricate, Dt-Af-Cl = dental affricate click, Al-Cl = alveolar click, Al-Af (Lt)-Cl = alveolar affricate lateral click, Pl = palatal, Pl-Cl = palatal click, Cl = velar, Uv = uvular, Gl = glottal)

Ts’ixa (Kalahari Khoe, Fehn 2014)

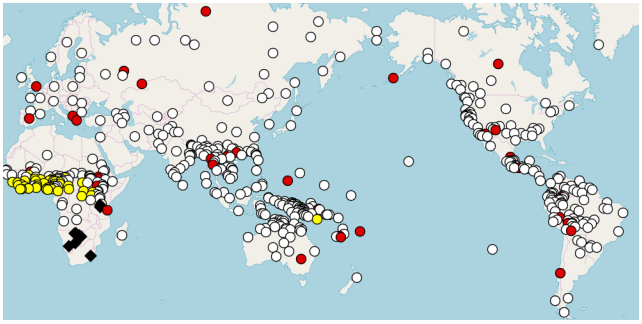
Most languages are somewhere in the middle

	bilabial	labio-dental	alveolar	post-alveolar	palatalized	palatal	velar	glottal
plosive	p		t		t^j		k	
nasal	m		n		n^j			
trill			r					
fricative		f v	s	ʃ	s^j			h
lateral			l		l^j			
approximant						j		

Estonian (Uralic)

Cross-linguistic distributions are skewed

Clicks, labiovelar stops, and front rounded vowels are cross-linguistically rare, ...

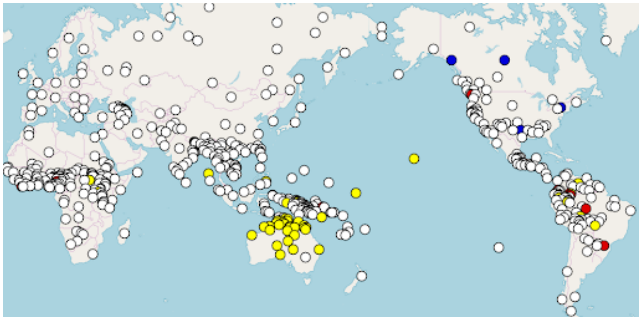


● front rounded vowels ● labiovelar plosives ◆ clicks

Maddieson 2013 (WALS Online)

Cross-linguistic distributions are skewed

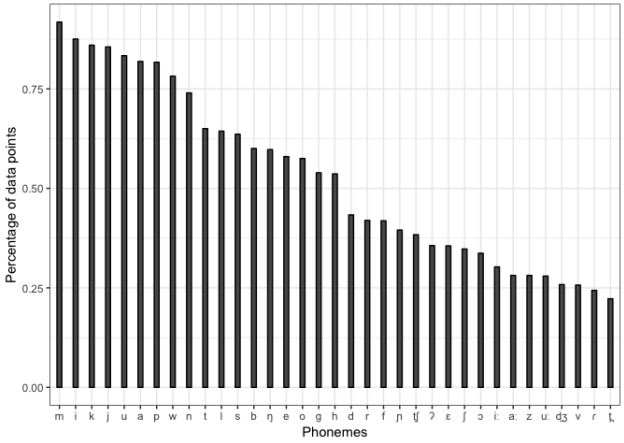
... while bilabials are present in nearly every language.



● no bilabials ● no fricatives ● no nasals

Maddieson 2013 (WALS Online)

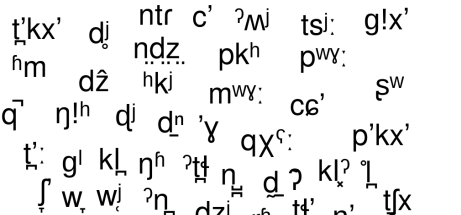
Some sounds are very common



The 35 most frequent segments in PHOIBLE 2.0

(Moran & McCloy 2019)

Some segments are attested only once



(Grossman & Moran 2019)

Explanations for skewed distributions

- Ease of production?
- Ease of perception?
- Structural properties of sound systems, e.g., symmetry, dispersion, feature economy?
- Acquisition biases?
- Universal grammar (e.g., markedness)?

From distributions to universals of language

But can we draw inferences about Language based directly on empirical frequencies or present-day distributions?

From distributions to ‘naturalness’

It would be ideal if linguists could infer the universal probability (or ‘learnability’) 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)

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)

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)

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)

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)

Point to the possibility that present-day distributions are at least to some extent artefacts of inheritance, on the one hand, and of language-external historical events, such as language contact, on the other.

Furthermore

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 & Coupé 2016)
4. Technology, in particular food production technology (Blasi et al. 2019)
5. Aspects of culture, including sexual mores (Ember & Ember 2007)

A brief example

Labiodental fricatives

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.

In present-day distributions, labiodental fricatives are among the most frequent sounds in the world's languages.

Labiodental fricatives

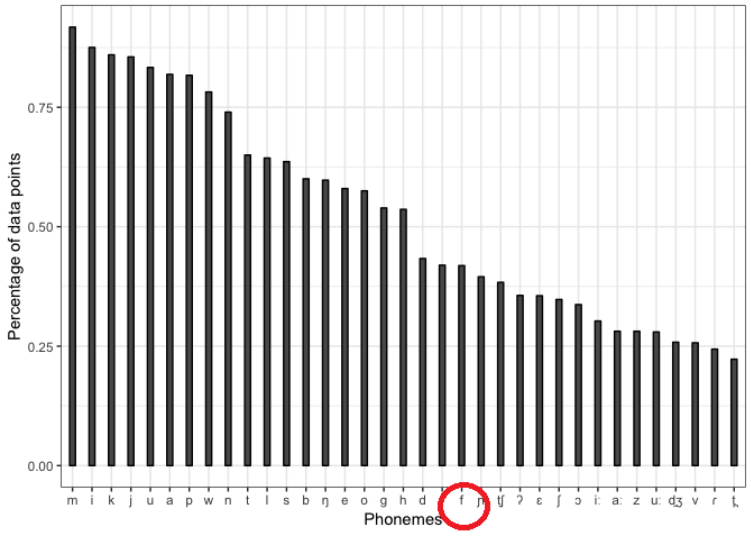
Two labiodental fricatives, /f/ and /v/, are among the most common phonological segments in the world's languages.

- /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.

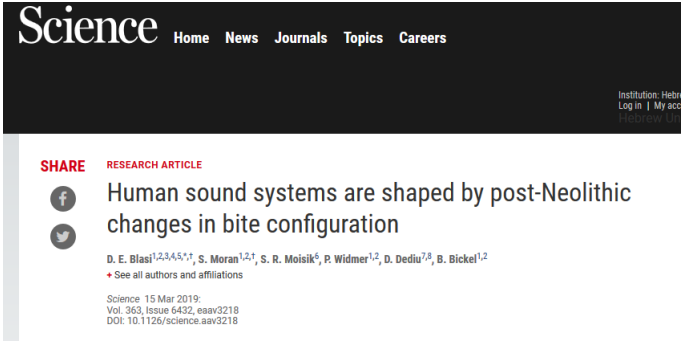
In present-day distributions, labiodental fricatives are among the most frequent sounds in the world's languages.

Maybe they're unmarked.

Recall the frequency of phonological segments



Back to the evidence



The main finding

- Labiodentals are late in human evolution.

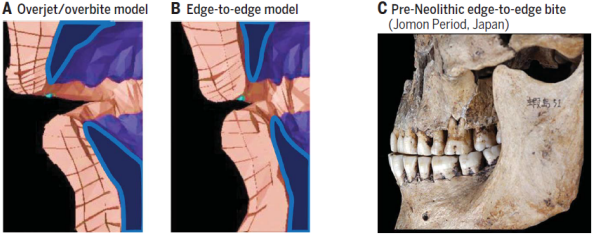
The main finding

- Labiodentals are late in human evolution.
- They postdate the advent and spread of agriculture several thousand years ago.

(Blasi et al. 2019)

Anatomical evidence

- Humans are born with overbite/overjet configuration of the jaws.
- In non-agricultural societies, this gives way to an edge-to-edge configuration.
- Overbite/overjet persists in humans with access to softer diets.



(Blasi et al. 2019)

Biomechanical evidence

Labiodentals require more effort to produce in the edge-to-edge configuration than in the overbite/overjet configuration.

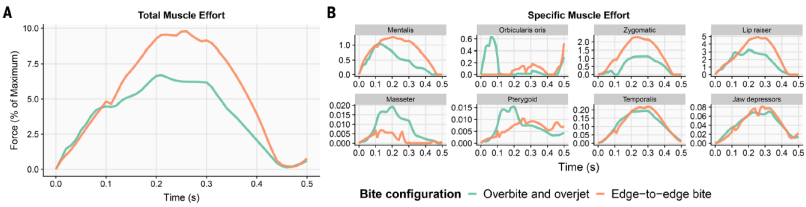
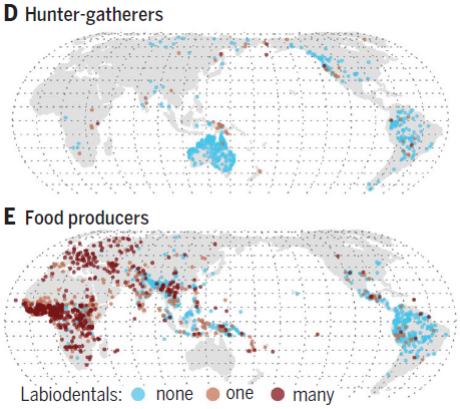


Fig. 3. Relative muscle effort in the production of labiodentals between the edge-to-edge and the overbite and overjet bite configurations.
(**A**) Sum of the total muscle force expressed as a percentage of the total maximum force of all muscles in the model. (**B**) Specific effort by muscle. Overall, labiodental articulation incurs less muscular effort in the overbite and overjet configuration than in the edge-to-edge configuration.

(Blasi et al. 2019)

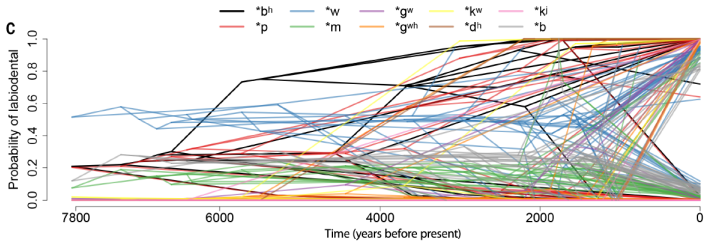
Contemporary ethnographic evidence

Languages spoken by hunter-gatherer societies tend not to have labiodentals, in comparison with agriculturalist societies.



Historical linguistic evidence

The probability of an Indo-European language to have a labiodental increases after the advent of agriculture and widespread access to softer diets.



(Blasi et al. 2019)

Historical linguistic evidence

- Labiodentals are rarely reconstructed for proto-languages, particularly those at relatively great time-depths.
- In some areas, labiodentals are largely absent, and where they are present, they are mostly the result of borrowing events.

(Blasi et al. 2019)

The bottom line

'Our studies reveal that the range and probabilities of speech sounds found across languages are not independent of large-scale changes in human ecology and biology, and thus we can no longer take for granted that the diversity of speech has remained stable since the emergence of *Homo sapiens*. **As such, claims of language universals, deep linguistic history, and language evolution cannot rely on a uniformitarian assumption without considering the wider anthropological context of language**.' (Blasi et al. 2019)

The bottom line

'Our studies reveal that the range and probabilities of speech sounds found across languages are not independent of large-scale changes in human ecology and biology, and thus we can no longer take for granted that the diversity of speech has remained stable since the emergence of *Homo sapiens*. **As such, claims of language universals, deep linguistic history, and language evolution cannot rely on a uniformitarian assumption without considering the wider anthropological context of language**.' (Blasi et al. 2019)

A finding that should make it into any phonology textbook.

Interim summary

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

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

Questions

To what extent are present-day cross-linguistic distributions the result of events of human history?

Specifically, are some cross-linguistically frequent sounds frequent because of the historically contingent events that led to language contact?

The hypotheses

- Present-day distributions may conceal substantive changes in human language.
- Borrowing events, ultimately due to contingent events of human history, have shaped phonological inventories in the relatively recent past, say, 500–1000 years before present.

Phonological segment borrowing

What do we mean by segment borrowing?

Segment borrowing:

- Modern Hebrew borrowed /tʃ/:
  English /tʃæns/ → Hebrew /tʃans/ ‘chance’

What do we mean by segment borrowing?

Segment borrowing:

- Modern Hebrew borrowed /tʃ/:
  English /tʃæns/ → Hebrew /tʃans/ ‘chance’

Working definition: a process where a sound becomes contrastive in a language as a result of lexical borrowing.

Not segment borrowing

- Northwestern California: Ejectives are an areal feature, though lexical borrowing is rare (BLEVINS 2017).
- Indian English realization of /t/ as retroflex [ʈ], due to substrate influence.

What we'd like to know

What's borrowed where?

- What segments or features are borrowed more frequently?
- Are there generalizations about borrowability?
- Does segment borrowing pattern areally? How?

A first worldwide typology of sound borrowing

Segbo: A World Survey of Phonological Segment Borrowing (Grossman et al. 2020)

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5316–5322
Marseille, 11–16 May 2020

SegBo: A Database of Borrowed Sounds in the World's Languages

Eitan Grossman[1], Elad Eisen[1], Dmitry Nikolaev[2], Steven Moran[3]
The Hebrew University of Jerusalem[1], University of Stockholm1[2], University of Zurich[3]
Jerusalem, Israel[1], Stockholm, Sweden[2], Zurich, Switzerland[3]

World Survey of Phonological Segment Borrowing (SegBo)

- The first large-scale cross-linguistic database of borrowed phonological segments, containing information on over 1600 borrowing events (+200 unique sounds) in 531 language varieties from 100+ families and +200 donor languages (Grossman et al. 2020).[1]
- Based on reports of borrowed sounds in grammars and secondary literature.
- **Importantly, most of the borrowing events documented in SEGBO are recent, dating to the past ∼500 years.**

[1] `https://github.com/segbo-db/segbo`

Where does our data come from?

TABLE 1 Consonant inventory

	Labial		Coronal			Dorsal		Glottal
	Bilabial	Labio-dental	Alveolar/ dental	Post-alveolar	Palatal	Velar	Uvular	Glottal
plosives	p b		t d			k g		ʔ[2]
affricates			t͡s	t͡ʃ[1] d͡ʒ[1]				
fricatives		f v	s z	ʃ ʒ[1]			χ	h[2]
nasals	m		n					
lateral approximants			l					
central approximants					j	w[1,3]	ɰ̊	

1 loan consonant; 2 phonological status is controversial (see §3.1.4); 3 labio-velar

Modern Hebrew (Semitic, ASHEROV & COHEN 2019)

Where does our data come from?

- In Diegueño (Cuchimi-Yuman, California), for example, the sound /ɡ/ is found only in words of foreign origin, such as *gaat* 'cat', from Spanish *gato* (Miller 1990).
- In Khwarshi (Nakh-Daghestanian, Daghestan), the voiceless velar fricative /x/ is found only in loanwords from the related language Avar, such as *xul* 'intention' (Khalilova 2009).

In such cases, an influx of loanwords has led to the introduction of a new segment into the phonological inventory of the "borrowing" language.

Structure of the data

1692 lines (1692 sloc) | 153 KB

Raw | Blame

Search this file...

	InventoryID	BorrowingLanguageGlottocode	BorrowedSound	SourceLanguageGlottocode	OnlyInLoanwords	Result	NewDistinction	PhonemeComments
2	1	abau1245	ʤ	tokp1240	yes	new phoneme	affricate	
3	1	abau1245	t	tokp1240	no	phonologization of allophone	no	allophone of /r/ in nat
4	2	abui1241	g	stan1306, indo1316	yes	new phoneme	no	Source language: Pap
5	2	abui1241	ɟ	stan1306, indo1316	yes	new phoneme	affricate	Source language: Pap
6	2	abui1241	ts	stan1306, indo1316	yes	new phoneme	affricate	QUESTION: the autho
7	3	abun1252	ʤ	stan1306	yes	new phoneme	affricate	Source language: Pap
8	3	abun1252	ɳʤ	stan1306	yes	new phoneme	affricate	prenasalisation is dist
9	4	achi1257	f	arab1395	yes	new phoneme	no	often as /β/
10	4	achi1257	ʃ	unknown	no	other distributional change	no	appears in two interje
11	5	agua1253	r	stan1288, huam1247	no	new phoneme	tap, rhotic	there is a native phon
12	6	musa1266	r	tokp1240	no	phonologization of allophone	flap, liquid	allophone of /d/ in nat
13	7	akaj1239	h	anda1280	unknown	new phoneme	glottal	aspirasion is phonemi
14	8	aleu1260	b	russ1263, stan1293	yes	new phoneme	voicing in plosives	in older loanwords rer
15	8	aleu1260	d	russ1263, stan1293	yes	new phoneme	voicing in plosives	alternating individuall

Geographical coverage of borrowing languages

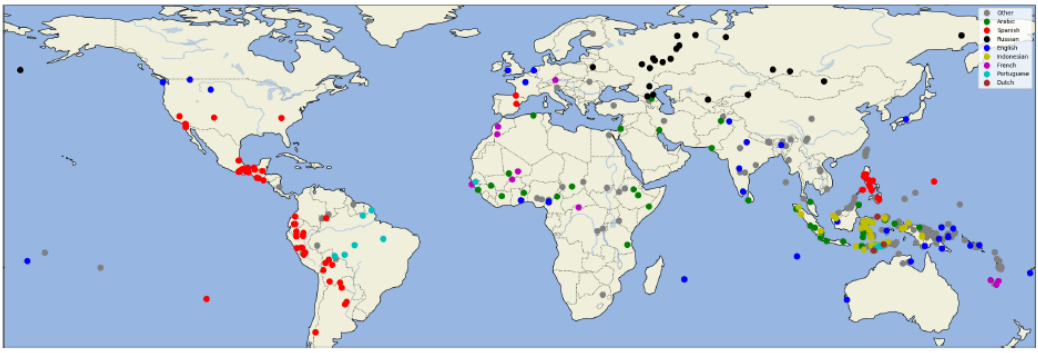
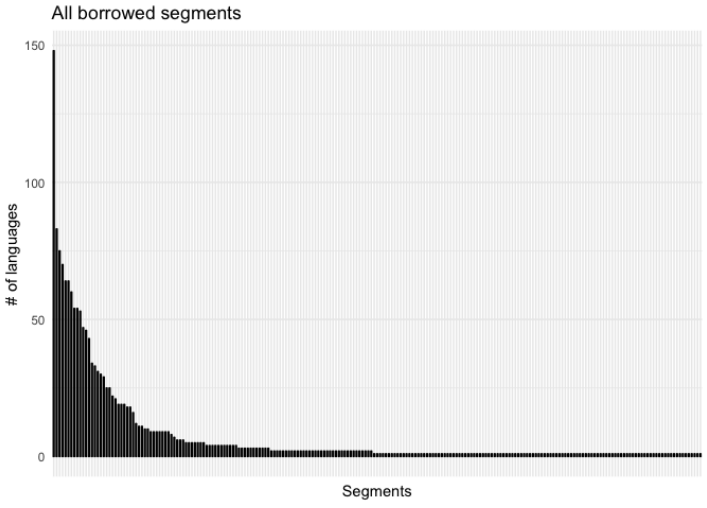
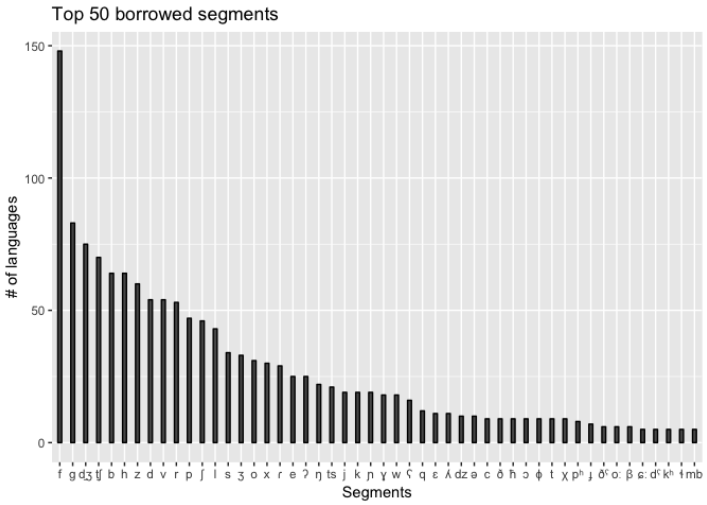


Figure 1: The geographical coverage of the SEGBO language sample. Languages are color-coded for donor languages.

What’s borrowed?



50 most frequently borrowed segments: f, g, dʒ, tʃ, h...



Areally

Eurasia	Papunesia	Africa	North America	South America	Australia
f	dʒ	p	g	b	o
ts	tʃ	h	b	f	f
x	h	z	f	d	s
tʃ	f	ʃ	d	g	t̪

Areally

Eurasia	Papunesia	Africa	North America	South America	Australia
f	dʒ	p	g	b	o
ts	tʃ	h	b	f	f
x	h	z	f	d	s
tʃ	f	ʃ	d	g	t̪

Areally

Eurasia	Papunesia	Africa	North America	South America	Australia
f	dʒ	p	g	b	o
ts	tʃ	h	b	f	f
x	h	z	f	d	s
tʃ	f	ʃ	d	g	t̪

Some preliminary observations

- An overall preference for phonologically simple sounds.
- Consonants > vowels - by far!
- In most areas, the most highly borrowed sounds are fricatives, affricates, and voiced stops, as well as rhotics and laterals.
- Clear area-specific preferences as well, often related to ‘gaps’ (e.g., African /p/, Eurasian /f/, American rhotics and voiced stops, Australian mid vowels)

Some distinctions introduced by borrowed segments

- **Affricates**: Abau, Abun, Asilulu, Dusner, East Makian, Koki, Konai, Maybrat, Nimboran, and another 40 languages.
- **Approximants**: Lewotobi, Blagar, Helong, Baukeno, Laiyolo...
- **Labiodentals**: Bierebo, Madurese, Kwaza, Sundanese, Sahu, Savosavo...

Interim summary

- Some sounds are borrowed more frequently than others.
- Some areal patterns can be observed.
- This may point to pre-contact areal profiles, and a post-contact homogenization of phonological segment inventories.

In other words, it might be the case that present-day distributions are more homogeneous due to relatively recent contact.

Frequency of borrowing and borrowability

Borrowability: a definition

“The likelihood that speakers will give up the separation between their ‘languages’ [...] in respect of a particular function-bearing structure”

(Matras 2007)

Borrowability hierarchies

nouns – adjectives – verbs – prepositions – coordinating conjunctions – quantifiers – determiners – [...]

(Muysken (1993), based on Haugen (1950)

Borrowability hierarchies

nouns – adjectives – verbs – prepositions – coordinating conjunctions – quantifiers – determiners – [...]

(Muysken (1993), based on Haugen (1950)

Nouns are “more borrowable” than adjectives.

Borrowability: vagueness

"Nouns are more borrowable than adjectives":

1. *Temporal*: Nouns are borrowed before adjectives
2. *Implicational*: If a language has borrowed adjectives, it must also have borrowed nouns
3. *Quantitative*: Languages have more borrowed nouns than adjectives
4. *Probabilistic*: A noun's chance to be borrowed is higher than that of an adjective

(Haspelmath 2008)

Probabilistic (and simplistic) model of borrowing (Eisen 2019)

For a language to borrow segment *S* from another language, these events have to occur:

Probabilistic (and simplistic) model of borrowing (Eisen 2019)

For a language to borrow segment *S* from another language, these events have to occur:

A. The contact situation has to permit borrowing:

Probabilistic (and simplistic) model of borrowing (Eisen 2019)

For a language to borrow segment *S* from another language, these events have to occur:

A. The contact situation has to permit borrowing:
   A.1 The source language has to *have* the segment *S*

Probabilistic (and simplistic) model of borrowing (Eisen 2019)

For a language to borrow segment *S* from another language, these events have to occur:

A. The contact situation has to permit borrowing:
   - A.1 The source language has to *have* the segment *S*
   - A.2 The target language has to *lack* the segment *S*

Probabilistic (and simplistic) model of borrowing (Eisen 2019)

For a language to borrow segment *S* from another language, these events have to occur:

A. The contact situation has to permit borrowing:
   - A.1 The source language has to *have* the segment *S*
   - A.2 The target language has to *lack* the segment *S*

B. The segment *S* has to be *borrowed*

Probabilistic (and simplistic) model of borrowing (Eisen 2019)

For a language to borrow segment S from another language, these events have to occur:

A. The contact situation has to permit borrowing:
   - A.1 The source language has to *have* the segment S (f_S)
   - A.2 The target language has to *lack* the segment S

B. The segment S has to be *borrowed*

f_S – typological frequency of segment S

Probabilistic (and simplistic) model of borrowing (Eisen 2019)

For a language to borrow segment S from another language, these events have to occur:

A. The contact situation has to permit borrowing:
   - A.1 The source language has to *have* the segment S (f_S)
   - A.2 The target language has to *lack* the segment S ($1 - f_S$)

B. The segment S has to be *borrowed*

f_S – typological frequency of segment S

Probabilistic (and simplistic) model of borrowing (Eisen 2019)

For a language to borrow segment S from another language, these events have to occur:

A. The contact situation has to permit borrowing:
   - A.1 The source language has to *have* the segment S (f_S)
   - A.2 The target language has to *lack* the segment S $(1 - f_S)$

B. The segment S has to be *borrowed* (b_S)

f_S – typological frequency of segment S

b_S – borrowability factor of segment S

Borrowability factor (Eisen 2019)

b_S – borrowability factor of segment S

Borrowability factor (Eisen 2019)

b_S – borrowability factor of segment S

Reflects **typological segment borrowability**:

Borrowability factor (Eisen 2019)

b_S – borrowability factor of segment S

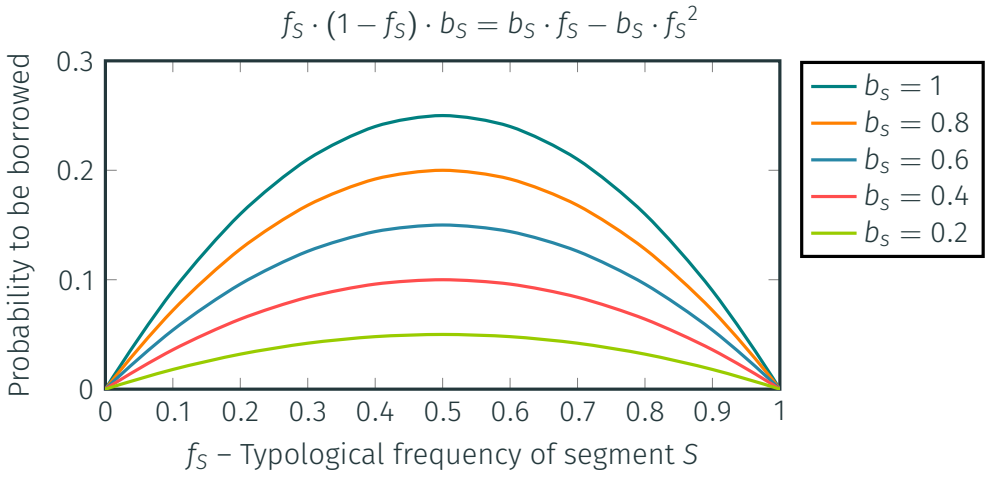
Reflects **typological segment borrowability**:

> *The probability of a segment to be borrowed from one language to another, given an appropriate contact situation.*

Model: predicted probability of borrowing (Eisen 2019)

$$f_S \cdot (1 - f_S) \cdot b_S = b_S \cdot f_S - b_S \cdot {f_S}^2$$

Model: predicted probability of borrowing (Eisen 2019)



Model: predictions (Eisen 2019)

1. Very rare segments ($f_S \to 0$) should be rarely borrowed.
2. Very common segments ($f_S \to 1$) should be rarely borrowed.
3. The most frequently borrowed segments should be ones whose typological frequency is around 50% ($f_S \approx 0.5$), since they are *more likely to appear in a contact situation that permits borrowing* in the first place.

Importantly, the relevant frequency is probably not the global frequency but rather within a particular area of interest.

Some results

- The segments with the highest borrowability score:
  - /ʕ/ (with a borrowability score of $b_S = 1.682$);
  - /f/ ($b_S = 1.129$);
  - /p/ ($b_S = 0.729$);
  - /d̪ʒ/ ($b_S = 0.713$);
  - /pʲʰ/ ($b_S = 0.672$);
  - /ħ/ ($b_S = 0.648$).
- The segments with the lowest borrowability score:
  - /ĩ/ (with a borrowability score of $b_S = 0.013$);
  - /ã/ ($b_S = 0.013$);
  - /a/ ($b_S = 0.016$);
  - /gb/ ($b_S = 0.017$);
  - /iː/ ($b_S = 0.017$);
  - /kʷ/ ($b_S = 0.017$).

The bottom line

The empirical frequency of borrowing of a segment is not identical to its borrowability.

The empirical frequency of borrowing has to be seen in light of a segment's **exposure**, i.e., chances of being borrowed in a particular contact situation.

Some sounds that are relatively infrequently borrowed turn out to be pretty borrowable.

Another operationalization

Which segments are most prone to being borrowed, given the chance?

Another operationalization

Which segments are most prone to being borrowed, given the chance?

Define **given the chance** as **(i) not having segment S and (ii) being in contact with languages with segment S**.

Testing set-up and assumptions

In order to conduct this analysis we need not only positive data (what has been borrowed where) but also negative data (what has not been borrowed).

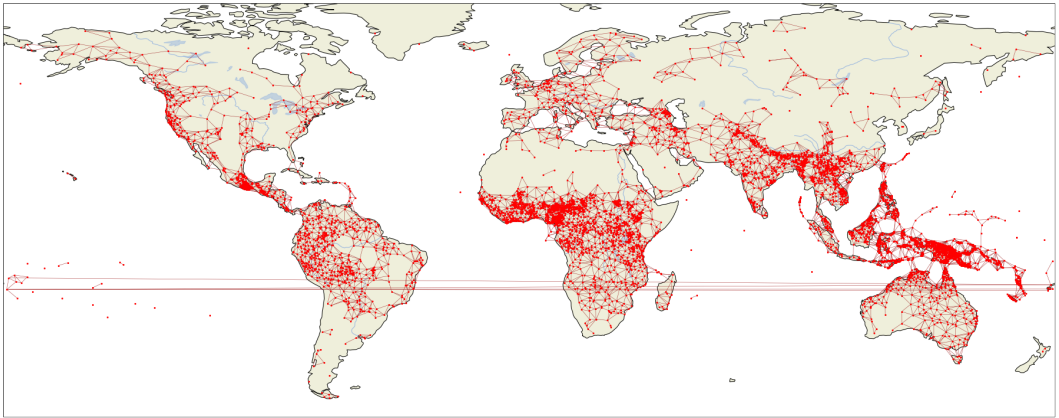
Testing set-up and assumptions

In order to conduct this analysis we need not only positive data (what has been borrowed where) but also negative data (what has not been borrowed).

SegBo provides the former type of data, while the difference between PHOIBLE (Moran & McCoy 2019) and SegBo provides the latter.

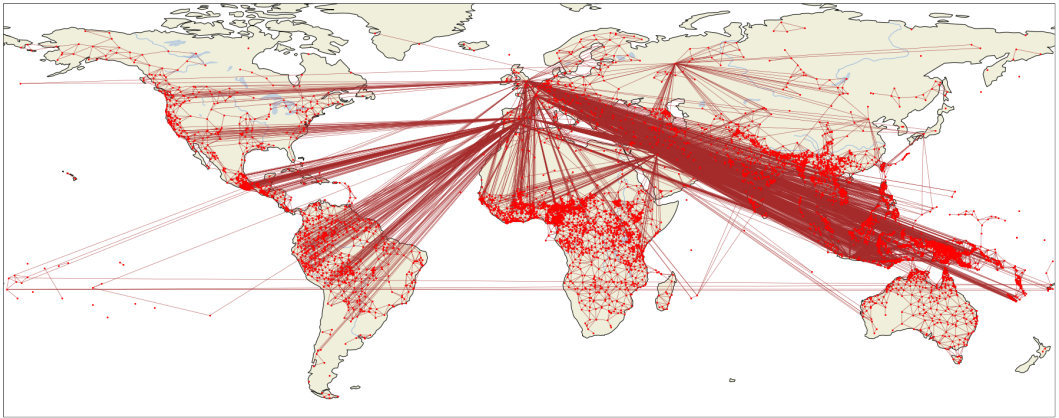
Neighbour graph used to extract contact data (Nikolaev 2019)

Before taking large languages into account:



Neighbour graph used to extract contact data (Nikolaev 2019)

After taking large languages into account:



Borrowability (Nikolaev 2019)

Top-10 most borrowable segments:

Segment	Exposure	Borrowed	Borrowability
ʕ	18	12	0.666667
f	200	112	0.560000
z	125	44	0.352000
ʒ	80	28	0.350000
v	113	39	0.345133
d̪ʒ	132	43	0.325758
ð	29	9	0.310345
g	219	63	0.287671
d	177	44	0.248588
χ	37	9	0.243243

Common segments with low borrowabilities

Segment	Exposure	Borrowed	Borrowability
p	272	34	0.125000
l	239	27	0.112971
e	191	19	0.099476
ʔ	152	14	0.092105
ŋ	185	16	0.086486
s	231	19	0.082251
ɲ	129	10	0.077519
w	235	11	0.046809
j	255	10	0.039216
t	215	8	0.037209

Greenbergian universals of borrowing?

The research question

Does the presence of certain phonological features in the native inventory of a language facilitate the borrowing of segments carrying those features?

For example, how likely are languages with no native ejectives to acquire this feature by means of lexical borrowing?

The method

Modify the borrowability set up by constructing a 2-by-2 contingency table for each feature or combination of features:

	−native	+native
−borrowed		
+borrowed		

Where +/−native = has/doesn't have non-borrowed segments with the feature and +/−borrowed = has borrowed segments with the feature.

The method

Modify the borrowability set up by constructing a 2-by-2 contingency table for each feature or combination of features:

	−native	+native
−borrowed		
+borrowed		

Where $+/-$native = has/doesn't have non-borrowed segments with the feature and $+/-$borrowed = has borrowed segments with the feature.

Perform Fisher's exact test.

Some results for individual features

ejective	−native	+native
−borrowed	423	172
+borrowed	0	6
p-value	0.006	

aspirated	−native	+native
−borrowed	1419	410
+borrowed	2	13
p-value	≈ 0	

pharyngealized	−native	+native
−borrowed	137	31
+borrowed	9	1
p-value	0.69	

interdental	−native	+native
−borrowed	619	149
+borrowed	13	0
p-value	0.145	

labialized	−native	+native
−borrowed	1285	358
+borrowed	2	6
p-value	0.002	

palatalized	−native	+native
−borrowed	865	211
+borrowed	6	5
p-value	0.048	

Some more results (and a sanity check)

labiodental fricative	−native	+native
−borrowed	2481	913
+borrowed	94	32
p-value	0.76	

postalveolar affricate	−native	+native
−borrowed	2519	981
+borrowed	47	14
p-value	0.471	

affricate	−native	+native
−borrowed	2841	1256
+borrowed	47	35
p-value	0.02	

retroflex	−native	+native
−borrowed	1278	549
+borrowed	3	7
p-value	0.011	

Universals of segment borrowing

- **Universal 1**: A language is much more likely to borrow an ejective if it already has ejective segments in its inventory.

Universals of segment borrowing

- **Universal 1**: A language is much more likely to borrow an ejective if it already has ejective segments in its inventory.
- **Universal 2:** A language is much more likely to borrow an aspirated segment if it already has aspirates in its inventory.

Universals of segment borrowing

- **Universal 1**: A language is much more likely to borrow an ejective if it already has ejective segments in its inventory.
- **Universal 2:** A language is much more likely to borrow an aspirated segment if it already has aspirates in its inventory.

The same holds for labialized and palatalized segments, as well as affricates and retroflex segments.

Segment borrowing and new features

Some features and combinations of features can be clearly borrowed even though they are not present in the inventory.

However, it seems that languages generally resist borrowing non-phonologized features and innovate them through other means first.

Classes

A few generalizations regarding classes emerge from the data:

- Consonants are more borrowable than vowels.
- Fricatives are more borrowable than affricates, which are more borrowable than stops.

However, other comparisons result in less significant differences:

- Nasalized vs. plain vowels.
- Voiced vs. voiceless consonants.

Evidence for event-based triggers (Bickel 2015)

Areally

Eurasia	Papunesia	Africa	North America	South America	Australia
f	dʒ	p	g	b	o
ts	tʃ	h	b	f	f
x	h	z	f	d	s
tʃ	f	ʃ	d	g	t̪

Eurasia	Papunesia	Africa	North America	South America	Australia
f	dʒ	p	g	b	o
ts	tʃ	h	b	f	f
x	h	z	f	d	s
tʃ	f	ʃ	d	g	t̪

Eurasia	Papunesia	Africa	North America	South America	Australia
f	dʒ	p	g	b	o
ts	tʃ	h	b	f	f
x	h	z	f	d	s
tʃ	f	ʃ	d	g	t̪

Areal phonological profiles (Moran et al. 2021)

Table 1. Cross-linguistically frequent voiced obstruents in South America versus worldwide coverage (in PHOIBLE versus PHOIBLE without South America).

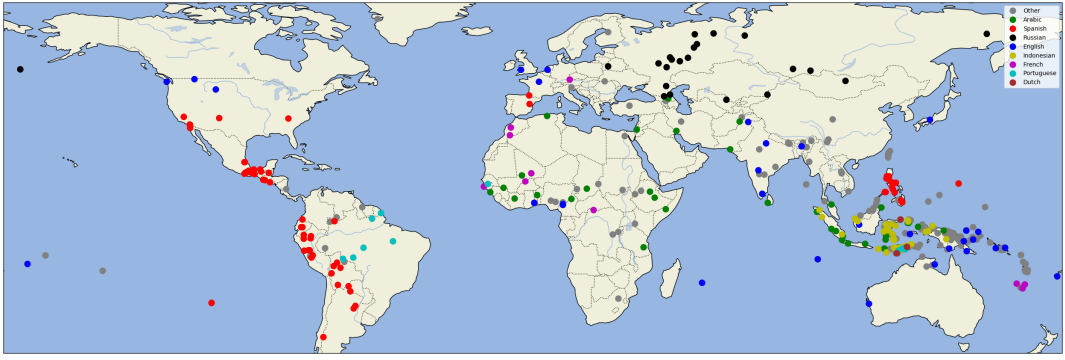
sound	frequency in SA	frequency in PHOIBLE	Δ	frequency in PHOIBLE (without SA)	Δ
b	0.44	0.63	−0.19	0.67	−0.23
d	0.37	0.46	−0.09	0.47	−0.11
g	0.29	0.57	−0.28	0.62	−0.33
β	0.17	0.10	+0.07	0.09	−0.08
v	0.03	0.27	−0.24	0.31	−0.28
ð	0.02	0.05	−0.03	0.06	−0.04
z	0.06	0.30	−0.24	0.34	−0.26
ʒ	0.06	0.16	−0.10	0.18	−0.12
ɣ	0.06	0.14	−0.10	0.16	−0.10

Which donor languages?

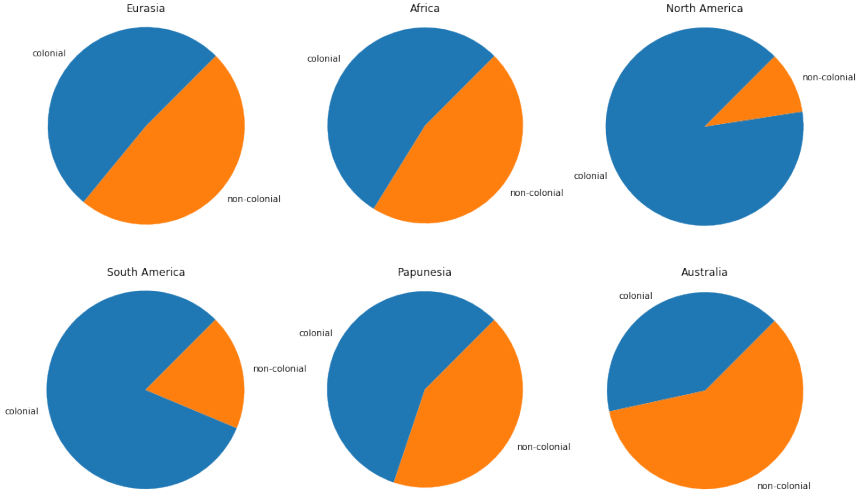
In each area, the phonological inventories of languages have been shaped by a small number of languages, mostly:

- Iberian Romance (Spanish, Portuguese)
- English, including English-lexifier pidgins and creoles
- Malayic (varieties of Indonesian/Malay)
- Arabic
- Russian

Borrowings from colonial languages (Grossman et al. 2020)



Prevalence of big donors by macro-area (Grossman et al. 2020)



Interim summary

The evidence points to the plausibility of segment borrowing playing a role in making pre-contact areal profiles more homogeneous.

Testing the hypothesis

Our question

Can we see if historical events, in particular those that led to wide-scale language contact, led to changes in cross-linguistic frequency distributions of speech sounds?

A forthcoming study

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]

(Moran et al. 2021)

Data: PHOIBLE

- A repository of cross-linguistic phonological inventory data, which have been extracted from source documents and tertiary databases and compiled into a single convenience sample (Moran 2012).[2]
- 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.

[2] `https://phoible.org/`

Data: 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).[3]

[3] https://github.com/bdproto/bdproto

The aim

- To investigate the impact of relatively recent contact events on present-day distributions, in order to see to what extent the Uniformitarian Assumption holds.
- An intuitive way of thinking about this is to compare the frequency distributions of speech sounds in PHOIBLE and BDPROTO, and to see how the differences, if there are any, are related to the frequency distributions of speech sounds in SEGBO.

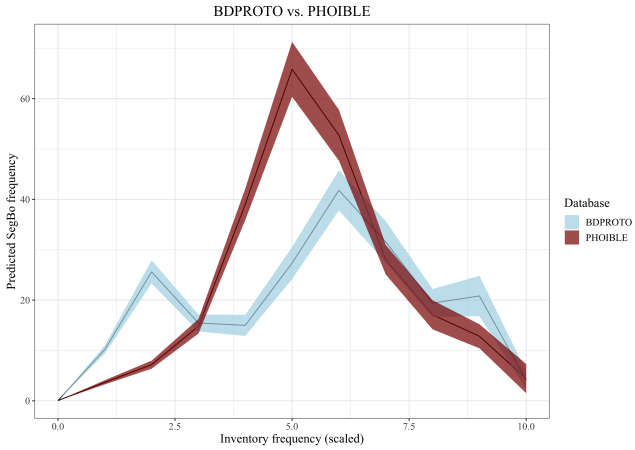
Analyses

We conducted a series of analyses that investigate the following:[4]

- (i) The overall similarity of each of the databases to each other by comparing the frequency distributions of segments in the three databases, and (ii) in what ways these similarity estimates might be affected by individual segments, geographic macro-areas, or segment class (consonants or vowels).
- How stable our estimates of similarity are between databases are, and possible sampling biases.
- The extent to which differences between frequency distribution in PHOIBLE (= present-day) and BDPROTO (=ancient) correlates with the frequency of borrowing in SEGBO.
- Which specific sounds drive the differences between PHOIBLE and BDPROTO.

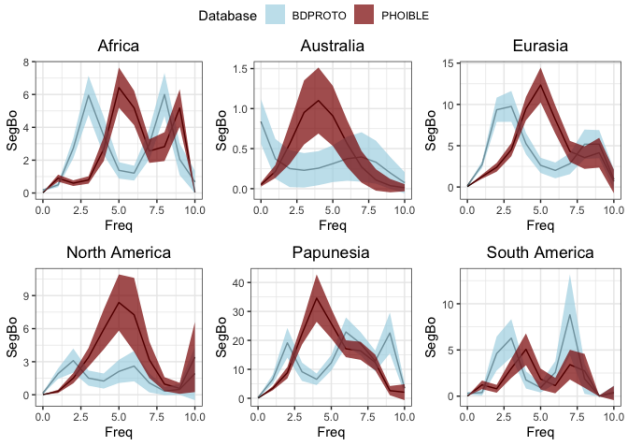
[4] https://github.com/uzling/inferring

Some results



(Moran et al. 2021)

Some results



(Moran et al. 2021)

Sounds driving differences between BDPROTO and PHOIBLE

Low-frequency				Mid-frequency			
BDPROTO		PHOIBLE		BDPROTO		PHOIBLE	
/f/	1.9 (10)	/ʒ/	1.7 (2.2)	/g/	5.8 (5.1)	/f/	5.1 (10)
/d̪ʒ/	0.7 (5.1)	/x/	1.9 (1.8)	/h/	5.4 (4.3)	/ð/	3.1 (5.1)
/t̪ʃ/	1.8 (4.8)	/ts/	2.4 (1.5)	/o/	5.9 (2)	/tʃ/	4.3 (4.8)
/v/	1.2 (3.7)	/ɣ/	1.6 (1.2)	/e/	5.9 (1.7)	/h/	6.0 (4.3)
/ʒ/	0.6 (2.2)	/ʕ/	0.2 (0.9)	/ts/	3.2 (1.5)	/z/	3.4 (4)
/ɾ/	0.5 (1.9)	/q/	0.8 (0.9)	/ŋ/	5.3 (1.5)	/v/	3.3 (3.7)
/ɣ/	1.6 (1.2)	/ə/	2.5 (0.7)	/ʔ/	4.7 (1.4)	/r/	4.6 (3.5)
/ʕ/	0.7 (0.9)	/dz/	1.1 (0.7)	/ɲ/	3.4 (1.2)	/d/	5.0 (3.3)
/q/	2.0 (0.9)	/ʎ/	0.5 (0.7)	/e:/	3.1 (0.3)	/ʃ/	3.8 (2.9)
/ɛ/	1.5 (0.7)	/ɸ/	0.4 (0.7)	/a:/	3.4 (0.2)	/ʔ/	3.7 (1.4)

Table 3: Comparing segment frequencies in the lower and middle bands of PHOIBLE/BDPROTO

(Moran et al. 2021)

Interpretation of results

- Over time, the segments that are drawn towards the median frequency band across languages are those that are reportedly most frequently borrowed. Conversely, the least frequently borrowed sounds are drawn to the extremes of the frequency range.
- This fits with the idea that both typologically rare and sounds that are typologically frequent are the least likely to be borrowed in contact scenarios; whereas those in the mid-range of the curve are the most likely to be borrowed.
- Furthermore, the largest discrepancy between BDPROTO and PHOIBLE arises for those sounds that are the most frequent in SegBo.

Interpretation of results

Importantly, these results hold both at a global level and for each macro-area tested.

As such, we conclude that our results support the alternative hypothesis: the most frequently borrowed segments in the recent past largely correspond to those segments that distinguish BDPROTO and PHOIBLE.

We take this as support for the plausible role of recent language contact in shaping the distribution of speech sounds in the languages of the world today.

Interpretation of results

In brief, our analyses do not support the null hypothesis of time-independence of cross-linguistic distributions.

Rather, they support the alternative hypothesis explored, which is that in terms of the distribution of their speech sounds, present-day languages are substantially different from ancient and reconstructed languages.

Interestingly, the sounds that drive the difference over time largely overlap with the most frequently borrowed sounds.

Summary

Our findings suggest that the Implicit Uniformitarian Hypothesis, at least with respect to the composition of phonological inventories, cannot be held uncritically.

Theories about Language may have to recognize that present-day distributions may be contingent on the historical events that led to new language contact situations to a larger extent than previously supposed.

Linguists who would like to draw inferences about human language based on cross-linguistic distributions may have to consider their theories in light of the pressures contributing to language evolution, even in the short term.

Conclusions

Conclusions

- It is possible and productive to study contact-induced change from a typological perspective.
- It can provide data for examining broader questions related to the time-dependence of present-day cross-linguistic distributions.
- This type of inquiry can probably be applied to other domains (e.g., morphosyntax), provided that data can be collected.

Vielen Dank!

Thank you!