Phonological features from a quantitative typological perspective

Steven Moran, Taras Zakharko & Balthasar Bickel

University of Zurich

Diversity Linguistics: Retrospect and Prospect

Max Planck Institute for Evolutionary Anthropology

Leipzig, Germany — Friday, May 1, 2015

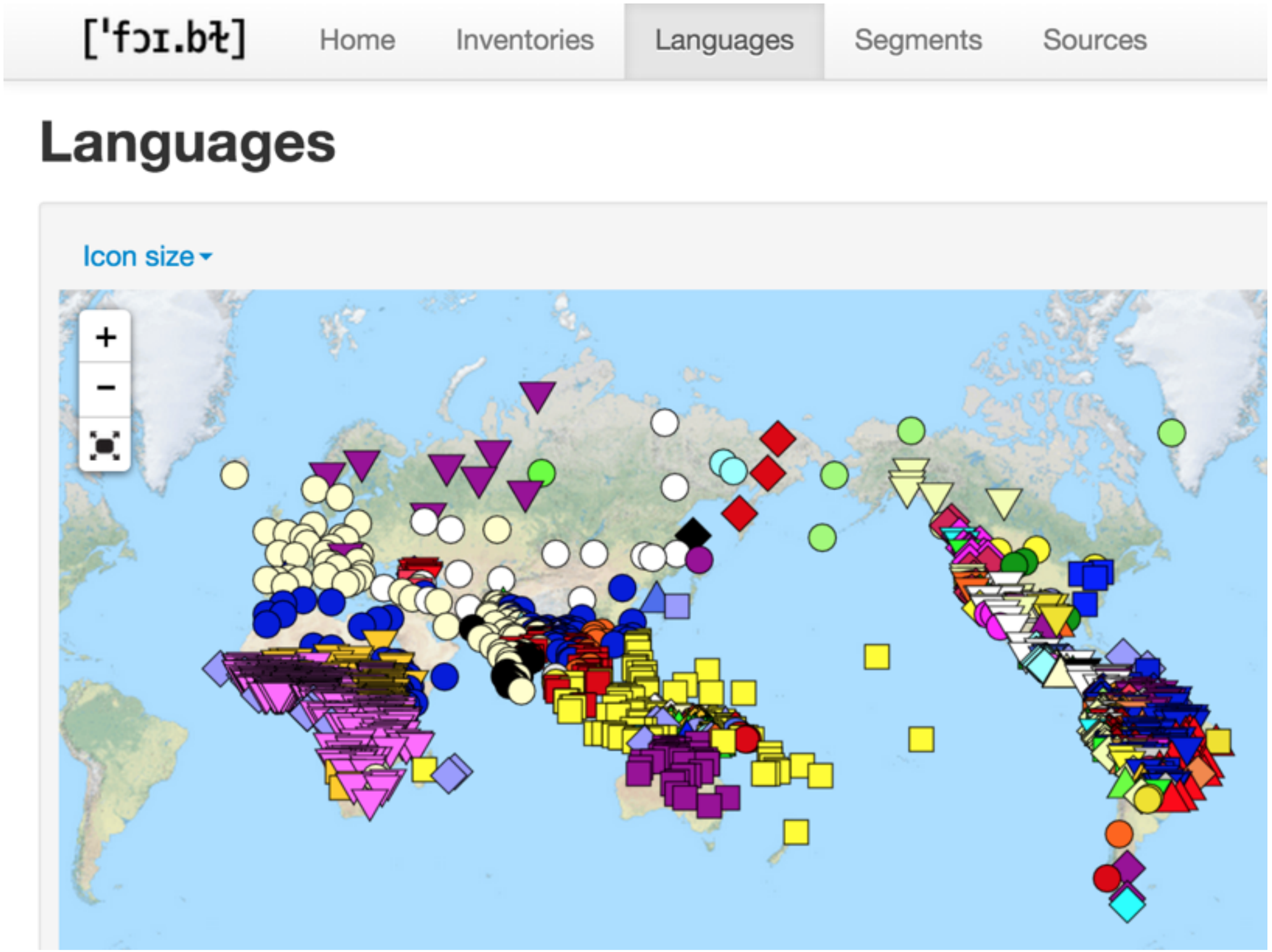
Outline

- Database and some descriptive properties of its contents
- Distinctive features and feature decomposition
- Results from some quantitative analyses

Database

- PHOIBLE Online
  - 2200+ segment inventories (1600+ distinct language descriptions)
  - distinctive feature system
  - additional linguistic and non-linguistic data
- http://phoible.org/
  - CLLD framework for user-friendly browsing of the data
- https://github.com/phoible/
  - Github repository
  - raw data, aggregation scripts, various code and documentation

Website and data



Website and data

phoible / phoible

Unwatch 5 | Star 8 | Fork 3

branch: master — phoible / data / SPA / SPA_Phones.tsv

drammock on May 8, 2014 fixing upstream changes

2 contributors

12168 lines (12167 sloc) | 251.19 kb — Raw | Blame | History

Search this file...

	spaLangNum	LanguageName	spaPhoneNum	spaDescription	spaAllophoneDescription	Notes
2	380	Korean	01	p		
3					[b]	60
4					[p-unreleased]	61
5			02	p-aspirated		
6			03	p-glottalized		
7			04	t		
8					[d]	60
9					[t-unreleased]	61

Segment inventories

- Stanford Phonology Archive (SPA, Crothers et al 1979)
- UCLA Phonological Segment Inventory Database (UPSID, Maddieson 1984, Maddieson & Precoda 1990)
- Alphabets of Africa (AA, Hartell 1993, Chanard 2006)
- PH inventories (Moran 2012)
- GM inventories (Africa and SE Asia; Green & Moran)
- South American Phonological Inventory Database (SAPHON, Michael et al 2012)
- RA (Common Linguistic Features in Indian Languages, Ramaswami 1999)
- Illustrations of the IPA (JIPA), UZ, STEDT, Handbooks (e.g. Australia, Oceania), individual collectors (C. Naumann, G. Segerer)

What's in these databases?

- "Factual claims" attributed to one or more linguists, including:
  - a linguist who described a language, or
  - a compiler of a typological database
- Guiding principles for PHOIBLE's development:
  - faithfulness to the field linguist's description of each language
  - faithfulness to linguists' interpretation of a phoneme inventory based on one or more languages
  - Stay as true to original grammar as possible (required several additions to IPA)
- >1 inventories for ~375 languages

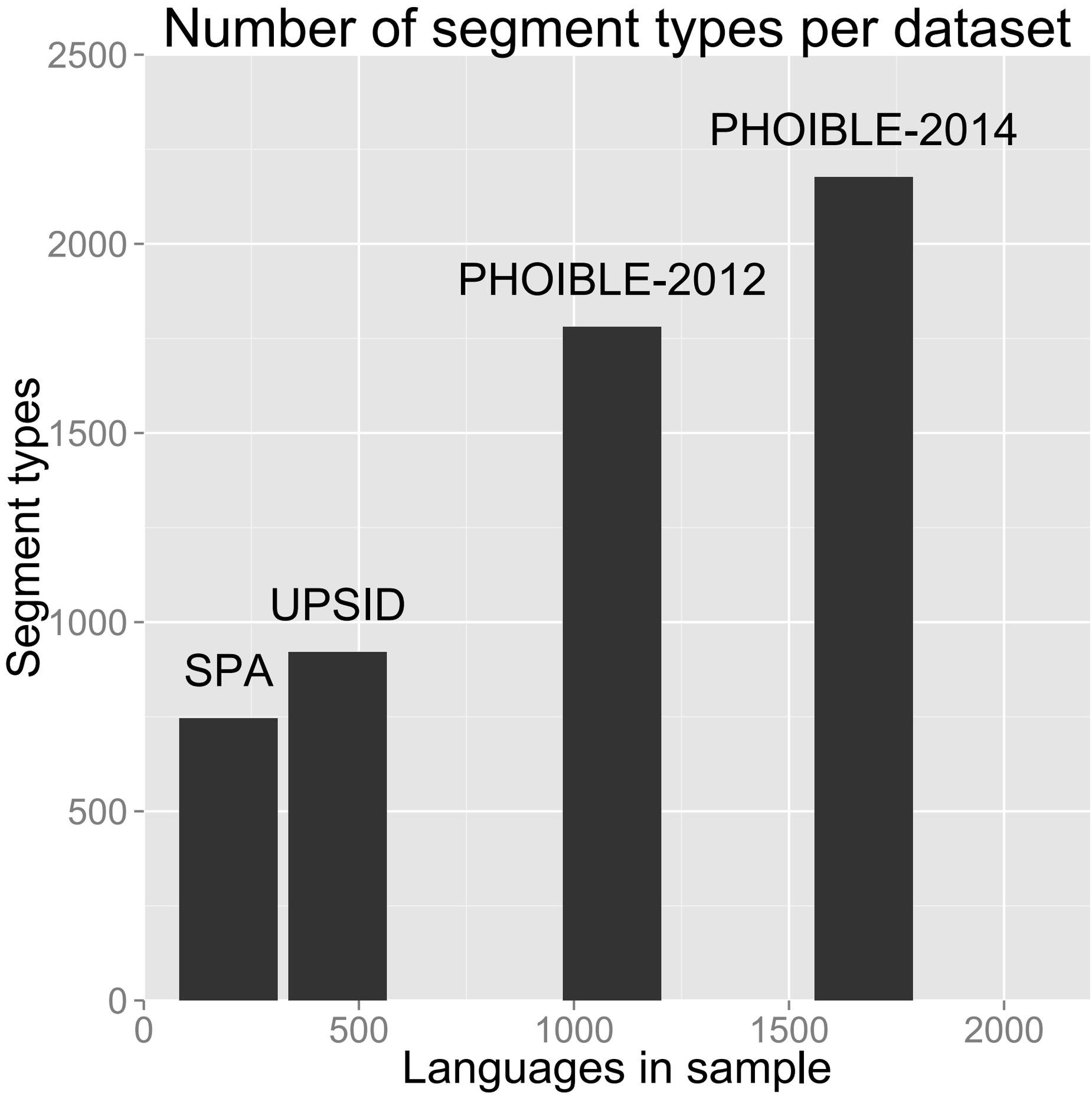
What's in these databases?

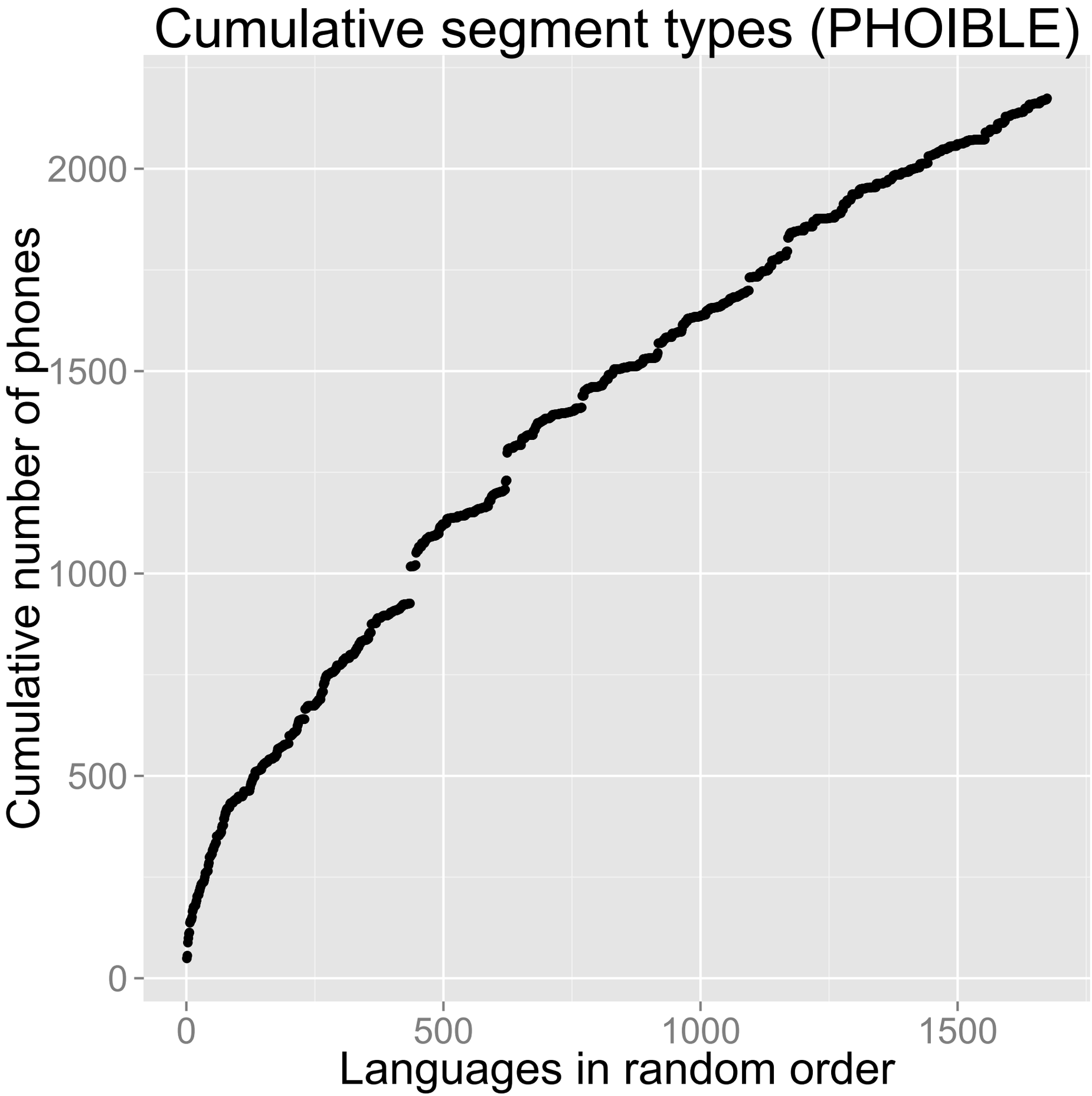
- Inventories include
  - Symbolic representation of phonemes (near-IPA)
  - Genealogical, geographic, & demographic data (e.g. Glottolog)
- Vector of feature values for each phoneme
  - Feature set mostly follows Hayes 2009 and Moisik & Esling 2011
- Goal: unique feature vector for each phoneme as described in source (regardless of within-language contrasts), e.g. feature vectors should distinguish between:

  English /s/ Spanish /s̪/ Basque /s̪/ Galician /s̪/

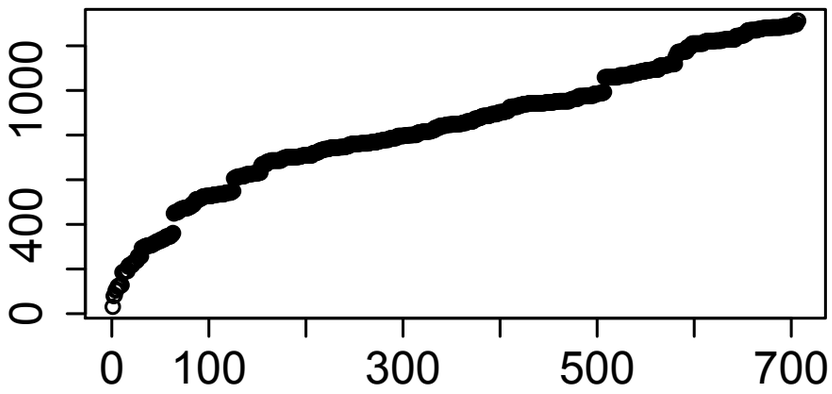
What's in these databases?

- 2000+ segment types (1000+ occur in only one language)
  - s̪̠ (non-strident voiceless retroflex fricative)
    - found in Sa'ban (Malayo-Polynesian, Austronesian)
  - ṵ̃ (nasalized creaky high back round vowel)
    - found in Mambay (Adamawa, Niger-Congo)
  - ʈˀ (glottalized voiceless retroflex stop)
    - found in Siona (Tucanoan)
  - ɬʟ̝̊ (simultaneous alveolar/velar voiceless lateral fricative)
    - found in Axluslay/Nivaclé (Matacoan)

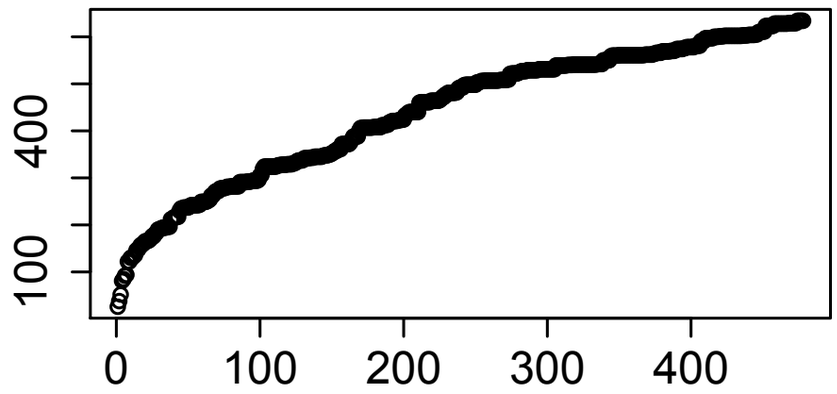

Number of segment types per dataset
2500
2000
1500
1000
500
0
Segment types
PHOIBLE-2014
PHOIBLE-2012
UPSID
SPA
0
500
1000
1500
2000
Languages in sample


Cumulative segment types (PHOIBLE)
Cumulative number of phones
2000
1500
1000
500
0
0
500
1000
1500
Languages in random order

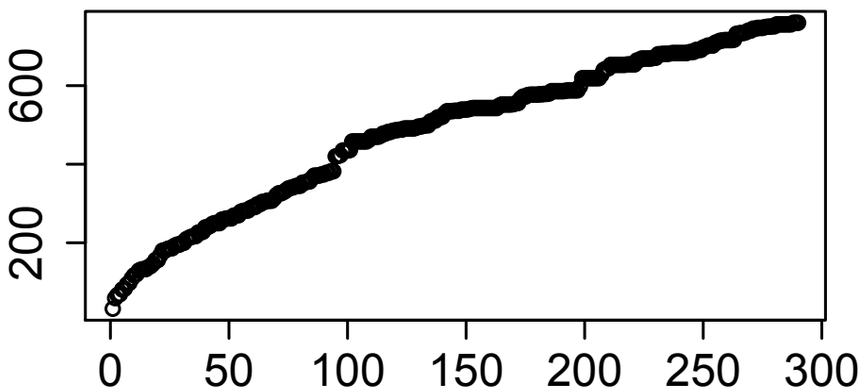
Africa



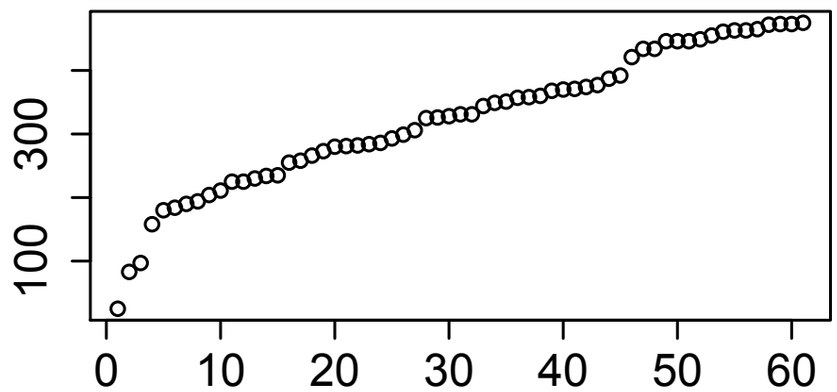
America



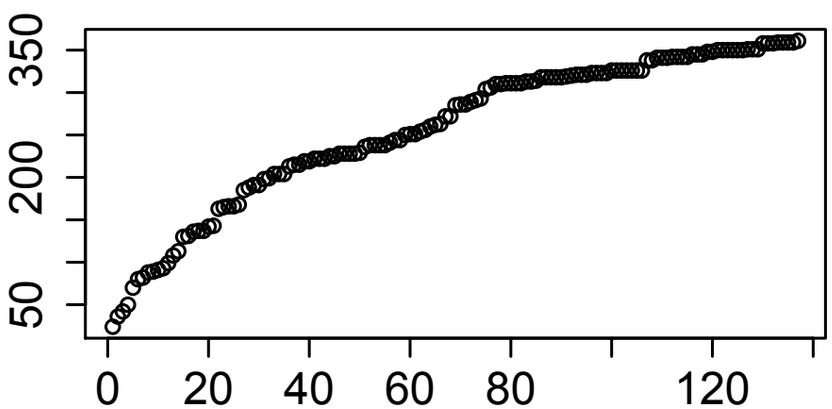
Asia

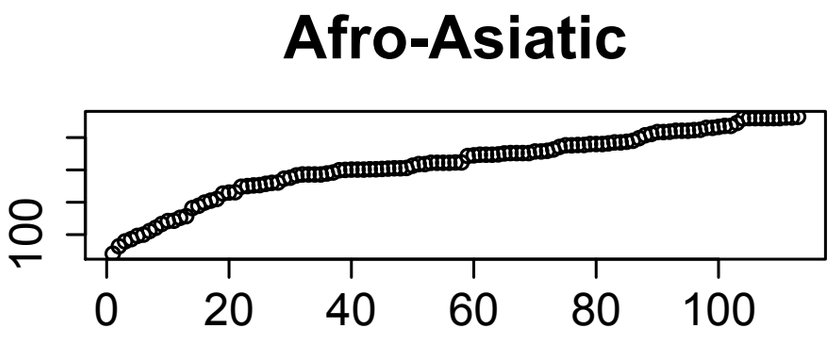


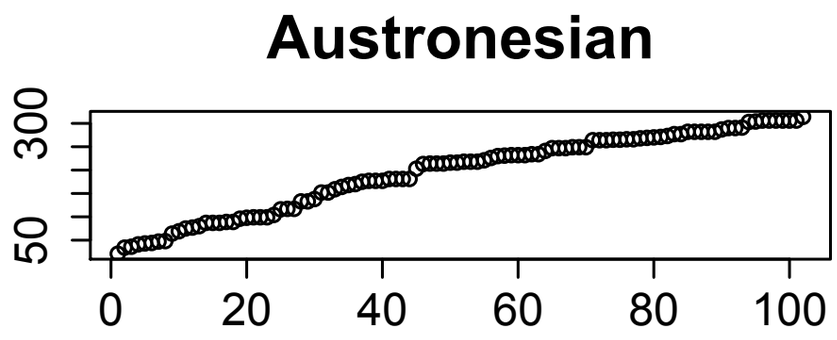
Europe

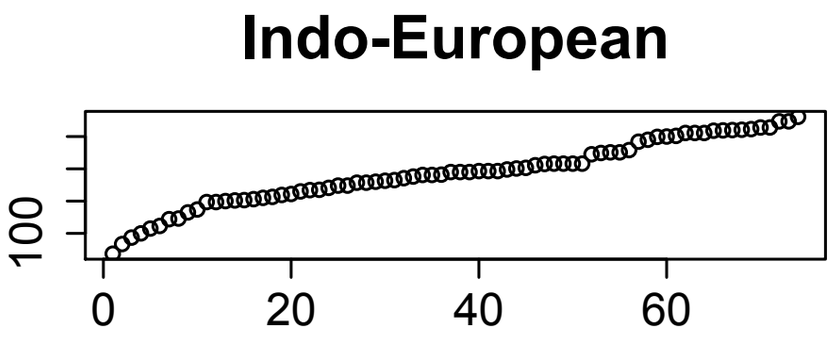


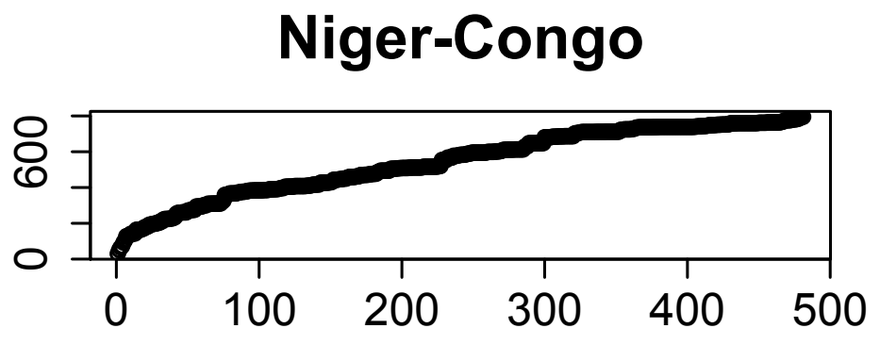
Pacific

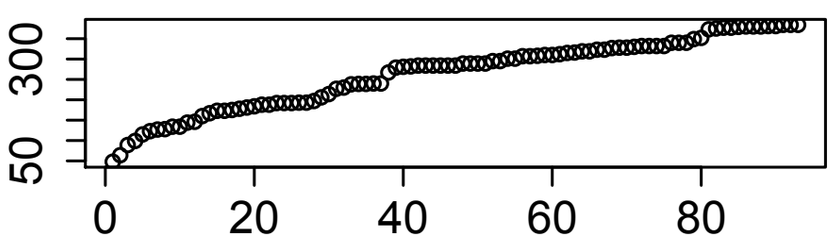


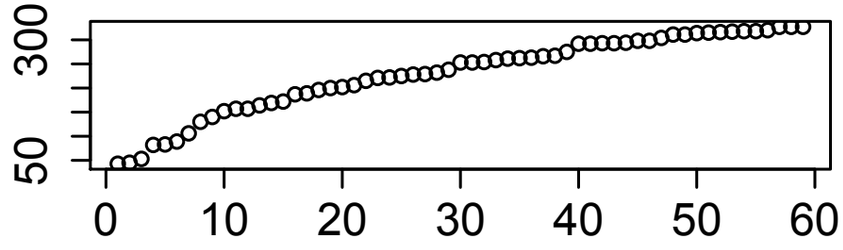

Afro-Asiatic
100
0
20
40
60
80
100

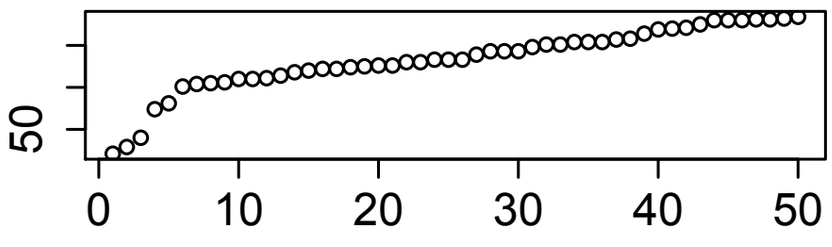

Austronesian
50
300
0
20
40
60
80
100


Indo-European
100
0
20
40
60


Niger-Congo
0
600
0
100
200
300
400
500

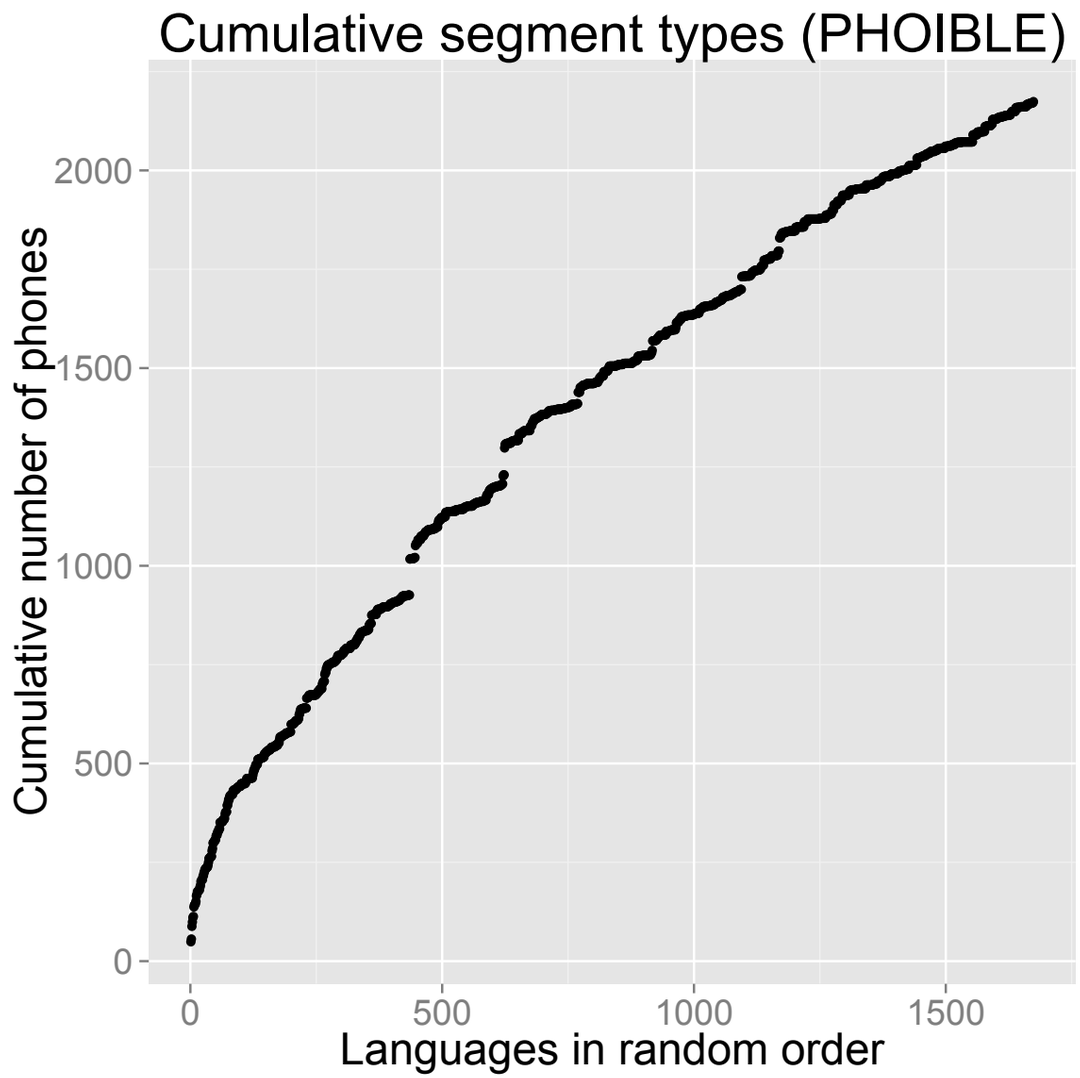

Nilo-Saharan
50
300
0
20
40
60
80


Sino-Tibetan
50
300
0
10
20
30
40
50
60


Tupi-Guarani
50
0
10
20
30
40
50

Cumulative segment types

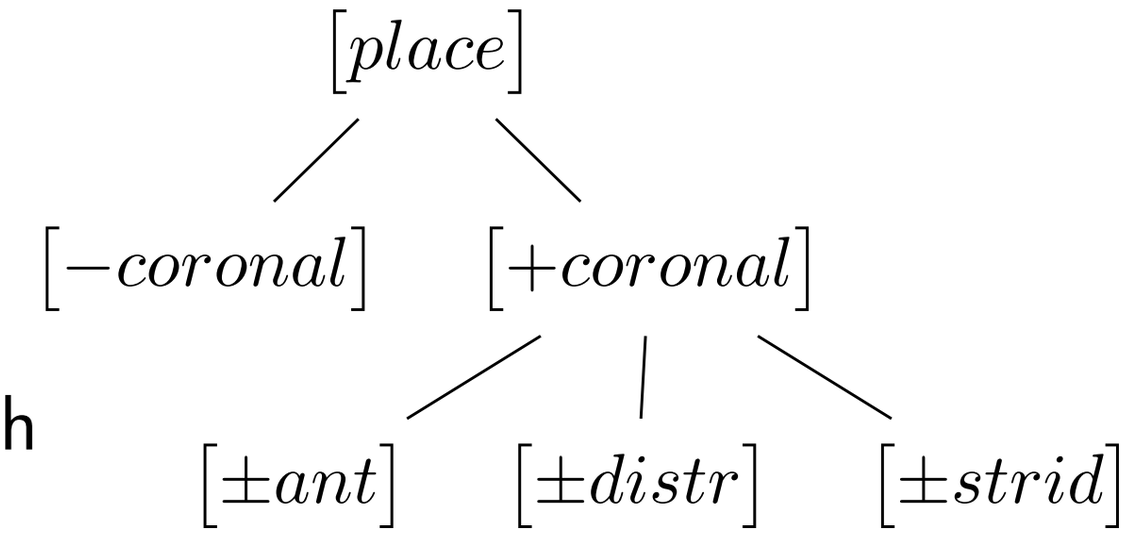
- No asymptotic limit in sight
- This makes sense because the phonetic space is nonfinite
- Do we gain any insight from features instead of segments?



Feature system

- PHOIBLE feature set has 37 features (Hayes 2009; Moisik & Esling 2011)
- Hierarchical organization: parent node [−value] ⇒ child node [0value]
  - *Example*: all [−coronal] segments are

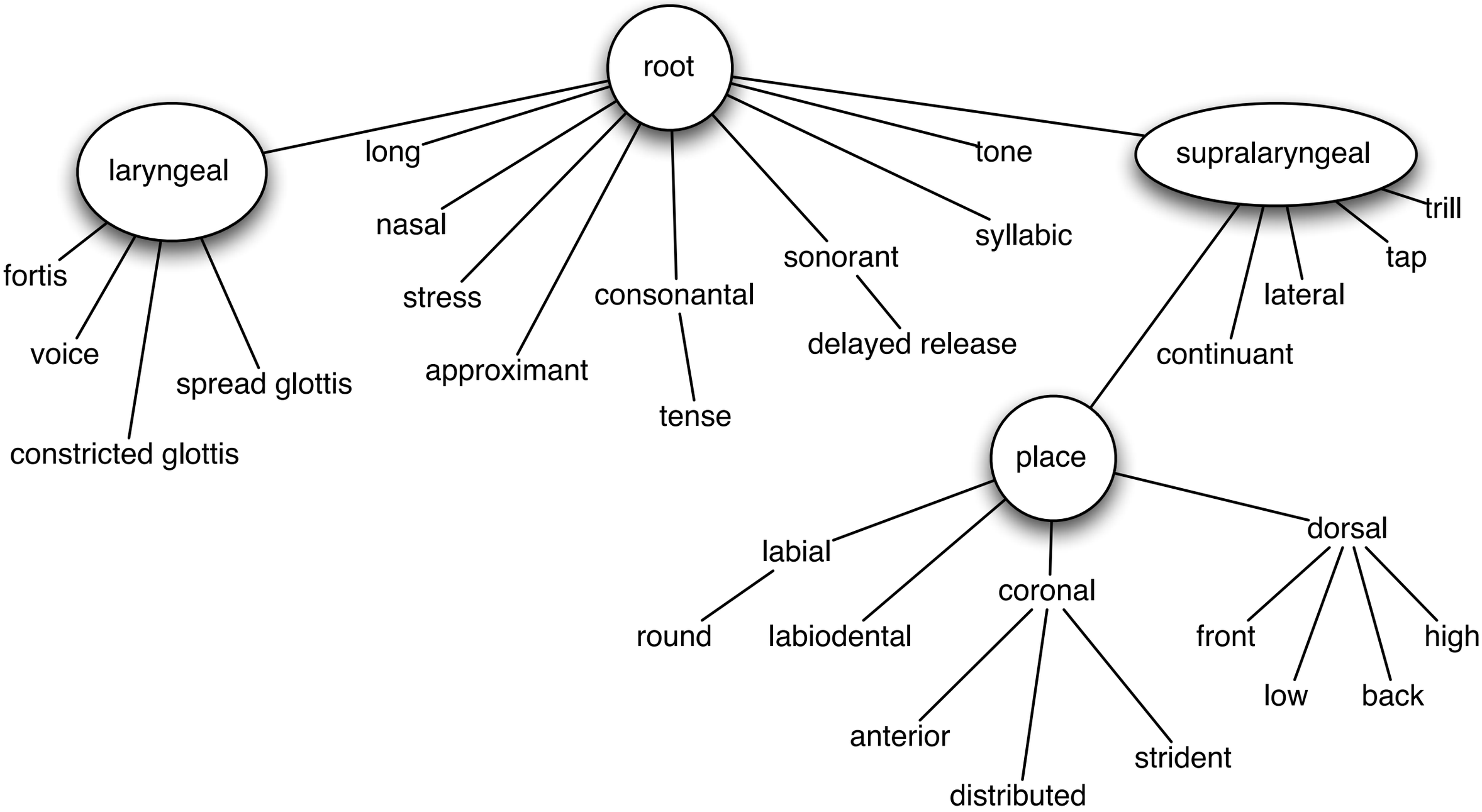
    [0anterior, 0distributed, 0strident]

  - 0 values treated as not contrasting with

    either + or −
- Contour segments: ordered tuple values for certain features

Feature geometry



Feature assignment

- Available distinctive feature sets lack broad typological coverage of segment inventories (Moran 2012) and natural and unnatural classes (Mielke 2004)

- Algorithm for assigning feature vectors for each (simple) segment type:

segment	+	+	-	-	0	+	0	-	0
diacritic	+	-	+	-	+	0	-	0	0
result:	+	-	+	-	+	+	-	-	0

- Complex segments (two or more simultaneous oral tract constrictions) typically assigned by hand

- Contour segments (temporal movement in phonetic features from a preceding segment to the following segment) cannot be captured in a single tier of distinctive features

  - prenasalized stop [nt] must be [+nasal], then [-nasal], so becomes [+,-]

Feature analysis

- This results in full, redundant descriptions
  Example: Pirahã segments: p, b, t, k, g, ʔ, s, h, i, o, a

	stress	syllabic	short	long	consonantal	sonorant	continuant	delayedRelease	approximant	tap	trill	nasal	lateral	labial	round	labiodental	coronal	anterior	distributed	strident	dorsal	high	low	front	back	tense	retractedTongueRoot	advancedTongueRoot	periodicGlottalSource	epilaryngealSource	spreadGlottis	constrictedGlottis	fortis	raisedLarynxEjective	loweredLarynxImplosive	click
p	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	-	0	0	0	-	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-
b	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	-	0	0	0	-	0	0	0	0	0	0	0	+	-	-	-	-	-	-	-
t	-	-	-	-	+	-	-	-	-	-	-	-	-	-	0	0	+	+	-	-	-	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-
k	-	-	-	-	+	-	-	-	-	-	-	-	-	-	0	0	-	0	0	0	+	+	-	-	-	0	0	0	-	-	-	-	-	-	-	-
g	-	-	-	-	+	-	-	-	-	-	-	-	-	-	0	0	-	0	0	0	+	+	-	-	-	0	0	0	+	-	-	-	-	-	-	-
ʔ	-	-	-	-	+	-	-	-	-	-	-	-	-	-	0	0	-	0	0	0	-	0	0	0	0	0	0	0	-	-	-	+	-	-	-	-
s	-	-	-	-	+	-	+	+	-	-	-	-	-	-	0	0	+	+	-	+	-	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-
h	-	-	-	-	-	-	+	+	-	-	-	-	-	-	0	0	-	0	0	0	-	0	0	0	0	0	0	0	-	-	+	-	-	-	-	-
i	-	+	-	-	-	+	+	0	+	-	-	-	-	-	0	0	-	0	0	0	+	+	-	+	-	+	-	-	+	-	-	-	0	-	-	0
o	-	+	-	-	-	+	+	0	+	-	-	-	-	+	+	-	-	0	0	0	+	-	-	-	+	+	-	-	+	-	-	-	0	-	-	0
a	-	+	-	-	-	+	+	0	+	-	-	-	-	-	0	0	-	0	0	0	+	-	+	-	-	0	-	-	+	-	-	-	0	-	-	0

Dimensionality reduction

- aka feature selection algorithms
  - data-driven models
  - generate more compact representations of a given data set
  - popular examples: PCA and MDS
- Goal is to detect meaningful dimensions in some input data
  - e.g. PCA identifies a sequence of best linear approximations
- We determine the minimal subset(s) of features needed to encode the phonemic contrasts in each segment inventory
  - basic heuristic-optimized version of a brute-force approach, which can be classified as a greedy best-first tree search

Dimensionality reduction

- Feature Reduction Algorithm: compute the minimal required set of features which are necessary to encode a phoneme inventory of a given language

	X	Y	Z
a	+	+	-
b	+	-	-
c	-	-	+

→

	X	Y
a	+	+
b	+	-
c	-	-

	Y	Z
a	+	-
b	-	-
c	-	+

Feature analysis

- After dimensionality reduction

	constrictedGlottis	continuant	coronal	labial	low	periodicGlottalSource
p	-	-	-	+	0	-
b	-	-	-	+	0	+
t	-	-	+	-	0	-
k	-	-	-	-	-	-
g	-	-	-	-	-	+
ʔ	+	-	-	-	0	-
s	-	+	+	-	0	-
h	-	+	-	-	0	-
i	-	+	-	-	-	+
o	-	+	-	+	-	+
a	-	+	-	-	+	+

	continuant	coronal	dorsal	labial	low	periodicGlottalSource
p	-	-	-	+	0	-
b	-	-	-	+	0	+
t	-	+	-	-	0	-
k	-	-	+	-	-	-
g	-	-	+	-	-	+
ʔ	-	-	-	-	0	-
s	+	+	-	-	0	-
h	+	-	-	-	0	-
i	+	-	+	-	-	+
o	+	-	+	+	-	+
a	+	-	+	-	+	+

	constrictedGlottis	continuant	coronal	front	labial	periodicGlottalSource
p	-	-	-	0	+	-
b	-	-	-	0	+	+
t	-	-	+	0	-	-
k	-	-	-	-	-	-
g	-	-	-	-	-	+
ʔ	+	-	-	0	-	-
s	-	+	+	0	-	-
h	-	+	-	0	-	-
i	-	+	-	+	-	+
o	-	+	-	-	+	+
a	-	+	-	-	-	+

	continuant	coronal	dorsal	front	labial	periodicGlottalSource
p	-	-	-	0	+	-
b	-	-	-	0	+	+
t	-	-	+	0	-	-
k	-	-	+	-	-	-
g	-	-	+	-	-	+
ʔ	-	-	-	0	-	-
s	+	+	-	0	-	-
h	+	-	-	0	-	-
i	+	-	+	+	-	+
o	+	-	+	-	+	+
a	+	-	+	-	-	+

	constrictedGlottis	continuant	coronal	high	labial	periodicGlottalSource
p	-	-	-	0	+	-
b	-	-	-	0	+	+
t	-	-	+	0	-	-
k	-	-	-	+	-	-
g	-	-	-	+	-	+
ʔ	+	-	-	0	-	-
s	-	+	+	0	-	-
h	-	+	-	0	-	-
i	-	+	-	+	-	+
o	-	+	-	-	+	+
a	-	+	-	-	-	+

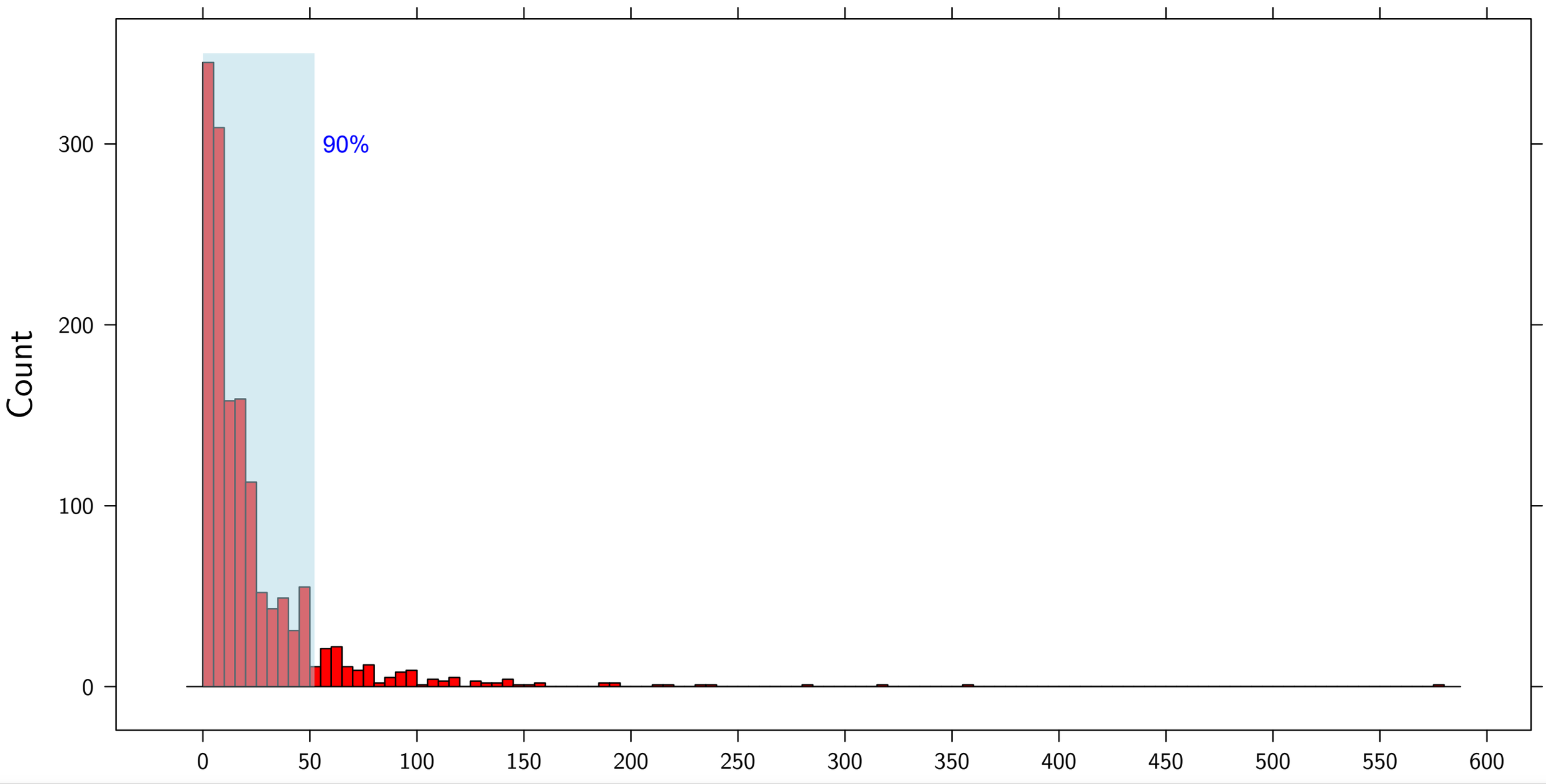
	continuant	coronal	dorsal	high	labial	periodicGlottalSource
p	-	-	-	0	+	-
b	-	-	-	0	+	+
t	-	+	-	0	-	-
k	-	-	+	+	-	-
g	-	-	+	+	-	+
ʔ	-	-	-	0	-	-
s	+	+	-	0	-	-
h	+	-	-	0	-	-
i	+	-	+	+	-	+
o	+	-	+	-	+	+
a	+	-	+	-	-	+

Findings: quantitative analyses of reduced feature sets

- while most inventories allow multiple (competing) feature analyses, 90% of the inventories allow less than 50 decompositions, making results manageable
- while a restricted set of features (based on a geometry with 37 features) allows coverage of almost all known segment inventories in the database, the phonetic and phonemic implementation in segments appear to constitute an extremely large inventory of which we do not know the limits
- phonetically informed and universally constrained feature geometries allow more efficient segment coding in languages than arbitrary and language-specific feature sets
- there is evidence for a universally preferred combination of segment inventory size and feature numbers, centered on about 36 segments and 12 features
- the number of features in phonological inventories may be stable genealogically
- certain features are diachronically preferred (e.g. high, front, labial) and dispreferred (e.g. labiodental, round, fortis)

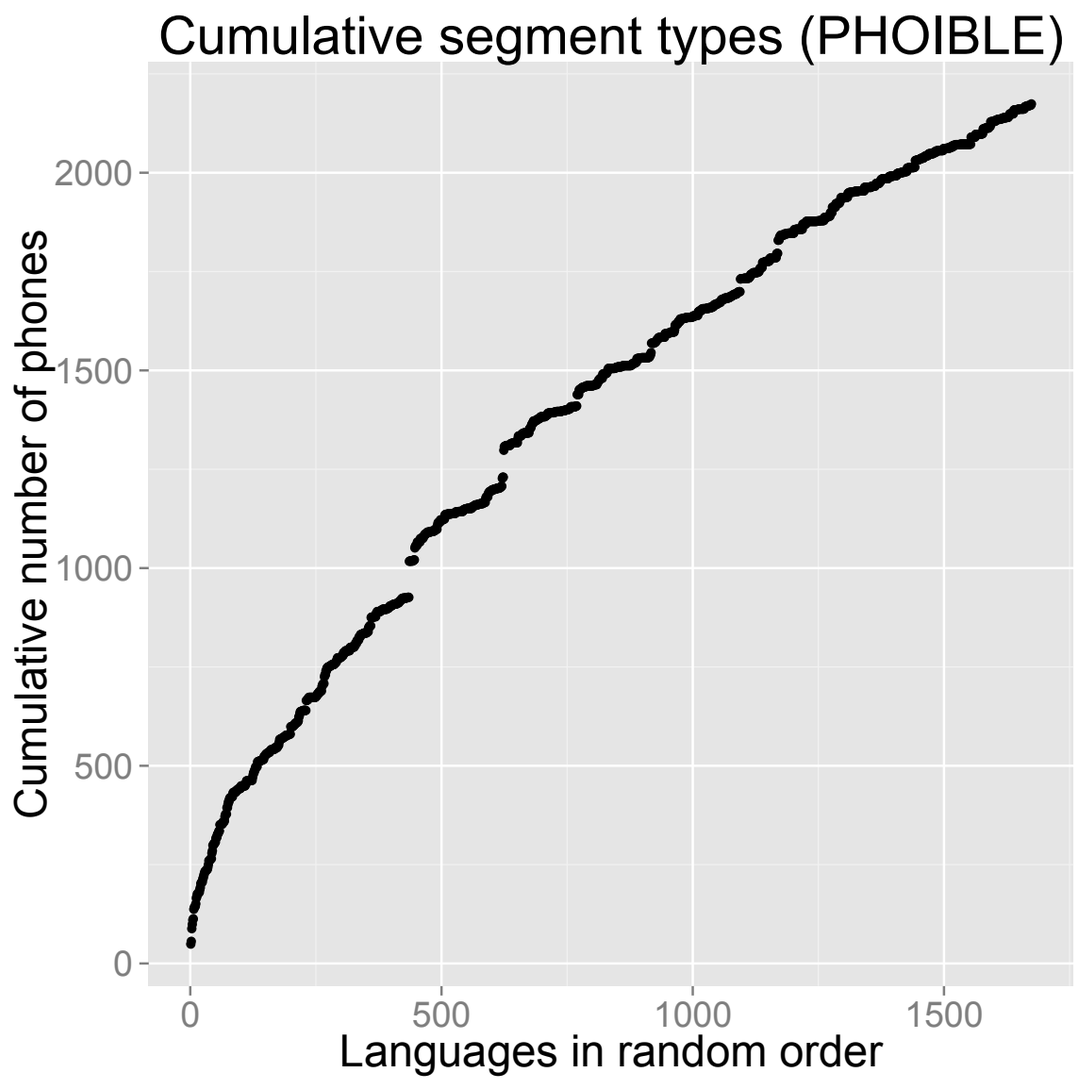
Feature decompositions by number

- inventories allow multiple (competing) feature analyses, but 90% of the inventories allow less than 50 decompositions



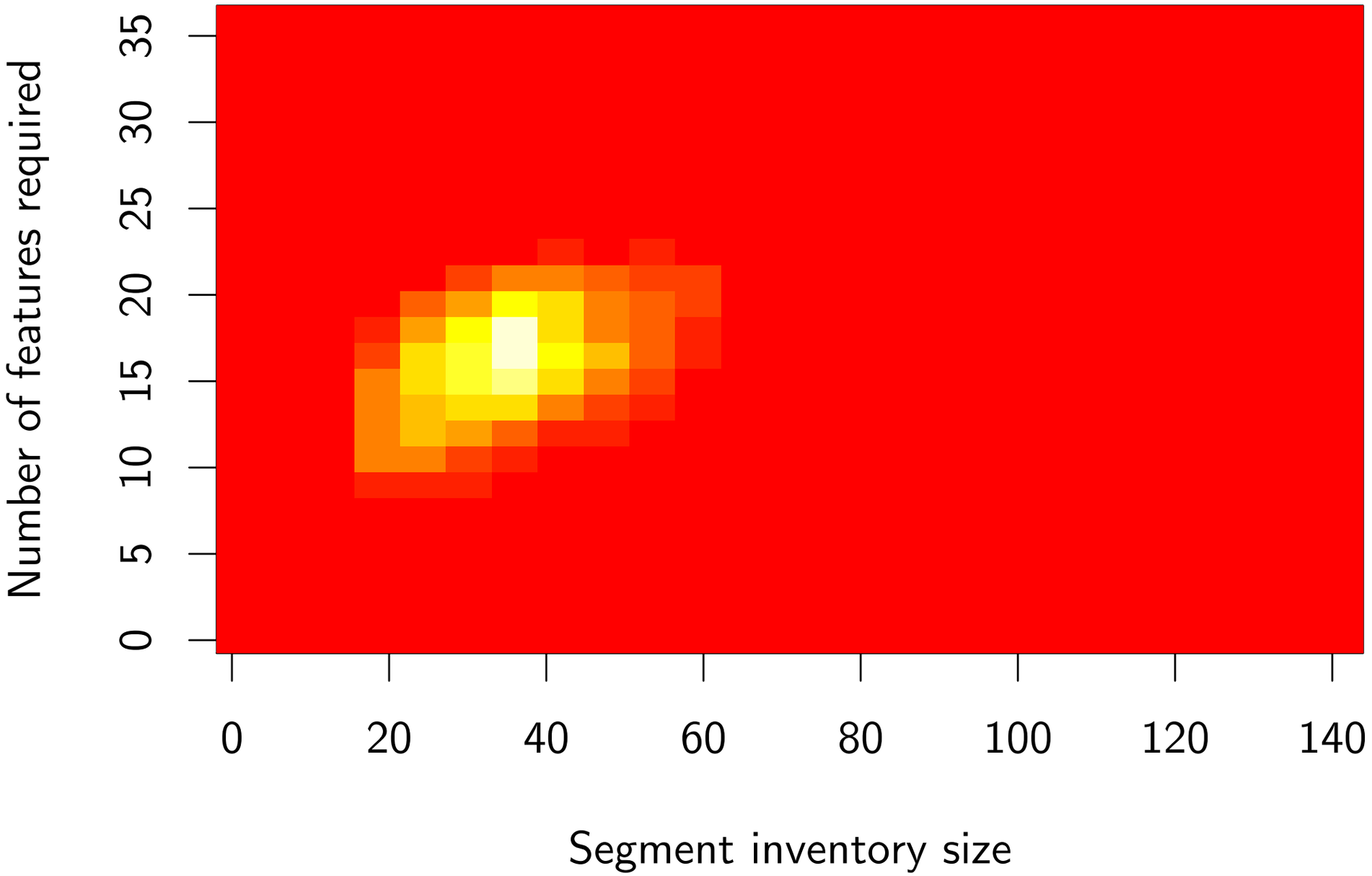
Duality of patterning

- while a restricted set of features (based on a geometry with 37 features) allows coverage of almost all known segment inventories in the database, the phonetic and phonemic implementation in segments appear to constitute an extremely large inventory of which we do not know the limits



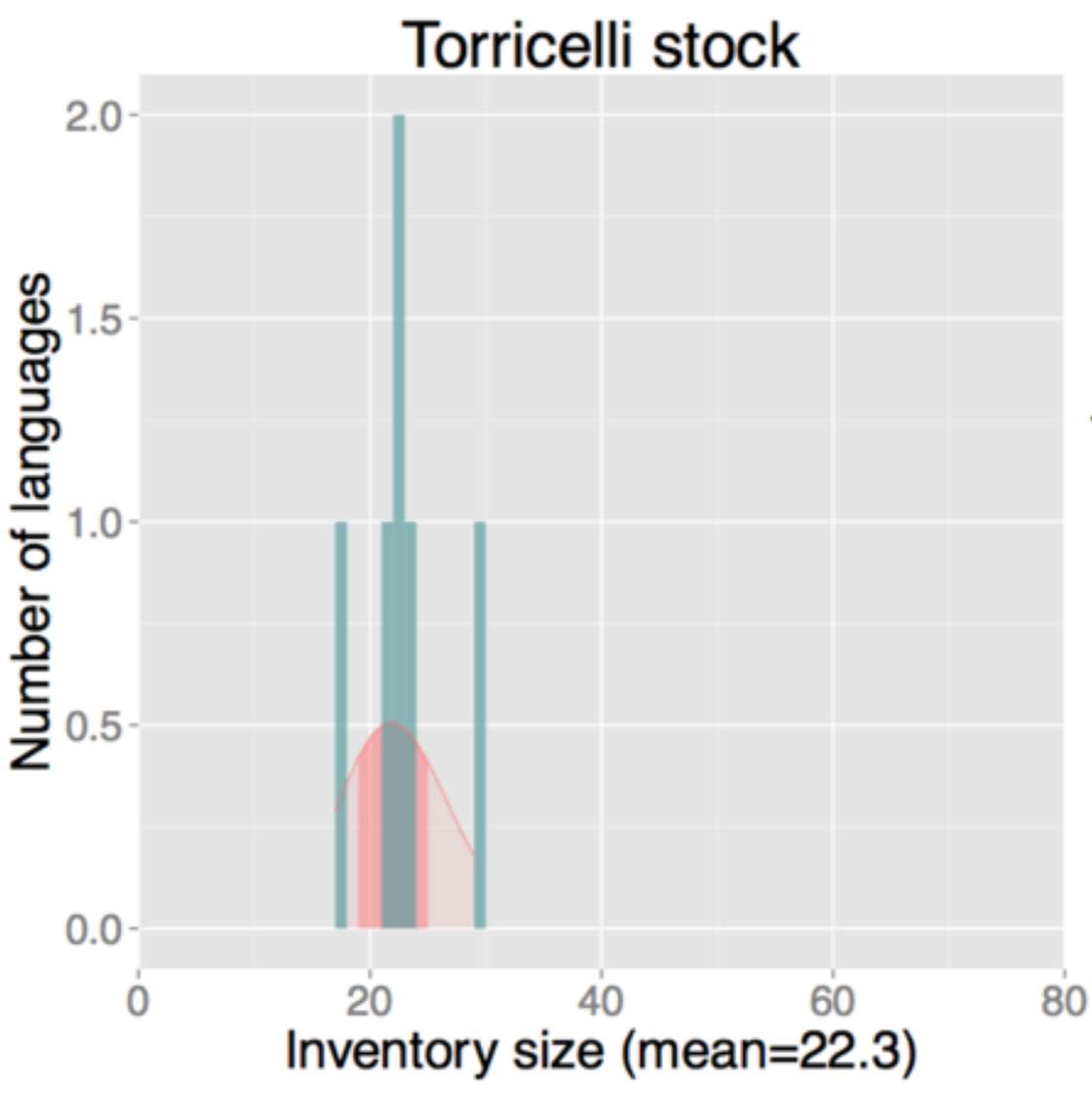
Languages converge on a universal feature set

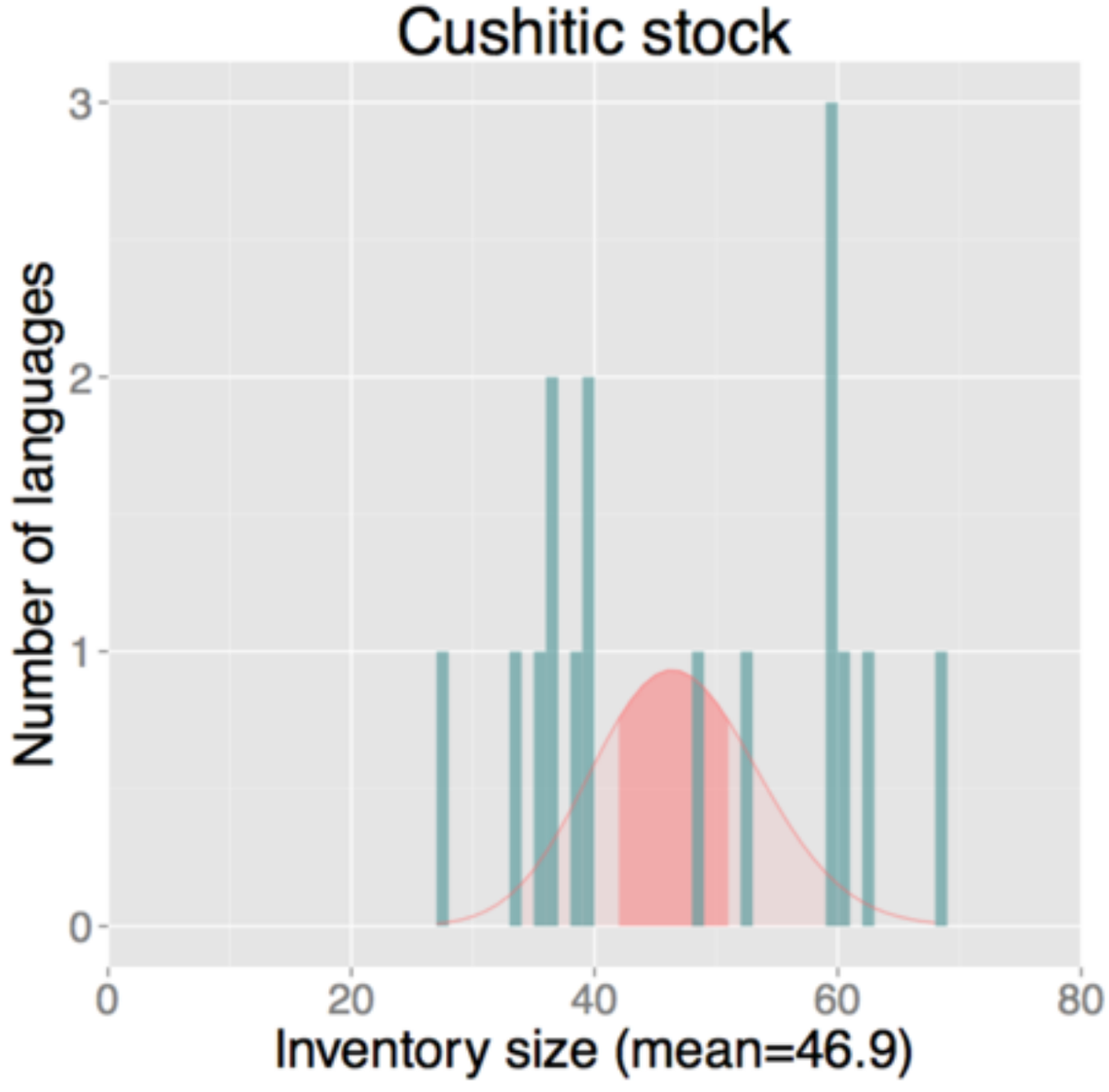
- there is evidence for a universally preferred combination of segment inventory size and feature numbers, centered on about 36 segments and 12 features



Capturing sizes

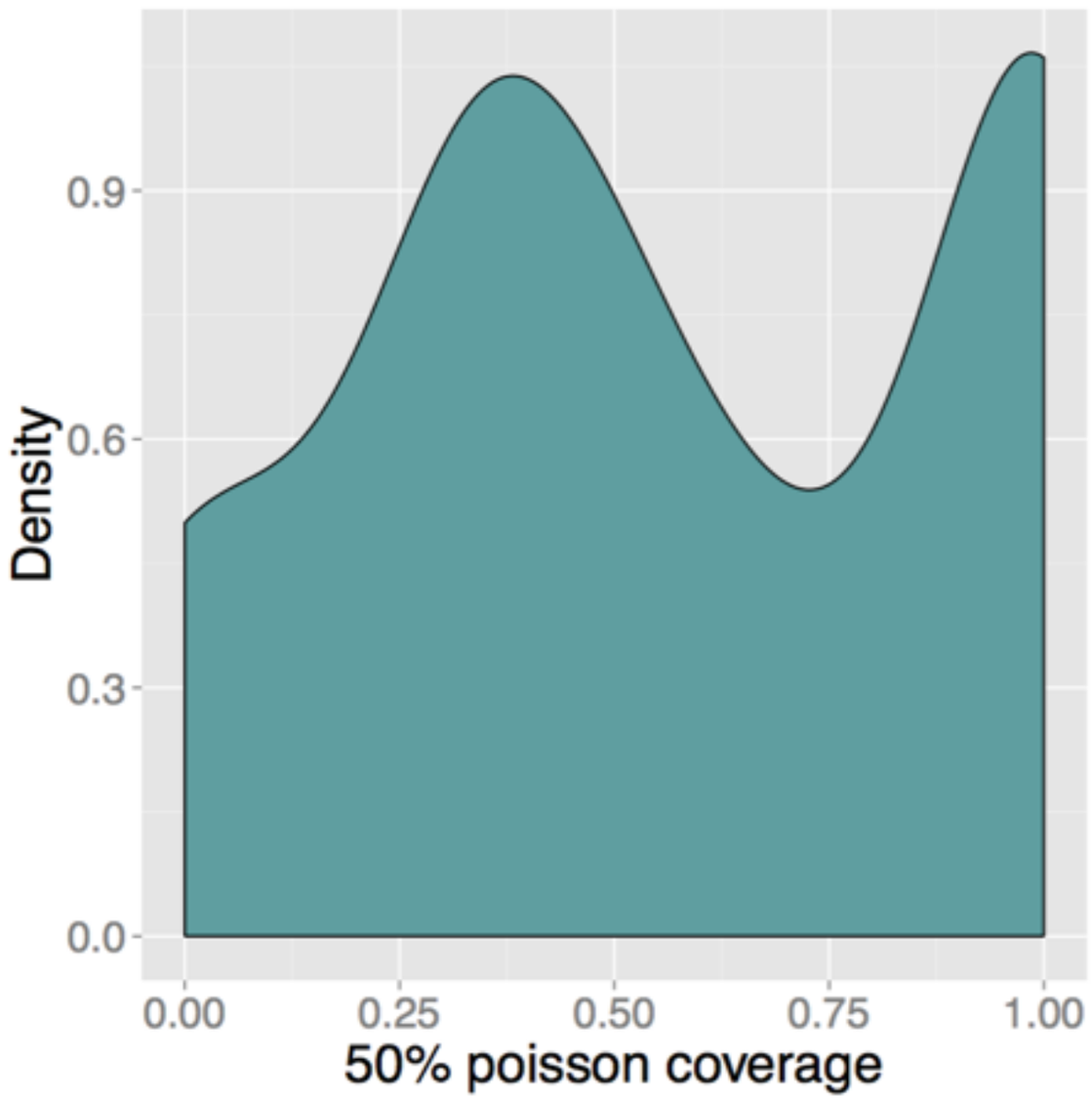
- the number of segments in phonological inventories vary greatly, but the number of features per inventory may be stable within language families

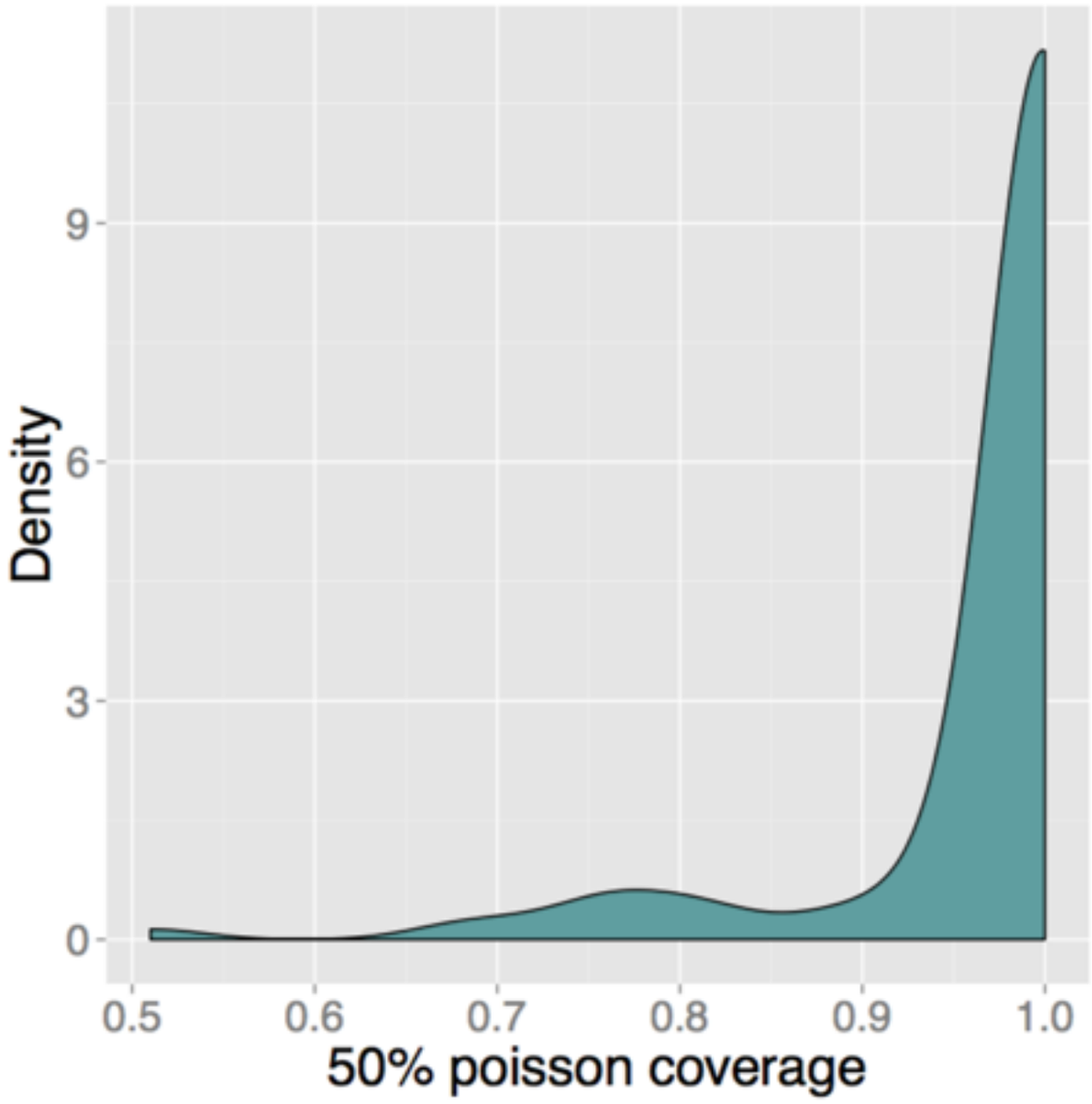




Capturing sizes

- the number of segments in phonological inventories vary greatly (left), but the number of features per inventory may be stable within language families (right)





Diachronic analysis

- application of Family Bias (Bickel 2011, 2013) to reduced feature sets
  - certain features are diachronically preferred
    - [continuant, coronal, dorsal, front, high, labial, nasal, voice, syllabic, sonorant]
  - others are dispreferred
    - [implosive, ejective, back, consonantal, spread_glottis, approximant, tap, long, round, labiodental, short, ATR, click, fortis]
- explained by specific combinations of phonetic efficiency or comprehension
- some broad areal patterns also appear
  - e.g. fricatives dispreferred in Australia

Conclusions

- Overall, these results show that algorithmically derived feature decompositions provide a fruitful but little-exploited terrain for phonological typology
- Even though segment space is nonfinite, a phonetically grounded system of 37 features suffices to code them all
- Our analyses reveal strong constraints on the organization and evolution of sound systems
- All these constraints are probabilistic, not categorical, in line with much recent work on the nature and emergence of phonology (Blevins 2004, Mielke 2004, Sandler et al 2011, Collier et al. 2014)
- Converging evidence from neuroscientific discovery of feature responses in the brain (Mesgarani et al. 2014)

Acknowledgements

- Conference organizers and colleagues
  - Claudia Bavero, Bernard Comrie, David Gil, Sven Grawunder, Martin Haspelmath, Paul Heggarty, Susanne Maria Michaelis, Frank Seifart
- Collaborators
  - Daniel McCloy, Richard Wright & Robert Forkel
- Data providers
  - Ian Maddieson, Lev Michael & Will Chang
- Organizations and support
  - MPI-EVA, U. Zurich, U. Washington, UW RRF