The Low-Complexity-Belt:

Evidence for Large-Scale Language Contact in Human Prehistory?

Christian Bentz
University of Tübingen

March 22, 2016

WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

EVOLAEMP
Language Evolution: The Empirical Turn

Overview

Data and Methods

Parallel Bible Corpus [Mayer & Cysouw, 2014]

```
# language_name:        English
# closest ISO 639-3:    eng
# year_short:           1890
# year_long:            Not available
# title:                The Bible in English, Darby Translation
# URL:                  http://unbound.biola.edu/index.cfm?method=downloa
# copyright_short:      © Public Domain
# copyright_long:       First published in 1890 by John Nelson Darby, an
Darby also published translations of the Bible in French and German.
01001001        In the beginning God created the heavens and the earth .
01001002        And the earth was waste and empty , and darkness was on t
01001003        And God said , Let there be light . And there was light .
01001004        And God saw the light that it was good ; and God divided
01001005        And God called the light Day , and the darkness he called
01001006        And God said , Let there be an expanse in the midst of th
01001007        And God made the expanse , and divided between the waters
01001008        And God called the expanse Heavens . And there was evenin
01001009        And God said , Let the waters under the heavens be gather
01001010        And God called the dry [ land ] Earth , and the gathering
```

DATA AND METHODS

Parallel Bible Corpus [Mayer & Cysouw, 2014]

```
# language_name:        Amharic
# closest ISO 639-3:    amh
# year_short:           1994
# year_long:            E-Text in transliterated ASCII format by Lapsley/Brooks
(www.nt-text.net). Revised Amharic Bible in XML ( 2003 ).
# title:                The New Testament in Amharic
# URL:                  http://unbound.biola.edu/index.cfm?method=downloads.show
# copyright_short:      © Printed Version by United Bible Societies 1962
# copyright_long:       Not available
40001001        የዳዊት ልጅ የአብርሃም ልጅ የኢየሱስ ክርስቶስ ትውልድ መጽሐፍ ።
40001002        አብርሃም ይስሐቅን ወለደ ፤ ይስሐቅም ያዕቆብን ወለደ ፤ ያዕቆብም ይሁዳንና ወንድሞቹን ወ
40001003        ይሁዳም ከትዕማር ፋሬስንና ዛራን ወለደ ፤ ፋሬስም ኤስሮምን ወለደ ፤
40001004        ኤስሮምም አራምን ወለደ ፤ አራምም አሚናዳብን ወለደ ፤ አሚናዳብም ነአሶንን ወለደ ፤ ነአሶ
40001005        ሰልሞንም ከራኬብ ቦኤዝን ወለደ ፤ ቦኤዝም ከሩት ኢዮቤድን ወለደ ፤ ኢዮቤድም እሴይን ወለ
40001006        እሴይም ንጉሥ ዳዊትን ወለደ ።
40001007        ሰሎሞንም ሮብዓምን ወለደ ፤ ሮብዓምም አቢያን ወለደ ፤ አቢያም አሣፍን ወለደ ፤
40001008        አሣፍም ኢዮሣፍጥን ወለደ ፤ ኢዮሣፍጥም ኢዮራምን ወለደ ፤ ኢዮራምም ዖዝያንን ወለደ ፤
40001009        ዖዝያንም ኢዮአታምን ወለደ ፤ ኢዮአታምም አካዝን ወለደ ፤
40001010        አካዝም ሕዝቅያስን ወለደ ፤ ሕዝቅያስም ምናሴን ወለደ ፤ ምናሴም አሞፅን ወለደ ፤
40001011        አሞፅም ኢዮስያስን ወለደ ፤ ኢዮስያስም በባቢሎን ምርኮ ጊዜ ኢኮንያንንና ወንድሞቹን ወለደ
```

Data and Methods

Parallel Bible Corpus [Mayer & Cysouw, 2014]

```
# language_name:        ភាសាខ្មែរ
# closest ISO 639-3:    khm
# year_short:           2011
# year_long:            Not available
# title:                Khmer Christian Bible<br>The New Testament in Khmer
# URL:                  https://www.bible.com/de/bible/315/mat.1.kcb
# copyright_short:      © Words of Life Ministries 2011
# copyright_long:       Khmer Christian Bible<br>Copyright © Holy Bible, Khmer Christian Bible
#                       copyright 2011 by Words of Life Ministries, P.O. Box 2581, Phnom Penh,
#                       3, Cambodia. All rights reserved.
40001001        កំណត់ត្រាវង្សត្រកូលរបស់ព្រះយេស៊ូគ្រិស្តដែលជាពូជពង្សរបស់ស្តេចដាវីឌ និងលោកអ័ប្រាហាំ :
40001002        លោកអ័ប្រាហាំបង្កើតលោកអ៊ីសាក លោកអ៊ីសាកបង្កើតលោកយ៉ាកុប លោកយ៉ាកុបបង្កើតលោកយូដា និ
40001003        លោកយូដា និងនាងតាម៉ារបង្កើតលោកពេរេស និងលោកសេរ៉ាស ឯលោកពេរេសបង្កើតលោកហេស្រុន
40001004        លោករ៉ាមបង្កើតលោកអ័មីណាដាប់ លោកអ័មីណាដាប់បង្កើតលោកណាសូន លោកណាសូនបង្កើតលោ
40001005        លោកសាលម៉ូន និងនាងរ៉ាហាប់បង្កើតលោកបូអូស ហើយលោកបូអូស និងនាងរស់បង្កើតលោកអូបិឌ ព
40001006        លោកអ៊ីសាយបង្កើតស្តេចដាវីឌ ស្តេចដាវីឌ និងប្រពន្ធលោកអ៊ូរីបង្កើតស្តេចសាឡូម៉ូន
40001007        ស្តេចសាឡូម៉ូនបង្កើតស្តេចរេហូបោម ស្តេចរេហូបោមបង្កើតស្តេចអ័ប៊ីយ៉ា ស្តេចអ័ប៊ីយ៉ាបង្កើតស្តេចអេស
40001008        ស្តេចអេសាបង្កើតស្តេចយ៉ូសាផាត ស្តេចយ៉ូសាផាតបង្កើតស្តេចយ៉ូរាម ស្តេចយ៉ូរាមបង្កើតស្តេចអូសៀស
40001009        ស្តេចអូសៀសបង្កើតស្តេចយ៉ូថាម ស្តេចយ៉ូថាមបង្កើតស្តេចអេហាស ស្តេចអេហាសបង្កើតស្តេចអេសេគ
40001010        ស្តេចអេសេគាសបង្កើតស្តេចម៉ាណាសេ ស្តេចម៉ាណាសេបង្កើតស្តេចអាំម៉ូន ស្តេចអាំម៉ូនបង្កើតស្តេចយ៉ូ
```

Data and Methods

Parallel Corpora

Corpus	Register	Size*	Size ∅*	Texts	Lang.
UDHR[1]	Legal	ca. 650K	1.831	372	348
PBC[2]	Religious	ca. 10M	261K	1136	890
EPC[3]	Political	ca. 150M	7M	21	21
	Total	**ca. 161M**		**1529**	**1050**

*in number of tokens

[1] [http://unicode.org/udhr/ translations.html]
[2] [Mayer & Cysouw, 2014]
[3] [Koehn, 2005]

Measuring Complexity

Information-theoretic account:
What is the distribution of information-encoding units given constant content of the message?

[Shannon & Weaver, 1949; Juola 1998, 2008; Ehret & Szmrecsanyi, 2015]

Measuring Complexity

Genesis 1:1

|in| the beginning god created the heavens and the earth and the earth was waste and empty and darkness was on the face of the deep and the spirit of god was hovering over the face of the waters and god said let there be light and there was light

Information encoding unit: |orthographic word|

Measuring Complexity

Linguistic/information-theoretic account:
What is the distribution of word types in different languages, given constant content of the message?

Measuring Complexity

Example: English and German definite articles

in the beginning god created the heavens and the earth

in dem anfang schuf gott den himmel und die erde

Measuring Complexity

Example: English and German definite articles

in the beginning god created the heavens and the earth

in dem anfang schuf gott den himmel und die erde

$$Entropy : H = -K \sum_{i=1}^{r} p(w_i) \log_2(p(w_i))$$

[Shannon & Weaver, 1948]

Measuring Complexity

Example: English and German definite articles

in the beginning god created the heavens and the earth

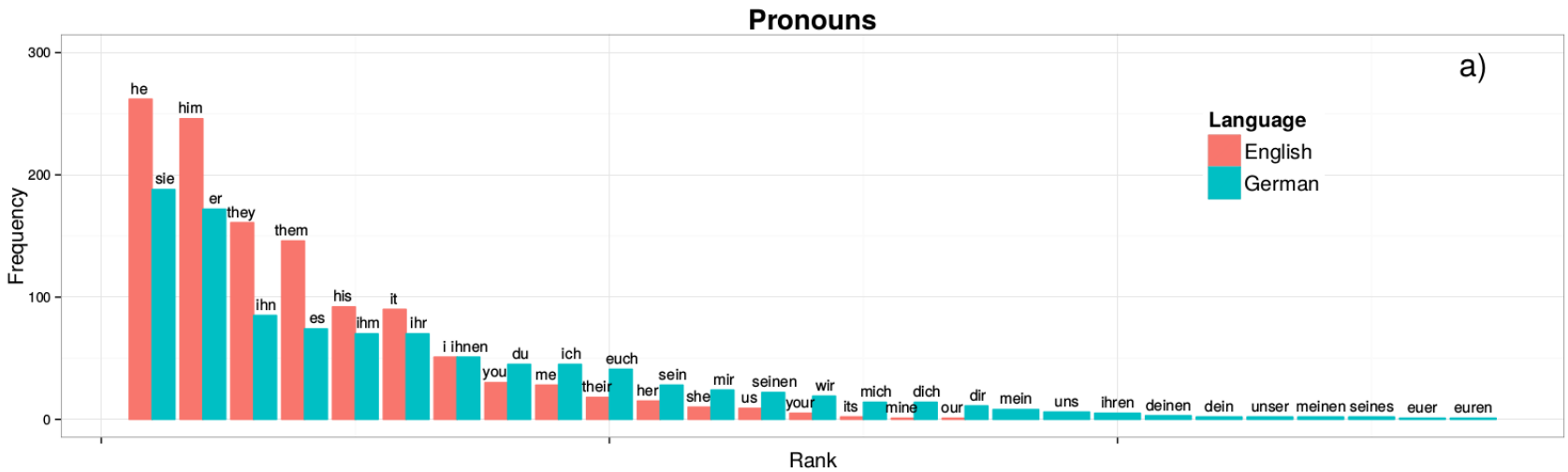
in dem anfang schuf gott den himmel und die erde

$$H_{eng} = -(\frac{3}{10}\log_2(\frac{3}{10}) + \frac{1}{10}\log_2(\frac{1}{10}) + \cdots + \frac{1}{10}\log_2(\frac{1}{10})) \approx 3.17$$

$$H_{deu} = -(\frac{1}{10}\log_2(\frac{1}{10}) + \cdots + \frac{1}{10}\log_2(\frac{1}{10})) \approx 3.32$$

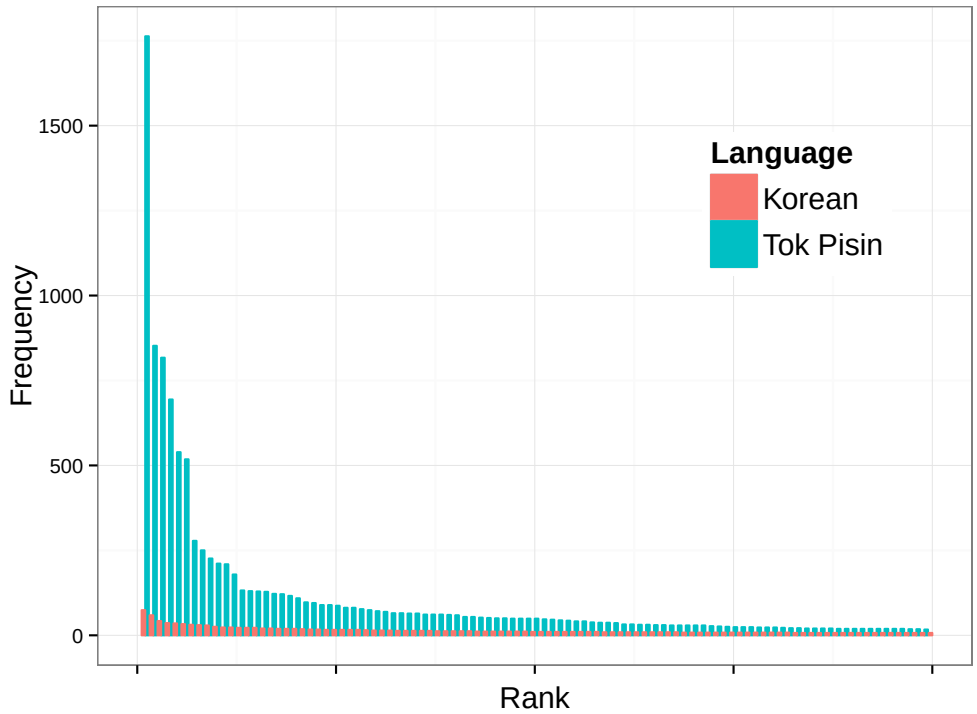
MEASURING COMPLEXITY

Example: English and German pronouns



$$H_{eng} = 2.16$$
$$H_{deu} = 2.64$$

Measuring Complexity



$$H_{kor} = 12.13$$
$$H_{tpi} = 6.83$$

Complexity and Iterated Learning

[Kirby et al. 2008, 2015]

Complexity and Iterated Learning

[Kirby et al. 2008, 2015]

Learnability only

Generation	0	1	2	3	4	5	6	7	8	9	10
○ Chain 1	27	17	9	6	5	4	4	2	2	2	2
□ Chain 2	27	17	15	8	7	6	6	6	5	5	4
△ Chain 3	27	24	8	6	6	5	6	5	5	5	5
◇ Chain 4	27	23	9	10	9	11	7	5	5	4	4

$$H_{start} = 4.75 \rightarrow H_{end} = 2.11$$

Complexity and Iterated Learning

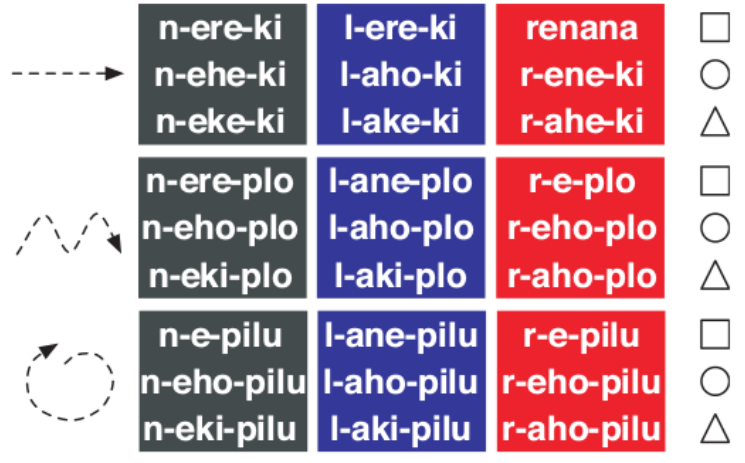
[Kirby et al. 2008, 2015]

Learnability only

Generation	0	1	2	3	4	5	6	7	8	9	10
○ Chain 1	27	17	9	6	5	4	4	2	2	2	2
□ Chain 2	27	17	15	8	7	6	6	6	5	5	4
△ Chain 3	27	24	8	6	6	5	6	5	5	5	5
◇ Chain 4	27	23	9	10	9	11	7	5	5	4	4

$$H_{start} = 4.75 \rightarrow H_{end} = 2.11$$

Expressivity + Learnability



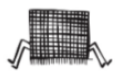
Generation	0	1	2	3	4	5	6	7	8	9	10
○ Chain 1	27	23	22	17	21	21	17	21	25	13	16
□ Chain 2	27	26	13	10	10	16	16	12	12	13	12
△ Chain 3	27	11	16	14	12	17	14	16	20	19	12
◇ Chain 4	27	19	19	17	19	17	22	23	21	27	23

$$H_{start} = 4.75 \rightarrow H_{end} = 4.75$$

COMPLEXITY AND ITERATED LEARNING

L2 speakers reduce complexity

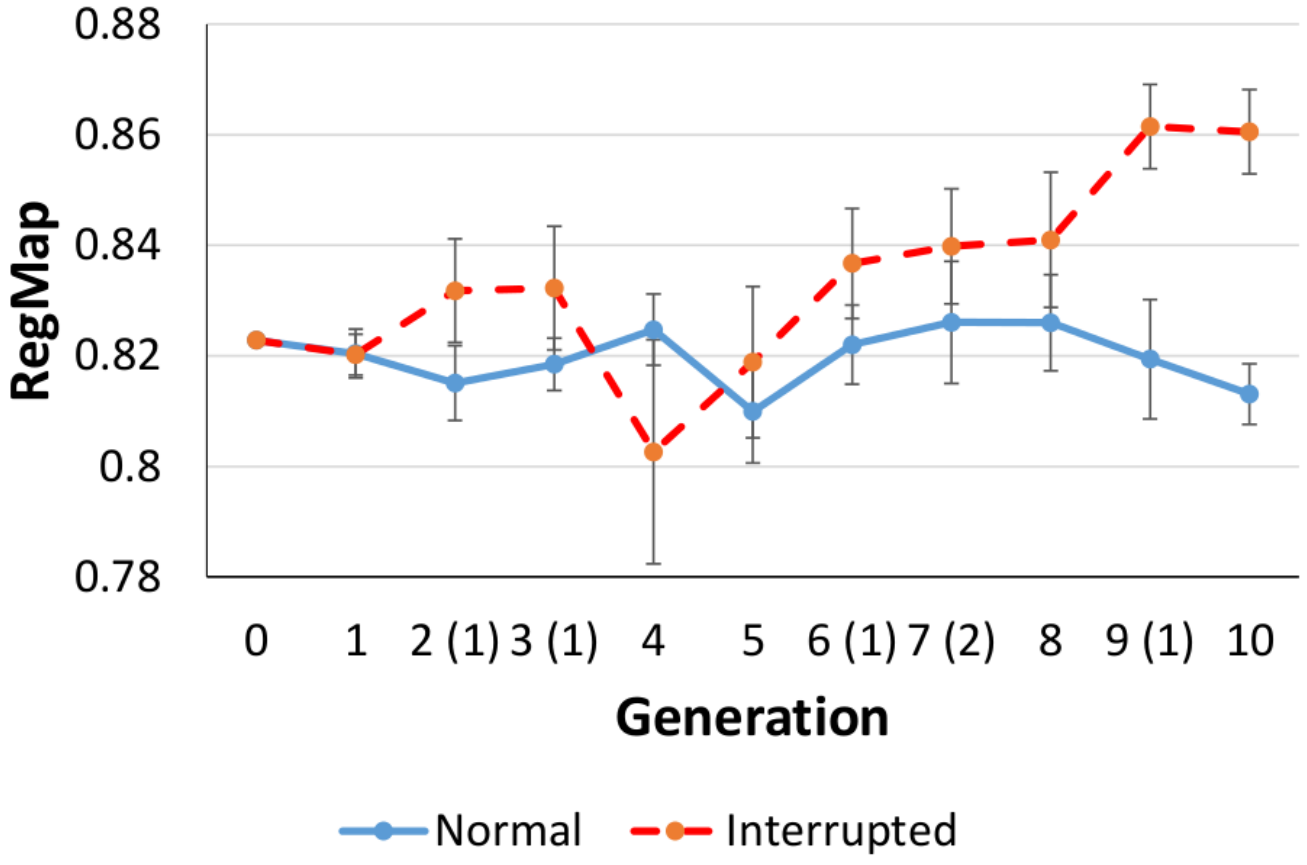
[Berdicevskis & Semenuks, under review]

		event: none	event: fall apart	event: grow antlers	event: fly
agent: round animal	number: singular	 seg_N	 seg_N m_V-o_{AGR}	 seg_N r_V-o_{AGR}	 seg_N b_V-o_{AGR}
	number: plural	 seg_N-l_{PL}	 seg_N-l_{PL} m_V-o_{AGR}	 seg_N-l_{PL} r_V-o_{AGR}	 seg_N-l_{PL} b_V-o_{AGR}
agent: square animal	number: singular	 fuv_N	 fuv_N m_V-i_{AGR}	 fuv_N r_V-i_{AGR}	 fuv_N b_V-i_{AGR}
	number: plural	 fuv_N-l_{PL}	 fuv_N-l_{PL} m_V-i_{AGR}	 fuv_N-l_{PL} r_V-i_{AGR}	 fuv_N-l_{PL} b_V-i_{AGR}

Complexity and Iterated Learning

L2 speakers reduce complexity

[Berdicevskis & Semenuks, under review]



Measuring Complexity

What about the word type entropies of natural languages?

Measuring Complexity

What about the word type entropies of natural languages?

1529 texts
1050 languages
140 families
23 areas

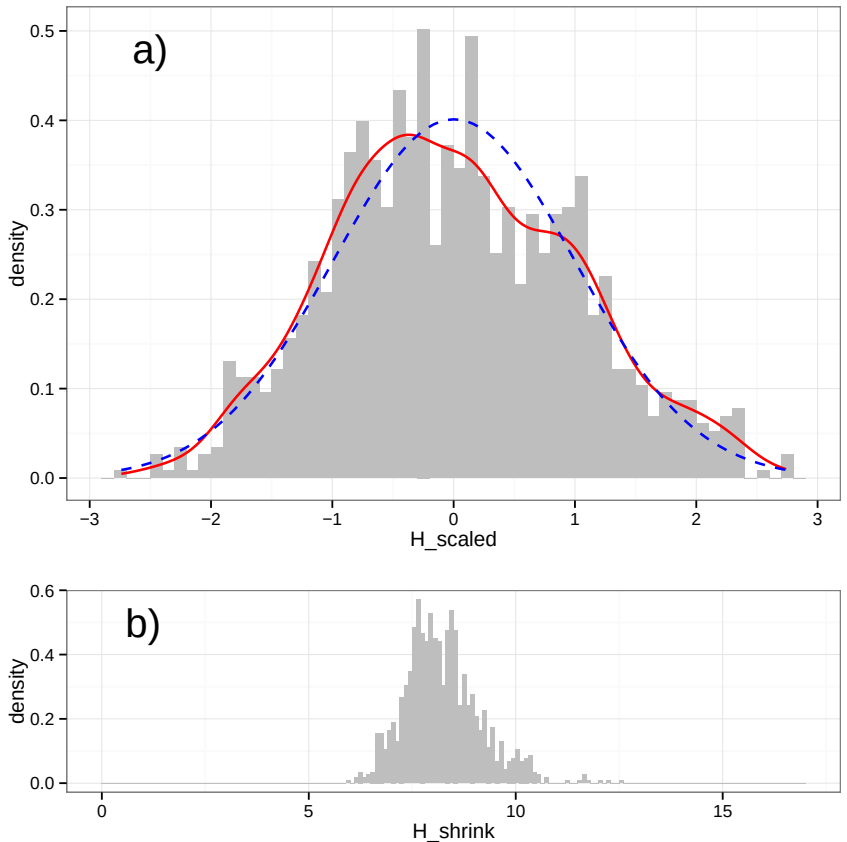
[AUTOTYP database, Bickel & Nichols, 1999]

Measuring Complexity

What about the word type entropies of natural languages?

1529 texts
1050 languages
140 families
23 areas

[AUTOTYP database, Bickel & Nichols, 1999]

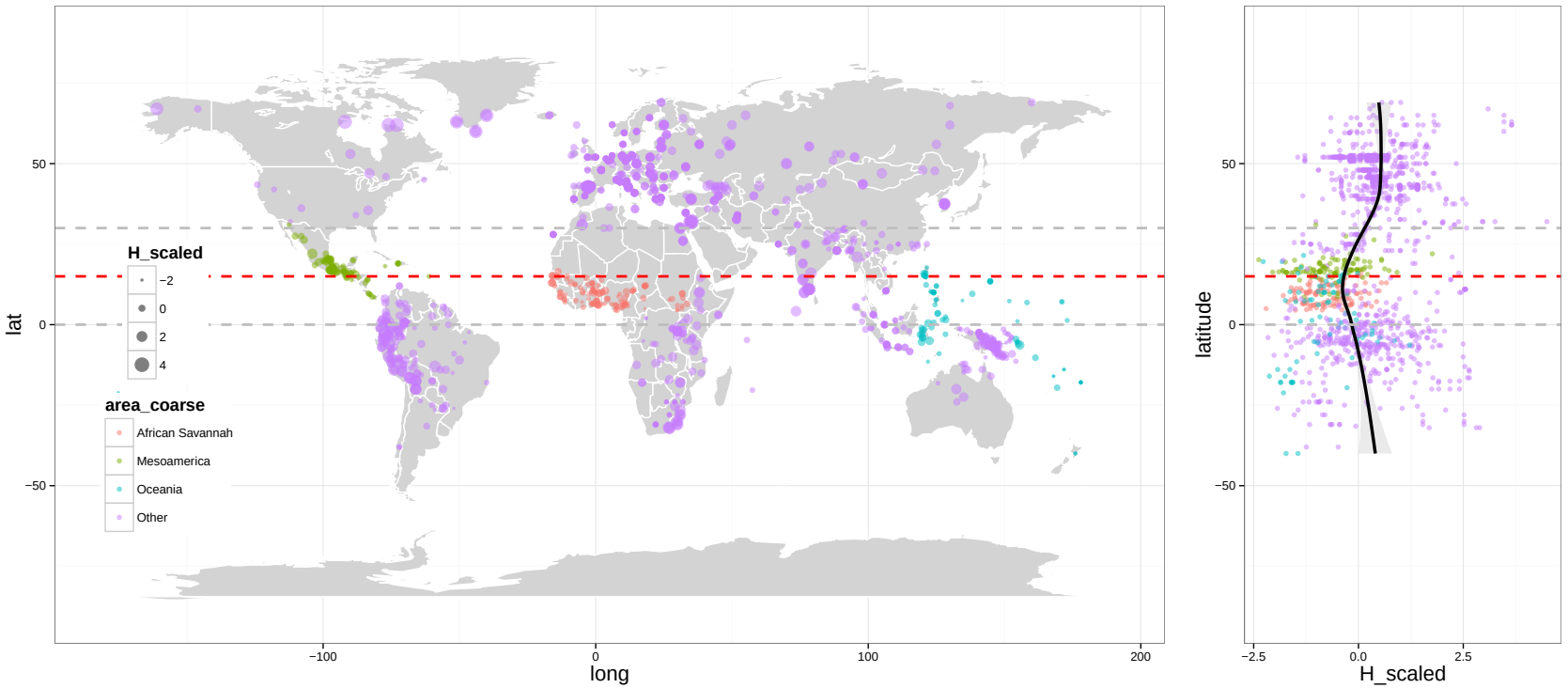


The Low-Complexity-Belt

Longitude and latitude data from *Glottolog 2.6*
[Hammarström, Forkel, Haspelmath & Bank, 2015]

The Low-Complexity-Belt

Longitude and latitude data from *Glottolog 2.6*
[Hammarström, Forkel, Haspelmath & Bank, 2015]



[R package ggmaps]

Latitude and Entropy relationship

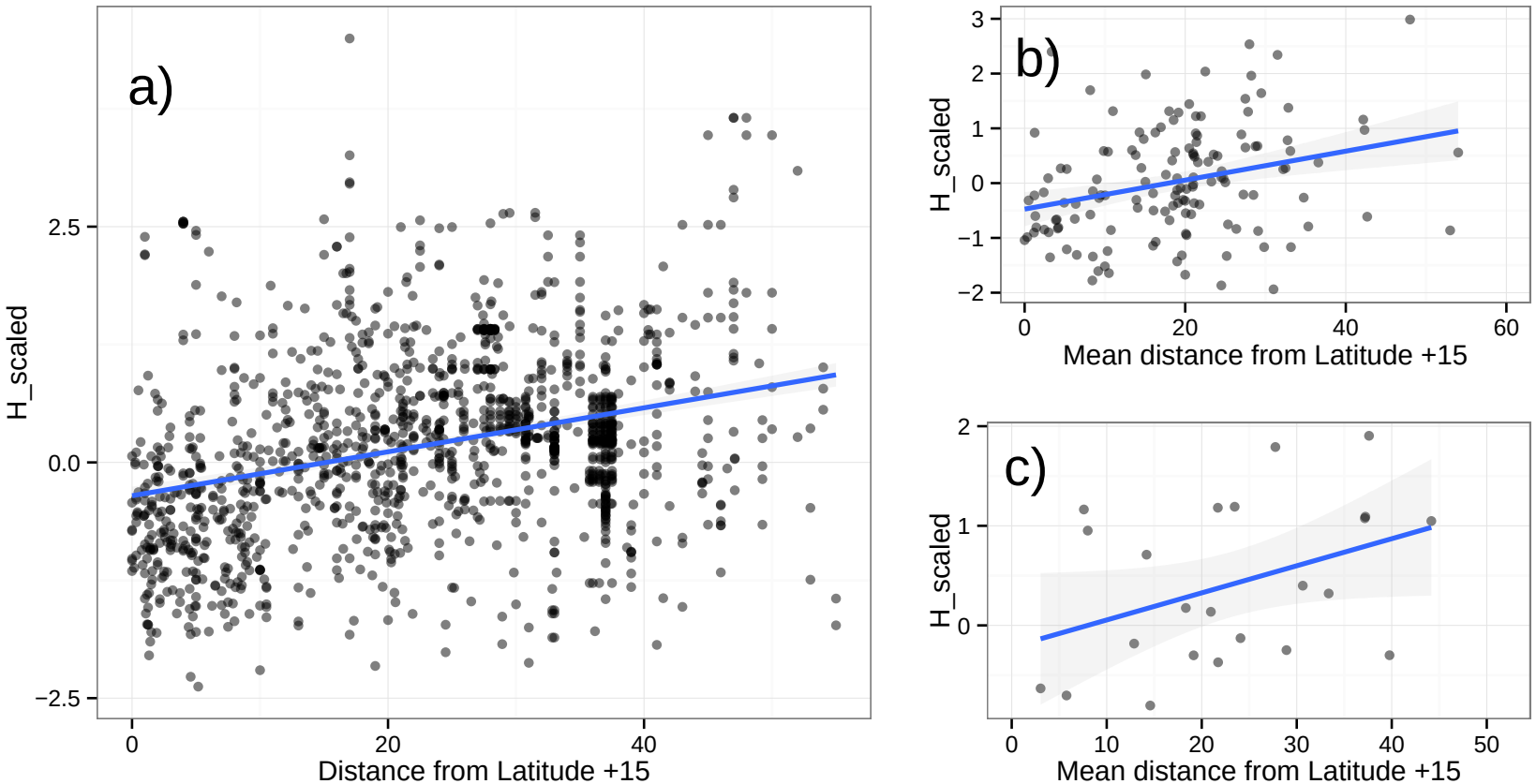
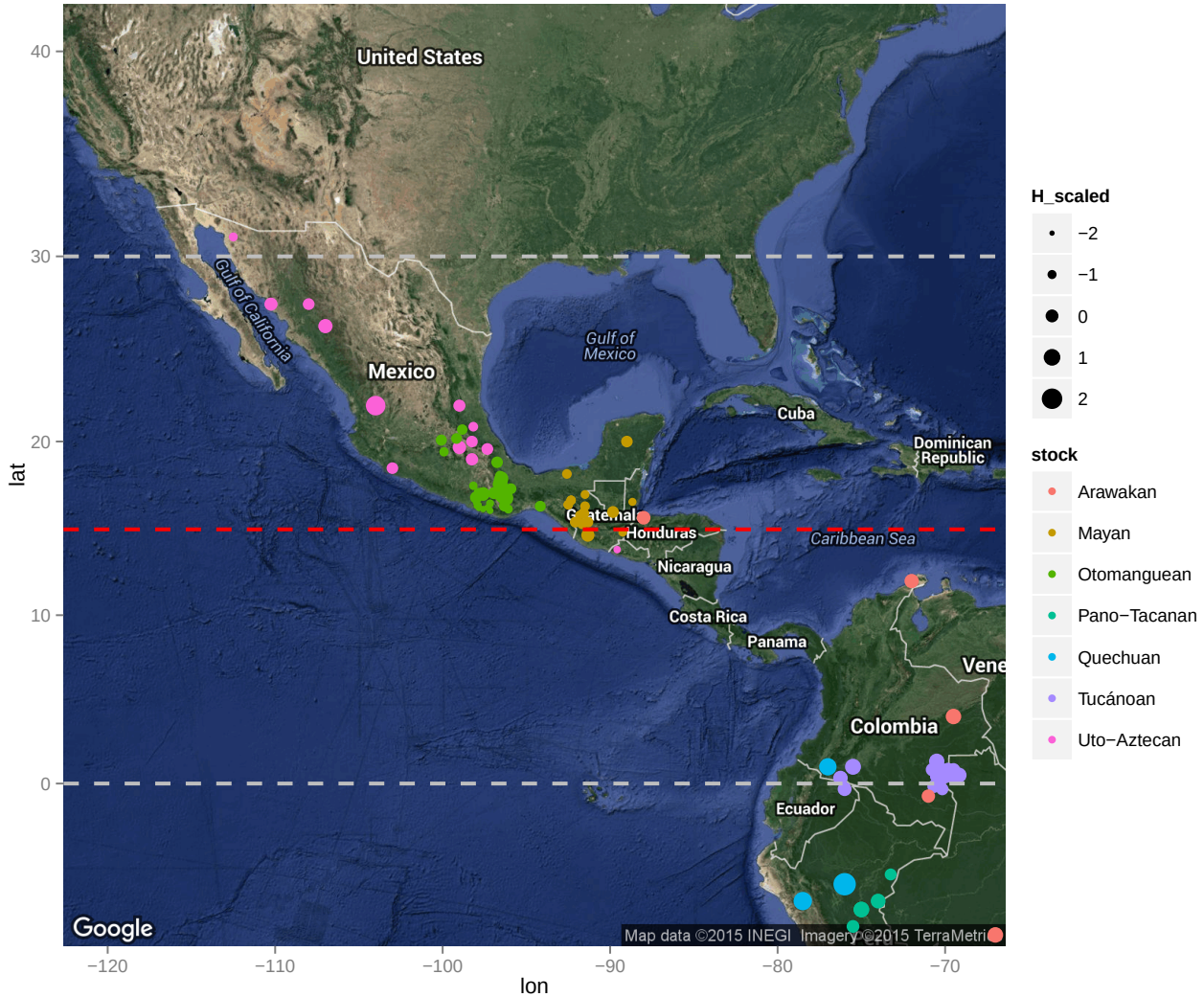


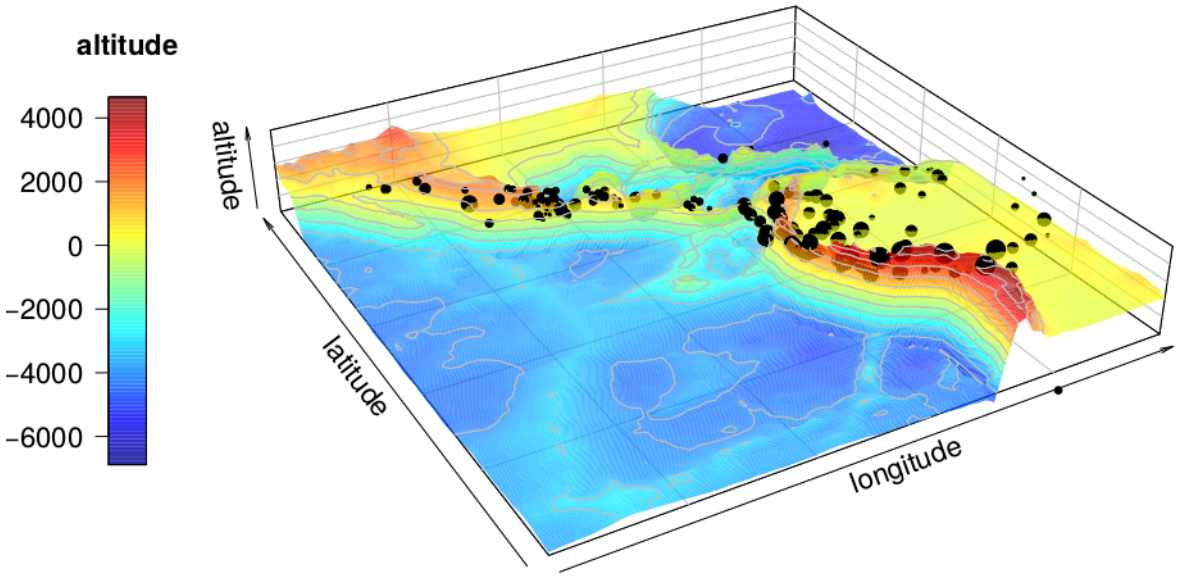
Figure : a) All languages, b) Family averages and c) Area averages.

Mesoamerica (*only families (stocks) with > 10 languages)



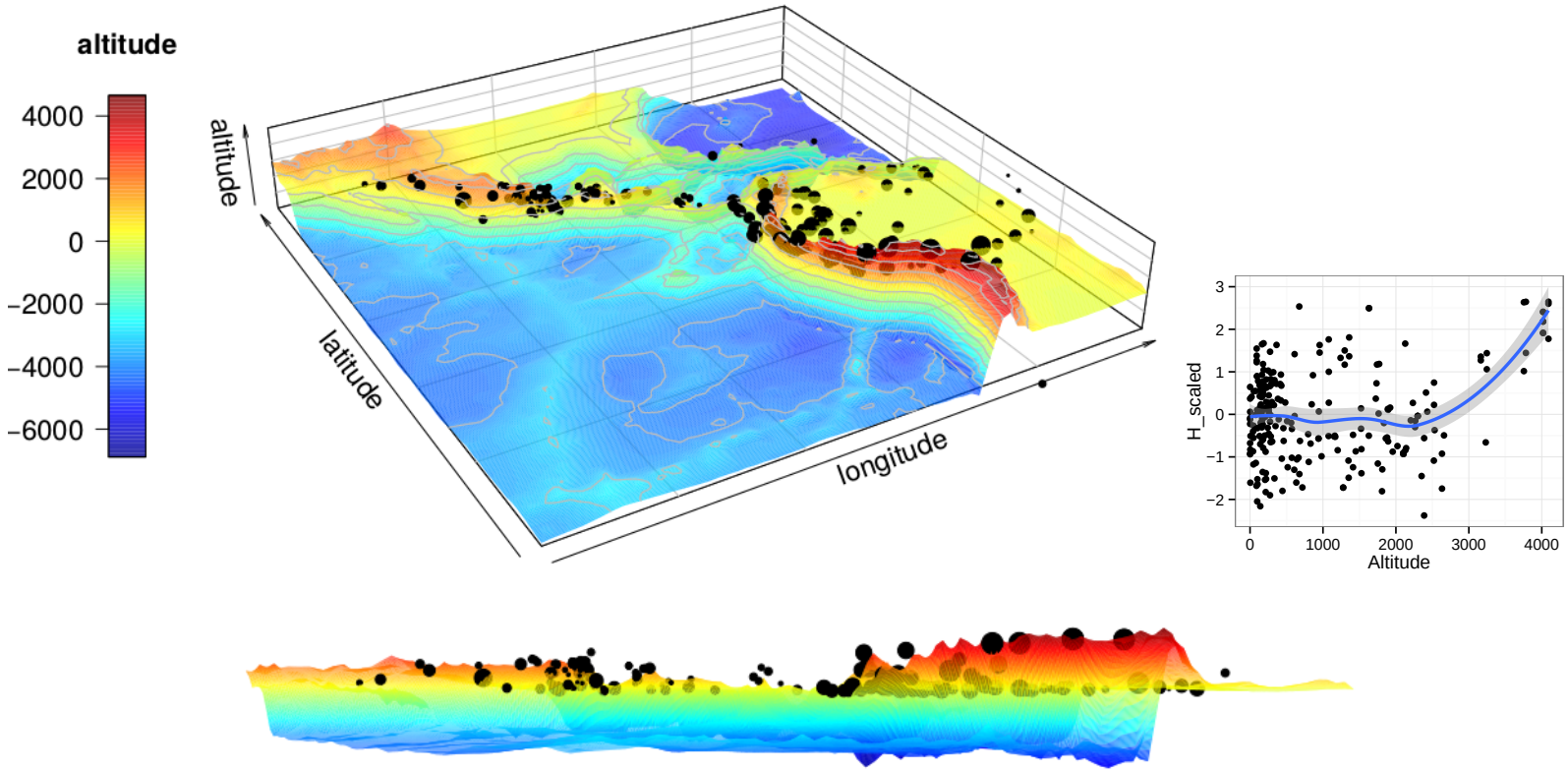
Mesoamerica and the Andes

183 languages, 90 families, 3 areas

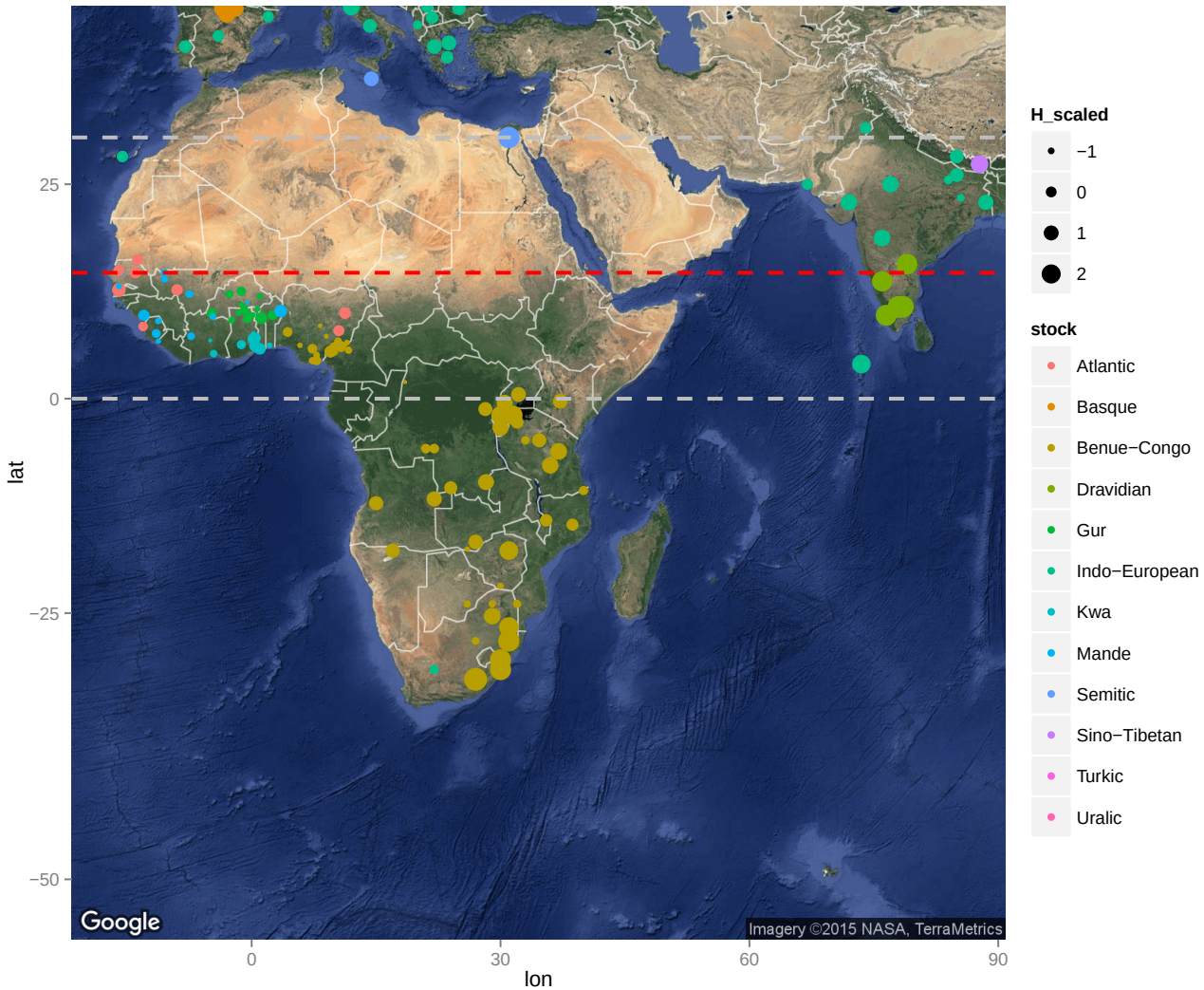


Mesoamerica and the Andes

183 languages, 90 families, 3 areas

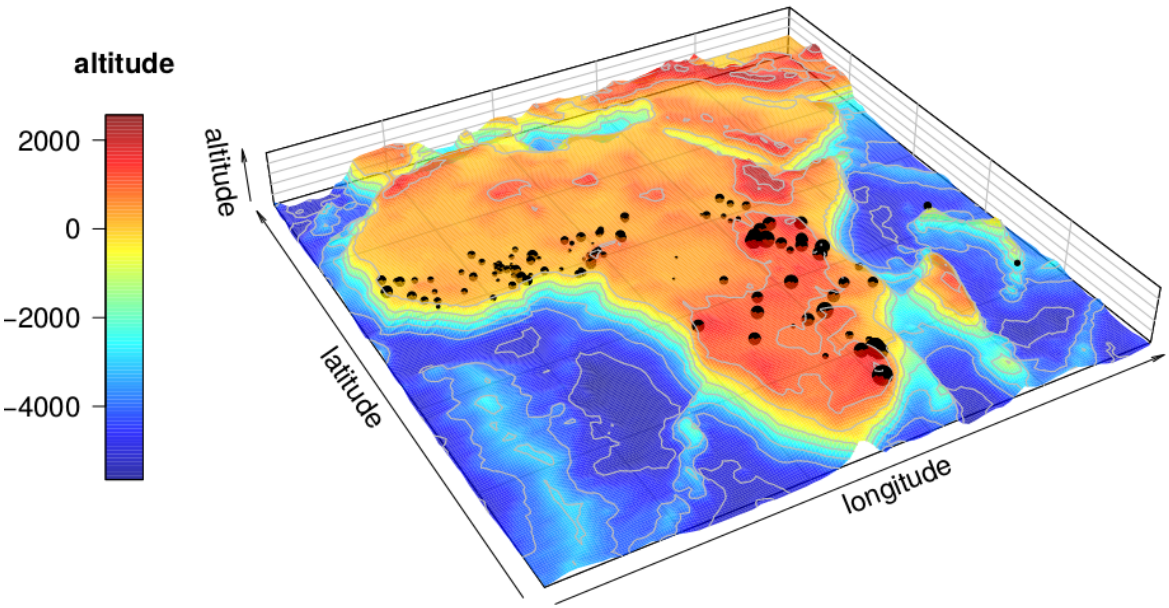


African Savannah (*only families (stocks) with > 10 languages)



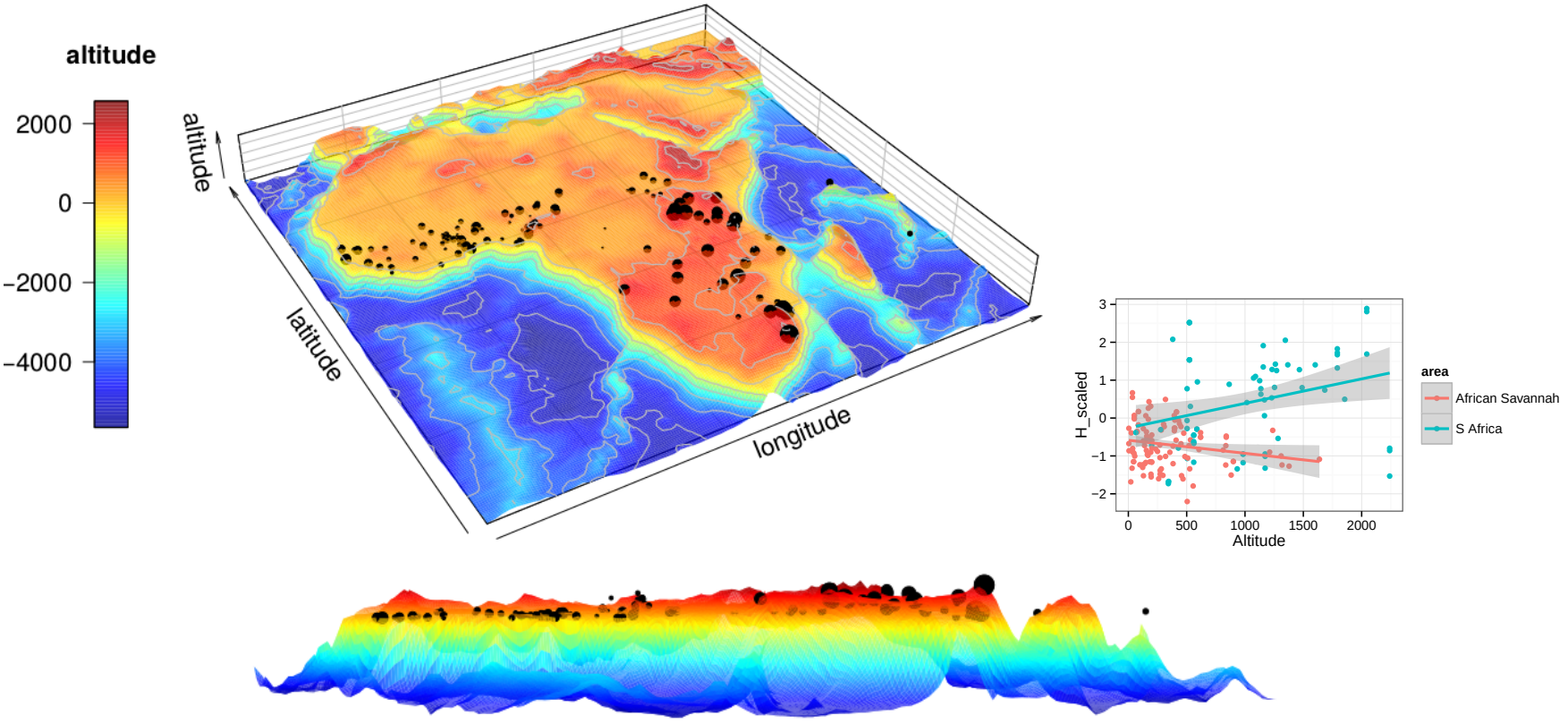
African Savannah and South Africa

127 languages, 21 families, 2 areas

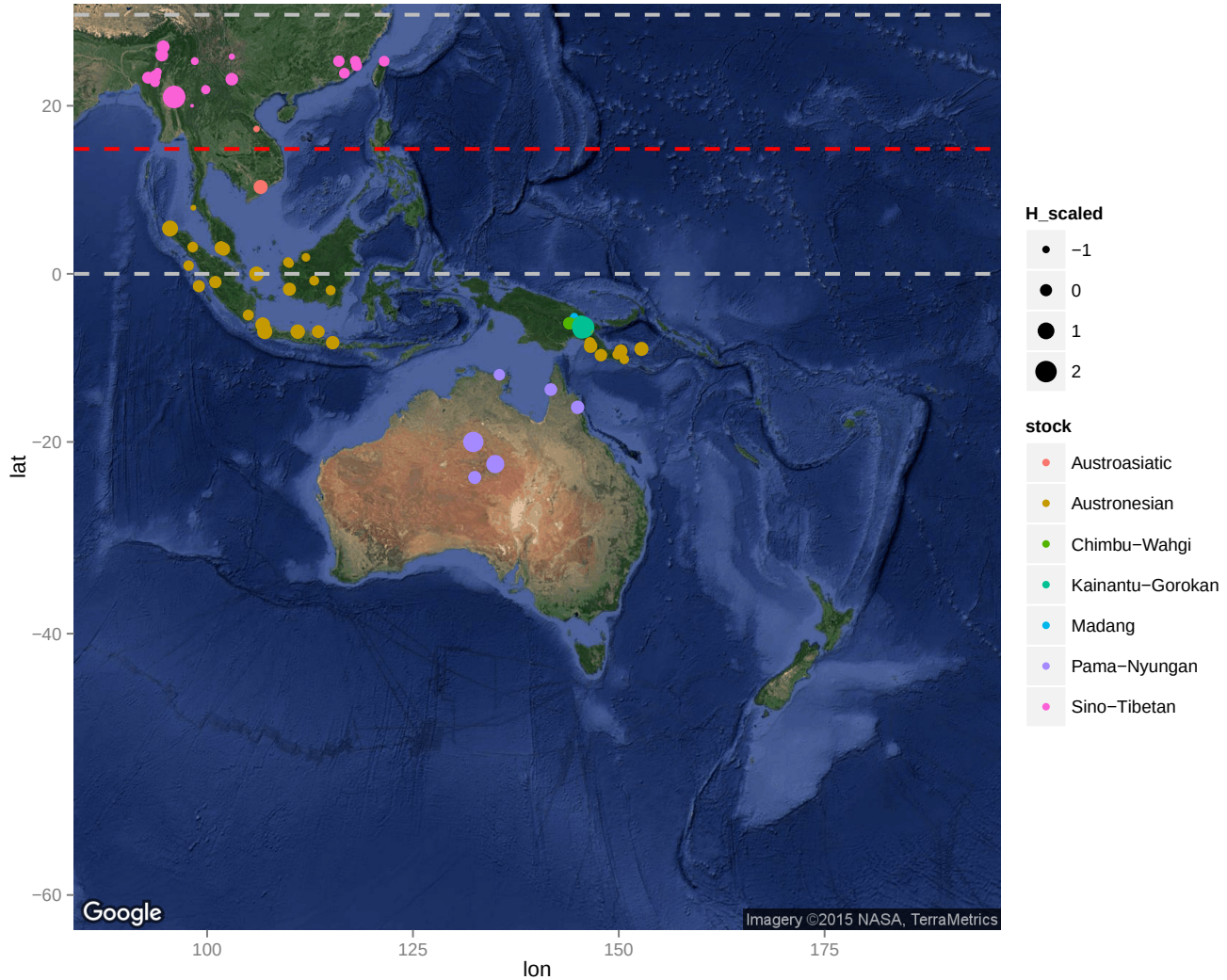


African Savannah and South Africa

127 languages, 21 families, 2 areas



Oceania (*only families (stocks) with > 10 languages)



Question

- Does the entropy/latitude relationship hold if we take family and area idiosyncrasies into account?

Mixed-Effects Regression

Table : Predicting entropy in linear mixed-effects regression.

Fixed	Random		AIC	p-value	$R^{2\ddagger}$	
	intercept	slope			f	f+r
Latitude	s,a,c,i	s,a,c	1931	**0.12**	**0.03**	**0.95**

† s: stock, a: area, c: corpus, i: iso
‡ f: fixed effect only, f+r: fixed and random effects

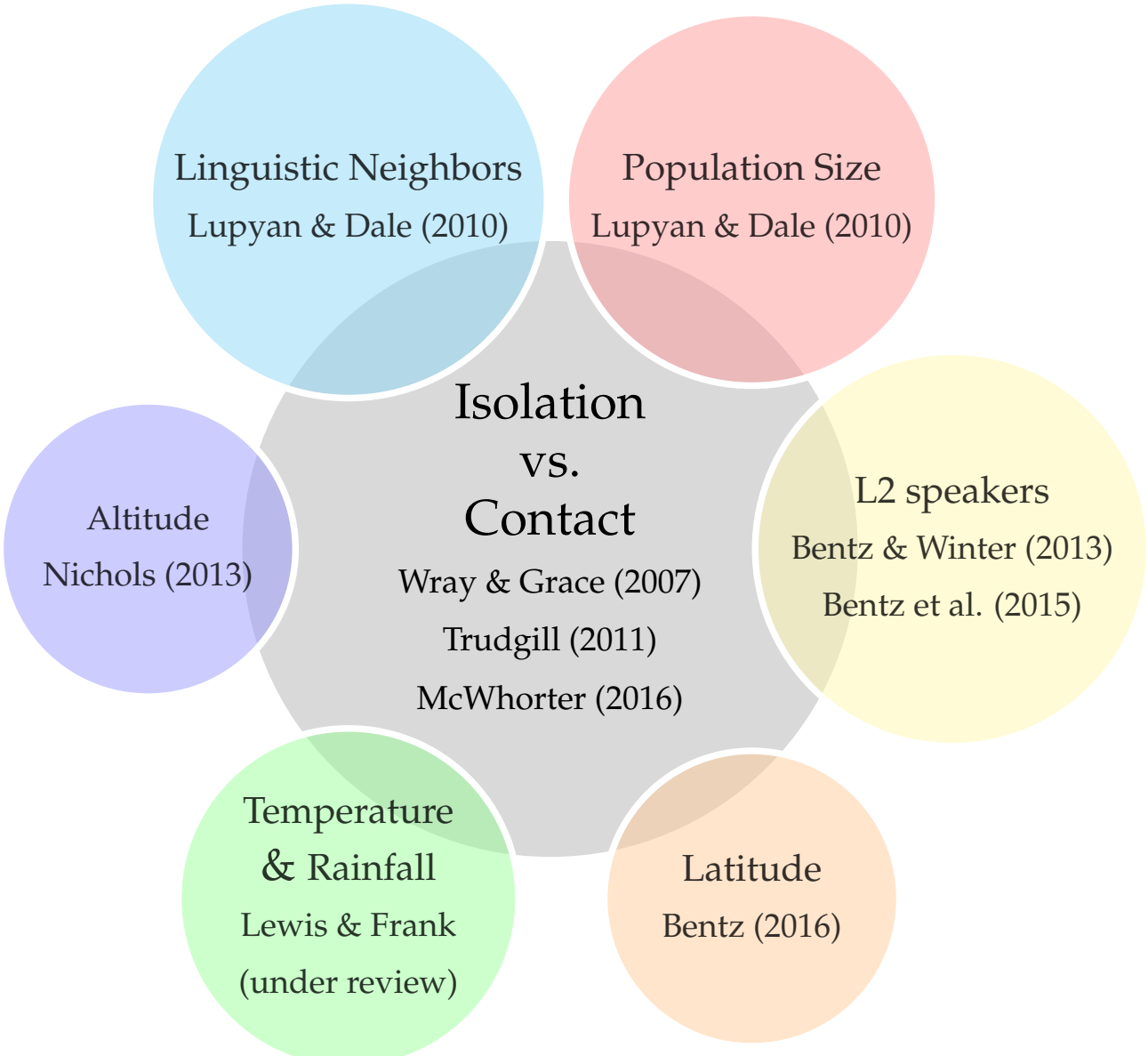
[R package *lme4*]
[R package *MuMIn*]

Summary

- The *Low-Complexity-Belt* is driven by specific macroareas: e.g. Mesoamerica, African Savannah, Oceania

Summary

- The *Low-Complexity-Belt* is driven by specific macroareas: e.g. Mesoamerica, African Savannah, Oceania
- Within these areas, there are families that drive the pattern: e.g. Otomanguean, Mayan, Gur, etc.


Linguistic Neighbors
Lupyan & Dale (2010)
Population Size
Lupyan & Dale (2010)
Isolation
vs.
Contact
Wray & Grace (2007)
Trudgill (2011)
McWhorter (2016)
Altitude
Nichols (2013)
L2 speakers
Bentz & Winter (2013)
Bentz et al. (2015)
Temperature
& Rainfall
Lewis & Frank
(under review)
Latitude
Bentz (2016)

Why Prehistoric Language Contact?

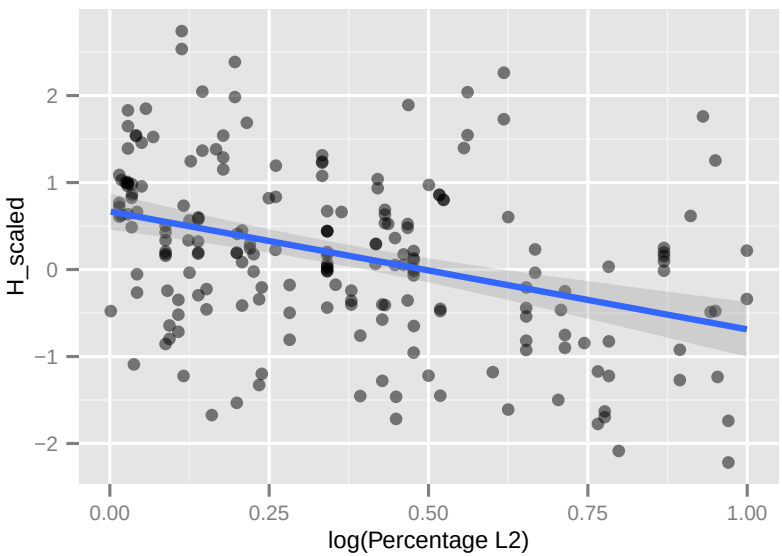
"[...] radical analyticity in certain West Benue-Congo languages of Niger-Congo (Gbe, Yoruboid, Nupoid), and all Chinese, Tai-Kadai, Hmong-Mien ones, most Austroasiatic ones and many Tibeto-Burman ones [...] is the **result of widespread adult acquisition in the past** [...]"

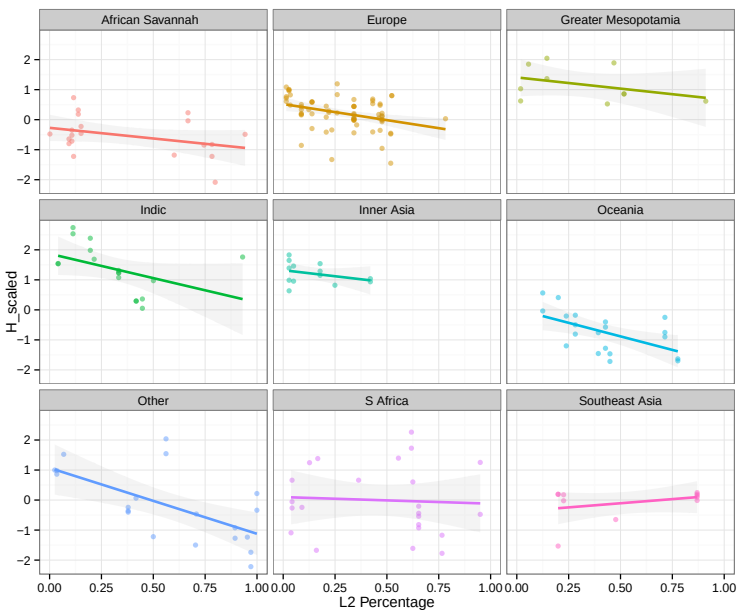
[McWhorter (2016) *Is radical analyticity normal?*, p. 52]

Why Prehistoric Language Contact?

Two main findings of
[Bentz, Verkerk, Kiela, Hill & Buttery 2015]:

- a) Entropies (complexities) of languages correlate negatively with non-native speaker proportions





Why Prehistoric Language Contact?

Two main findings of Bentz et al. 2015:

- b) Entropies (complexities) have a deep phylogenetic signal, while non-native speaker proportions do not

Table 5. Results for the phylogenetic signal analysis (mean λ).

Family	Text	α	H_w	TTR
Austronesian	UDHR	0.98	1	1
Austronesian	PBC	0.94	0.82	1
Bantu	UDHR	0.46	0.85	0.58
Indo-European	UDHR	1	0.64	1

doi:10.1371/journal.pone.0128254.t005

Conclusions

- The complexity of languages measured by the entropy of parallel corpora

Conclusions

- The complexity of languages measured by the entropy of parallel corpora
- Languages are less complex around the equator

Conclusions

- The complexity of languages measured by the entropy of parallel corpora
- Languages are less complex around the equator
- This potentially reflects deep contact, i.e. non-native language learning

Collaborators





Collaborators








Thank you!