Adaptive Languages

An Information-Theoretic Account of Linguistic Diversity

Christian Bentz
University of Tübingen

January 17, 2017

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

EVOLAEMP
LANGUAGE EVOLUTION: THE EMPIRICAL TURN

Overview

Intuitive Terminology

- order $\leftrightarrow$ disorder
- regularity $\leftrightarrow$ irregularity
- predictability $\leftrightarrow$ unpredictability
- certainty $\leftrightarrow$ uncertainty

$\Big\}$ Entropy

"*Entropy* as *possibility* is my favorite short description of entropy because (...) unlike *uncertainty* and *missing information*, it has positive connotation."

"*Entropy* as *possibility* is my favorite short description of entropy because (...) unlike *uncertainty* and *missing information*, it has positive connotation."

"Entropy is an additive measure of the **number of possibilities** available to a system."

Lemons, 2013

Entropy as Choice

Minimal

aaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaa

Entropy as Choice

Minimal

```
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
```

Maximal

```
fcbihspm hkgiwlelbj
sdmkfuuf cvkym cfcsqdvcc
trdgjmpnkjhujril
unnapsfmgbk ggqvntxprl
kfkmpsgjetn
grycfjuxxcusejlexfhkfrmh
jknecxjqg isonkqcwmxr
ymwwuieumi brlromnqyq
yclvlkmtgfd fcmvulfkyawa
```

Entropy as Choice

Minimal

```
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaa
```

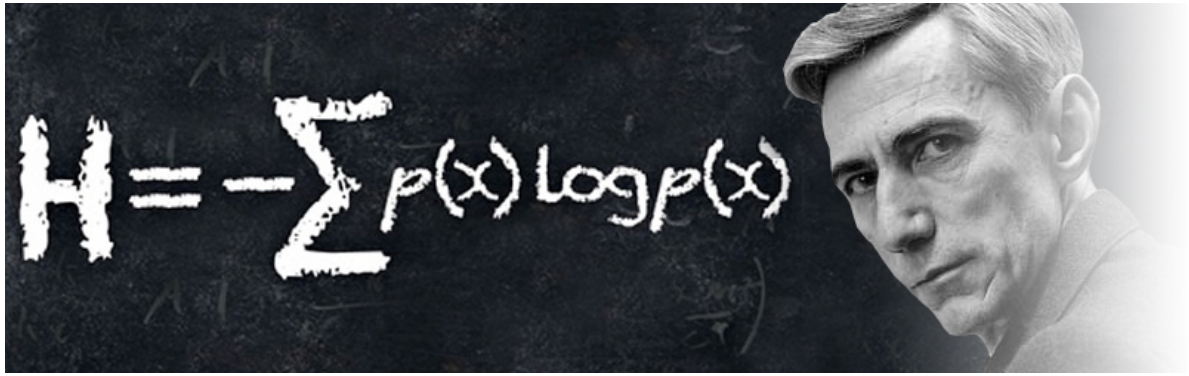
Maximal

```
fcbihspm hkgiwlelbj
sdmkfuuf cvkym cfcsqdvcc
trdgjmpnkjhujril
unnapsfmgbk ggqvntxprl
kfkmpsgjetn
grycfjuxxcusejlexfhkfrmh
jknecxjqg isonkqcwmxr
ymwwuieumi brlromnqyq
yclvlkmtgfd fcmvulfkyawa
```

Intermediate

```
all human beings are born
free and equal in dignity
and rights they are endowed
with reason and conscience
and should act towards one
another in a spirit of
brotherhood everyone is
entitled to all the rights
and freedoms
```

Shannon Entropy



Shannon & Weaver (1949) The mathematical theory of communication

Shannon Entropy

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i) \tag{1}$$

- $-\log_2 p(x_i)$ is the **information content** of a unit x_i (e.g. word type).

Shannon Entropy

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i) \tag{1}$$

- $-\log_2 p(x_i)$ is the **information content** of a unit x_i (e.g. word type).
- For example:
  "human" in the UDHR: $-\log_2(\frac{13}{2000}) \sim 7.27$
  "the" in the UDHR: $-\log_2(\frac{121}{2000}) \sim 4.05$

Shannon Entropy

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i) \tag{2}$$

- The entropy is the **average information content** of information encoding units: $H(X) = [0, \infty[$

MINIMAL H(X)

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

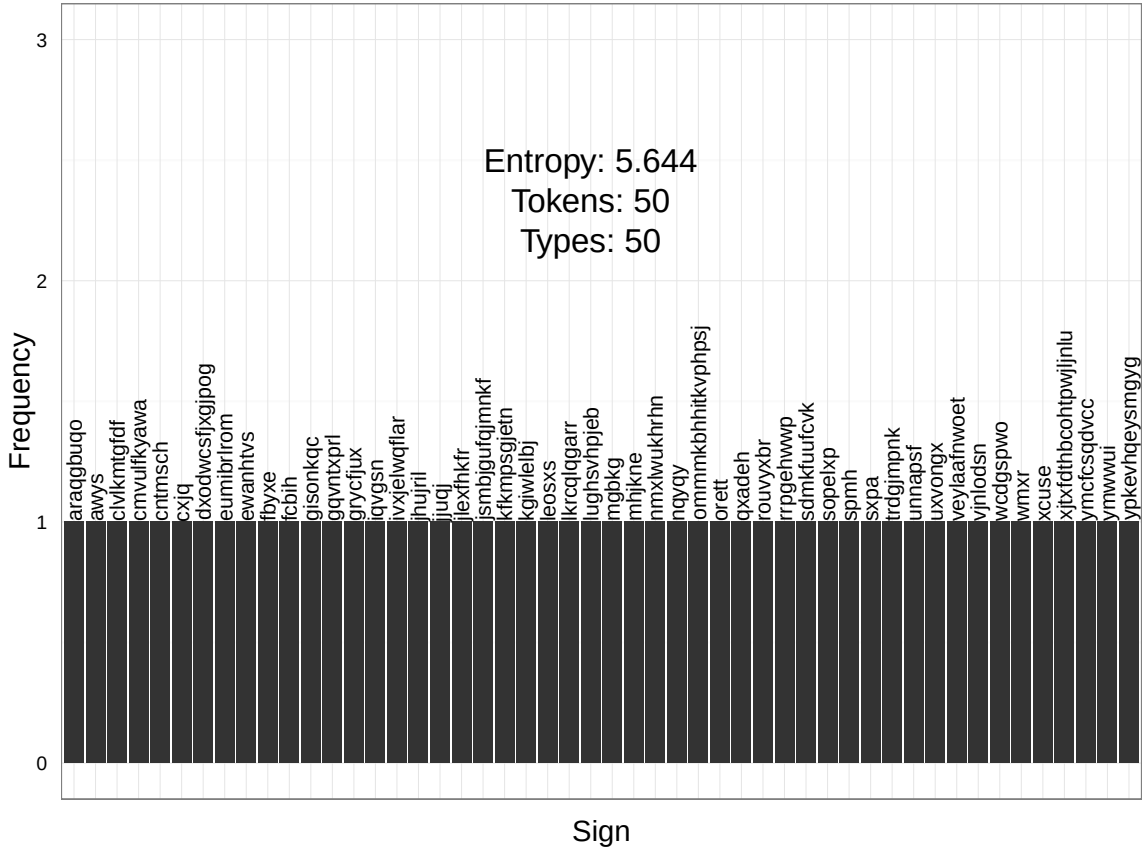
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

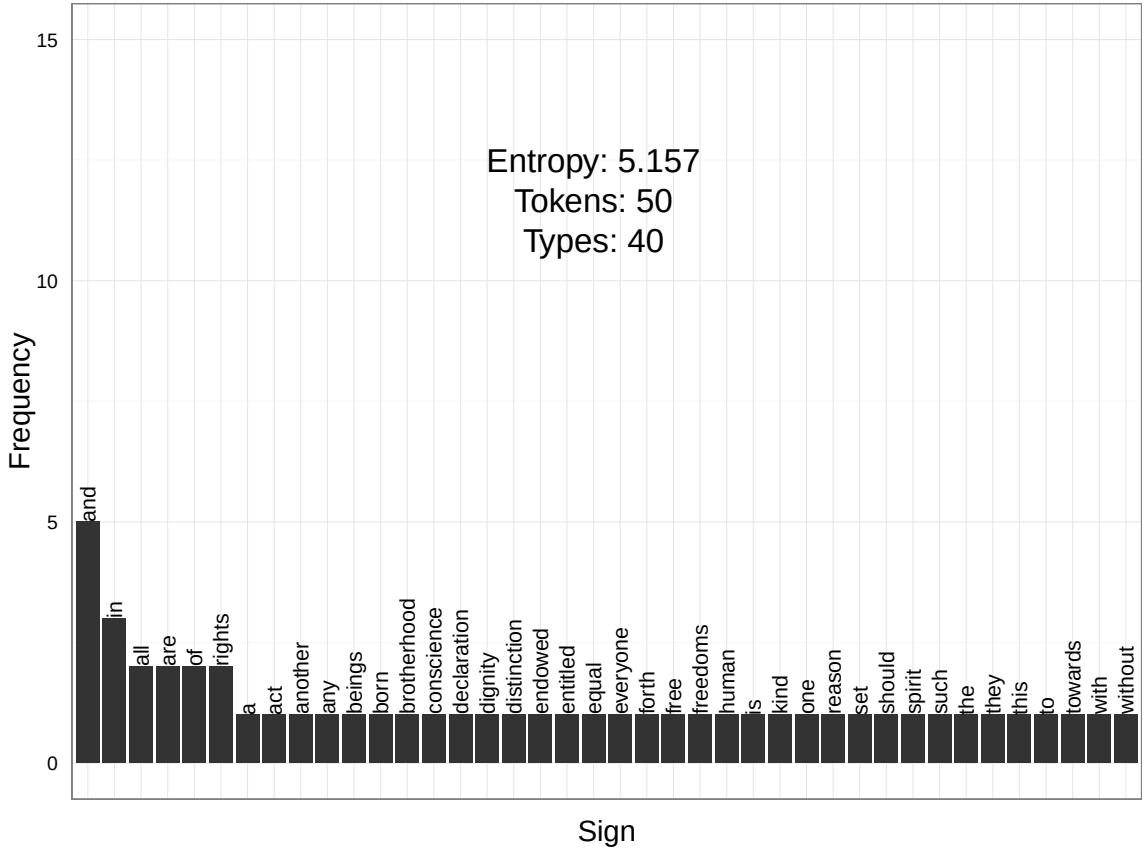
$$H(X) = 1 \times \log_2(1) = 0 \tag{3}$$

Random Text



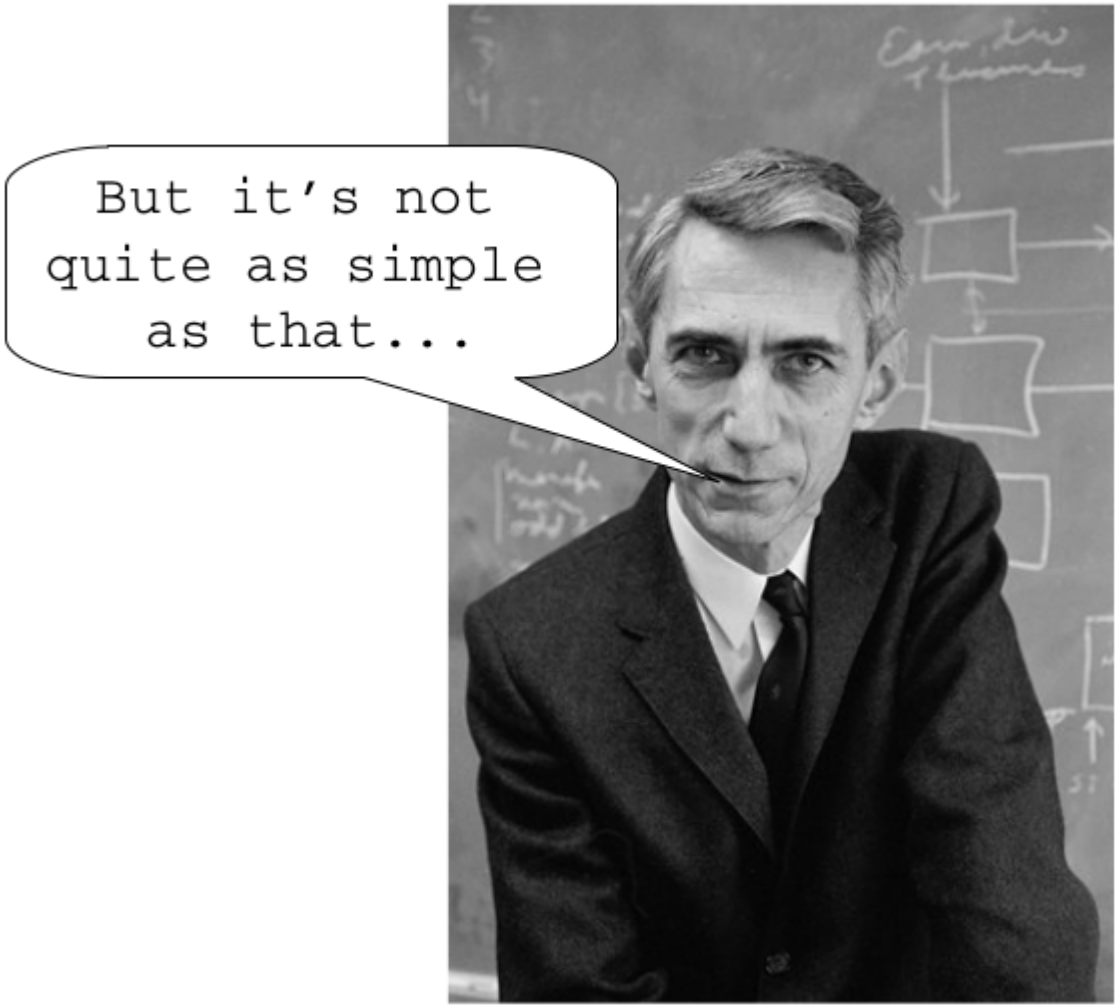
$$H(X) = -(\frac{1}{50}\log_2(\frac{1}{50}) + ... + \frac{1}{50}\log_2(\frac{1}{50})) = 5.644 \quad (4)$$

English UDHR



$$H(X) = -(\frac{5}{50}\log_2(\frac{5}{50}) + \frac{3}{50}\log_2(\frac{3}{50}) + ... + \frac{1}{50}\log_2(\frac{1}{50})) = 5.157 \quad (5)$$

Methodological Issues



Methodological Issues

- $H(X)$ depends on **text size** (number of tokens)

Methodological Issues

- $H(X)$ depends on **text size** (number of tokens)
- the **probability** of words is not a simple function of their **frequency**, they depent on **co-text**

Methodological Issues

- $H(X)$ depends on **text size** (number of tokens)
- the **probability** of words is not a simple function of their **frequency**, they depent on **co-text**
- what are **"words"** anyways?

Haspelmath (2011) The indeterminacy of word segmentation
Wray (2014) Why are we so sure we know what a word is?

Methodological Issues

Article

The entropy of words – Estimations across more than 1000 languages

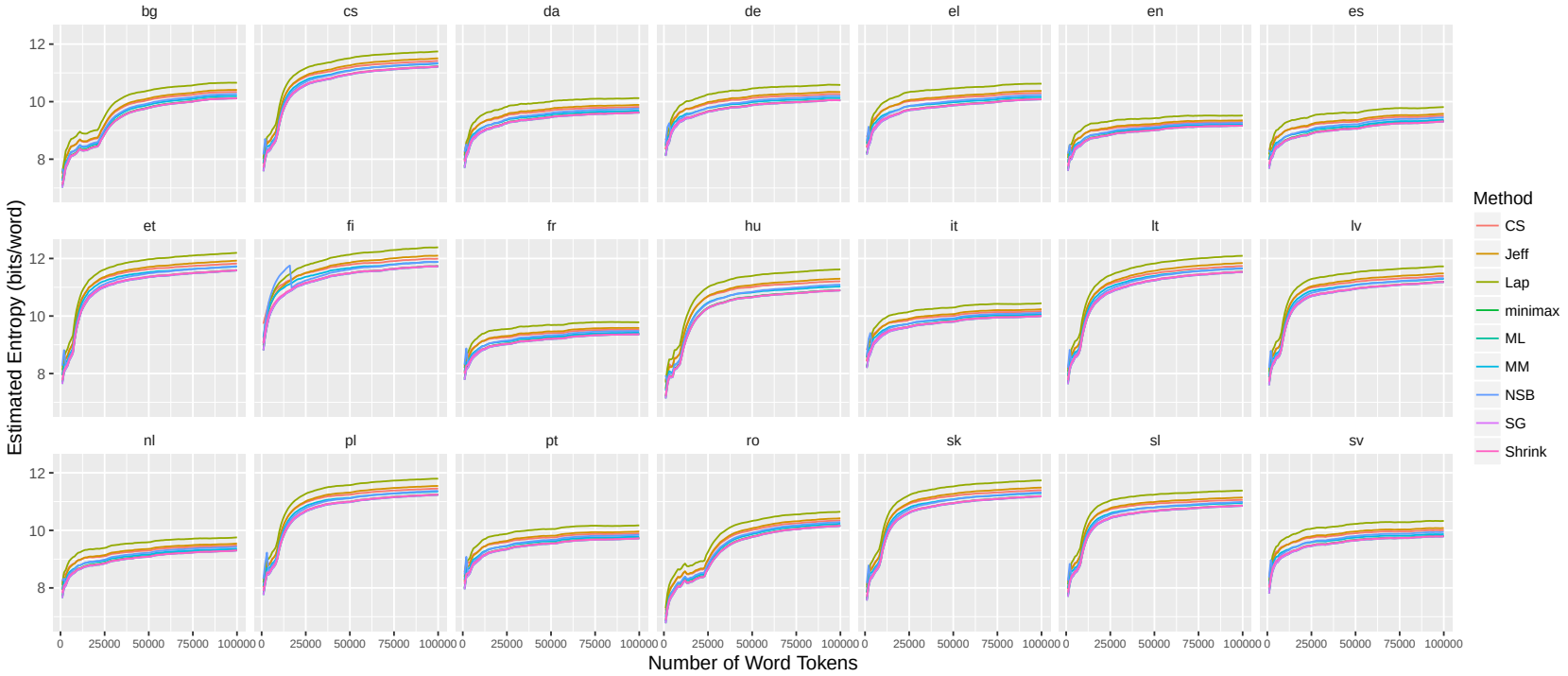
Christian Bentz [1,2]*, Dimitrios Alikaniotis [3], Michael Cysouw [4] and Ramon Ferrer-i-Cancho [5]

1 DFG Center for Advanced Studies, University of Tübingen, Rümelinstraße 23, D-72070 Tübingen, Germany; chris@christianbentz.de
2 Department of General Linguistics, University of Tübingen, Wilhemlstraße 19-23, D-72074 Tübingen, Germany
3 Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, United Kingdom
4 Forschungszentrum Deutscher Sprachatlas, Philipps-Universität Marburg, Deutschhausstrasse 3, 35037 Marburg
5 Complexity and Quantitative Linguistics Lab, LARCA Research Group, Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain
* Correspondence: chris@christianbentz.de

Academic Editor: name
Version January 9, 2017 submitted to Entropy; Typeset by LaTeX using class file mdpi.cls

Abstract: The uncertainty associated with words is a fundamental property of natural languages. It lies at the heart of quantitative linguistics, computational linguistics, and language sciences more generally. Information-theory gives us tools at hand to measure precisely this uncertainty – the word entropy. We here use three parallel corpora – encompassing ca. 450 million words in

Text Size Dependence



Bentz, Alikaniotis, Cysouw & Ferrer-i-Cancho (forthcoming)

Misconception

"(...) as many critics have since noted, and as Shannon was well aware, this model is not appropriate as a model of human communication, because "information" in Shannon's technical sense is not equivalent to "meaning" in *any* sense."

Fitch (2011) The Evolution of Language

The Mathematical Theory of Communication

Level A. How accurately can the symbols of communication be transmitted? (The technical problem.)

Level B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)

Level C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

Shannon & Weaver (1949), p.4

"The mathematical theory of the engineering aspect of communication (...) admittedly applies in the first instance only to problem A (...)"

However

"The mathematical theory of the engineering aspect of communication (...) admittedly applies in the first instance only to problem A (...)"

However

"(...) levels B and C, above, can make use only of those signal accuracies which turn out to be possible when analyzed at Level A. Thus any limitations discovered in the theory at Level A necessarily apply to levels B and C."

Shannon & Weaver (1949), p.6

"(...) the analysis at Level A discloses that this level overlaps the other levels more than one could possible naively suspect. Thus the theory of Level A is, at least to a significant degree, also a theory of levels B and C."

Shannon & Weaver (1949), p.6

Entropy is a **necessary** but **not sufficient** condition for communication

Applications to Typology

- What is the **information encoding potential** (entropy share) of different linguistic features?

Applications to Typology

- What is the **information encoding potential** (entropy share) of different linguistic features?
- How do **information encoding strategies** differ across languages?

Parallel Corpora

```
# 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
```

Parallel Corpora

```
# 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
```

Parallel Corpora

```
# language_name:        ภาสาขฺฐร
# 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 ឯលោកពេរេសបង្កើតលោកហេស្រុ
40001004        លោករ៉ាមបង្កើតលោកអ័មីណាដាប់ លោកអ័មីណាដាប់បង្កើតលោកណាសូន លោកណាសូនបង្កើតលោ
40001005        លោកសាលម៉ូន និងនាងរ៉ាហាប់បង្កើតលោកបូអូស ហើយលោកបូអូស និងនាងរស់បង្កើតលោកអូបិ
40001006        លោកអ៊ីសាយបង្កើតស្តេចដាវីឌ ស្តេចដាវីឌ និងប្រពន្ធលោកអ៊ូរីបង្កើតស្តេចសាឡូម៉ូន
40001007        ស្តេចសាឡូម៉ូនបង្កើតស្តេចរេហូបោម ស្តេចរេហូបោមបង្កើតស្តេចអ័ប៊ីយ៉ា ស្តេចអ័ប៊ីយ៉ាបង្កើតស្តេចអេស
40001008        ស្តេចអេសាបង្កើតស្តេចយ៉ូសាផាត ស្តេចយ៉ូសាផាតបង្កើតស្តេចយ៉ូរាម ស្តេចយ៉ូរামបង្កើតស្តេចអូសៀស
40001009        ស្តេចអូសៀសបង្កើតស្តេចយ៉ូថាម ស្តេចយ៉ូថាមបង្កើតស្តេចអេហាស ស្តេចអេហាសបង្កើតស្តេចអេសេគ
40001010        ស្តេចអេសេគាសបង្កើតស្តេចម៉ាណាសេ ស្តេចម៉ាណាសេបង្កើតស្តេចអាំម៉ូន ស្តេចអាំម៉ូនបង្កើតស្តេចយ៉ូ
```

Parallel Corpora

- *Parallel Bible Corpus* (PBC): ca. **1200** lang.
  Mayer & Cysouw (2014)
- *Universal Declaration of Human Rights* (UDHR): ca. **400** lang.
  www.unicode.org/udhr
- *Open Subtitles Corpus* (OSC): ca. **100** lang.
  Tiedemann (2012)
- *Europarl Parallel Corpus* (EPC): **21** lang.
  Koehn (2005)

Applications to Typology

JOURNAL OF QUANTITATIVE LINGUISTICS, 2016
http://dx.doi.org/10.1080/09296174.2016.1265792

Routledge
Taylor & Francis Group

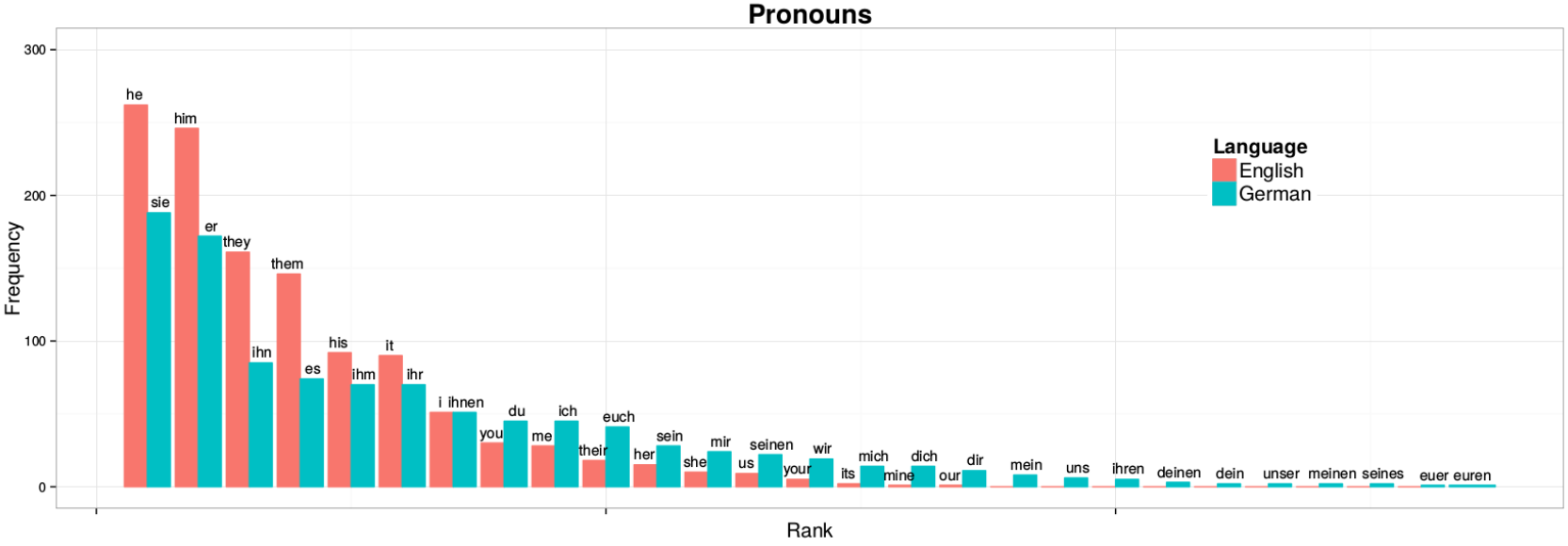
Variation in Word Frequency Distributions: Definitions, Measures and Implications for a Corpus-Based Language Typology

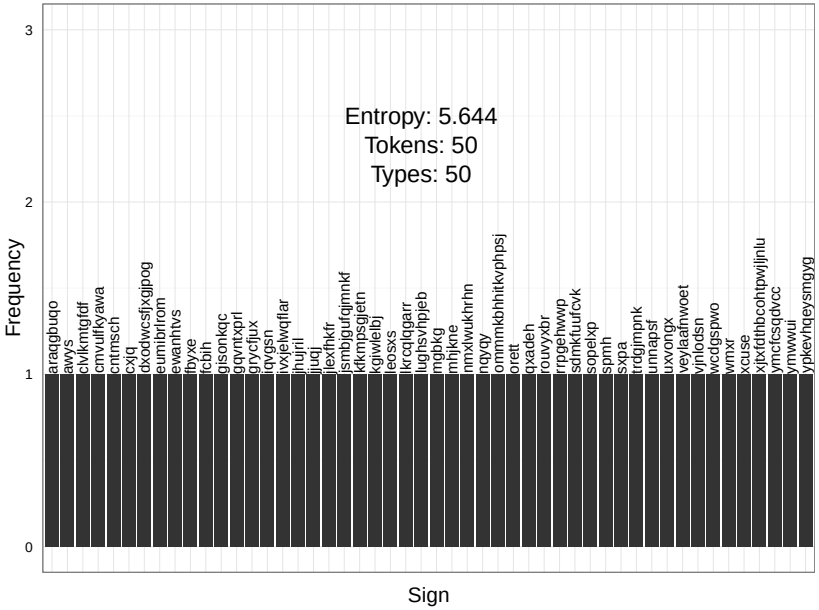
Christian Bentz[a,b], Dimitrios Alikaniotis[a], Tanja Samardžić[c] and Paula Buttery[a]

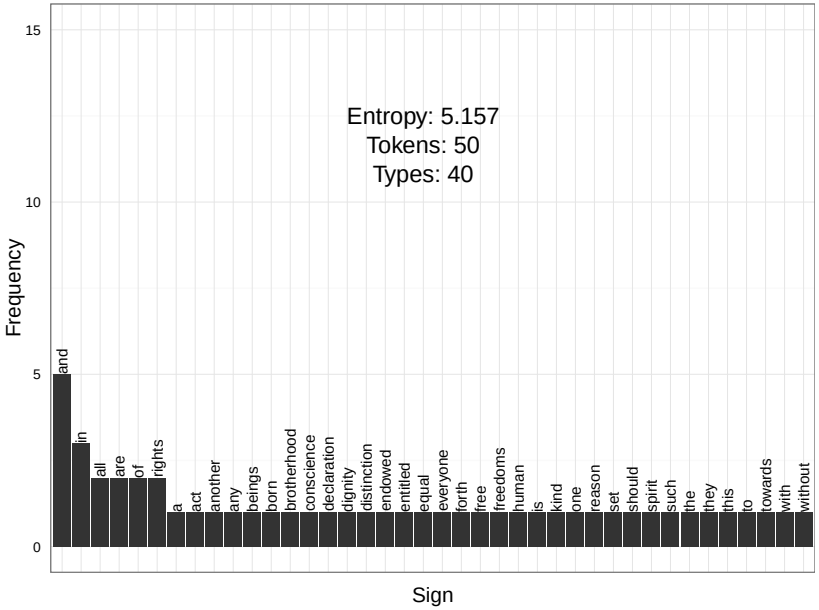
[a]Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, United Kingdom; [b]Department of General Linguistics, University of Tübingen, Tübingen, Germany; [c]URPP on Language and Space, University of Zürich, Zürich, Switzerland

ABSTRACT
Word frequencies are central to linguistic studies investigating processing difficulty, learnability, age of acquisition, diachronic transmission and the relative weight given to a concept in society. However, there are few cross-linguistic studies on entire distributions of word frequencies, and even less on systematic changes within them. Here, we first define and test an exact measure for the relative

Entropy Differences







Analysis 1

- Languages: English and German
- Corpora: OSC, Bible, UDHR
- **manually neutralize** inflections, derivations, compounds, clitics/contractions

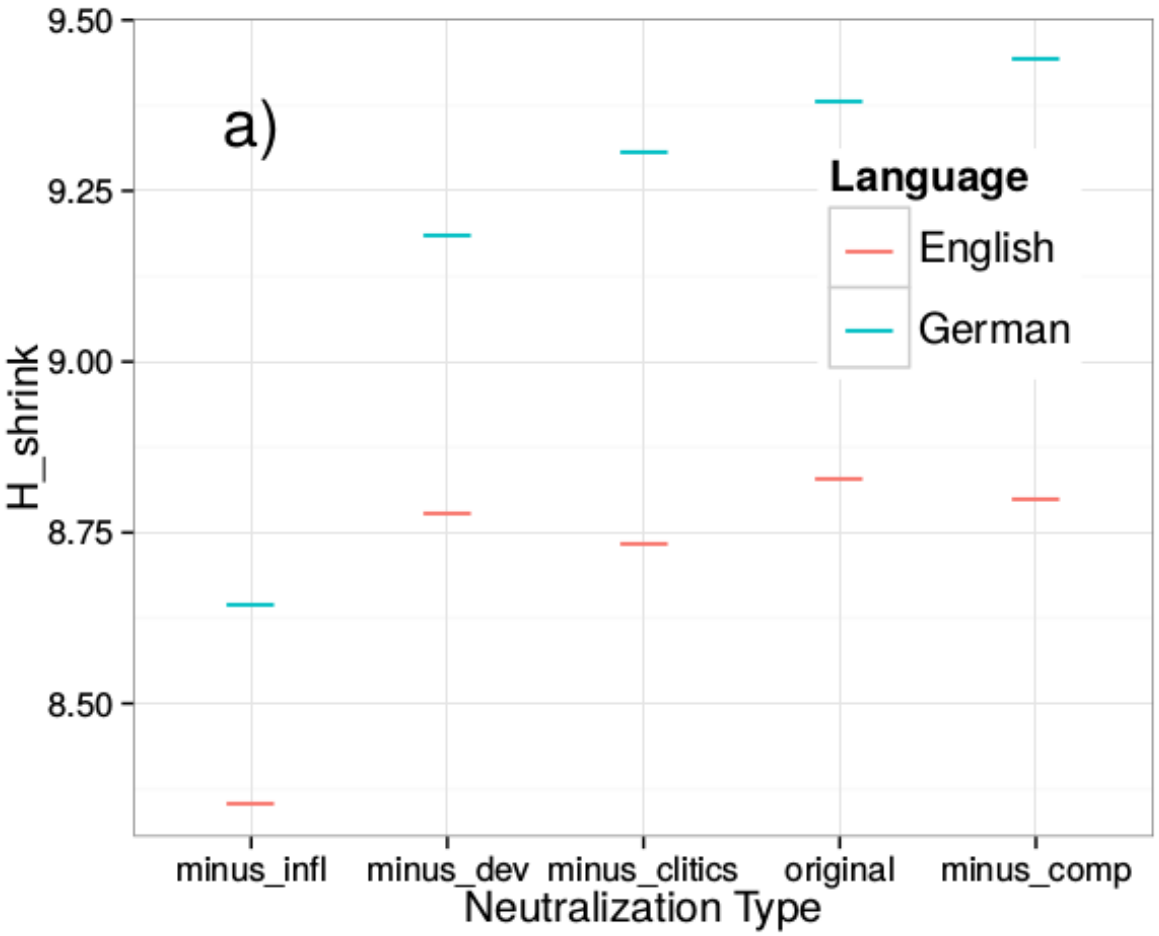
Analysis 1

Open Subtitles Corpus

Go back to Oslo and meet up with your boss, report what you've seen here, and ask him to send you back.
It's about Hellfjord.
There is someone in this room who has read too much Donald Duck.
It's quiet without Salmander.
Oh my god.
Hey, what are you doing?
What is this?
This is no fish eye.
Chop off an arm and a leg, and blame the sea serpent.

Fahr nach Oslo, triff dich mit deinem Chef und erstatte ihm Bericht, - - was du hier gesehen hast, und bitte ihn, dich zurückzuschicken.
Es geht um Hellfjord.
In diesem Raum befindet sich jemand, der zu viel Donald Duck gelesen hat.
Es ist so still hier ohne Salmander.
Oh mein Gott.
Hey, was machst du in meiner Küche?
Was ist das?
Das sind keine Fischaugen.
Hackt ihr einen Arm und ein Bein ab und beschuldigt das Seeungeheuer.

Analysis 1


a)
Language
English
German
9.50
9.25
9.00
8.75
8.50
H_shrink
minus_infl
minus_dev
minus_clitics
original
minus_comp
Neutralization Type

Analysis 1: Conclusions

- **Inflectional marking** has the biggest **entropy share** in both languages, i.e. ca. 0.8 bits/word (9%) in German, and ca. 0.5 bits/word (6%) in English

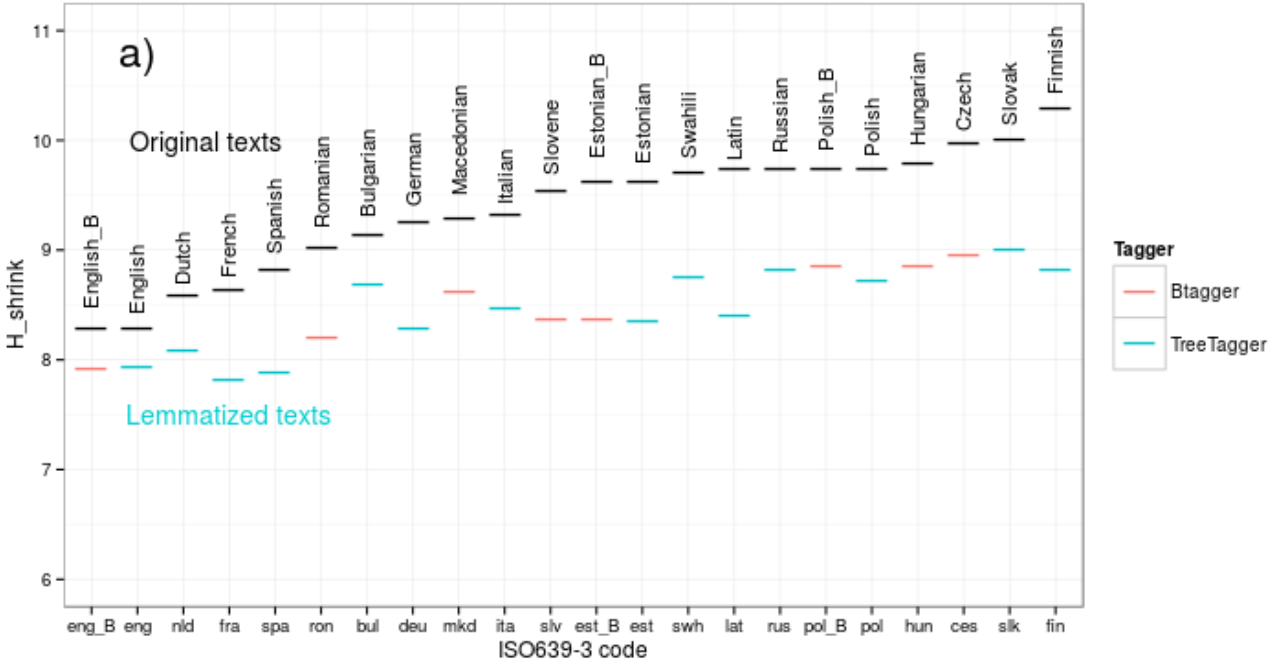
Analysis 1: Conclusions

- **Inflectional marking** has the biggest **entropy share** in both languages, i.e. ca. 0.8 bits/word (9%) in German, and ca. 0.5 bits/word (6%) in English
- ΔH between **English and German** is due to inflections (48% decrease), derivations (27% decrease), compounds (14% increase), contractions/clitics (2% increase)

Analysis 2

- Languages: 19
- Corpora: PBC, UDHR
- **automatically neutralize** inflections via lemmatization

Analysis 2


a)
Original texts
Lemmatized texts
H_shrink
11
10
9
8
7
6
English_B
English
Dutch
French
Spanish
Romanian
Bulgarian
German
Macedonian
Italian
Slovene
Estonian_B
Estonian
Swahili
Latin
Russian
Polish_B
Polish
Hungarian
Czech
Slovak
Finnish
eng_B eng nld fra spa ron bul deu mkd ita slv est_B est swh lat rus pol_B pol hun ces slk fin
ISO639-3 code
Tagger
Btagger
TreeTagger


b)
ΔH_shrink
2.0
1.5
1.0
0.5
0.0
English
English_B
Bulgarian
Dutch
Macedonian
French
Romanian
Italian
Polish_B
Russian
Hungarian
Spanish
Swahili
German
Polish
Slovak
Czech
Slovene
Estonian_B
Estonian
Latin
Finnish
eng eng_B bul nld mkd fra ron ita pol_B rus hun spa swh deu pol slk ces slv est_B est lat fin
ISO639-3 code
Tagger
Btagger
TreeTagger

Analysis 2: Conclusions

- The **entropy share** of inflections across 19 languages ranges from 0.4 bits/word (5%) in English to 1.5 bits/word (15%) in Finnish.

Analysis 2: Conclusions

- The **entropy share** of inflections across 19 languages ranges from 0.4 bits/word (5%) in English to 1.5 bits/word (15%) in Finnish.
- ΔH between **19 languages** (Indo-European, Uralic, Atlantic-Congo) is explained to **55% by inflectional marking** differences

Further Analyses: Tone marking?

Usila Chinantec (Otomanguean)

```
40001001        I⁴la³ ti²ton³ la⁴jang³⁴ sa¹jeun³ quian¹ Jesucristo a³lang⁴³ jon⁴³tyie¹ A³br
40001002        A³brang²³ lang⁴³ jmai³ I³sa²³ . Jian³ i²con²³ I³sa²³ a⁴hyon²³ Ja³co²³ . I²
40001003        I² i²con²³ Judá jian²³ A¹ta³mar²³ ra⁵sian³ Fares jian³ Zara . Jian³ i²con²³
40001004        I² i²con²³ Aram ja³⁴ Aminadab . I² Aminadab ja³⁴ Naasón i²con²³i³ . Jian³ i
40001005        Jian³ i²con²³ Salmón jian²³ A¹ra³hab²³ ra⁵sian³ Booz . Jian³ i²con²³ Booz j
40001006        Jian³ i²con²³ Isaí hain⁴ ra⁵sian³ re¹ Da³vei²³ . Jian³ i²con²³ re¹ Da³vei²³
40001007        Jian³ i²con²³ Salomón ra⁵sian³ Roboam . Jian³ i²con²³ Roboam ra⁵sian³ Abías
40001008        Jian³ i²con²³ Asa ra⁵sian³ Josafat . Jian³ i²con²³ Josafat ja³⁴ Joram . Jia
40001009        Jian³ i²con²³ Uzías ja³⁴ Jotam . Jian³ i²con²³ Jotam ja³⁴ Acaz . Jian³ i²co
40001010        Jian³ i²con²³ Ezequías ra⁵sian³ Manasés . Jian³ i²con²³ Manasés ra⁵sian³ A³
40001011        Jian³ i²con²³ Josías hain⁴ cuan³⁴ Jeconías jian²³ si³reunh¹ ma²a⁴han⁵ Israe
40001012        I² coh⁵ a⁴húan³i³ chion³² jeu³ Babilonia jon³ , jon³ ra⁵sian³ a³jon⁴³ Jecon
40001013        Jian³ i²con²³ Zorobabel ra⁵sian³ Abiud . Jian³ i²con²³ Abiud ja³⁴ Eliaquim
40001014        Jian³ i²con²³ Azor ja³⁴ Sadoc . Jian³ i²con²³ Sadoc ja³⁴ Aquim . Jian³ i²co
40001015        Jian³ i²con²³ Eliud ra⁵sian³ Eleazar . Jian³ i²con²³ Eleazar ra⁵sian³ Matán
40001016        I² i²con²³ Ja³co²³ hain⁴ ra⁵sian³ Se¹ , a³hain⁴ i³cúa³ Ma³rei² a³cuan³⁴ Jes
40001017        I² la⁴ne³ ja³⁴ quia⁵quin⁴ sa¹jeun³ liah⁴ma²sian³ A³brang²³ la⁴teg⁴ Da³vei²³
chion³² ta⁵ Babilonia ; conh⁴liah⁴ i²quia⁵quin⁴ liah⁴ma²tionh²i³ chion³² Babilonia la⁴teg⁴
40001018        La⁴la³ ra⁵sian³ Jesucristo : A⁴leg³⁴ re³ i⁴sanh⁴ Ma³rei² sie²³i³ jian²³ Se¹
```

Further Analyses: Word Order?

Word Order Typology through Multilingual Word Alignment

Robert Östling
Department of Linguistics
Stockholm University
SE-106 91 Stockholm, Sweden
robert@ling.su.se

Abstract

With massively parallel corpora of hundreds or thousands of translations of the same text, it is possible to automatically perform typological studies of language structure using very large language samples. We investigate the domain of word order using multilingual word alignment and high-precision annotation transfer in a corpus with 1144 translations in 986 languages of the New Testament. Results are jection of part-of-speech (PoS) tags and dependency parse trees to investigate five different word order properties in 986 different languages, through a corpus of New Testament translations. The results are validated through comparison to relevant chapters in the World Atlas on Language Structures, WALS (Dryer and Haspelmath, 2013), and we find a very high level of agreement between this database and our method.

We identify two primary applications of this method. First, it provides a new tool for basic research in linguistic typology. Second, it has been

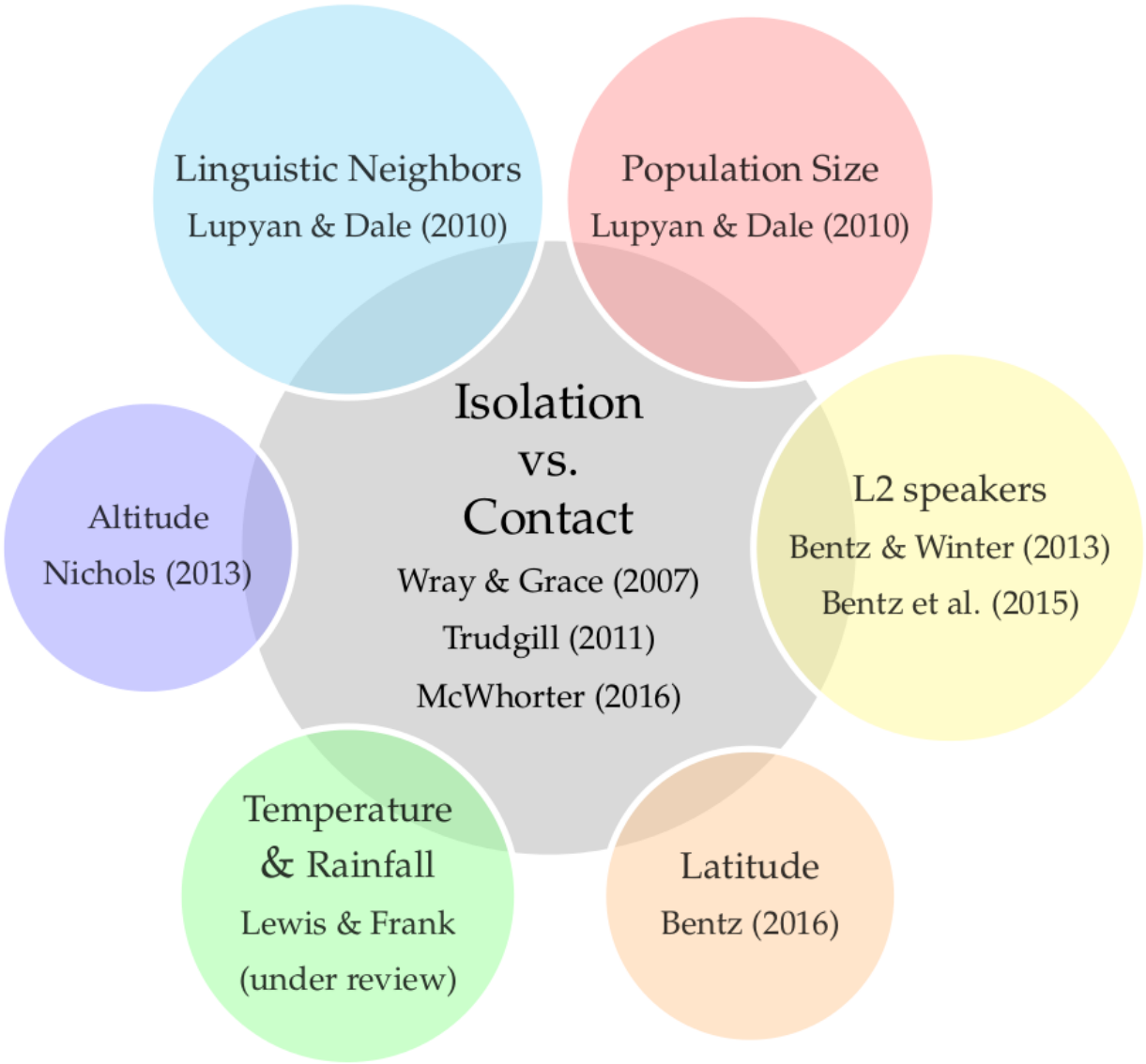
Applications to Typology: Conclusions

- The **information encoding potential** of *any* **linguistic feature** can be measured, if there is a systematic way of manipulating it in texts

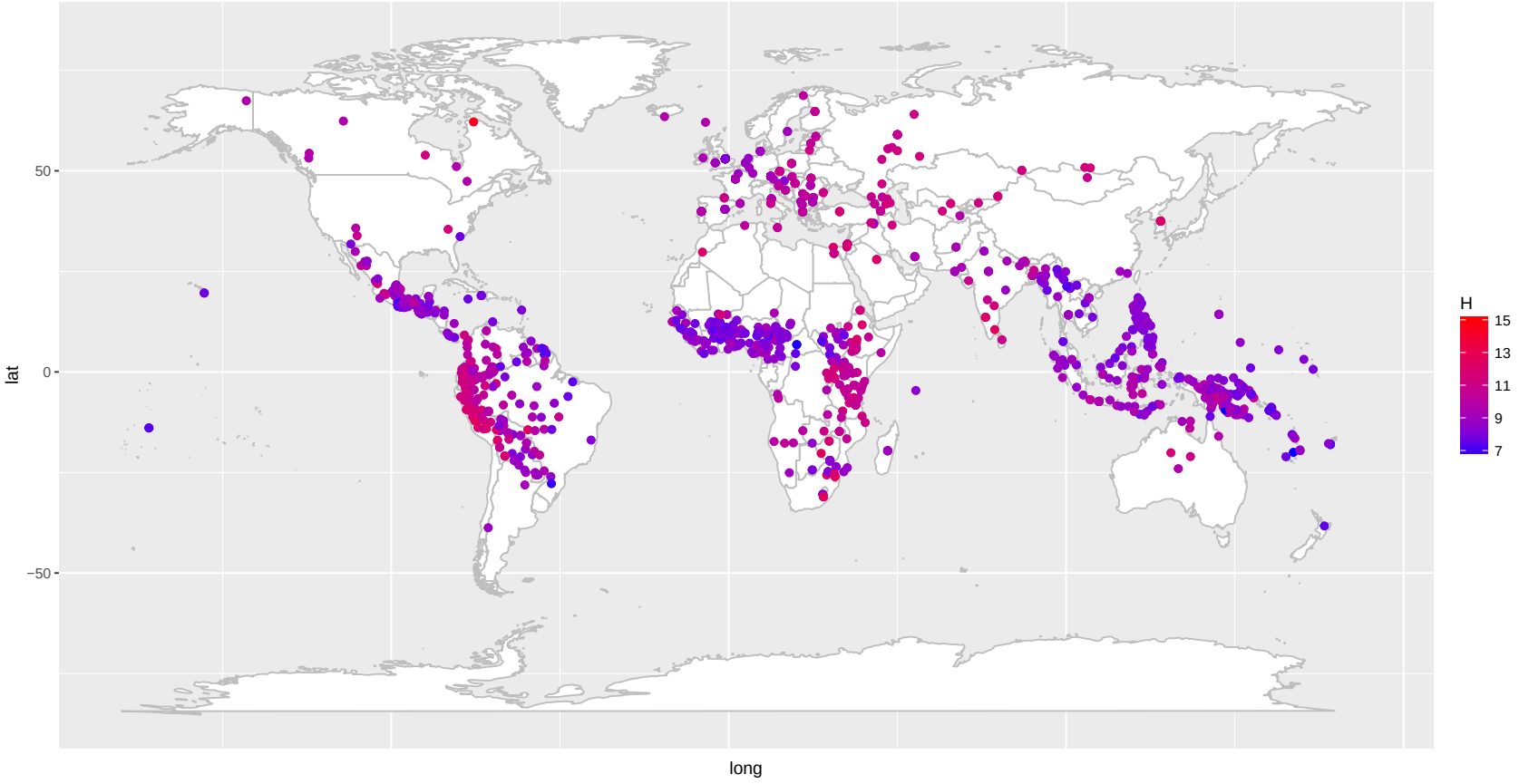
Applications to Typology: Conclusions

- The **information encoding potential** of *any* **linguistic feature** can be measured, if there is a systematic way of manipulating it in texts
- The analyses do **not have to depend** on **word types**, they could also use characters, morphemes, constructions, etc.

Explanations of Diversity



Word Entropy across 1092 languages



Bentz, Alikaniotis, Cysouw & Ferrer-i-Cancho (forthcoming)

Non-native language learning

PLOS | ONE

RESEARCH ARTICLE

Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms

Christian Bentz[1]*, Annemarie Verkerk[2], Douwe Kiela[3], Felix Hill[3], Paula Buttery[1,3]

1 Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, United Kingdom, **2** Reading Evolutionary Biology Group, School of Biological Sciences, University of Reading, Reading, United Kingdom, **3** Computer Laboratory, University of Cambridge, Cambridge, United Kingdom

* chris@christianbentz.de

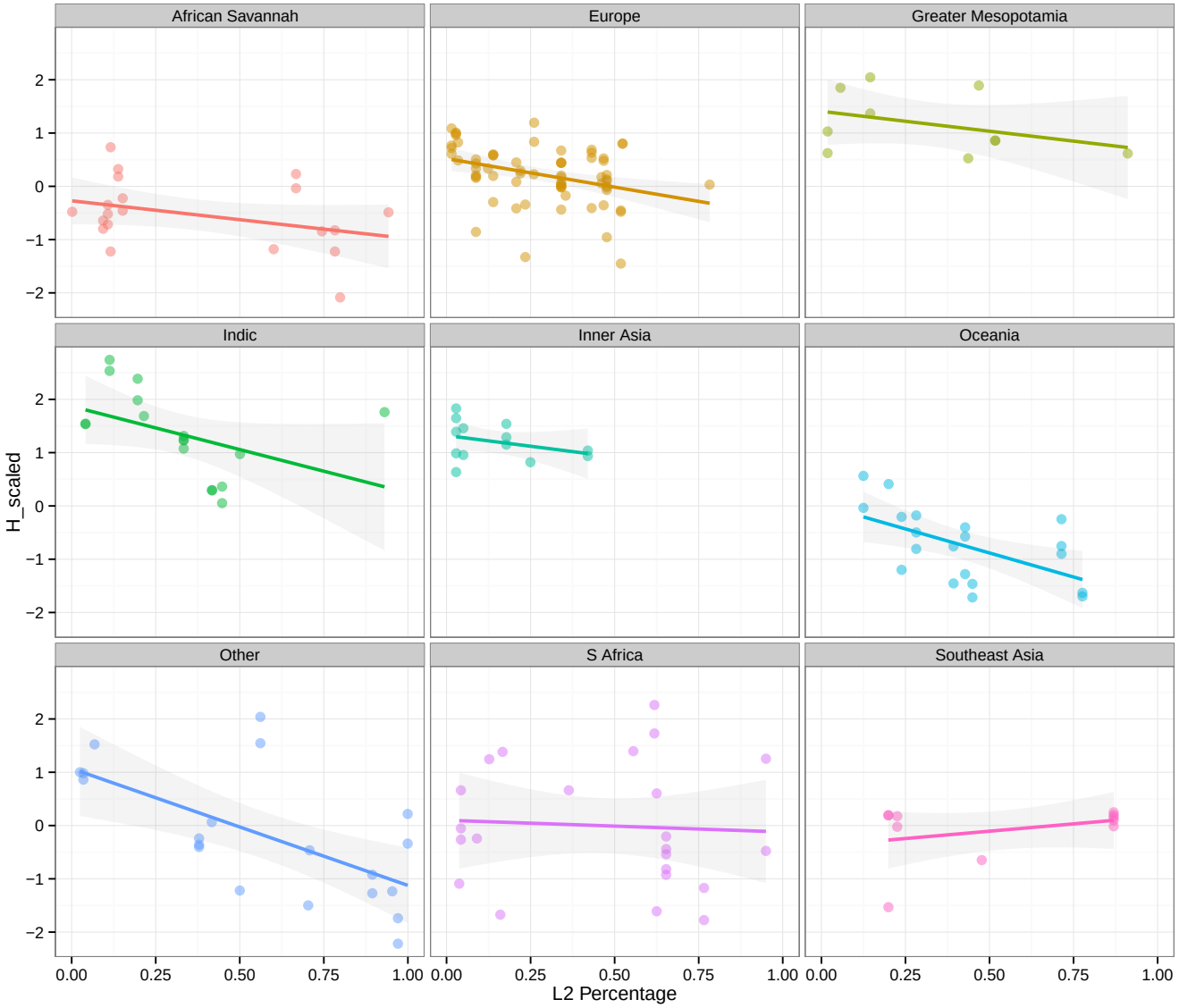
OPEN ACCESS

Citation: Bentz C, Verkerk A, Kiela D, Hill F, Buttery P (2015) Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer

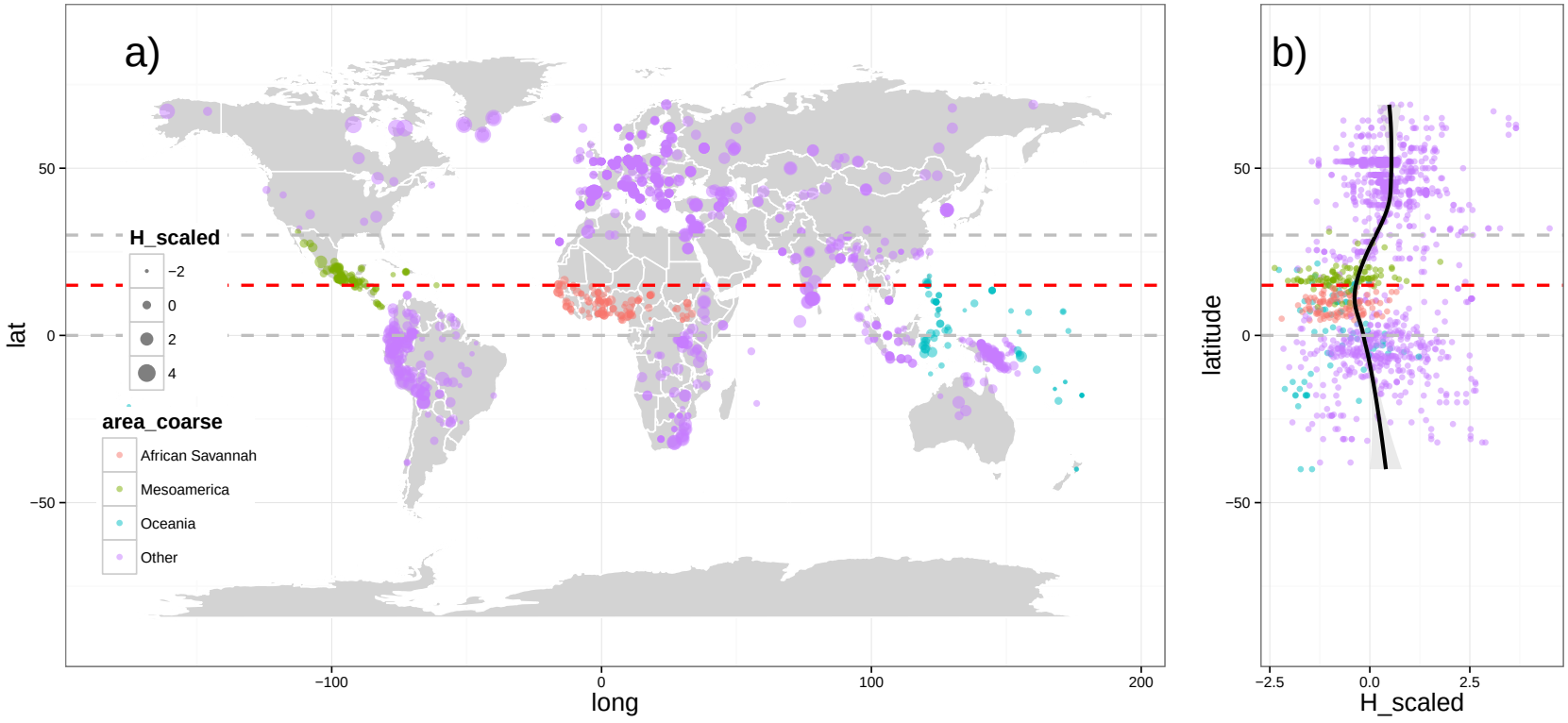
Abstract

Explaining the diversity of languages across the world is one of the central aims of typological, historical, and evolutionary linguistics. We consider the effect of *language contact*-the number of non-native speakers a language has-on the way languages change and evolve. By analysing hundreds of languages within and across language families, regions, and text types, we show that languages with greater levels of contact typically employ fewer word

Non-native language learning

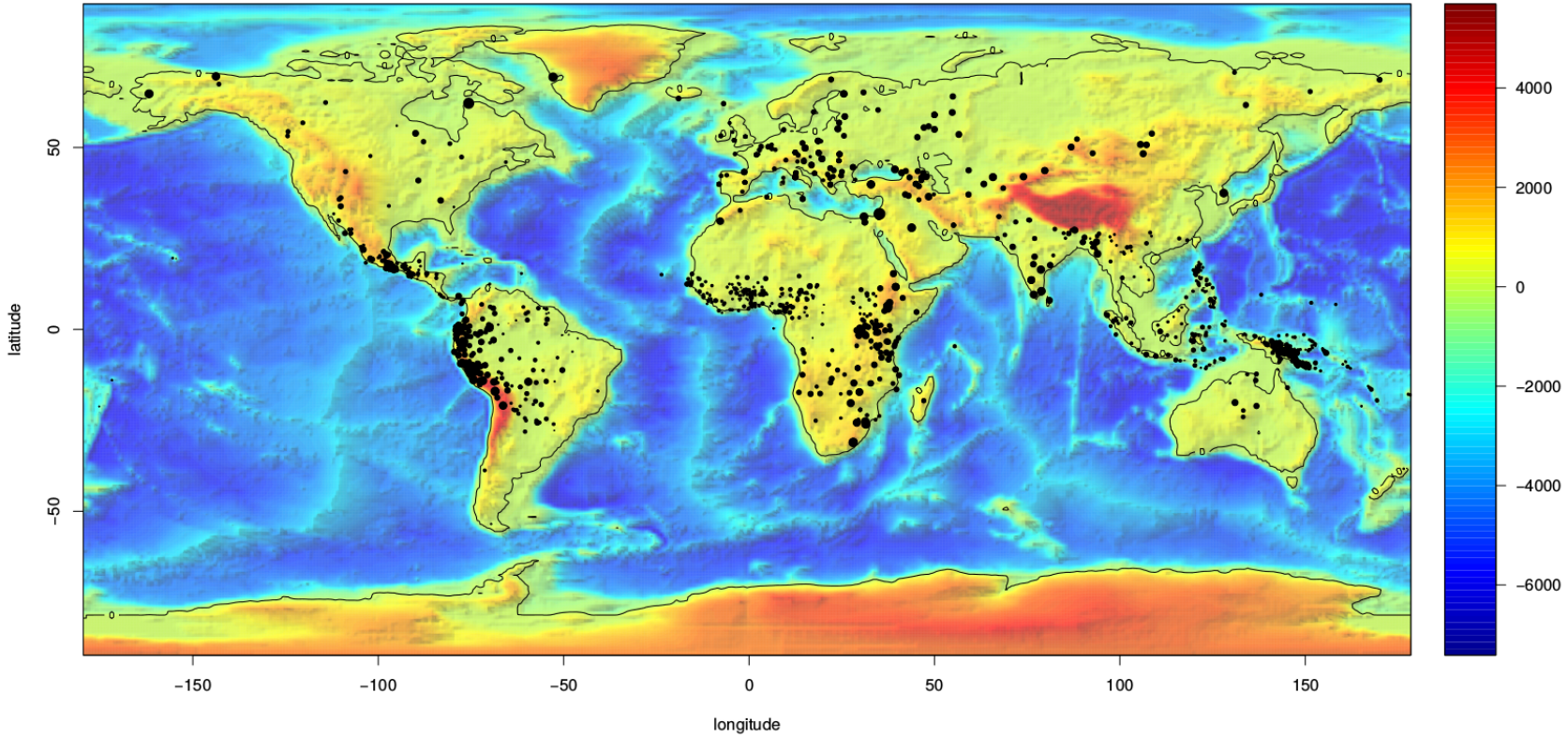


Latitude



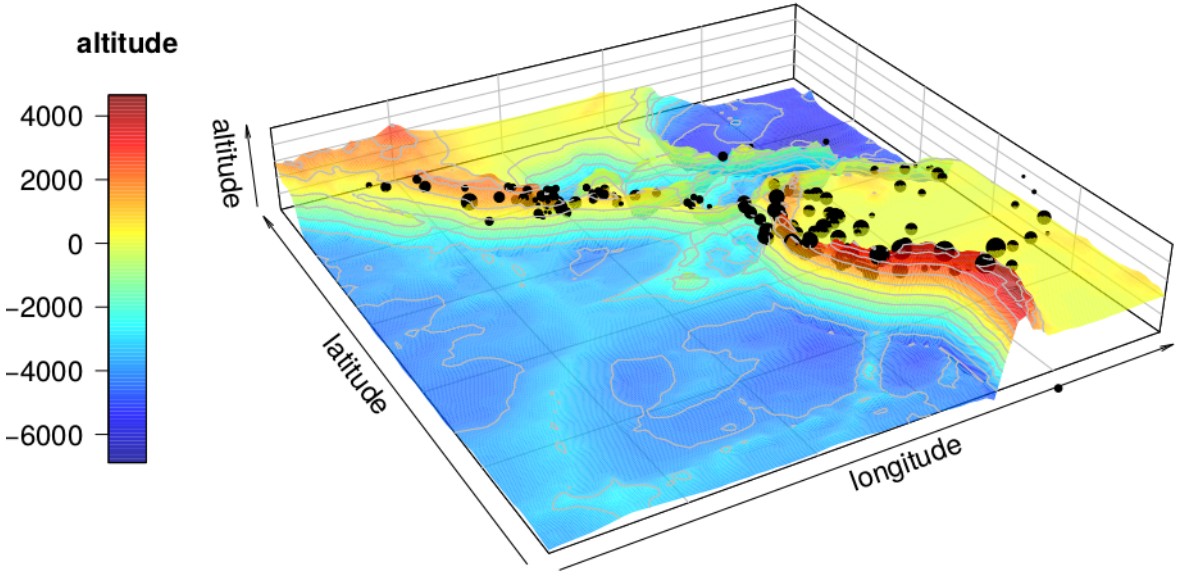
Bentz (2016) The Low-Complexity-Belt

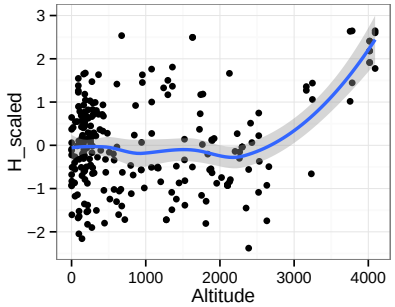
Altitude

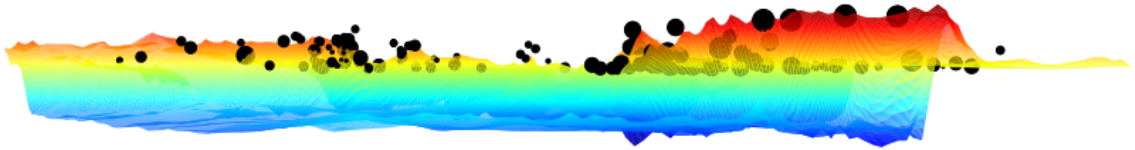

latitude
50
0
−50
longitude
−150
−100
−50
0
50
100
150
4000
2000
0
−2000
−4000
−6000

Mesoamerica and the Andes

183 languages, 90 families, 3 areas

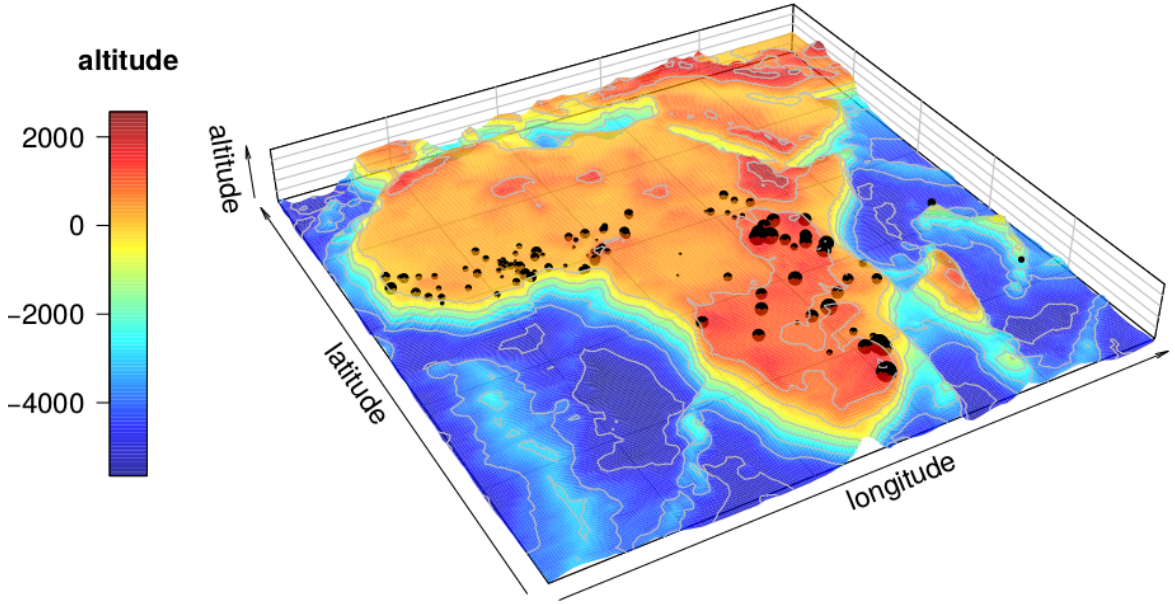


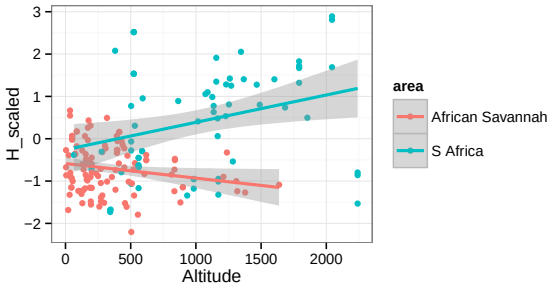


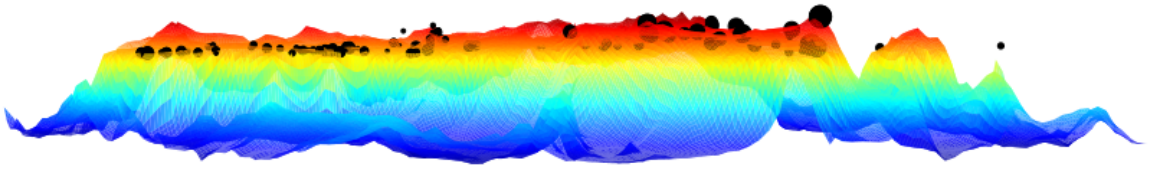


African Savannah and South Africa

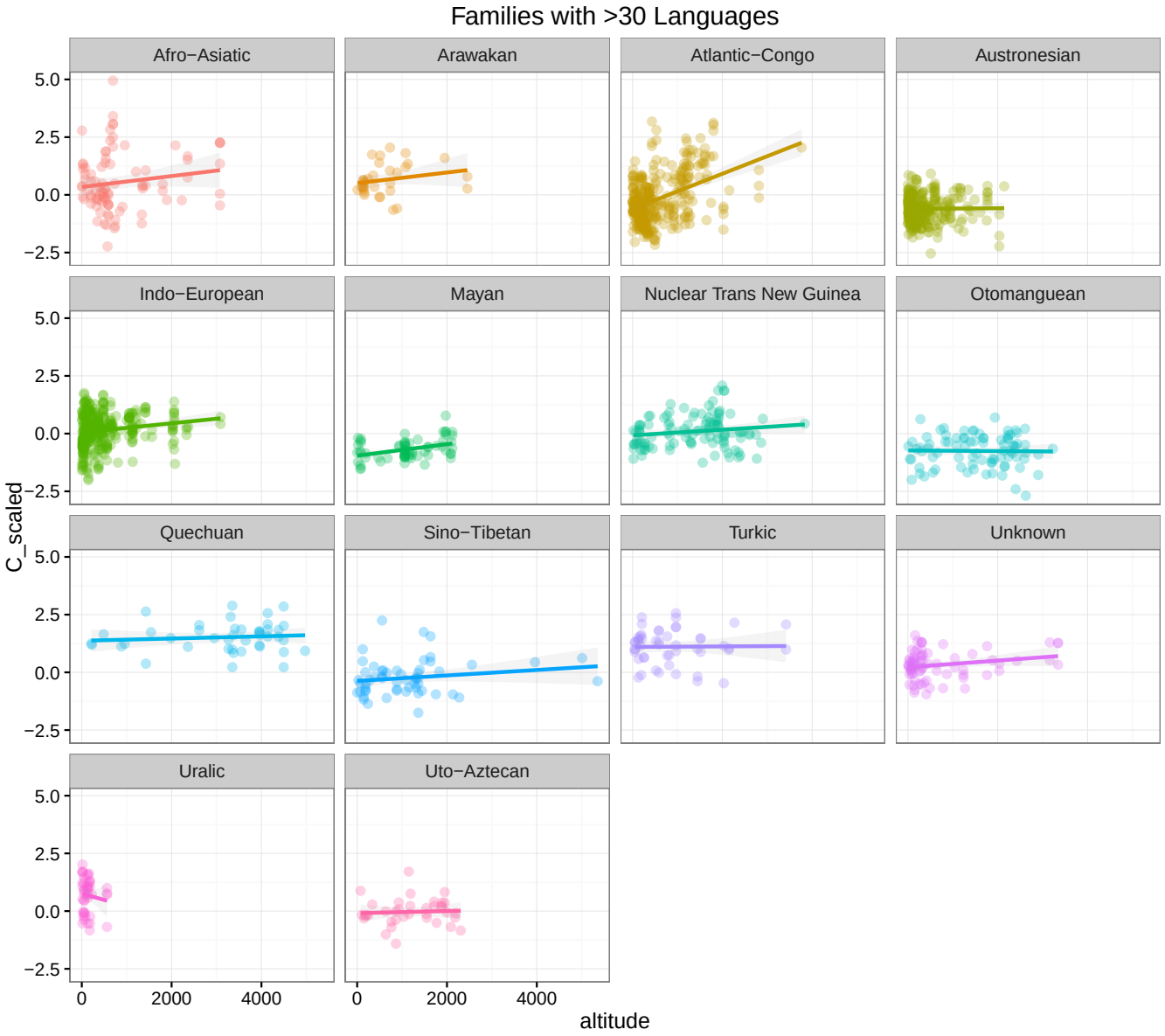
127 languages, 21 families, 2 areas





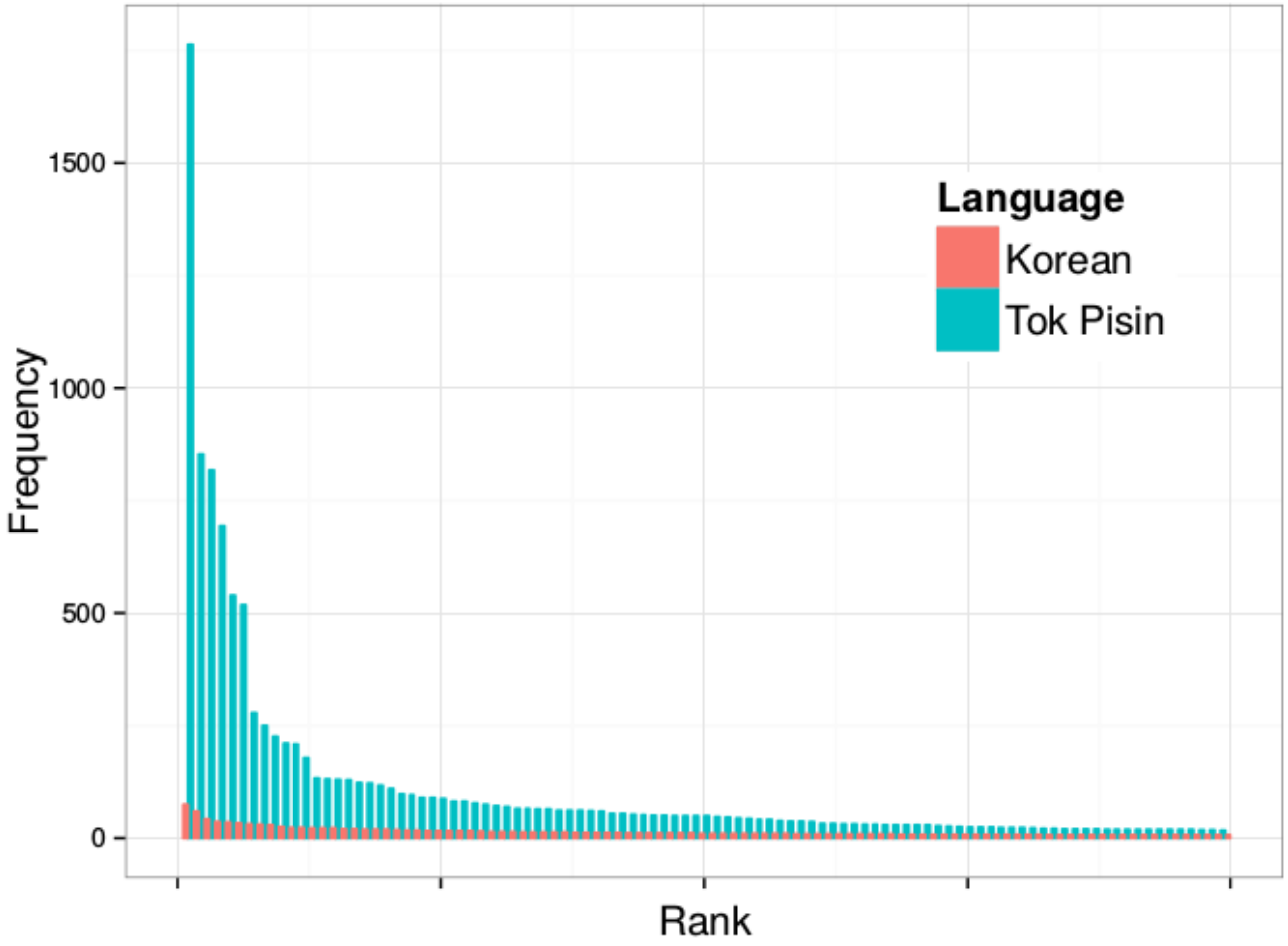


Altitude


Families with >30 Languages
Afro−Asiatic
Arawakan
Atlantic−Congo
Austronesian
Indo−European
Mayan
Nuclear Trans New Guinea
Otomanguean
Quechuan
Sino−Tibetan
Turkic
Unknown
Uralic
Uto−Aztecan
C_scaled
5.0
2.5
0.0
−2.5
0
2000
4000
altitude

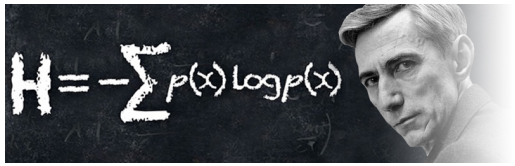
Disclaimer

What's better?



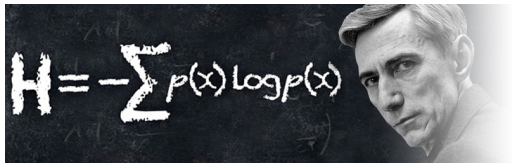
Conclusions

- Information theory *is* relevant to natural languages

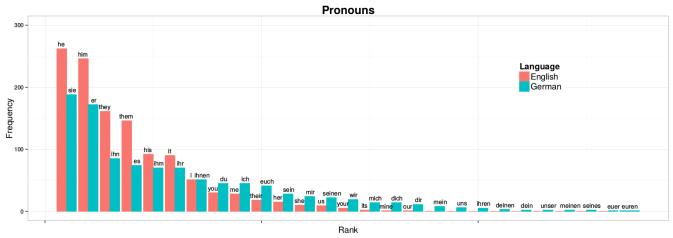


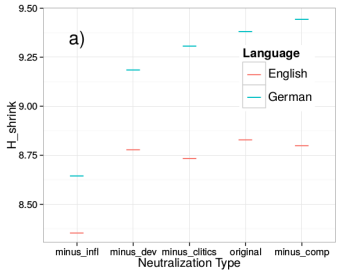
Conclusions

- Information theory *is* relevant to natural languages



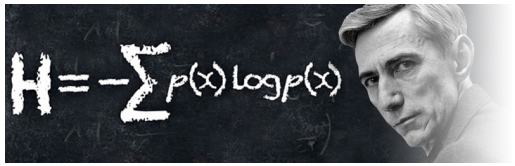
- We can determine the *entropy share* of linguistic features



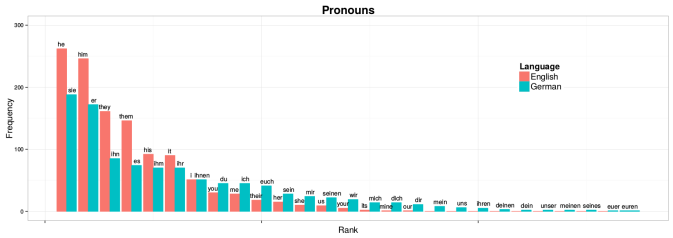


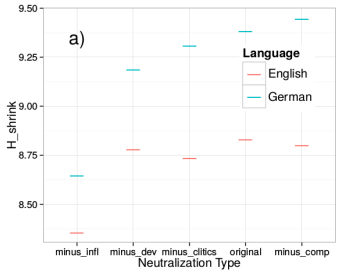
Conclusions

- Information theory *is* relevant to natural languages

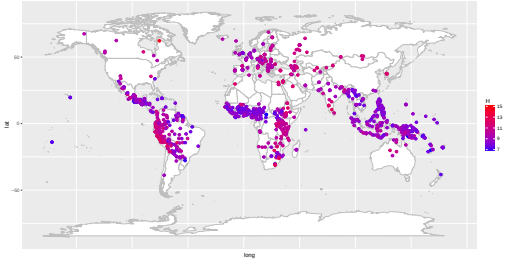


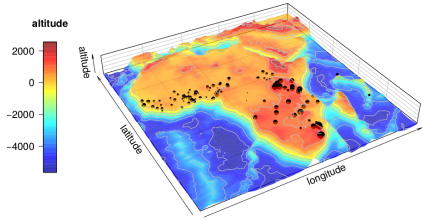
- We can determine the *entropy share* of linguistic features

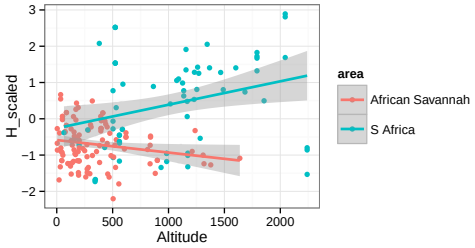




- We start to understand *entropy diversity*







Thank you