

The Distribution of Information-Theoretic Complexity across Languages of The World

Evidence for Large-Scale Pre-Historic Contact?

Christian Bentz
University of Tübingen

December 8, 2015



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

EVOLAEMP
LANGUAGE EVOLUTION: THE EMPIRICAL TURN

OVERVIEW

LANGUAGE COMPLEXITY

Data and Methods

Measuring Complexity

GEOGRAPHICAL DISTRIBUTION

Latitude

Altitude

Statistical Models

Discussion

Pre-Historic Contact?

DATA AND METHODS

How to measure the **complexity of languages**?

- ▶ Phonetic complexity
- ▶ Morphological complexity
- ▶ Syntactic complexity
- ▶ Learning and usage complexity
- ▶ **Information-theoretic complexity**

[Miestamo & Sinnemäki, 2008; Sampson, Trudgill & Gil, 2009;
Newmeyer & Preston, 2014]

DATA AND METHODS

Parallel Corpora [Mayer & Cysouw, 2014]

```
# language_name: English
# closest ISO 639-3: eng
# year_short: 2011
# year_long: Not available
# title: The Books of the Bible<br>The New Testament in English
# URL: https://www.bible.com/bible/31/lka.1.books
# copyright_short: © Biblica Inc.® 2011
# copyright_long: The Books of the Bible<br><br>Copyright © 2011 by Biblica, Inc.® Used by permission. All rights reserved worldwide.<br>The Books of the Bible™ is copyrighted and has been made available on the internet for your personal use only. Any other use including, but not limited to, copying or reposting on the internet is prohibited. These Scriptures may not be altered or modified in any form and must remain in their original context. These Scriptures may not be sold or otherwise offered for sale. These Scriptures are not shareware and may not be duplicated. These scriptures are not public domain.
40001081 This is the genealogy of Jesus the Messiah the son of David , the son of Abraham :
40001082 Abraham was the father of Isaac , Isaac the father of Jacob , Jacob the father of Judah and his brothers ,
40001083 Judah the father of Perez and Zerah , whose mother was Tamar , Perez the father of Hezron , Hezron the father of Ram ,
40001084 Ran the father of Amminadab , Amminadab the father of Nahshon , Nahshon the father of Salmon ,
40001085 Salmon the father of Boaz , whose mother was Rahab , Boaz the father of Obed , whose mother was Ruth , Obed the father of Jesse ,
40001086 and Jesse the father of King David . David was the father of Solomon , whose mother had been Uriah's wife ,
40001087 Solomon the father of Rehoboam , Rehoboam the father of Abijah , Abijah the father of Asa ,
40001088 Asa the father of Jehoshaphat , Jehoshaphat the father of Jehoram , Jehoram the father of Uzziah ,
40001089 Uzziah the father of Jotham , Jotham the father of Ahaz , Ahaz the father of Hezekiah ,
40001090 Hezekiah the father of Manasseh , Manasseh the father of Anan , Anan the father of Josiah ,
40001091 and Josiah the father of Jeconiah and his brothers at the time of the exile to Babylon .
40001092 After the exile to Babylon : Jeconiah was the father of Shealtiel , Shealtiel the father of Zerubbabel ,
40001093 Zerubbabel the father of Abihud , Abihud the father of Elakim , Elakim the father of Azor ,
40001094 Azor the Father of Zadok , Zadok the father of Akim , Akim the father of Elihud ,
40001095 Elihud the Father of Eleazar , Eleazar the father of Matthan , Matthan the father of Jacob ,
40001096 and Jacob the father of Joseph , the husband of Mary , and Mary was the mother of Jesus Who is called the Messiah .
40001097 Thus there were fourteen generations in all from Abraham to David , fourteen from David to the exile to Babylon , and fourteen from the exile to the
Messiah .
40001098 This is how the birth of Jesus the Messiah came about : His mother Mary was pledged to be married to Joseph , but before they came together , she was
found to be pregnant through the Holy Spirit .
40001099 Because Joseph her husband was faithful to the law , and yet did not want to expose her to public disgrace , he had in mind to divorce her quietly .
40001100 But after he had considered this , an angel of the Lord appeared to him in a dream and said , " Joseph son of David , do not be afraid to take Mary home
as your wife , because what is conceived in her is from the Holy Spirit .
40001101 She will give birth to a son , and you are to give him the name Jesus , because he will save his people from their sins . "
40001102 All this took place to fulfill what the Lord had said through the prophet :
40001103 " The virgin will conceive and give birth to a son , and they will call him Immanuel " ( which means " God with us " ) .
40001104 When Joseph woke up , he did what the angel of the Lord had commanded him and took Mary home as his wife .
40001105 But he did not consummate their marriage until she gave birth to a son . And he gave him the name Jesus .
40002001 After Jesus was born in Bethlehem in Judea , during the time of King Herod , Magi from the east came to Jerusalem
and asked , " Where is the one who has been born king of the Jews ? We saw his star when it rose and have come to worship him . "
40002002 When King Herod heard this he was disturbed , and all Jerusalem with him .
40002003 When he had called together all the people's chief priests and teachers of the law , he asked them where the Messiah was to be born .
40002004 " In Bethlehem in Judea " Ezra 9:22 said - For this is where the Messiah has written .
```

Plain Text • Tab Width: 8 • Ln 1, Col 1 INS

DATA AND METHODS

Parallel Corpora [Mayer & Cysouw, 2014]

```
# language_name: Hawaiian
# closest ISO 639-3: haw
# year_short: 1994
# year_long: Not available
# title: The New Testament in Hawaiian
# author: Not available
# copyright_short: Not available
# copyright_long: Not available
40001001  O ke kauhanau na ka hananua o Iesu Kristo , ka mamo a Davida , ka nane a Aberahama .
40001002  Na Aberahama o Isaaka ; na Isaaka o Iakoba ; na Iakoba o Iuda a ne kona poe hoahanau ;
40001003  Na Iuda laua me Tanara o Paresa a ne Zara ; na Paresa o Hezerona ; na Hezerona o Arana ;
40001004  Na Arana o Amlnabeda ; na Aminadoba o Naasona ; na Naasona o Salemona ;
40001005  Na Salemona laua me Rehaba o Boaza ; na Boaza laua me Ruta o Obeda ; na Obeda o Iese ;
40001006  Na Iese o Davida ke alli ; na Davida na ke alii laua me ka wahine a Uriia o Solomonia ;
40001007  Na Solomonia o Rehoboma ; na Rehoboma o Abia ; na Abia o Asa ;
40001008  Na Asa o Iosepata ; na Iosepata o Iorana ; na Iorana o Ozia ;
40001009  Na Ozia o Totama ; na Totama o Ahaza ; na Ahaza o Hezekia ;
40001010  Na Hezekia o Manase ; na Manase o Anona ; na Anona o Iosia ;
40001011  Na Iosia o Tekonika a me kona poe hoahanau .. i ka nanaawa o ka lawe ana i Babulona :
40001012  A mahope nai o ka lawe ana i Babulona , na Iekonia o Saletiela ; na Saletiela o Zerubabelia ;
40001013  Na Zerubabelia o Abtuda ; na Abtuda o Eliakima ; na Eliakima o Azora ;
40001014  Na Azora o Sadoka ; na Sadoka o Akima ; na Akina o Eliuda ;
40001015  Na Elluda o Eleazar ; na Eleazar o Mahetana ; na Mahetana o Iakoba ;
40001016  Na Iakoba o Iosepae ke kane a Maria nana l hanau o Iesu , i kapala o Kristo .
40001017  O na hananua a pau nai ia Aberahama nai a hiki ia Davida , he umi ia hananua a me kumanaha ; a mai ia Davida mai a hiki i ka lawe ana i Babulona , he unti la hananua a me kumanaha ; a mat ka lawe ana aku i Babulona mat , a hiki la Kristo he unti la hananua a me kumanaha .
unl la hananua a me kumanaha ; a mat ka lawe ana aku i Babulona mat , a hiki la Kristo he unti la hananua a me kumanaha .
40001018  Penet hot ka hanau ana o Iesu Kristo : I hoopalau e i kona nakuwahine o Maria na Iosepa , aole nae lai i pilii , a ike oia , ua hapai na ka Uhane Hemolele .
40001019  Aka , he kanaka pono kana kane o Iosepa , aole la i makenake e holino la ia na ke akea ; manao iho la ia e kipaku malu la ia .
40001020  A i kona nooono ana na ia mau nea , aia hol , iketak'ku ka anela a ka Haku e ia ma ka noe , i mai la , E Iosepa , e ka namo a Davida , mai makau oe ke lawe la Maria i wahine nau ; no ka nea , ua hapai ola na ka Uhane Hemolele .
40001021  A e hanau nai ota i kekiktane , a e kapa aku oe i kona inoa o IESU ; no ka nea , e hooala ia i kona poe kanaka mat ko lakou hewa .
40001022  Ma ia mau nea i ko ai a ka mea a ka Haku i olelo mat ai me ka kaula , i ka i ana mai .
40001023  Ala hot , e hapat ana kekahil wahine puupaa , a e hanau nai la i kekiktane , a e kapila kona inoa o Emanuela ; o ke ano kela , o ke Akua me kakou .
40001024  A ala ae la o Iosepa mat ka hianoe ana , hana aki la ia e like me ka anela a ka Haku i kaunoa mai la ia , a lawe mai la ia i kana wahine ;
40001025  Aole nae i noe aki la , a hiki i ka w hanau at ota i kana makahilo kane , a e kapila kona inoa o IESU .
40002001  A hananua o Iesu ma Betelhehena i Iudea , i ke kau la Herode ke alli , aia hot , hele mat la na nagoi mal ka atina hikina mai a Ierusalema ,
40002002  I ka l ana mat , At la ihea ka mea i hanau tho net i alli nei ka poe Iudalo ? No ka mea , ua ike nakou na ka atina hikina i kona hoku , a ua hele mai nei nakou i kukull hoonaiikal la .
40002003  A lohe ae la o Herode ke alii , apaoapo ae la kona olli , oia a me i Kierusalem a pau .
40002004  Alalau , hoululu ae la ia i ka poe kahuna nut a ne ka poe kakauuelo a na kanaka , ntinau mai la oia ia lakou , Al ia thea kahi e hanau ali o ka Mesta ?
40002005  Hot aku la betelhehena i Iudea net ; no ka mea , ua palapalala e ke kaula penel ,
40002006  O oe , e Betelhehena , i ka aino o Iuda , aole no oe ka nea vuuk loa twaeno o ko Iuda pote alli ; no ka mea , mailoko mal ou e hele mai ana kekaht alli ,
ezza a hananua i kau a nea brakas a ka tepelema .
```

Plain Text • Tab Width: 8 • Ln 1, Col 1 IN5

DATA AND METHODS

Parallel Corpora [Mayer & Cysouw, 2014]

language_name: 한국어
closest ISO 639-3: kor
year_short: Not available
year_long: Not available
title: 새번역
subtitle: New Translation of The Bible in Korean
URL: https://www.bible.com/de/bible/142/gen.1.rnksv
copyright_short: e Foundation for the Bible Society
copyright_long: Here's a new translation of the Bible with the copyright owned by the Foundation for the Bible Society, used by permission of hayeoteun.
이기여 사용한 성경에서 새번역의 저작권은 재단법인 대동서성교회 소유이며 재단법인 대동서성교회의 협력을 받고 사용하였음

40001081 아브라함 의 자손이요 다윗 의 자손인 예수 그리스도 의 계보를 미리하니
40001082 아브라함은 이삭 을 낳고, 이삭은 야곱 을 낳고, 야곱은 유다 와 그의 형제들을 낳고,
40001083 유다는 그의 형제들 중에서 베레스 와 세라 를 낳고, 베레스는 헤스론 을 낳고, 헤스론은 람 을 낳고,
40001084 람은 아미나답 을 낳고, 아미나답은 나순 을 낳고, 나순은 살론 을 낳고,
40001085 살론은 리암 에게서 보아스 를 낳고, 보아스는 룸 에게서 오벳 을 낳고, 오벳은 이세 을 낳고,
40001086 이세는 디단 왕을 낳았다. 디단은 우리민족 의 아내였던 이에게서 솔로몬 을 낳고,
40001087 솔로몬은 르호보암 을 낳고, 르호보암은 아비야 을 낳고, 아비야는 아삽 을 낳고,
40001088 아삽은 여호사밧 을 낳고, 여호사밧은 요암 을 낳고, 요암은 옷사마야 을 낳고,
40001089 옷사마야는 요담 을 낳고, 요담은 아하스 을 낳고, 아하스는 히스지야 을 낳고,
40001090 히스지야는 므낫세 을 낳고, 므낫세는 아모스 을 낳고, 아모스는 요시야 을 낳고,
40001091 예루살렘 주민이 비벌론 으로 끌려갈 무렵에, 요시야는 어고나 와 그의 형제들을 낳았다.
40001092 어고나는 주민이 비벌론 으로 끌려간 뒤에, 어고나는 스알디엘 을 낳고, 스알디엘은 아모스바벨 을 낳고,
40001093 스알디엘은 아비후온 을 낳고, 아비후온은 엘리睚온 을 낳고, 엘리睚온은 아소르 을 낳고,
40001094 아소르은 사동 을 낳고, 사동은 아킬 을 낳고, 아킬은 엘리羞 을 낳고,
40001095 엘리羞은 엘리실 을 낳고, 엘리실은 맘단 을 낳고, 맘단은 아곱 을 낳고,
40001096 아곱은 마리야 의 남편 요셉 을 낳았다. 마리야 에게서 그리스도 라고 하는 예수 가 태어나셨다.
40001097 그러므로 그 모든 수는 아브라함 으로부터 디단 까지 열네 대요, 디단 으로부터 바빌론 을 끌려간 때로부터 그리스도 까지 열네 대이다.
40001098 예수 그리스도 의 태어나심은 미리하니 . 그의 어머니 마리야 가 요셉 과 약혼하고 나서, 같이 살기 전에, 마리야 가 성령으로 임태한 사실이 드러났다.
40001099 마리야 의 날짜 묵은 은 유통 사증이어서 악혼자에게 부끄러움을 주지 않으려고, 기만해 파운더하 하였다.
40001100 묵은 이 일에 대해서 생각하는데, 주님의 천사가 꿈에 대해서 나타나며 말하였다 . “ 디단 의 자손 요셉 아, 두려워하지 말고, 마리야 를 네 아내로 맞아 들어라 . 그 태중에 있는 아기는 성령으로胎마음인 것이다 . ”
40001101 마리야 가 아들을 낳을 것이니, 너는 그 이름을 예수 라고 하여라 . 그가 자가 백성을 그들의 죄에서 구원하실 것이다 . ”
40001102 이 모든 일이 일어난 것은, 주님께 예언자를 시켜서 이르시기였다.
40001103 “보아라, 동정녀가 임태하여 아들을 낳을 것이다 , 그의 이름을 임마누엘 이라고 할 것이다 ”
40002001 해롯 원에, 예수께서 유대 베들레헴 에서 나섰다 . 그곳에 동쪽으로부터 박사들이 예루살렘 에 와서
40002002 말하였다 . “ 유대 인의 왕으로 나신 이가 어디로 계십니까 ? ” 우리가 동방에서 그의 별을 보고, 그에게 경배하려 왔습니다 . ”
40002003 해롯 원은 그 말을 듣고 달았습니다 . 예루살렘 사람들은 그와 함께 달았습니다 . ”
40002004 원은 백성의 대사장들과 율법 교사들을 만나고, 예루살렘 사람들도 그와 함께 달았습니다 . ”
40002005 그들이 원에게 말하였다 . “ 유대 베들레헴입니다 . 예언자가 이렇게 기록하여 좋았습니다 . ”
40002006 “너 웃을 땐에 있는 베들레헴 아, 너는 웃을 고마운가 . 가운데서는 아무, 찾지가 았다 . 너에게서 통치자가 나온 것이다 . ”
40002007 “너 웃을 땐에 있는 베들레헴 아, 너는 웃을 고마운가 . 가운데서는 아무, 찾지가 았다 . 너에게서 통치자가 나온 것이다 . ”

DATA AND METHODS

Parallel Corpora [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 Foundation 1994. <br> Unicode UTF-8 transformation and XML-tagging by Dirk Röckmann 2003
(www.nt-text.net). Revised Amharic Bible in XML ( 2003 ). 
# title: The New Testament in Amharic
# URL: http://ibound.biola.edu/index.cfm?method=downloads.showDownloadMain
# copyright_short: e Printed Version by United Bible Societies 1962
# copyright_long: Not available
40001001 የኢትዮጵያ ፌዴራል ቢሮንድ ክፍል ተመልከቻ መሬዳኝ፡
40001002 ነገርም ይከተላቸው መለያ፤ ስጋፍነው ያደረግኝ መለያ፤ ቁጥር-109 ይ-የታኝ መጋቢትም ይ-ሰጠኝ መለያ፤
40001003 ምስታዊ ክፍልና ፍትህር ንብረት መለያ፤ ምስታዊ ክፍልና ፍትህር ንብረት መለያ፤
40001004 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም እኩልም መለያ፤ ነጋግጧም እኩልም እኩልም መለያ፤
40001005 ነጋግጧም እኩልም እኩልም መለያ፤ ነጋግጧም እኩልም እኩልም እኩልም መለያ፤
40001006 ነጋግጧም ተታኝ መለያ፤
40001007 ነጋግጧም ሲ-ማይኝ መለያ፤ ሲ-ማይኝ ነጋግጧም ነጋግጧም መለያ፤ ነጋግጧም እኩልም መለያ፤
40001008 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም እኩልም መለያ፤ ነጋግጧም እኩልም እኩልም መለያ፤
40001009 ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤
40001010 ነጋግጧም ሲ-ማይኝ መለያ፤ ስጋፍነው የሚገኘው መለያ፤
40001011 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤
40001012 ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤
40001013 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40001014 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40001015 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40001016 ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤
40001017 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40001018 ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤
40001019 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40001020 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40001021 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40001022 ምስታዊ እኩልም መለያ፤
40001023 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40001024 ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤
40001025 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40002001 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40002002 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40002003 ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤
40002004 ምስታዊ እኩልም መለያ፤ ምስታዊ እኩልም መለያ፤
40002005 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
40002007 ነጋግጧም እኩልም መለያ፤ ነጋግጧም እኩልም መለያ፤
```

Plain Text • Tab Width: 8 • Ln 1, Col 1 IN5

DATA AND METHODS

Parallel Corpora [Mayer & Cysouw, 2014]

```
# language_name: Kannada
# closest ISO 639-3: kan
# year_short: 1992
# year_long: Not available
# title: ಇದ್ದ ದ್ವಿತೀಯ, ಅಥ ಉದ್ದೇಶ ಪ್ರಸಂಗ<br>The Holy Bible, Easy Reading Version, in the Kannada language of India
# URL: http://www.bible.is/ERWTC/Matt/1<br>http://dbs.org/bible/content/texts/KANVTC/
# copyright_short: © 1992-2010 World Bible Translation Center
# copyright_long: License Agreement for Bible Texts World Bible Translation Center Last Updated: September 21, 2006 Copyright © 2006 by World Bible Translation Center All Rights reserved. These Scriptures: * Are copyrighted by World Bible Translation Center. * Are not public domain. * May not be altered or modified in any form. * May not be sold or offered for sale in any form. * May not be used for commercial purposes (including, but not limited to, use in advertising or Web banners used for the purpose of selling online add space). * May be distributed without modification in electronic form for non-commercial use. However, they may not be hosted on any kind of server (including a Web or ftp server) without written permission. A copy of this license (without modification) must also be included. * May be quoted for any purpose, up to 1,000 verses, without written permission. However, the extent of quotation must not comprise a complete book nor should it amount to more than 50% of the work in which it is quoted. A copyright notice must appear on the title or copyright page using this pattern: "Taken from the HOLY BIBLE: EASY-TO-READ VERSIONTM © 2006 by World Bible Translation Center, Inc. and used by permission." If the text quoted is from one of WBTC's non-English versions, the printed title of the actual text quoted will be substituted for "HOLY BIBLE: EASY-TO-READ VERSIONTM." The copyright notice must appear in English or be translated into another language. When quotations from WBTC's text are used in non-sellable media, such as church bulletins, orders of service, posters, transparencies or similar media, a complete copyright notice is not required, but the initials of the version (such as "ERV" for the Easy-to-Read VersionTM in English) must appear at the end of each quotation. Any use of these Scriptures other than those listed above is prohibited. For additional rights and permission for usage, such as the use of WBTC's text on a Web site, or for clarification of any of the above, please contact World Bible Translation Center in writing or by email at distribution@wtc.com. World Bible Translation Center P.O. Box 820648 Fort Worth, Texas 76182, USA Telephone: 1-817-595-1664 Toll-Free in US: 1-888-54-BIBLE E-mail: info@wtc.com WBTC's web site - World Bible Translation Center's web site: http://www.wtbc.org
40001001 ಮಾನು ಶ್ರದ್ಧೆ ಪಡುವಿದ್ದಾರು . ಅದು ಹಿನ್ನೆಲೆ ಪಡುವಿದ್ದಾರು .
40001002 ಅಪ್ಪಣಿಯ ಪ್ರಾಣಕ ವರ್ತ . ಏಷಾದ ಯಾವಾನಿಯ ವರ್ತ . ಯಾವಾನಿಯ ಏಷಾದ ವರ್ತ ಅಪ್ಪಣಿಯ ವರ್ತ .
40001003 ಅಪ್ಪಣಿಯ ಪ್ರಾಣ ಮಾನು ಪಾರು ವರ್ತ . ( ಅದು ಹಿನ್ನೆಲೆ , ) ಕಿಂತ ಸಿಗ್ನಿಕ್ ವರ್ತ . ಸಿಗ್ನಿಕ್ ಅಪ್ಪಣಿ ವರ್ತ .
40001004 ಅಪ್ಪಣಿಯ ಅಪ್ಪಣಿಯ ವರ್ತ . ಅಪ್ಪಣಿಯ ಅಪ್ಪಣಿಯ ವರ್ತ . ಅಪ್ಪಣಿಯ ಅಪ್ಪಣಿಯ ವರ್ತ .
40001005 ಸರ್ವಾಂಗ ಪ್ರಾಣಕ ವರ್ತ . ( ಪ್ರಾಣಕ ಹಿನ್ನೆಲೆ . ) ಪ್ರಾಣಕ ಪ್ರಾಣಕ ವರ್ತ . ( ಪ್ರಾಣಕ ಹಿನ್ನೆಲೆ . ) ಪ್ರಾಣಕ ಹಿನ್ನೆಲೆ ವರ್ತ .
40001006 ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿ ಅಪ್ಪಣಿ ವರ್ತ . ಯಾವಾನಿ ಹಿನ್ನೆಲೆ ವರ್ತ . ( ಅಪ್ಪಣಿಯ ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿಯ ಹಿನ್ನೆಲೆಗ್ರಹಿತ . )
40001007 ಸಿಗ್ನಿಕ್ ಹಿನ್ನೆಲೆ ವರ್ತ . ಸಿಗ್ನಿಕ್ ಅಪ್ಪಣಿ ವರ್ತ . ಅಪ್ಪಣಿ ಅಪ್ಪಣಿ ವರ್ತ .
40001008 ಅಪ್ಪಣಿ ಅಪ್ಪಣಿಯ ವರ್ತ . ಅಪ್ಪಣಿಯ ಅಪ್ಪಣಿಯ ವರ್ತ . ಅಪ್ಪಣಿಯ ಅಪ್ಪಣಿಯ ವರ್ತ .
40001009 ಸಿಗ್ನಿಕ್ ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿ ವರ್ತ . ಅಪ್ಪಣಿ ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿ ವರ್ತ .
40001010 ಸಿಗ್ನಿಕ್ ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿ ವರ್ತ . ಸಿಗ್ನಿಕ್ ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿ ವರ್ತ . ಅಪ್ಪಣಿ ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿ ವರ್ತ .
40001011 ಸಿಗ್ನಿಕ್ ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿ ವರ್ತ . ( ಈ ಸಿಗ್ನಿಕ್ ಹಿನ್ನೆಲೆ ಯಾವಾನಿ ಅಪ್ಪಣಿ ಸಿಗ್ನಿಕ್ ಹಿನ್ನೆಲೆಗ್ರಹಿತ . )
40001012 ಸಿಗ್ನಿಕ್ ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿ ವರ್ತ . ಸಿಗ್ನಿಕ್ ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿ ವರ್ತ . ಸಿಗ್ನಿಕ್ ಹಿನ್ನೆಲೆ ಅಪ್ಪಣಿ ವರ್ತ .
40001013 ನಿತ್ಯಾದಿ ಮಹಾವಿಷ್ಣು ಜೀವಿ . ನಿತ್ಯಾದಿ ಮಹಾವಿಷ್ಣು ಜೀವಿ . ನಿತ್ಯಾದಿ ಮಹಾವಿಷ್ಣು ಜೀವಿ .
```

Plain Text • Tab Width: 8 •

Ln 1, Col 1

INS

DATA AND METHODS

Parallel Corpora

Corpus	Register	Size*	Size Ø*	Texts	Lang.
<i>UDHR</i> ¹	Legal	ca. 650K	1.831	372	348
<i>PBC</i> ²	Religious	ca. 10M	261K	1136	890
<i>EPC</i> 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, 1949; Juola 1998, 2008; Ehret & Szmrecsanyi, 2015]

MEASURING COMPLEXITY

Genesis 1:1

In the beginning God created the heavens and the earth . [eng]

Im Anfang schuf Gott den Himmel und die Erde ; [deu]

In principio Dio creò il cielo e la terra . [ita]

Au commencement Dieu créa les cieux et la terre . [fra]

MEASURING COMPLEXITY

Genesis 1:1

|In|the|beginning|God|created|the|heavens|and|the|earth|. [eng]

|Im|Anfang|schuf|Gott|den|Himmel|und|die|Erde|; [deu]

|In |principio|Dio|creò|il|cielo|e|la|terra|. [ita]

|Au|commencement|Dieu|créa|les|cieux|et|la|terre|. [fra]

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|. [eng]

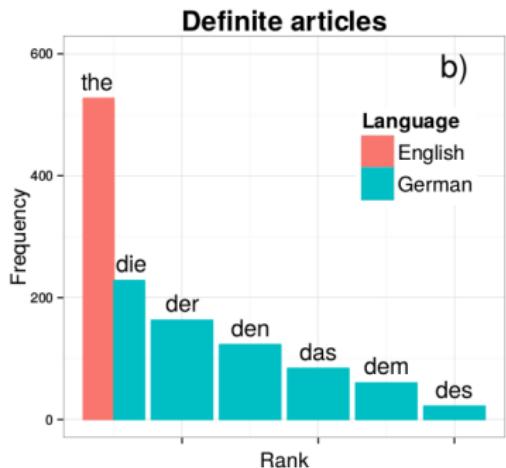
|**Im**|Anfang|schuf|Gott|**den**|Himmel|und|**die**|Erde|; [deu]

MEASURING COMPLEXITY

Example: English and German definite articles

|In|**the**|beginning|God|created|**the**|heavens|and|**the**|earth|. [eng]

|Im|Anfang|schuf|Gott|**den**|Himmel|und|**die**|Erde|; [deu]

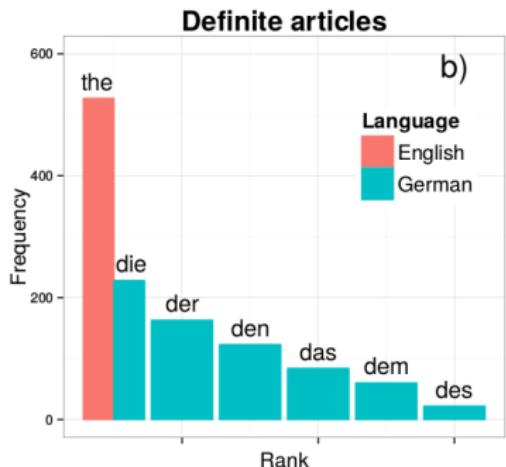


MEASURING COMPLEXITY

Example: English and German definite articles

|In|**the**|beginning|God|created|**the**|heavens|and|**the**|earth|. [eng]

|Im|Anfang|schuf|Gott|**den**|Himmel|und|**die**|Erde|; [deu]



Entropy [Shannon, 1949]:

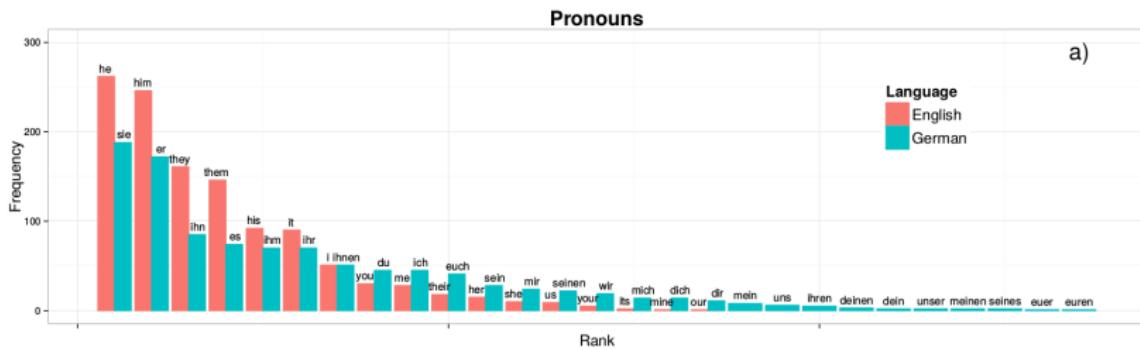
$$H_w = -K \sum_{i=1}^r p_i \log_2(p_i) \quad (1)$$

$$H_w^{eng} = 0$$

$$H_w^{deu} = 1.6$$

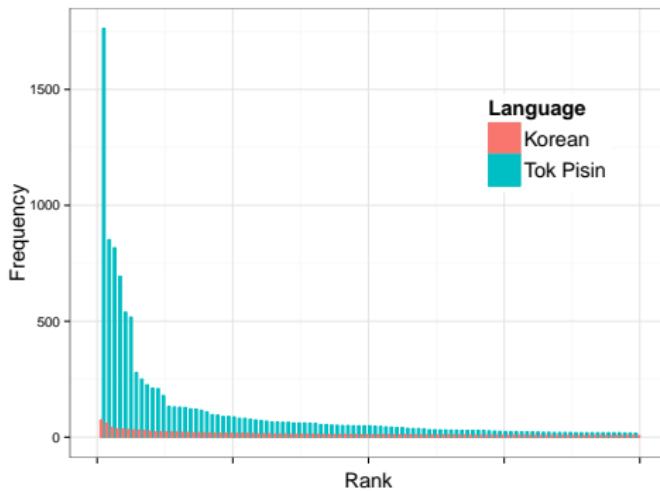
MEASURING COMPLEXITY

Example: English and German pronouns



$$H_w^{eng} = 2.16$$
$$H_w^{deu} = 2.64$$

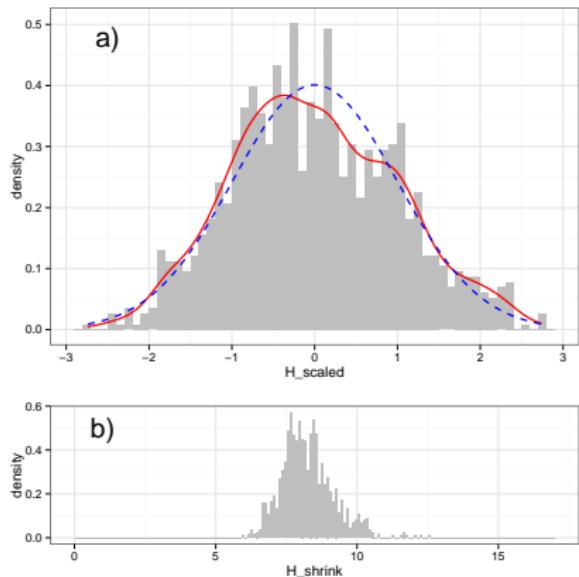
MEASURING COMPLEXITY



$$H_w^{kor} = 12.13$$
$$H_w^{tpi} = 6.83$$

MEASURING COMPLEXITY

The entropies of natural languages

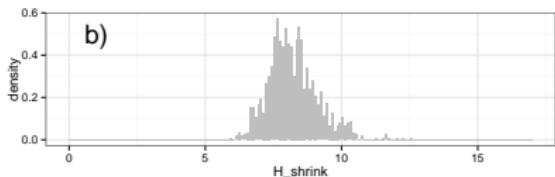
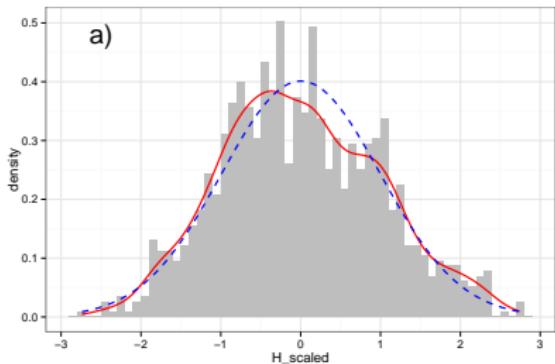


1529 texts
1050 languages
140 families
23 areas

[AUTOTYP
database, Bickel &
Nichols, 1999]

MEASURING COMPLEXITY

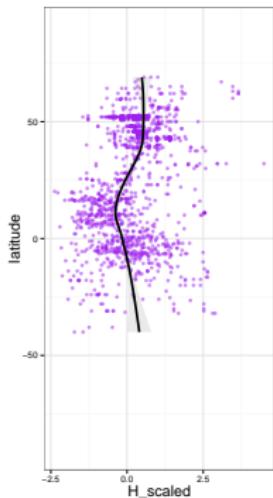
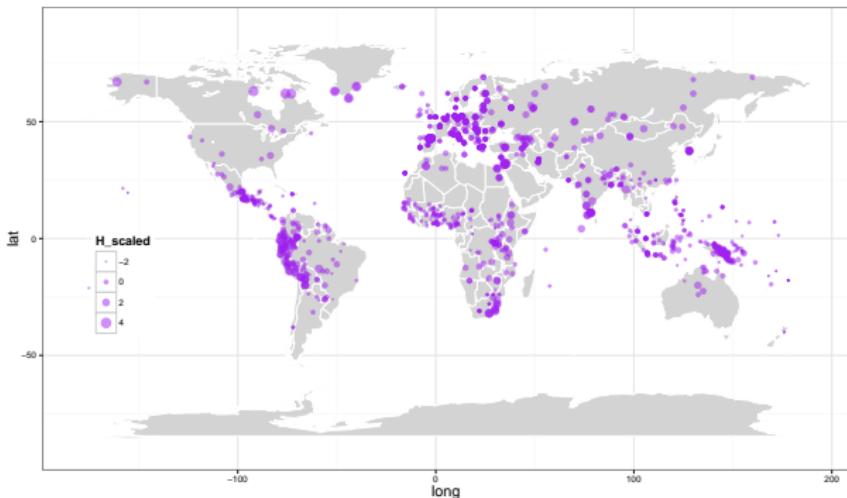
The entropies of natural languages



Why is there such a wide range of entropies, i.e. complexities?

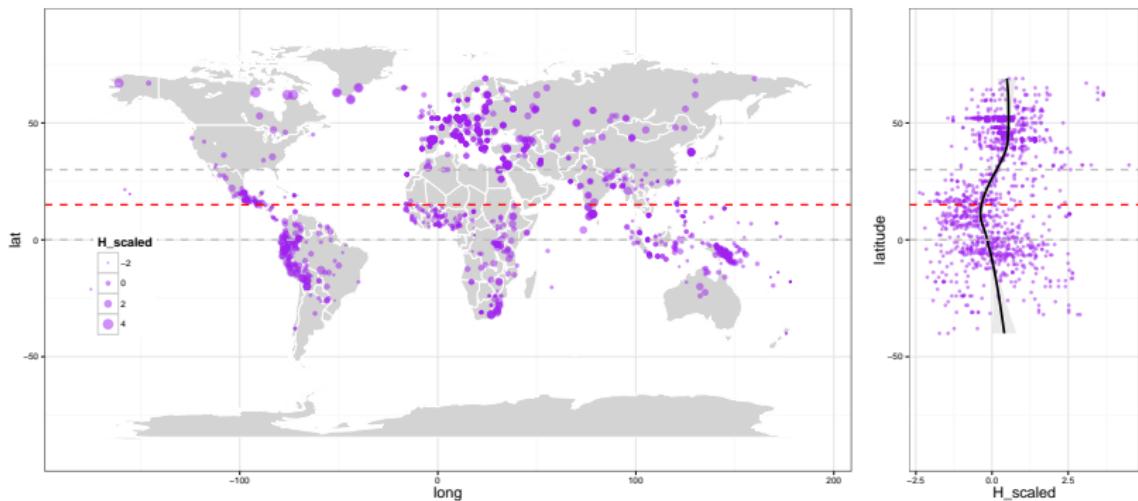
GEOGRAPHICAL DISTRIBUTION

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



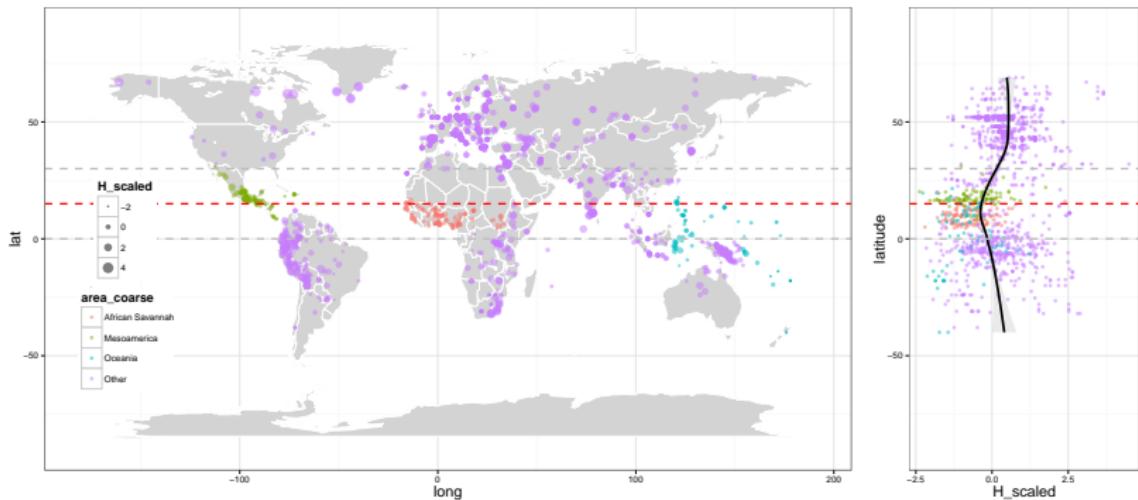
[R package ggmapping]

GEOGRAPHICAL DISTRIBUTION



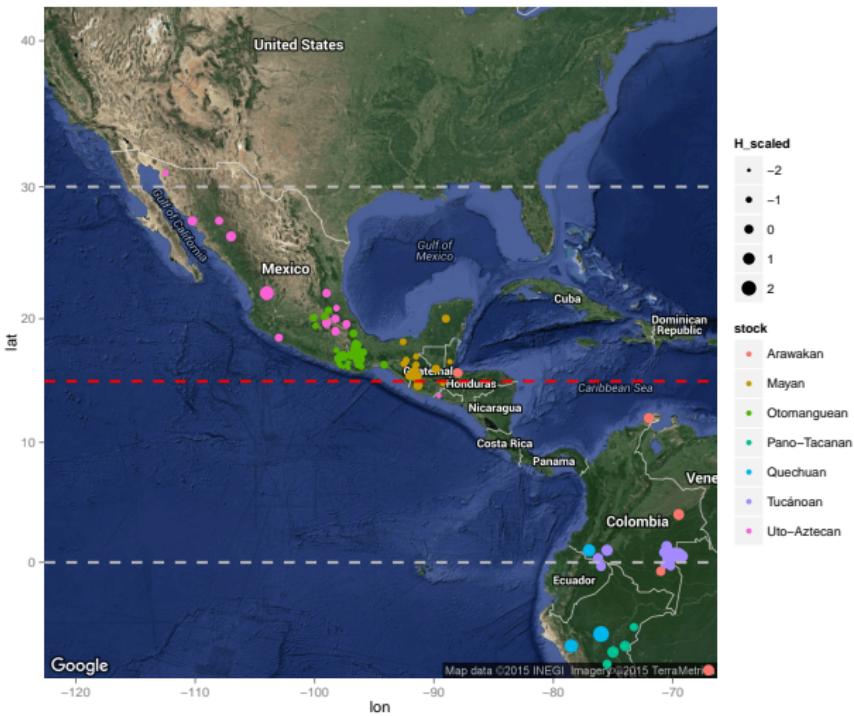
Low-Complexity-Belt (LCB)

GEOGRAPHICAL DISTRIBUTION

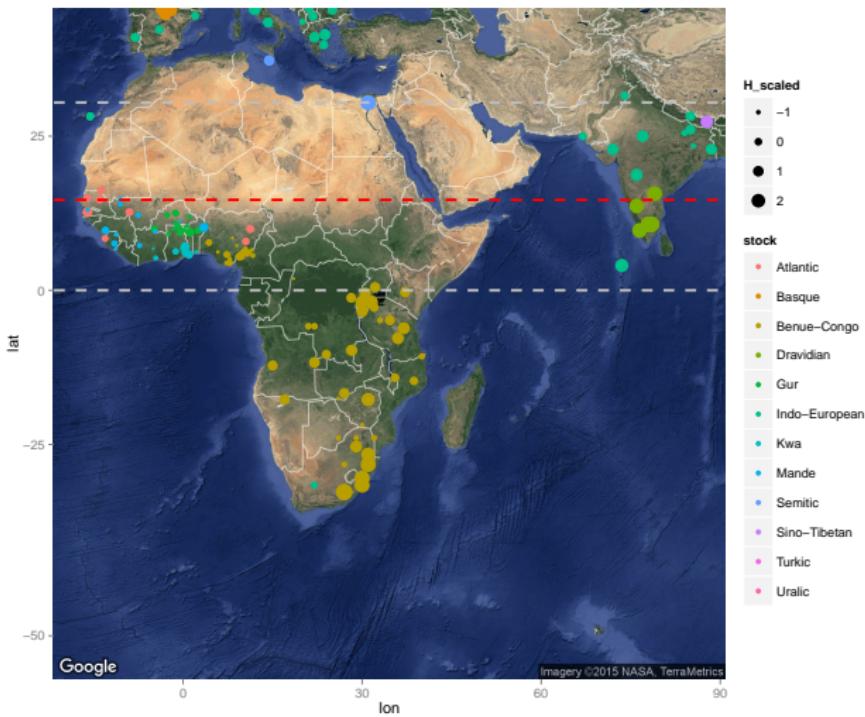


Low-Complexity-Belt (LCB)

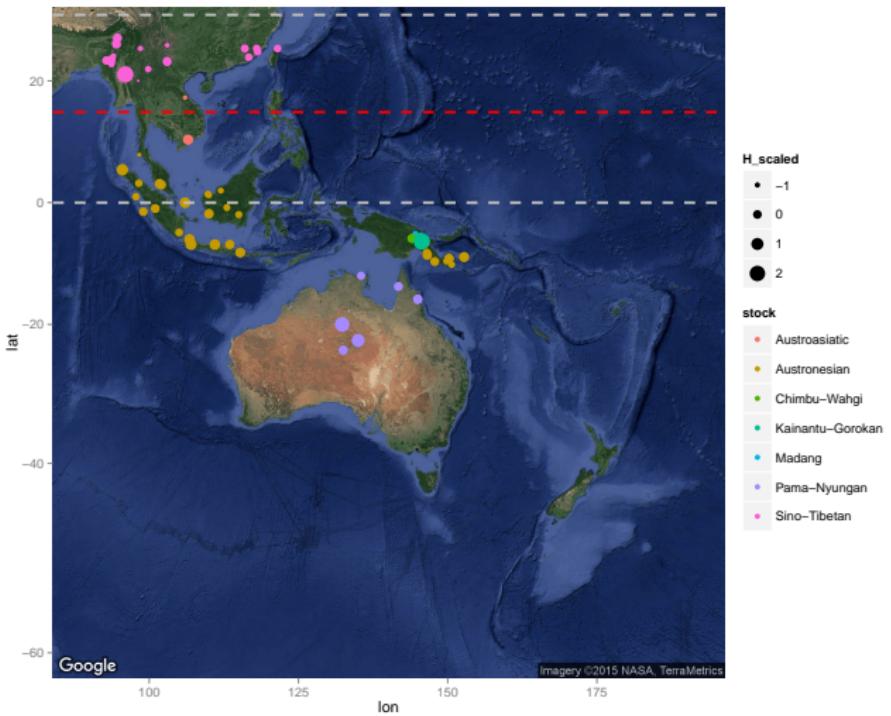
MESOAMERICA (*ONLY FAMILIES (STOCKS) WITH > 10 LANGUAGES)



AFRICAN SAVANNAH (*ONLY FAMILIES (STOCKS) WITH > 10 LANGUAGES)



OCEANIA (*ONLY FAMILIES (STOCKS) WITH > 10 LANGUAGES)



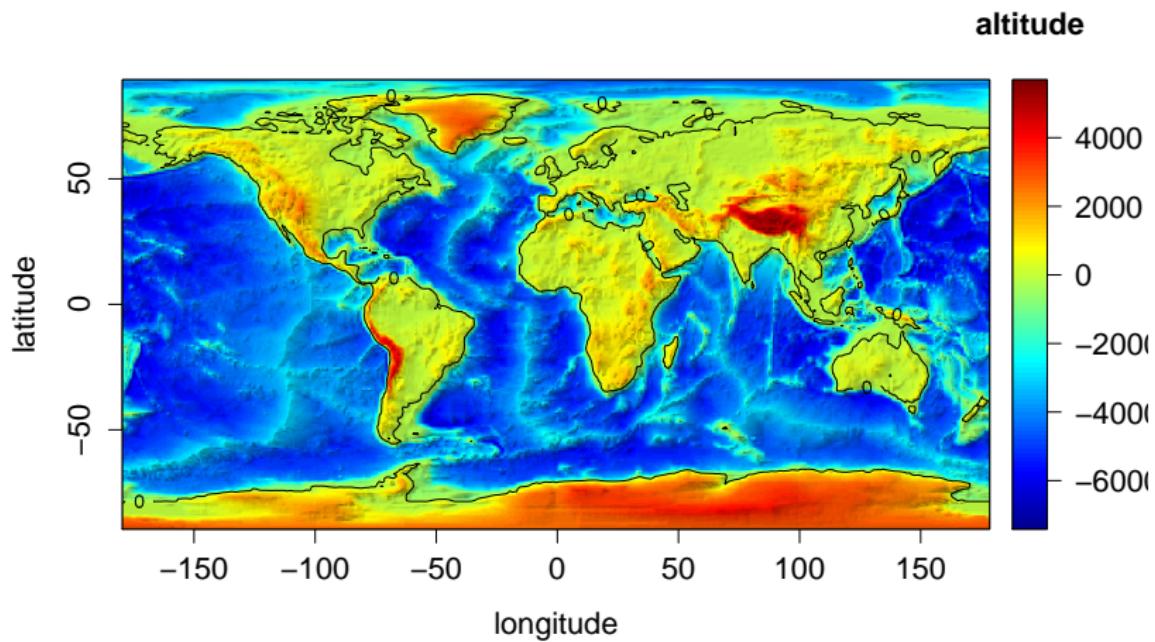
SUMMARY

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

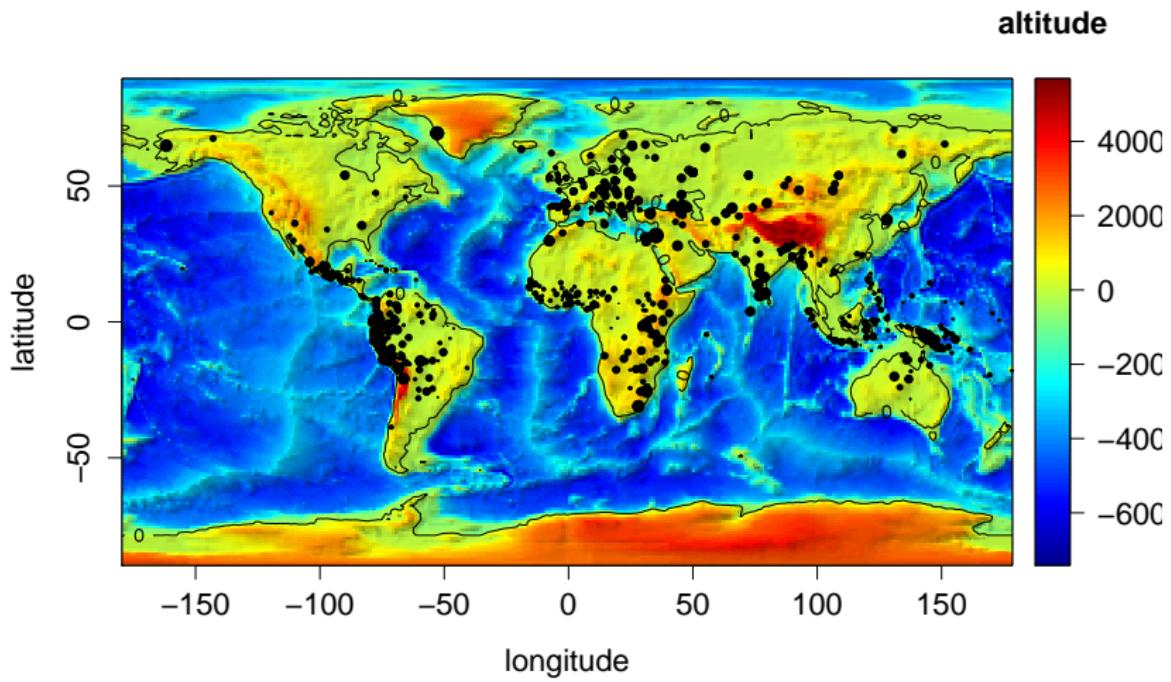
SUMMARY

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

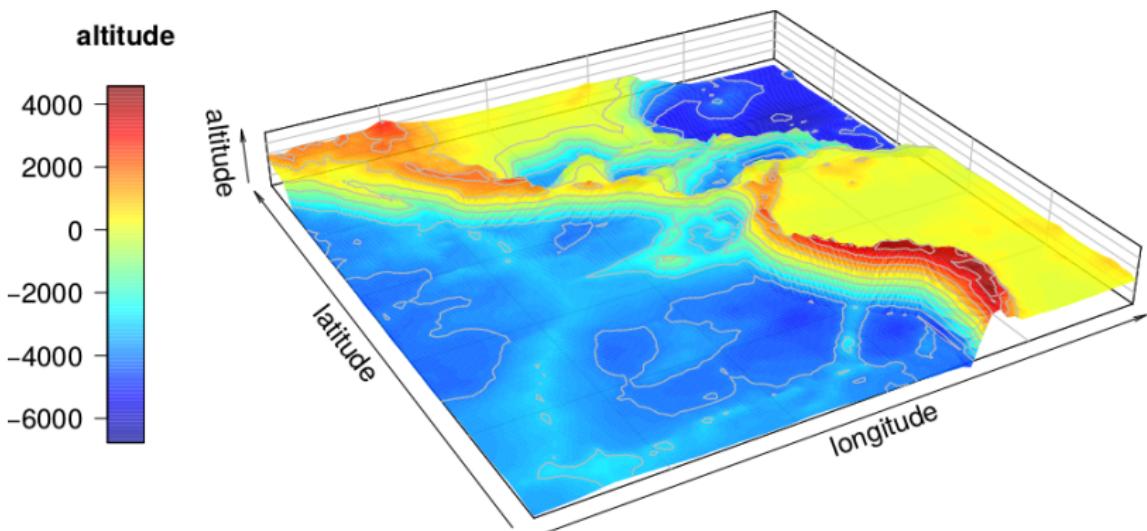
ALTITUDE [R PACKAGE PLOT3D]



ALITUDE AND COMPLEXITY



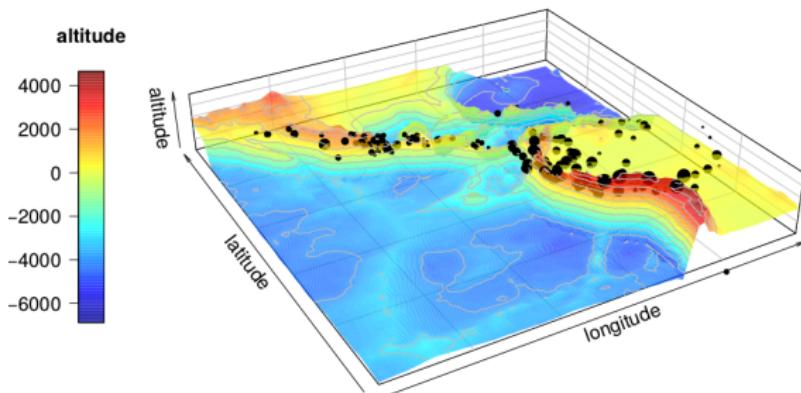
MESOAMERICA AND THE ANDES



[R package *plot3D*]

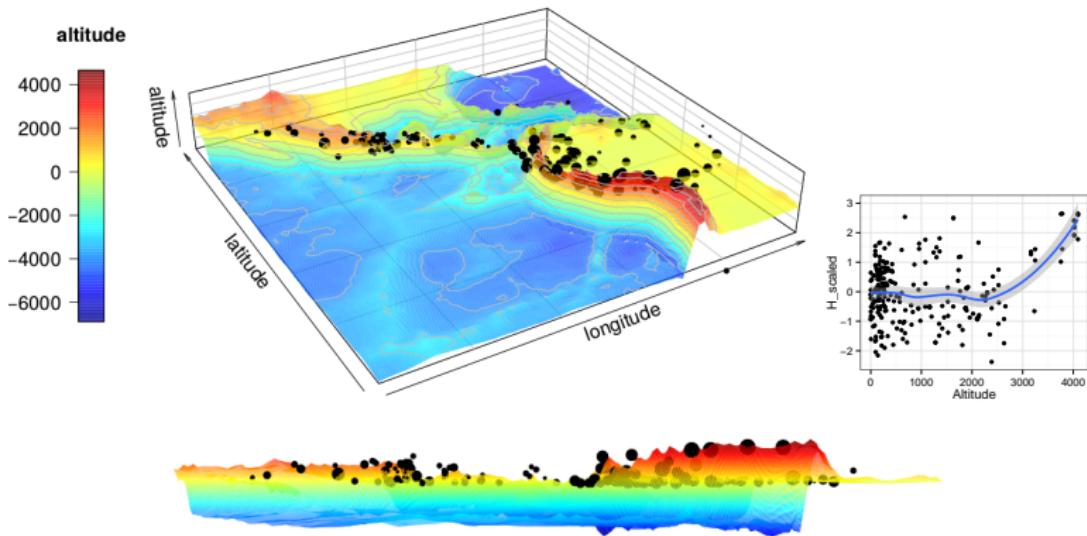
MESOAMERICA AND THE ANDES

183 LANGUAGES, 90 FAMILIES, 3 AREAS



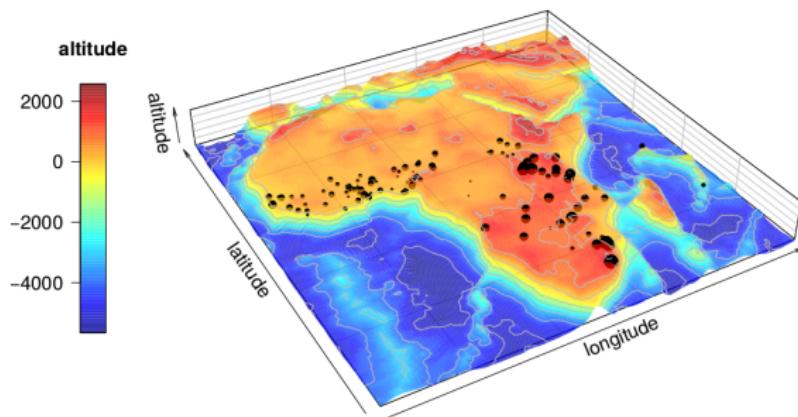
MESOAMERICA AND THE ANDES

183 LANGUAGES, 90 FAMILIES, 3 AREAS



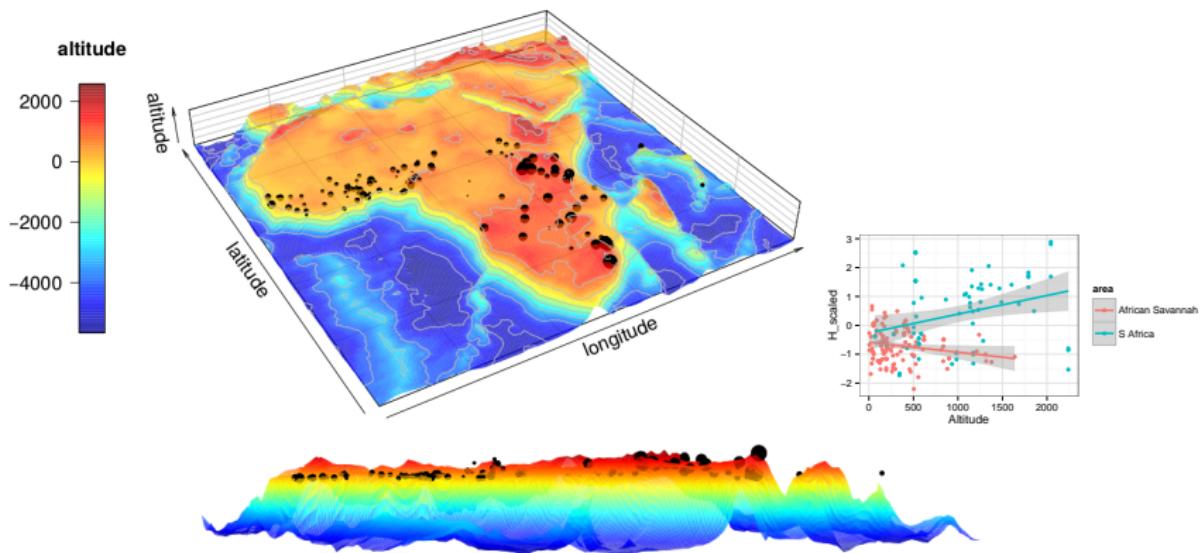
AFRICAN SAVANNAH AND SOUTH AFRICA

127 LANGUAGES, 21 FAMILIES, 2 AREAS



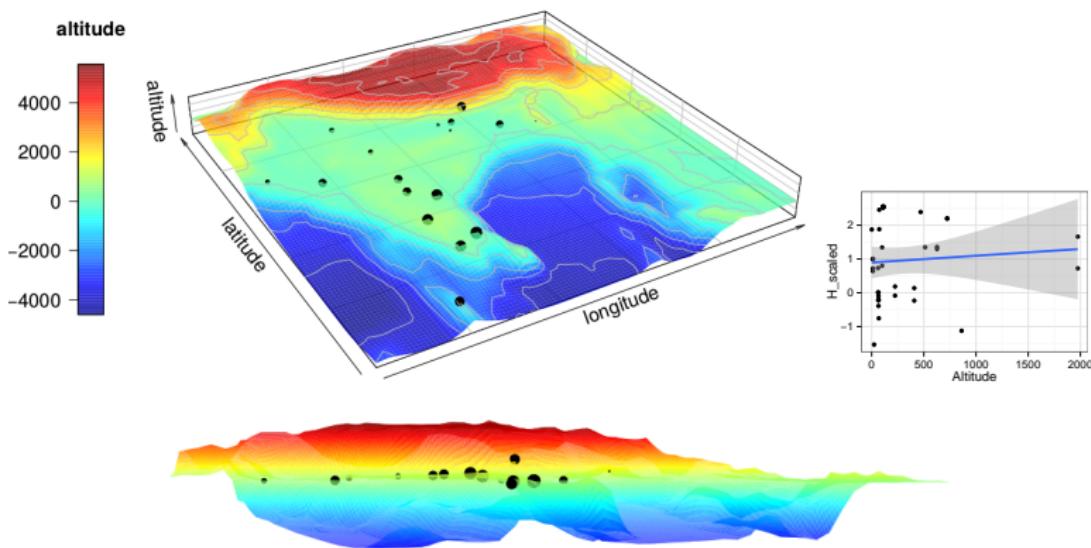
AFRICAN SAVANNAH AND SOUTH AFRICA

127 LANGUAGES, 21 FAMILIES, 2 AREAS

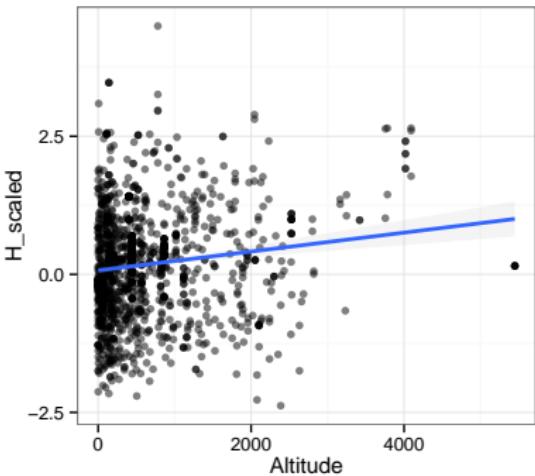
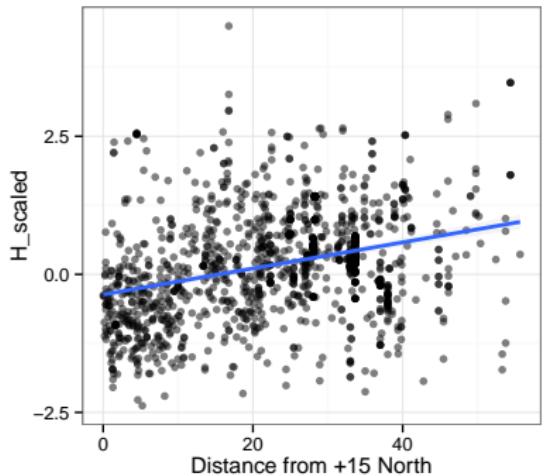


INDIC

18 LANGUAGES, 4 FAMILIES, 1 AREA



SIMPLE REGRESSIONS



Latitude: $\beta = 0.023, p = 2.2e^{-16}, R^2 = 0.0993$
Altitude: $\beta = 1.712e - 04, p = 2.24e^{-07}, R^2 = 0.01824$

MIXED-EFFECTS REGRESSIONS

Table : Results of stepwise linear mixed-effects regression.

Fixed	Random		AIC	t-value	$R^2\dagger$	
	intercept	slope			f	f+r
Lat,Alt	s [†]	-	2998	2.7,5.9	0.02	0.65
Lat,Alt	s,a	-	2848	-1.5,2.9	0.01	0.71
Lat,Alt	s,a,c	-	2776	-1.8,3	0.01	0.75
Lat,Alt	s,a,c,i	-	1931	0.96,2.7	0.01	0.94
Lat,Alt	s,a,c,i	c	1908	1.4,0.34	0.01	0.96
	s,a,c,i	c	1900	0,0	0	0.99

[†] s: stock, a: area, c: corpus, i: iso

[‡] f: fixed effect only, f+r: fixed and random effects

DISCUSSION

- ▶ The effects of latitude and altitude on complexity hold **between** - but not necessarily **within** different families and areas.

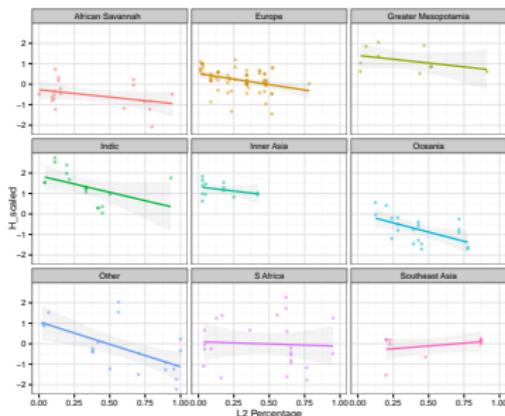
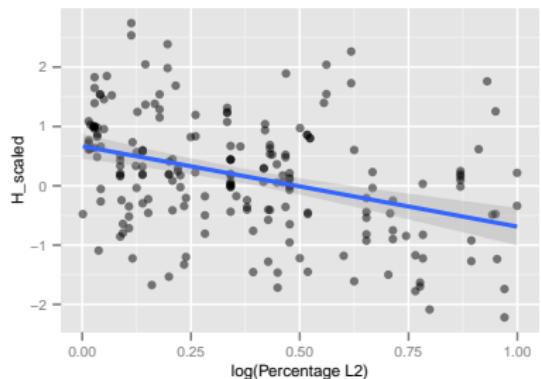
WHY PRE-HISTORIC LANGUAGE CONTACT?

- ▶ language contact and simplification [Wray & Grace, Trudgill 2011, McWhorter 2002]
- ▶ altitude and complexification via isolation [Nichols, 2013]
- ▶ population size and morphology [Lupyan & Dale, 2010, 2012, 2015]
- ▶ non-native speakers and case-marking [Bentz & Winter, 2012, 2013]
- ▶ non-native speakers and entropy [Bentz, Verkerk, Kiela, Hill & Buttery, 2015]

WHY PRE-HISTORIC LANGUAGE CONTACT?

Two main findings of Bentz et al. 2015:

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



WHY PRE-HISTORIC 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

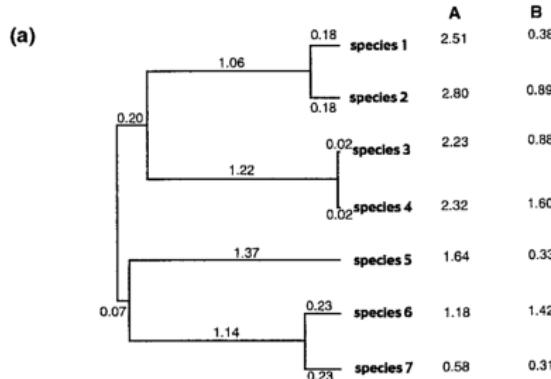
Table 5. Results for the phylogenetic signal analysis (mean λ).

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

PHYLOGENETIC SIGNAL ANALYSIS

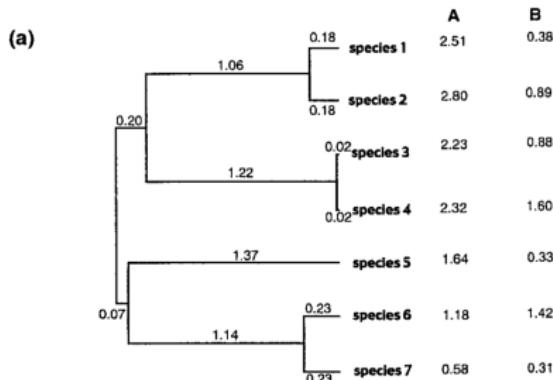
Pagel, 1999; Freckleton, Harvey & Pagel (2002)



(b)

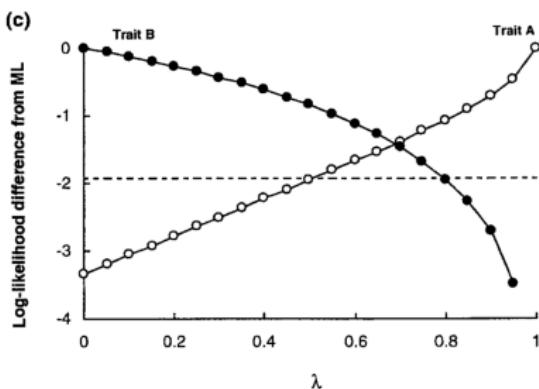
	species 1	species 2	species 3	species 4	species 5	species 6	species 7
species 1	1.44	1.26	0.20	0.20	0	0	0
species 2	1.26	1.44	0.20	0.20	0	0	0
species 3	0.20	0.20	1.44	1.42	0	0	0
species 4	0.20	0.20	1.44	1.44	0	0	0
species 5	0	0	0	0	1.44	0.07	0.07
species 6	0	0	0	0	0.07	1.44	1.21
species 7	0	0	0	0	0.07	1.21	1.44

PHYLOGENETIC SIGNAL ANALYSIS



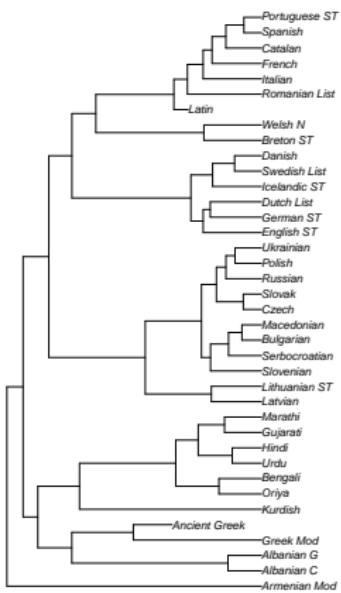
(b)

	species 1	species 2	species 3	species 4	species 5	species 6	species 7
species 1	1.44	1.26	0.20	0.20	0	0	0
species 2	1.26	1.44	0.20	0.20	0	0	0
species 3	0.20	0.20	1.44	1.42	0	0	0
species 4	0.20	0.20	1.44	1.44	0	0	0
species 5	0	0	0	0	1.44	0.07	0.07
species 6	0	0	0	0	0.07	1.44	1.21
species 7	0	0	0	0	0.07	1.21	1.44

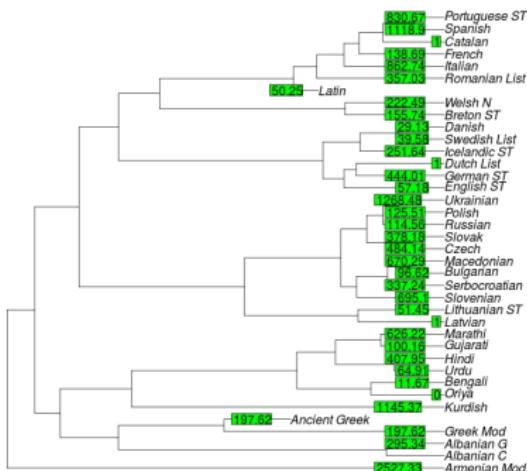
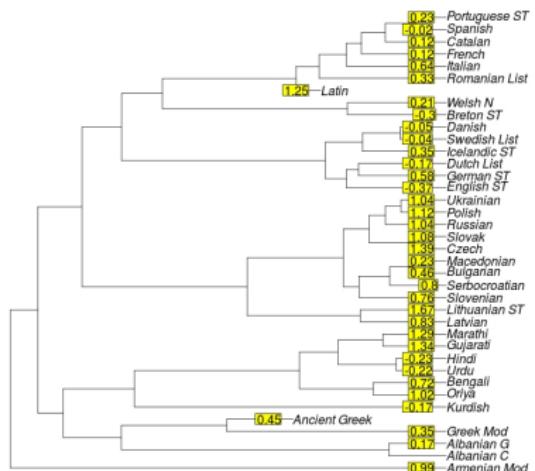


EXAMPLE: ENTROPY AND ALTITUDE

- ▶ 38 Indo-European languages, pruned tree from Bouckaert et al. 2012, BayesTraits V2 (Meade & Pagel 2013)



PHYLOGENETIC SIGNAL: ENTROPY AND ALTITUDE



Analysis with *BayesTraits V2* (Meade & Pagel 2013):

Entropy: $Lh = -26.62, \lambda = 0.999995$

Altitude: $Lh = -265.65, \lambda = 0.754488$

PHYLOGENETIC GENERALIZED LEAST SQUARES REGRESSION: ENTROPY AND ALTITUDE

- ▶ Correlation between entropy on altitude, but taking into account the phylogenetic covariance matrix

$L_h = -292.39, \mu_H = 0.50, \mu_{Alt} = 657.69, r=0.016, \lambda = 0.81$

$L_h = -292.40, \mu_H = 0.50, \mu_{Alt} = 658.24, r= 0, \lambda = 0.81$

POTENTIAL APPLICATIONS

- ▶ Using ASJP, IElex or EuraLex instead of cognate data can give better coverage
- ▶ *Phylogenetic signal* of language "external" factors: population size, non-native speakers, altitude, cultural traits?
- ▶ Replication of Lupyan & Dale (2010) with PGLS: population size and morphology
- ▶ Altitude, latitude, non-native speakers and language complexity with PGLS
- ▶ Language "internal" trade-offs, e.g. word order and case-marking. Is there evidence for phylogenetic correlation?
- ▶ Other WALS features?

WHAT ABOUT BONES, GENES, AND TOOLS?

