Measuring and Modelling Lexical Diversity across Languages

Christian Bentz & Douwe Kiela
cb696@cam.ac.uk

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Outline

Background

- Definition, measures
- **Qualitative results**: lexical diversity space, lexical diversity scales

Statistical Modelling

- Sociolinguistic factors, language families and regions
- **Quantitative results**: simple linear model, mixed-effects model

Future Directions

- Potential repercussions on word order flexibility

Definition

Lexical diversity

The **distribution of word forms** used to encode a **constant information content**

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Quantitative measures

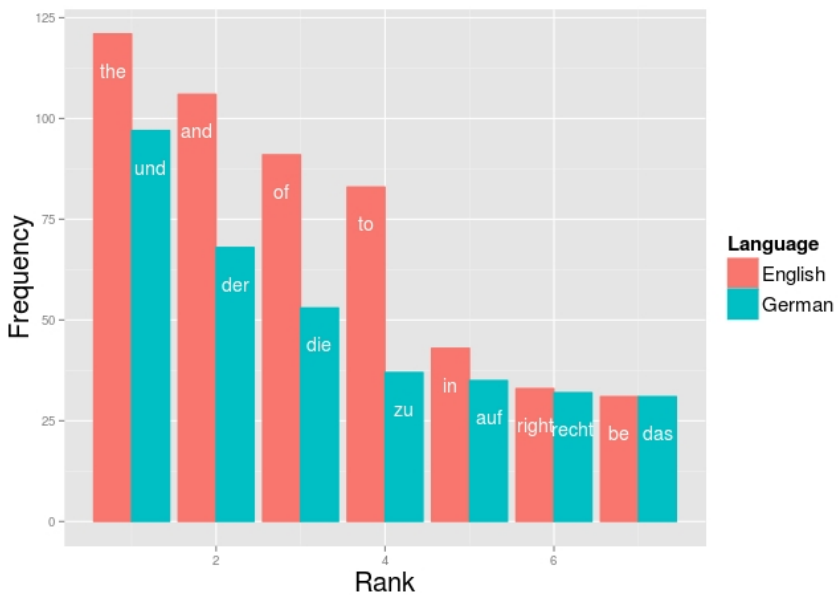
Frequency distributions

Order types (word forms delimited by white spaces) according to their token frequencies (Zipf,1932,1949)

Quantitative measures

Frequency distributions

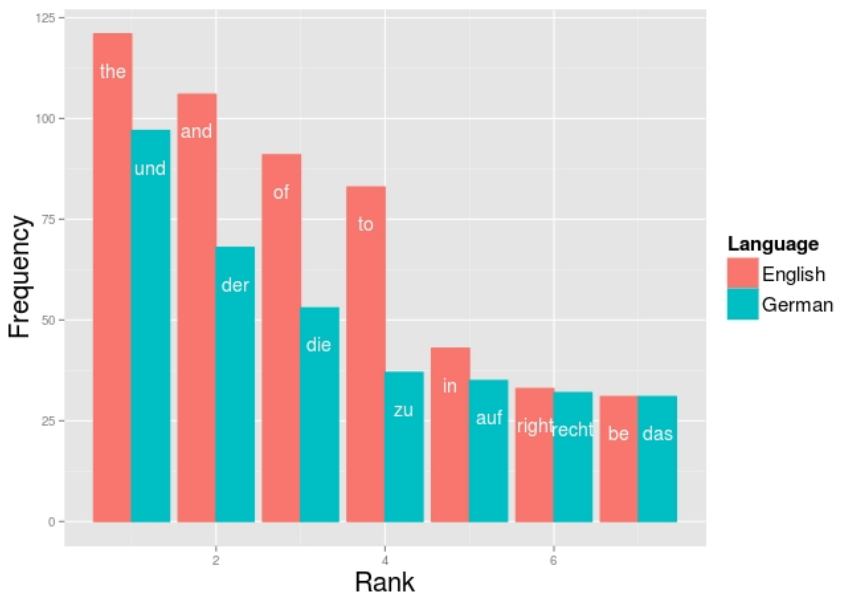
Order types (word forms delimited by white spaces) according to their token frequencies (Zipf,1932,1949)

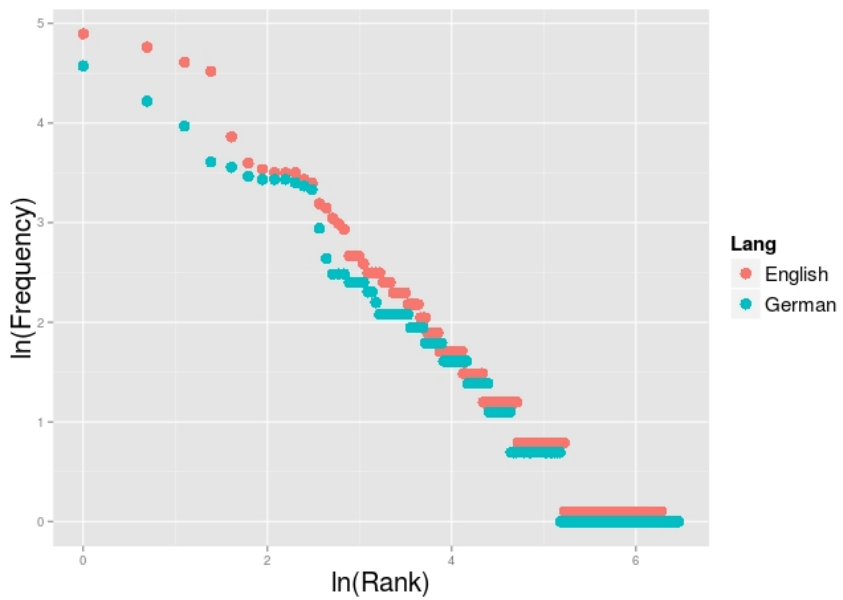


Quantitative measures

Frequency distributions

Order types (word forms delimited by white spaces) according to their token frequencies (Zipf,1932,1949)





Quantitative measures

Zipf-Mandelbrot's law (Zipf,1949; Mandelbrot, 1953)

$$f(r_i) = \frac{C}{\beta + r_i^{\alpha}},$$

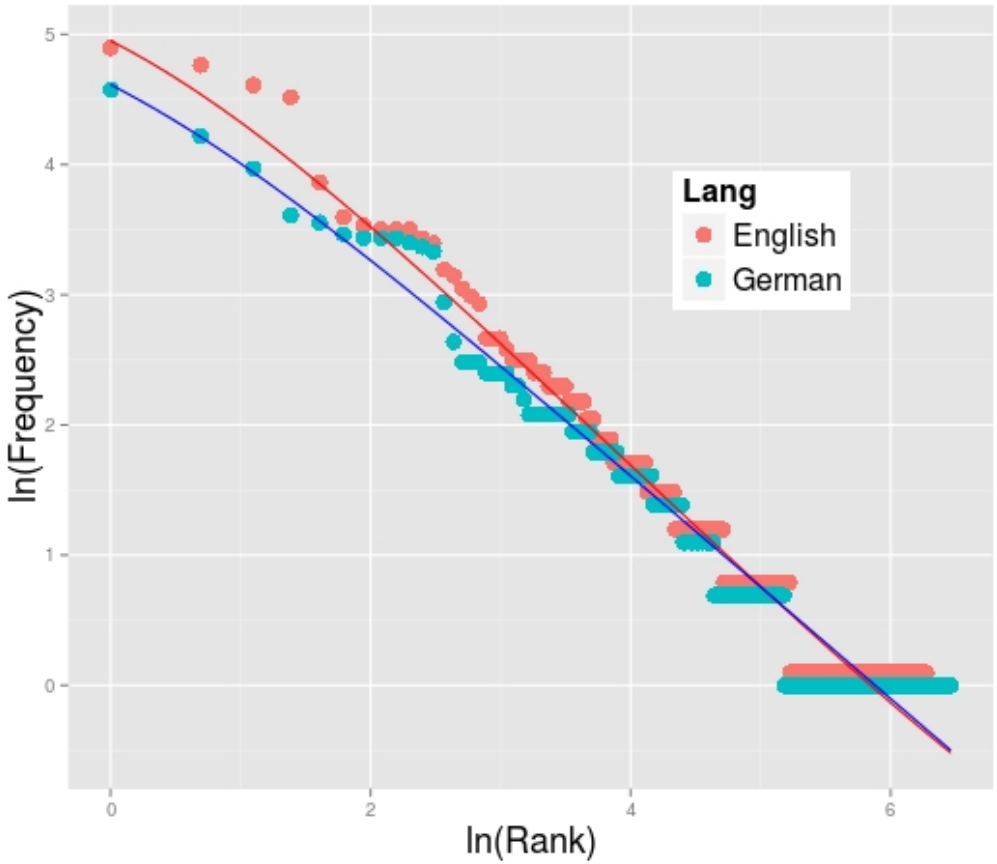
$C > 0,$
$\alpha > 0,$
$\beta > -1,$
$i = 1, 2, \ldots, n$

Quantitative measures

Zipf-Mandelbrot's law (Zipf,1949; Mandelbrot, 1953)

$$f(r_i) = \frac{C}{\beta + r_i^{\alpha}},$$

$C > 0,$
$\alpha > 0,$
$\beta > -1,$
$i = 1, 2, \ldots, n$



Quantitative measures

Shannon entropy (Shannon & Weaver, 1949)

$$H = -K \sum_{i=1}^{k} p_i \times log_2(p_i)$$

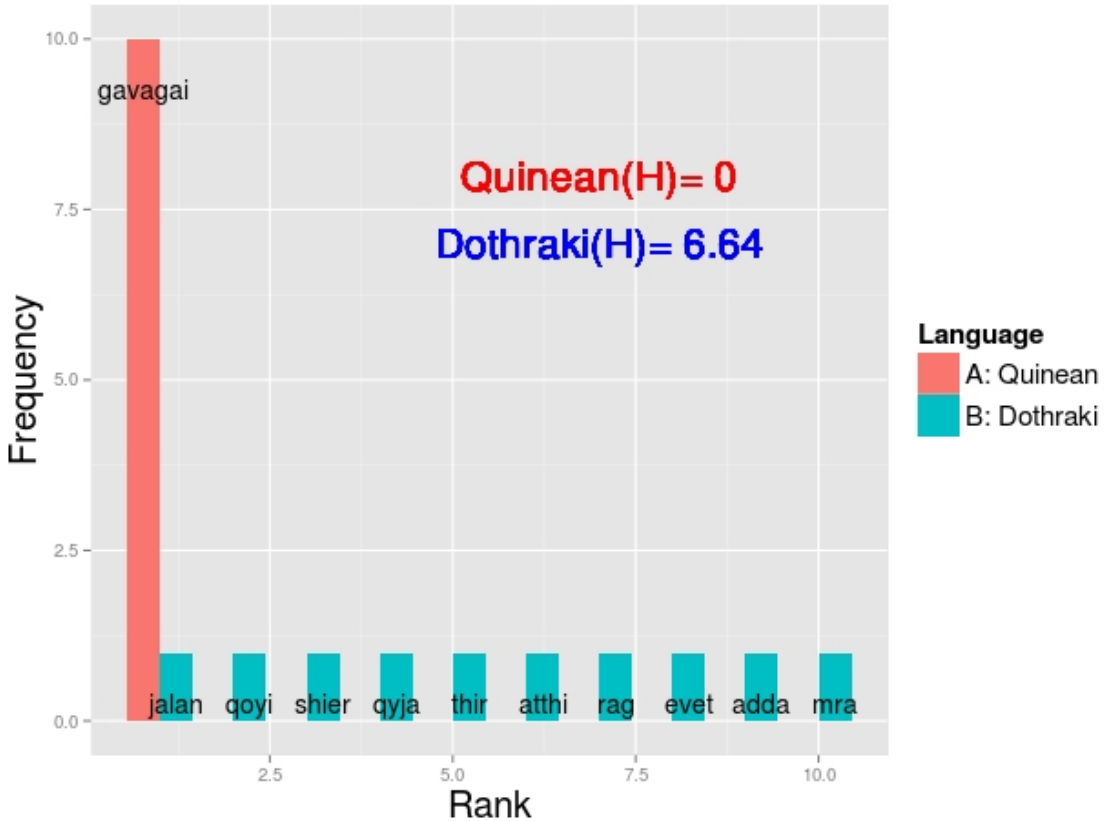
$$p_i : \frac{frequency\ of\ w_i}{total\ number\ of\ tokens}$$

Quantitative measures

Shannon entropy (Shannon & Weaver, 1949)

$$H = -K \sum_{i=1}^{k} p_i \times log_2(p_i)$$

$$p_i : \frac{frequency\ of\ w_i}{total\ number\ of\ tokens}$$

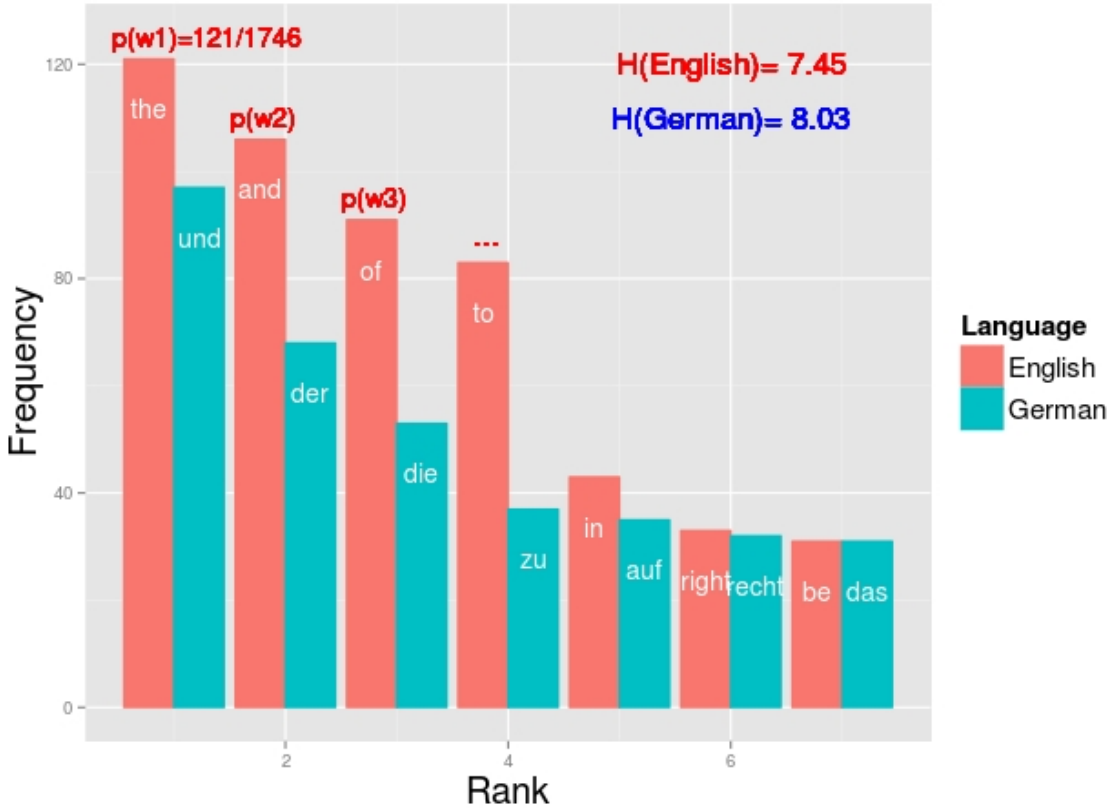


Quantitative measures

Shannon entropy (Shannon & Weaver, 1949)

$$H = -K \sum_{i=1}^{k} p_i \times log_2(p_i)$$

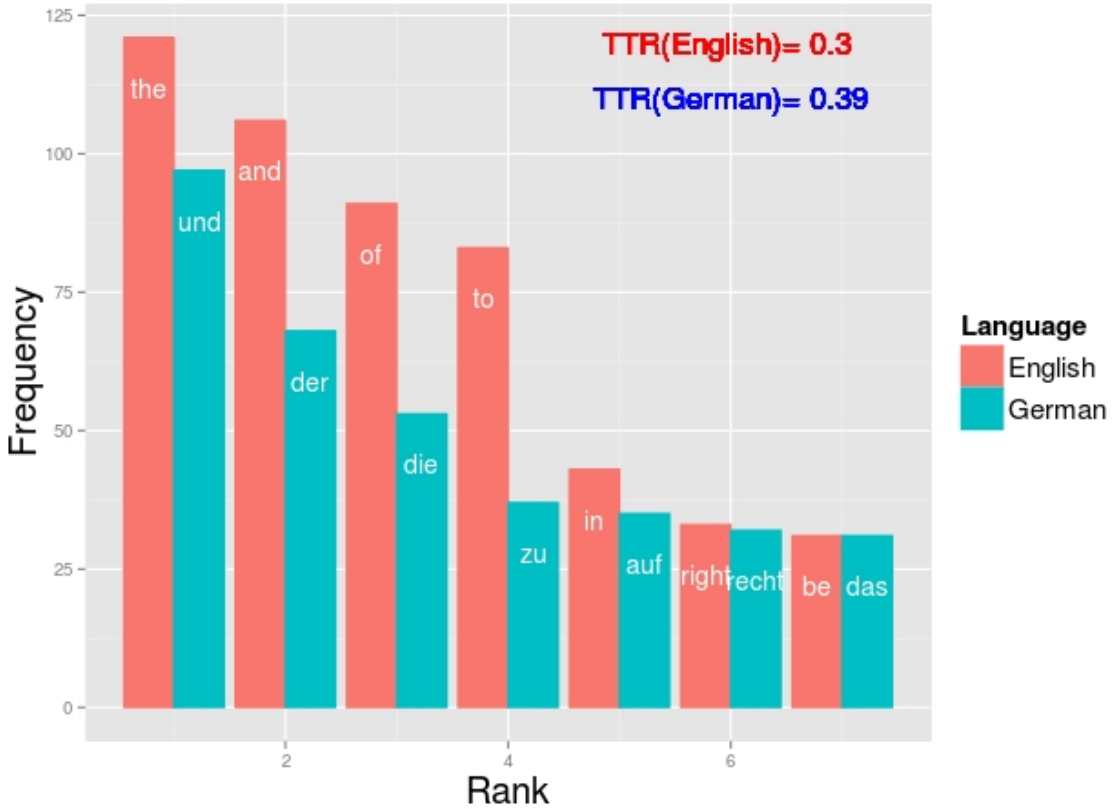
$$p_i : \frac{frequency\ of\ w_i}{total\ number\ of\ tokens}$$



Quantitative measures

Type-Token Ratio (TTR)

$$TTR = \frac{number\ of\ types}{number\ of\ tokens}$$



Measures: Summary

Linguistic Studies

- **ZM-law**
  Zipf (1932, 1935, 1949), Ha et al. (2006), Baroni (2009), Popescu et al. (2008,2009, 2010), Baixeries et al. (2013), Bentz et al. (2014)

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Measures: Summary

Linguistic Studies

- **ZM-law**
  Zipf (1932, 1935, 1949), Ha et al. (2006), Baroni (2009), Popescu et al. (2008,2009, 2010), Baixeries et al. (2013), Bentz et al. (2014)
- **Shannon H**
  Gries 2012 $\rightarrow$ Possible connection with language learning?

Measures: Summary

Linguistic Studies

- **ZM-law**
  Zipf (1932, 1935, 1949), Ha et al. (2006), Baroni (2009), Popescu et al. (2008,2009, 2010), Baixeries et al. (2013), Bentz et al. (2014)
- **Shannon H**
  Gries 2012 $\rightarrow$ Possible connection with language learning?
- **Type-Token Ratio**
  Tweedie & Baayen (1998), Baayen (2001)

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Measures: Summary

Linguistic Studies

- **ZM-law**
  Zipf (1932, 1935, 1949), Ha et al. (2006), Baroni (2009), Popescu et al. (2008,2009, 2010), Baixeries et al. (2013), Bentz et al. (2014)
- **Shannon H**
  Gries 2012 $\rightarrow$ Possible connection with language learning?
- **Type-Token Ratio**
  Tweedie & Baayen (1998), Baayen (2001)

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Qualitative results

LDT in Diachrony: Old English and Modern English

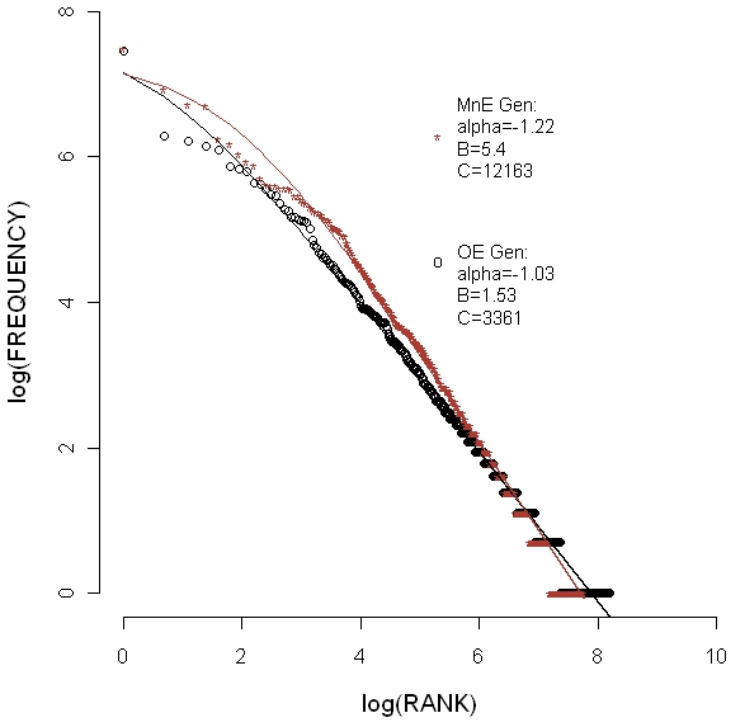
Bentz, Kiela, Hill & Buttery (2014) Zipf's law and the grammar of languages. A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory.* aop.

Qualitative results

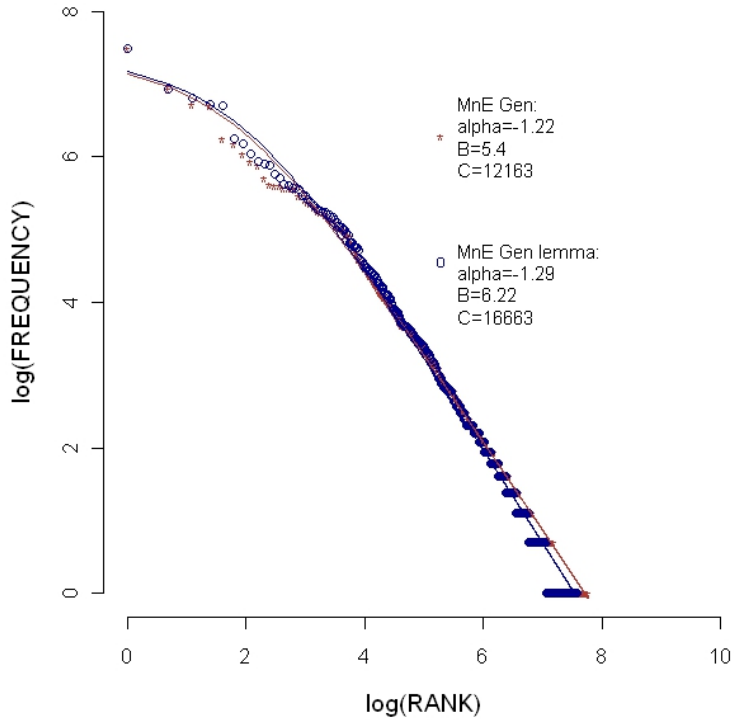
LDT in Diachrony: Old English and Modern English

Bentz, Kiela, Hill & Buttery (2014) Zipf's law and the grammar of languages. A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory.* aop.

10a) MnE and OE Genesis



10b) MnE Genesis + lemmatized version



Qualitative results

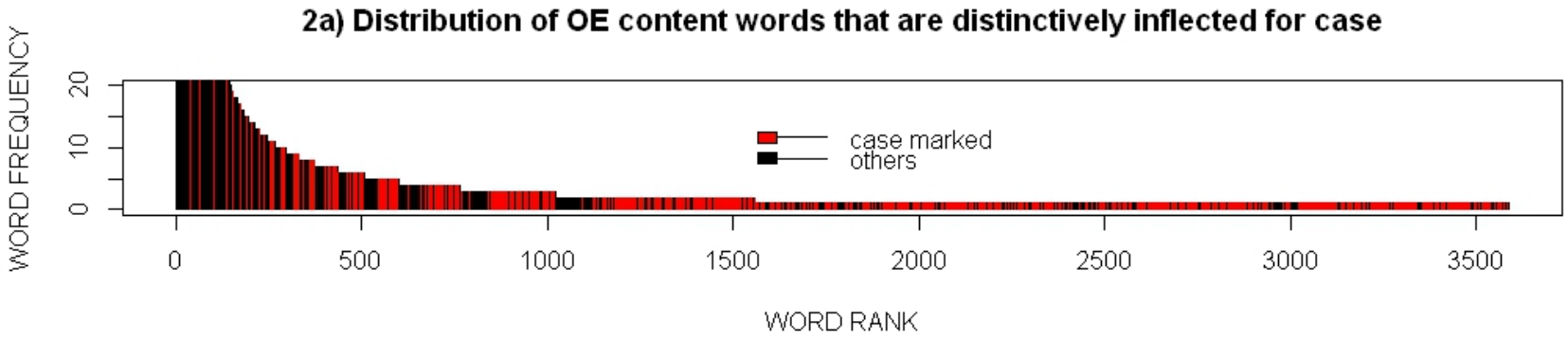
What drove this change in frequencies?

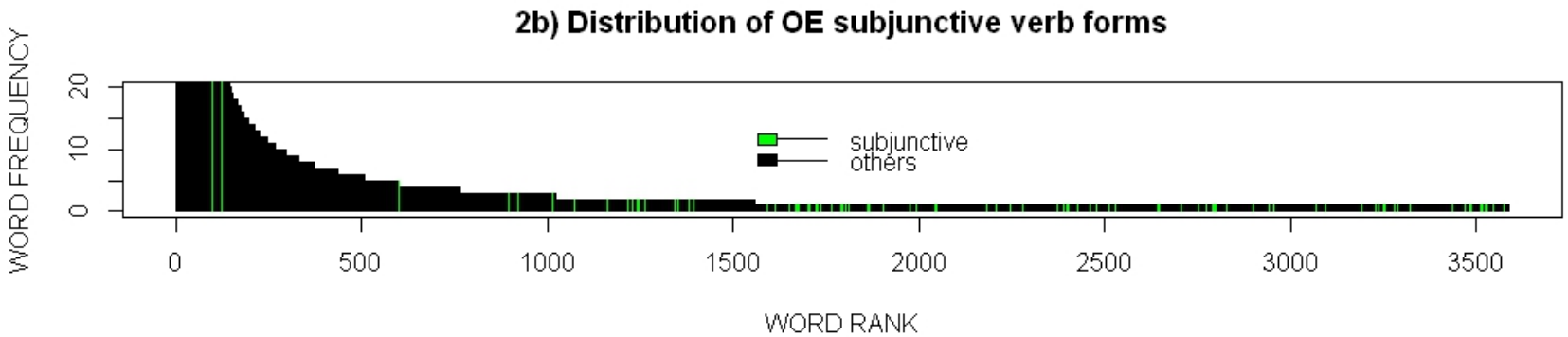
Inflections: OE *land, lande, landes* → ModE *land, in the land, of the land*

Qualitative results

What drove this change in frequencies?

Inflections: OE *land, lande, landes* → ModE *land, in the land, of the land*





Qualitative results

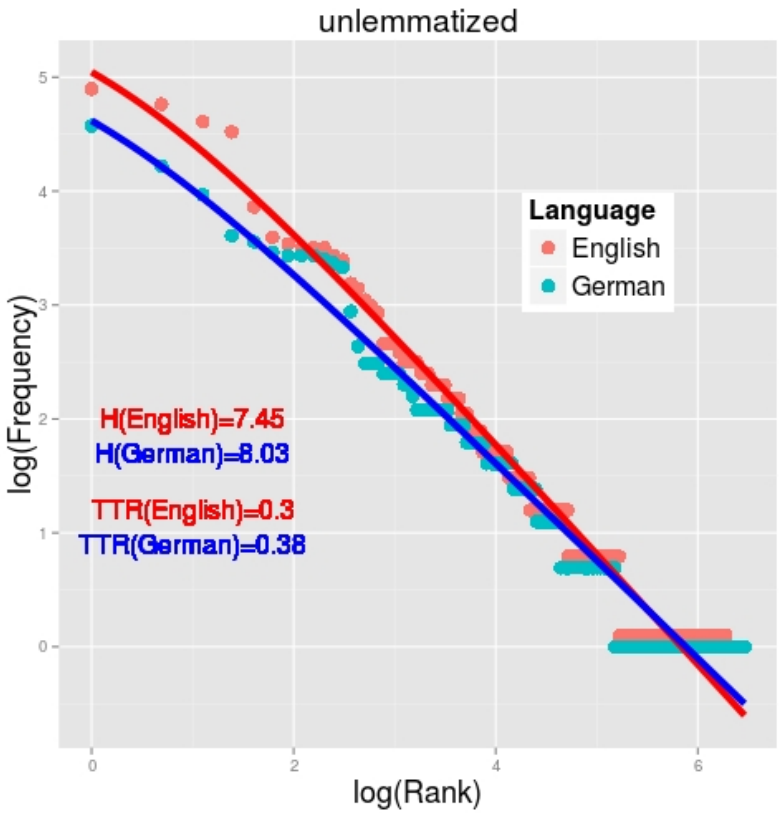
Inflections in German and English

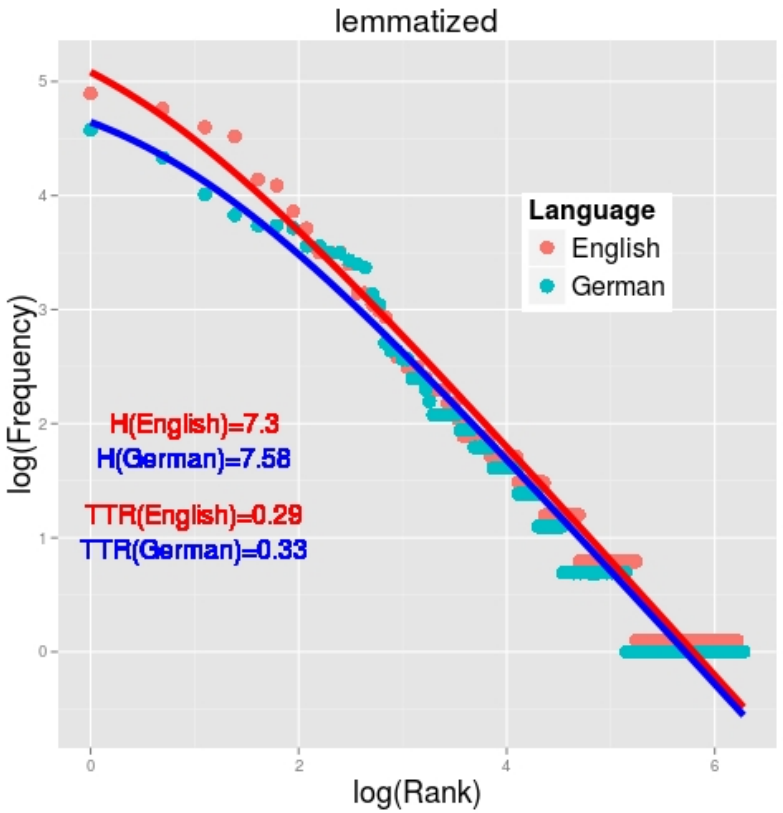
Lemmatization: yields frequency distributions without inflected types for English and German

Qualitative results

Inflections in German and English

Lemmatization: yields frequency distributions without inflected types for English and German





Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Qualitative results

Inflections and frequencies

Having inflections in a language automatically creates lower frequency items

$\rightarrow$ higher entropy (higher uncertainty)

$\rightarrow$ higher type-token ratios

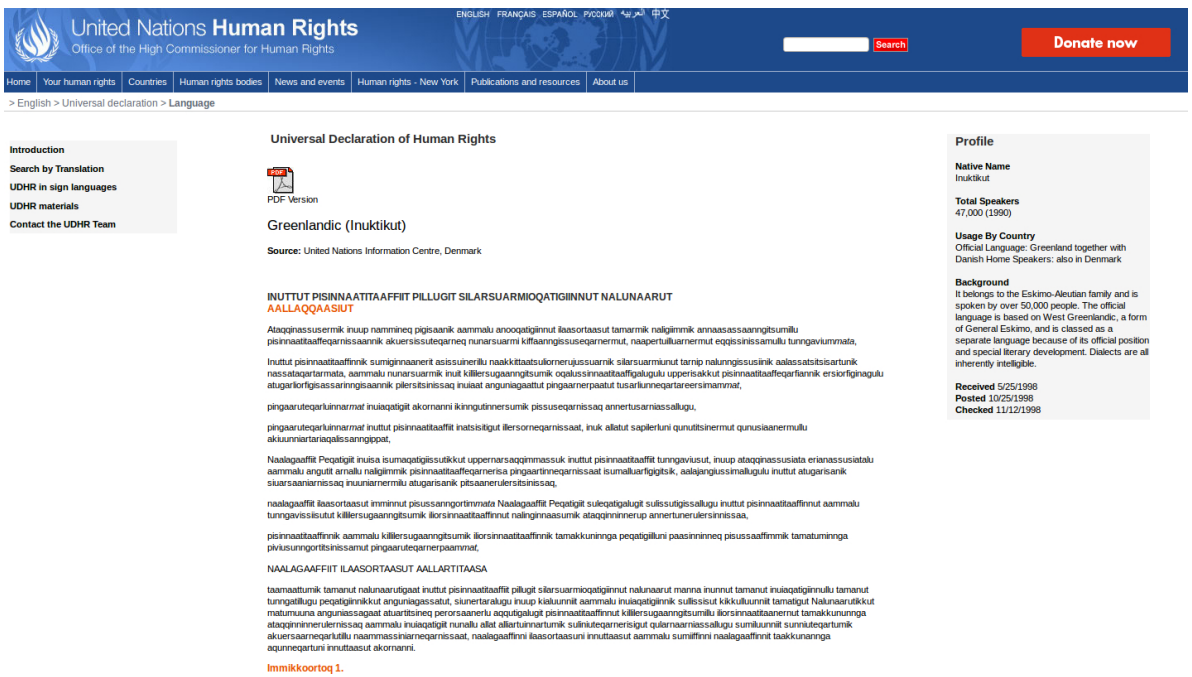
$\rightarrow$ lower ZM-parameters

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Qualitative results

Synchronic analyses: LDT across languages of today

Parallel texts
- Universal Declaration of Human Rights (UDHR) in ∼400 languages
- Parallel Bible Corpus (∼800 languages)
- Europarl Corpus (21 languages)



United Nations **Human Rights**
Office of the High Commissioner for Human Rights

ENGLISH FRANÇAIS ESPAÑOL РУССКИЙ العربية 中文

Search | Donate now

Home | Your human rights | Countries | Human rights bodies | News and events | Human rights - New York | Publications and resources | About us

> English > Universal declaration > **Language**

Introduction
Search by Translation
UDHR in sign languages
UDHR materials
Contact the UDHR Team

Universal Declaration of Human Rights

PDF Version

Greenlandic (Inuktikut)

Source: United Nations Information Centre, Denmark

INUTTUT PISINNAATITAAFFIT PILLUGIT SILARSUARMIOQATIGIINNUT NALUNAARUT
AALLAQQAASIUT

Ataqqinassusermik inuup nammineq pigisaanik aammalu anooqatigiinnut ilaasortaasut tamarmik naligiimmik annaasassaanngitsumillu pisinnaatitaaffeqarnissaannik akuerisissuteqarneq nunarsuarmi kiffaanngissuseqarnermut, naapertuilluarnermut eqqissinissamullu tunngaviu*mmata*,

Inuttut pisinnaatitaaffinnik sumiginnaanerit asissuinerillu naakkittaatsuliornerujussuarnik silarsuarmiunut tarnip nalunngissusiinik aalassatsitsisartunik nassataqartarmata, aammalu nunarsuarmik inuit killilersugaanngitsumik oqalussinnaatitaaffigalugulu upperisakkut pisinnaatitaaffeqarfiannik ersiorfiginagulu atugarliorfigisassaringisaannik pilersitsinissaq inuiaat anguniagaattut pingaarnerpaatut tusarliunneqartareersimam*mat*,

pingaaruteqarluinnar*mat* inuiaqatigiit akornanni ikinngutinnersumik pissuseqarnissaq annertusarniassallugu,

pingaaruteqarluinnar*mat* inuttut pisinnaatitaaffiit inatsisitigut illersorneqarnissaat, inuk allatut sapilerluni qunutitsinermut qunusiaanermullu akiuunniartariaqalissanngippat,

Naalagaaffiit Peqatigiit inuisa isumaqatigiissutikkut uppernarsaqqinnassuk inuttut pisinnaatitaaffit tunngaviusut, inuup ataqqinassusiata erianassusiatalu aammalu angutit arnallu naligiimmik pisinnaatitaaffeqarnerisa pingaartinneqarnissaat isumalluarfigigitsik, aalajangiussimallugulu inuttut atugarisanik siuarsaaniarnissaq inuuniarnermilu atugarisanik pitsaanerulersitsinissaq,

naalagaaffiit ilaasortaasut imminnut pisussanngortin*mmata* Naalagaaffiit Peqatigiit suleqatigalugit sulissutigissallugu inuttut pisinnaatitaaffinnut aammalu tunngavissiisutut killilersugaanngitsumik ilorsinnaatitaaffinnut nalinginnaasumik ataqqinninnerup annertunerulersinnissaa,

pisinnaatitaaffinnik aammalu killilersugaanngitsumik ilorsinnaatitaaffinnik tamakkuninnga peqatigilluni paasinninneq pisussaaffimmik tamatuminnga piviusunngortitsinissamut pingaaruteqarnerpaa*mmat*,

NAALAGAAFFIIT ILAASORTAASUT AALLARTITAASA

taamaattumik tamanut nalunaarutigaat inuttut pisinnaatitaaffiit pillugit silarsuarmioqatigiinnut nalunaarut manna inunnut tamanut inuiaqatigiinnullu tamanut tunngatillugu peqatigiinnikkut anguniagassatut, siunertaralugu inuup kialuunnit aammalu inuiaqatigiinnik sullissisut kikkulluunniit tamatigut Nalunaarutikkut maturnuuna anguniassagaat atuartitsineq perorsaanerlu aqqutigalugit pisinnaatitaaffinnut killilersugaanngitsumillu ilorsinnaatitaanernut tamakkununnga ataqqinninnerulernissaq aammalu inuiaqatigiit nunallu allat allartuinnartumik suliniuteqarnerisigut qularnaarniassallugu sumiluunniit sunniuteqartumik akuersaarneqarlutillu naammassiniarneqarnissaat, naalagaaffinni ilaasortaasuni innuttaasut aammalu sumiiffinni naalagaaffinnit taakkunannga aqunneqartuni innuttaasut akornanni.

Immikkoortoq 1.

Profile

Native Name
Inuktikut

Total Speakers
47,000 (1990)

Usage By Country
Official Language: Greenland together with Danish Home Speakers: also in Denmark

Background
It belongs to the Eskimo-Aleutian family and is spoken by over 50,000 people. The official language is based on West Greenlandic, a form of General Eskimo, and is classed as a separate language because of its official position and special literary development. Dialects are all inherently intelligible.

Received 5/25/1998
Posted 10/25/1998
Checked 11/12/1998

Arts & Humanities Research Council

CAMBRIDGE ASSESSMENT

UNIVERSITY OF CAMBRIDGE

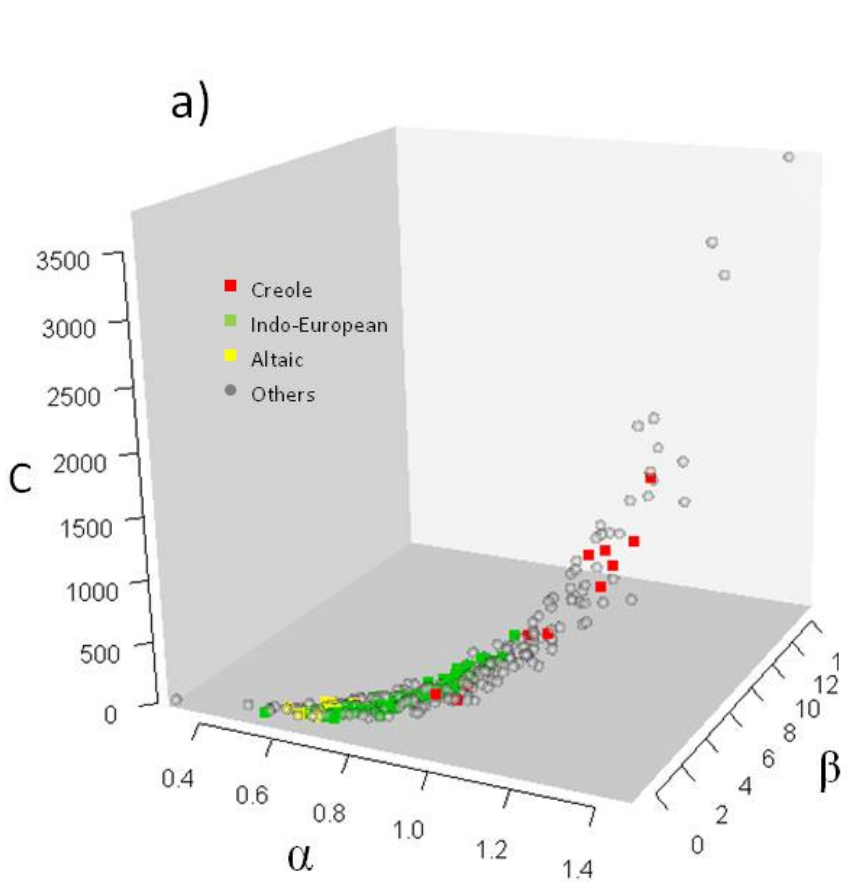
Lexical Diversity Space

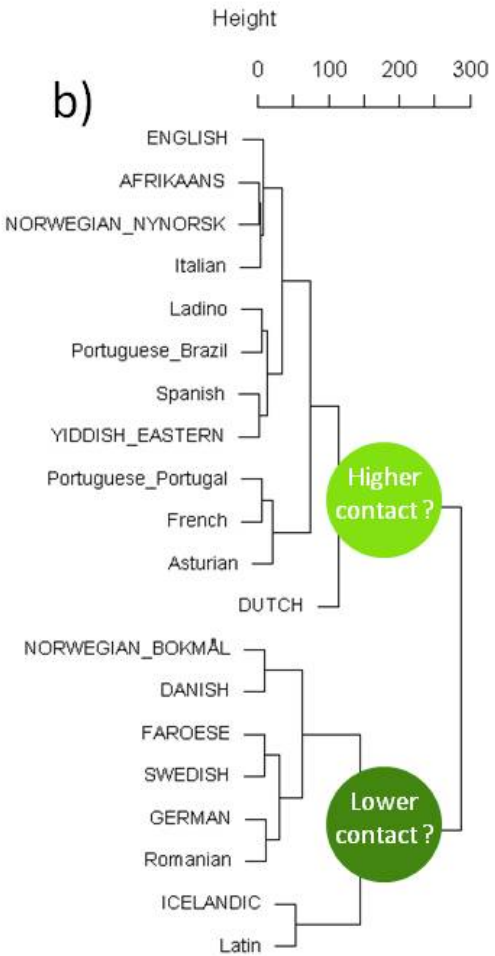
Calculating **Zipf-Mandelbrot parameters** for 359 languages (words delimited by white spaces)

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Lexical Diversity Space

Calculating **Zipf-Mandelbrot parameters** for 359 languages (words delimited by white spaces)





Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

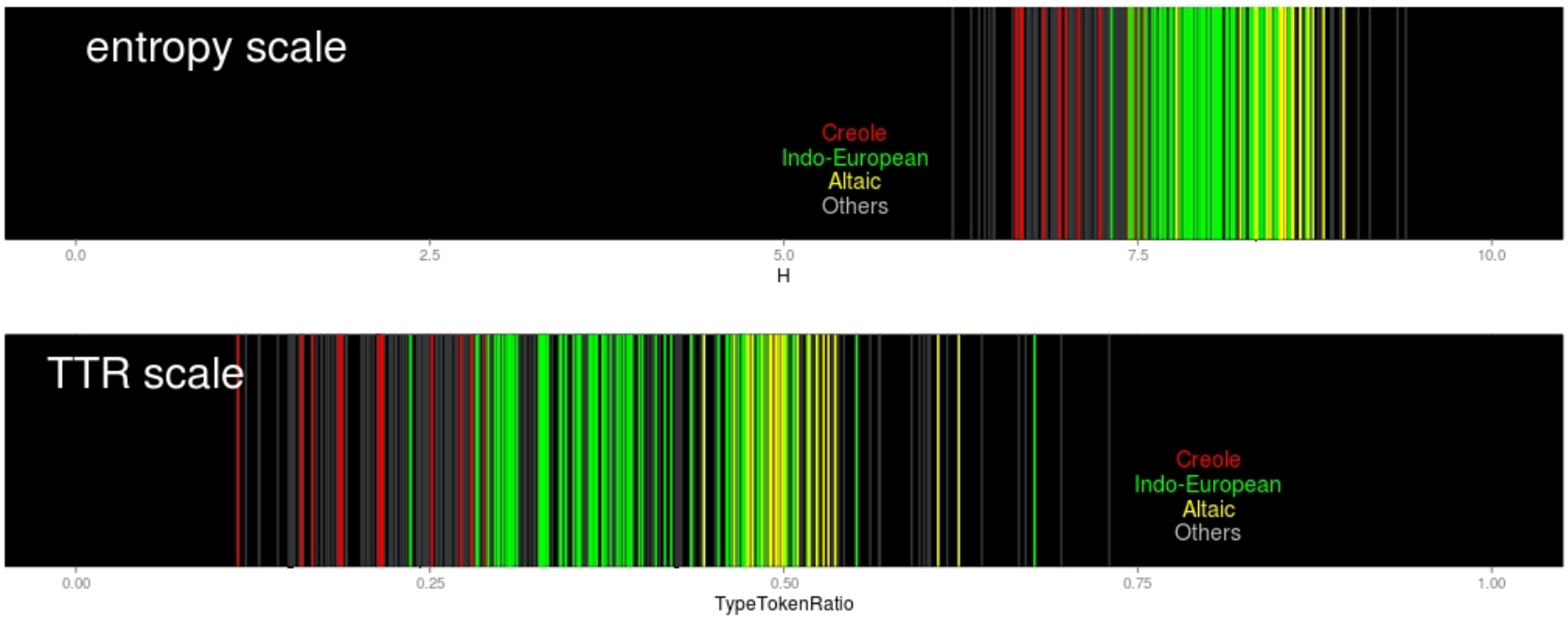
Lexical Diversity Scales

Calculating **Shannon entropy H** and **Type-Token Ratios** for 359 languages

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Lexical Diversity Scales

Calculating **Shannon entropy H** and **Type-Token Ratios** for 359 languages



Quantitative Analyses

Statistical Modelling of Lexical Diversity

Is it possible to predict LDTs of languages?

Bentz, Verkerk, Kiela, Hill & Buttery (submitted) Adpative languages: Modelling the co-evolution of population structure and lexical diversity

Quantitative Analyses

Statistical Modelling of Lexical Diversity

Is it possible to predict LDTs of languages?

Bentz, Verkerk, Kiela, Hill & Buttery (submitted) Adpative languages: Modelling the co-evolution of population structure and lexical diversity

Potential factors

Quantitative Analyses

Statistical Modelling of Lexical Diversity

Is it possible to predict LDTs of languages?

Bentz, Verkerk, Kiela, Hill & Buttery (submitted) Adpative languages: Modelling the co-evolution of population structure and lexical diversity

Potential factors

- language families

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Quantitative Analyses

Statistical Modelling of Lexical Diversity

Is it possible to predict LDTs of languages?

Bentz, Verkerk, Kiela, Hill & Buttery (submitted) Adpative languages: Modelling the co-evolution of population structure and lexical diversity

Potential factors

- language families
- language regions

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Quantitative Analyses

Statistical Modelling of Lexical Diversity

Is it possible to predict LDTs of languages?

Bentz, Verkerk, Kiela, Hill & Buttery (submitted) Adpative languages: Modelling the co-evolution of population structure and lexical diversity

Potential factors

- language families
- language regions
- writing systems

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Quantitative Analyses

Statistical Modelling of Lexical Diversity

Is it possible to predict LDTs of languages?

Bentz, Verkerk, Kiela, Hill & Buttery (submitted) Adpative languages: Modelling the co-evolution of population structure and lexical diversity

Potential factors

- language families
- language regions
- writing systems
- **non-native influence** (McWhorter, 2002, 2007, 2011; Wray & Grace, 2007; Lupyan & Dale 2010, 2012; Trudgill, 2011; Bentz & Winter, 2013)

Simple Linear Regression

- LDT measures (ZM-parameters, entropy H, TTR) for 359 language

Simple Linear Regression

- LDT measures (ZM-parameters, entropy H, TTR) for 359 language
- L2 ratio information on 231 languages

Simple Linear Regression

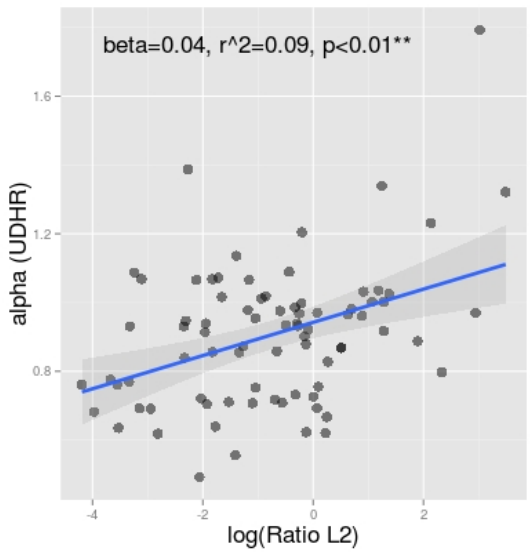
- LDT measures (ZM-parameters, entropy H, TTR) for 359 language
- L2 ratio information on 231 languages

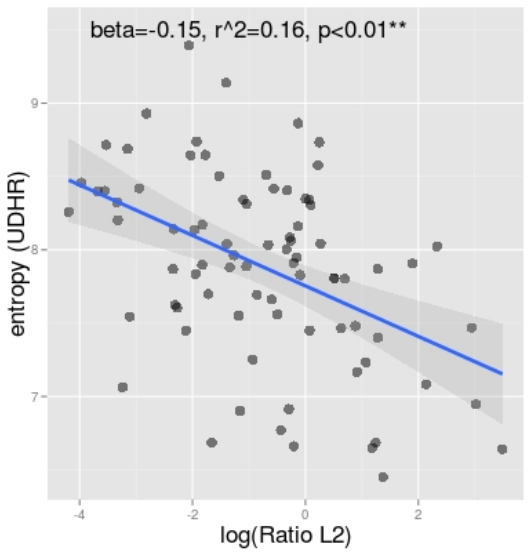
→ **Sample: 99 overlapping languages**

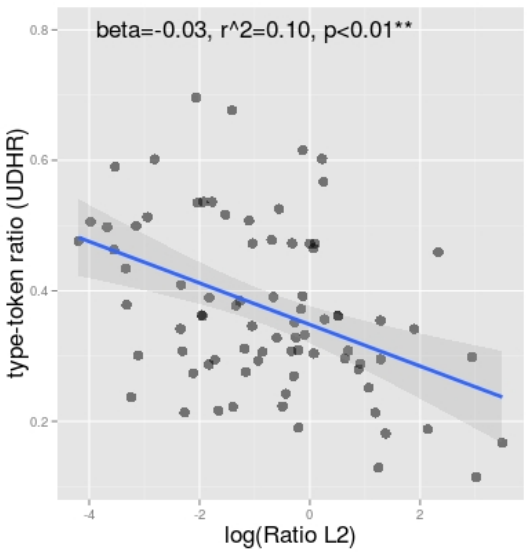
Simple Linear Regression

- LDT measures (ZM-parameters, entropy H, TTR) for 359 language
- L2 ratio information on 231 languages

→ **Sample: 99 overlapping languages**







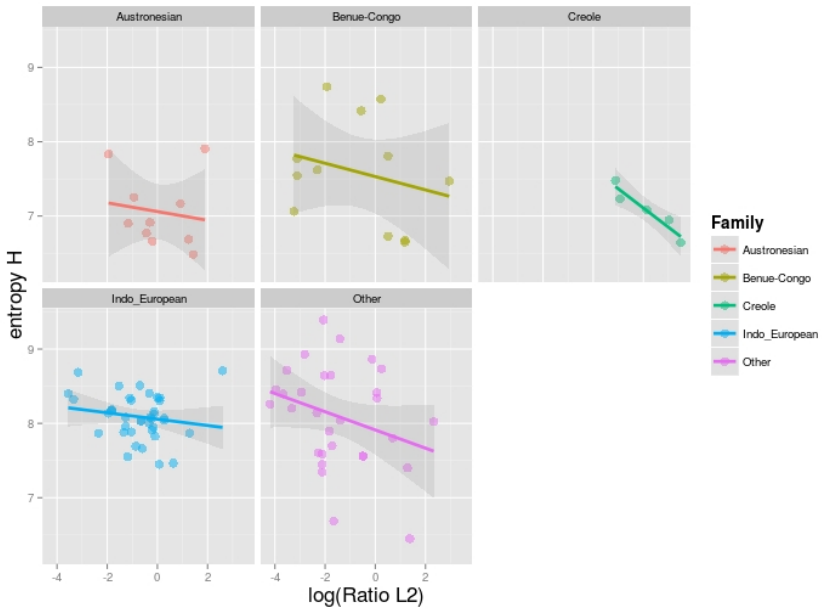
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

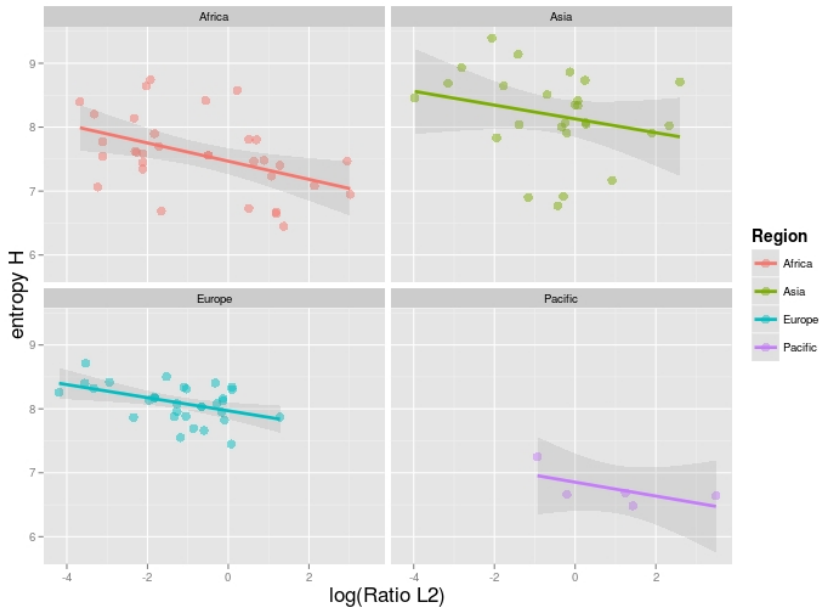
Mixed-effects regression

Regressing LDT on L2 ratios but controlling for non-independence of datapoints (language families, regions, writing systems)
Results: α ($\chi^2(1) = 1.38, p > 0.5$), H ($\chi^2(1) = 9.28, p < 0.001$), TTR ($\chi^2(1) = 7.11, p < 0.01$)

Mixed-effects regression

Regressing LDT on L2 ratios but controlling for non-independence of datapoints (language families, regions, writing systems)
Results: α ($\chi^2(1) = 1.38, p > 0.5$), H ($\chi^2(1) = 9.28, p < 0.001$), TTR ($\chi^2(1) = 7.11, p < 0.01$)





Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Interpretation/ Future directions

Questions

- Does that mean non-native speakers ruin languages?

Interpretation/ Future directions

Questions

- Does that mean non-native speakers ruin languages?
- Are there more and less efficient/expressive languages?

Interpretation/ Future directions

Questions

- Does that mean non-native speakers ruin languages?
- Are there more and less efficient/expressive languages?
- Are there simplex and complex languages?

Interpretation/ Future directions

Questions

- Does that mean non-native speakers ruin languages?
- Are there more and less efficient/expressive languages?
- Are there simplex and complex languages?

Suggestions

- Languages are **complex adaptive systems** shaped by the cognitive and conceptual needs of their speakers (Croft, 2000; Ritt, 2004; Christiansen & Chater, 2008; Beckner et al., 2009)
- A lack of lexical diversity might be made up for by encoding of information at a different level (constructions, fixed word order, multi word expressions)

Lexical Diversity and Word Order

- **Permutation entropy**: Reflects the mutual dependence of words in ngrams, i.e. word sequences (Zhang et al. 2006; Ramisch et al. 2008)

Lexical Diversity and Word Order

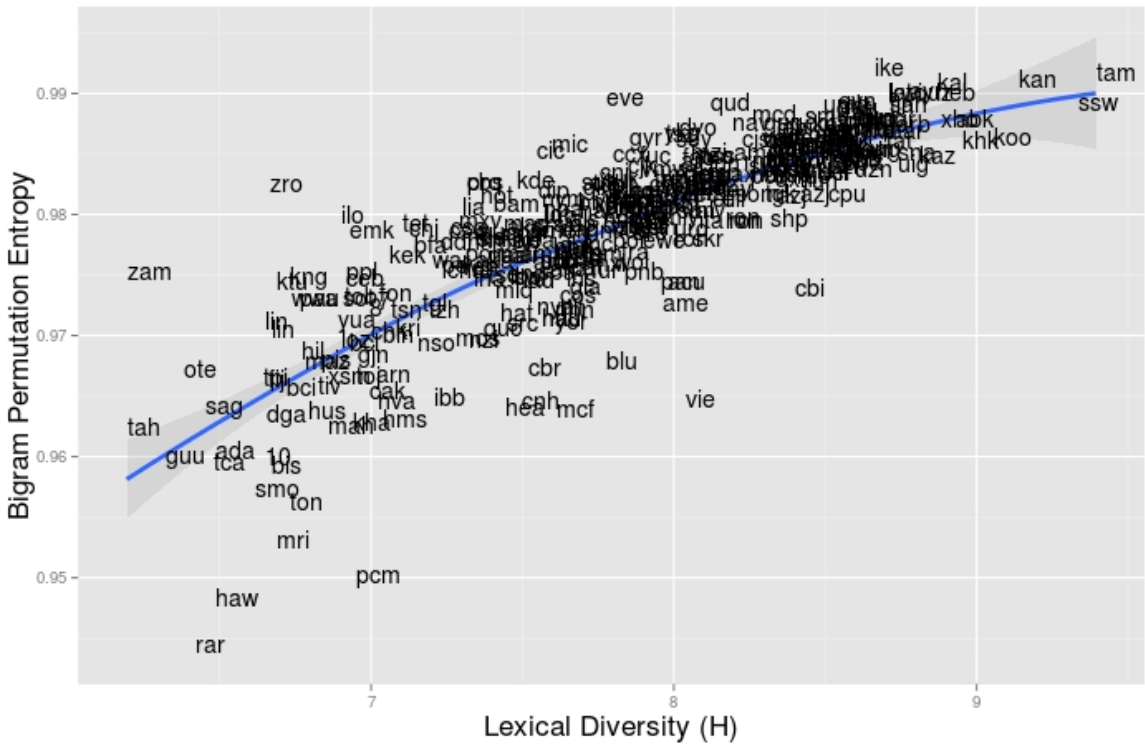
- **Permutation entropy**: Reflects the mutual dependence of words in ngrams, i.e. word sequences (Zhang et al. 2006; Ramisch et al. 2008)
- PE $= 0 \rightarrow$ words are completely **dependent (fixed order)**

Lexical Diversity and Word Order

- **Permutation entropy**: Reflects the mutual dependence of words in ngrams, i.e. word sequences (Zhang et al. 2006; Ramisch et al. 2008)
- PE $= 0 \rightarrow$ words are completely **dependent (fixed order)**
- PE $= 1 \rightarrow$ words are completely **independent (free order)**

Lexical Diversity and Word Order

- **Permutation entropy**: Reflects the mutual dependence of words in ngrams, i.e. word sequences (Zhang et al. 2006; Ramisch et al. 2008)
- PE $= 0 \rightarrow$ words are completely **dependent (fixed order)**
- PE $= 1 \rightarrow$ words are completely **independent (free order)**



Conclusion

Lexical Diversity (LDT)

- Reflects **information encoding** strategies (inflection, compounding, base vocabulary)

Conclusion

Lexical Diversity (LDT)

- Reflects **information encoding** strategies (inflection, compounding, base vocabulary)
- It can be measured **quantitatively** (ZM-law, Shannon entropy, type-token ratios)

Conclusion

Lexical Diversity (LDT)

- Reflects **information encoding** strategies (inflection, compounding, base vocabulary)
- It can be measured **quantitatively** (ZM-law, Shannon entropy, type-token ratios)
- It is linked to language learning via **frequency effects**

Conclusion

Lexical Diversity (LDT)

- Reflects **information encoding** strategies (inflection, compounding, base vocabulary)
- It can be measured **quantitatively** (ZM-law, Shannon entropy, type-token ratios)
- It is linked to language learning via **frequency effects**
- It can be **statistically predicted** across languages (language families, language regions, writing systems, potential non-native influence)

Collaborators

Douwe Kiela

Felix Hill

Andrew Caines

Dimitrios Alikaniotis

Paula Buttery

Arts & Humanities Research Council

CAMBRIDGE ASSESSMENT

UNIVERSITY OF CAMBRIDGE

Thank You!

¡chris@christianbentz.de¿