Causality in historical language change

Christian Bentz
cb696@cam.ac.uk

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Outline

Part 1: Theory

- Classical theories and the **internal/external** distinction
- Adaptive systems and the **causal/referential** distinction

Part 2: Analyses

- **Lexical diveristy** and L2 speaker proportions

Problems

- inferring causality
- information encoding efficiency of languages
- etc.

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

PART I

The distinction of **internal** and **external** causes of change

- **Sociolinguistics** (Croft, 2000; Jones & Esch, 2002; Jones & Singh, 2005)
- **Genetic Linguistics** (Thomason & Kaufman, 1988)
- **Principles & Parameters** (Briscoe, 2000a, 2000b; Clark & Roberts, 1993; Lightfoot, 1979; Pintzuk, Tsoulas, & Warner, 2000; Yang, 2000)

Sociolinguistics and Genetic Linguistics

- **internal**: languages follow 'natural','normal' and 'regular' paths of change, according to general principles such as *assimilation*, *analogical extension* and *analogical leveling* (Thomason & Kaufman, 1988: 22pp.; Jones & Singh 2005: 18-19)
- **external**: language contact, i.e. child bilingualism or adult second language learning (L2)

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Examples

- **internal:** OE stana**s**, scip**u**, sorg**a**, nam**an** $\rightarrow$ PDE stone**s**, ship**s**, sorrow**s**, name**s** (analogical extension)
- **external**: OE pronouns replaced by Old Norse pronouns hie/heo, him/hira, heom/heora $\rightarrow$ θeir, θeim, θeira (borrowing)

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Principle & Parameter view

- **internal**: innate set of parameters (UG) that limits the space of possible grammars (Yang, 2000: 232; Clark & Roberts, 1993: 340; Biberauer, Holmberg, Roberts, & Sheehan, 2010)
- **external**: refers to the varying language input (causing parameter setting) during acquisition

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Problems (see Jones & Singh, 2005: 25-26)

- **Sociolinguistics/Genetic Linguistics**: Explain *what* is happening in great detail, but not *why* it is happening (i.e. causation). How do 'internal' causes work, what are the triggering events?
- **Principles & Parameters**: UG as 'internal' cause is too broad, i.e. it is by definition universal across all languages, and does not predict anything specific about language change

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

The CAS Model

Language as a Complex Adaptive System

"The **structures of language** emerge from interrelated patterns of experience, **social interaction**, and **cognitive mechanisms**." (Beckner et al., 2009)

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

The CAS Model

Language as a Complex Adaptive System

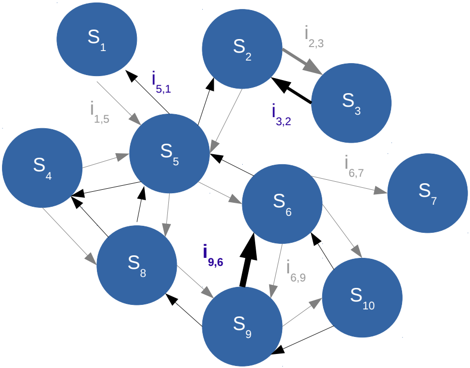
"The **structures of language** emerge from interrelated patterns of experience, **social interaction**, and **cognitive mechanisms**." (Beckner et al., 2009)

Earlier studies

Gell-Mann, 1992; Croft, 2000; Kirby & Hurford, 2002; Ritt, 2004; Christiansen & Chater, 2008

The CAS Model

S: Speaker
i: linguistic interaction
c: linguistic competence



Definitions:

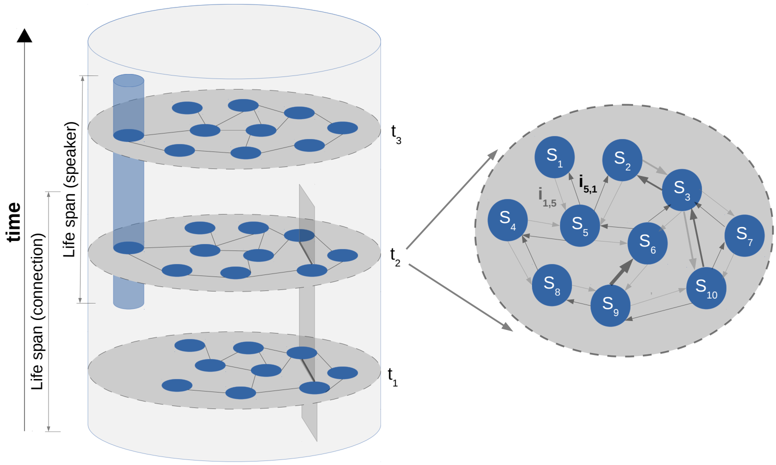
Population
$P = \{S_1, S_2, \ldots, S_{10}\}$

Language (accumulative):
$L_{acc} = (i_{1,2}, i_{2,1}, \ldots, i_{10,9})$

Language (minimalist):
$L_{min} = c_1 \cap c_2 \cap \ldots \cap c_{10}$

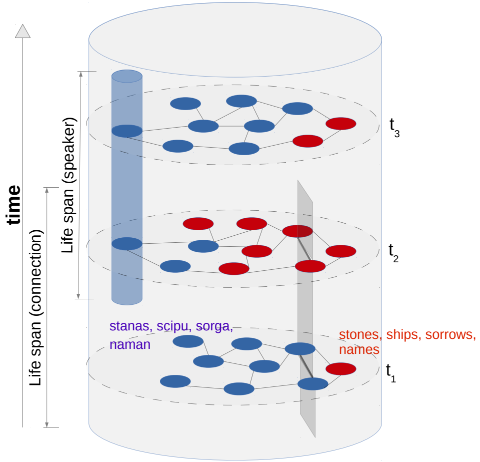
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

The CAS Model



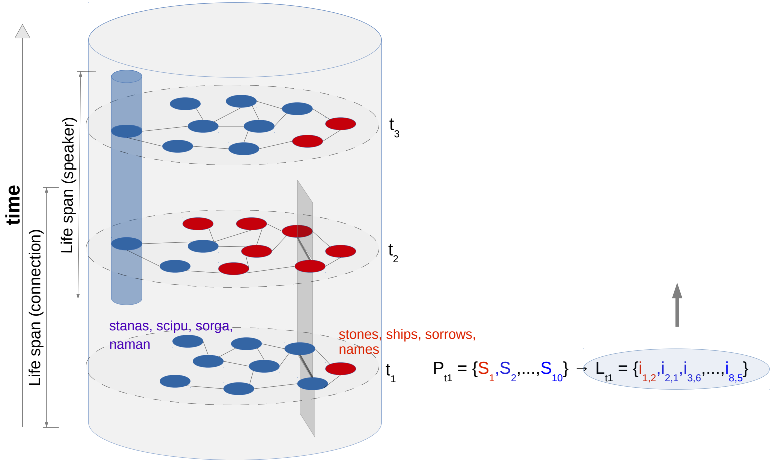
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Language Change in the CAS Model



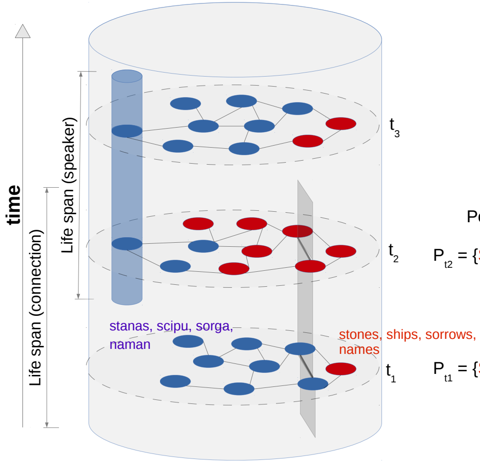
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Language Change in the CAS Model



Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

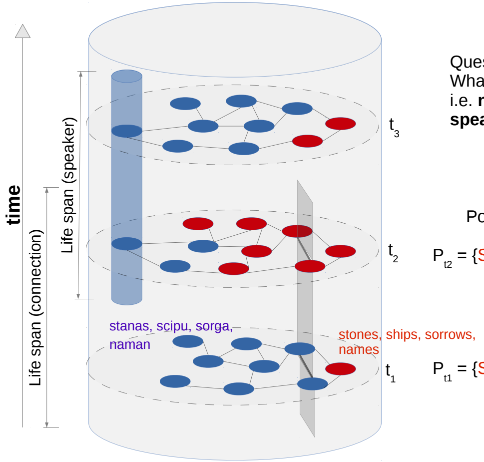
Language Change in the CAS Model



Population	Language
$P_{t2} = \{S_1, S_2, \ldots, S_{10}\} \rightarrow$	$L_{t2} = \{i_{1,2}, i_{2,1}, \ldots, i_{10,9}\}$
$P_{t1} = \{S_1, S_2, \ldots, S_{10}\} \rightarrow$	$L_{t1} = \{i_{1,2}, i_{2,1}, i_{3,6}, \ldots, i_{8,5}\}$

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Language Change in the CAS Model



Question:
What about the **internal/external**, i.e. **native speaker/non-native speaker** distinction?

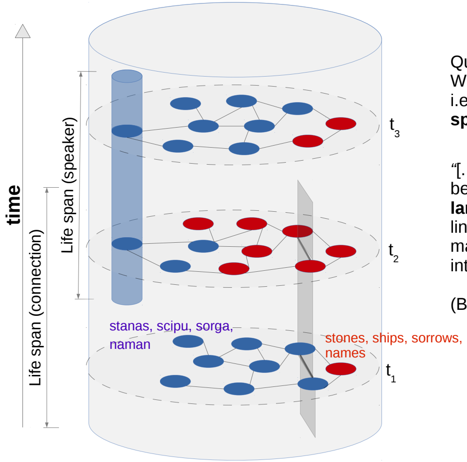
Population | Language

$P_{t2} = \{S_1, S_2, \ldots, S_{10}\} \rightarrow L_{t2} = \{i_{1,2}, i_{2,1}, \ldots, i_{10,9}\}$

$P_{t1} = \{S_1, S_2, \ldots, S_{10}\} \rightarrow L_{t1} = \{i_{1,2}, i_{2,1}, i_{3,6}, \ldots, i_{8,5}\}$

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Language Change in the CAS Model



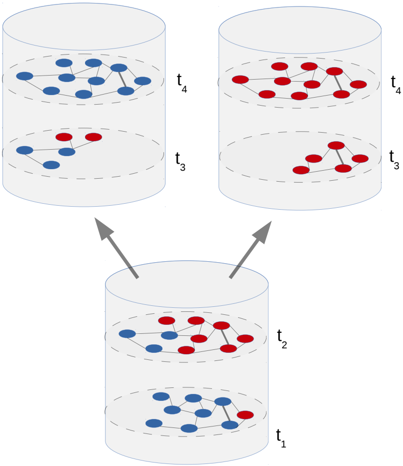
Question:
What about the **internal/external**, i.e. **native speaker/non-native speaker** distinction?

"[...] the traditional distinction between **language-external** and **language-internal causes** for linguistic change and evolution may turn out to be of little interest in the end."

(Bickel, 2013: 13)

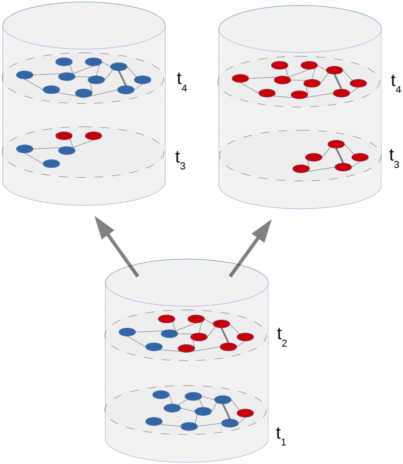
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Population drift



Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Population drift

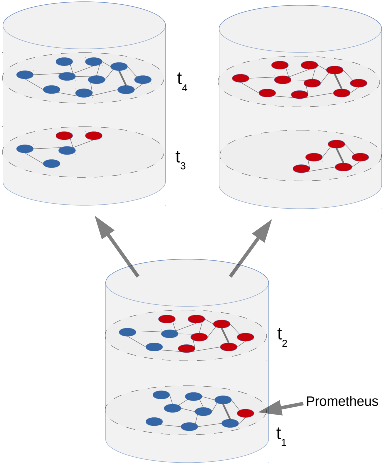


Question:

Is the population drift the **cause** for Language A to use *stanas, scipu, sorga, naman*, and Language B to use *stones, ships, sorrows and names*?

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Population drift



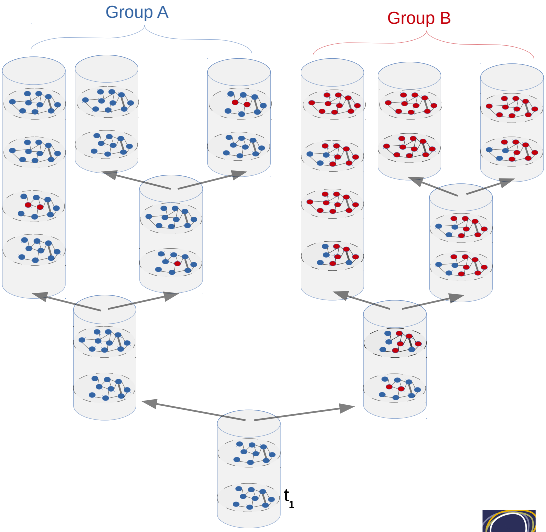
Question:

Is the population drift the **cause** for Language A to use *stanas, scipu, sorga, naman,* and Language B to use *stones, ships, sorrows and names*?

(No). The population drift might be an amplifier for the change but not the cause.

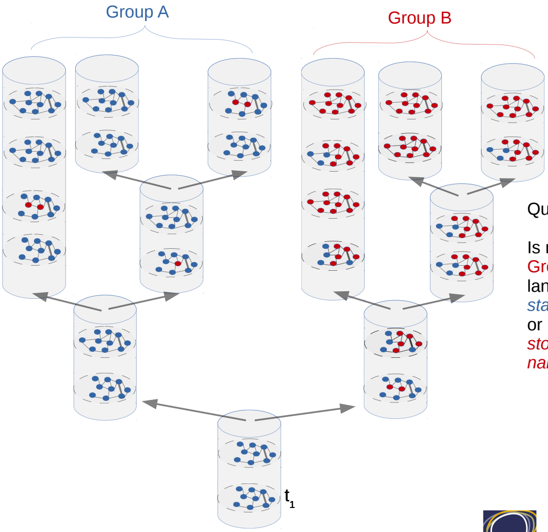
Arts & Humanities Research Council CAMBRIDGE ASSESSMENT UNIVERSITY OF CAMBRIDGE

Phylogenetic Groups



Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Phylogenetic Groups

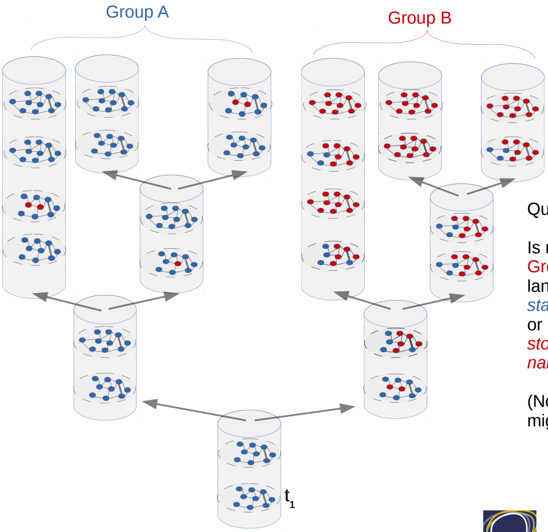


Question:

Is membership in Group A or Group B the **cause** for languages to use either *stanas, scipu, sorga, naman*, or *stones, ships, sorrows and names*?

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Phylogenetic Groups



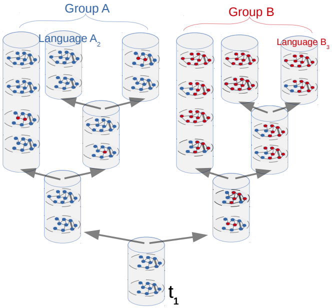
Question:

Is membership in Group A or Group B the **cause** for languages to use either *stanas, scipu, sorga, naman*, or *stones, ships, sorrows and names*?

(No). However, membership might predict usage.

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Implications of the CAS Model



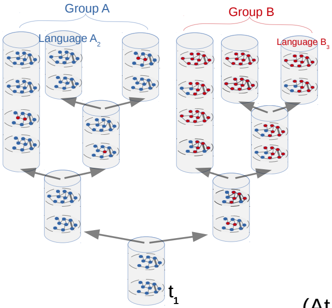
Implications for statistical modelling:

Why is a specific feature (to a specific extent) present in a language or not?

Example:
Why is the -s plural more strongly represented in Language B_3 than in Language A_2?

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Implications of the CAS Model



Implications for statistical modelling:

Why is a specific feature (to a specific extent) present in a language or not?

Example:
Why is the -s plural more strongly represented in Language B_3 than in Language A_2?

(At least) two kinds of predictors:

primary (causal) predictors:
- processing biases/ constraints of the speakers/hearers

secondary (relational) predictors:
- language family, genus, regions, etc.

The CAS Model

Conclusions

- the distinction between **internal** and **external** causes of change is somewhat misleading
- within a CAS account it might make more sense to think about **primary (causal)** and **secondary (relational)** predictors.

PART II ANALYSES

What can we predict about languages using the CAS model?

- We need quantitative, cross-linguistic data that reflect the structures of languages we are interested in

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

PART II ANALYSES

What can we predict about languages using the CAS model?

- We need quantitative, cross-linguistic data that reflect the structures of languages we are interested in
- We need quantitative, cross-linguistic data that reflect potential processing/comprehension constraints of speakers

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Examples in earlier studies

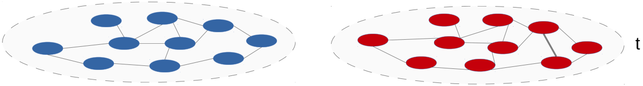
Qualitative hypothesis

- Higher proportions of non-native speakers tend to simplify morphology (Wray& Grace, 2007; McWhorter, 2002, 2007; Trudgill, 2011)

Quantitative evidence

- Bigger language populations $\rightarrow$ less morphological elaboration (Lupyan& Dale 2010, 2012)
- More non-native speakers $\rightarrow$ less case marking (Bentz& Winter, 2012, 2013)

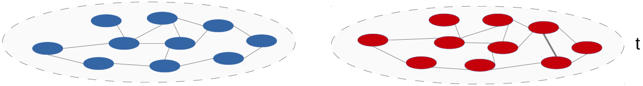
Language comparison in the CAS model



$L^{A}_{t} = \{i_{1,2}, i_{2,1}, i_{1,3}, \ldots, i_{j,j-1}\}$ $L^{B}_{t} = \{i_{1,2}, i_{2,1}, i_{1,3}, \ldots, i_{j,j-1}\}$

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Language comparison in the CAS model



$L^{A}_{t} = \{i_{1,2}, i_{2,1}, i_{1,3}, \ldots, i_{j,j-1}\}$ $L^{B}_{t} = \{i_{1,2}, i_{2,1}, i_{1,3}, \ldots, i_{j,j-1}\}$

This is impossible.

However, it is possible to sample from these sets and (roughly) approximate them.

Important: Keep the content of the interactions **constant**!

↓

Parallel texts

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Zipfian/Information theoretic approach: How do **word form distributions (lexical diversities)** differ across languages, considering that the **content is constant**?

Data: Parallel texts

Zipfian/Information theoretic approach: How do **word form distributions (lexical diversities)** differ across languages, considering that the **content is constant**?

Data: Parallel texts

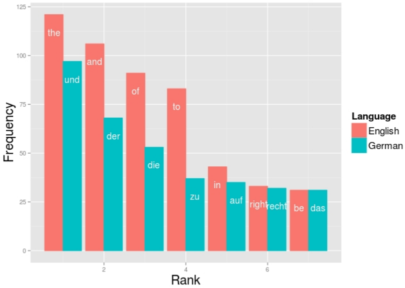
- **Parallel Bible Corpus** (810 languages, ca. 20000 words per language)
- **Universal Declaration of Human Rights** (376 languages, ca. 2000 words per language)
- **European Parliament Corpus** (21 languages, ca. 7 million words per language)

Zipfian approach: Analysis of word form distributions across languages (lexical diversities)

Method: Order types, i.e. word forms delimited by white spaces and non-word characters (see Haspelmath 2011 and Wray 2014 for critical review), according to their token frequencies (Zipf,1932,1949)

Zipfian approach: Analysis of word form distributions across languages (lexical diversities)

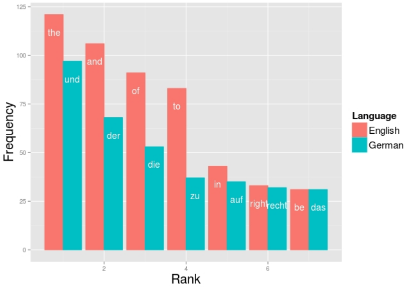
Method: Order types, i.e. word forms delimited by white spaces and non-word characters (see Haspelmath 2011 and Wray 2014 for critical review), according to their token frequencies (Zipf,1932,1949)

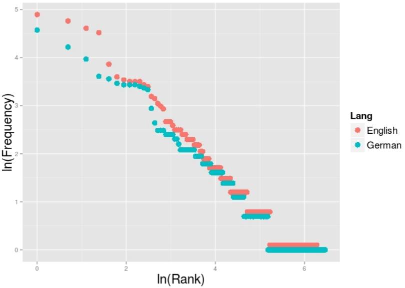


Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Zipfian approach: Analysis of word form distributions across languages (lexical diversities)

Method: Order types, i.e. word forms delimited by white spaces and non-word characters (see Haspelmath 2011 and Wray 2014 for critical review), according to their token frequencies (Zipf,1932,1949)





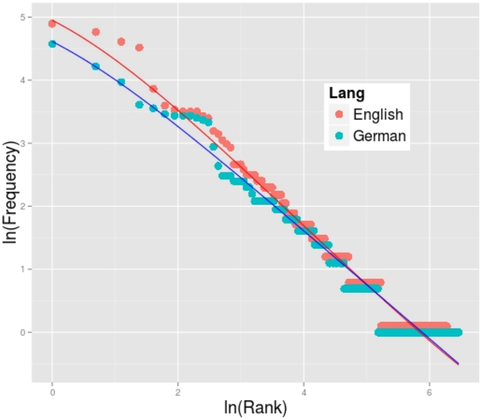
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Quantitative measures

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

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

$$i = 1, 2, \ldots, n$$



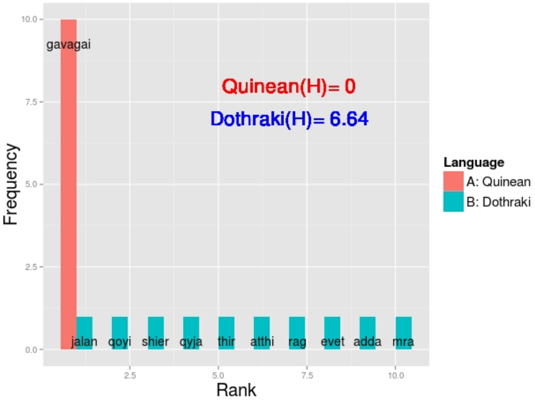
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Quantitative measures

Shannon entropy (Shannon & Weaver, 1949)

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

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

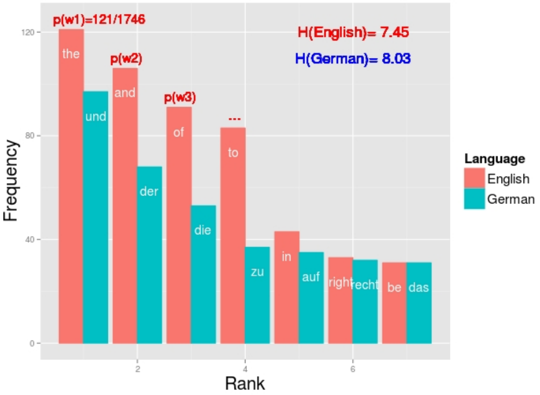


Quantitative measures

Shannon entropy (Shannon & Weaver, 1949)

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

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

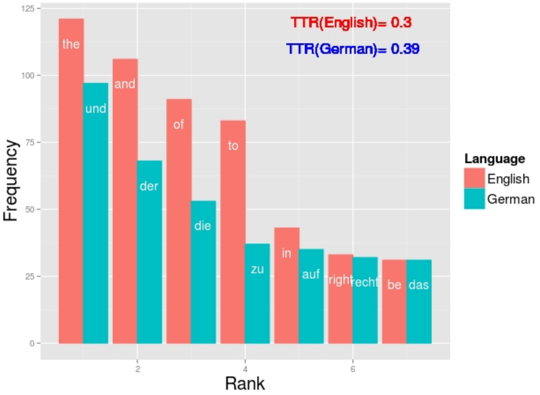


Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Quantitative measures

Type-Token Ratio (TTR)

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



Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Scaled values for 647 languages of 83 families (PBC, UDHR, EPC)

Altaic
Indo-European
Creole

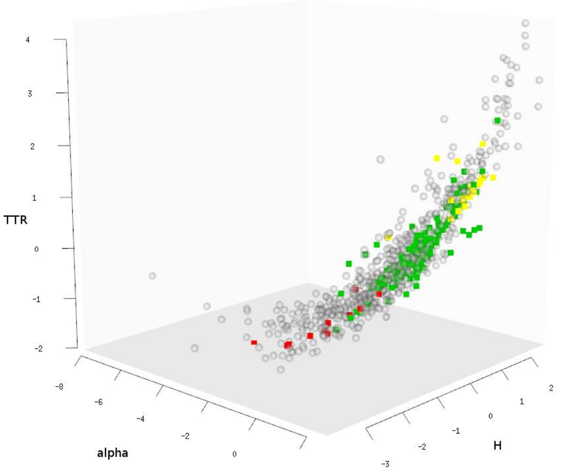
Bentz, Verkerk, Kiela, Hill & Buttery (submitted)

Lexical Diversity Space

Scaled values for 647 languages of 83 families (PBC, UDHR, EPC)

Altaic
Indo-European
Creole

Bentz, Verkerk, Kiela, Hill & Buttery (submitted)



Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

What causes languages to have higher/lower LDTs?

Hypothesis

- Languages with **higher lexical diversities** might be those languages with lower non-native speaker proportions.

What causes languages to have higher/lower LDTs?

Hypothesis

- Languages with **higher lexical diversities** might be those languages with lower non-native speaker proportions.
- Potential causal link: There is evidence in applied linguistics that lexical diversity is systematically lower for L2 speakers (Jarvis 2002, Treffers-Daller 2013).

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Statistical Model

Predicting lexical diversity from L2 proportions requires a **linear mixed-effects model** (Baayen et al., 2008; Bates et al., 2014; Jäger et al., 2011)

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Statistical Model

Predicting lexical diversity from L2 proportions requires a **linear mixed-effects model** (Baayen et al., 2008; Bates et al., 2014; Jäger et al., 2011)

- continuous dependent/outcome variable: LDTs scaled

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

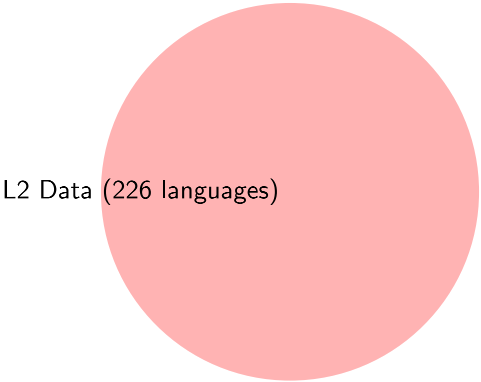
Statistical Model

Predicting lexical diversity from L2 proportions requires a **linear mixed-effects model** (Baayen et al., 2008; Bates et al., 2014; Jäger et al., 2011)

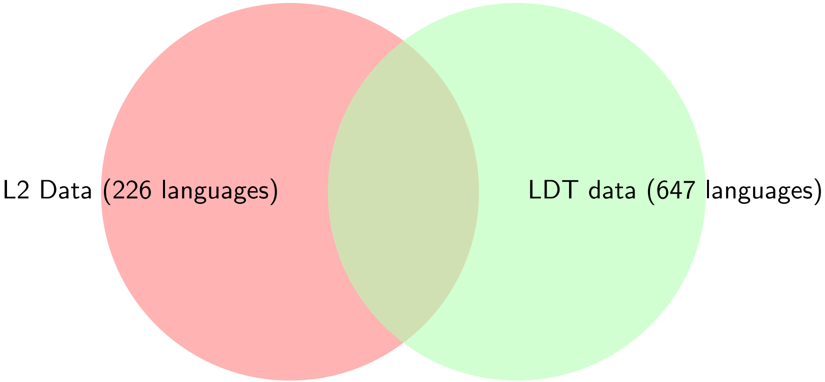
- continuous dependent/outcome variable: LDTs scaled
- (potentially) **causal predictor**:
  L2 proportions as (fixed effect)
- **referential predictors** (random effects), accounting for non-independence of data points (family, region, text type, LDT measure)

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

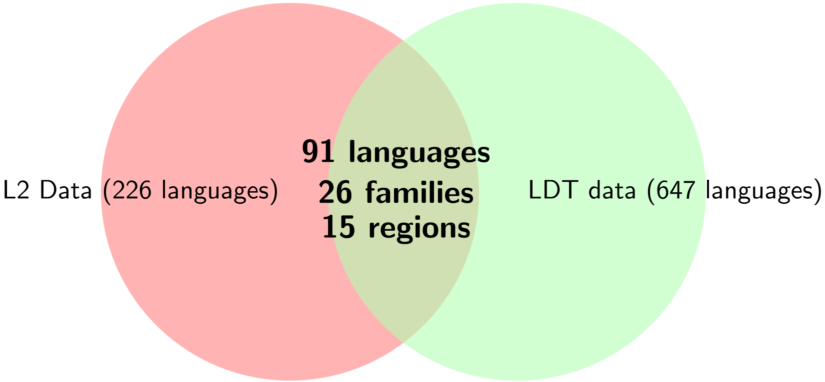
Statistical Model: Data Overlap



Statistical Model: Data Overlap



Statistical Model: Data Overlap

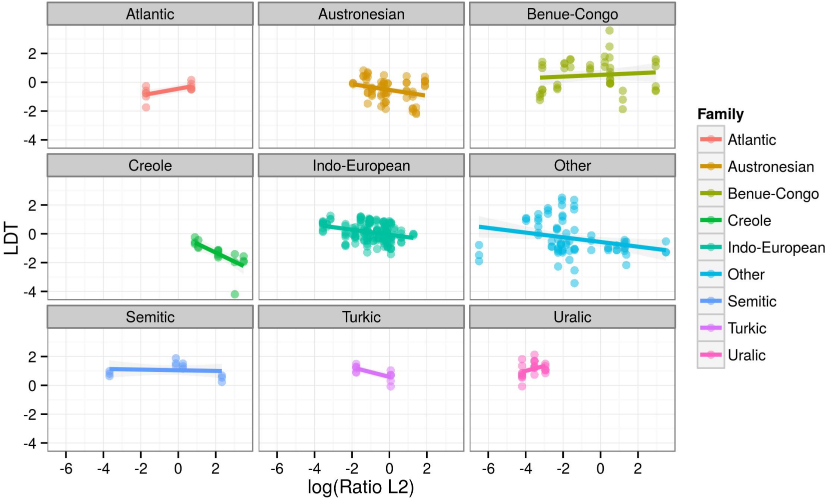


Results

Dependent	Fixed	Random	β	SE	p-value
LDT scaled	log(L2/L1)	family	-0.2772	0.1329	0.0375
		region			
		measure			
		text type			
		ISO code			

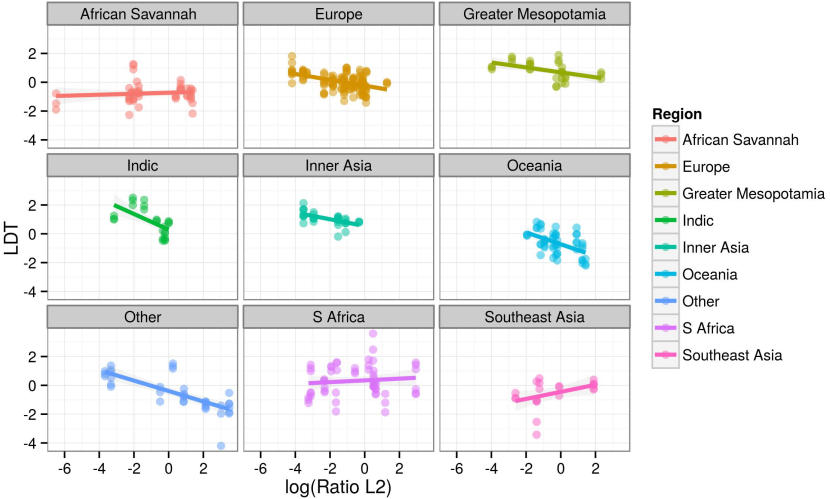
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

LDT and L2 proportions across families



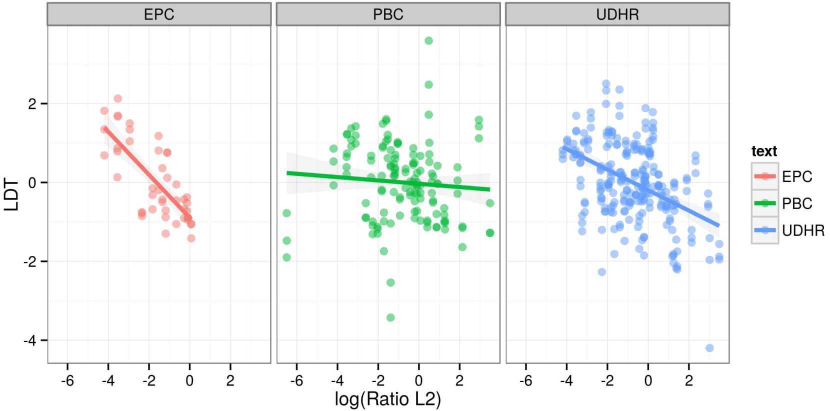
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

LDT and L2 proportions across regions



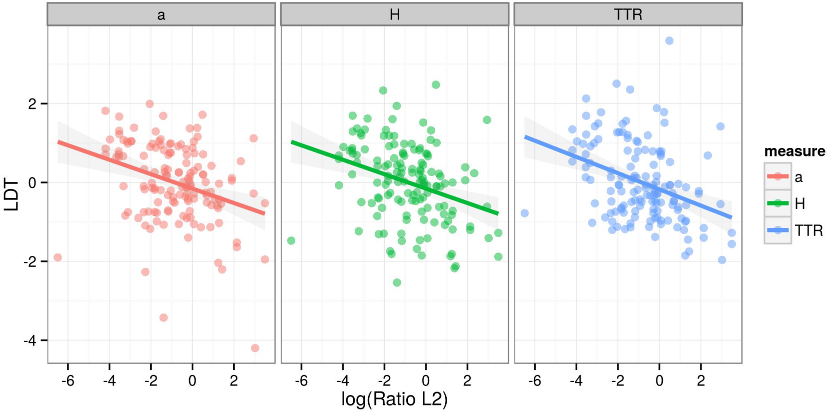
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

LDT and L2 proportions across text types



Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

LDT and L2 proportions across measures



Lexical diversity: Conclusions

- Languages with more non-native speakers tend to have *lower* lexical diversity

Lexical diversity: Conclusions

- Languages with more non-native speakers tend to have *lower* lexical diversity
- These trends hold across *most* families, regions, text types and the LDT measures used

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Problem: Causality

- The mixed-effects model still *does not* prove that the causality runs from L2 proportions to language structure

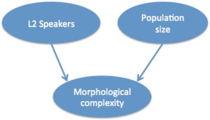
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Problem: Causality

- The mixed-effects model still *does not* prove that the causality runs from L2 proportions to language structure

Suggestion

- Sean Roberts did a preliminary causal graph (Nihat Ay's talk) analysis of an earlier dataset (Bentz & Winter 2013) and found some evidence for an L2 to language structure causality. (http://www.replicatedtypo.com)



Arts & Humanities Research Council

CAMBRIDGE ASSESSMENT

UNIVERSITY OF CAMBRIDGE

Problem: Encoding efficiency

- Are some languages **more/less efficient** at encoding information?

Problem: Encoding efficiency

- Are some languages **more/less efficient** at encoding information?

Suggestion

- 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)
- Fermin's talk (see also Mosocoso del Prado 2011, Ehret & Szmrecsanyi (to appear))

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Collaborators

Douwe Kiela

Felix Hill

Andrew Caines

Annemarie Verkerk

Paula Buttery

Arts & Humanities Research Council

CAMBRIDGE ASSESSMENT

UNIVERSITY OF CAMBRIDGE

Thank You!

¡chris@christianbentz.de¿