Towards measuring and modelling the (potential) impact of non-native speakers on language structures

Christian Bentz
cb696@cam.ac.uk

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Outline

Background

- Language as a **Complex Adaptive System**
- **Non-native speakers (L2)** as drivers of language change

Statistical Modeling

- **Case marking** and L2 speaker proportions
- **Lexical diversity** and L2 speaker proportions

Conclusions

- Problems and future directions

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

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)

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)

Linguistic Niche Hypothesis

"The level of **morphological specification** is a product of languages adapting to the learning constraints [...] of the speaker population. Complex morphological paradigms [...] present particular learning challenges for **adult learners** [...]"
(Lupyan & Dale, 2010)

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)

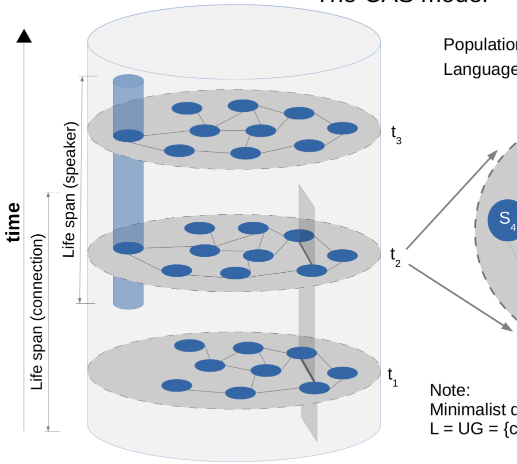
Linguistic Niche Hypothesis

"The level of **morphological specification** is a product of languages adapting to the learning constraints [...] of the speaker population. Complex morphological paradigms [...] present particular learning challenges for **adult learners** [...]"
(Lupyan & Dale, 2010)

Earlier studies

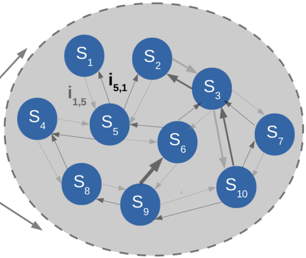
Gell-Mann, 1992; Croft, 2000; Kirby & Hurford, 2002; Ritt, 2004; Christiansen & Chater, 2008

The CAS model



Population: $P_{t2} = \{S_1, S_2, S_3, \ldots, S_8\}$

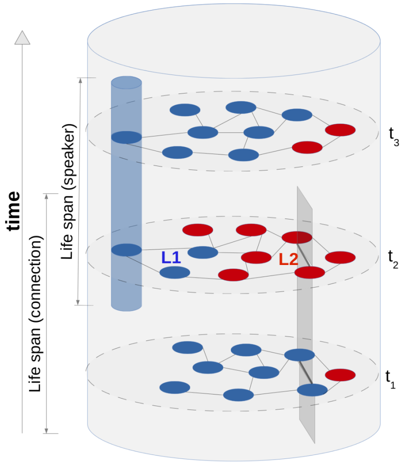
Language: $L_{t2} = \{i_{1,5}, i_{2,5}, i_{3,6}, \ldots, i_{8,5}\}$



Note:
Minimalist definition
$L = UG = \{c_1 \cap c_2 \cap c_3 \cap \ldots \cap c_n\}$

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Language contact in the CAS model



Prediction of the CAS model:

Population	Language
$P_{t1} = \{S_1, S_2, S_3, \ldots, S_8\} \rightarrow$	$L_{t2} = \{i_{1,5}, i_{2,5}, i_{3,6}, \ldots, i_{8,5}\}$
$P_{t1} = \{S_1, S_2, S_3, \ldots, S_7\} \rightarrow$	$L_{t1} = \{i_{1,5}, i_{2,5}, i_{3,6}, \ldots, i_{8,5}\}$

Collecting L2 Data

Project with Søren Wichmann, Bodo Winter
(at MPI for Evolutionary Anthropology)

Max Planck Institute
for Evolutionary Anthropology

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Collecting L2 Data

Project with Søren Wichmann, Bodo Winter
(at MPI for Evolutionary Anthropology)

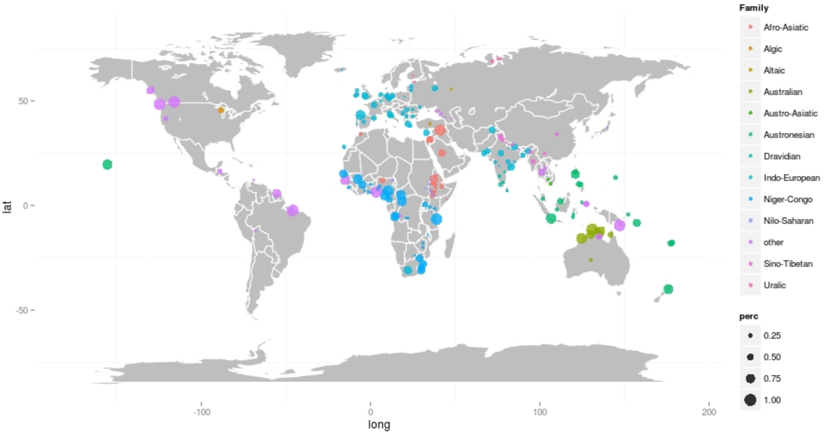
Max Planck Institute
for Evolutionary Anthropology

Dataset of L2 and L1 numbers for **231 languages** (56 families, 27 regions)

Language	SILCode	Stock(Autotyp)	Region(Au	FAM(WALS	Genus(WALS)	L1 Ethnologue	L1 Encarta	Other	NativeSpeak(	L2 Ethnologue	L2 Others	L2 Estimation	L2Ratio
Kutenai	kut	Kutenai	Basin and	Ktn	Kutenai	12	NA	NA	12	s 1990 Canada+USA: ~310		310	25.83333333
Kongo	kon	Benue-Congo	S Africa	Niger-Congo, Atlantic-Co		5955908	NA	NA	5955908	5000000	NA	5000000	0.839502558
Aari	aiw	Omotic	Greater Ab	AA	South Omotic	155000	NA	NA	155000	13319	NA	13319	0.085929032
Afar	aar	Cushitic	Greater Ab	AA	Eastern Cushi	1078200	NA	1.4 mi	1239100	22848	NA	22848	0.01843919
Alaba-K'abeena	alw	Cushitic	Greater Ab	AA	Eastern Cushi	162000	NA	NA	162000	29699	NA	29699	0.18332716
Amharic	amh	Semitic	Greater Ab	AA	Semitic	17528500	17400000	Officia	17464250	4000000	7000000	5500000	0.314929069
Arabic	arb	Semitic	N Africa	AA	Semitic	221000000	150000000	206,0	192300000	246000000	NA	246000000	1.27925117
Arabic, Algerian (	arq	Semitic	N Africa	AA	Semitic	22397000	NA	NA	22397000	3000000	NA	3000000	0.133946511
Arabic, southern	pga	Semitic	N Africa	AA	Semitic	20000	NA	NA	20000	44000	NA	44000	2.2
Arbore	arv	Cushitic	Greater Ab	AA	Eastern Cushi	4440	NA	NA	4440	3108	NA	3108	0.7
Argobba	agj	Semitic	Greater Ab	AA	Semitic	10900	NA	NA	10900	3236	NA	3236	0.296880734
Awngi	awn	Cushitic	Greater Ab	AA	Central Cushi	500000	NA	###	428490	64425	NA	64425	0.150353567
Basketo	bst	Omotic	Greater Ab	AA	North Omotic	57800	NA	NA	57800	8961	NA	8961	0.155034602
Bench (Gimira)	bcq	Omotic	Greater Ab	AA	North Omotic	174000	NA	NA	174000	22640	NA	22640	0.130114943
Borna (Shinassha	bwo	Omotic	Greater Ab	AA	North Omotic	19900	NA	NA	19900	2276	NA	2276	0.114371859
Bussa	dox	Cushitic	Greater Ab	AA	Eastern Cushi	6620	NA	NA	6620	920	NA	920	0.13897281
Dime Dima	dim	Omotic	Greater Ab	AA	South Omotic	6500	NA	NA	6500	529	NA	529	0.081384615
Dirasha (Gidole)	gdl	Cushitic	Greater Ab	AA	Eastern Cushi	90000	NA	NA	90000	7000	NA	7000	0.077777778
Dizi	mdx	Omotic	Greater Ab	AA	North Omotic	21100	NA	NA	21100	2054	NA	2054	0.097345972
Dorze	doz	Omotic	Greater Ab	AA	North Omotic	20800	NA	NA	20800	3597	NA	3597	0.172932692
Gamo-Gofa-Dawr	gmo	Omotic	Greater Ab	AA	North Omotic	1240000	NA	NA	1240000	77883	NA	77883	0.062808871
Gawwada (Dullay	gwd	Cushitic	Greater Ab	AA	Eastern Cushi	32700	NA	NA	32700	1367	NA	1367	0.041804281
Gedeo Darasa	drs	Cushitic	Greater Ab	AA	Eastern Cushi	637000	NA	NA	637000	47950	NA	47950	0.075274725
HadiyyaAdea	hdy	Cushitic	Greater Ab	AA	Eastern Cushi	924000	NA	NA	924000	15889	NA	15889	0.017195887
Hamer-Banna	amf	Omotic	Greater Ab	AA	South Omotic	42800	NA	NA	42800	7120	NA	7120	0.16635514
Harari Adare	har	Semitic	Greater Ab	AA	Semitic	21300	NA	NA	21300	7766	NA	7766	0.364600939
Hausa	hau	Chadic	African	AA	West Chadic	24988000	24200000	Officia	24594000	15000000	15000000	15000000	0.609904855
Hebrew	heb	Semitic	Greater Me	AA	Semitic	5316700	NA	Up to	5316700	NA	4683300	4683300	0.880865951
Kachama-Ganjule	kcx	Omotic	Greater Ab	AA	North Omotic	4070	NA	NA	4070	419	NA	419	0.102948403
Kafa	kbr	Omotic	Greater Ab	AA	South Omotic	570000	NA	NA	570000	46720	NA	46720	0.081964912
Kambaata	ktb	Cushitic	Greater Ab	AA	Eastern Cushi	570000	NA	NA	570000	79332	NA	79332	0.139178947
Kistane (Soddo)	gru	Semitic	Greater Ab	AA	Semitic	255000	NA	NA	255000	60538	NA	60538	0.237403922

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

L2 Data Distribution



Arts & Humanities Research Council

CAMBRIDGE ASSESSMENT

UNIVERSITY OF CAMBRIDGE

Case Marking and L2 Ratios (Bentz & Winter, 2013)

Why case marking?

Case Marking and L2 Ratios (Bentz & Winter, 2013)

Why case marking?

- case marking is **hard to learn for adults**, irrespective of whether their native languages employ case or not (Papadopoulou et al., 2011)

Case Marking and L2 Ratios (Bentz & Winter, 2013)

Why case marking?

- case marking is **hard to learn for adults**, irrespective of whether their native languages employ case or not (Papadopoulou et al., 2011)
- there is **psycholinguistic evidence** for case reduction (Gürel, 2000; Haznedar, 2006)

Arts & Humanities Research Council CAMBRIDGE ASSESSMENT UNIVERSITY OF CAMBRIDGE

Case Marking and L2 Ratios (Bentz & Winter, 2013)

Why case marking?

- case marking is **hard to learn for adults**, irrespective of whether their native languages employ case or not (Papadopoulou et al., 2011)
- there is **psycholinguistic evidence** for case reduction (Gürel, 2000; Haznedar, 2006)
- there is **historical, qualitative evidence** for case loss (Trudgill, 2011; Herman& Wright, 2000)

Papadopoulou et al., 2011

- Case marking by Greek native speakers learning Turkish as L2

Papadopoulou et al., 2011

- Case marking by Greek native speakers learning Turkish as L2
- "Cloze task" with gaps in text

Papadopoulou et al., 2011

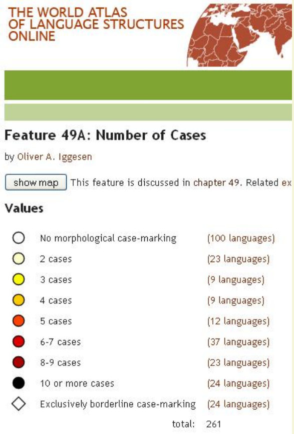
- Case marking by Greek native speakers learning Turkish as L2
- "Cloze task" with gaps in text

Table 2 Case suffixes: Correct scores per proficiency level

Cases	Level I (N = 35)	Level II (N = 37)	Level III (N = 39)
Specific object (accusative)	21% (29/140)	39% (58/148)	49% (77/156)
Non-specific object (unmarked)	76% (53/70)	64% (47/74)	62% (48/78)
Other cases	28% (253/910)	41% (393/962)	58% (588/1014)
Total	30% (335/1120)	42% (498/1184)	57% (713/1248)

Case marking in the **World Atlas of Language Structures** (Dryer& Haspelmath, 2011)

Case marking in the **World Atlas of Language Structures** (Dryer& Haspelmath, 2011)


THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

Feature 49A: Number of Cases

by Oliver A. Iggesen

show map This feature is discussed in chapter 49. Related ex

Values

○	No morphological case-marking	(100 languages)
●	2 cases	(23 languages)
●	3 cases	(9 languages)
●	4 cases	(9 languages)
●	5 cases	(12 languages)
●	6-7 cases	(37 languages)
●	8-9 cases	(23 languages)
●	10 or more cases	(24 languages)
◇	Exclusively borderline case-marking	(24 languages)
	total:	261

Case marking in the **World Atlas of Language Structures** (Dryer& Haspelmath, 2011)

THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

Feature 49A: Number of Cases

by Oliver A. Iggesen

show map This feature is discussed in chapter 49. Related ex

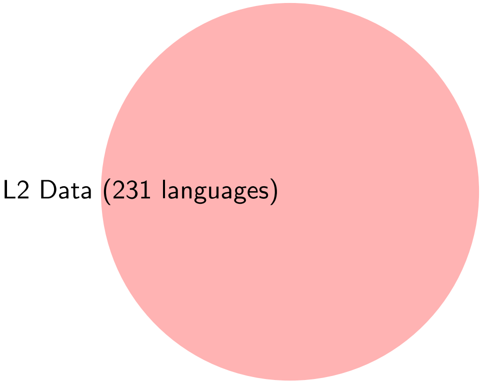
Values

○	No morphological case-marking	(100 languages)
○	2 cases	(23 languages)
●	3 cases	(9 languages)
●	4 cases	(9 languages)
●	5 cases	(12 languages)
●	6-7 cases	(37 languages)
●	8-9 cases	(23 languages)
●	10 or more cases	(24 languages)
◇	Exclusively borderline case-marking	(24 languages)
	total:	261

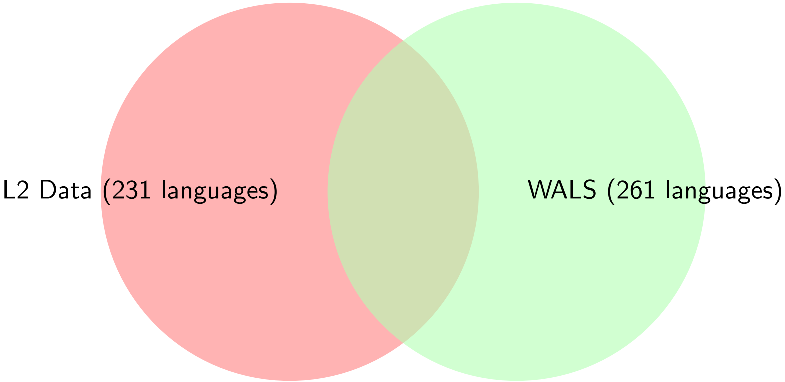
Hungarian (Tompa 1968: 206-209)

Nominative:	*hajó*
Accusative:	*hajó-t*
Inessive:	*hajó-ban*
Elative:	*hajó-ból*
Illative:	*hajó-ba*
Superessive:	*hajó-n*
Delative:	*hajó-ról*
Sublative:	*hajó-ra*
Adessive:	*hajó-nál*
Ablative:	*hajó-tól*
Allative:	*hajó-hoz*
Terminative:	*hajó-ig*
Dative:	*hajó-nak*
Instrumental-Comitative:	*hajó-val*
Formal:	*hajó-képp*
Essive:	*hajó-ul*
Essive-Formal(-Similitive):	*hajó-ként*
Translative-Factitive:	*hajó-vá*
Causal-Final:	*hajó-ért*
Distributive:	*hajó-nként*
Sociative:	*hajó-stul*

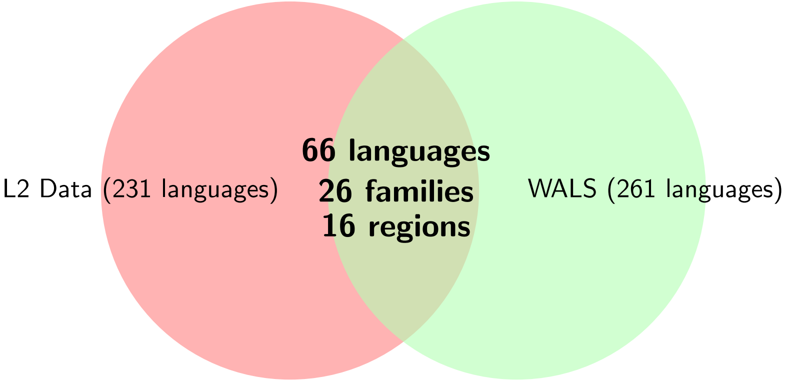
Statistical Model: Data Overlap



Statistical Model: Data Overlap



Statistical Model: Data Overlap



Statistical Models

Two separate models:

Statistical Models

Two separate models:

- a) Are languages **without case** those languages with higher L2 percentages?

Arts & Humanities Research Council CAMBRIDGE ASSESSMENT UNIVERSITY OF CAMBRIDGE

Statistical Models

Two separate models:

- a) Are languages **without case** those languages with higher L2 percentages?
- b) Do languages with more L2 speakers have **fewer case** paradigms?

Model A

Case as a binary variable (case/no case)

- requires **logistic regression** (binary dependent/outcome variable)
- Requires **mixed-effects** (random and fixed effects) due to non-independence of data points (family and area clusters) (Baayen et al., 2008; Bates et al., 2014; Bickel & Nichols, 2009; Jäger et al., 2011)

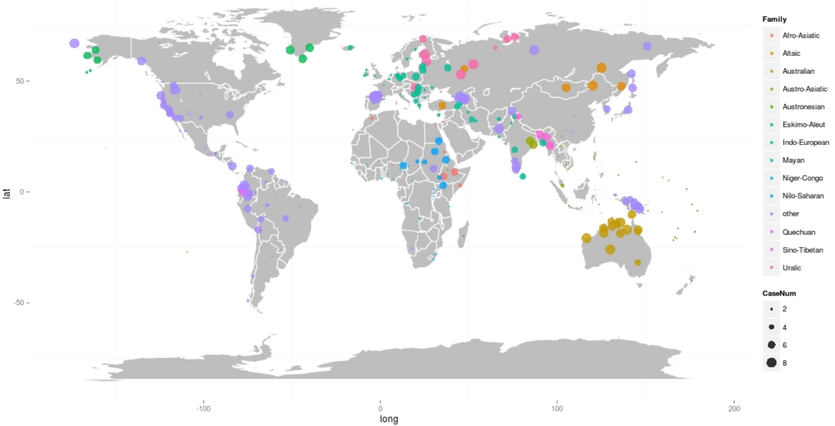
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Model A

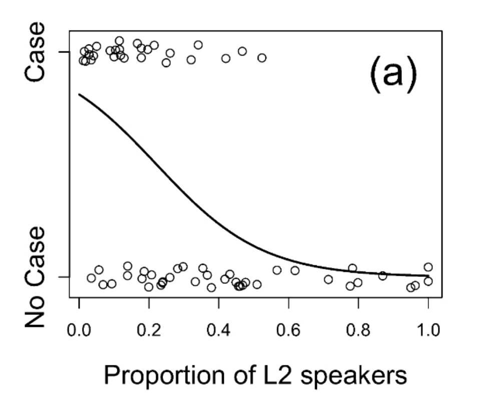
Case as a binary variable (case/no case)

- requires **logistic regression** (binary dependent/outcome variable)
- Requires **mixed-effects** (random and fixed effects) due to non-independence of data points (family and area clusters) (Baayen et al., 2008; Bates et al., 2014; Bickel & Nichols, 2009; Jäger et al., 2011)
- **Model specification**:
  $P(y_i = 1) = f^{-1}(\alpha_0 + \alpha_{jk_i} + (\beta_0 + \beta_{jk_i}) \times x_i + e_{jk_i})$

WALS Chapter 49: Number of Cases



Model A: Outcome



Are languages **without case** those languages with higher L2 percentages?
-**Yes.**

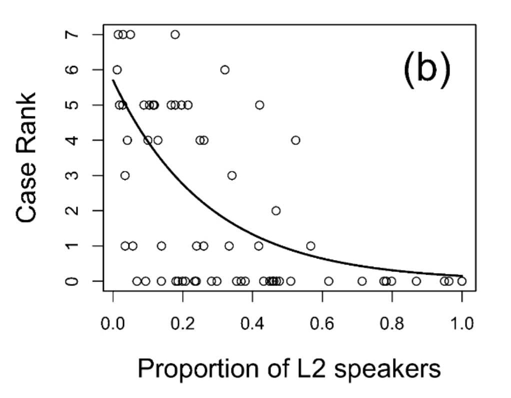
Statistical Significance
coefficient estimates:
-6.57 ± 2.03;
$p = 0.00014$

Model B

Case as a continuous variable (no case, 2 cases, 3 cases, etc.)

- requires **Poisson or negative binomial regression** (continuous dependent/outcome variable)
- Requires **mixed-effects** (random and fixed effects) due to non-independence of data points (family and area clusters) (Baayen et al., 2008; Bates et al., 2014; Bickel & Nichols, 2009; Jäger et al., 2011)

Model B: Outcome



Are languages with **fewer cases** those languages with higher L2 percentages?
-**Yes.**

Statistical Significance
coefficient estimates:
-3.6$\pm$ 1.06;
p $=$ 0.00062

Arts & Humanities Research Council CAMBRIDGE ASSESSMENT UNIVERSITY OF CAMBRIDGE

Case Marking: Conclusions

- Languages with more L2 speakers tend to have **fewer** cases or **no** case marking at all (in our sample)

Case Marking: Conclusions

- Languages with more L2 speakers tend to have **fewer** cases or **no** case marking at all (in our sample)
- These trends hold even if family and areal relationships are accounted for

General Problems

- WALS chapters are only very **coarse grained** descriptions of linguistic structures

General Problems

- WALS chapters are only very **coarse grained** descriptions of linguistic structures
- They tell us nothing about the **actual productivity** of morphological markers

General Problems

- WALS chapters are only very **coarse grained** descriptions of linguistic structures
- They tell us nothing about the **actual productivity** of morphological markers
- overall morphological productivity in a language is driven by a multitude of **different markers**

Example: German cases

- According to WALS German has four nominal cases (Nom, Acc, Dat, Gen)

Example: German cases

- According to WALS German has four nominal cases (Nom, Acc, Dat, Gen)
- But there is a lot of **case syncretism** for individual noun declensions

Example: German cases

- According to WALS German has four nominal cases (Nom, Acc, Dat, Gen)
- But there is a lot of **case syncretism** for individual noun declensions
- **Frequencies** of case marked forms might differ strongly

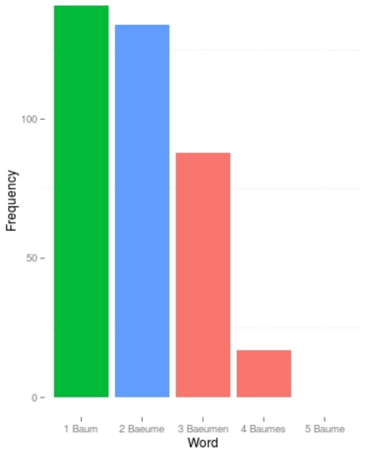
Case Syncretism

	SG	PL
NOM	Baum (Eng. tree)	Bäume (Eng. trees)
ACC	Baum	Bäume
DAT	Baum(**e**)	Bäume**n**
GEN	Baum**es**	Bäume

Case Syncretism

	SG	PL
NOM	Baum (Eng. tree)	Bäume (Eng. trees)
ACC	Baum	Bäume
DAT	Baum(**e**)	Bäume**n**
GEN	Baum**es**	Bäume

Word Frequencies (CELEX)



Towards a cross-linguistic measure of morphological productivity

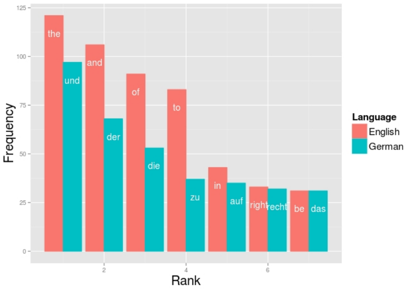
- Data: **whole corpora** with constant information content (parallel texts)
- Method: **frequency distributions** across languages

Measuring overall morphological productivity in corpora

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

Measuring overall morphological productivity in corpora

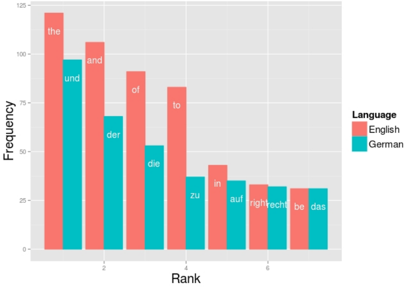
Frequency distributions: Order types (word forms delimited by white spaces) according to their token frequencies (Zipf,1932,1949)

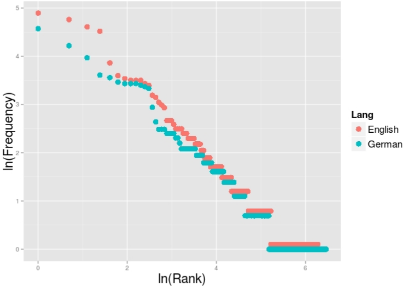


Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Measuring overall morphological productivity in corpora

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





What drives differences in frequency distributions?

Experiment:

- Balanced Parallel Corpus of English and German (ca. 10000 words; OpenSubTitles, Europarl, Bible, UDHR)

What drives differences in frequency distributions?

Experiment:

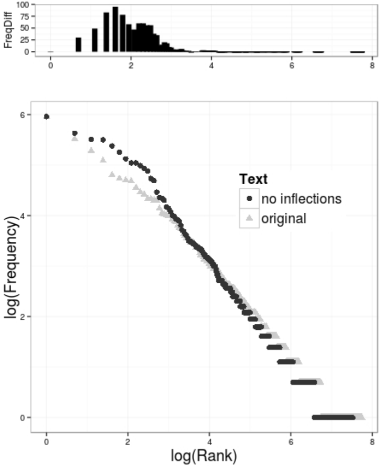
- Balanced Parallel Corpus of English and German (ca. 10000 words; OpenSubTitles, Europarl, Bible, UDHR)
- Remove successively: Inflections, derivations, compounds, clitics

What drives differences in frequency distributions?

Experiment:

- Balanced Parallel Corpus of English and German (ca. 10000 words; OpenSubTitles, Europarl, Bible, UDHR)
- Remove successively: Inflections, derivations, compounds, clitics
- Compute the percentage of change in frequency difference

Example:



German inflections

Baum 141
Bäume 134
Bäumen 88
Baumes 17
Baume 0

Baum 380

What drives differences in frequency distributions?

- **inflectional morphology:** ca. 48% (also Bentz et al., 2014)

What drives differences in frequency distributions?

- **inflectional morphology:** ca. 48% (also Bentz et al., 2014)
- **derivational morphology:** ca. 28%

What drives differences in frequency distributions?

- **inflectional morphology:** ca. 48% (also Bentz et al., 2014)
- **derivational morphology:** ca. 28%
- compounds: ca. 15%

What drives differences in frequency distributions?

- **inflectional morphology:** ca. 48% (also Bentz et al., 2014)
- **derivational morphology:** ca. 28%
- compounds: ca. 15%
- clitics: ca. 4%

What drives differences in frequency distributions?

- **inflectional morphology:** ca. 48% (also Bentz et al., 2014)
- **derivational morphology:** ca. 28%
- compounds: ca. 15%
- clitics: ca. 4%
- others (base vocabulary, orthography, etc.): ca. 5%

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Morphological productivity and lexical diversity

Finding: Productive morphology creates **new word types**, more **low frequency items**, and hence high **lexical diversity**

Morphological productivity and lexical diversity

Finding: Productive morphology creates **new word types**, more **low frequency items**, and hence high **lexical diversity**

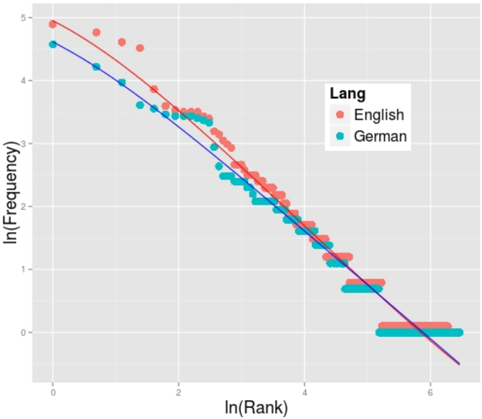
$\Downarrow\Downarrow\Downarrow$

We can use lexical diversity measures as proxy for overall morphological productivity (Bentz et al., 2014; Popescu et al., 2009; Ha et al., 2006)

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Lexical diversity measures

- Zipf-Mandelbrot's α
- Shannon entropy (H)
- Type-Token Ratios (TTR)



Quantitative measures

Shannon entropy (Shannon & Weaver, 1949)

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

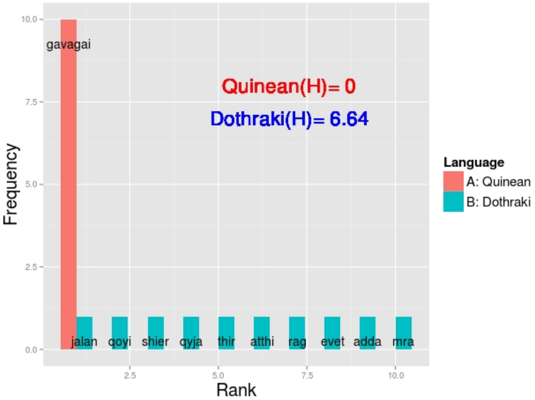
$$p_i : \frac{frequency\ of\ w_i}{total\ number\ of\ tokens}$$

Quantitative measures

Shannon entropy (Shannon & Weaver, 1949)

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

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

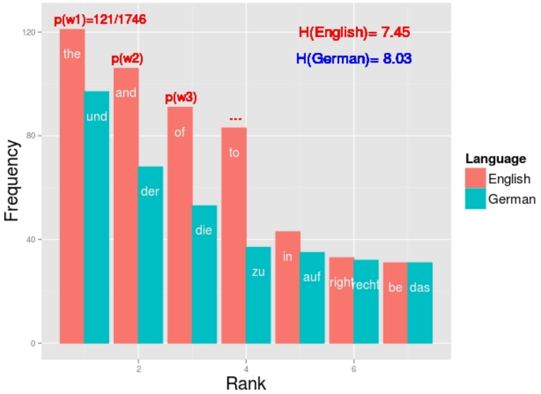


Quantitative measures

Shannon entropy (Shannon & Weaver, 1949)

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

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

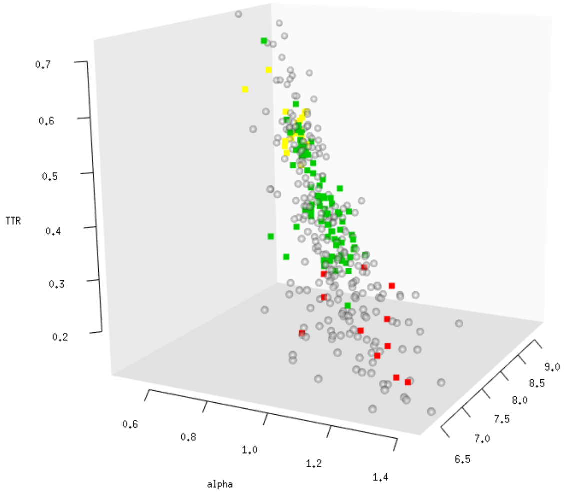


Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Lexical diverstiy measures

Productive morphology creates higher lexical diversity
$\rightarrow$ **higher** entropy (higher uncertainty)
$\rightarrow$ **higher** type-token ratios
$\rightarrow$ **lower** ZM's α

Lexical Diversity Space



369 texts the Universal Declaration of Human Rights (UDHR)

Altaic
Indo-European
Creole

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Statistical Model

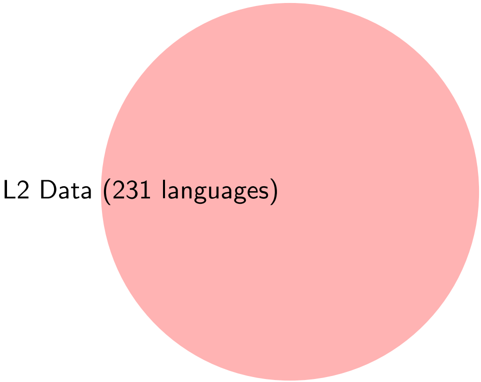
- Are languages with **higher lexical diversities** (i.e. higher morphological productivity) those languages with lower L2 proportions?

Statistical Model

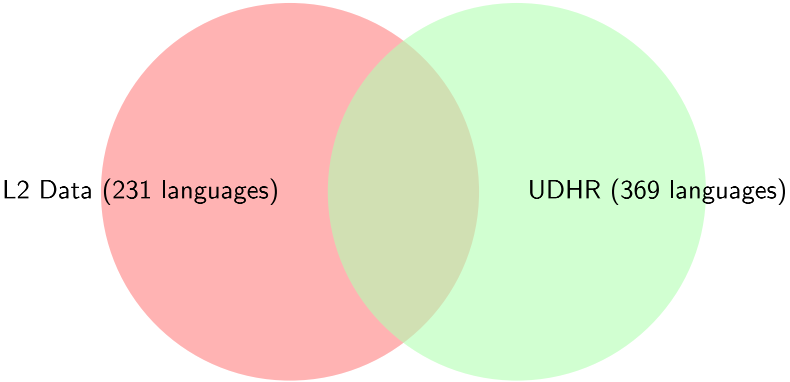
Lexical diversity measures as continuous variables

- requires **linear regression**:
  continuous dependent/outcome variables: α,H,TTR
  continuous predictors: L2 proportions (fixed effect)
- requires **mixed-effects** (random and fixed effects) due to non-independence of data points (family and area clusters) (Baayen et al., 2008; Bates et al., 2014; Jäger et al., 2011)

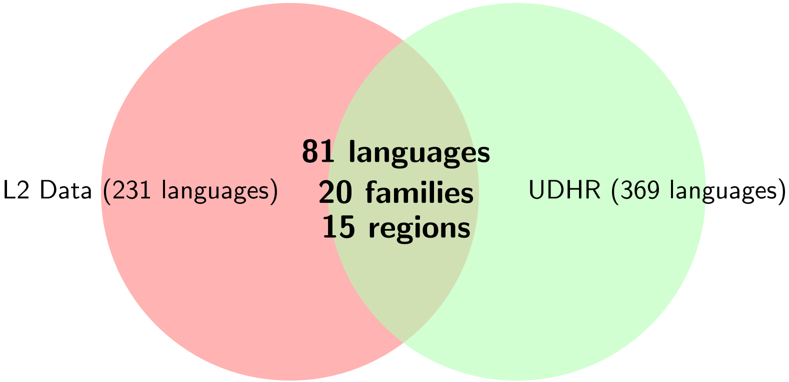
Statistical Model: Data Overlap



Statistical Model: Data Overlap



Statistical Model: Data Overlap



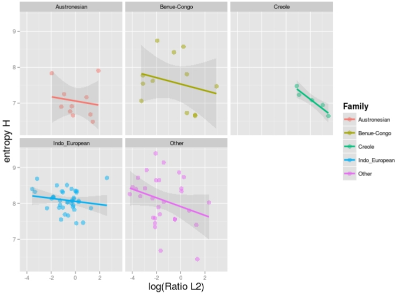
Results

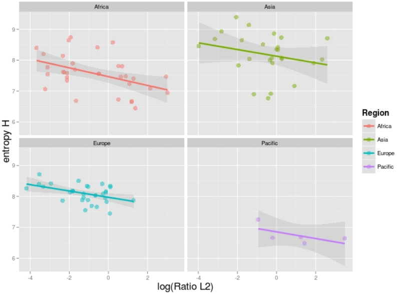
All coefficients point in the right direction. However, only coefficients for H and TTR are significant

Dependent variable	Fixed effects	Random effects	Coefficient (L2 ratio)	Likelihood ratio test	
				df (L2 ratio)	χ^2 (L2 ratio)
ZM's α	log (L2), script	family, region	0.023	1	1.38
Entropy H	log (L2), script	family, region	-0.14	1	9.28***
TTR	log (L2), script	family, region	-0.026	1	7.11**

*p < 0.05; **p < 0.01; ***p < 0.001

L2 effect across families and regions





Arts & Humanities Research Council CAMBRIDGE ASSESSMENT UNIVERSITY OF CAMBRIDGE

Lexical diversity: Conclusions

- Languages with more L2 speakers tend to have *lower* lexical diversity (at least in the UDHR)

Lexical diversity: Conclusions

- Languages with more L2 speakers tend to have *lower* lexical diversity (at least in the UDHR)
- These trends hold even if family and areal relationships are accounted for

Problems

Problems

- **Correlation is not causation** (Roberts & Winters, 2012,2013) $\rightarrow$ but there is independent psycholinguistic and historical evidence.

Arts & Humanities Research Council CAMBRIDGE ASSESSMENT UNIVERSITY OF CAMBRIDGE

Problems

- **Correlation is not causation** (Roberts & Winters, 2012,2013) → but there is independent psycholinguistic and historical evidence.
- **Synchronic data and diachronic implications** → Diachronic study on frequency distributions in Old English and Modern English (Bentz et al., 2014)

Problems

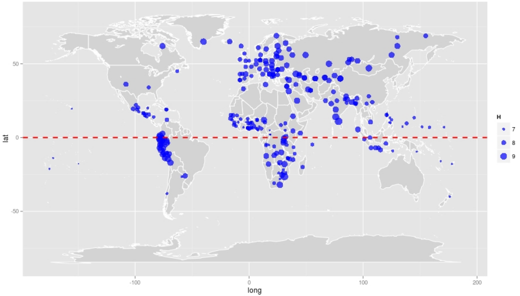
- **Correlation is not causation** (Roberts & Winters, 2012,2013) → but there is independent psycholinguistic and historical evidence.
- **Synchronic data and diachronic implications** → Diachronic study on frequency distributions in Old English and Modern English (Bentz et al., 2014)
- **Parallel texts use doculects** → Frequency distributions show similar behavior with regards to inflection across different types of texts (Bentz et al., 2014; Corral et al. ,2014; Popescu et al., 2009; Ha et al., 2006)

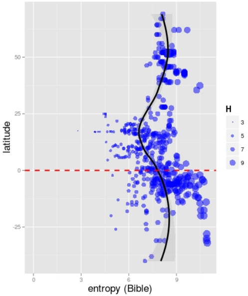
Geographical Distribution of Lexical Diversity

Parallel Bibel Corpus (ca. 800 languages; Mayer & Cysouw, 2014)

Geographical Distribution of Lexical Diversity

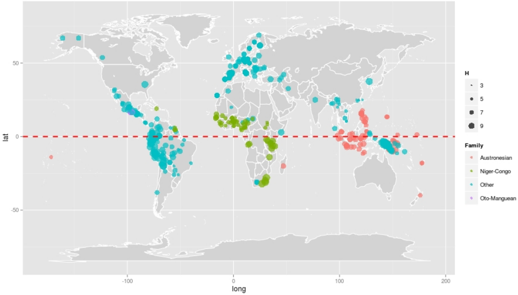
Parallel Bibel Corpus (ca. 800 languages; Mayer & Cysouw, 2014)
Lexical diversity seems lower around the equator. - Why?

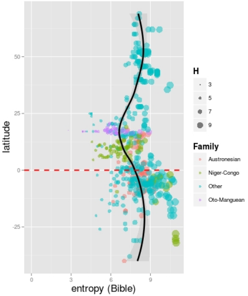




Geographical Distribution of Lexical Diversity

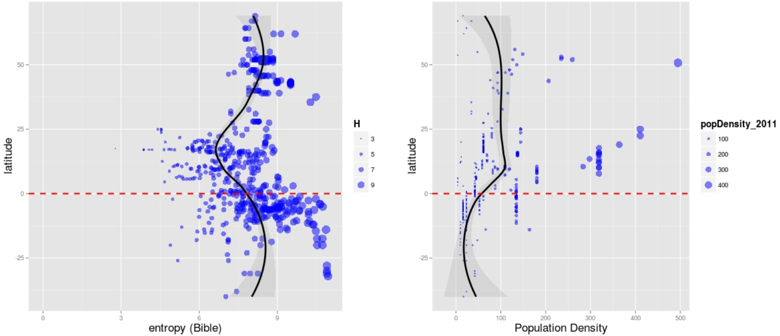
Language Families





Geographical Distribution of Lexical Diversity

(?) Population Density → More Contact → Lower Lexical Diversity (?)



Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Questions

What is the relationship between **language areas**, **families** and **contact phenomena**? What is **cause** and **effect**?

Questions

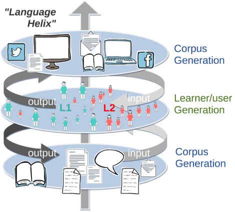
What is the relationship between **language areas**, **families** and **contact phenomena**? What is **cause** and **effect**?

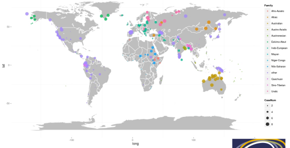
- family clustering $\leftrightarrow$ linguistic structure
- areal clustering $\leftrightarrow$ linguistic structure

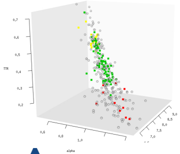
Conclusions

Our statistical analyses suggest:

- Languages with **higher L2 proportions** have **fewer** cases or **no case** marking at all
- Languages with **higher L2 proportions** have **lower lexical diversities** (at least when measured with entropy H or TTR)
- Both effects are stable across families and regions
- This is evidence that languages **adapt** to **learning constraints** of speaker populations







Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Collaborators

Douwe Kiela

Felix Hill

Andrew Caines

Dimitrios Alikaniotis

Paula Buttery

Thank You!

¡chris@christianbentz.de¿