

Measuring and modeling the impact of non-native speakers on morphological productivity

Christian Bentz
cb696@cam.ac.uk

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

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

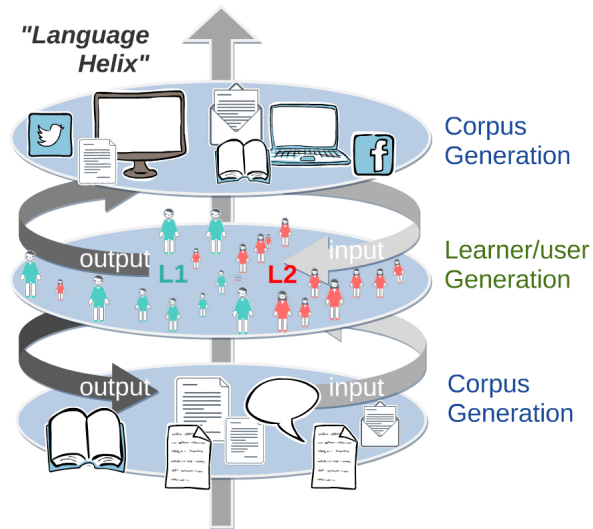
"The **structures of language** emerge from interrelated patterns of experience, **social interaction**, and **cognitive mechanisms**."
(Beckner et al., 2009)

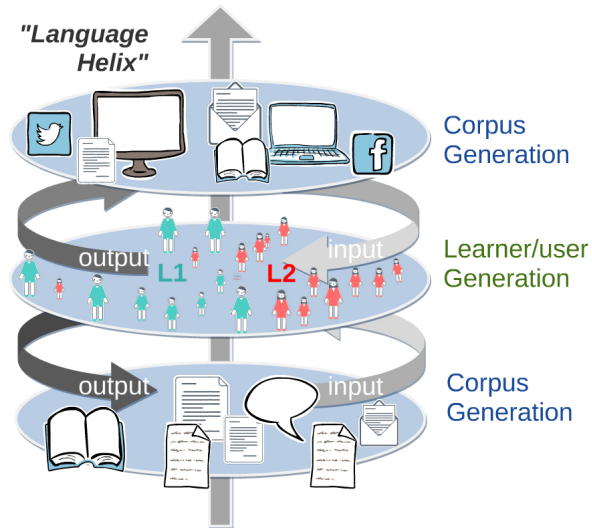
Linguistic Niche Hypothesis

"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)

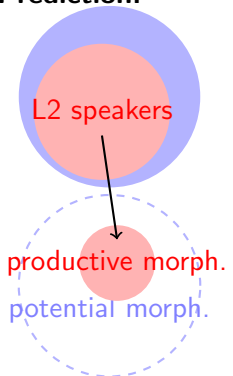
Earlier studies

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





Prediction:



Collecting L2 Data

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



Max Planck Institute
for Evolutionary Anthropology

Collecting L2 Data

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



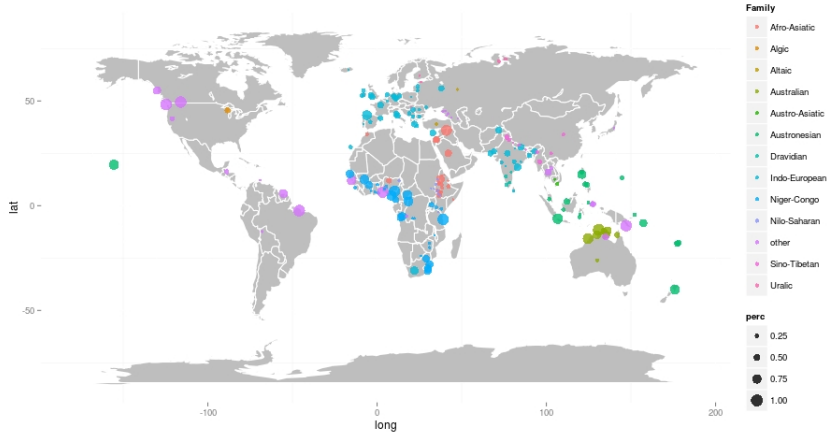
Max Planck Institute
for Evolutionary Anthropology

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	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	Omoti	Greater Ab-AA	South Omoti		155000	NA	NA	155000	13319	NA	13319	0.085929032
Afar	aar	Cushitic	Greater Ab-AA	Eastern Cush		1078200	NA	1.4 m	1239100	22848	NA	22848	0.01843919
Alaba-K'abeena	alw	Cushitic	Greater Ab-AA	Eastern Cush		162000	NA	NA	162000	29699	NA	29699	0.18332716
Amharic	amh	Semitic	Greater Ab-AA	Semitic		17528500	174000000	Official	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	arv	Semitic	N Africa	AA	Semitic	20000	NA	NA	20000	44000	NA	44000	2.2
Arbore	arv	Cushitic	Greater Ab-AA	Eastern Cush		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 Cush		500000	NA	###	428490	64425	NA	64425	0.150353567
Basketo	bst	Omoti	Greater Ab-AA	North Omoti		57800	NA	NA	57800	8961	NA	8961	0.155034602
Bench (Gimira)	bcq	Omoti	Greater Ab-AA	North Omoti		174000	NA	NA	174000	22640	NA	22640	0.130114943
Borna (Shinassha)	bwo	Omoti	Greater Ab-AA	North Omoti		19900	NA	NA	19900	2276	NA	2276	0.114371859
Bussa	dox	Cushitic	Greater Ab-AA	Eastern Cush		6620	NA	NA	6620	920	NA	920	0.13897281
Dime Dima	dim	Omoti	Greater Ab-AA	South Omoti		6500	NA	NA	6500	529	NA	529	0.081384615
Dirasha (Gidole)	gdl	Cushitic	Greater Ab-AA	Eastern Cush		90000	NA	NA	90000	7000	NA	7000	0.077777778
Dizi	mdx	Omoti	Greater Ab-AA	North Omoti		21100	NA	NA	21100	2054	NA	2054	0.097345972
Dorze	dor	Omoti	Greater Ab-AA	North Omoti		20800	NA	NA	20800	3597	NA	3597	0.172932692
Gamo-Gofa-Dawo	gmo	Omoti	Greater Ab-AA	North Omoti		1240000	NA	NA	1240000	77883	NA	77883	0.062808871
Gawwada (Dullay)	gwd	Cushitic	Greater Ab-AA	Eastern Cush		32700	NA	NA	32700	1367	NA	1367	0.041804281
Gedeo Darasa	drs	Cushitic	Greater Ab-AA	Eastern Cush		637000	NA	NA	637000	47950	NA	47950	0.075274725
Hadiyya Adea	hdy	Cushitic	Greater Ab-AA	Eastern Cush		924000	NA	NA	924000	15889	NA	15889	0.017195887
Hamar-Banna	amf	Omoti	Greater Ab-AA	South Omoti		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	Official	24594000	15000000	15000000	15000000	0.609904855
Hebrew	heb	Semitic	Greater Me-AA	Semitic		5316700	NA	Up to	5316700	NA	4683300	4683300	0.880865951
Kachama-Ganjul	kcc	Omoti	Greater Ab-AA	North Omoti		4070	NA	NA	4070	419	NA	419	0.102948403
Kafa	kbr	Omoti	Greater Ab-AA	South Omoti		570000	NA	NA	570000	46720	NA	46720	0.081964912
Kambaata	ktb	Cushitic	Greater Ab-AA	Eastern Cush		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



L2 Data Distribution



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)

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

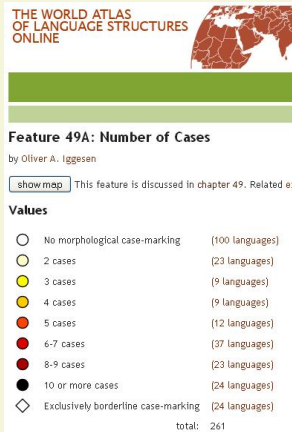
- Case marking by Greek native speakers learning Turkish as L2
- "Cloze task" with gaps in text

Table 2 Case suffixes: Correct scores per proficiency level

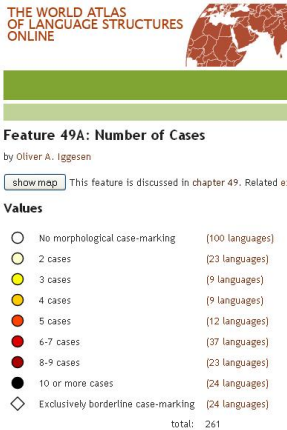
Cases	Level I (N = 35)	Level II (N = 37)	Level III (N = 39)
Specific object (accusative)	21% (29/140)	39% (58/148)	49% (77/156)
Non-specific object (unmarked)	76% (53/70)	64% (47/74)	62% (48/78)
Other cases	28% (253/910)	41% (393/962)	58% (588/1014)
Total	30% (335/1120)	42% (498/1184)	57% (713/1248)

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

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



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




Hungarian (Tompá 1968: 206–209)

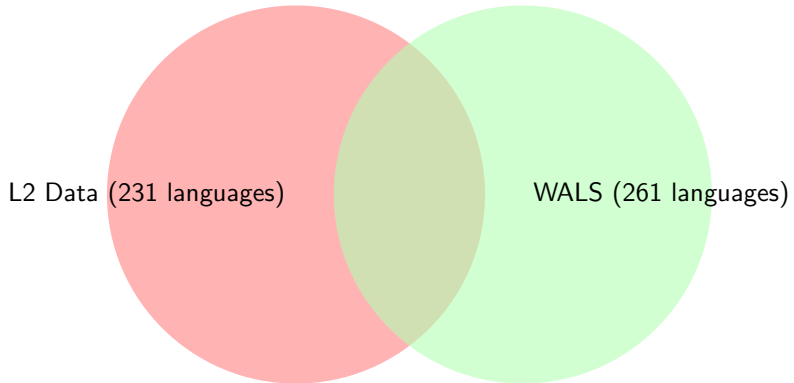
Nominative:	<i>hajó</i>
Accusative:	<i>hajó-t</i>
Inessive:	<i>hajó-ban</i>
Elativ:	<i>hajó-ból</i>
Illative:	<i>hajó-ba</i>
Superessive:	<i>hajó-n</i>
Delative:	<i>hajó-ról</i>
Sublative:	<i>hajó-ra</i>
Adessive:	<i>hajó-nál</i>
Ablative:	<i>hajó-tól</i>
Allative:	<i>hajó-hoz</i>
Terminative:	<i>hajó-ig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

Statistical Model: Data Overlap

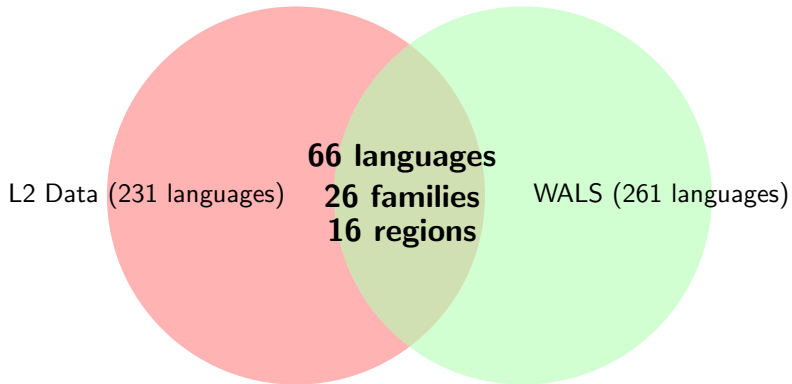
L2 Data (231 languages)

A large, solid red circle is centered on the slide. To its left, the text "L2 Data (231 languages)" is written in a black, sans-serif font.

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 ratios?

Statistical Models

Two separate models:

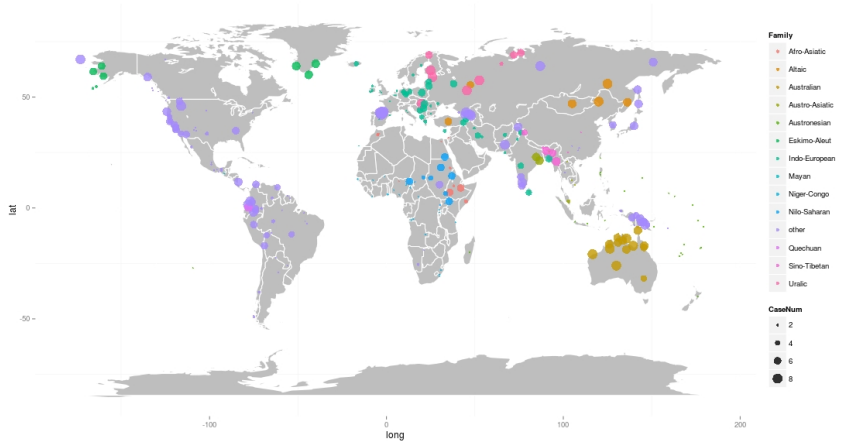
- a) Are languages **without case** those languages with higher L2 ratios?
- b) Do languages with more L2 speakers have **fewer case** paradigms?

Model A

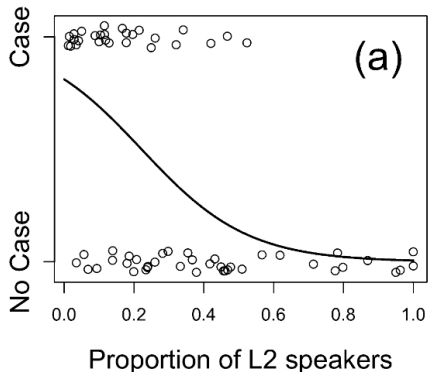
Case as a binary variable (case/no case)

- 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; Jäger et al., 2011)

WALS Chapter 49: Number of Cases



Model A: Outcome



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

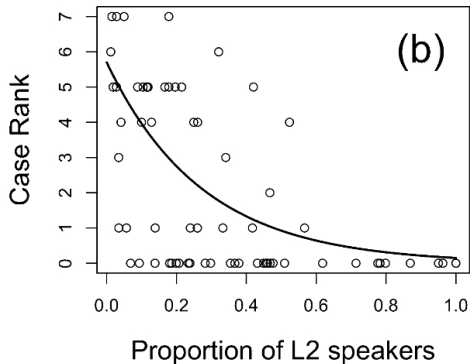
**Statistical
Significance**
coefficient estimates:
 -6.57 ± 2.03 ;
 $p = 0.00014$

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; 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 ± 1.06 ;
 $p = 0.00062$

Case Marking: Conclusions

- Languages with more L2 speakers tend to have **fewer** cases or **no** case marking at all

Case Marking: Conclusions

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

Excuse: How about verbal inflections?

- Kaius Sinnemäki is currently investigating WALS chapter 22A (inflectional synthesis of the verb)
<http://www.helsinki.fi/~ksinnema/presentations.html>

Excuse: How about verbal inflections?

- Kaius Sinnemäki is currently investigating WALS chapter 22A (inflectional synthesis of the verb)
<http://www.helsinki.fi/~ksinnema/presentations.html>
- There is a slight negative trend when correlated with L2 percentages, but **not significant** in mixed-effects models

(Potential) Reasons

- Differences in learning difficulties between case marking and verbal inflection (suggested by Papadopoulou et al. 2011)?

Excuse: How about verbal inflections?

- Kaius Sinnemäki is currently investigating WALS chapter 22A (inflectional synthesis of the verb)
<http://www.helsinki.fi/~ksinnema/presentations.html>
- There is a slight negative trend when correlated with L2 percentages, but **not significant** in mixed-effects models

(Potential) Reasons

- Differences in learning difficulties between case marking and verbal inflection (suggested by Papadopoulou et al. 2011)?
- Smaller sample size (36 languages) undermines significant p-values?

Excuse: How about verbal inflections?

- Kaius Sinnemäki is currently investigating WALS chapter 22A (inflectional synthesis of the verb)
<http://www.helsinki.fi/~ksinnema/presentations.html>
- There is a slight negative trend when correlated with L2 percentages, but **not significant** in mixed-effects models

(Potential) Reasons

- Differences in learning difficulties between case marking and verbal inflection (suggested by Papadopoulou et al. 2011)?
- Smaller sample size (36 languages) undermines significant p-values?
- Categories in WALS chapter 22A do not reflect learnability?

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

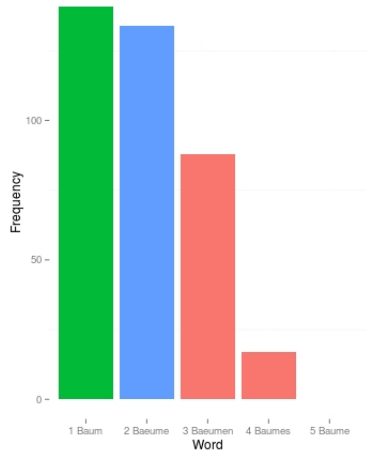
Case Syncretism

	SG	PL
NOM	Baum (Eng. tree)	Bäume (Eng. trees)
ACC	Baum	Bäume
DAT	Baum(e)	Bäumen
GEN	Baumes	Bäume

Word Frequencies (CELEX)

Case Syncretism

	SG	PL
NOM	Baum (Eng. tree)	Bäume (Eng. trees)
ACC	Baum	Bäume
DAT	Baum(e)	Bäumen
GEN	Baumes	Bäume



Towards a cross-linguistic measure of morphological productivity

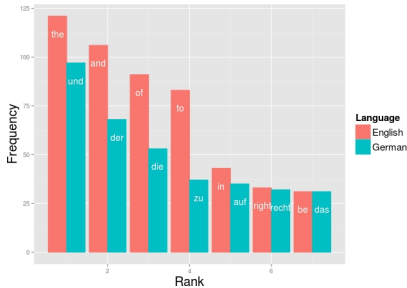
- Data: **whole corpora** with constant information content (parallel texts)
- Method: **frequency distributions** across languages

Measuring overall morphological productivity in corpora

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

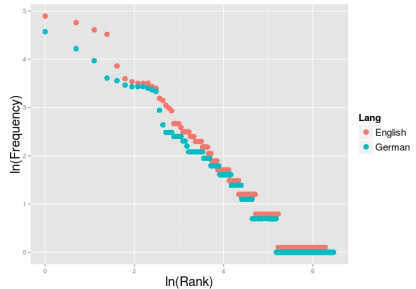
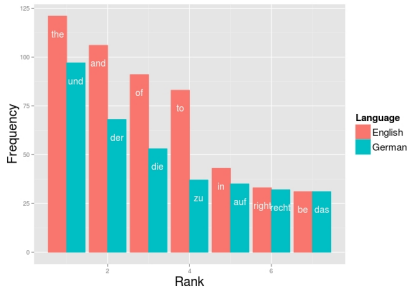
Measuring overall morphological productivity in corpora

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



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?

- **inflectional morphology** (Bentz et al., 2014)

What drives differences in frequency distributions?

- **inflectional morphology** (Bentz et al., 2014)
- derivational morphology

What drives differences in frequency distributions?

- **inflectional morphology** (Bentz et al., 2014)
- derivational morphology
- compounds

What drives differences in frequency distributions?

- **inflectional morphology** (Bentz et al., 2014)
- derivational morphology
- compounds
- clitics

What drives differences in frequency distributions?

- **inflectional morphology** (Bentz et al., 2014)
- derivational morphology
- compounds
- clitics
- base vocabulary

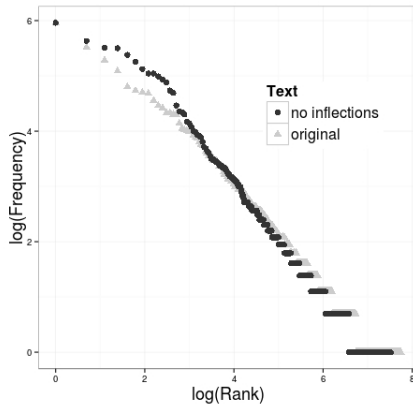
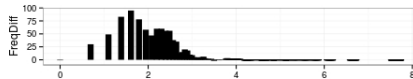
What drives differences in frequency distributions?

- **inflectional morphology** (Bentz et al., 2014)
- derivational morphology
- compounds
- clitics
- base vocabulary
- orthography

What drives differences in frequency distributions?

- **inflectional morphology** (Bentz et al., 2014)
- derivational morphology
- compounds
- clitics
- base vocabulary
- orthography
- etc.

German inflections



Baum 141	}	Baum 380
Bäume 134		
Bäumen 88		
Baumes 17		
Baume 0		

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

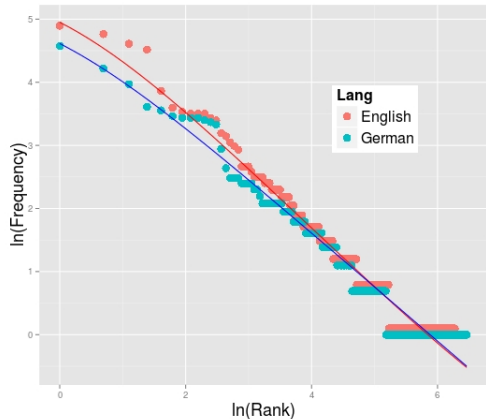
Finding: Productive morphology creates **new word types**, more **low frequency items**, and hence high **lexical diversity**



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

Lexical diversity measures

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

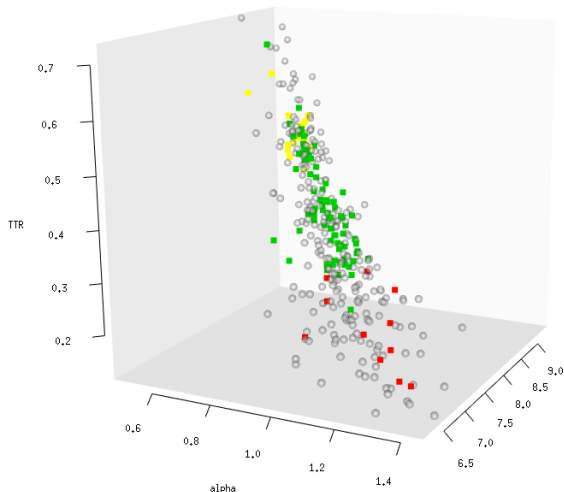


Lexical diversity measures

Inflection creates higher lexical diversity

- **higher** entropy (higher uncertainty)
- **higher** type-token ratios
- **lower** ZM's α

Lexical Diversity Space



369 texts the
Universal Declaration
of Human Rights
(UDHR)

Altaic
Indo-European
Creole

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

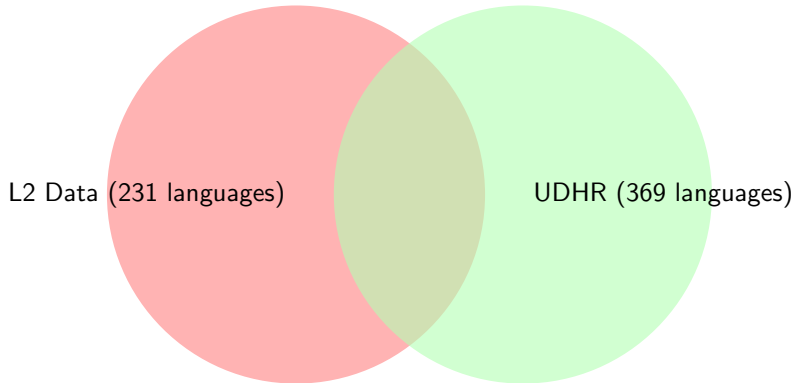
- requires **linear regression**:
continuous dependent/outcome variables: α , H, TTR
continuous predictors: L2 proportions (fixed effect)
- requires **mixed-effects** (random and fixed effects) due to non-independence of data points (family and area clusters) (Baayen et al., 2008; Bates et al., 2014;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

Statistical Model: Data Overlap

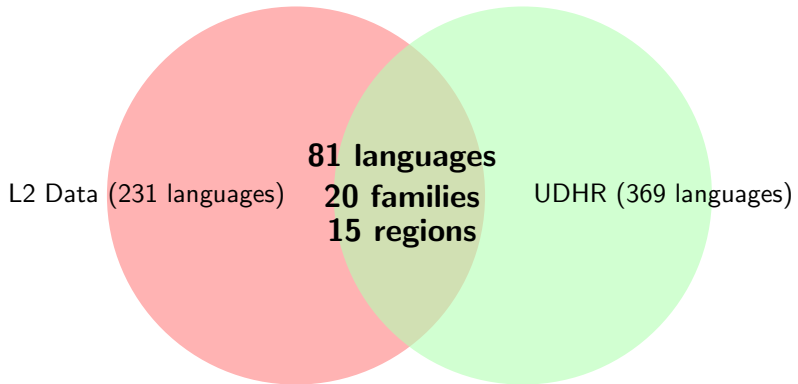
L2 Data (231 languages)

A large, solid red circle is centered on the slide. To its left, the text "L2 Data (231 languages)" is written in a black, sans-serif font.

Statistical Model: Data Overlap



Statistical Model: Data Overlap



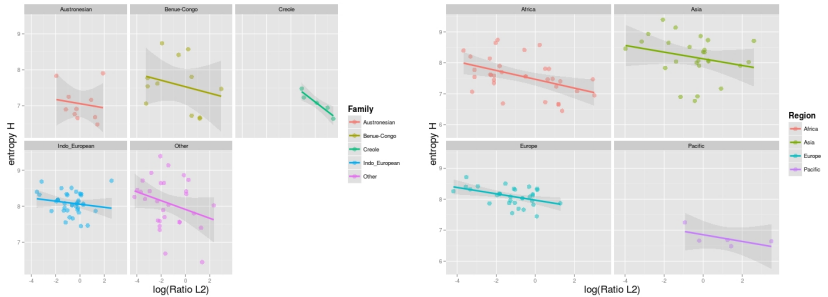
Results

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



Problems

Problems

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

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)

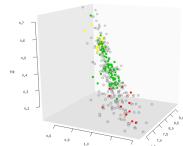
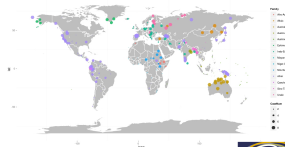
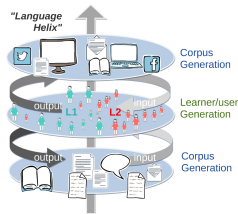
Problems

- **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)

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



Collaborators



Douwe Kiela



Felix Hill



Andrew Caines



Dimitrios Alikaniotis



Paula Buttery

Thank You!

jchris@christianbentz.dej