

The impact of non-native speakers on word forms and (potentially) word order

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
- **Word forms** and L2 speaker proportions

Future Directions

- **Word forms** and **word order**
- **Conclusions**

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)

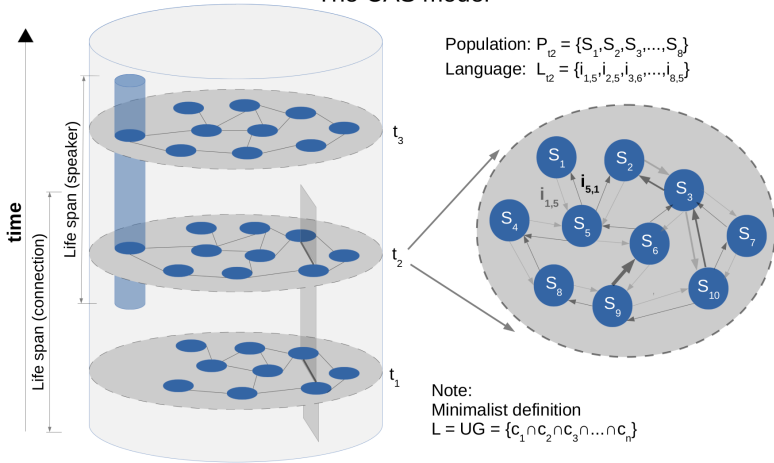
Linguistic Niche Hypothesis

"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)

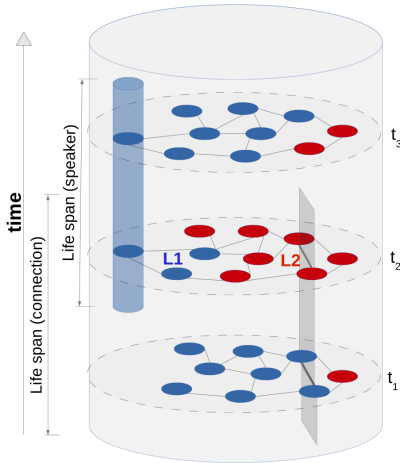
Earlier studies

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

The CAS model



Language contact in the CAS model



Prediction of the CAS model:

Population

Language

$$P_{t_1} = \{S_1, S_2, S_3, \dots, S_8\} \rightarrow L_{t_2} = \{i_{1,5}, i_{2,5}, i_{3,6}, \dots, i_{8,5}\}$$

$$P_{t_1} = \{S_1, S_2, S_3, \dots, S_7\} \rightarrow L_{t_1} = \{i_{1,5}, i_{2,5}, i_{3,6}, \dots, i_{8,5}\}$$

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)



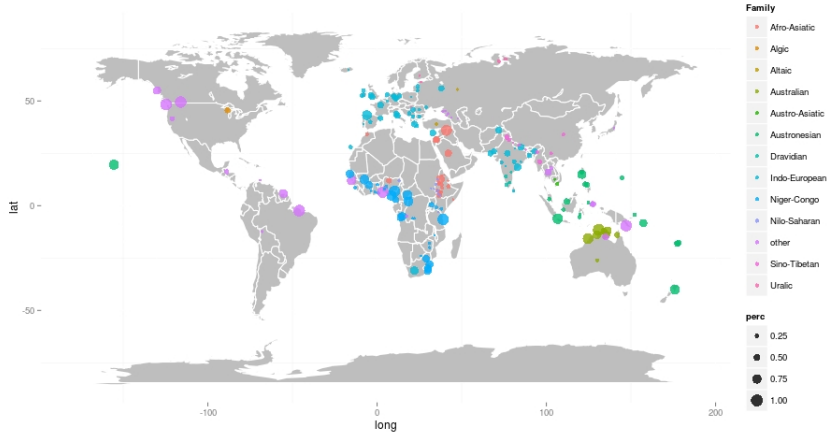
Max Planck Institute
for Evolutionary Anthropology

Dataset of L2 and L1 numbers for 226 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	5000000	0.839502558	
Aari	aiw	Omoti	Greater Ab-AA	South Omoti		155000	NA	NA	155000	13319	13319	0.085929032	
Afar	aar	Cushitic	Greater Ab-AA	Eastern Cush		1078200	NA	1.4 m	1239100	22848	22848	0.01843919	
Alaba-K'abeena	alw	Cushitic	Greater Ab-AA	Eastern Cush		162000	NA	NA	162000	29699	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	pga	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 Adeaa	hdy	Cushitic	Greater Ab-AA	Eastern Cush		924000	NA	NA	924000	15889	NA	15889	0.017195887
Hamer-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



What can we predict using non-native speaker data?

Qualitative hypothesis

- Languages with more non-native speakers tend to simplify morphological marking (Wray& Grace, 2007; McWhorter, 2002, 2007; Trudgill, 2011)

Quantitative evidence

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

General problems

- Both Lupyan & Dale (2010) and Bentz & Winter (2012, 2013) used the **World Atlas of Language Structures (WALS)**

General problems

- Both Lupyan & Dale (2010) and Bentz & Winter (2012, 2013) used the **World Atlas of Language Structures** (WALS)
- WALS is a helpful but coarse grained source for typological data

General problems

- Both Lupyan & Dale (2010) and Bentz & Winter (2012, 2013) used the **World Atlas of Language Structures** (WALS)
- WALS is a helpful but coarse grained source for typological data

Example

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

General problems

- Both Lupyan & Dale (2010) and Bentz & Winter (2012, 2013) used the **World Atlas of Language Structures** (WALS)
- WALS is a helpful but coarse grained source for typological data

Example

- According to WALS German has four **nominal cases** (Nom, Acc, Dat, Gen)
- But there are up to 37 different **declension classes**
- **case syncretism** in individual noun declensions

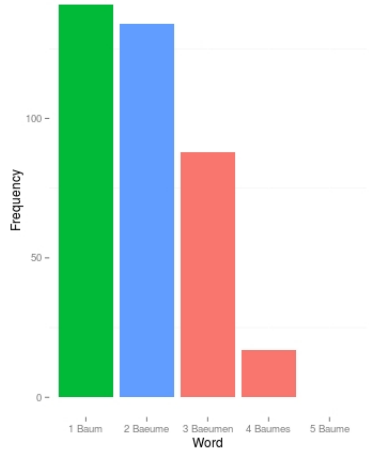
Case Syncretism

	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



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

Data: Parallel texts (constant content)

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

Data: Parallel texts (constant content)

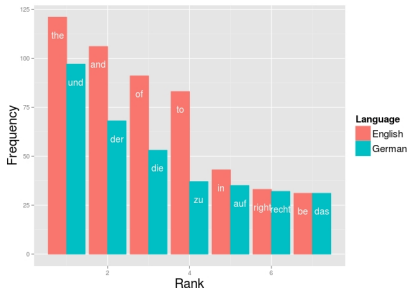
- **Parallel Bible Corpus** (810 languages, ca. 20000 words per language)
- **Universal Declaration of Human Rights** (376 languages, ca. 2000 words per language)
- **European Parliament Corpus** (21 languages, ca. 7 million words per language)

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

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

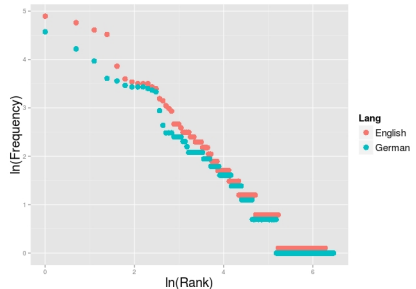
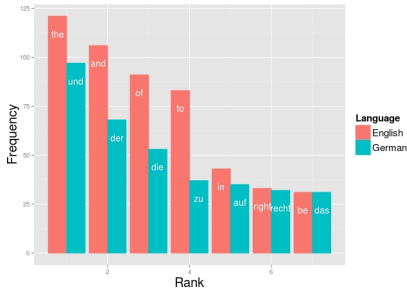
Zipfian approach: Analysis of word form distributions across languages (lexical diversities)

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



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

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

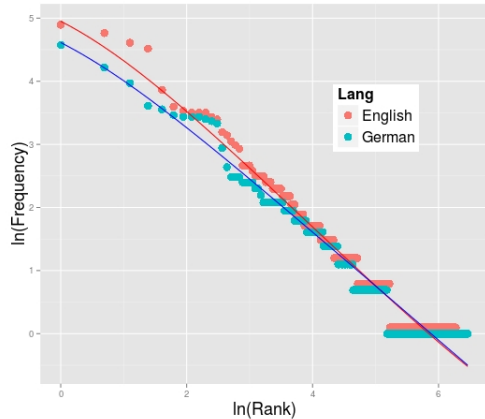


Quantitative measures

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

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

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

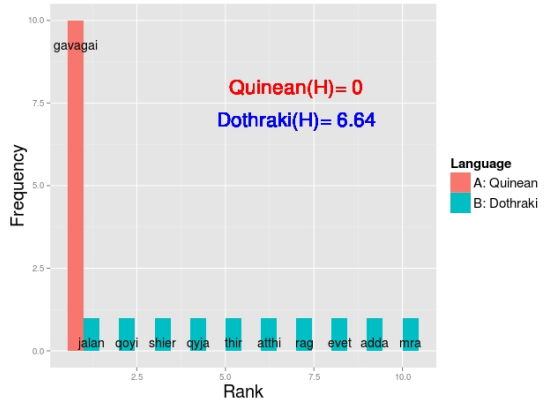


Quantitative measures

Shannon entropy
(Shannon & Weaver,
1949)

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

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

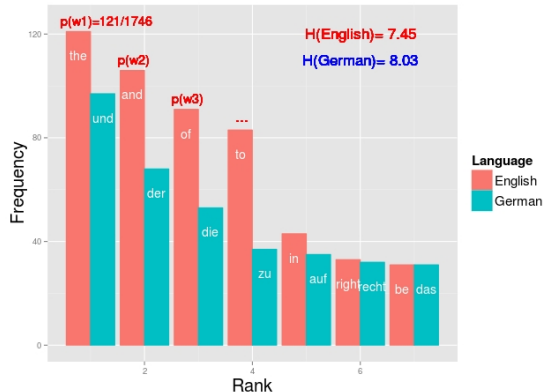


Quantitative measures

Shannon entropy
(Shannon & Weaver,
1949)

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

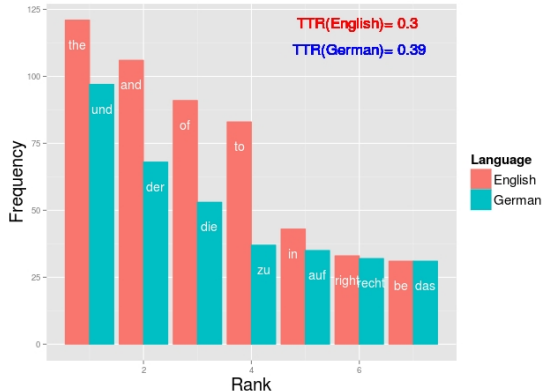
p_i : $\frac{\text{frequency of } w_i}{\text{total number of tokens}}$



Quantitative measures

Type-Token Ratio (TTR)

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



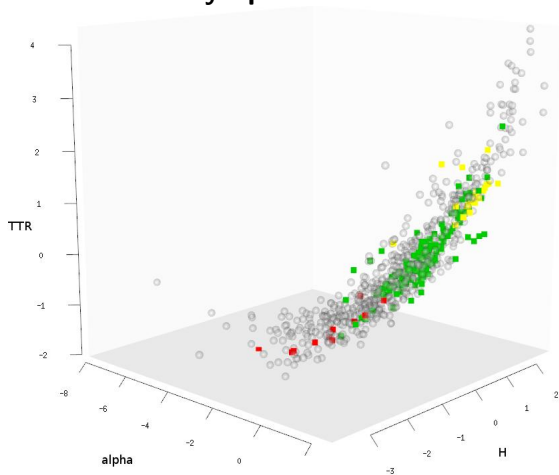
Scaled values for 647
languages of 83 families
(PBC, UDHR, EPC)

Altaic
Indo-European
Creole

Lexical Diversity Space

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

Altaic
Indo-European
Creole



Hypothesis

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


Statistical Model

Predicting lexical diversity from L2 proportions

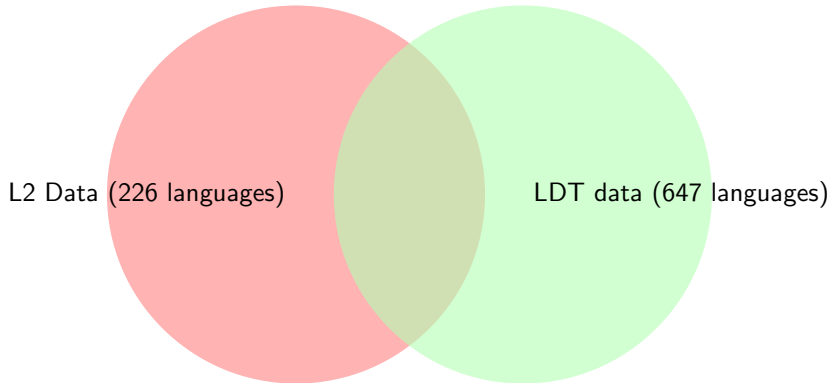
- requires **linear regression**:
continuous dependent/outcome variable: LDTs scaled
continuous predictors: L2 proportions (fixed effect)
- requires **mixed-effects** (random and fixed effects) due to non-independence of data points (family, region, text type, LDT measure) (Baayen et al., 2008; Bates et al., 2014; Jäger et al., 2011)

Statistical Model: Data Overlap

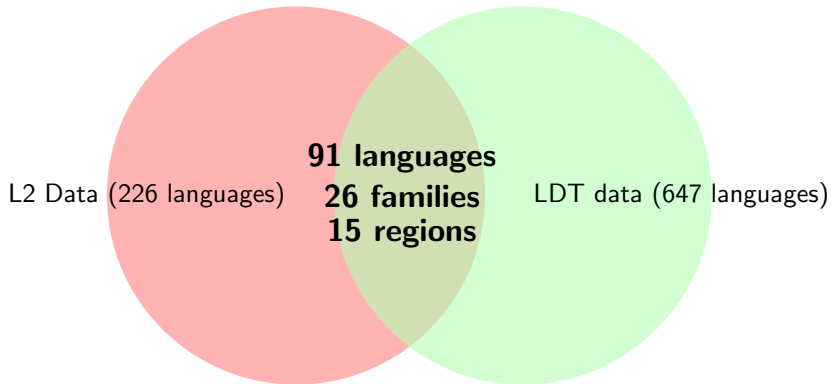
L2 Data (226 languages)

A large, solid red circle is centered on the slide. To its left, the text "L2 Data (226 languages)" is written in a black, sans-serif font. The circle is a uniform light red color and has no internal details or borders.

Statistical Model: Data Overlap



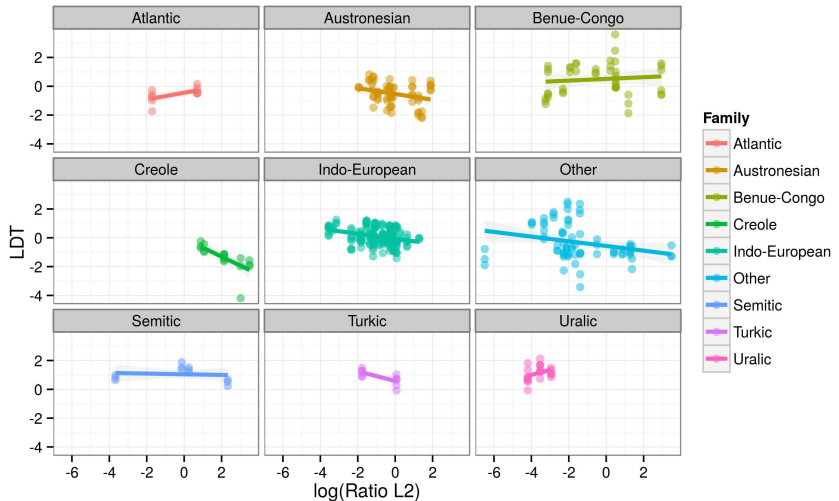
Statistical Model: Data Overlap



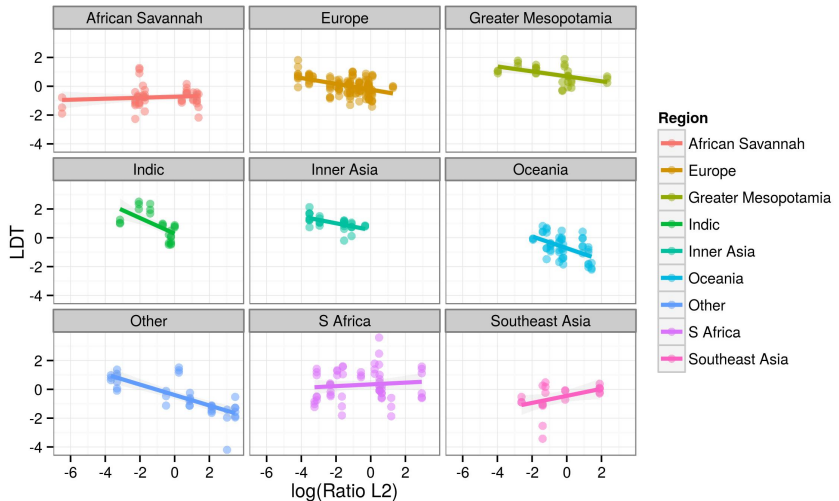
Results

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

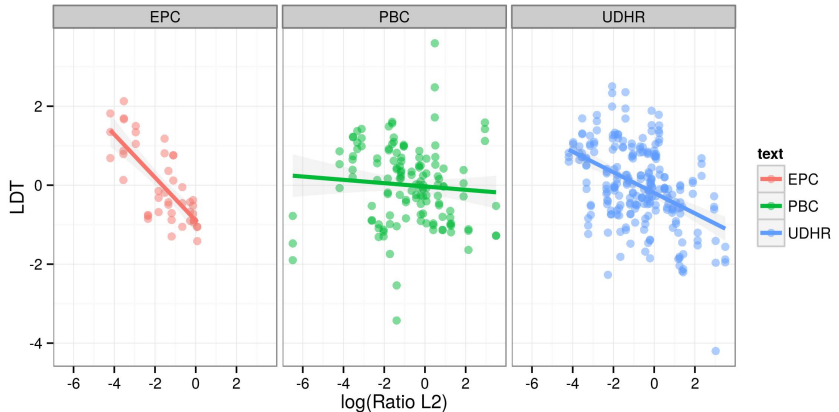
LDT and L2 proportions across families



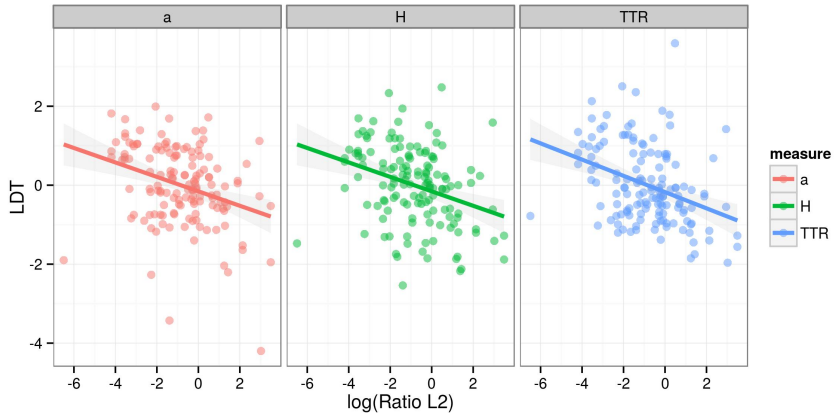
LDT and L2 proportions across regions



LDT and L2 proportions across text types



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

Further Question

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

Further Question

- 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)

Further Question

- 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)
- Is there a trade-off between range of word forms and flexibility of word order?

Lexical Diversity and Word Order

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

Lexical Diversity and Word Order

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

Lexical Diversity and Word Order

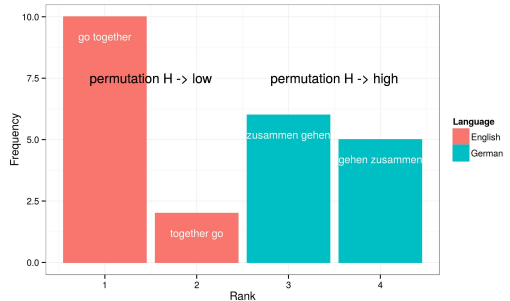
- **Permutation entropy:** Reflects the mutual word order flexibility in ngrams, i.e. word sequences (Zhang et al. 2006; Ramisch et al. 2008)
- $PE = 0 \rightarrow$ word order is *fixed*
- $PE = 1 \rightarrow$ word order is *free*

Permutation entropy (PE)

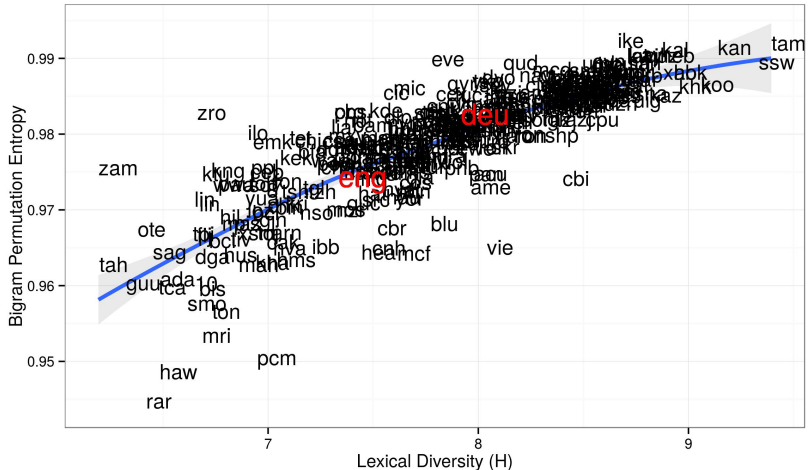
- $PE = 0$
→ word order is *fixed*
- $PE = 1$
→ word order is *free*

Permutation entropy (PE)

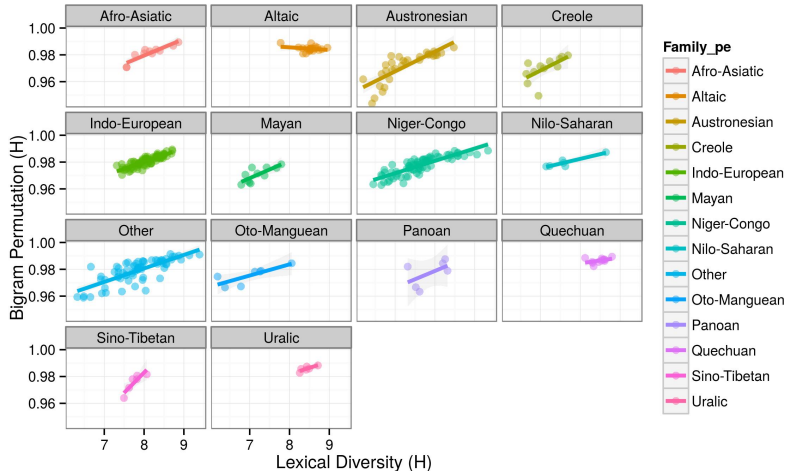
- $PE = 0$
→ word order is *fixed*
- $PE = 1$
→ word order is *free*



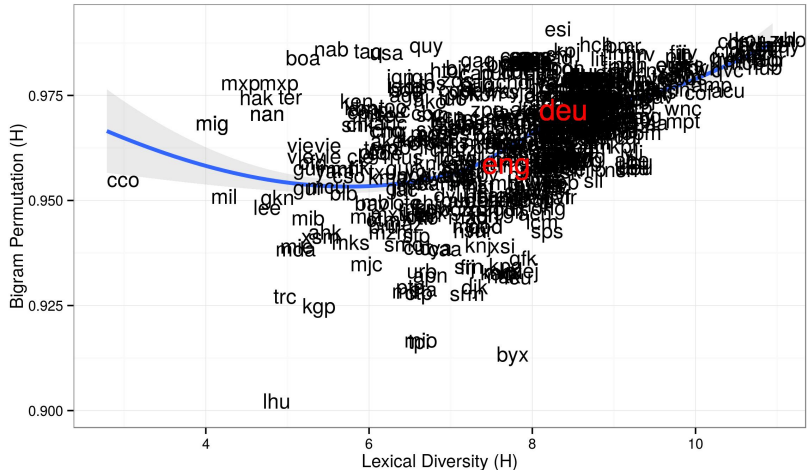
LDT (measured in H) versus PE (average) for UDHR



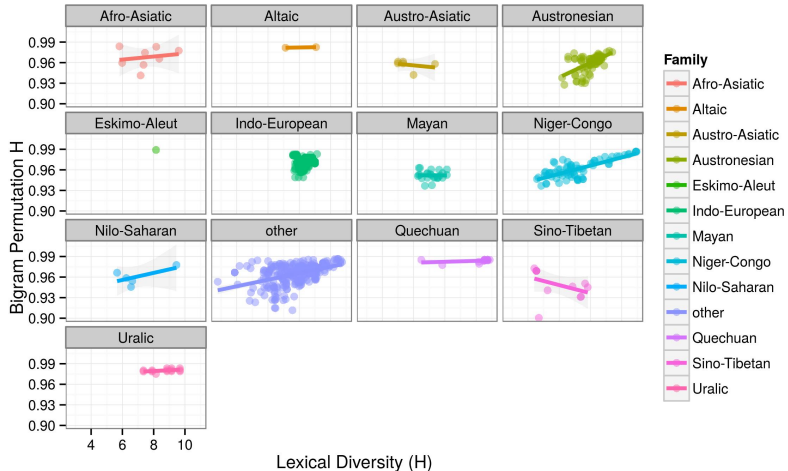
LDT (measured in H) versus PE (average) for UDHR (Families)



LDT (measured in H) versus PE (average) for Bible



LDT (measured in H) versus PE (average) for Bible (Families)



Conclusions

Our statistical analyses suggest:

Conclusions

Our statistical analyses suggest:

- Languages with **higher L2 proportions** tend to have fewer word forms, i.e. **lower lexical diversities**

Conclusions

Our statistical analyses suggest:

- Languages with **higher L2 proportions** tend to have fewer word forms, i.e. **lower lexical diversities**
- This effect holds for most **families** and **regions**, and for all **text types** and **measures** (in our sample)

Conclusions

Our statistical analyses suggest:

- Languages with **higher L2 proportions** tend to have fewer word forms, i.e. **lower lexical diversities**
- This effect holds for most **families** and **regions**, and for all **text types** and **measures** (in our sample)
- There is evidence that lexical diversities are strongly, positively correlated with **permutation entropies** (word orders?)

Conclusions

Potential implications for language change:

Conclusions

Potential implications for language change:

- Higher **L2 proportions** → fewer **word forms** → fixed **word order**?

Conclusions

Potential implications for language change:

- Higher **L2 proportions** → fewer **word forms** → fixed **word order**?
- Languages trade off information encoding strategies (word forms vs. word order) (?)
- Languages are **adaptive systems** shaped by the cognitive niche of **speaker populations**

Collaborators



Douwe Kiela



Felix Hill



Andrew Caines



Dimitrios Alikaniotis



Paula Buttery

Thank You!

jchris@christianbentz.dej