Learning pressures simplify morphology: Corpus, computational and experimental evidence

Christian Bentz
University of Tübingen

Aleksandrs Berdicevskis
The Arctic University of Norway

December 8, 2016

WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

EVOLAEMP
Language Evolution: The Empirical Turn

Overview

Introduction

Corpus Analyses
- Quantitative Measures
- Corpora
- Entropy Estimation

Lemmatization
- Method
- Corpora
- Results

Iterated Learning
- Experiments
- Results

Morphological Diversity

(1) Hawaiian (Austronesian)

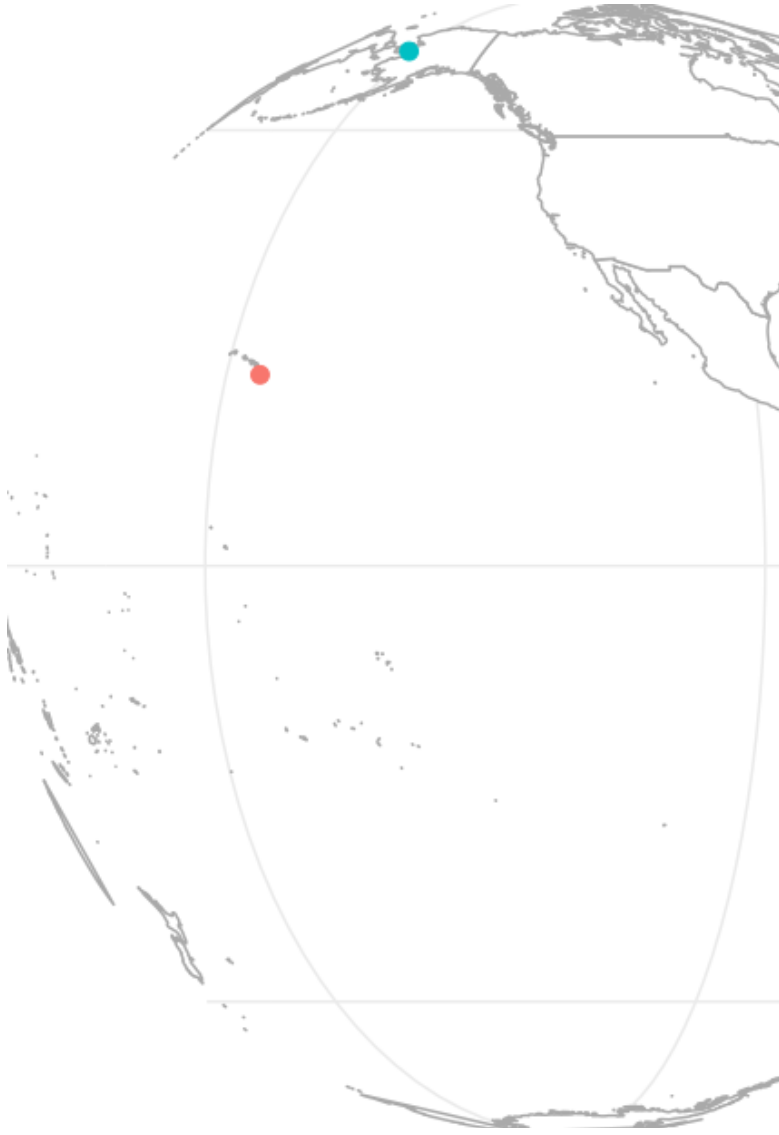
A ua olelo aku o Ioane ia ia [...]
Then PERF say.to SUBJ.Johan he.DAT [...]

“Then Johan said to him [...]"

(2) Iñupiatun (Eskimo-Aleut)

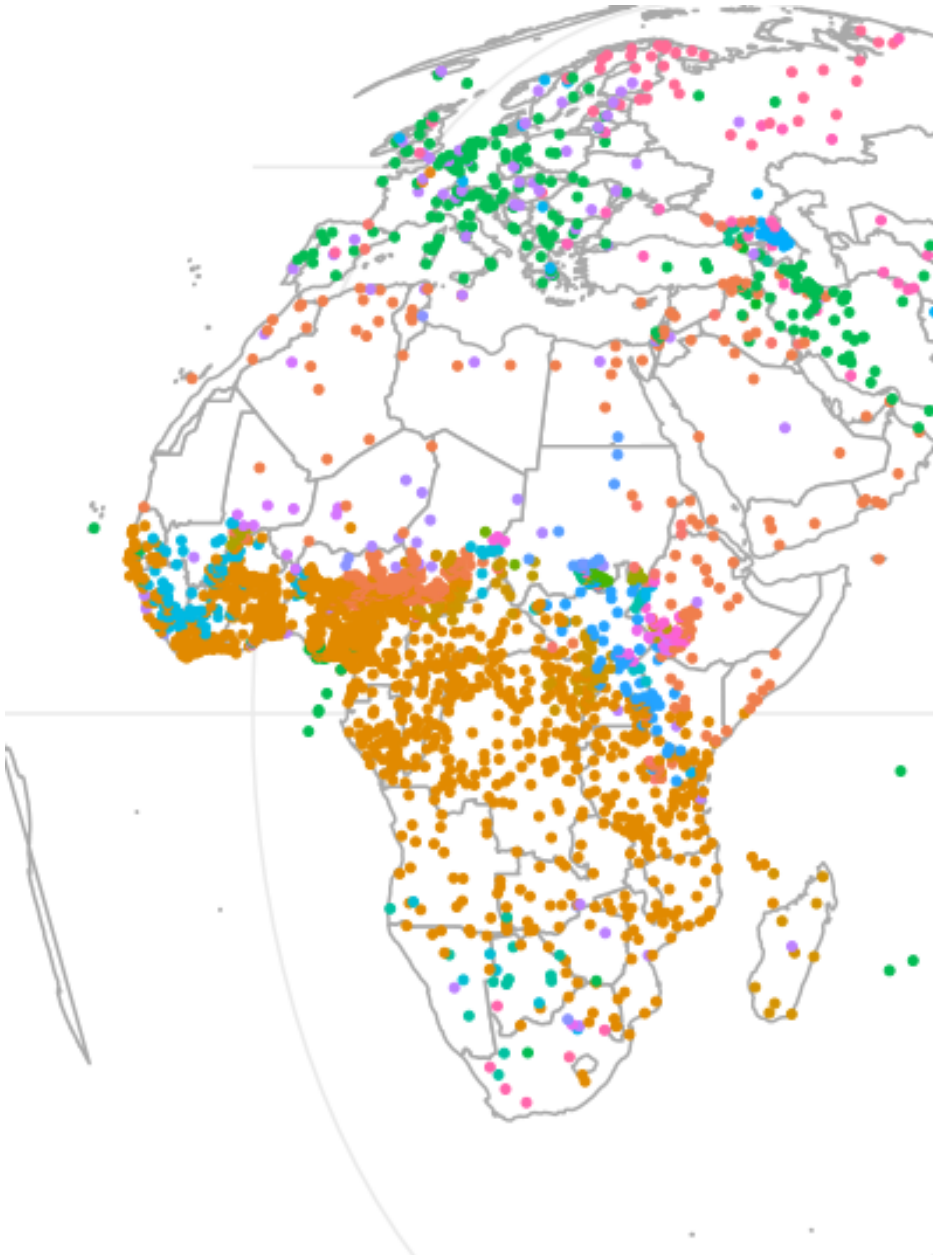
Aglaan Jesus-ngum itnaġniġai [...]
But Jesus-ERG this.say.report.3S.to.3PL

"But Jesus said to them [...]"

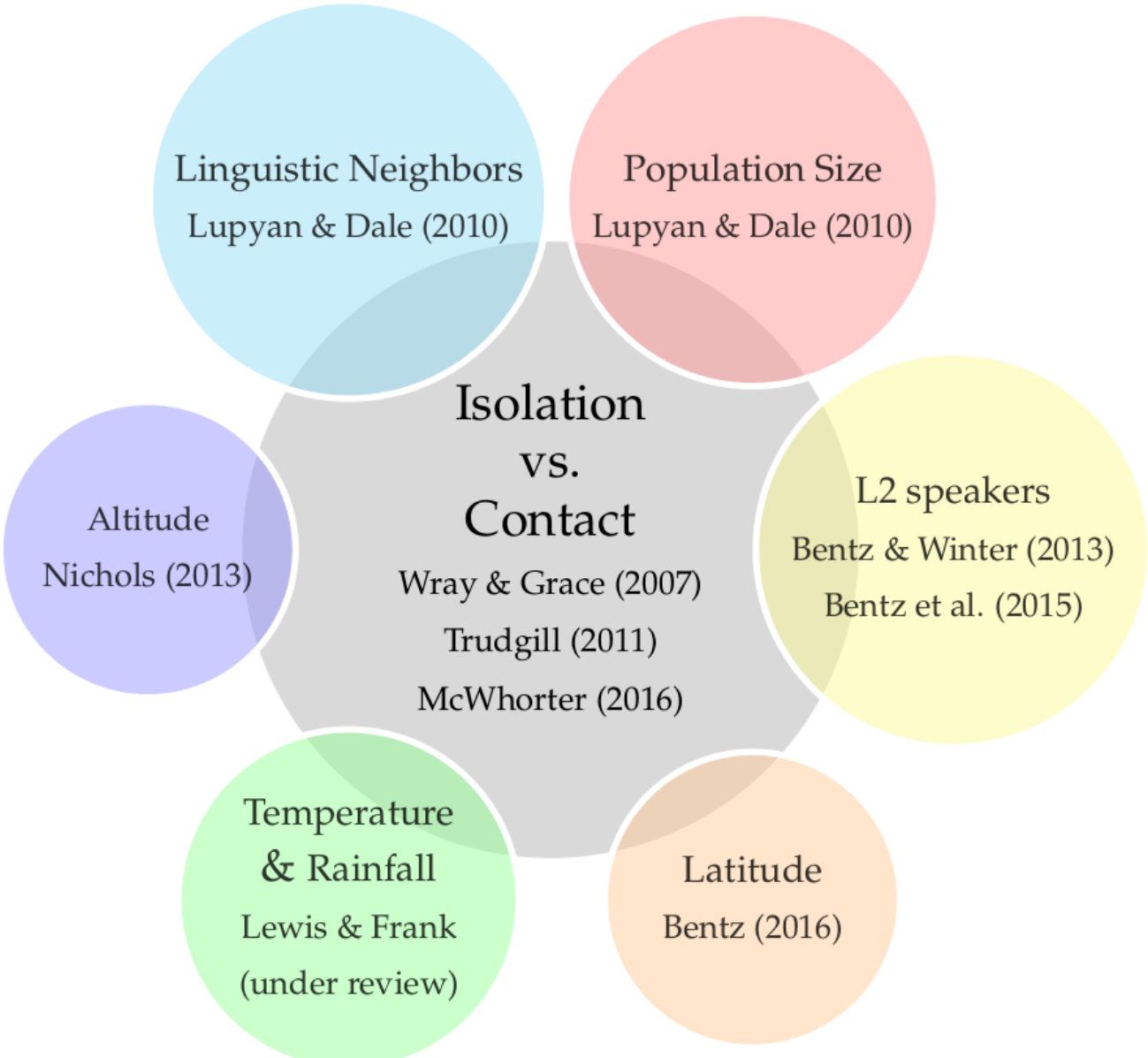


Typological Questions

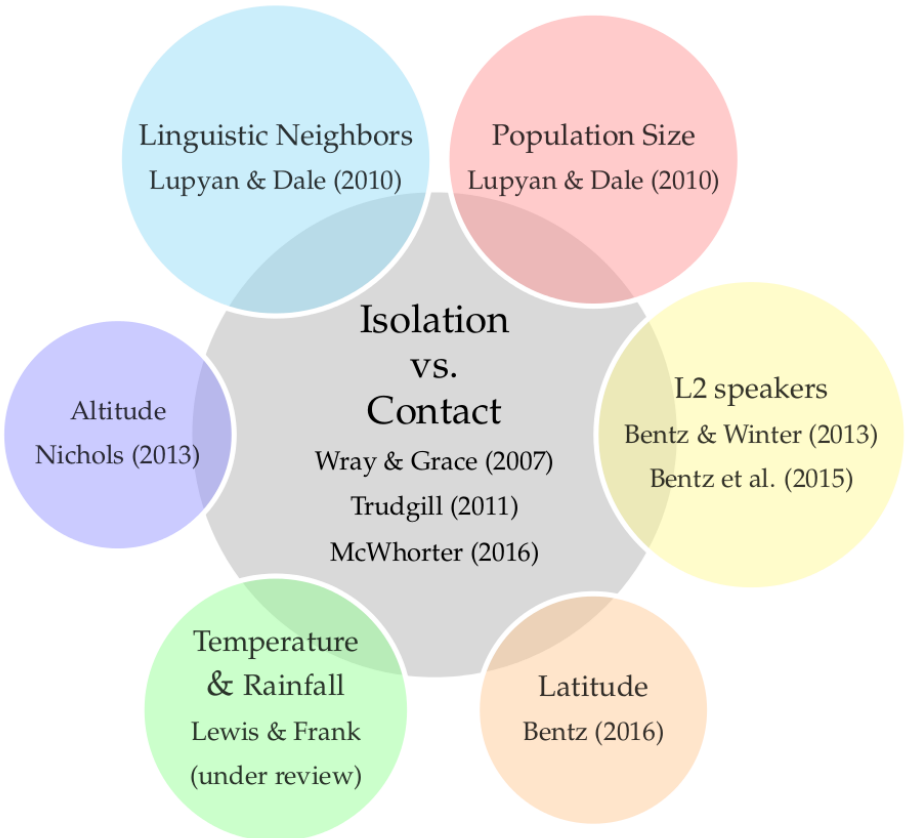
- How do we measure these differences in complexity?
- Are there systematic explanations for the patterns we find across the world?



"External" Factors



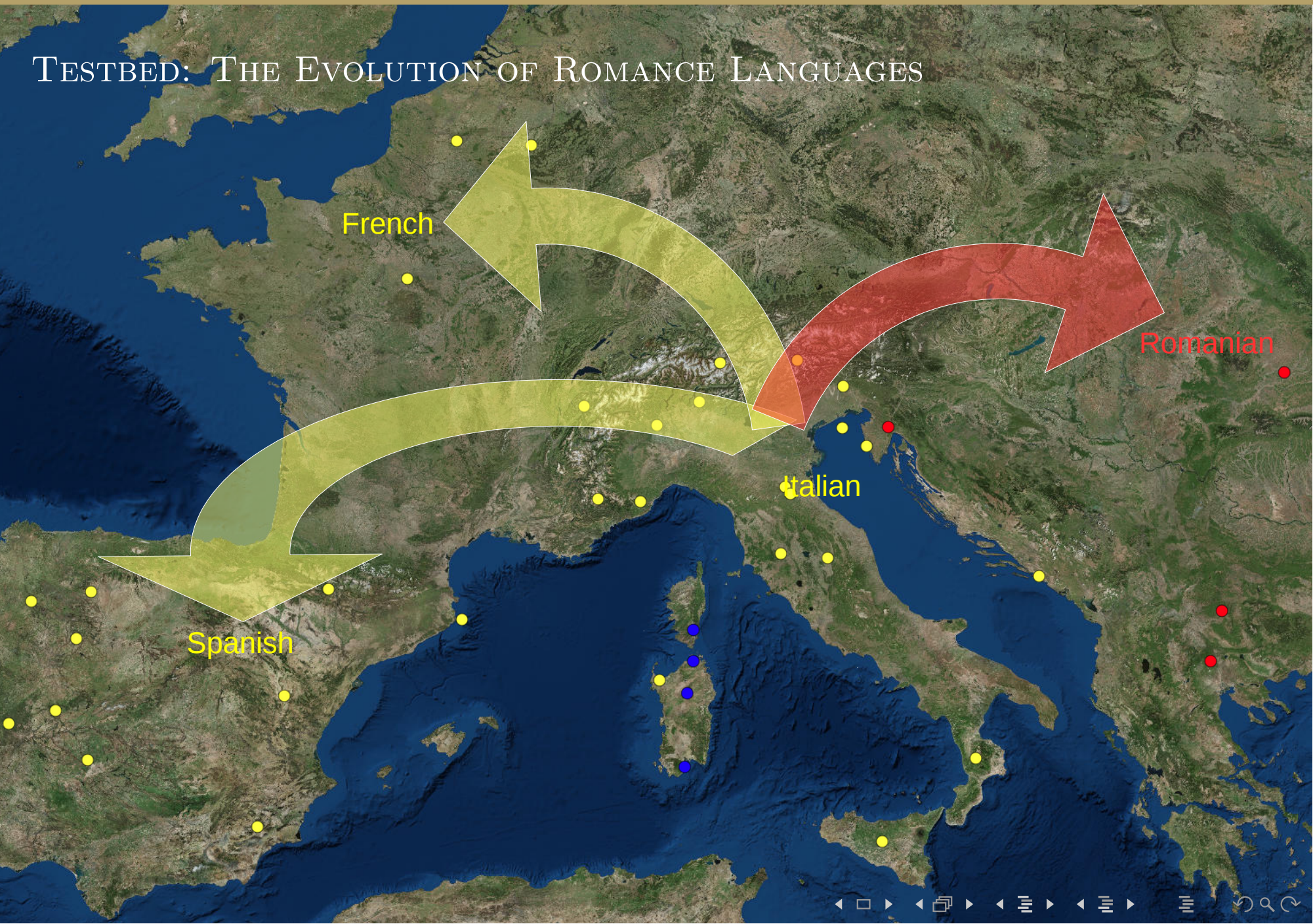
Missing Link



Quantitative Studies
(Corpora, Typological Data)

Language Learning and
Transmission (Experiments)

Testbed: The Evolution of Romance Languages



Testbed: The Evolution of Romance Languages

- As **Vulgar Latin** varieties spread throughout Europe, their **morphological complexity** was reduced (Herman & Wright 2000)
- This is argued to be (partly) due to **L2 contact** (Herman & Wright 2000, Bentz & Christiansen 2013)

Simple Example: Word for "Brother" in the Bible

- **Latin**

  01004008 Dixitque Cain ad Abel **fratrem** suum [...]
  01004009 Ubi est Abel **frater** tuus?
  01004011 [...] suscepit sanguinem **fratris** tui de manu tua!

Simple Example: Word for "Brother" in the Bible

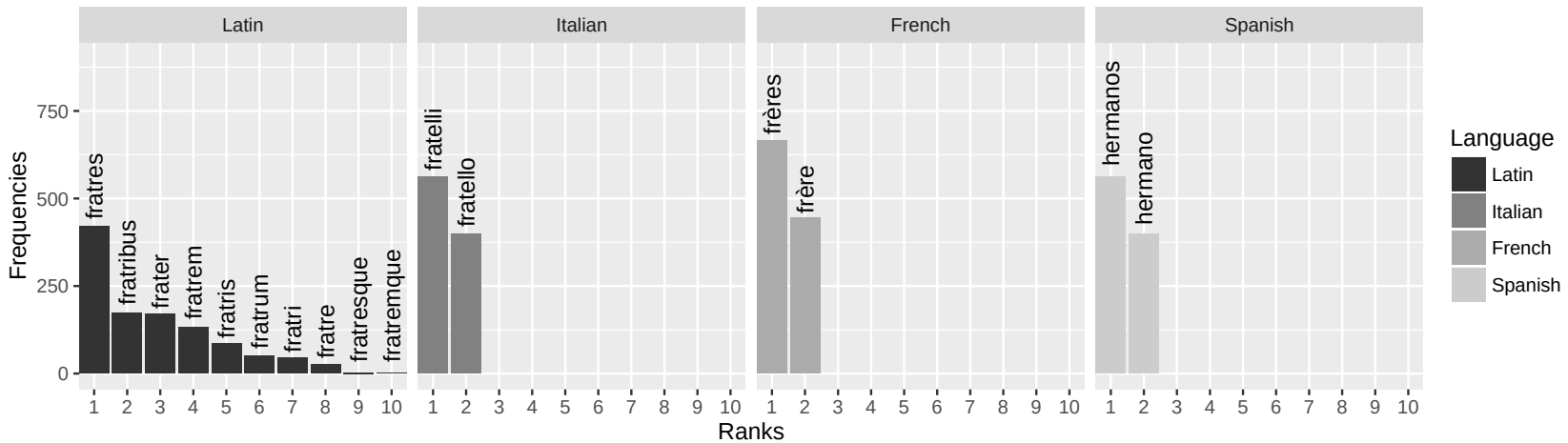
- **Latin**

  01004008 Dixitque Cain ad Abel **fratrem** suum [...]
  01004009 Ubi est Abel **frater** tuus?
  01004011 [...] suscepit sanguinem **fratris** tui de manu tua!

- **Italian**

  01004008 Caino disse al **fratello** Abele [...]
  01004009 Dov'è Abele , tuo **fratello**?
  01004011 [...] ha bevuto il sangue di tuo **fratello!**

Simple Example: Word for "Brother" in the Bible

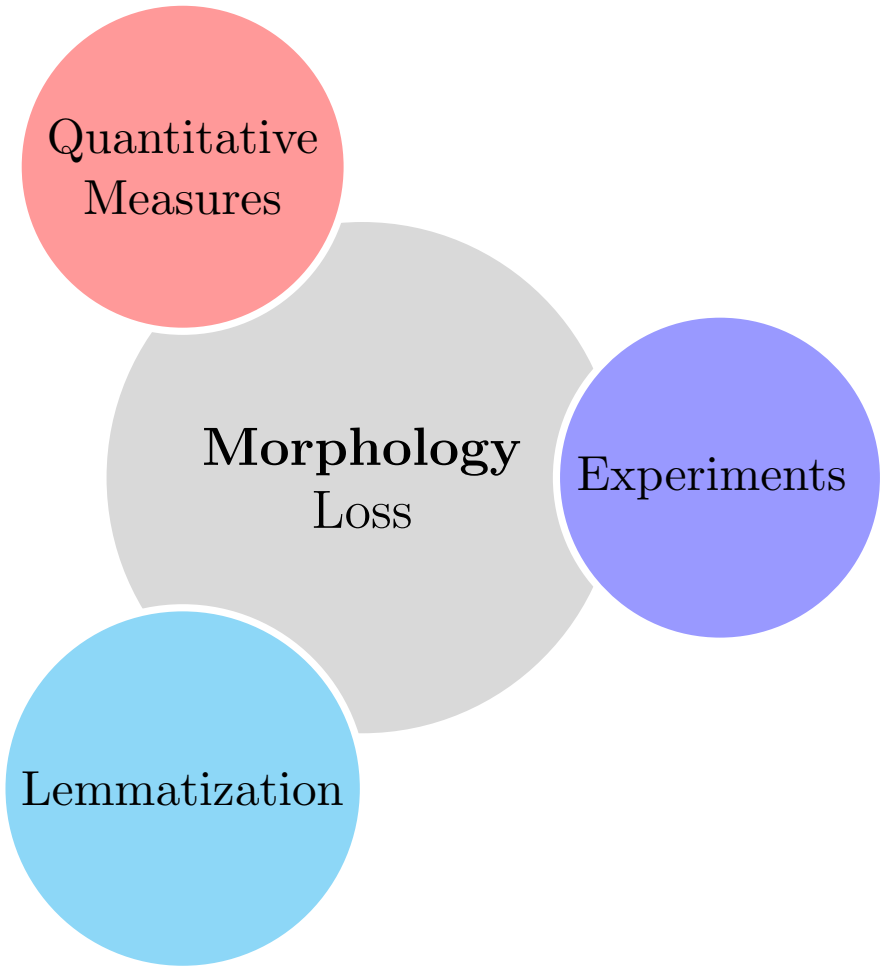


Converging Evidence

Bentz et al. (2015)

Bentz & Alikaniotis (2016)

Ferrer-i-Cancho & Bentz (forthcoming)



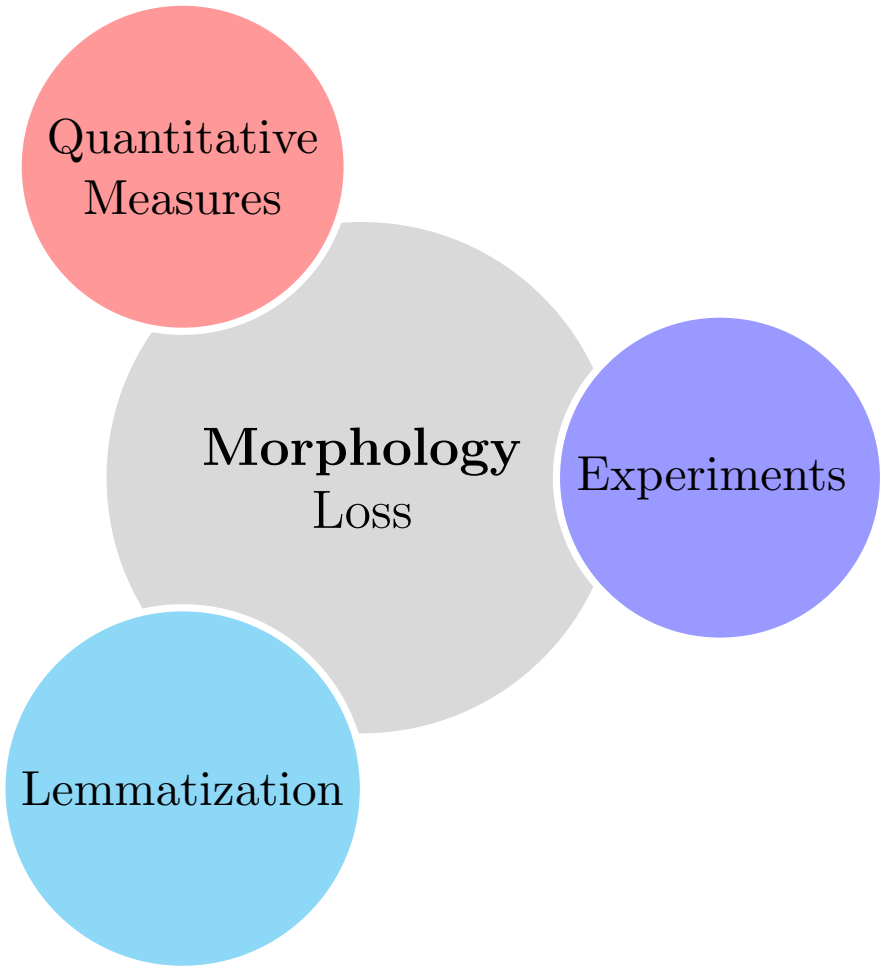
Bentz, Alikaniotis, Samardžić & Buttery (forthcoming)

Converging Evidence

Bentz et al. (2015)

Bentz & Alikaniotis (2016)

Ferrer-i-Cancho & Bentz (forthcoming)



Bentz, Alikaniotis, Samardžić & Buttery (forthcoming)

Kirby et al. (2008, 2015)

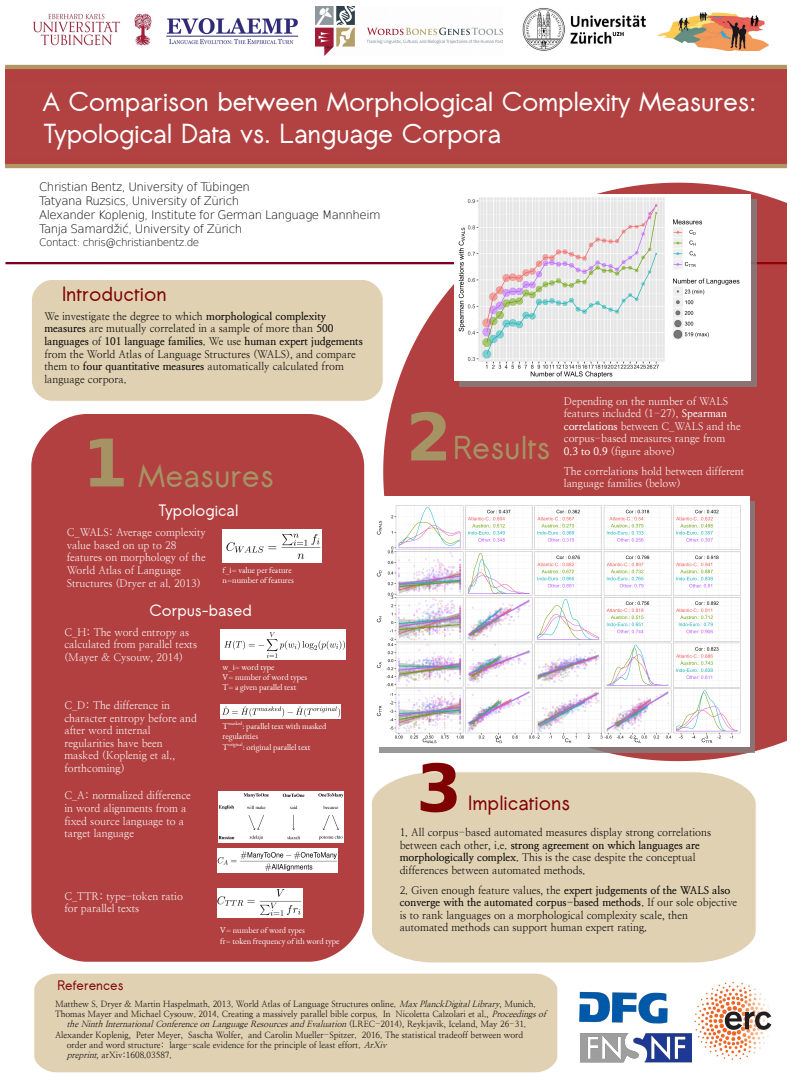
Berdicevskis (2012)

Berdicevskis & Semenuks (forthcoming)

Quantitative Measures

- sample of more than **500 languages** of 101 language families
- **4 corpus-based measures** compared to a measure based on **typological data (WALS)**
- strong Spearman correlations (up to 0.9) between all of them

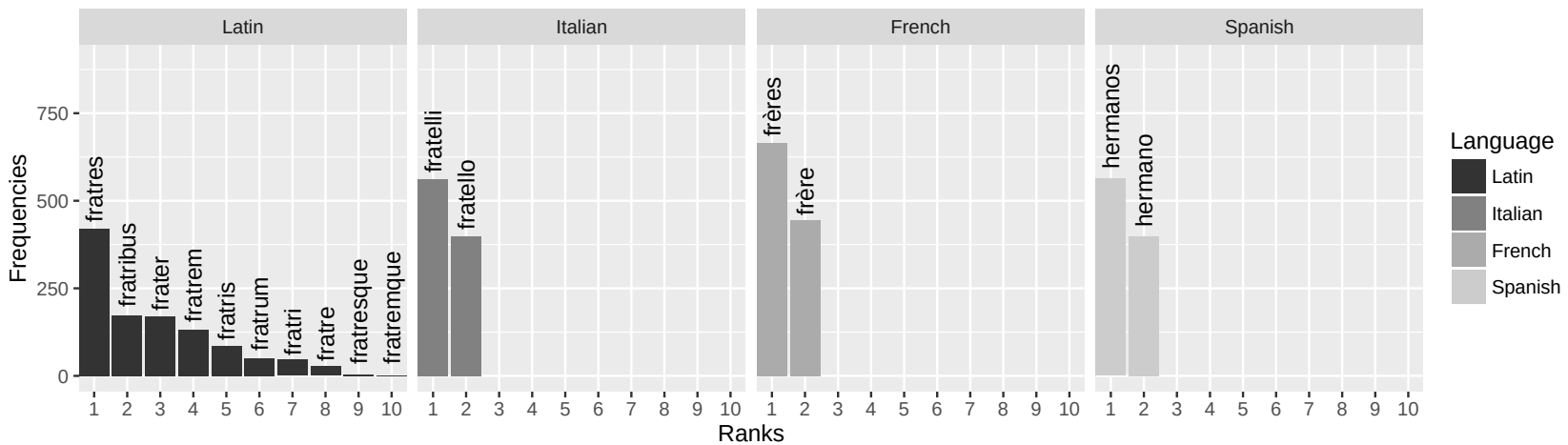
Bentz, Ruzsics, Koplenig & Samardžić (2016)



Shannon Entropy

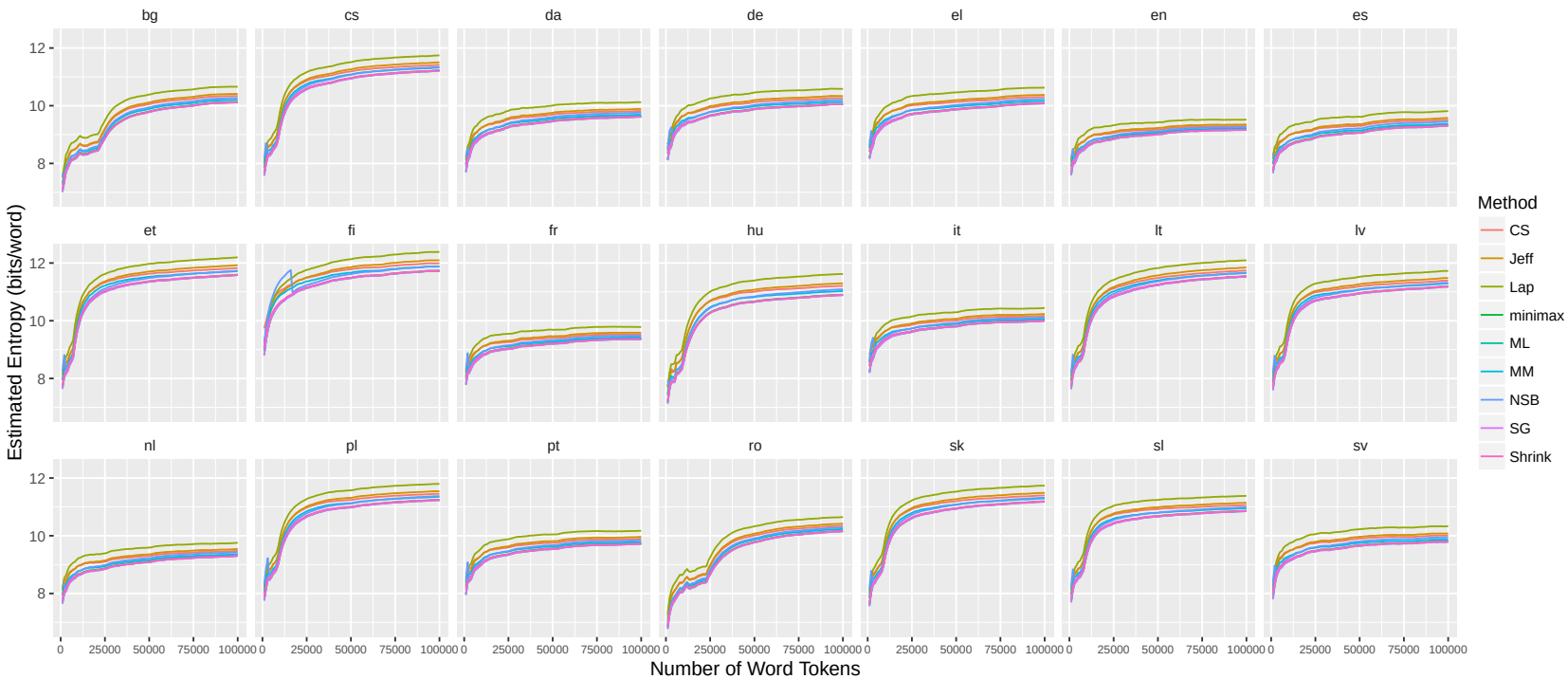
- Measure the skewness of the word form distribution via the **entropy H** according to Shannon (1949):

$$H(T) = -\sum_{i=1}^{V} p(w_i) \log_2(p(w_i)). \quad (1)$$



Conceptual Problem

- word entropy depends on text size
- $H(T)$ converges onto stable value at ca. 50K tokens



Bentz et al. (forthcoming)

Corpus Analyses

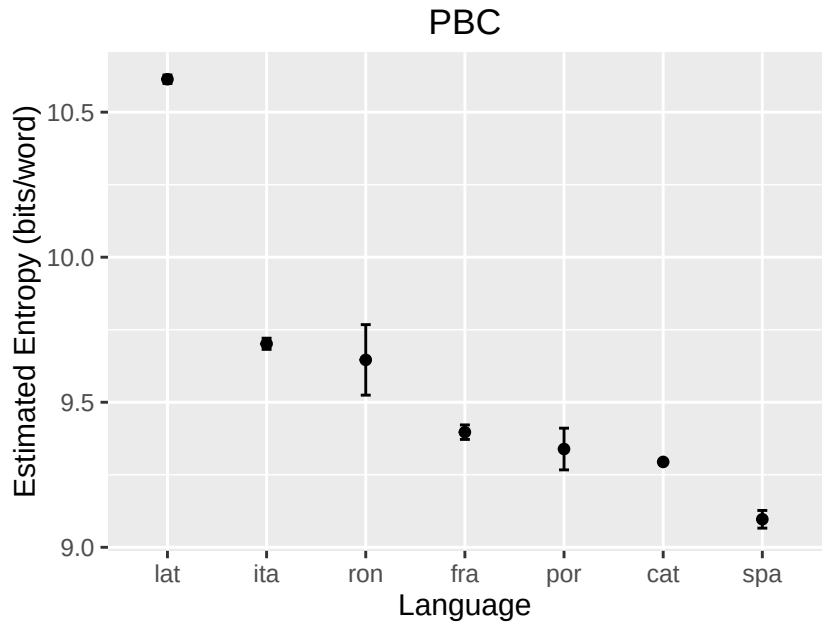
First Analysis

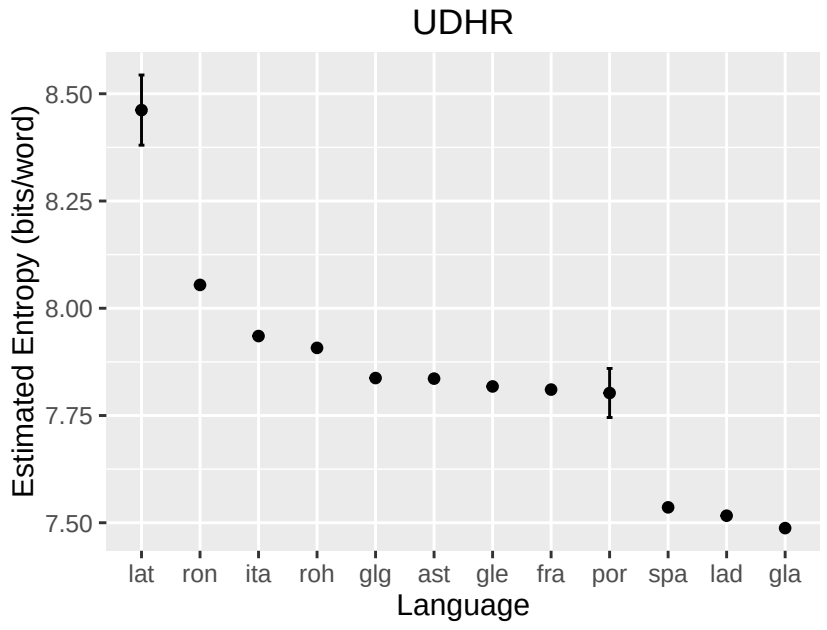
Measure the **entropy (bits/word) change** for **Latin towards Romance languages** using the so-called *James-Stein shrinkage estimator* (Hausser & Strimmer, 2014)

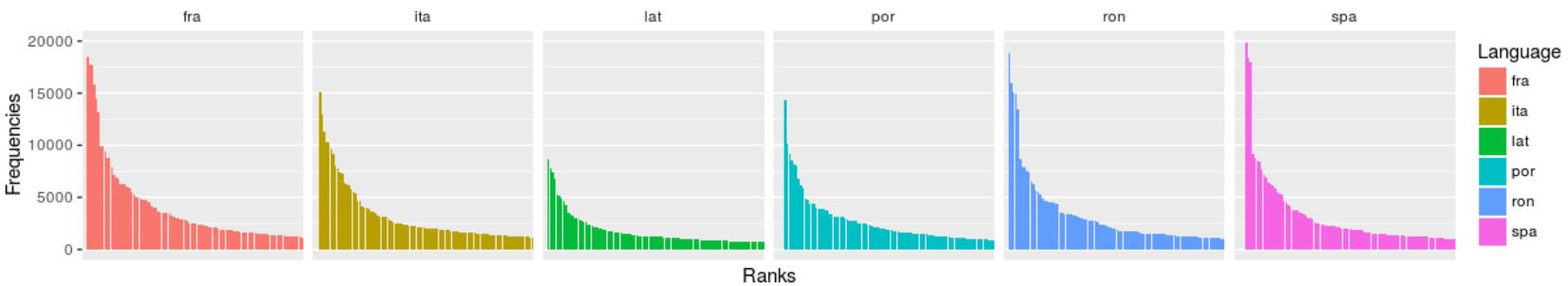
Corpus Analyses

Select all the **Romance languages** (+ Latin) from the *Parallel Bible Corpus* (7 languages), and the *Universal Declaration of Human Rights* (10 languages).

Results


PBC
Estimated Entropy (bits/word)
10.5
10.0
9.5
9.0
lat
ita
ron
fra
por
cat
spa
Language


UDHR
Estimated Entropy (bits/word)
8.50
8.25
8.00
7.75
7.50
lat
ron
ita
roh
glg
ast
gle
fra
por
spa
lad
gla
Language


fra
ita
lat
por
ron
spa
Frequencies
20000
15000
10000
5000
0
Ranks
Language
fra
ita
lat
por
ron
spa

Conclusion (Corpus Analyses)

- Entropy is reduced from **ca. 11 (bits/word) in Latin** to **ca. 9.75-9.0 (bits/word) in 6 Romance languages** of the PBC, i.e. by around **10-15%**
- A similar pattern is found for the UDHR though with overall lower entropy values (due to differences in text size)

Lemmatization

Second Analysis

Neutralize **inflectional marking** in Latin and the Romance languages to measure the effect of **inflectional differences** (Bentz, Alikaniotis, Samardžić & Buttery, in print)

TreeTagger (Schmidt 1994, 1995)

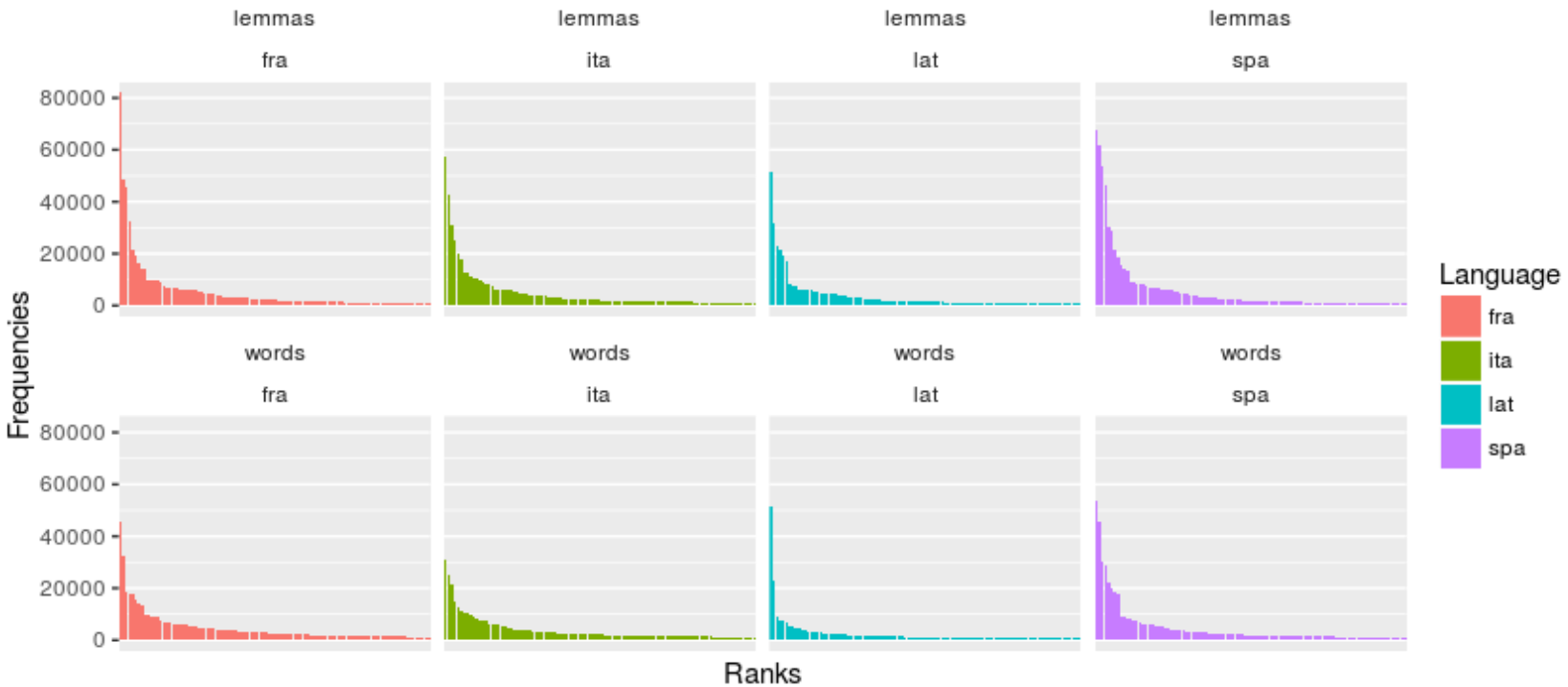
Input	→	**Tag**	**Lemma**
fratrem	→	N:acc	frater
fratris	→	N:gen	frater
fratribus	→	N:dat	frater
vivit	→	V:IND	vivo
movetur	→	V:IND	moveo
humanum	→	ADJ	humanus

...

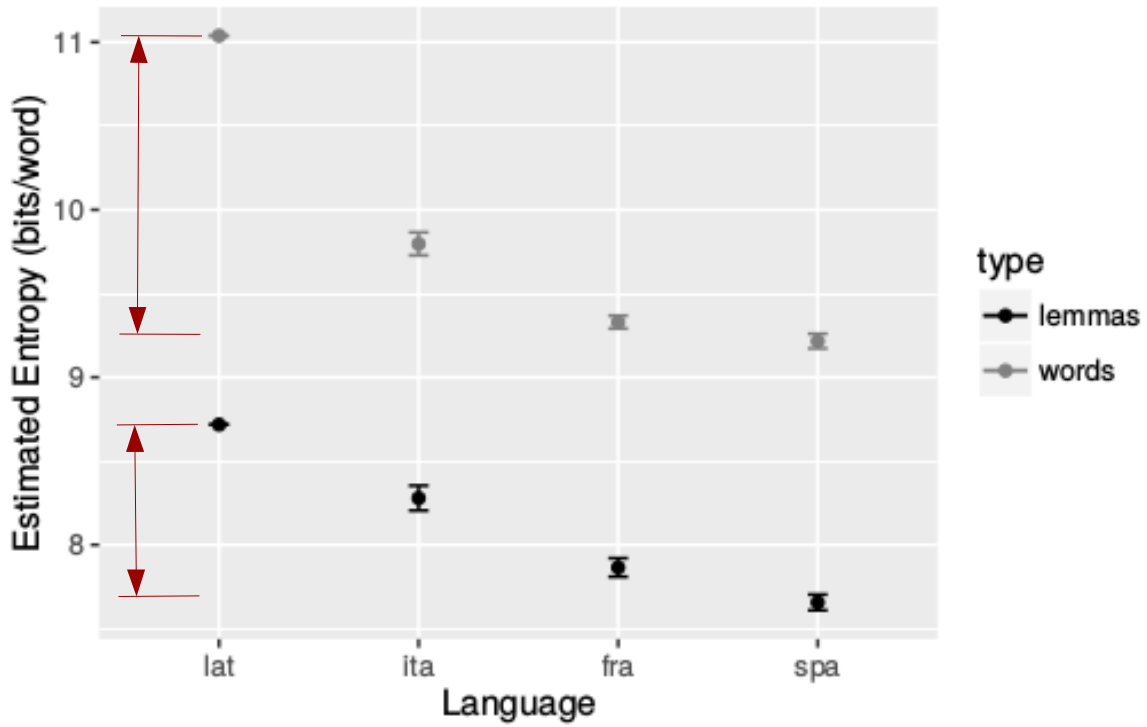
Corpora

Select the **Romance languages** (+ Latin) - which can be **lemmatized with the TreeTagger** - from the *Parallel Bible Corpus* (4 languages: Latin, French, Spanish, Italian)

Results (Lemmatization)



Results (Lemmatization)



Conclusions (Lemmatization)

- In **Latin, Italian, Spanish and French** entropy (bits/word) is reduced by **ca. 15-20% through lemmatization**, i.e. when inflectional marking is neutralized.
- The **entropy difference to Latin** is reduced via **lemmatization by ca. 50%**

Iterated Learning Experiments

Third Analysis

Illustrate via **iterated learning experiments** how inflectional marking is lost through **learning pressures (non-native, i.e. L2)** over several generations

Experimental Design (Berdicevskis & Semenuks, forthcoming)

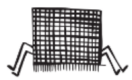
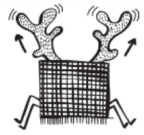
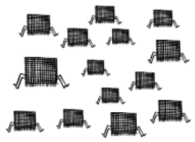
Setup

- **artificial language learning** task
- Overall **300 participants**
  (3 types of chains × 10 generations × 10 subjects)
- native speakers of **Russian**
- experiment on **webpage** (*jsPsych* javascript)

Artificial Language: Epsilon

(Berdicevskis & Semenuks, forthcoming)

Artificial Language: Epsilon

		event: none	event: fall apart	event: grow antlers	event: fly
agent: round animal	number: singular	 seg_N	 seg_N m_V-o_{AGR}	 seg_N r_V-o_{AGR}	 seg_N b_V-o_{AGR}
	number: plural	 seg_N-l_{PL}	 seg_N-l_{PL} m_V-o_{AGR}	 seg_N-l_{PL} r_V-o_{AGR}	 seg_N-l_{PL} b_V-o_{AGR}
agent: square animal	number: singular	 fuv_N	 fuv_N m_V-i_{AGR}	 fuv_N r_V-i_{AGR}	 fuv_N b_V-i_{AGR}
	number: plural	 fuv_N-l_{PL}	 fuv_N-l_{PL} m_V-i_{AGR}	 fuv_N-l_{PL} r_V-i_{AGR}	 fuv_N-l_{PL} b_V-i_{AGR}

(Berdicevskis & Semenuks, forthcoming)

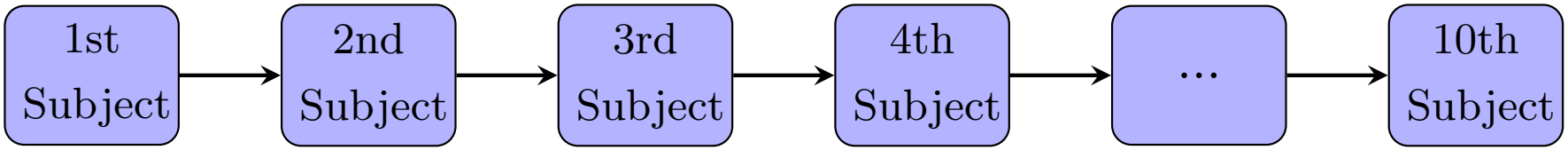
Artificial Language: Epsilon

"This system **resembles the more complex Russian morphosyntactic system** where nouns are marked for number, and adjectives and verbs agree with nouns in number and gender. The bottom line is that agreement is salient and pervasive in Russian morphosyntax, and thus the **mother tongue is not imposing pressure** on the participants to shed agreement." (Berdicevskis & Semenuks, forthcoming)

Experimental Design (Berdicevskis & Semenuks, forthcoming)

Condition I

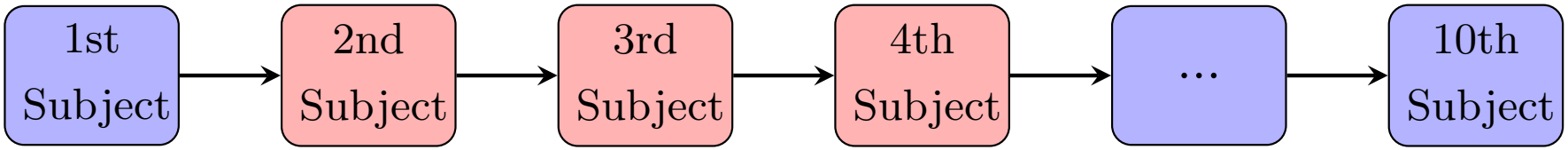
Set of **10 uninterrupted chains** of 10 "generations", i.e. subjects



Experimental Design (Berdicevskis & Semenuks, forthcoming)

Condition II

Set of **10 temporarily interrupted chains**

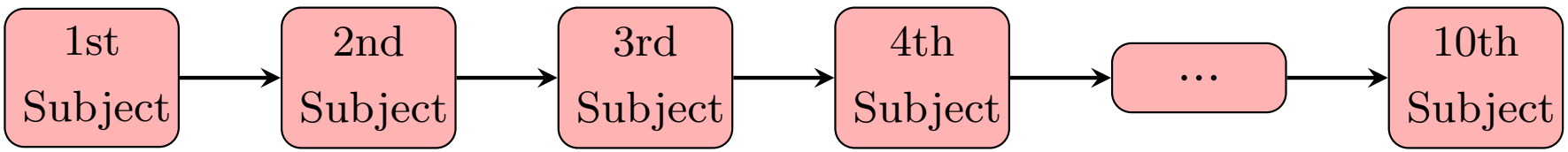


Interruption (i.e. L2 influence) is here introduced via less exposure to the target language in the learning phase (6 versus 3 learning blocks).

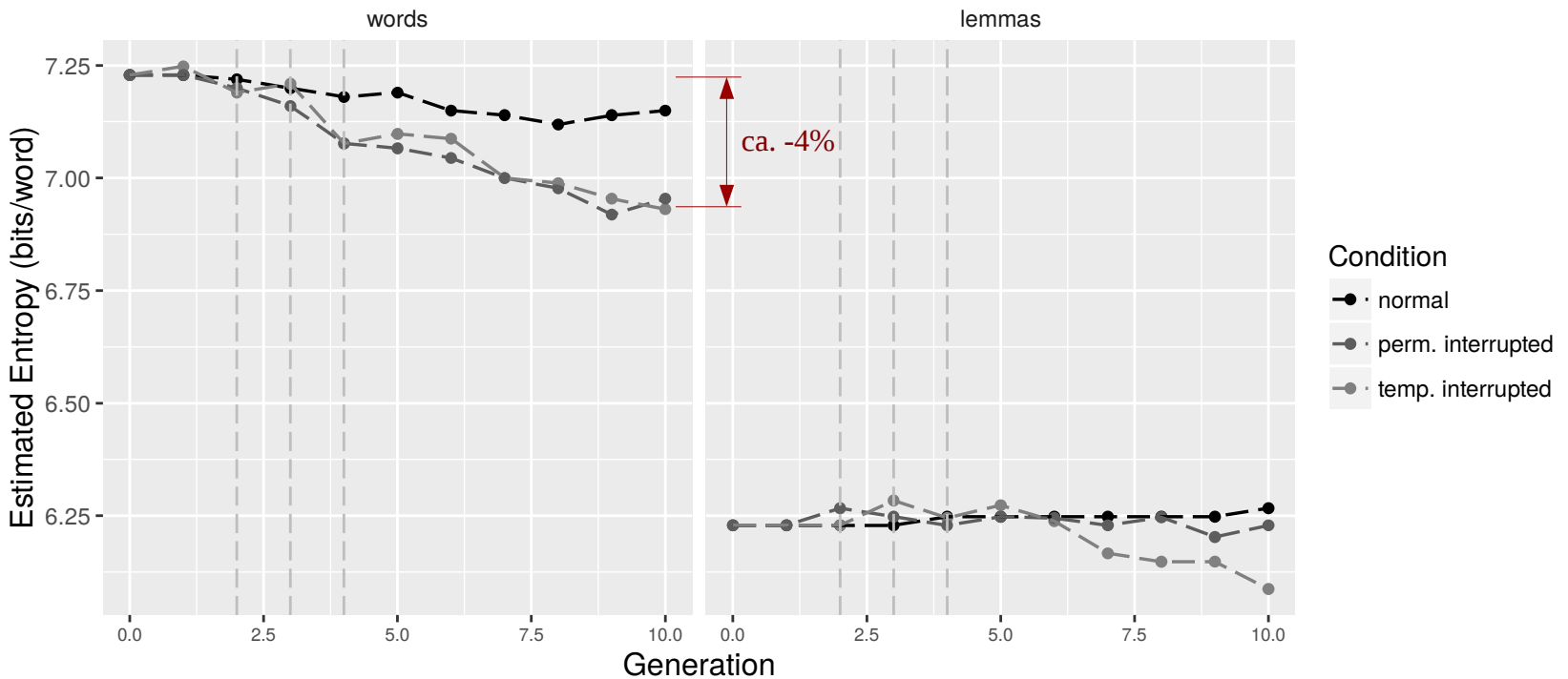
Experimental Design (Berdicevskis & Semenuks, forthcoming)

Condition III

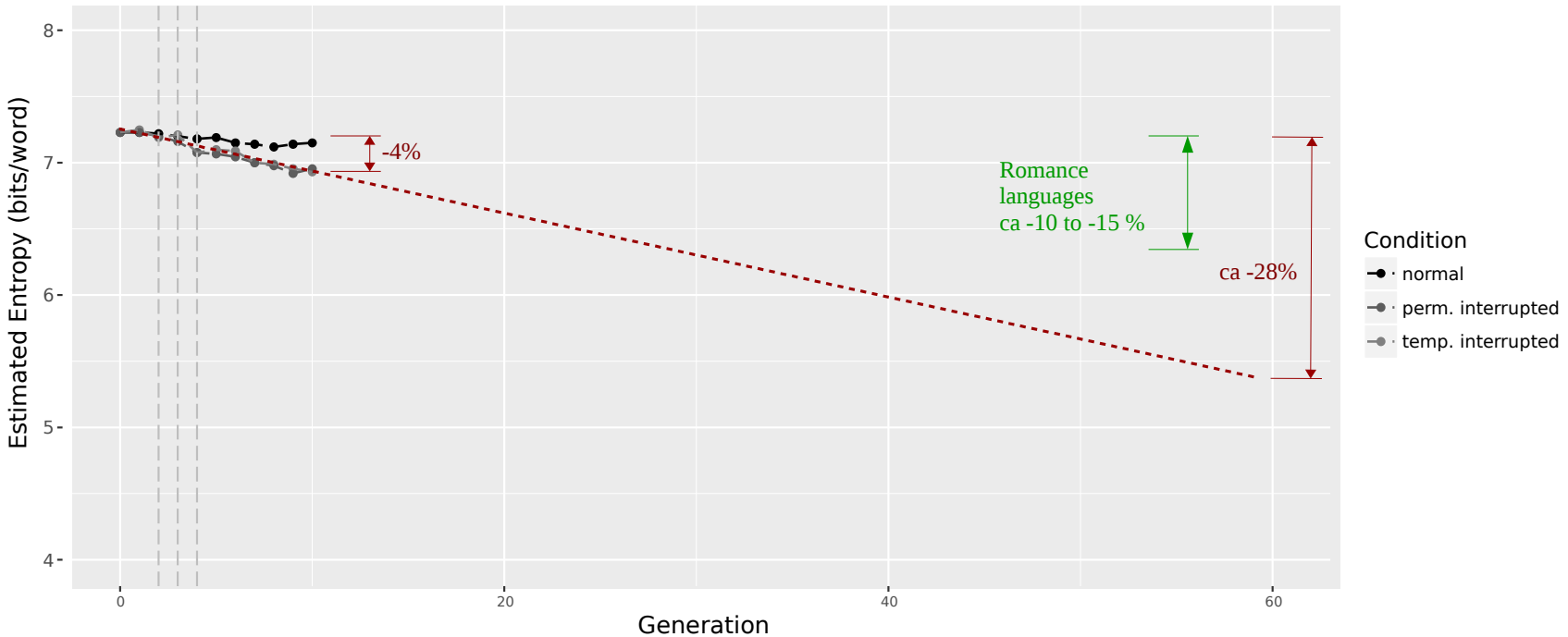
Set of **10 permanently interrupted chains**



Results: Entropy in Epsilon



Results: "Extrapolation"

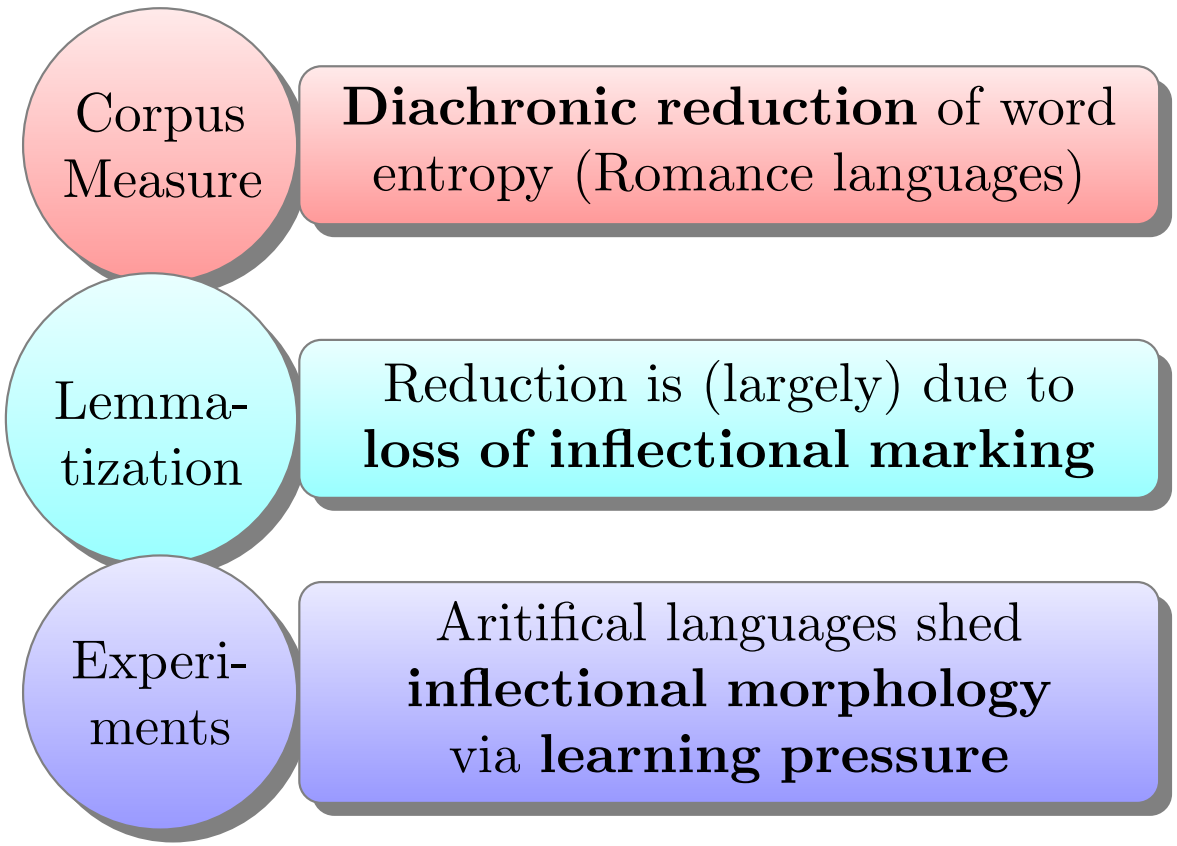


60 generations $\sim$ 1800 years

Conclusions (Iterated Learning)

- in the interrupted conditions word **entropy is reduced by around 4%** in 10 generations
- this reduction is mostly due to **loss of inflection** rather than **base vocabulary**

Conclusions



Thank You!