Towards a Computational Model of Grammaticalization and Lexical Diversity

Christian Bentz & Paula Buttery
cb696@cam.ac.uk

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Outline

Background

- Lexical diversity
- Grammaticalization

Computational Model

- Architecture
- Outcome

Future Directions

- Model improvement

Lexical diversity

Definition

Definition: The **distribution of word forms** used to encode a **constant information content**

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Lexical diversity

Definition

Definition: The **distribution of word forms** used to encode a **constant information content**

Parallel texts:

- Universal Decalaration of Human Rights ($\sim$ 400 languages)
- Parallel Bible Corpus ($\sim$ 1000 languages)
- Europarl (21 languages)

Lexical diversity

Driving factors

- **Morphological marking**
  English: *the ship*
  German: *das Schiff, dem Schiff(e), des Schiffes*

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Lexical diversity

Driving factors

- **Morphological marking**
  English: *the ship*
  German: *das Schiff, dem Schiff(e), des Schiffes*
- **Compounding**
  English: *key to the cabin of the captain of the ship*
  German: *Schifffahrtskapitaenkabinenschluessel*

Lexical diversity

Driving factors

- **Morphological marking**
  English: *the ship*
  German: *das Schiff, dem Schiff(e), des Schiffes*
- **Compounding**
  English: *key to the cabin of the captain of the ship*
  German: *Schifffahrtskapitaenkabinenschluessel*
- **Lexicon**
  English: *close*
  German: *zuschliessen, abschliessen*

Lexical diversity

Driving factors

- **Morphological marking**
  English: *the ship*
  German: *das Schiff, dem Schiff(e), des Schiffes*
- **Compounding**
  English: *key to the cabin of the captain of the ship*
  German: *Schifffahrtskapitaenkabinenschluessel*
- **Lexicon**
  English: *close*
  German: *zuschliessen, abschliessen*
- **Orthography**
- **etc.**

Lexical diversity

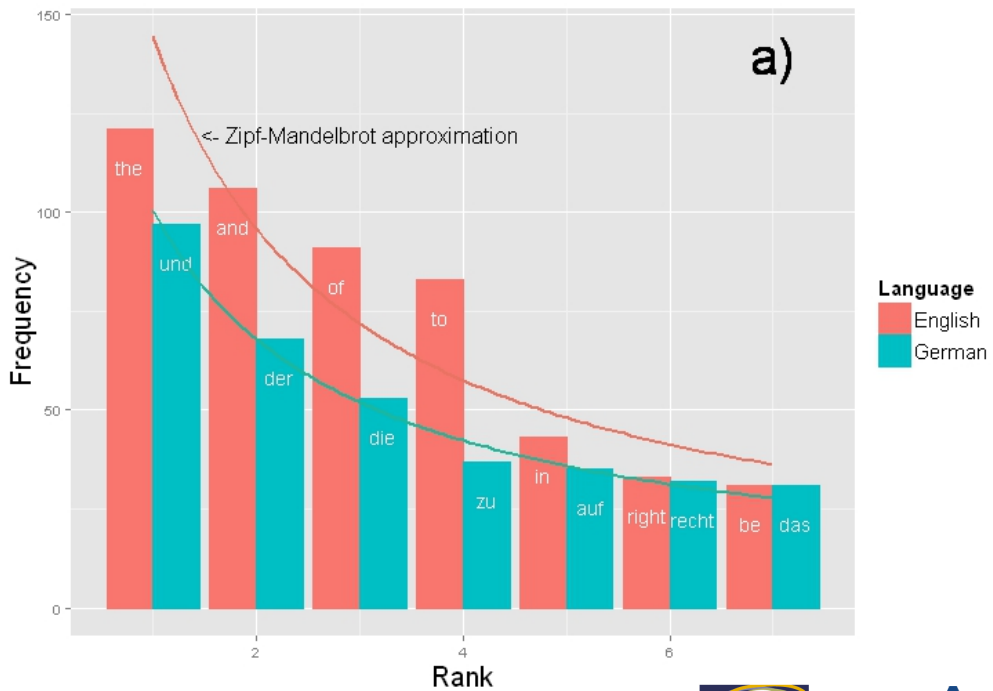
Quantitative measure

Zipf-Mandelbrots law: Order types (word forms delimited by white spaces) according to their token frequencies (Zipf, 1949; Mandelbrot, 1953)

Lexical diversity

Quantitative measure

Zipf-Mandelbrots law: Order types (word forms delimited by white spaces) according to their token frequencies (Zipf, 1949; Mandelbrot, 1953)



Lexical diversity

Zipf-Mandelbrot's law

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

$C > 0,$
$\alpha > 0,$
$\beta > -1,$
$i = 1, 2, \ldots, n$

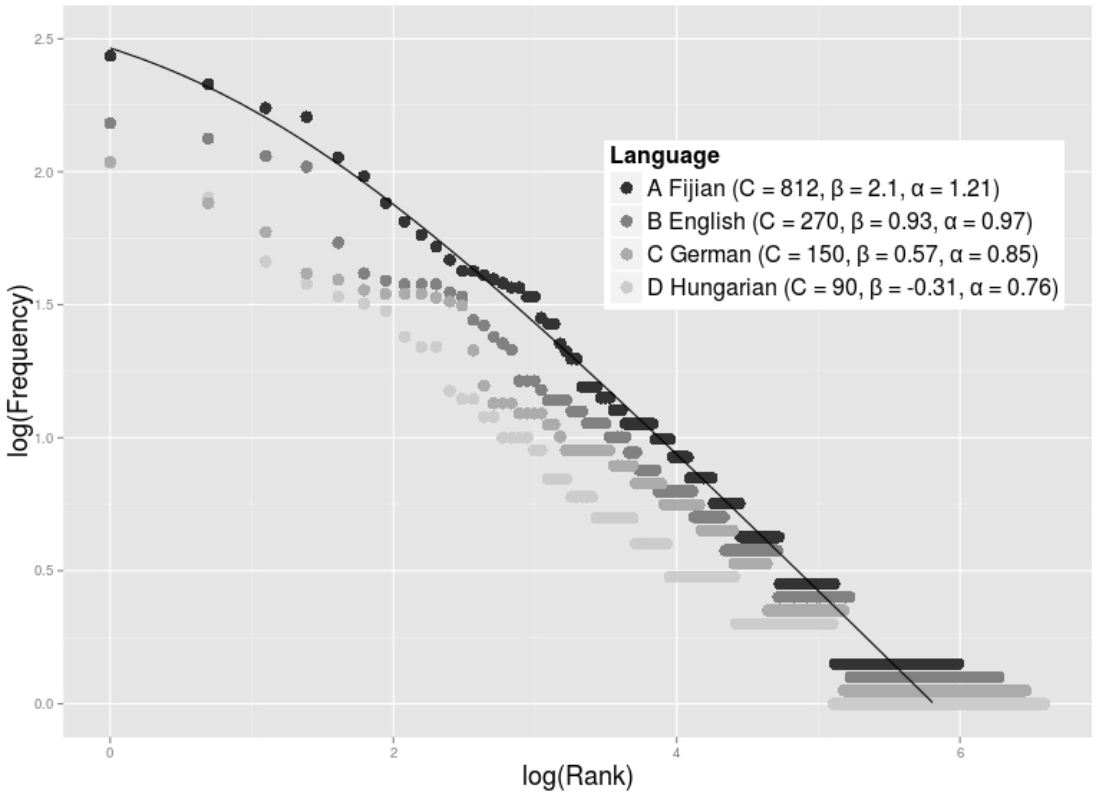
Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Lexical diversity

Zipf-Mandelbrot's law

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

$C > 0,$
$\alpha > 0,$
$\beta > -1,$
$i = 1, 2, \ldots, n$



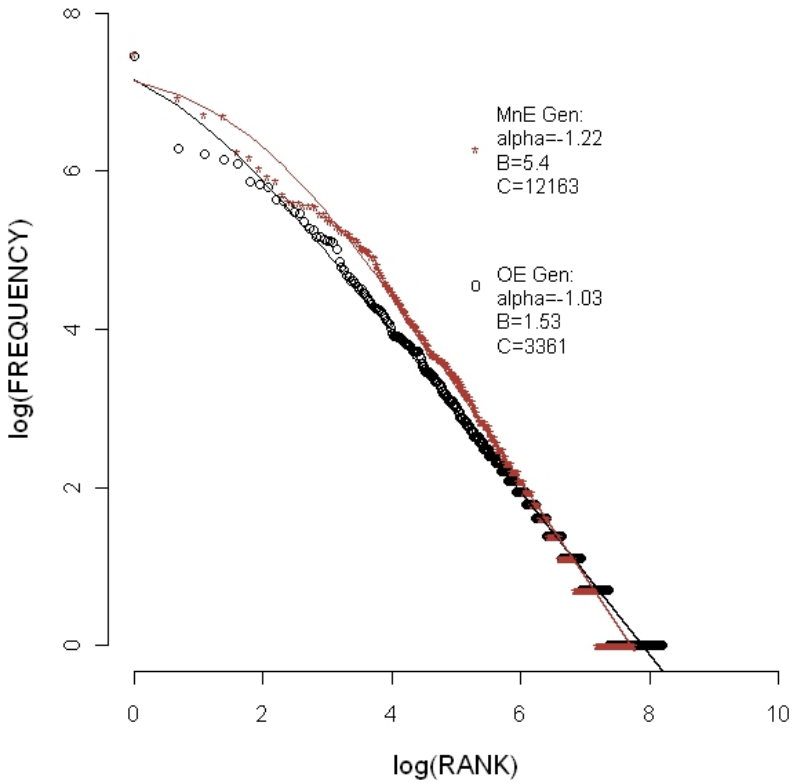
Diachrony

Bentz, Kiela, Hill & Buttery (2014) Zipf's law and the grammar of languages. *Corpus Linguistics and Linguistic Theory.*

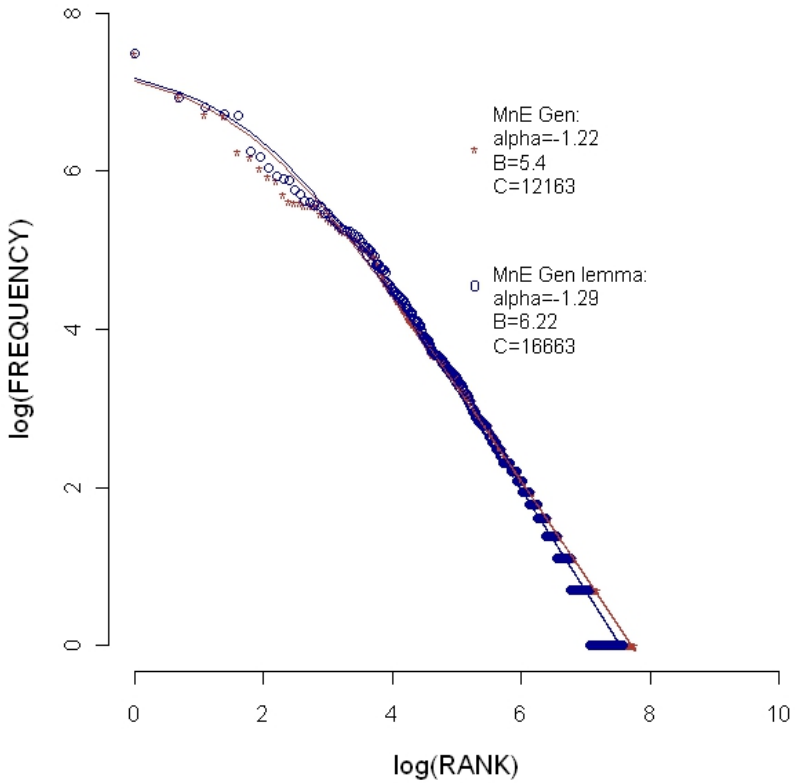
Diachrony

Bentz, Kiela, Hill & Buttery (2014) Zipf's law and the grammar of languages. *Corpus Linguistics and Linguistic Theory.*

10a) MnE and OE Genesis



10b) MnE Genesis + lemmatized version

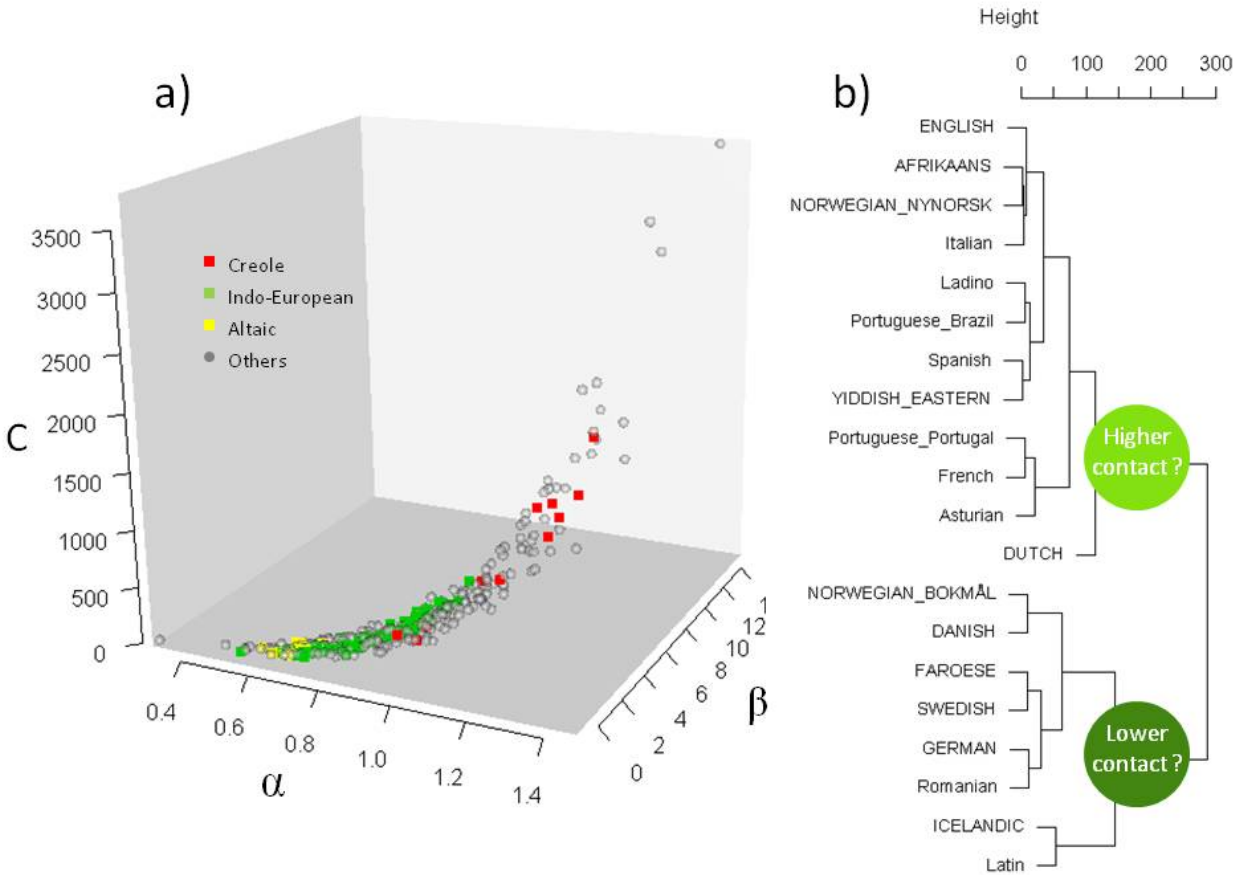


Synchrony

Bentz, Kiela, Hill & Buttery (in preparation) Adaptive languages: Modeling lexical diversity cross-linguistically.

Synchrony

Bentz, Kiela, Hill & Buttery (in preparation) Adaptive languages: Modeling lexical diversity cross-linguistically.



Lexical diversity

- Seems to be reduced in contact scenarios (non-native language learning, see Trudgill, 2011; Lupyan & Dale 2011; McWhorter, 2007)

Lexical diversity

- Seems to be reduced in contact scenarios (non-native language learning, see Trudgill, 2011; Lupyan & Dale 2011; McWhorter, 2007)

Question

WHY DO LANGUAGES GET HIGH LEXICAL DIVERSITIES IN THE FIRST PLACE?

Grammaticalization

Definition

In the final stage of grammaticalization **frequently co-occurring** words **merge** by means of phonological fusion (Bybee, 2003: 617) and hence 'morphologize' to built **inflection**s and **derivations**

Grammaticalization

Definition

In the final stage of grammaticalization **frequently co-occurring** words **merge** by means of phonological fusion (Bybee, 2003: 617) and hence 'morphologize' to built **inflection**s and **derivations**

Cline

content item >grammatical word >clitic >inflectional affix
(Hopper and Traugott, 2003: 7)

Grammaticalization

Definition

In the final stage of grammaticalization **frequently co-occurring** words **merge** by means of phonological fusion (Bybee, 2003: 617) and hence 'morphologize' to built **inflection**s and **derivations**

Cline

content item > *grammatical word* > *clitic* > *inflectional affix*
(Hopper and Traugott, 2003: 7)

Example

Old English *līc* 'body' → *-ly*
Latin *cantare habeo* 'I have to sing' → Italian *canterò*

Grammaticalization

Hypothesis

Grammaticalization → **increasing** lexical diversity

Deflexion → **decreasing** lexical diversity

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Grammaticalization

Hypothesis

Grammaticalization → **increasing** lexical diversity
Deflexion → **decreasing** lexical diversity

Question

Can we computationally model the impact of grammaticalization on lexical diversity?

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Computational Model

Starting point: Fijian UDHR

- parallel text, control for constant information content
- analytic language with low lexical diversity

Computational Model

Starting point: Fijian UDHR

- parallel text, control for constant information content
- analytic language with low lexical diversity

Process

- **merge** a given percentage (p_m) of **frequently co-occurring words** over several generations (n_G)

Computational Model

Starting point: Fijian UDHR

- parallel text, control for constant information content
- analytic language with low lexical diversity

Process

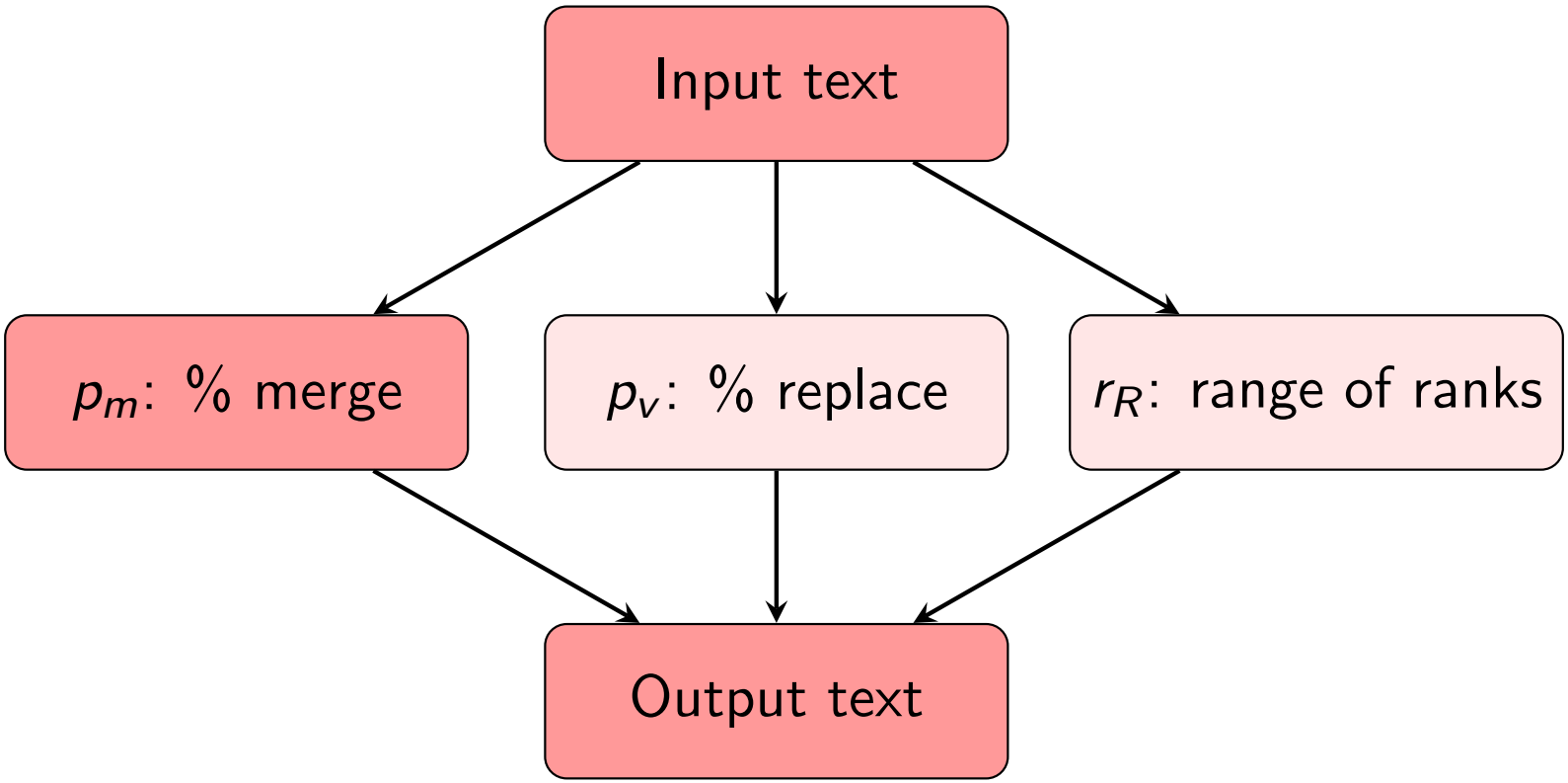
- **merge** a given percentage (p_m) of **frequently co-occurring words** over several generations (n_G)

Endpoint

- Do we arrive at lexical diversities similar to the ones for German or Hungarian?

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Architecture

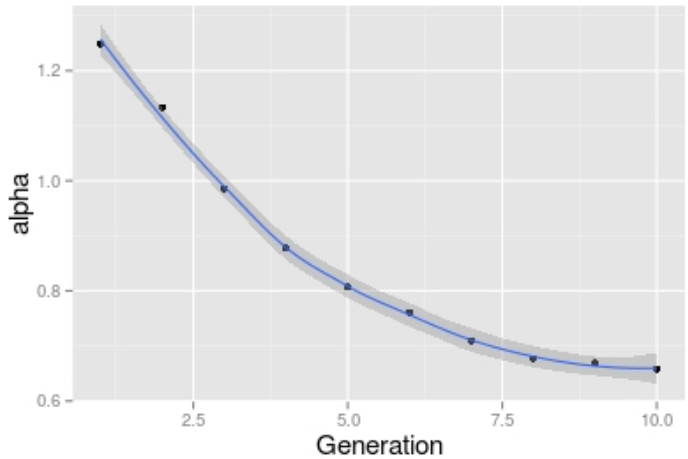


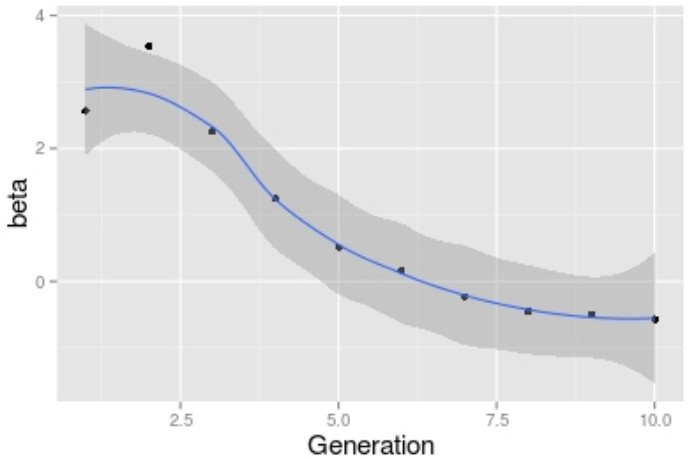
Output

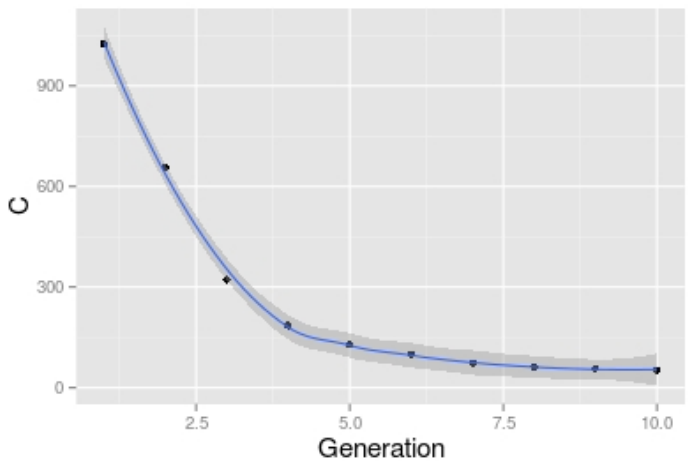
$p_m = 2.5, p_v = 0; r_R = 0; n_G = 10$

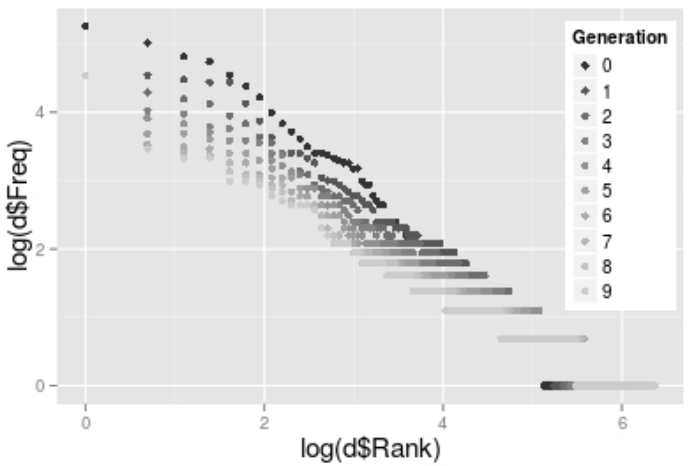
Output

$p_m = 2.5, p_v = 0; r_R = 0; n_G = 10$









Output

Words created in English

- *ofthe* $\rightarrow$ genitive marked article, German: *des*

Output

Words created in English

- *ofthe* $\rightarrow$ genitive marked article, German: *des*
- *inthe* $\rightarrow$ preposition merged with article, Italian: *in* + *il* rendering *nel*

Output

Words created in English

- *ofthe* → genitive marked article, German: *des*
- *inthe* → preposition merged with article, Italian: *in* + *il* rendering *nel*
- *topromote* → preposition + verb, German: *zusehen, zuschliessen*

Output

Words created in English

- *ofthe* $\rightarrow$ genitive marked article, German: *des*
- *inthe* $\rightarrow$ preposition merged with article, Italian: *in* + *il* rendering *nel*
- *topromote* $\rightarrow$ preposition + verb, German: *zusehen, zuschliessen*
- *ofsociety* $\rightarrow$ preposition + noun (case prefix?)

Output

Words created in English

- *ofthe* → genitive marked article, German: *des*
- *inthe* → preposition merged with article, Italian: *in* + *il* rendering *nel*
- *topromote* → preposition + verb, German: *zusehen, zuschliessen*
- *ofsociety* → preposition + noun (case prefix?)
- *humanrights, humanbeing* → compounding, German: *Menschenrechte*

Output

Words created in English

- *ofthe* → genitive marked article, German: *des*
- *inthe* → preposition merged with article, Italian: *in* + *il* rendering *nel*
- *topromote* → preposition + verb, German: *zusehen, zuschliessen*
- *ofsociety* → preposition + noun (case prefix?)
- *humanrights, humanbeing* → compounding, German: *Menschenrechte*
- *everyonehastherighttofreedomof, withoutanydiscrimination*

Future Directions

Model improvement

- Exploring models with varying parameters for vocabulary replacement and merging of bigrams (comparison to actual language change data)

Future Directions

Model improvement

- Exploring models with varying parameters for vocabulary replacement and merging of bigrams (comparison to actual language change data)
- More realistic model by parsing and POS tagging

Future Directions

Model improvement

- Exploring models with varying parameters for vocabulary replacement and merging of bigrams (comparison to actual language change data)
- More realistic model by parsing and POS tagging
- Considering frequency measures beyond bigram frequencies

Collaborators

Douwe Kiela

Felix Hill

Andrew Caines

Arts & Humanities Research Council
CAMBRIDGE ASSESSMENT
UNIVERSITY OF CAMBRIDGE

Thank You!

¡chris@christianbentz.de¿