Is loanword phonology simpler?

A statistical investigation

Neige Rochant[1] & Marwan Kilani[2]

[1]Sorbonne Nouvelle University / CNRS: Lacito – LLACAN
[2]University of Basel

First Swiss Workshop on Sociolinguistics, Language Contacts and Historical Linguistics in the Ancient World
9–10 Feb. 2023

Outline for section 1

1. Introduction
2. Hypothesis
3. An explicative model
4. Discussion

- Loanwords = source of information for study of interactions in ancient contexts

- Loanwords = source of information for study of interactions in ancient contexts
- Distinguishing potential loanwords from inherited words = crucial & controversial

- Loanwords = source of information for study of interactions in ancient contexts
- Distinguishing potential loanwords from inherited words = crucial & controversial
  - ⋆ Language-specific approaches (Kang 2011)

- Loanwords = source of information for study of interactions in ancient contexts
- Distinguishing potential loanwords from inherited words = crucial & controversial
  - ⋆ Language-specific approaches (Kang 2011)
  - ⋆ Quantitative methods to automatically identify loanwords (Haspelmath and Tadmor 2009a; Zhang, Fabri, and Nerbonne 2021; Miller et al. 2020; Nath et al. 2022)

- Loanwords = source of information for study of interactions in ancient contexts
- Distinguishing potential loanwords from inherited words = crucial & controversial
  - ⋆ Language-specific approaches (Kang 2011)
  - ⋆ Quantitative methods to automatically identify loanwords (Haspelmath and Tadmor 2009a; Zhang, Fabri, and Nerbonne 2021; Miller et al. 2020; Nath et al. 2022)
- But little attention has been paid to potentially statistically relevant patterns/features distinguishing loanwords from inherited words

Initial hunch

- Distribution of phonemes in loanwords ≠ distribution of phonemes in inherited words.

Initial hunch

- Distribution of phonemes in loanwords ≠ distribution of phonemes in inherited words.

- Loanwords seem to pick from poorer phonemic inventories than inherited words

Data: the WOLD database (Haspelmath and Tadmor 2009b; 2021)

Data: the WOLD database (Haspelmath and Tadmor 2009b; 2021)

What is it?

- 64,289 lexical entries from 41 languages

Data: the WOLD database (Haspelmath and Tadmor 2009b; 2021)

What is it?

- 64,289 lexical entries from 41 languages
- coded for source (loanword vs. inherited) and, if relevant, source language(s)

Data: the WOLD database (Haspelmath and Tadmor 2009b; 2021)

What is it?

- 64,289 lexical entries from 41 languages
- coded for source (loanword vs. inherited) and, if relevant, source language(s)
- Total of 15,213 likely loanwords from about 300 different languages

Data: the WOLD database (Haspelmath and Tadmor 2009b; 2021)

What is it?

- 64,289 lexical entries from 41 languages
- coded for source (loanword vs. inherited) and, if relevant, source language(s)
- Total of 15,213 likely loanwords from about 300 different languages
- Miller et al. 2020 added IPA to the target languages ← we used their dataset

Data: the WOLD database (Haspelmath and Tadmor 2009b; 2021)

What is it?

- 64,289 lexical entries from 41 languages
- coded for source (loanword vs. inherited) and, if relevant, source language(s)
- Total of 15,213 likely loanwords from about 300 different languages
- Miller et al. 2020 added IPA to the target languages ← we used their dataset
- Issues:
  - ⋆ Source words in orthography / various transcriptions (not IPA)

Data: the WOLD database (Haspelmath and Tadmor 2009b; 2021)

What is it?

- 64,289 lexical entries from 41 languages
- coded for source (loanword vs. inherited) and, if relevant, source language(s)
- Total of 15,213 likely loanwords from about 300 different languages
- Miller et al. 2020 added IPA to the target languages ← we used their dataset
- Issues:
  - ⋆ Source words in orthography / various transcriptions (not IPA)
  - ⋆ Source language(s) ≠ reliable (e.g., unsystematic choice between closest vs. ultimate source)

Data: the WOLD database

What can be done with it?

- ✔ Compute statistics of phonemes in loanwords and inherited words

Outline for section 2

1. Introduction

2. Hypothesis

3. An explicative model

4. Discussion

Hypothesis:

Loanwords feature a specific phonemic distribution:

→ they pick from poorer phonemic inventories

viz.:

Loanwords tend to contain **fewer**:

- rare phonemes
- co-articulated phonemes

Hypothesis:

Loanwords feature a specific phonemic distribution:

→ they pick from poorer phonemic inventories

viz.:

Loanwords tend to contain **fewer**:

- rare phonemes = attested in <= 6 languages
- co-articulated phonemes, e.g., labial-velars, ejectives, etc.

Hyp. verified by a quick calculation

Hyp. verified by a quick calculation

Freq. of co-articulated phonemes in total: 0.092
Freq. of co-articulated phonemes in loanwords: 0.055

Ratio of freq. of co-articulated phonemes in loanwords / all words:

→ 0.593

Hyp. verified by a quick calculation

Freq. of co-articulated phonemes in total: 0.092
Freq. of co-articulated phonemes in loanwords: 0.055

Ratio of freq. of co-articulated phonemes in loanwords / all words:

→ 0.593

Freq. of rare phonemes in total: 0.110
Freq. of rare phonemes in loanwords: 0.075

Ratio of freq. of rare phonemes in loanwords / all words:

→ 0.683

Outline for section 3

1. Introduction
2. Hypothesis
3. An explicative model
4. Discussion

2 hypotheses:

- Loanwords tend to contain fewer rare phonemes *simply because they are rare*
- Loanwords tend to contain fewer co-articulated phonemes because *phonemes of source words tend to be replaced by mono-articulated phonemes*

A model to verify these hypotheses

Hyp 1: Loanwords contain fewer rare phonemes simply because they are rare

A model to verify these hypotheses

Hyp 1: Loanwords contain fewer rare phonemes simply because they are rare

- When a target language TL borrows a word with a phoneme ϕ:
  - ⋆ if it has ϕ, it preserves it
  - ⋆ if it does not have ϕ, it replaces it.

A model to verify these hypotheses

Hyp 1: Loanwords contain fewer rare phonemes simply because they are rare

- When a target language TL borrows a word with a phoneme ϕ:
  - ⋆ if it has ϕ, it preserves it
  - ⋆ if it does not have ϕ, it replaces it.
- A phoneme in a source word has a probability p_1 of being ϕ

A model to verify these hypotheses

Hyp 1: Loanwords contain fewer rare phonemes simply because they are rare

- When a target language TL borrows a word with a phoneme ϕ:
  - ⋆ if it has ϕ, it preserves it
  - ⋆ if it does not have ϕ, it replaces it.
- A phoneme in a source word has a probability p_1 of being ϕ
- The target language TL has a probability p_2 of having ϕ in its inventory

A model to verify these hypotheses

Hyp 1: Loanwords contain fewer rare phonemes simply because they are rare

- When a target language TL borrows a word with a phoneme ϕ:
  - ⋆ if it has ϕ, it preserves it
  - ⋆ if it does not have ϕ, it replaces it.
- A phoneme in a source word has a probability $p1$ of being ϕ
- The target language TL has a probability $p2$ of having ϕ in its inventory
- ⟶ The probability for ϕ to be in the output of the borrowing process in the target word = $p1$x$p2$.

A model to verify these hypotheses

Hyp 1: Loanwords contain fewer rare phonemes simply because they are rare

- When a target language TL borrows a word with a phoneme ϕ:
  - ⋆ if it has ϕ, it preserves it
  - ⋆ if it does not have ϕ, it replaces it.
- A phoneme in a source word has a probability p_1 of being ϕ
- The target language TL has a probability p_2 of having ϕ in its inventory
- ⟶ The probability for ϕ to be in the output of the borrowing process in the target word = $p_1 \times p_2$.
- ⟶ We expect ϕ to have a frequency $p_1 \times p_2$ in loanwords

A model to verify these hypotheses

Hyp 2: Loanwords tend to contain fewer co-articulated phonemes because phonemes of the source words tend to be replaced by mono-articulated phonemes

A model to verify these hypotheses

Hyp 2: Loanwords tend to contain fewer co-articulated phonemes because phonemes of the source words tend to be replaced by mono-articulated phonemes

- When a phoneme ϕ is co-articulated (> 1 grapheme),
  - ⋆ if it is not preserved during borrowing
  - ⋆ and if there exists a corresponding **mono-articulated phoneme** (=1 grapheme)

  then it is replaced by the latter.

Model performance

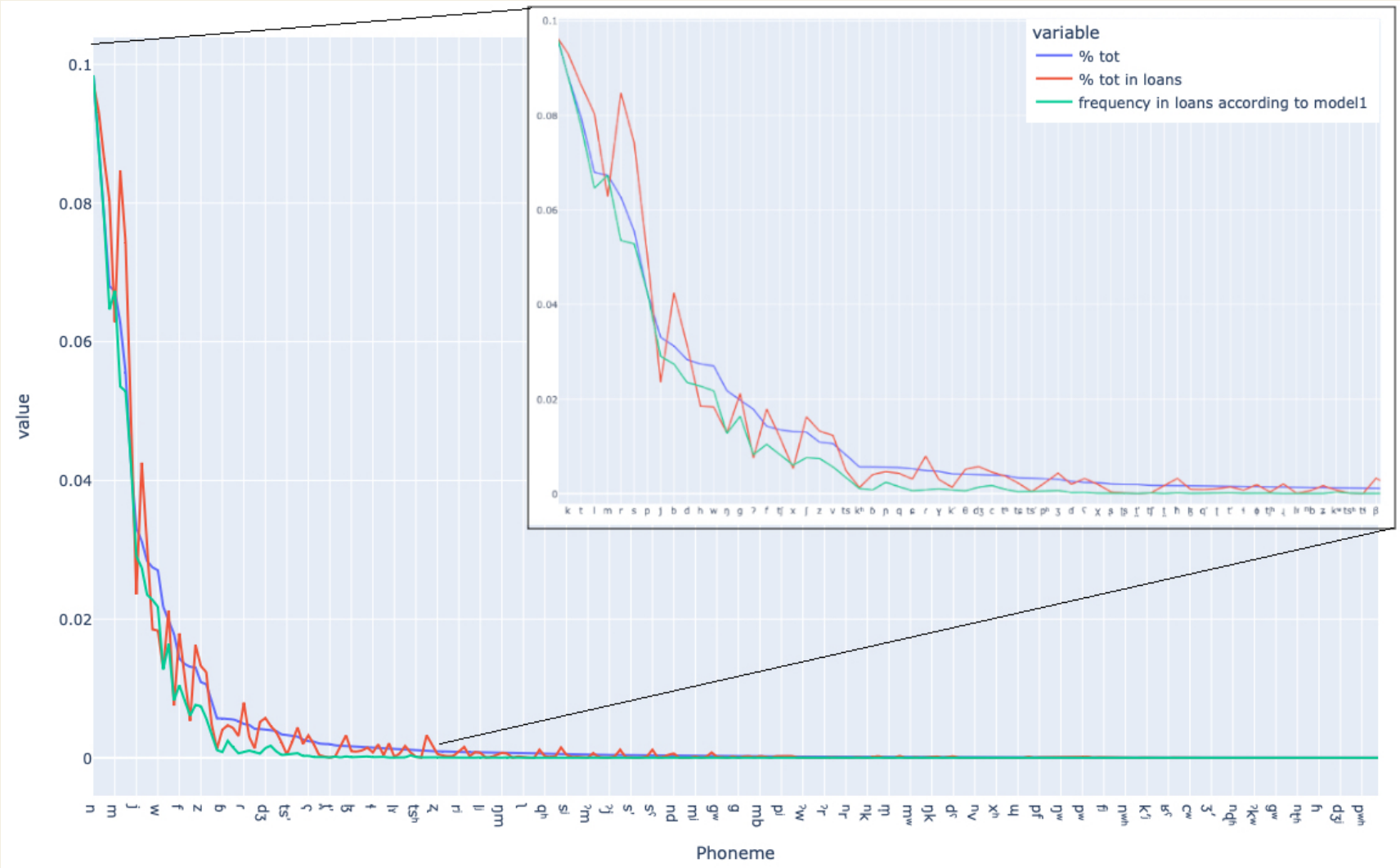

Figure 1: Real vs. expected frequency of phonemes in loanwords according to model 1 (= just hypothesis 1)

Model performance

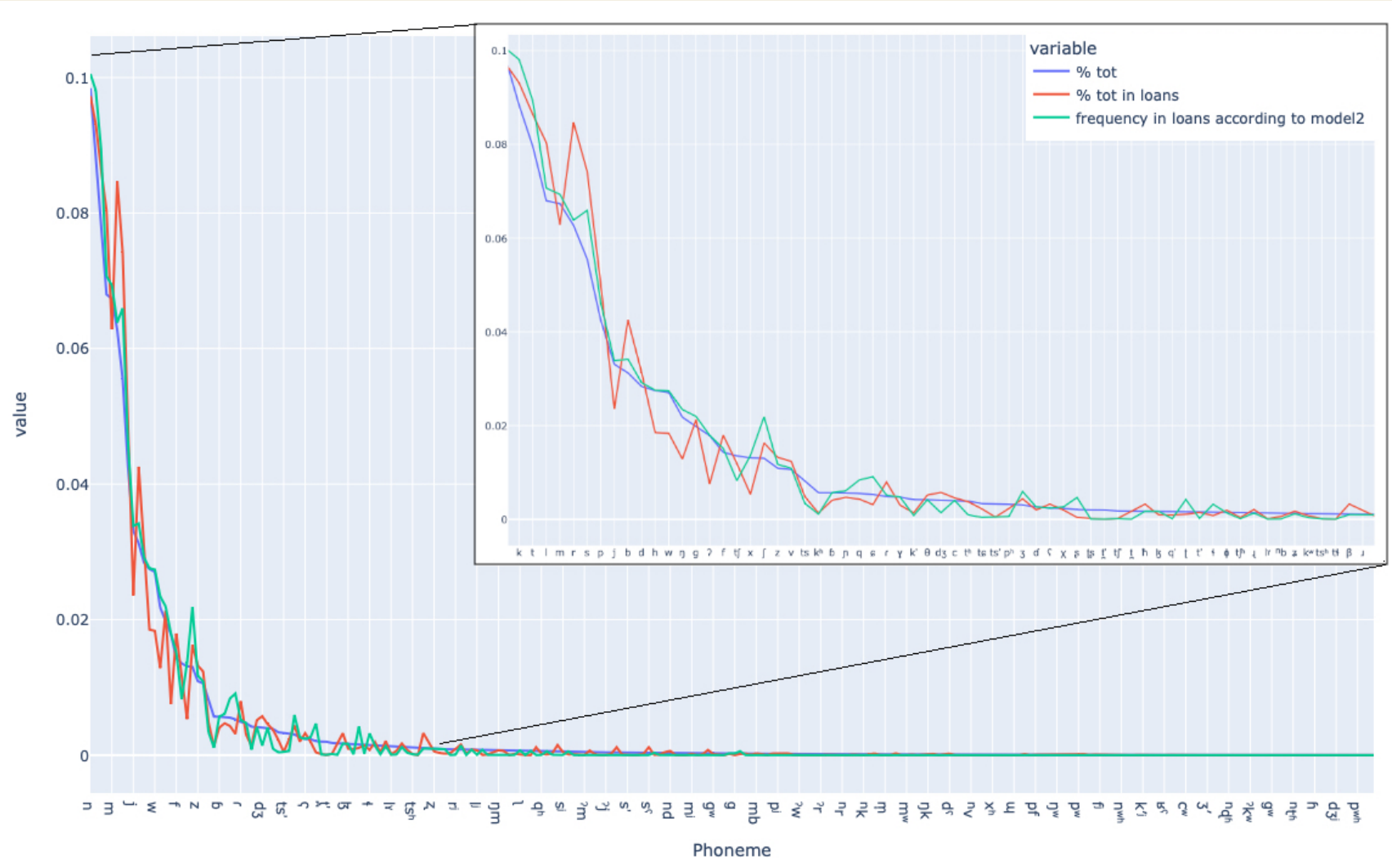

Figure 2: Real vs. expected frequency of phonemes in loanwords according to model 2 (= hypothesis 1 + hypothesis 2)

Test of model accuracy

Which of these models yields the closest output:

- m_1 (just Hyp1)
- m_2 (Hyp1 + hyp2)
- 'Equal to All' model: a baseline assuming that the phonemic distribution in loanwords is the same as in inherited words

Test of model accuracy

Which of these models yields the closest output:

- m_1 (just Hyp1)
- m_2 (Hyp1 + hyp2)
- 'Equal to All' model: a baseline assuming that the phonemic distribution in loanwords is the same as in inherited words

l2-based error test: the lower the error, the more accurate the model.

Test of model accuracy - Results

Error of model 'Equal to all words':	0.0300
Error of model 1:	0.0384
Error of model 2:	0.0295

Model 1 is worse than model "Equal to All" at predicting the frequency of phonemes in loans, which is understandable since it does not grasp phoneme replacements.

Model 2, which includes a prediction of replacement phonemes, is more accurate than model 1, and more accurate than "Equal to All".

Outline for section 4

1. Introduction
2. Hypothesis
3. An explicative model
4. Discussion

Conclusions

Model 2 succeeds at predicting the frequency of phonemes in loanwords with a lower error (0.0295) and hence a higher accuracy than the other models.

Conclusions

Model 2 succeeds at predicting the frequency of phonemes in loanwords with a lower error (0.0295) and hence a higher accuracy than the other models.

+ It is much better than Model “Equal to All” at predicting the relative frequency scores of phonemes in loanwords (cf. Fig. 2).

Conclusions

Model 2 succeeds at predicting the frequency of phonemes in loanwords with a lower error (0.0295) and hence a higher accuracy than the other models.

+ It is much better than Model "Equal to All" at predicting the relative frequency scores of phonemes in loanwords (cf. Fig. 2).

⟶ This supports our 2 hypotheses:

- ⋆ rare phonemes are rarer in loanwords *just because they are rare*
- ⋆ co-articulated phonemes are frequently replaced by mono-articulated equivalents upon borrowing

Conclusions

It would be interesting to check which specific phonemes are affected, because the data present two additional points of interest, viz.:

Conclusions

1. While overall, the most frequent mono-articulated phonemes are more frequent in loanwords, there are a few exceptions:

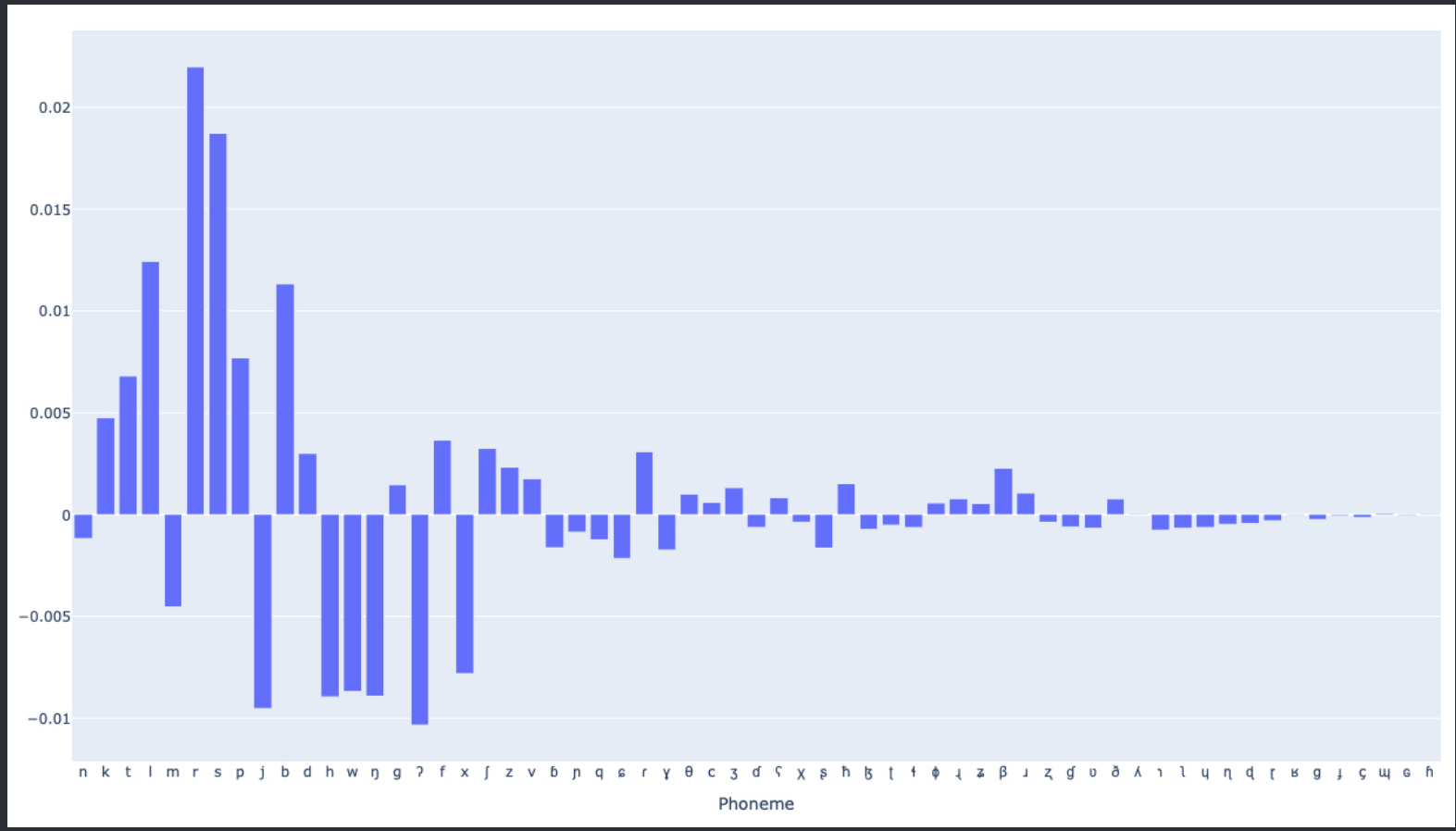

Figure 3: Difference of frequency of mono-articulated phonemes in all words combined vs. loanwords only

Conclusions

2. The more frequent a co-articulated phoneme is in the corpus, the less frequent it is in loanwords

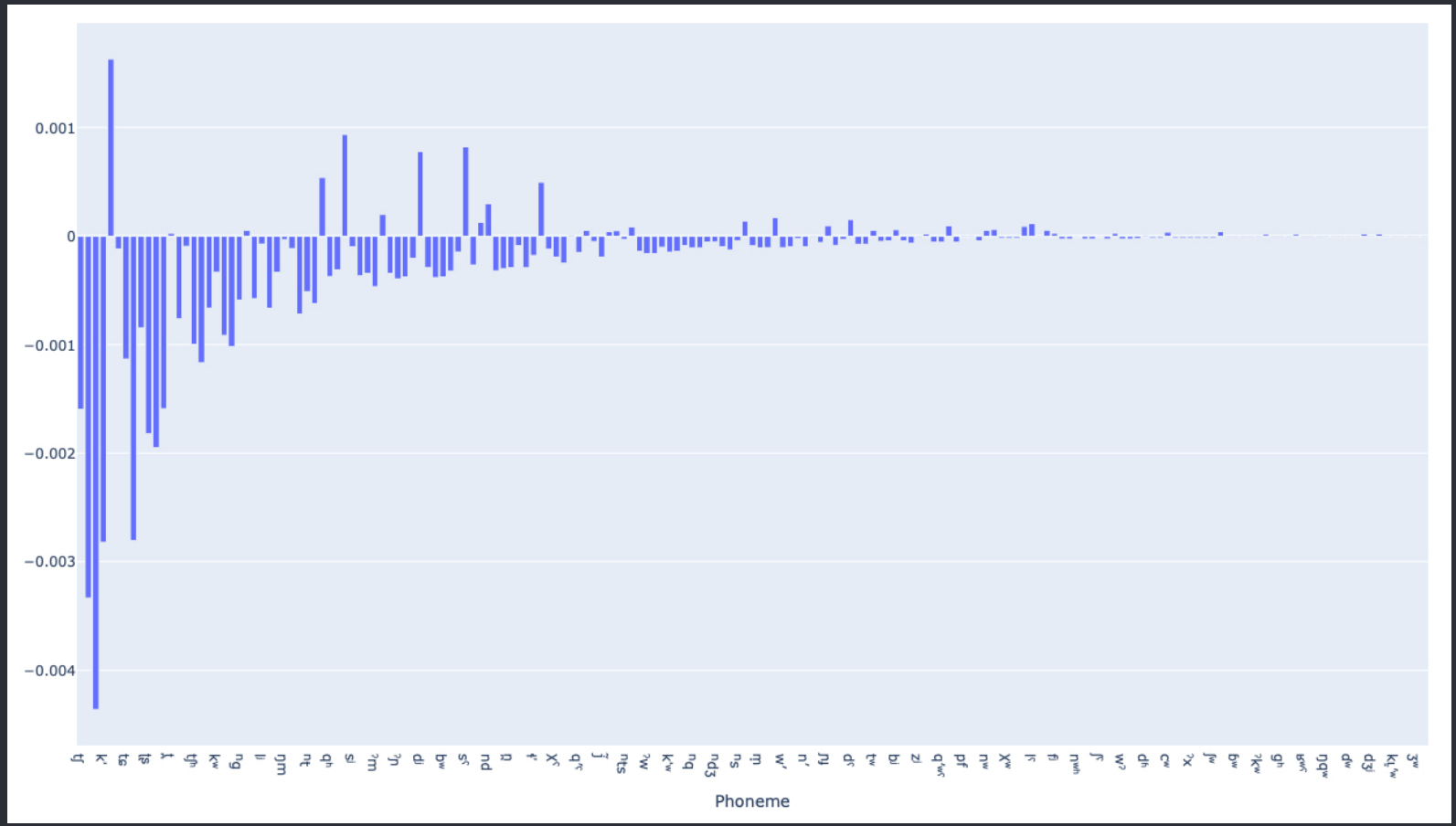

Figure 4: Difference of frequency of coarticulated phonemes in all words vs. loanwords only

Significance

Significance

Our results suggest that: when borrowed, words are statistically expected to lose phonological rareness/complexity and hence “phonological distinctiveness”

Significance

Our results suggest that: when borrowed, words are statistically expected to lose phonological rareness/complexity and hence "phonological distinctiveness"

→ Borrowed words are more likely to look alike by chance, esp. if borrowed several times

Significance

Our results suggest that: when borrowed, words are statistically expected to lose phonological rareness/complexity and hence "phonological distinctiveness"

→ Borrowed words are more likely to look alike by chance, esp. if borrowed several times

→ Potential implications for the study of loanwords and Wanderwörter esp. in prehistoric contexts.

Significance

Our results suggest that: when borrowed, words are statistically expected to lose phonological rareness/complexity and hence "phonological distinctiveness"

→ Borrowed words are more likely to look alike by chance, esp. if borrowed several times

→ Potential implications for the study of loanwords and Wanderwörter esp. in prehistoric contexts.

E.g. *How can we set apart loanwords which look alike because they stem from the same source word and loanwords from different source words that happen to look alike because of the phenomenon described in this paper?*

➤ First requirement to investigate these questions: better datasets

Gramarzé

Dangge

Grazia fitg

Grant marci

Bibliography

Haspelmath, M. and U. Tadmor (2009a). *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.

— eds. (2009b). *WOLD*. Leipzig: Max Planck Institute for Evolutionary Anthropology. url: https://wold.clld.org/.

— (2021). *CLDF dataset derived from Haspelmath and Tadmor's "World Loanword Database" from 2009*. type: dataset. doi: 10.5281/ZENODO.5139859. url: https://zenodo.org/record/5139859 (visited on 12/13/2022).

Kang, Y. (2011). "Loanword Phonology: Loanword Phonology". en. In: *The Blackwell Companion to Phonology*. Ed. by M. van Oostendorp et al. Oxford: John Wiley & Sons, Ltd, pp. 1–25. doi: 10.1002/9781444335262.wbctp0095. url: https://onlinelibrary.wiley.com/doi/10.1002/9781444335262.wbctp0095 (visited on 12/13/2022).

Miller, J. E. et al. (Dec. 2020). "Using lexical language models to detect borrowings in monolingual wordlists". en. In: *PLOS ONE* 15.12. Ed. by S. Wichmann, e0242709. doi: 10.1371/journal.pone.0242709. url: https://dx.plos.org/10.1371/journal.pone.0242709 (visited on 12/13/2022).

Nath, A. et al. (Oct. 2022). "A Generalized Method for Automated Multilingual Loanword Detection". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 4996–5013. url: https://aclanthology.org/2022.coling-1.442.

Zhang, L., R. Fabri, and J. Nerbonne (2021). "Detecting loan words computationally". en. In: *Contact Language Library*. Ed. by E. O. Aboh and C. B. Vigouroux. Vol. 59. Amsterdam: John Benjamins, pp. 269–288. doi: 10.1075/coll.59.11zha. url: https://benjamins.com/catalog/coll.59.11zha (visited on 12/13/2022).

TEX, LATEX, and Beamer