Conflicting trigger effects in Uyghur backness harmony

Connor Mayer[1], Travis Major[1], & Mahire Yakup[2]
[1] University of California, Los Angeles
[2] Nazarbayev University

Tu+5
University of Delaware
February 9th, 2020

Overview

In this presentation we will:

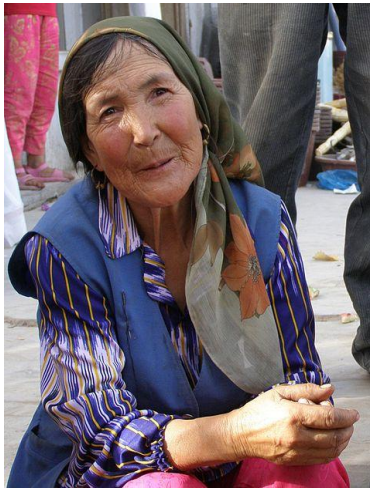
- Present new empirical data from a corpus study and wug tests of Uyghur backness harmony
- Examine differences between the corpus pattern and wug pattern
- Speculate on reasons for these discrepancies

The Uyghur Language

Uyghur is a Southeastern Turkic language

- Spoken by roughly 10 million people, primarily in China, Uzbekistan, Kazakhstan, Kyrgyzstan

- Exhibits backness and rounding harmony



https://commons.wikimedia.org/wiki/File:Uyghur-Dopa-Maker.jpg
https://commons.wikimedia.org/wiki/File:Uyghur_woman_Xinjiang.jpg

Uyghur backness harmony

Standard descriptions of Uyghur backness harmony requires suffix forms to agree in backness with vowels and certain consonants in the stem (e.g., Lindblad 1990; Hahn 1991; Engsaeth et al. 2010).

- We use the locative suffix /-DA/ as a prototypical example

- Backness agreement is reflected in the vowel: /**a**/ or /**æ**/

- Voicing changes in the initial segment are not relevant: /t/ or /d/

Relevant segments

	Front		Back	
	Unrounded	Round	Unrounded	Round
High	i	y		u
Mid	e	ø		o
Low	æ		a	

	Front	Back
Voiceless	k	q
Voiced	g	ʁ

- The front vowels /i/ and /e/ are reported to be **transparent** to harmony
- **Velars** pattern as **front**
- **Uvulars** pattern as **back**

Relevant segments

	Front		Back	
	Unrounded	Round	Unrounded	Round
High	i	y	ɯ?	u
Mid	e	ø	ɤ?	o
Low	æ		ɑ	

	Front	Back
Voiceless	k	q
Voiced	g	ʁ

Some researchers (e.g., McCollum 2019) claim Uyghur has phonemic back counterparts to /i/ and /e/.

- Phonetic evidence for phonemic status is murky
- Better evidence for (non-categorical) allophonic variation
- I assume /i/ and /e/ are **phonologically transparent**, not phonetically

Vowel harmony

Most suffixes must match the backness of the final harmonizing vowel in the stem

	Front	Back
1	køz-dæ “on/in an eye”	at-ta “on/in a horse”
2	syt-tæ “on/in milk”	orun-da “on/in a place”
3	xæmit-tæ “on/in Xemit”	tarix-ta “on/in history”
4	halæt-tæ “on/in a situation”	ævlad-da “on/in a generation”

Consonant harmony

In the absence of a harmonizing vowel, dorsal consonants appear to serve as triggers for harmony.

Front	**k**iʃi-d**æ** “on/in a person”
	gezit-t**æ** “on/in a newspaper”
Back	**q**iz-d**a** “on/in a girl”
	qir**ʁ**iz-d**a** “on/in the Kyrgyz”

Conflicting triggers

In cases of backness conflict between vowels and consonants in the same word, a more distant vowel overrides a less distal consonant

Front	mæʃq-tæ “on/in an exercise”
	tæqdir-dæ “on/in fate”
Back	rak-ta “on/in a shrimp”
	taksi-da “on/in a taxi”

Apparent exceptions

Lindblad (1990) reports several exceptions to the general pattern of conflict resolution.

Front	mumkin-lik “possibility”
Back	mæntiq-liq “logical”

Note that all of these exceptions involve derivational suffixes.

Transparent Vowels

Stems with no harmonizers are arbitrarily specified for backness

- Statistical preference for back suffixes

Front	biz-dæ “on/in us”
	siz-dæ “on/in you”
Back	it-ta “on/in a dog”
	pil-da “on/in an elephant”

Prior Phonological Analyses (non-exhaustive)

Lindblad (1990) and Hahn (1991) propose underlying contrasts between /**i**/ ~ /**ɯ**/ and /**e**/ ~ /**ɤ**/ to account for spreading and transparent words.

- /**ɯ**/ and /**ɤ**/ **neutralize** to /i/ and /e/ post-lexically
- Harmony is only triggered by vowels
- Dorsals **undergo** harmony (e.g., [-li**q**]~[-li**k**]) but do not **trigger** it
- The harmony value of stems with apparent conflicts or no harmonizers is determined by the backness of their underlying vowels, e.g.:
  - /q**ɯ**z-DA/ → [qizd**a**] /k**i**sh**i**-DA/ → [kishid**æ**]
  - /m**æ**nt**ɯ**q-lIK/ → [m**æ**ntiqli**q**] /m**u**mk**i**n-lIK/ → [m**u**mkinli**k**]

Derivation of *[mæntiqliq]*

UR	/mæntɯq-lIK/
1. Spreading	mæntɯq-lɯK
2. Spreading	mæntɯq-lɯq
3. Fronting	mæntiqliq
SR	[mæntiqliq]

Why revisit Uyghur backness harmony?

- Covert contrast analysis requires a large amount of hidden structure
  - Backness of dorsals serves as clear cue to backness of stem
  - Might learners converge on a more surface-true grammar?

- The ‘fallback’ pattern in Uyghur backness harmony is typologically unusual

- The pattern is computationally more complex than most segmental phonological patterns (Mayer and Major 2018)

The present study

We use copus data, and real and wug elicitation to provide an empirical phonological description of the basic pattern of Uyghur backness harmony.

We focus on:

- Differences in the strength of front/back vowels and front/back dorsals
- Resolution of conflicts between vowels and dorsals
- Distance-based decay

Corpus study

We created a corpus of about 24,000 articles from the Kazakh Uyghur newspaper *Uyghur Awazi* using a custom webscraper.

- ~6.2 million words

Searching the corpus

We searched for inflected forms of word stems taken from two sources:

- Vocabulary lists from Uyghur textbooks (Nazarova and Niyaz 2013, 2016)

- Dictionary entries from an online Uyghur dictionary (http://www.uighurdictionary.com/)
  - Retrieved using a webscraper
  - Multi-word entries were omitted

Total of **15,632 stems**

Searching the corpus

We restricted our stems to nouns, verbs, and adjectives.

For each stem, we chose a set of inflectional suffixes that harmonize and searched the corpus for occurrences of front and back variants of each.

E.g., for nouns:

- **Dative**: [-**k**e], [-**g**e], [-**q**e], [-**q**a], [-**gh**a]
- **Locative:** [-t**e**], [-d**e**], [-t**a**], [-d**a**]
- **Plural:** [-l**e**r], [-l**a**r]
- **Delimiting:** [-**k**iche], [-**g**iche], [-**q**iche], [-**gh**iche]
- **Comparative**: [-chili**k**], [-chili**q**]

Limitations

We do not count forms with multiple inflectional affixes:

- E.g., [qiz-i-**gha**] ‘to her daughter’

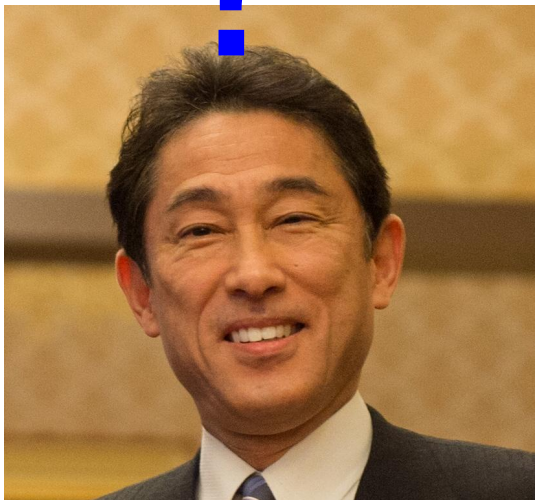
We do not count derivational morphology:

- E.g., [bext] ‘happiness’ ~ [bext-li**k**] ‘happy’
- These derived forms are often listed in the dictionary

Some false positives may occur:

- [kishida]: “on/in a person” or the former Japanese Minister of Foreign Affairs?

?



Corpus results

We retrieved a total of **~475k tokens** from **11,460** unique inflected types with **1,066** unique stem forms.

We consider a subset of the forms with at most one harmonizing vowel and at most one intervening, conflicting dorsal.

- Aggregate stems based on templatic representations by omitting transparent segments and replacing harmonizing segments with their categories.

E.g.: [b**e**r] → **F**; [m**e**sh**q**] → **FQ**; [ti**k**] → **K** [bir] → **N**

[**o**t] →**B**; [r**a**k] → **BK**; [chi**q**] → **Q** [chish] → **N**

N: transparent vowel; **F**: front vowel; **B**: back vowel; **K**: front dorsal; **Q**: back dorsal

Calculating percent back responses

To calculate the percent of back responses for each template type we:

1. Calculating the proportion of suffixes for each word that are back
2. Take the mean of these proportions for all words for each template

This weights each word equally, regardless of frequency.

- Prevents inflections of high frequency words from overwhelming low frequency ones.

Neutral stems

Stems with no harmonizing elements (**n=69**) vary in their backness:

- Majority are back **(75%)**

- **22%** of neutral stems occur with both front and back suffix forms
  - chish-l**a**r (80% of tokens)
  - chish-l**æ**r (20% of tokens)

Dorsal forms

For stems containing only transparent vowels and one harmonizing velar (**n=21**) or uvular (**n=25**):

- **Q** forms: **97%** back tokens.
- **K** forms: **15%** back tokens.

The high proportion of back responses for **K forms** is largely due to *gezit* “newspaper”.

- Exclusively takes back suffixes!

No stem-suffix pairs alternate.

Front and back forms

Stems with a single front (**n=78**) or back (**n=123**) vowel harmonize as expected:

- **F** forms: **2%** back tokens

- **B** forms: **98%** back tokens

No stem-suffix pairs alternate.

Forms with dorsal conflicts

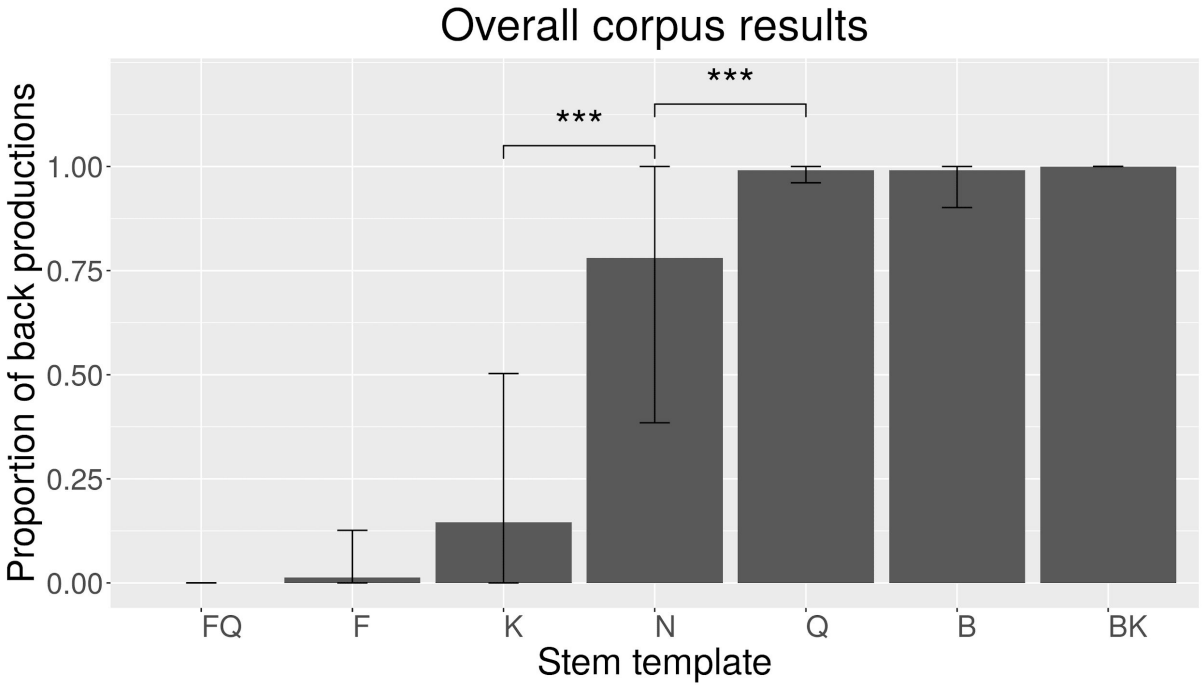
In forms with a front vowel and following uvular (**n=5**) or a back vowel and following dorsal (**n=3**):

- **FQ forms:** 0% back responses

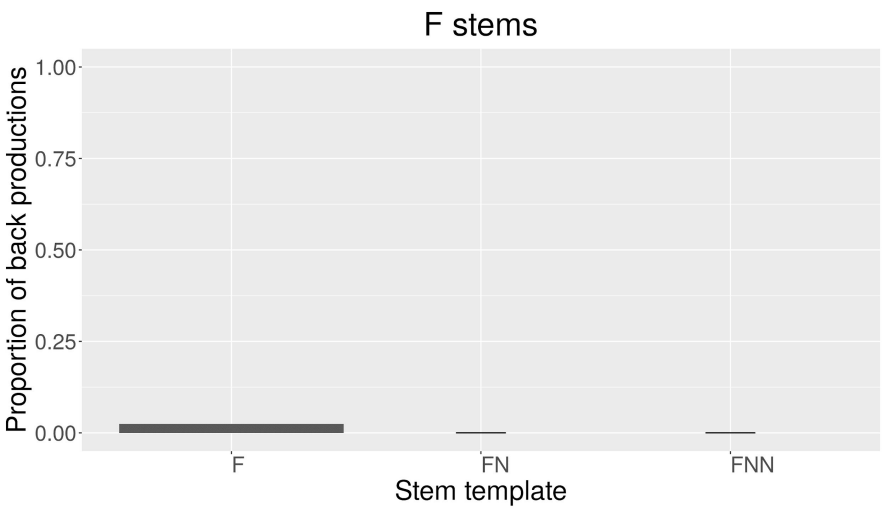
- **BK forms:** 100% back responses

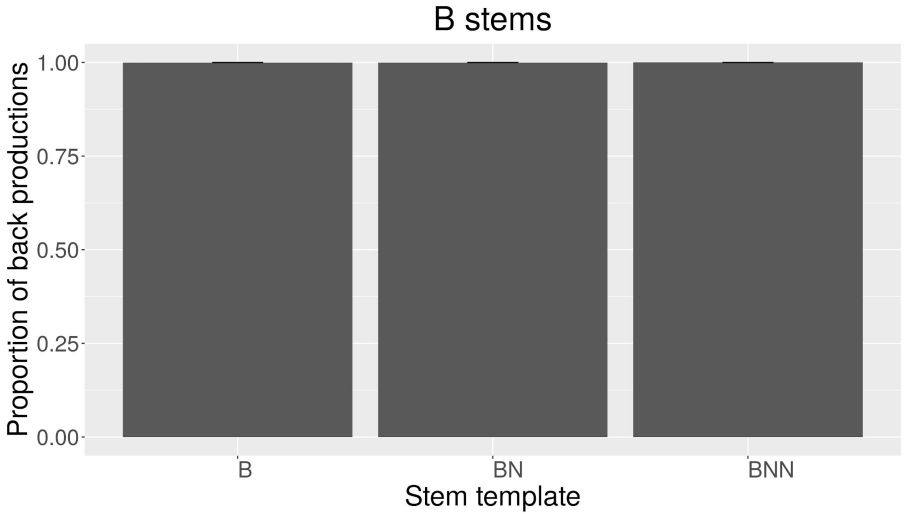
Summary of corpus results

The corpus results pattern essentially as expected from the traditional analysis.



Distance-based decay?



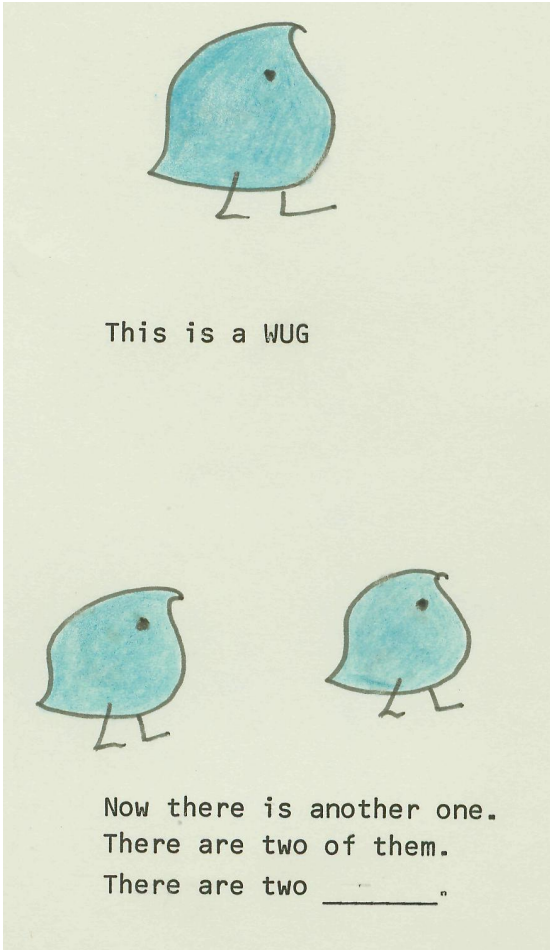


Wug testing backness harmony

Wug tests (Berko 1958) involve asking speakers to inflect unattested word forms.

- Tests (morpho)phonological productivity
- Controls for lexical effects

We use wug tests of Uyghur backness harmony to investigate the productivity of the pattern in speakers' learned grammars.



Creating wug words

We used a custom Python script to generate a large set of wug words matching the following 15 word templates in 5 categories:

- **Non-harmonizers**: CNC, CNCNC, CNCNCNC
- **Front vowels**: C**F**C, C**F**CNC, C**F**CNCNC
- **Back vowels**: C**B**C, C**B**CNC, C**B**CNCNC
- **Front vowels with back dorsal**: C**F**C**Q**, C**F**CN**Q**, C**F**CNCN**Q**
- **Back vowels with front dorsal**: C**B**C**K**, C**B**CN**K**, C**B**CNCN**K**

C: transparent consonant; **N**: transparent vowel;

F: front vowel; **B**: back vowel; **K**: front dorsal; **Q**: back dorsal

Creating wug words

A native Uyghur speaker selected **four words per template**:

- Based on phonological plausibility and balance of vowel qualities
- Consonants were not carefully controlled
- Resulted in a total of **60 wug words**

A few examples:

- **CVC**: [nir], [des], [wiw], [ref]
- **CFQ**: [d͡ʒøʁ], [møʁ], [ryq], [pæq]
- **CBK**: [tug], [mok], [zak], [nuk]

Frame sentences

We embedded wug words in one of **three frame paragraphs**:

- Elicit both **unsuffixed** and **locative suffixed** forms
- Provides a relatively naturalistic context

Example frame:

Ular _____ bir kona sheher dédi. Hazir kishiler _____ (orun kélish) yashimaydu.

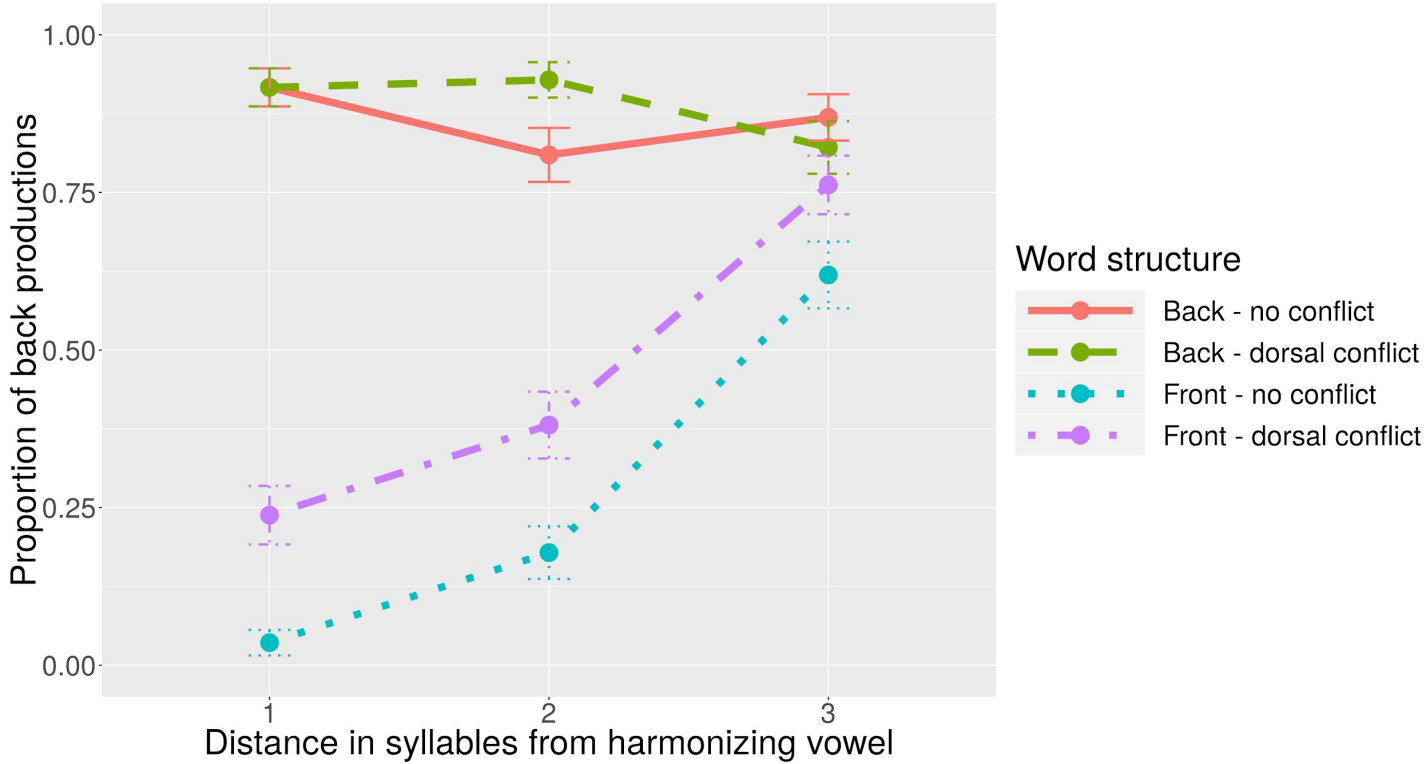
“They say _____ was an old city. Nowadays, people don’t live in _____. “

Participants

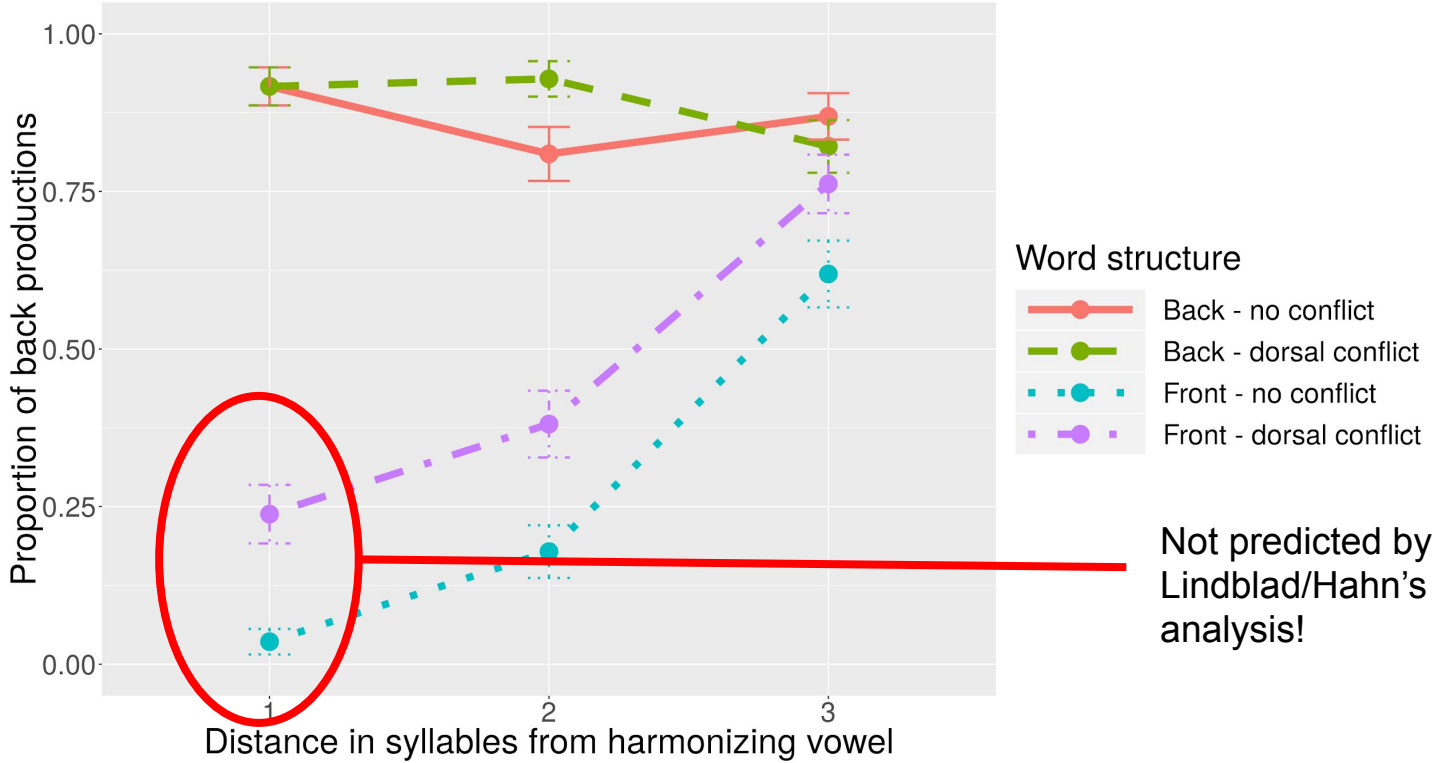
We elicited from **23 native speakers** of Uyghur living in Almaty, Kazakhstan (ages 19-62; mean 40).

- Stimuli were presented in one of two random orders
- 1/3rd of words elicited in each frame paragraph
- Participants were recorded reading stimuli
- Choice of /-Dæ/ vs. /-Da/ form of locative was coded

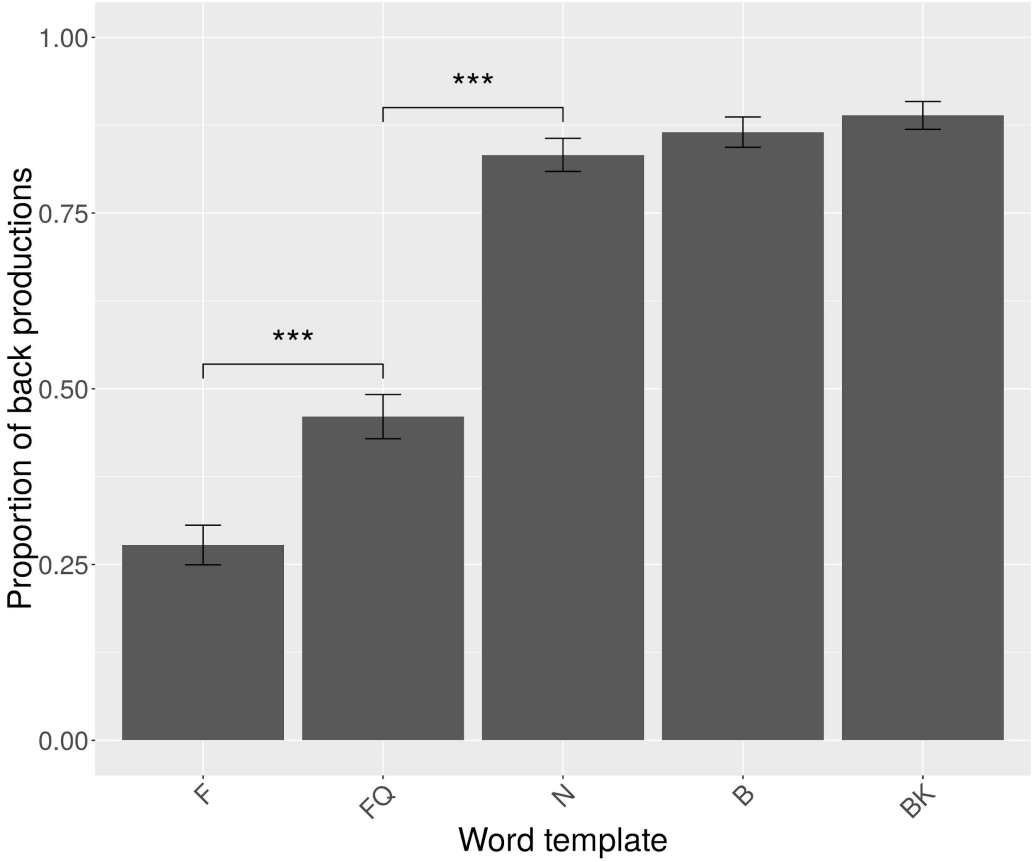
Interactions between distance and conflict



Interactions between distance and conflict



By template (ignoring distance)



Takeaway points

- **Neutral stems** skew heavily towards **back suffixes**

- **Trigger distance** effects only significant for **front vowel triggers**

- **Conflicting trigger** effects only significant for **front vowel triggers**

Attested words

We also elicited a set of attested words from the same group of speakers.

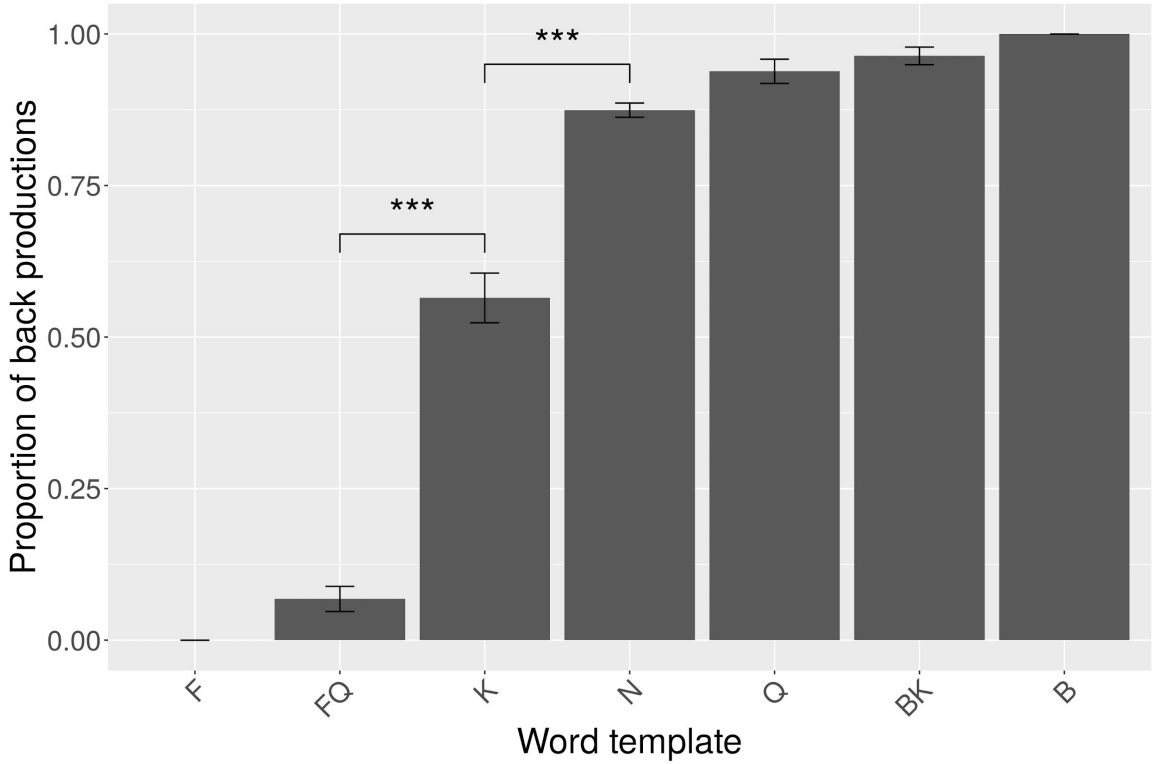
Elicited in random order in a single frame sentence:

> Mahinur _____ deydu
>
> “Mahinur will say _____”

An accompanying phrase indicated whether the word should be produced in the locative form or not.

Set of words was similar (but not perfectly matched) to wug words

Attested word results



Comparing corpus and elicitation results

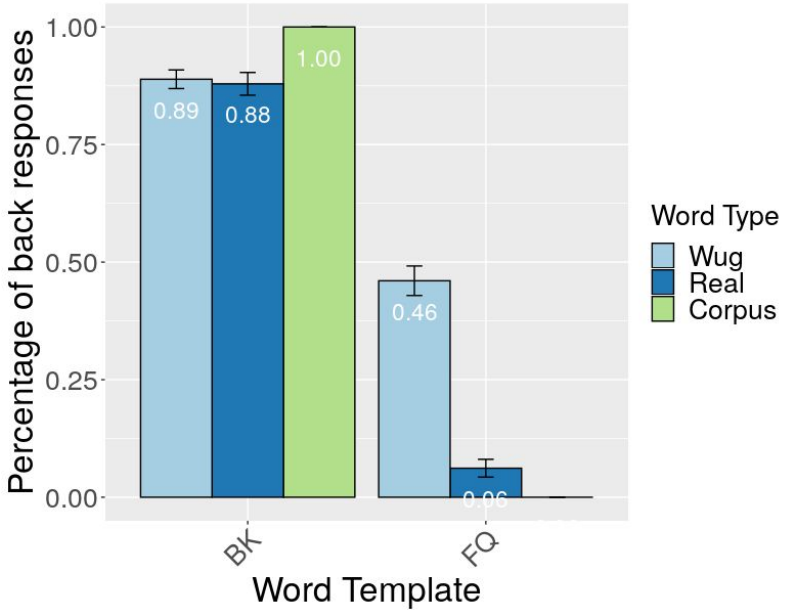
Neutral stems behave **like back forms** in wug and attested word elicitations, but intermediate between front and back in the corpus data.

- When there are no obvious clues, choose the most likely response.

Velars (/k, g/) are much weaker cues for frontness in attested word elicitations than in the corpus data.

Comparing corpus and elicitation results

When conflicting triggers are present, **uvulars** (/**q**/ and /**ʁ**/) **skew suffix choice towards back forms** in wug tests, but not in the corpus or attested word elicitations.



Why do these differences exist?

The grammars suggested by speakers' responses to wug tests and attested word elicitations **do not fit the corpus data well.**

All three data sets can be modeled using maximum entropy Optimality Theory.

- A probabilistic variant of weighted harmonic grammar (Goldwater & Johnson 2003).
- The same set of Agree constraints with different weights can capture all three datasets well

Why do we see this discrepancy?

Previous work on attested/wug discrepancies

Previous work has explained discrepancies between language data and speaker performance by suggesting that statistical patterns may be more effectively learned if they are:

- Phonologically natural (“surfeit of the stimulus”) (e.g., Hayes et al. 2009, Becker et al. 2011)
- Computationally tractable (e.g., Lai 2015, McMullin & Hansson 2019)
- Easily recoverable from surface forms (e.g., Bowers 2019)

Working hypothesis

The computational complexity of the Uyghur pattern poses problems for learnability.

- Corpus patterns are (largely) lexicalized/allomorphy.

The learned grammar is mediated by phonetic and statistical patterns:

- Overall bias towards back suffixes
- Uvulars exert phonetic backing on nearby vowels (e.g., Gallagher 2016)
- This phonetic property influences influences the learned grammar

Conclusions

Uyghur speakers perform backness harmony in a way that does not follow straightforwardly from statistics in the language data.

- Uvulars appear to serve as triggers across the board
- Speaker performance is incompatible with previous analyses that capture corpus data

This discrepancy provides insight into the biases that shape language learning.

- Lots of work remains to be done!

көп рəhmət! Thank you! كۆپ رەھمەت!

Thanks to all of the wonderful Uyghurs who participated in our study in Almaty, Kazakhstan. In particular, we thank Ruslan Arziyov, Shawket Omerov, and Narzigam Makhmudova, without whom this study would have never happened.

Thanks to our research assistants Azadeh Safakish, Kat Vlach, Aiden Jung, Daniela Zokaeim, and Tyler Carson.

Thanks to Bruce Hayes, Kie Zuraw, Tim Hunter, Adam McCollum, Jesse Zymet, and the attendees of UCLA phonology seminar for their valuable comments.

This work was supported by the Social Sciences & Humanities Research Council of Canada, and the UCLA Harry and Yvonne Lenart Graduate Travel Fellowship.

References

Becker, M., Ketrez, N., and Nevins, A. (2011). The surfeit of the stimulus: analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language, 87*(1), 84-125.

Bowers, D. (**2019**). The Nishnaabemwin restructuring controversy: New empirical evidence. *Phonology, 36*(2), 187-224.

Engesæth, T., Yakup, M., and Dwyer, A. (**2010**). *Greetings from the Teklamakan: a Handbook of Modern Uyghur*. University of Kansas Scholarworks, Lawrence.

Gallagher, G. (**2016**). Vowel height allophony and dorsal place contrasts in Cochabamba Quechua. *Phonetica, 73*, 101-119.

Goldwater, S., and Johnson, M. (**2003**). Learning OT constraint rankings using a maximum entropy model. In Spenader, J., Eriksson, A., and Dahl, O. (Eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111–120. Stockholm: Stockholm University, Department of Linguistics.

References

Hahn, R. F. (**1991**). Diachronic aspects of regular disharmony in modern Uyghur. In: Boltz, W. and Shapiro, M. (Eds.), *Studies in the Historical Phonology of Asian Languages*. John Benjamins.

Hayes, B., Zuraw, K., Siptar, P., and Londe, Z. (**2009**). Natural and unnatural constraints in Hungarian vowel harmony. *Phonology, 23,* 59-104.

Lai, R. (**2015**). Learnable vs. unlearnable harmony patterns. *Linguistic Inquiry, 46*, 425-451.

Lindblad, V. M. (**1990**). *Neutralization in Uyghur*. PhD Thesis, University of Washington.

Mayer, C., and Major, T. (**2018**). A challenge for tier-based strict locality from Uyghur backness harmony. In Foret, A., Kobele, G., Pogodalla, S. (Eds). *Formal Grammar 2018. FG 2018. Lecture Notes in Computer Science, vol 10950*. Springer, Berlin, Heidelberg.

McCollum, A. (**2019**). Transparency, locality, and contrast in Uyghur backness harmony (ms.). Rutgers.

References

McMullin, K., and Hansson, G.O. (**2019**). Inductive learning of locality relations in segmental phonology. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 10,* 14.

Nazarova, G., & Niyaz, K. (**2013**). *Uyghur: An elementary textbook.* Washington, DC. Georgetown University Press.

Nazarova, G., & Niyaz, K. (**2016**). *Uyghur: An intermediate textbook.* Washington, DC. Georgetown University Press.

Appendices

Statistics

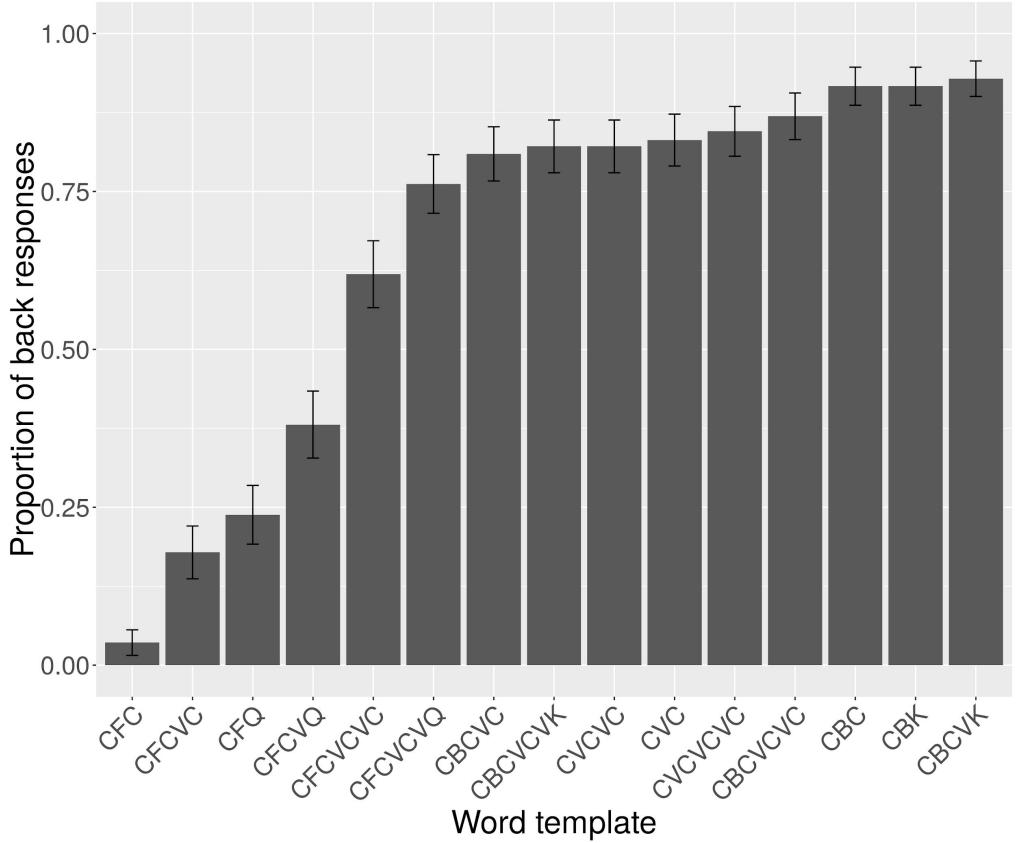
We used linear-mixed effects models (*lme4* in R) to perform statistical tests:

- **Dependent variable**: Suffix choice
- **Independent variables**:
  - Word template ignoring transparent segments
    - **B**, **F**, **BK**, **FQ**, **N**
  - Distance in segments from trigger vowel to suffix
    - Not applicable for **N** stems
- **Random effects:** Subject

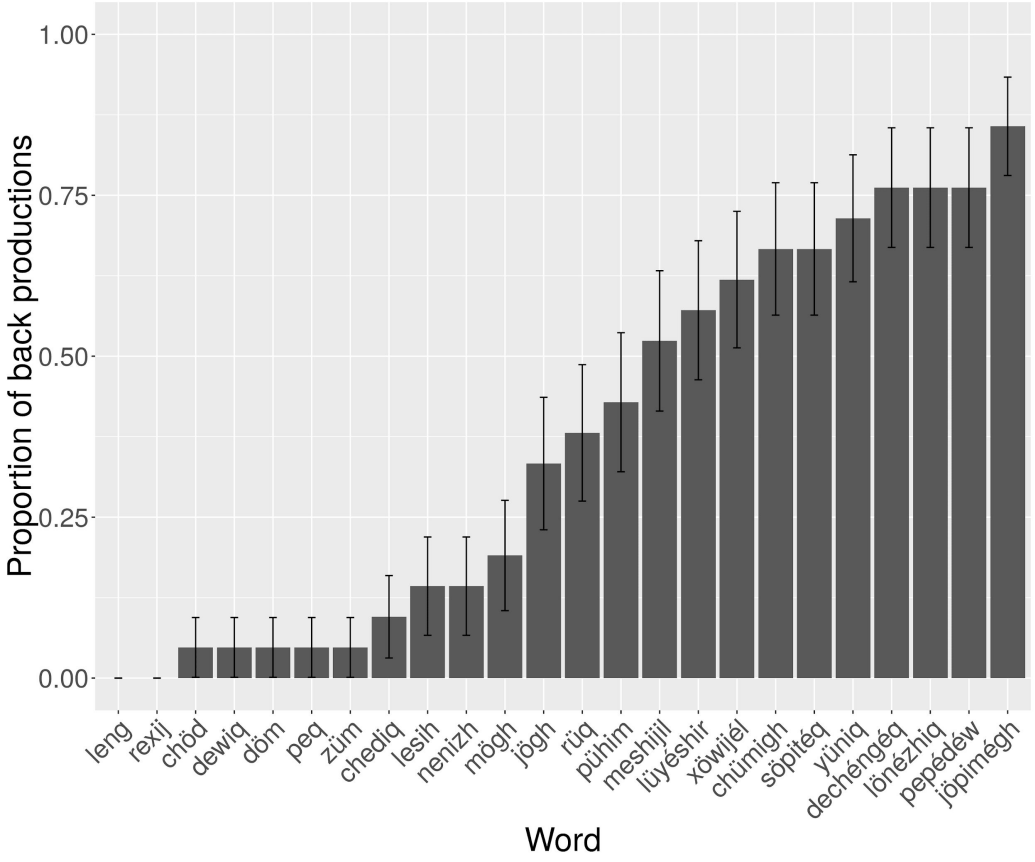
Post-hoc pairwise comparisons done using the *emmeans* package

p-values calculated using *lmerTest* package

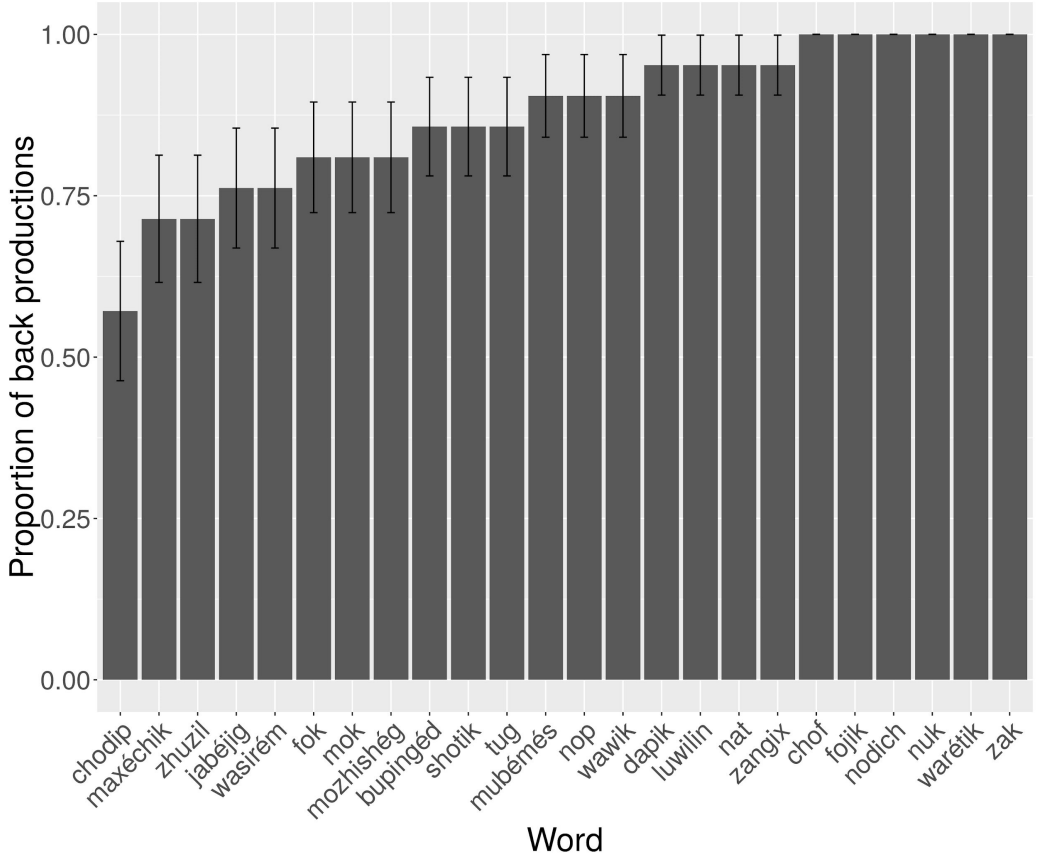
Wugs by detailed template



Wugs by word (front triggers)



Wugs by word (back triggers)



Wugs by word (no trigger)