Different [back][round]s

Lexical origin and vowel harmony in Turkish

Michael Dow & Jeffrey Lamontagne

Université de Montréal & Indiana University Bloomington

June 17th, 2024

Roadmap

1. Introduction
   - Turkish vowel (dis)harmony
   - Phonology on Twitter
   - The current study
2. Methodology
   - Corpus collection and extraction
   - Statistical analysis
3. Results
   - Overview
   - Harmony classes
   - Etymological origin
   - Parasitic harmony
4. Discussion
   - Phonological behaviour of borrowings
   - Revisiting lexical strata

Introduction

Turkish vowel harmony

- The Turkish vowel system makes use of phonemic height, rounding, and backness, with considerable symmetry:

	Front		Back	
	Non-round	Round	Non-round	Round
High	/i/	/y/	/ɯ/	/u/
Non-high	/e/	/ø/	/a/	/o/

- NATIVE roots are mostly harmonic
  - Backness harmony targeting all vowels
  - Rounding harmony targeting high vowels
- QUASI-BORROWINGS
  - From Persian and Arabic
  - Main source of disharmonic words in Turkish
  - Constitute older borrowings or share an etymological source with them
- BORROWINGS
  - Mainly from English, French or Italian
  - Mostly more recent borrowings

Turkish vowel harmony

- Backness harmony in non-high vowels is analysed as productive in suffixes (e.g., Bennink, 1992):

[+fr,-rd]	[+fr,+rd]	[-fr,-rd]	[-fr,+rd]	
/gelir/	/kyj/	/kɯz/	/okul/	(Root)
[ɟelir+ler]	[cyj+ler]	[kɯz+lar]	[okul+lar]	Nom. pl.
‘income’	‘village’	‘girl’	‘school’	(Gloss)

- Rounding and backness harmonies in high vowels are analysed as productive in suffixes (e.g., Bennink, 1992):

[+fr,-rd]	[+fr,+rd]	[-fr,-rd]	[-fr,+rd]	
/gelir/	/kyj/	/kɯz/	/okul/	(Root)
[ɟelir+im]	[cyj+ym]	[kɯz+ɯm]	[okul+um]	1SG poss.
‘income’	‘village’	‘girl’	‘school’	(Gloss)

Disharmony in Turkish

- There's some debate about whether harmony is synchronically productive in Turkish roots, but productive application in suffixes is uncontroversial (e.g. Bennink, 1992; Polgárdi, 1999; Chong, 2019)
- Historical borrowings, especially from Arabic and Persian (11th century+), have introduced unrepaired disharmonic roots in the past (e.g. Lanfranca, 2012):
  - /mysade/ 'permission', /kitap/ 'book', /kalem/ 'library', /meny/ 'menu'
- Some suffixes, borrowed and not, do not exhibit (full) harmony (e.g. Baturay, 2012; Lanfranca, 2012):
  - /-izm/ '-ism', /-Ijor/ 'pres. prog.', /-ki/ 'loc. nom.'
- Borrowings may not consistently trigger suffix harmony (e.g. Baturay, 2012)
- No large-scale study has yet been conducted

Twitter for phonology

- Spelling – especially standardised – can obviously pose challenges to analysing phonological variation
- Despite this, work on phonological variation in written media, specifically Twitter, shows evidence for overt and covert representation of phonological processes (e.g. Eisenstein, 2013; Tatman, 2016; Lamontagne and McCulloch, 2022; Dalola, 2022; Law, 2022)
- Prior cases largely require the stylistic intent, even if not the phonological phenomenon under investigation
- Turkish vowel harmony is a perfect case:
  - The contrast is unambiguously encoded in the spelling
  - No alternative (clear “default”) spelling is available, so a form communicating (dis)harmony *must* be selected

Goals

1. How productive is vowel harmony in suffixes?
2. Can we identify a “default” vowel quality or an active feature through decreased productivity?
3. Does the etymological source of the root affect harmony application?
4. Do borrowing properties (potential harmony trigger vowel, root length) affect harmony application?

Methodology

Corpus creation: Target roots

- Extraction of all nouns with transcriptions and etymological information from the Turkish Wiktionary lexicon provided by the Kaikki archive (Ylonen, 2022)
- Elimination of words of 4+ syllables and vowel-final roots
- Classification of noun by harmony class based on vowel of root-final syllable (front unrounded, front rounded, back unrounded, back rounded)
- Classification of etymological source based on language tags with manual verification (NATIVE, QUASI-BORROWINGS, and BORROWINGS)
- Extraction of lexical frequency (2016 Open Subtitles database)
- Restriction to ten most frequent roots for each permutation of harmony class (front unrounded, front rounded, back unrounded, back rounded), etymological source (NATIVE, QUASI-BORROWING, BORROWING), and number of syllables (1, 2, 3)

Corpus creation: Token extraction

- 391 target roots were combined with every possible first-person possessive suffix
  - *-im* [im], *-üm* [ym], *-ım* [ɯm], *-um* [um]
  - Selected to have maximal harmony options (both rounding and backness harmonies)
  - Variants with root-final <ğ> were additionally generated for all borrowings ending in /k/ to account for the possibility of velar softening (e.g., *müzik* > *müziğim*)
- Query tweets for each resulting target word (*academictwitteR* package; Barrie and Chun-ting Ho 2021) with following settings:
  - Turkish-language tweets
  - Exact phrase (i.e. exact target query)
  - Not a retweet (avoids duplicates)
- Date ranges scaled per etymological group out of necessity (BORROWINGS, 10 years; QUASI-BORROWINGS, 1 year; NATIVE WORDS, 1 month)
- Technical caveat: *-im* and *-ım* are conflated by Twitter API; the search *-im* returns both. The distinction has to be reconstituted from the actual text of each tweet

Extraction and coding

- Number of syllables defined as number of vowel symbols in the root
- Presence of disharmonic vowel sequences within root used to categorize roots for each type of harmony
- Roots were classified by rounding and backness of syllable-final vowel
- Presence or absence of backness harmony and rounding harmony was defined for each query between root and suffix variant
- Number of query results summed according to word length, origin, and harmony profile

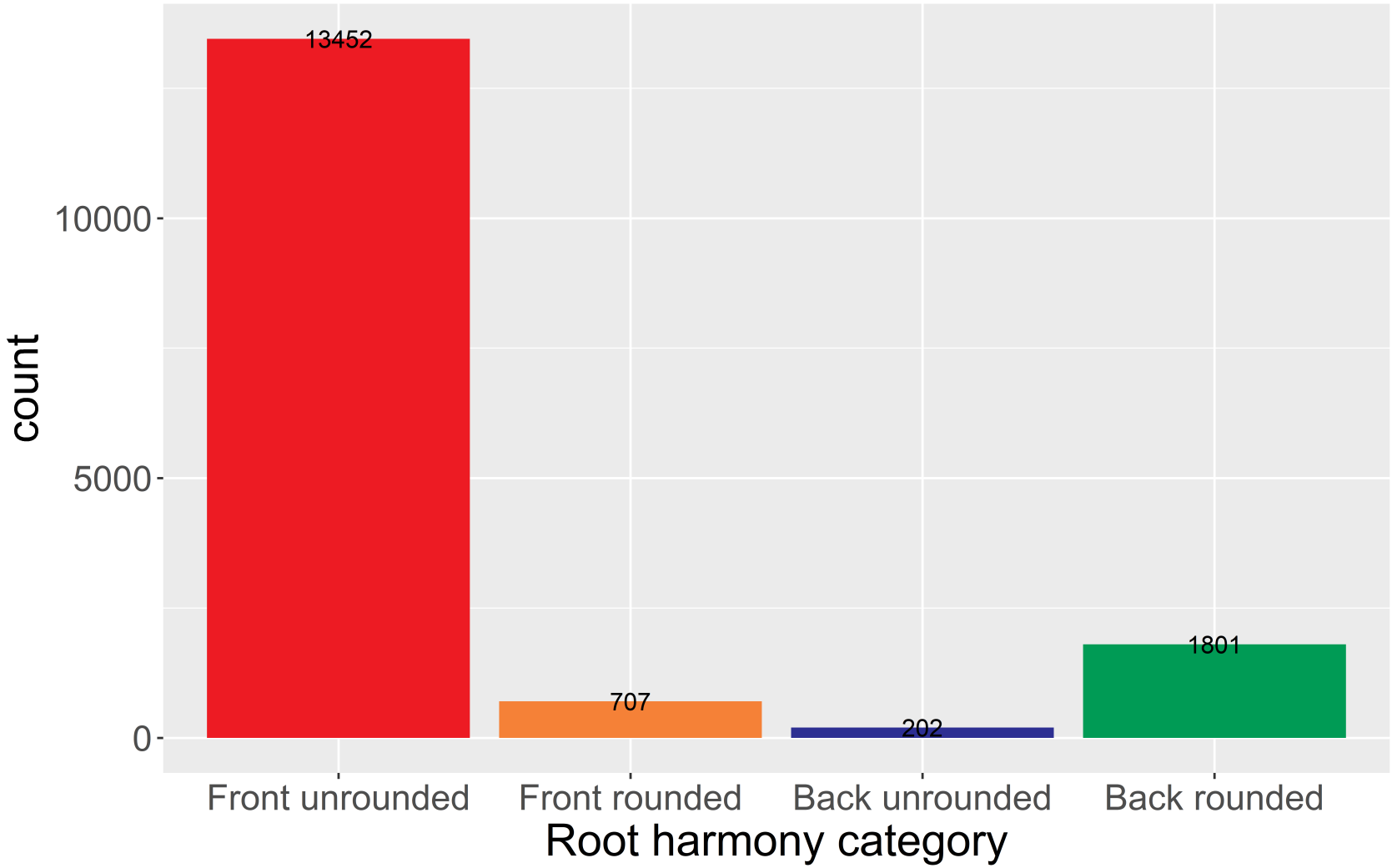
Statistical analysis

- Mixed-effects logistic regression (lme4; Bates et al., 2015)
  - 26 182 tokens
  - Dependent variables: Rounding harmony, Backness harmony
  - Random intercept: root
  - Fixed effects:
    - Lexical origin (NATIVE, QUASI-BORROWING, BORROWING)
    - Root harmony class (front unrounded, front rounded, back rounded, back unrounded)
    - Root profile (Monosyllabic, Harmonic, types of disharmony)
    - Pairwise interactions

Results

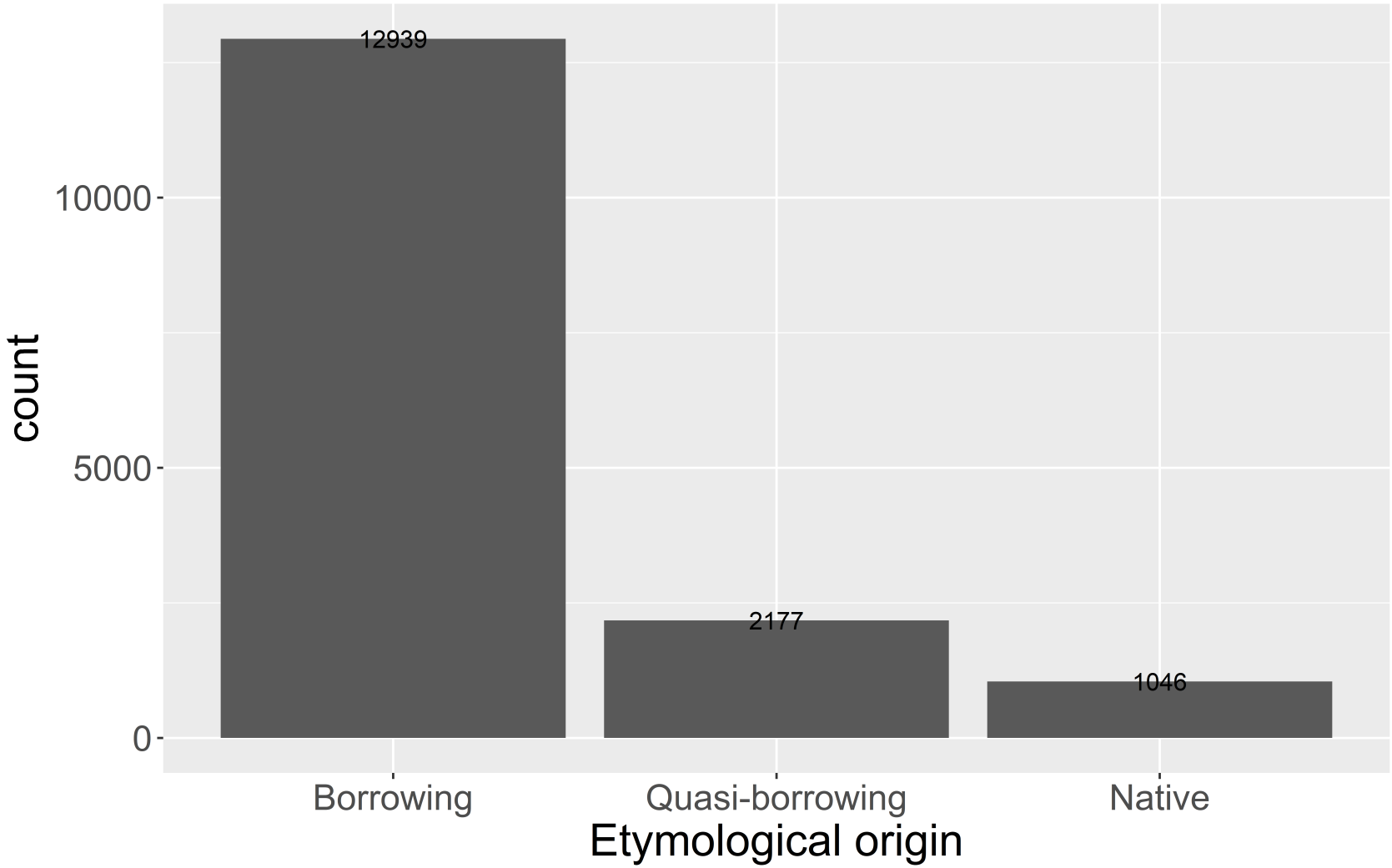
Overview

16 162 tokens: Back unrounded roots (202) are least frequent in the dataset, while front unrounded roots (13 452) are especially frequent.



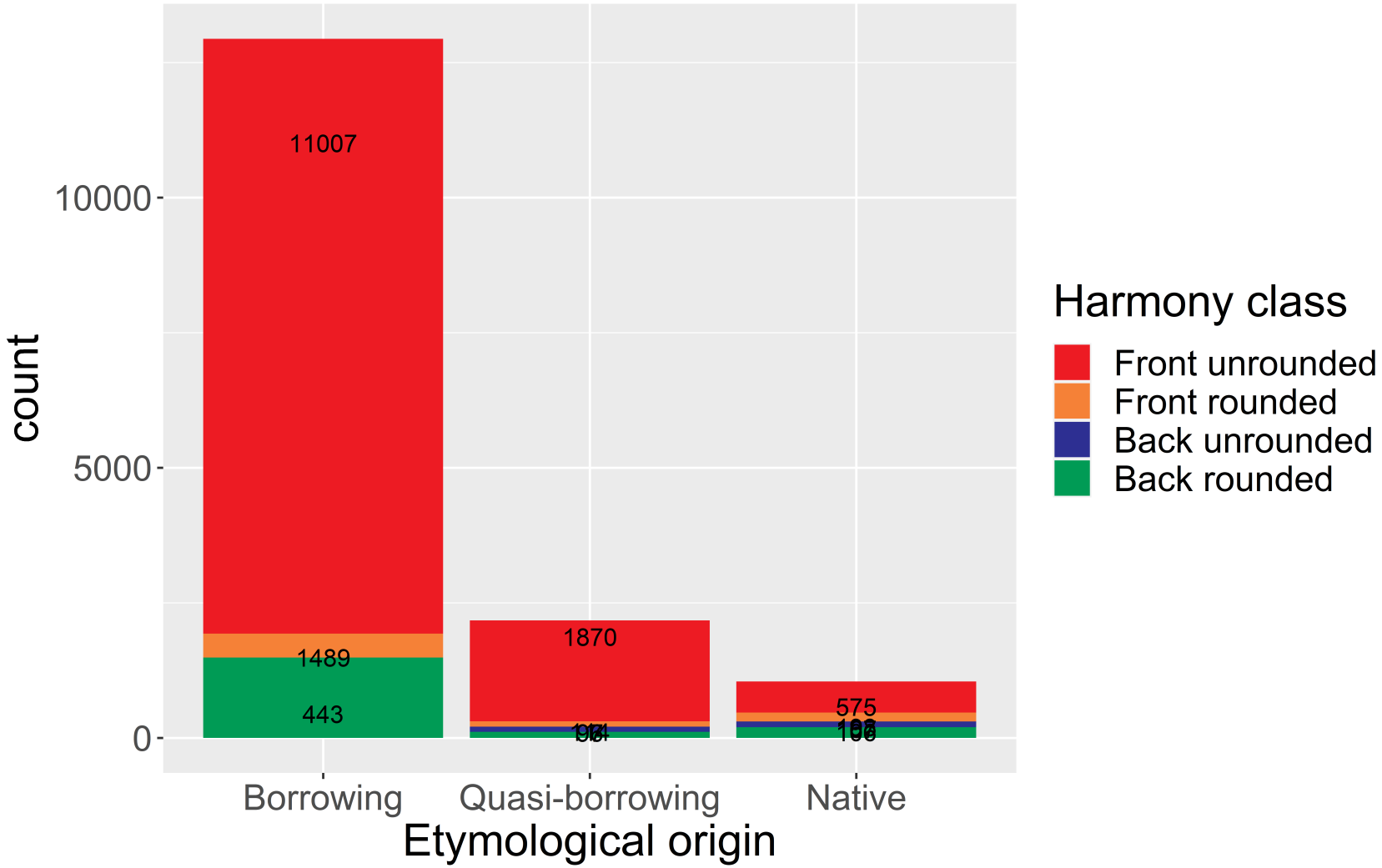
Overview

12 939 Borrowings, 2 177 Quasi-borrowings, 1 046 Native roots.



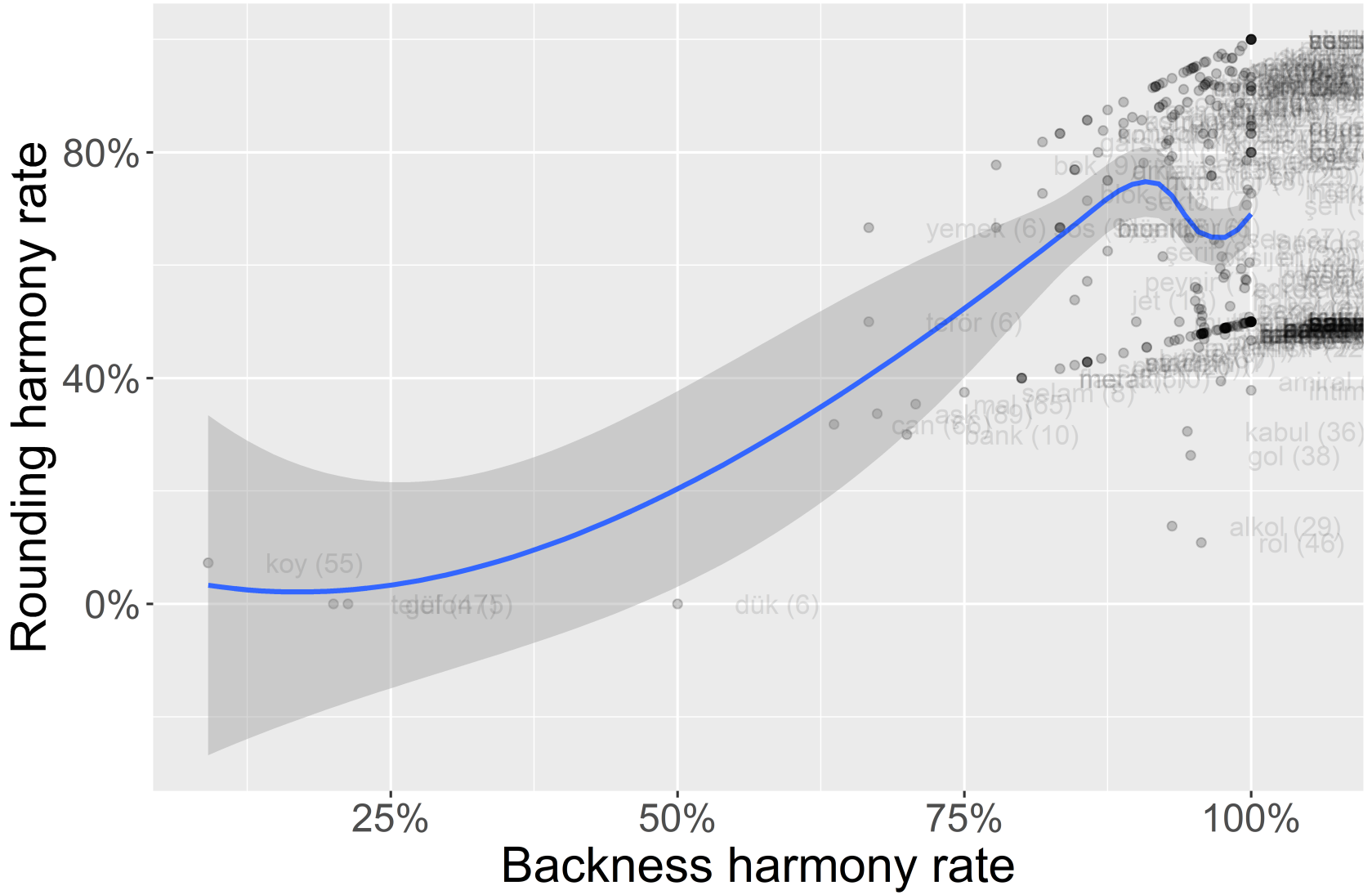
Overview

All harmony classes are represented for all origins, though back unrounded roots are relative rare for all harmony classes.



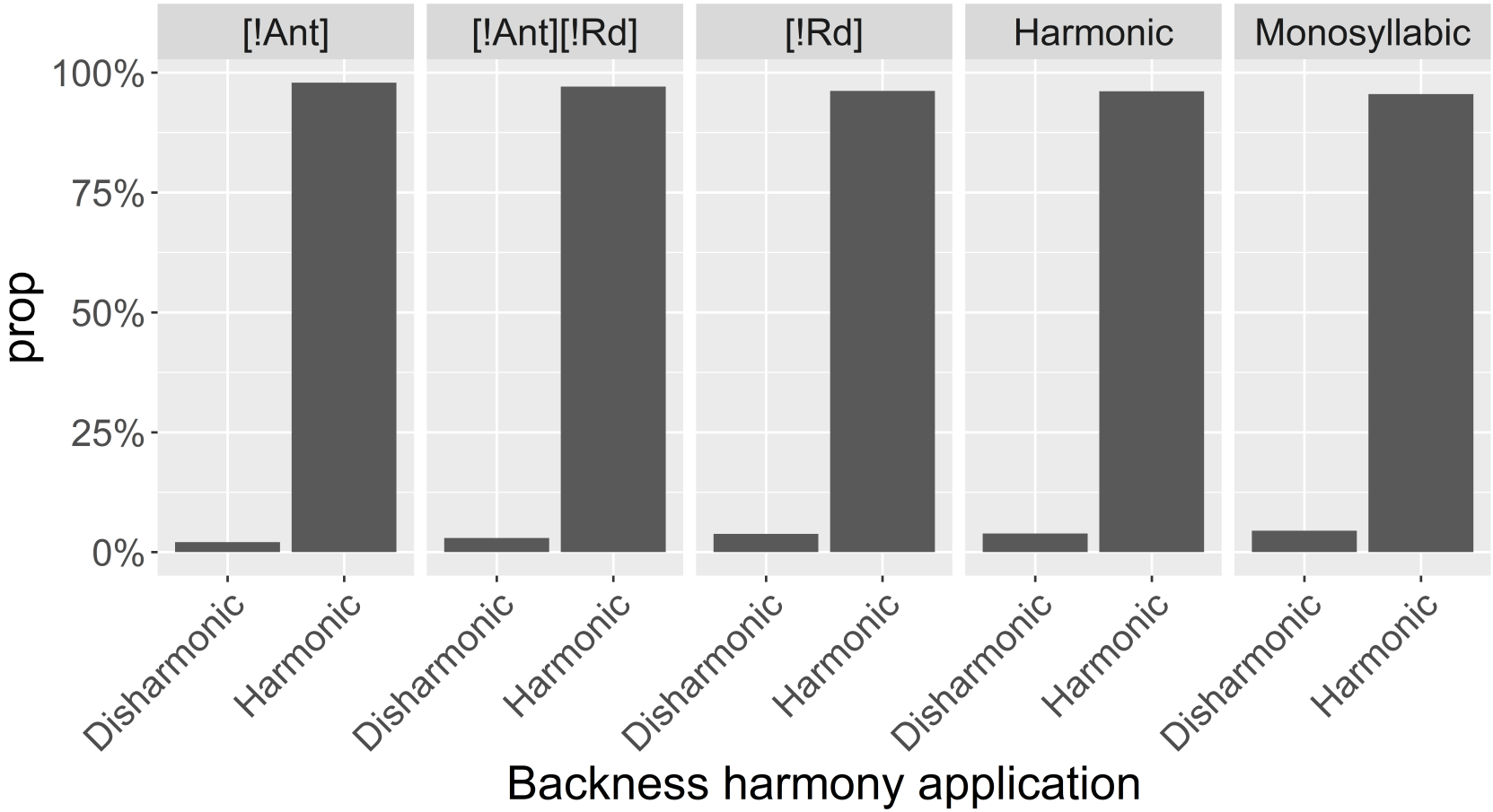
Relationship between harmonies

Exhibiting backness harmony makes rounding harmony more likely, and exhibiting rounding harmony makes backness harmony more likely.



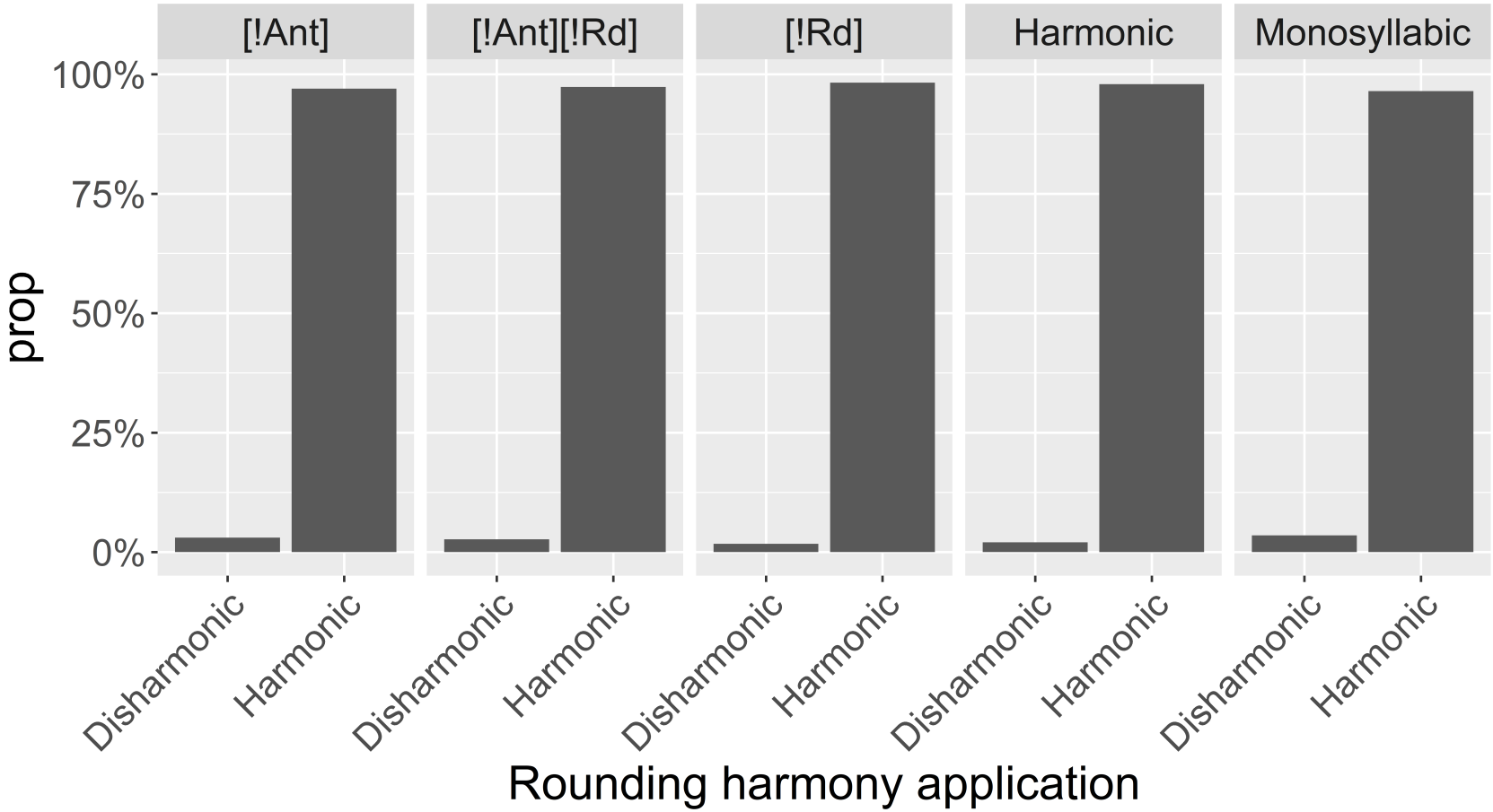
Root profile

- There's a small but significant increase in backness *dis*harmony for monosyllabic roots.
- Disharmonic and harmonic roots pattern together overall; backness *dis*harmony is not readily analysed as harmony triggered by a non-final vowel in the root.



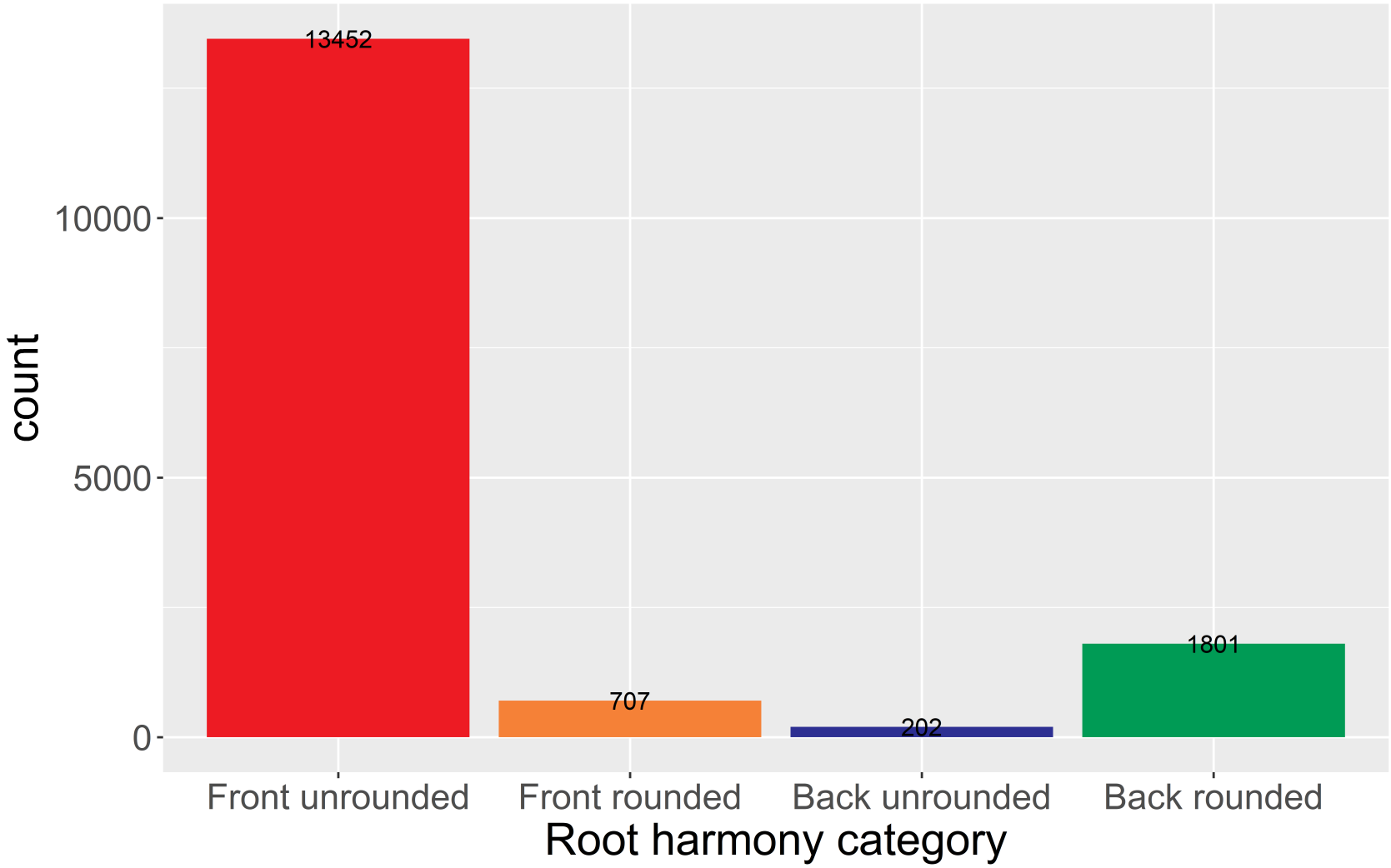
Root profile

- There's a small but significant increase in rounding *dis*harmony for monosyllabic roots.
- Disharmonic and harmonic roots pattern together overall; backness *dis*harmony is not readily analysed as harmony triggered by a non-final vowel in the root.



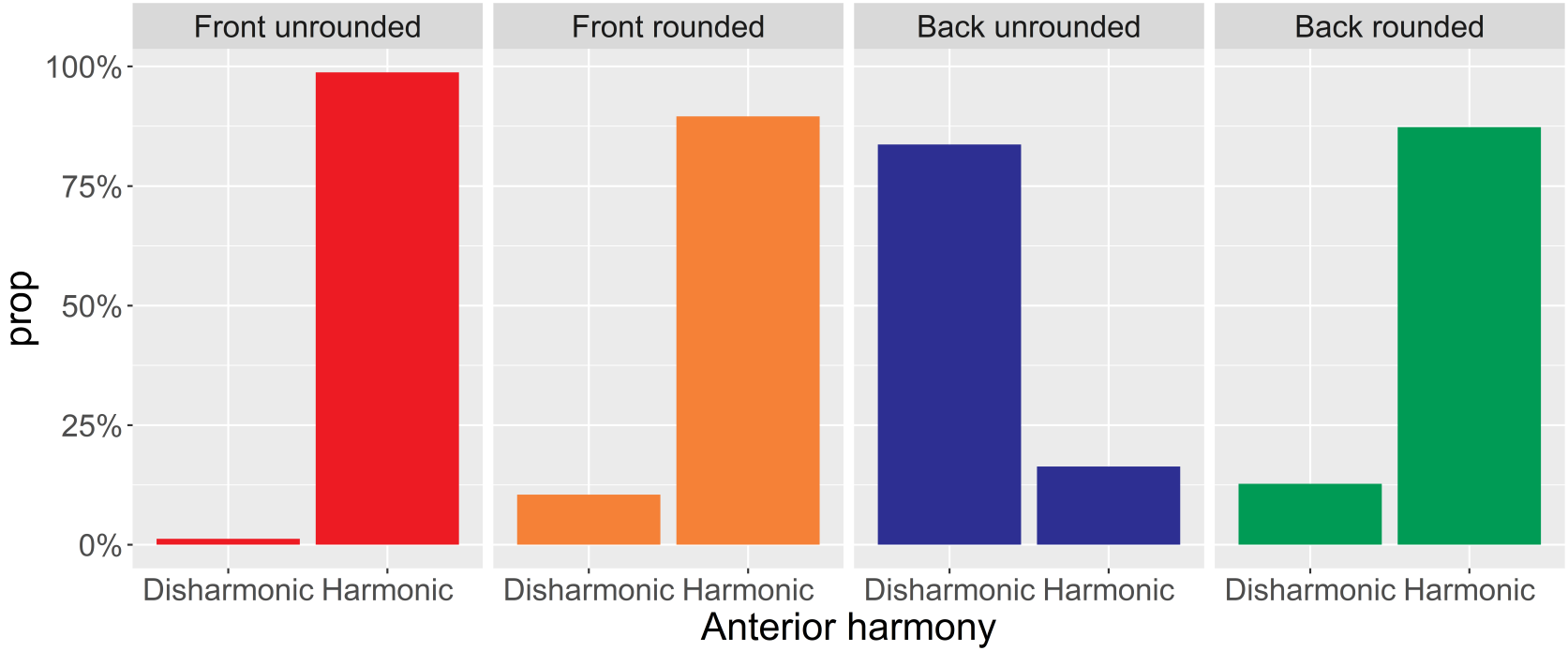
Harmony class

Returning to vowel class on its own.



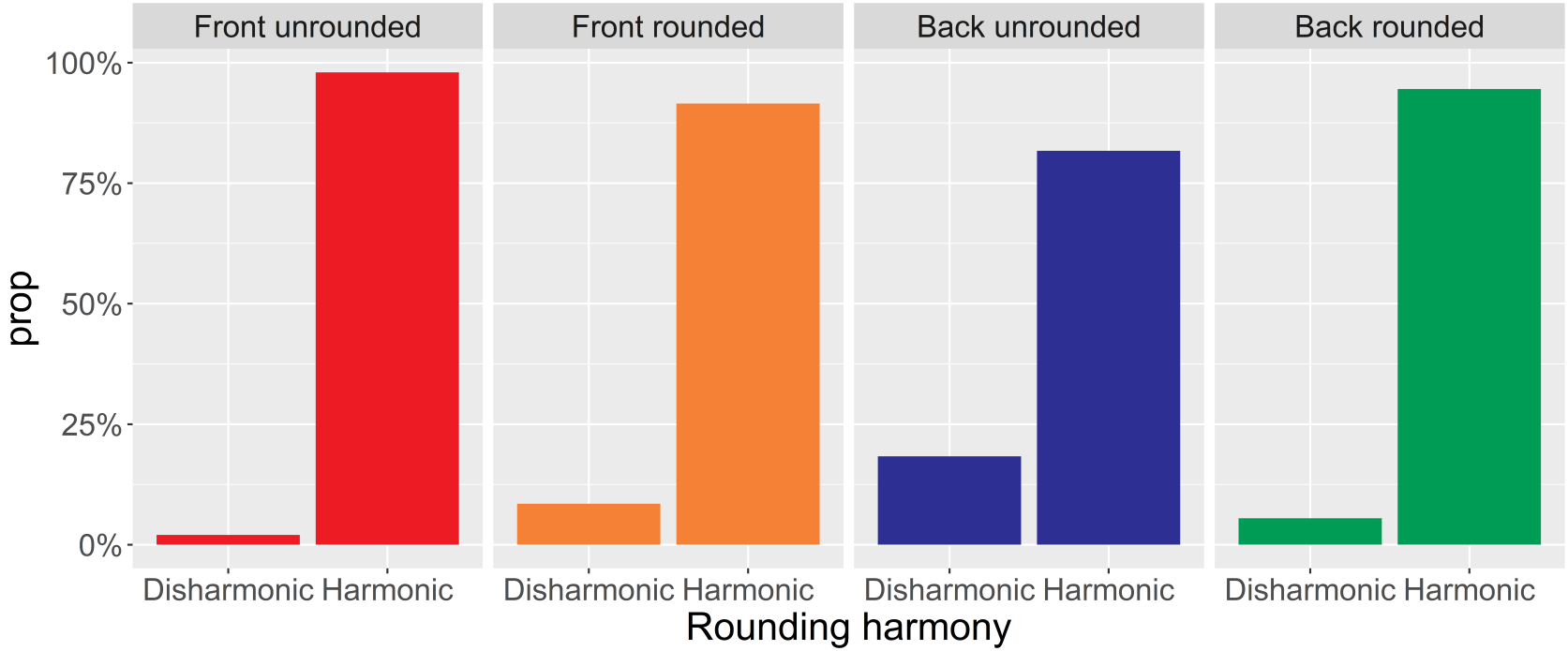
Harmony class

Back unrounded roots exhibit backness *dis*harmony in a majority of cases, otherwise backness harmony is favoured.



Harmony class

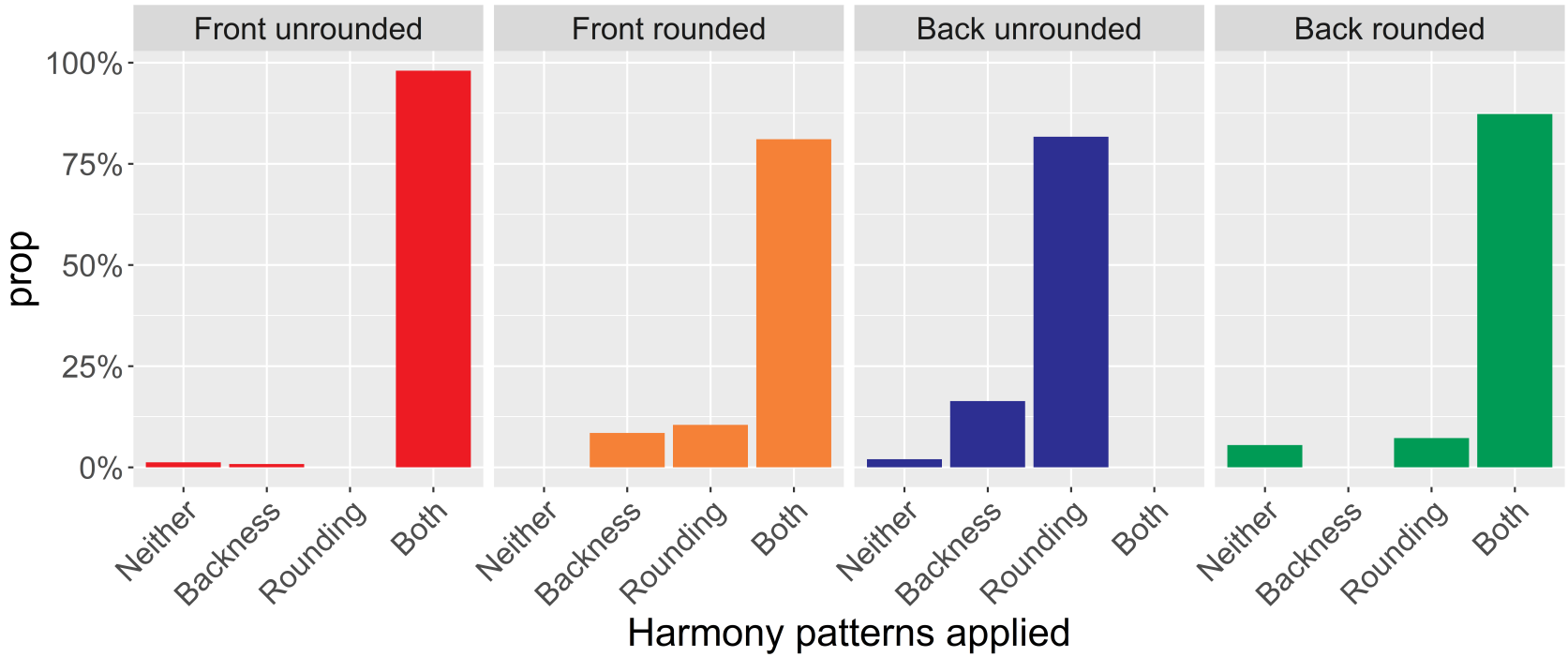
Rounding harmony dominates across the board, though lower rates are found in front rounded roots and especially in back unrounded roots



Harmony class

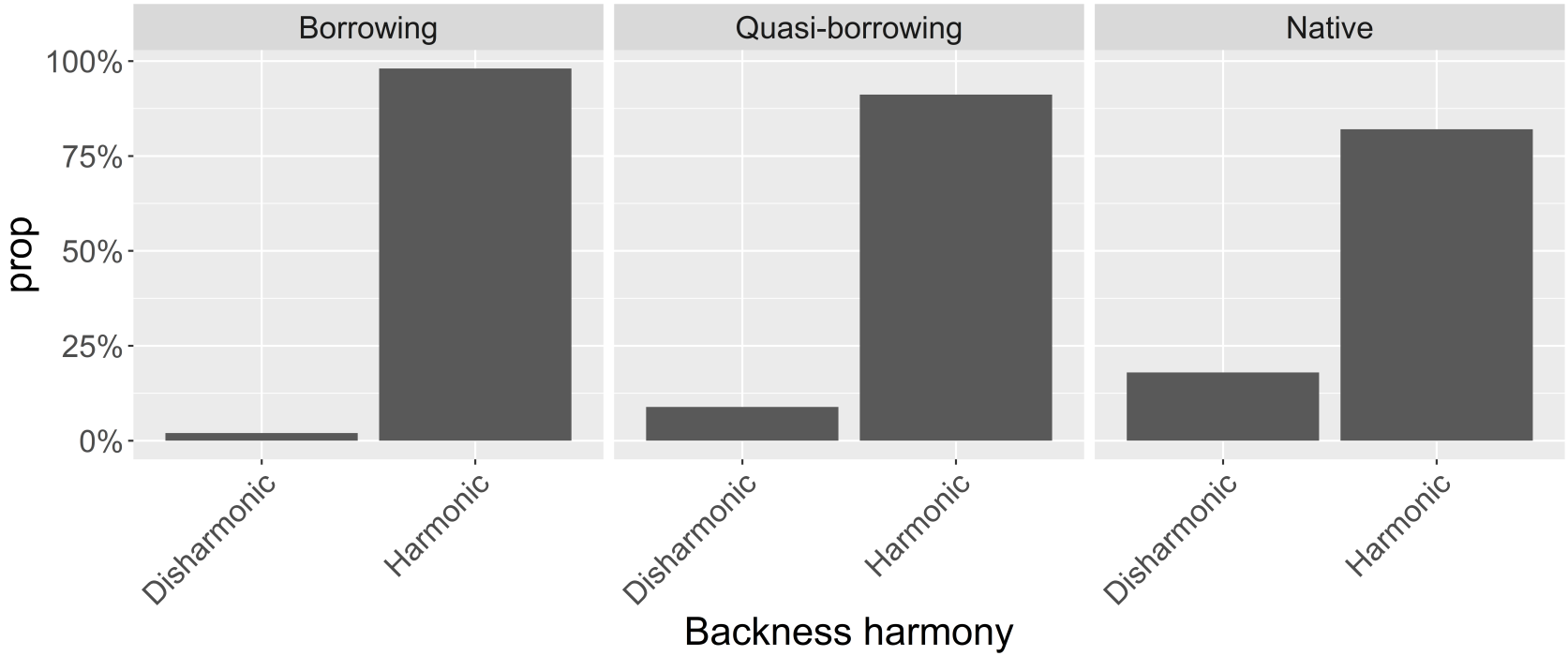
The two harmony patterns do not strictly coincide:

- Front unrounded roots nearly categorically undergo both harmonies. (97.98%)
- Front rounded roots never exhibit *no* harmony, but 18.95% only harmonise one feature.
- Back unrounded roots predominantly exhibit only rounding harmony (81.68%), and no token exhibited both harmonies.
- Back rounded roots predominantly harmonise both features (87.28%), and no token exhibits only backness harmony.



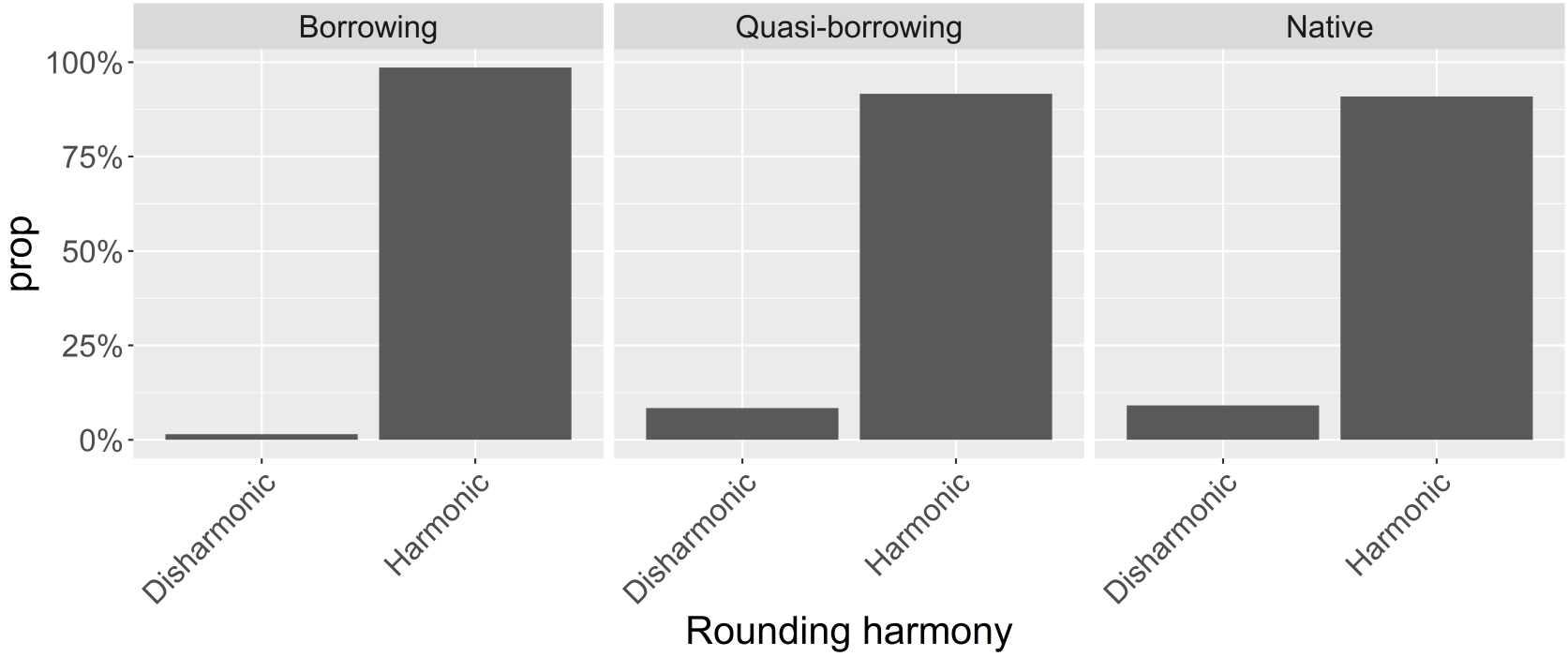
Etymological origin

Returning now to etymological origins – with a warning that a caveat is upcoming!
BORROWINGS are most commonly harmonic for backness, while NATIVE ROOTS exhibit the most backness *dis*harmony.



Etymological origin

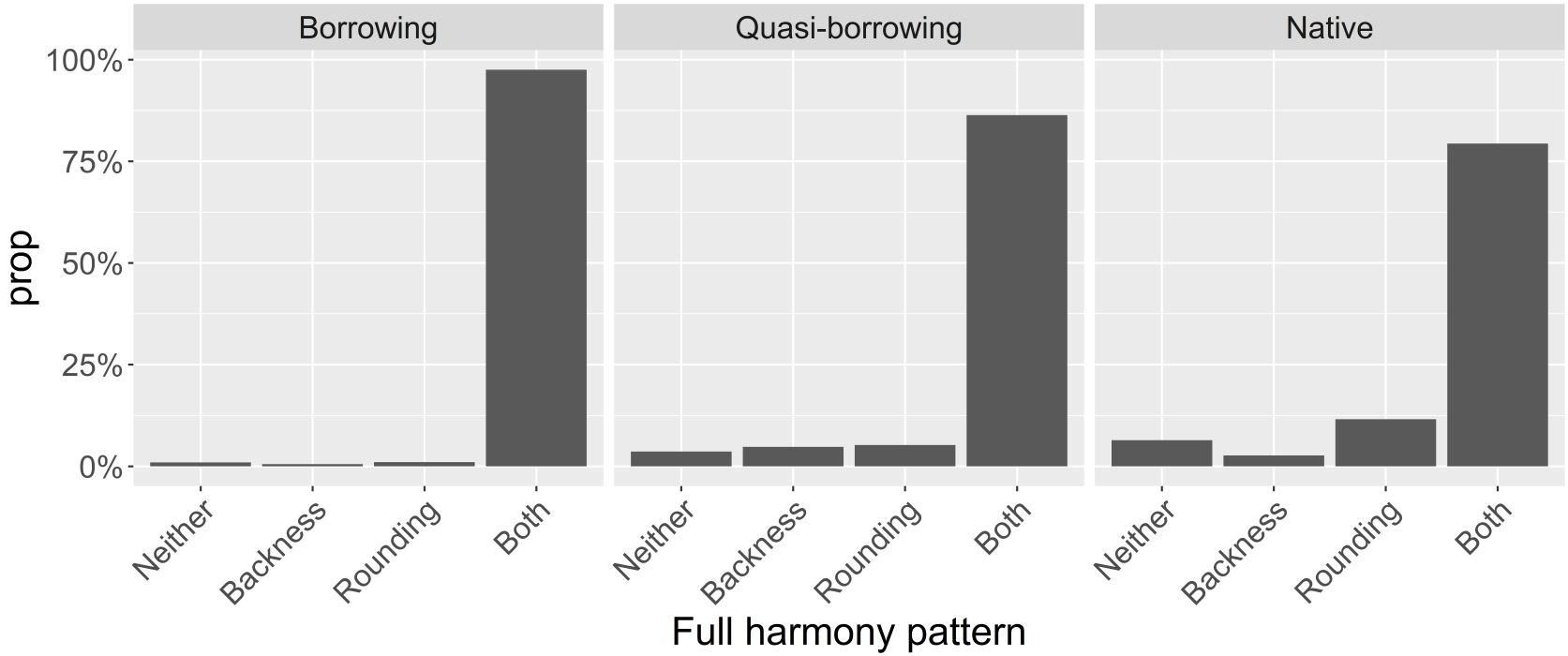
BORROWINGS are slightly more commonly harmonic for rounding, rounding harmony is nearly categorical across etymological origins.



Etymological origin

Combined,

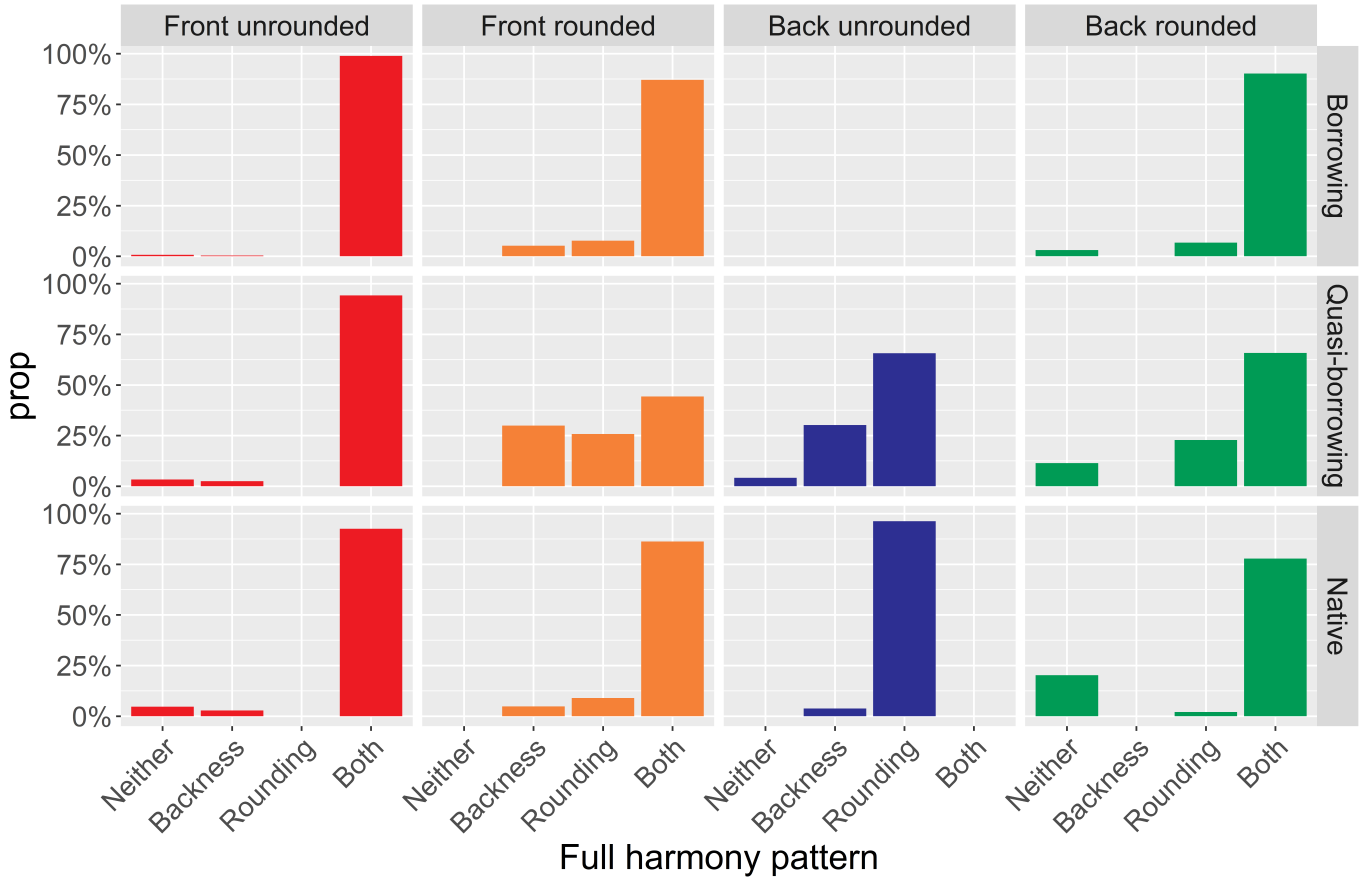
- NATIVE ROOTS exhibit the most full *dis*harmony, but the norm is to apply both harmony patterns.
- QUASI-BORROWINGS exhibit the most (partial) *dis*harmony.



Etymological origin & Harmony class

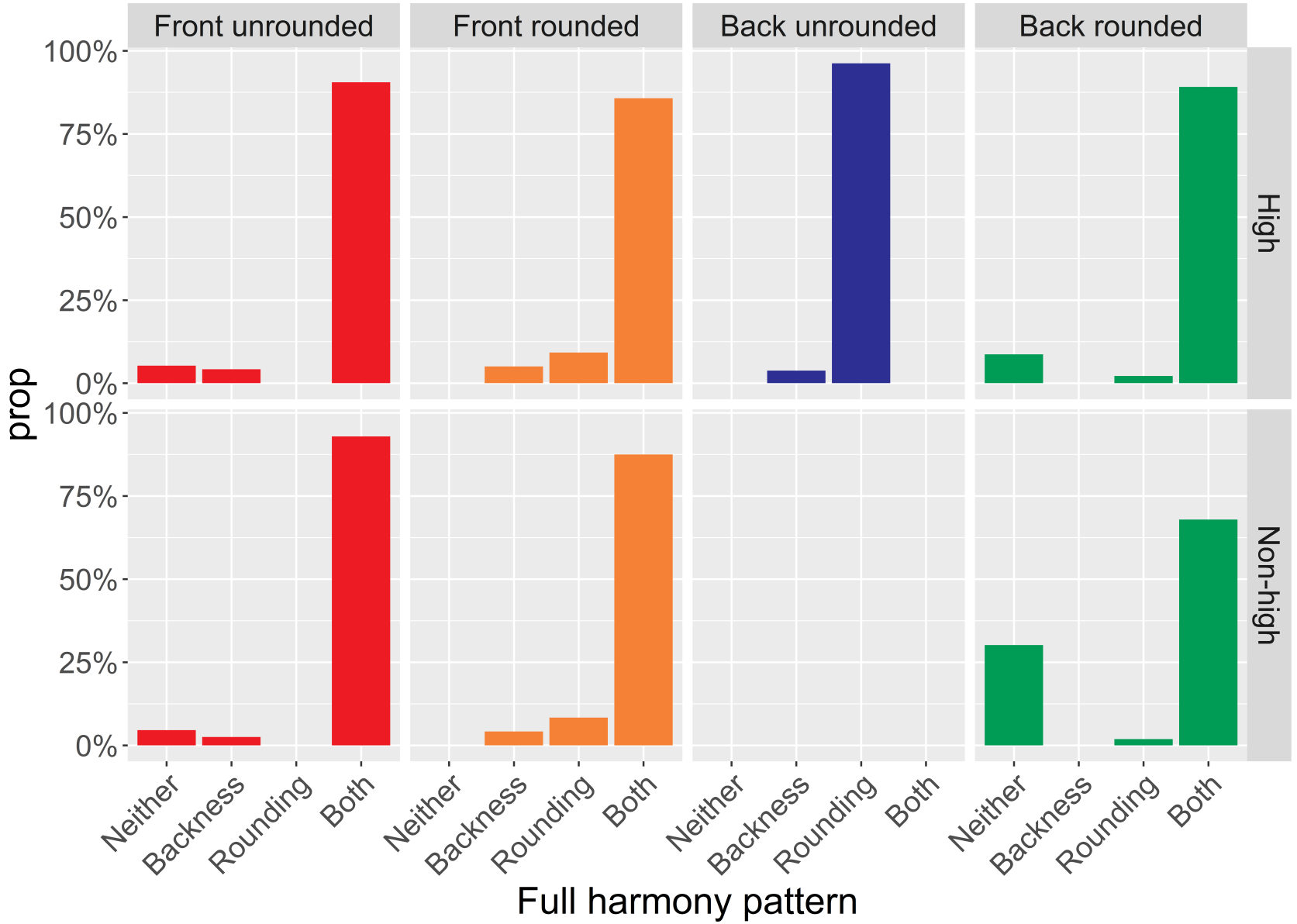
Rates for etymological origin *partly* reflect harmony class distributions:

- QUASI-BORROWINGS exhibit much more rounding *dis*harmony (front rounded and back rounded roots)
- QUASI-BORROWINGS exhibit much more backness *dis*harmony (front rounded and back unrounded roots)
- Back unrounded roots are absent from BORROWINGS because of source language inventories



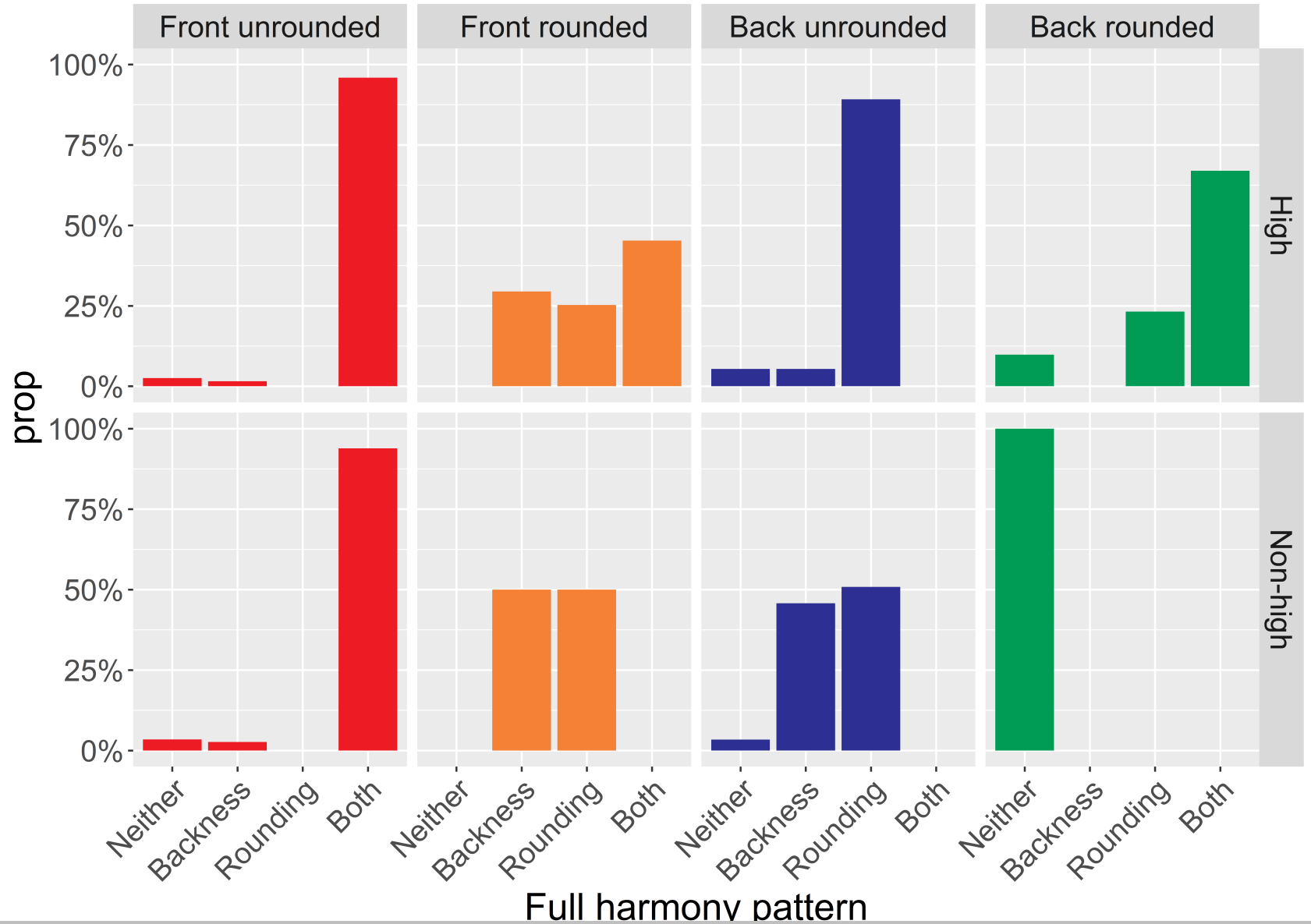
Parasitic harmony (NATIVE ROOTS)

In NATIVE ROOTS, there's more *dis*harmony when the root's last vowel is a non-high back rounded vowel.



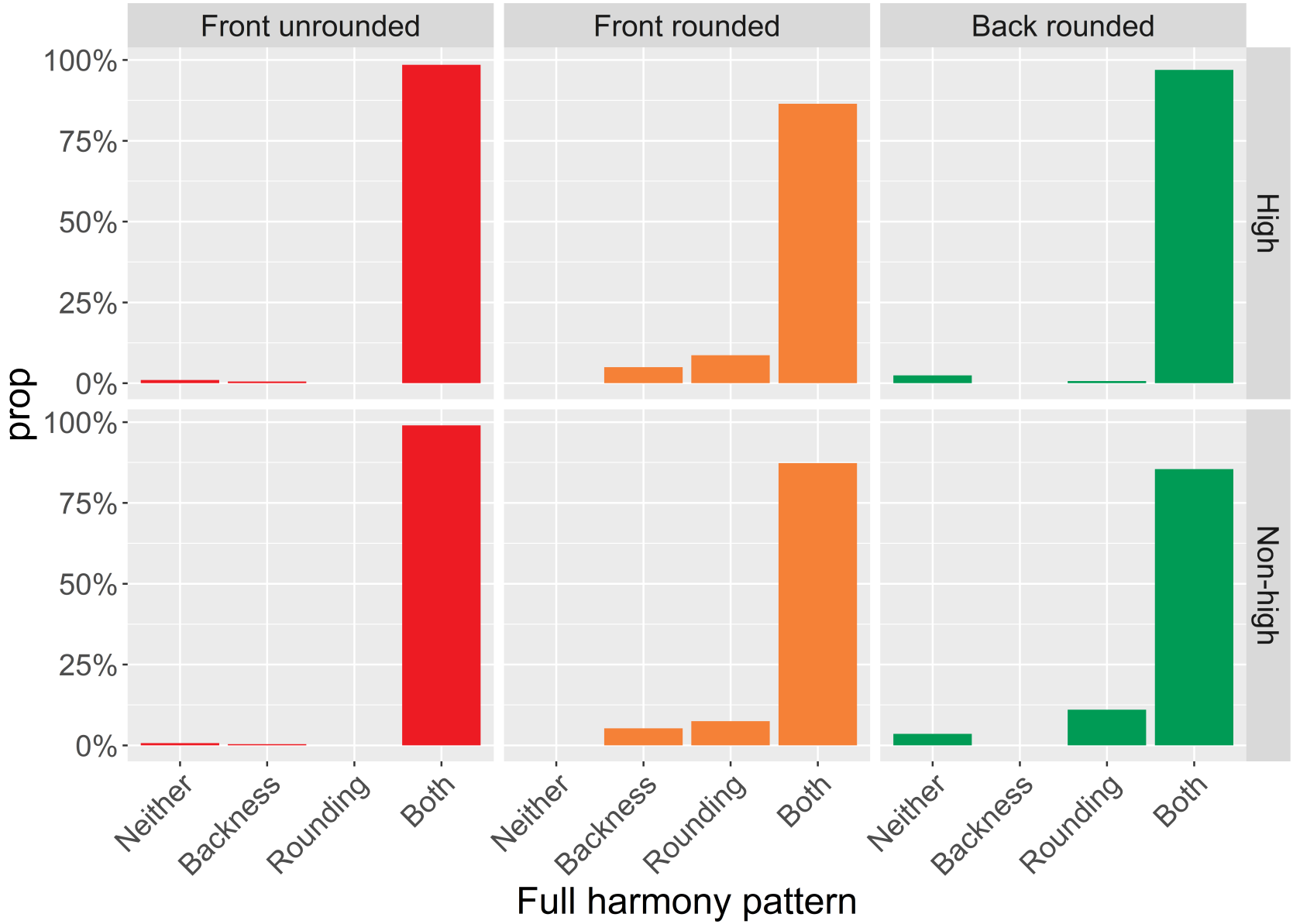
Parasitic harmony (QUASI-BORROWINGS)

In QUASI-BORROWINGS, no *full* harmony with non-high vowels back round and front rounded vowels, introduction of backness *dis*harmony with non-high back unrounded vowels.



Parasitic harmony (BORROWINGS)

In BORROWINGS, slightly less rounding harmony with non-high back round vowels compared to high vowels.



Discussion

Summary

- Harmony dominates, but is not categorical
- Applying one type of harmony predicts applying the other
- Back unrounded vowels are least likely to trigger harmony, particularly with respect to rounding
- Front unrounded vowels are especially likely to trigger harmony
  - Plausible default value for the suffix under analysis? (Apparent only thanks to QUASI-BORROWINGS!)
  - Minimally marked form?
- BORROWINGS are especially likely to be harmonic
- QUASI-BORROWINGS are especially likely to be (at least partially) *dis*harmonic

Borrowing distinctiveness

- Phonological integration typically correlates with morphological integration (Bessett, 2016)
- This is formalised as the *Derived Environment Condition* or as *Non-Derived Environment Blocking*: many processes are described as *requiring* affixation for application *even when the morpheme does not contribute to the required phonological context* (e.g. Kiparsky, 1982; Inkelas, 2014; Chong, 2019)
- It would therefore be expected that *all* words under analysis would undergo harmony
- Disharmony could be a (rarely present!) phonological cue to being a borrowing

Revisiting lexical strata

- Considerable work describes *strata* in lexicons (LaCharité and Paradis, 1993; Itô and Mester, 1999, e.g.)
  - Strata are generally nested
  - More embedded strata are subject to more phonological processes
  - Stratum processes follow the subset principle; a stratum is subject to the processes of all less nested strata
  - Stratum nesting depth reflects stratum age in the lexicon
- For Turkish vowel harmony, this does not appear to hold:
  - The older non-NATIVE stratum (QUASI-BORROWINGS) is *least* subject to harmony (matching lexical patterns in roots?)
  - The newest stratum (BORROWINGS) is *most* subject to harmony
  - While harmony processes are correlated, neither categorically predicts the other
  - When digging into the parasitic nature of high-vowel harmony, QUASI-BORROWINGS exhibit distinctive behaviour.

Future directions

- Increasing the pool of suffixes and roots:
  - Testing high vs. non-high vowels (parasitic vs. non-parasitic harmony patterns)
  - Wider range of affixes (phonological shapes, morphological status)
  - Including lower-frequency roots
- Expanding to speech
  - E.g. querying Youtube video transcriptions to create a corpus
  - Production and/or judgment study

Thanks

- Nathan Samson, for help with queries and planning stages
- The Social Sciences and Humanities Research Council of Canada for research funding (*Harnessing Twitter for morphophonological variation*)

References I

Barrie, Christopher and Justin Chun-ting Ho. 2021. academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint. **Journal of Open Source Software** 6(62): 3272. doi:10.21105/joss.03272. URL https://github.com/cjbarrie/academictwitteR.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. **Journal of Statistical Software** 67(1): 1–48. doi:10.18637/jss.v067.i01. URL https://www.jstatsoft.org/index.php/jss/article/view/v067i01.

Baturay, Semra. 2012. Loan word adaptation and vowel harmony in Turkish: A government phonology account. In **Proceedings of ConSOLE XX**. 1–22.

Bennink, Clemens. 1992. Harmony and disharmony in Turkish. **Linguistics in the Netherlands** 9(1): 25–36.

Bessett, Ryan M. 2016. The role of phonology in discerning lone other-language items as borrowing or codeswitches. In **New waves of analyzing variation (NWAV) 45**. Vancouver, BC, 0–1.

Chong, A. J. 2019. Exceptionality and derived-environment effects: A comparison of Korean and Turkish. **Phonology** 36(4): 543–572.

Dalola, Amanda. 2022. #youarewhatyoutweetchhh: Identity and fricative epithesis in French-language tweets. **Journal of French Language Studies** 32(2): 243–266.

Eisenstein, Jacob. 2013. Phonological factors in social media writing. In **Proceedings of the workshop on language analysis in social media**. 11–19.

Inkelas, Sharon. 2014. **The interplay of morphology and phonology**. Oxford: Oxford University Press.

Itô, Junko and Armin Mester. 1999. The phonological lexicon. In **The handbook of Japanese linguistics**, ed. Natsuko Tsumijura. Oxford: Blackwell Publishers Ltd, 62–100.

Kiparsky, Paul. 1982. Lexical morphology and phonology. In **Linguistics in the morning calm**, ed. I. S. Yang. Seoul: Hanshin, 3–91.

LaCharité, Darlene and Carole Paradis. 1993. Introduction: The emergence of constraints in generative phonology and a comparison of three current constraint-based models. **Canadian Journal of Linguistics/Revue canadienne de linguistique** 38(2): 127–153.

References II

Lamontagne, Jeffrey and Gretchen McCulloch. 2022. Phonological variation on twitter: Evidence from letter repetition in three french dialects. **Journal of French Language Studies** 32(2): 165–196.

Lanfranca, Mark. 2012. **An acoustic study of underspecified vowels in Turkish**. Ph.D. thesis, University of Kansas.

Law, James. 2022. Reflections of the French nasal vowel shift in orthography on Twitter. **Journal of French Language Studies** 32(2): 187–215.

Polgárdi, Krisztina. 1999. Vowel harmony and disharmony in Turkish. **The Linguistic Review** 16(2): 187–204.

Tatman, Rachel. 2016. "i'm a spawts guay": Comparing the use of sociophonetic variables in speech and Twitter. **U Penn Working Papers in Linguistics: Selected Papers from NWAV 44** 18(2): 1–10.

Ylonen, Tatu. 2022. Wiktextract: Wiktionary as machine-readable structured data. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**. 1317–1325.

Bonus

Harmony rate table:

	Front unrounded	Front rounded	Back unrounded	Back rounded
Neither	166 (1.23%)	0 (0.00%)	4 (1.98%)	99 (5.50%)
Backness	105 (0.78%)	60 (8.49%)	33 (16.34%)	0 (0.00%)
Rounding	0 (0.00%)	74 (10.47%)	165 (81.68%)	130 (7.22%)
Both	13181 (97.99%)	573 (81.05%)	0 (0.00%)	1572 (87.28%)
Total	**13452**	**707**	**202**	**1801**