Introduction

Post-Obstruent Tensification (POT) and Cluster Simplification (CS)

Post-Obstruent Tensification (POT) (Kim-Renaud, 1974; Sohn, 1999)

- lax C —> tense C / [p, t, k] _
  - 받고 /pat-ko/ → [pat-k*o] ‘to receive-and’
  - 잡다 /cap-ta/ → [cap-t*a] ‘to hold-DECL’

Cluster Simplification (CS) (Kim-Renaud, 1974; Sohn, 1999)

- CC —> C / _{#, C}
  - 닭 /talk/ → [tak] ‘chicken’
  - 앉는 /anc-nɨn/ → [an-nɨn] ‘to sit-COMP’
  - 굶나 /kulm-na/ → [kum-na] ‘to starve-INTERROGATIVE’

Post-Obstruent Tensification (POT) and Cluster Simplification (CS)

Verbs with coda -lC provide an environment where the rules can both apply.

The majority outcome is opaque:

- 맑고 /malk-ko/ → [mal-k*o] ‘to be clear-and’
- 낡고 /nalk-ko/ → [nal-k*o] ‘to be old-and’

UR	/pat-ko/ to receive-and	/anc-nɨn/ to sit.COMP	/malk-ko/ to be clear-and
POT	pat-k*o	—	malk-k*o
CS	—	an-nɨn	mal-k*o
SR	[pat-k*o]	[an-nɨn]	[mal-k*o]

- CS applies too late to bleed POT → **counter-bleeding opacity**

Post-Obstruent Tensification (POT) and Cluster Simplification (CS)

We should also check environments for POT and CS do not overlap, a "base rate" of process application.

POT should apply to obstruent-final verbs with obstruent-initial affixes:

- 익다 /ik-ta/ [ik-t*a] 'to ripen-DECL'
- 뱉고 /pet^{h}-ko/ [pet-k*o] 'to spit-and'

CS should apply to -lC-final verbs with sonorant-initial affixes:

- 맑나 /malk-na/ [maɲ-na] 'to be clear-INTERROGATIVE'
- 밟는 /palp-nɨn/ [pam-nɨn] 'to step on-COMP'

Reason for further inquiry

Reported cases of variation in opaque context

- 밟고 /palp-ko/ [pal-k*o]~[pap-k*o] ‘to step on-and’ (Kim, 2003)
- 낡지 /nalk-ci/ [nal-c*i]~[nak-c*i] ‘to be old-CONN’ (Kim, 2003)

  N.B. no general post-liquid tensification (/cul-ta/ [cul-ta], *[cul-t*a] ‘to decrease-DECL’)

Nouns have transparent outcomes:

- 닭도 /talk-to/ [tak-t*o] ‘chicken-also’
- 흙과 /hɨlk-kwa/ [hɨk-k*wa] ‘soil-and’

Reason for further inquiry

In line with other evidence that textbook-opaque cases are often quite nuanced (Sanders, 2003; Mayer, 2021; Bowers, 2019; White, 2013)

A thorough, open dataset can be used by many parties to address big topics in phonology:

- Structure of the grammar (rules vs. constraints?)
- Proper treatment of opacity (listing vs. generation?)
- Ecologically-valid, full-scale learning models

→SIGMORPHON Shared Task

Corpus studies

Adult corpus study

- NIKL Korean Dialogue Corpus: ~900,000 Intonational Phrases (National Institute of Korean Language, 2021)
- Semi-spontaneous speech with phonemic transcriptions
- Extracted all affixed -lC verbs from the corpus (7,570 tokens, 1,395 types), manually annotated for pronunciation.
- Excluded words including -lh-final stems because they participate in additional processes (e.g. /ilh-ta/ [il-t^{h}a] 'to lose-DECL') (Kim-Renaud, 1974; Sohn, 1999)

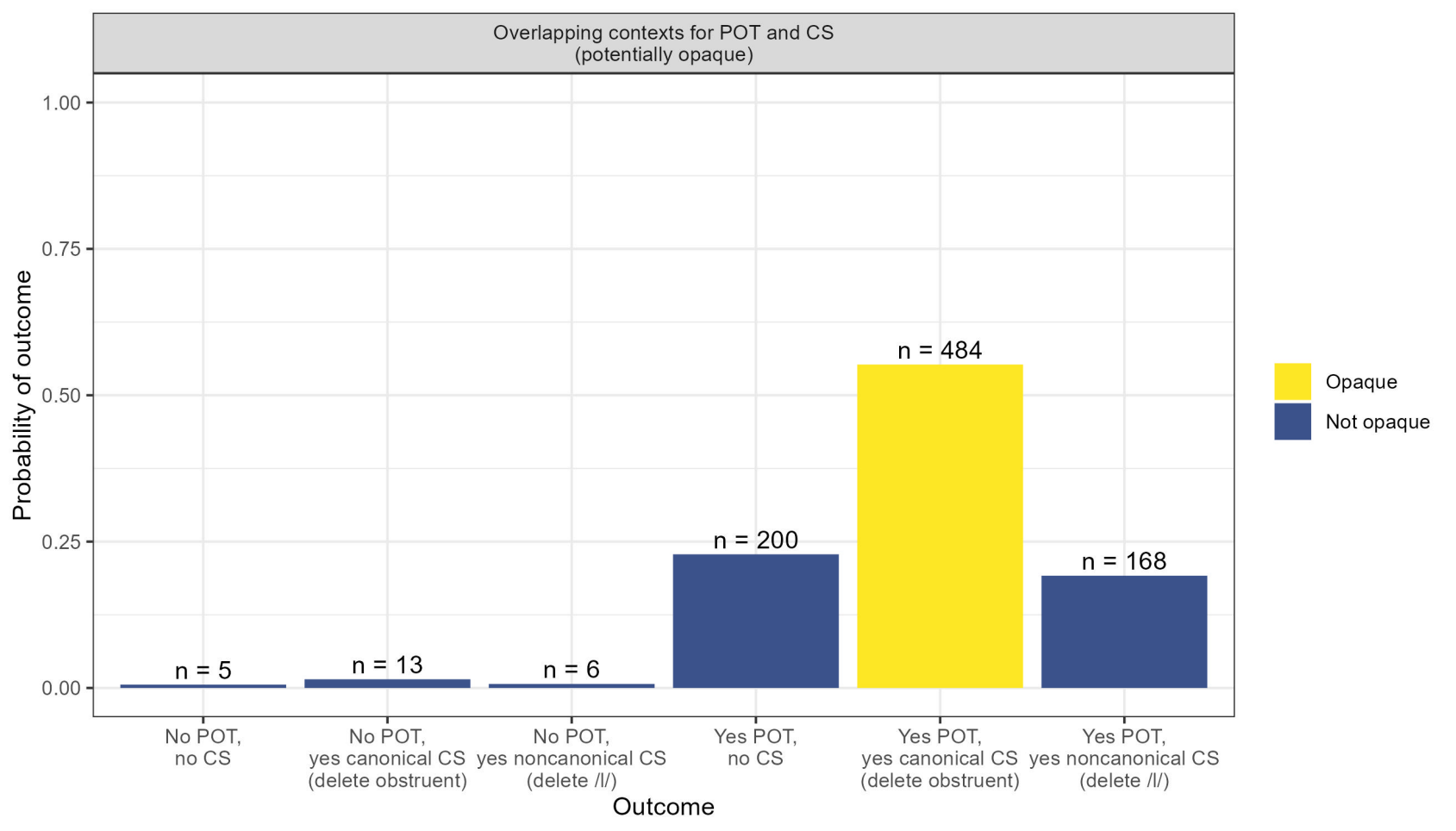
Adult corpus study

Three relevant contexts:

- LC-Obstruent - 1,045 tokens (236 types): /ilk-ta/ ‘to read-DECL’
- LC-Sonorant - 99 tokens (34 types): /halt^{h}-nɨn/ ‘to lick-COMP’
- LC-Vowel - 6,462 tokens (1,125 types): /nʌlp-ɨn/ ‘to be wide-COMP’

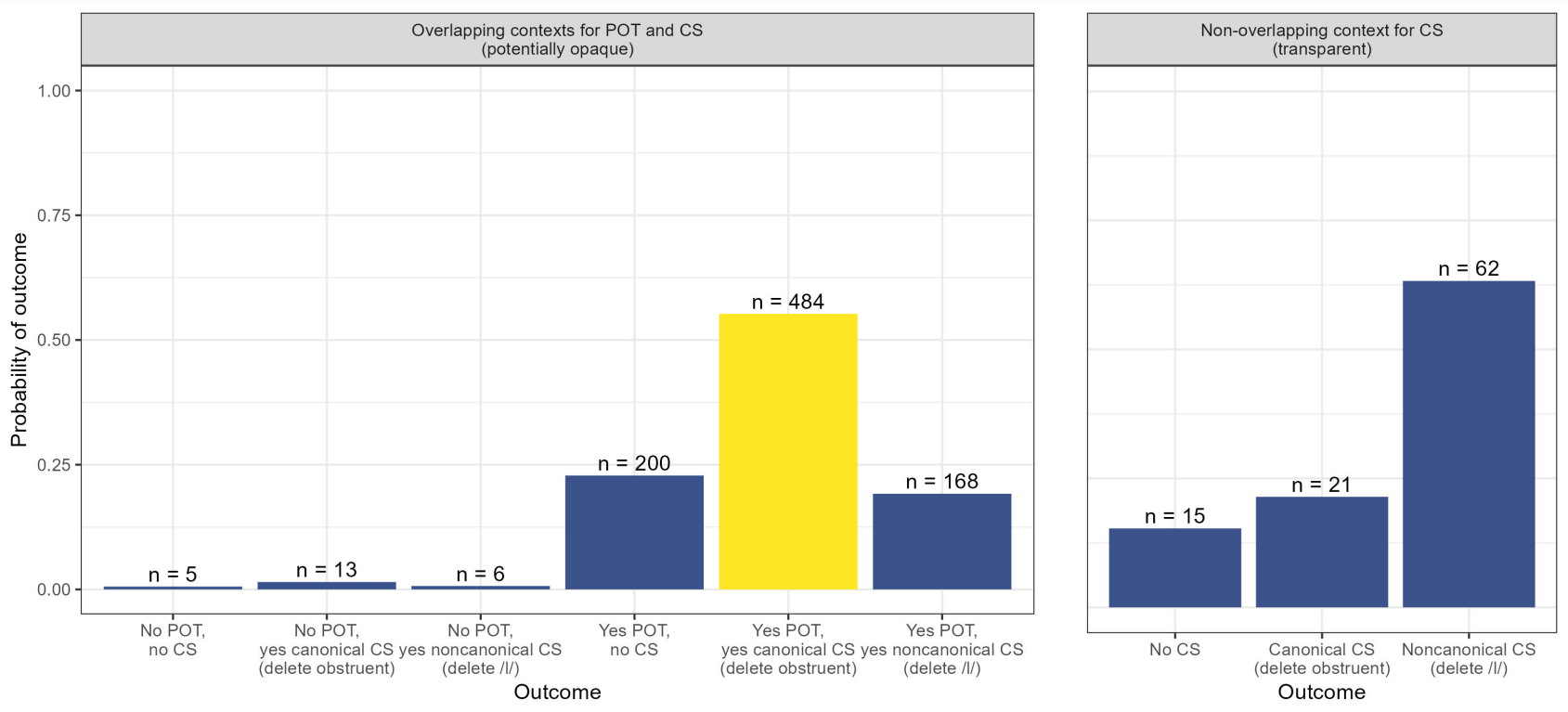
Did not annotate non-potentially-opaque POT cases due to frequency and near-obligatory POT application within the Accentual Phrase (Jun, 1998)

Adult corpus: Results



- ~53% of potentially-opaque items are actually opaque
- Most variability due to different targets of CS

Adult corpus: Results

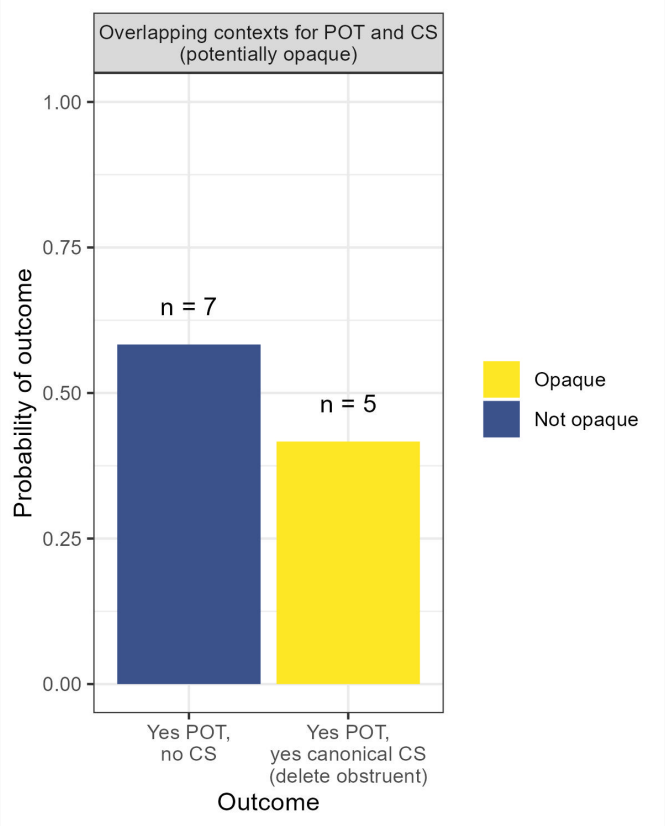


- Preference for opacity-creating version of CS in potentially-opaque overlapping contest(!)
- Negligible "over-application" of CS before vowel-initial affixes (not shown) → UR-restructuring unlikely.

Infant corpus study

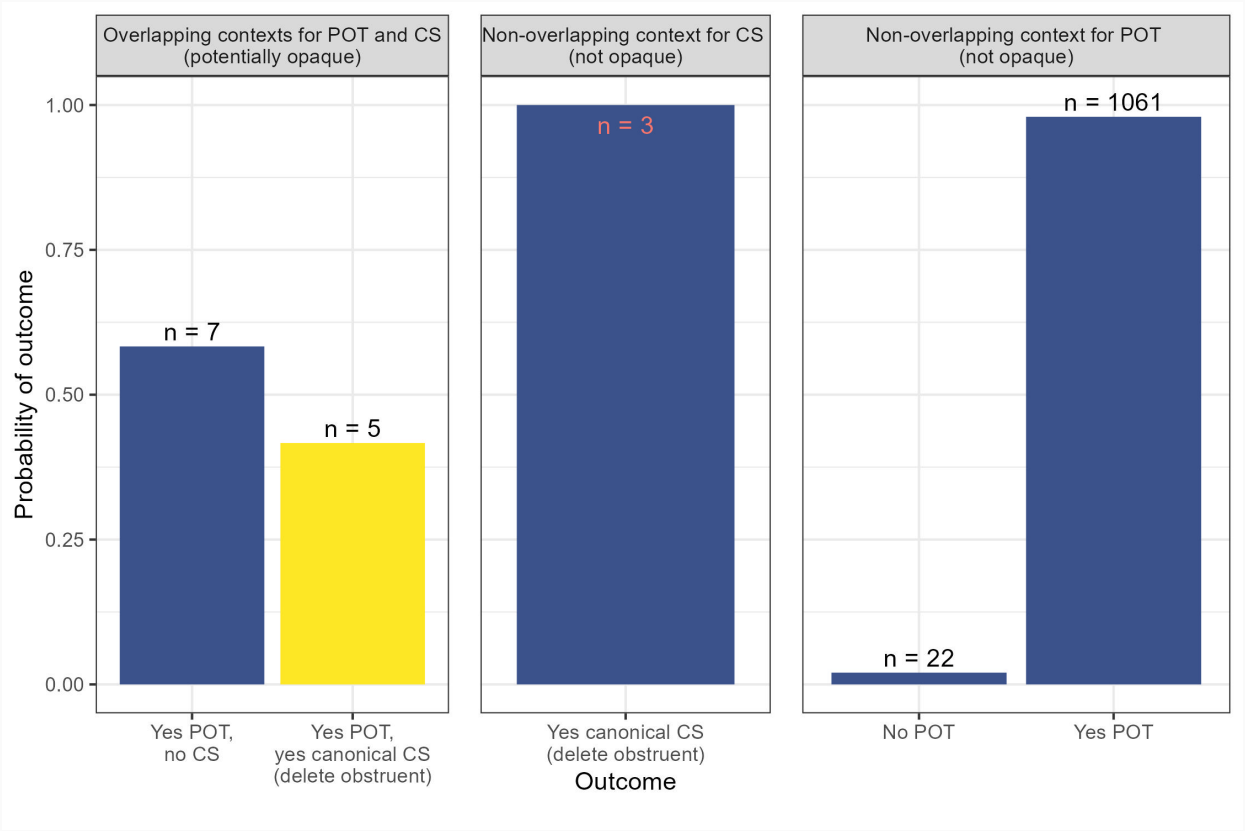
- Ko corpus (Ko et al., 2020); ~53,000 words of child-directed speech
- Spontaneous speech with phonemic transcription
- Extracted and hand-checked all affixed -lC verbs from the corpus (289 tokens, 38 types)
- Also extracted and hand-checked all -C.C- verbs from the corpus (1,083 tokens, 171 types)
- Similar exclusions as before.

Corpus study: infant-directed speech



- ~41% of potentially-opaque items are actually opaque.
- Only “canonical” CS
- So few cases!!

Corpus study: infant-directed speech



- POT and CS play by the rules
- Learning data is sparse!

Interim summary

- Textbook descriptions get the plurality of the data right for adults, not for kids → variation between opaque and non-opaque outcomes.
- When something gives, it's CS (POT always applies)
- Opacity-creating CS preferred in potentially-opaque cases!

How representative is the corpus data of the generalizations in the speakers' heads? → a *wug*-test!

Experiment

Experiment: Stimuli

Two verb shapes (-lC, -C)

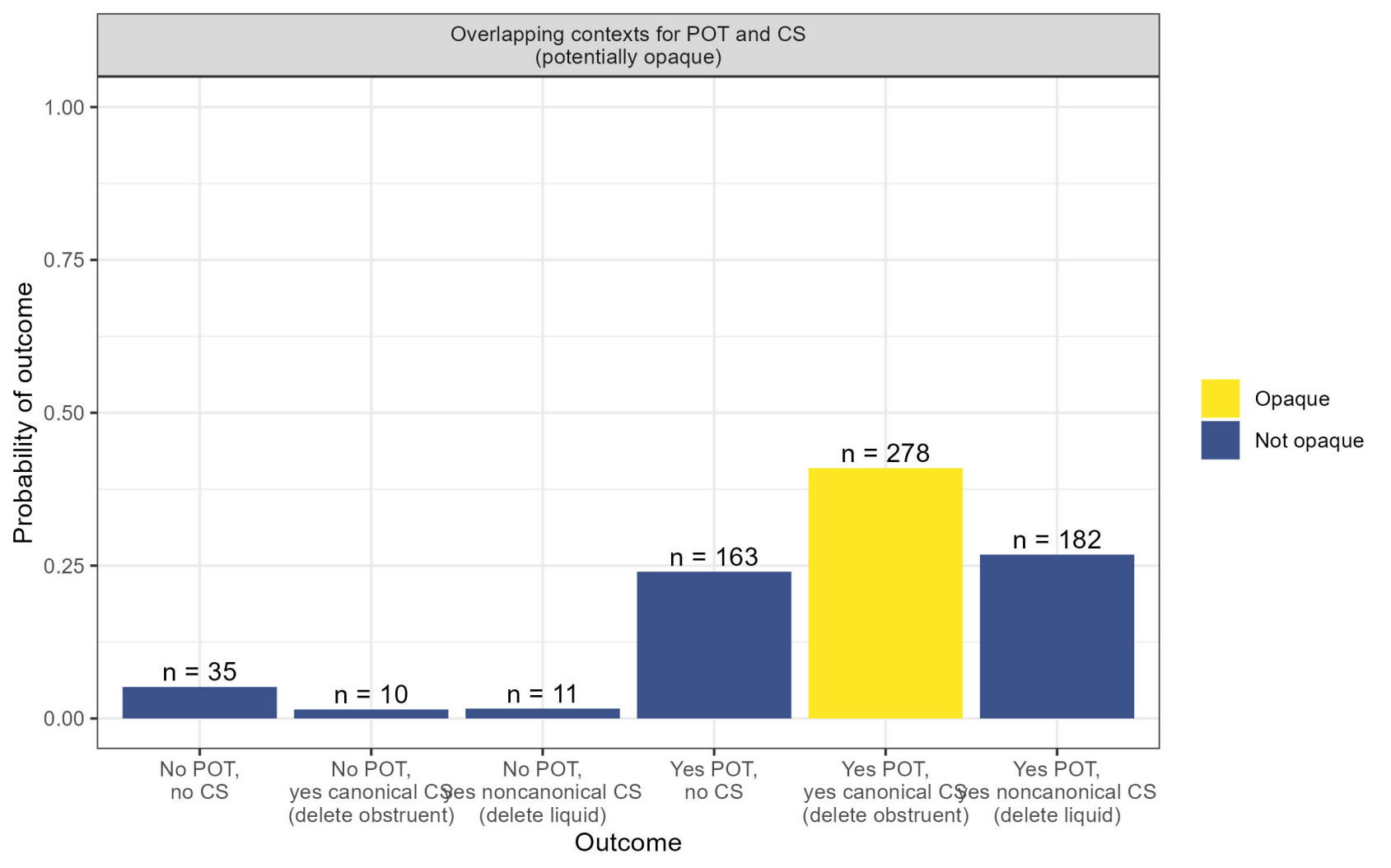
- × **three affixes** (-a/ʌ, -na, -ta)
- × **three frequency levels** (high-freq, low-freq, nonce)
- × **10 words** in each bin
- → **180 stimuli** per person.

Targets “baseline” non-overlapping contexts, as well as critical potentially-opaque environment.

Experiment: Procedure

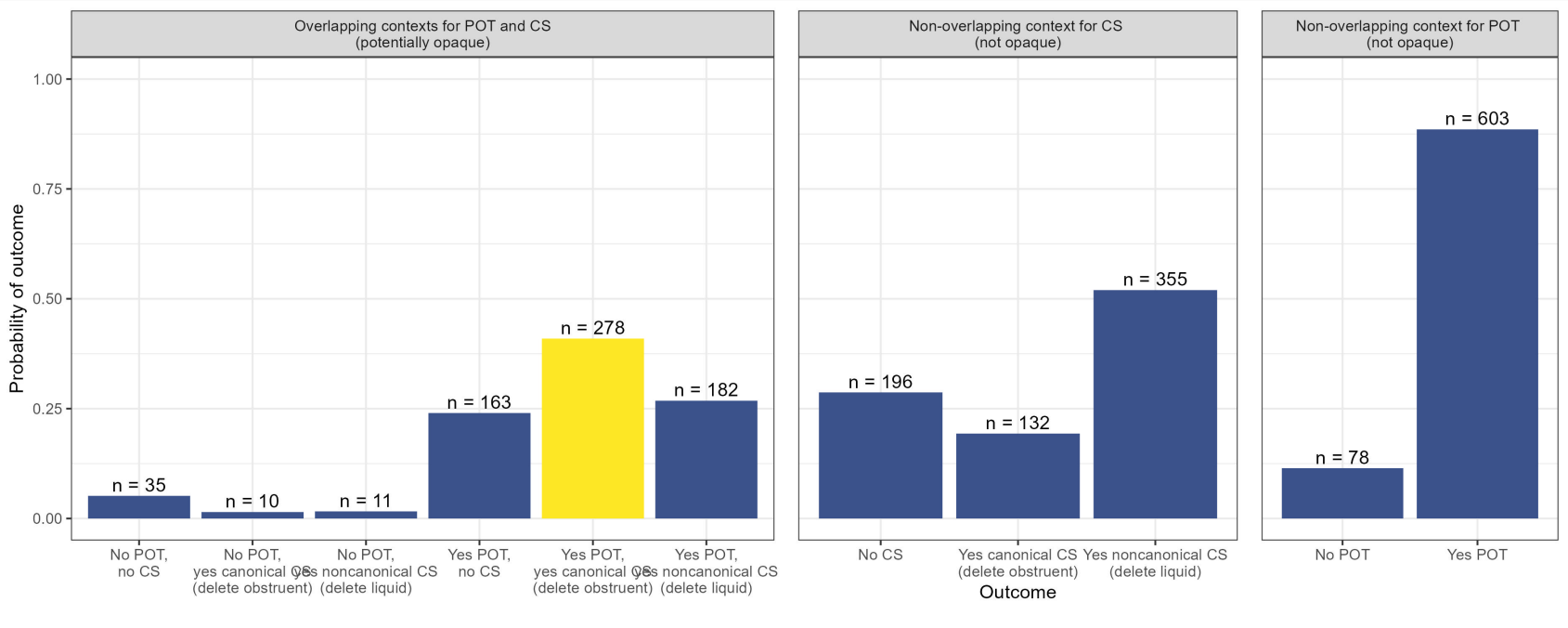
- Each participant saw the stem in a semantically-neutral phrase (주어진 단어는 XXX입니다. 'The given word is XXX.') with a V-initial affix not used in the experiment (-ajo/ʌjo)
- Instructed to read a sentence (녹음할 단어는 XXX입니다. 'The word to be recorded is XXX.') containing the stem suffixed with one of the three affixes (the "target word") out loud; speech recorded.
- Once completed, "self-transcribed" their response
- Then asked for word-familiarity, language-background
- We re-categorized existing words that were not known as nonce words for each participant.
- Recruited online and word-of-mouth; 23 subjects.

Experiment: Results



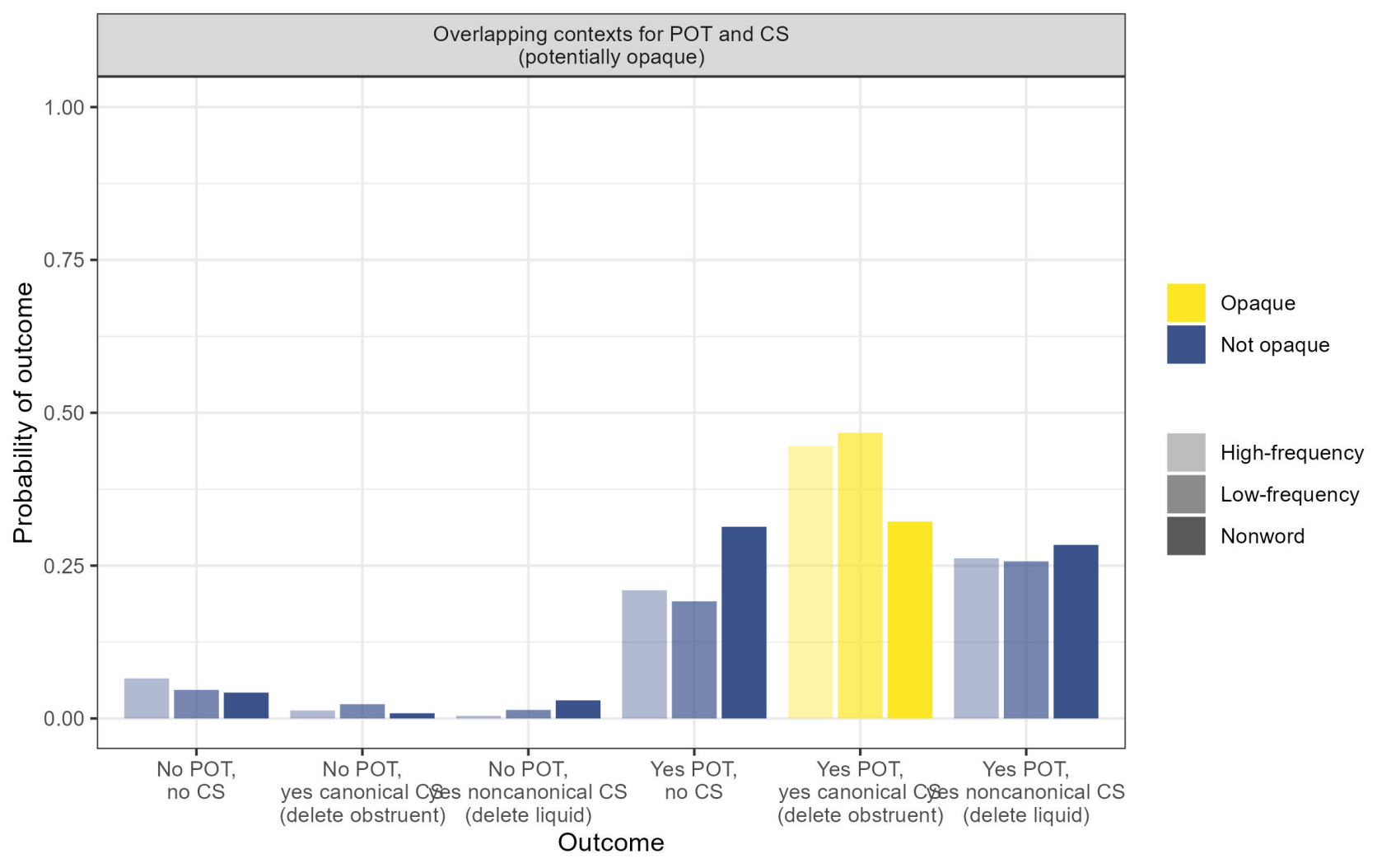
- ~40% of potentially-opaque items are actually opaque
- Most variability due to non-opaque application of CS

Experiment: Results



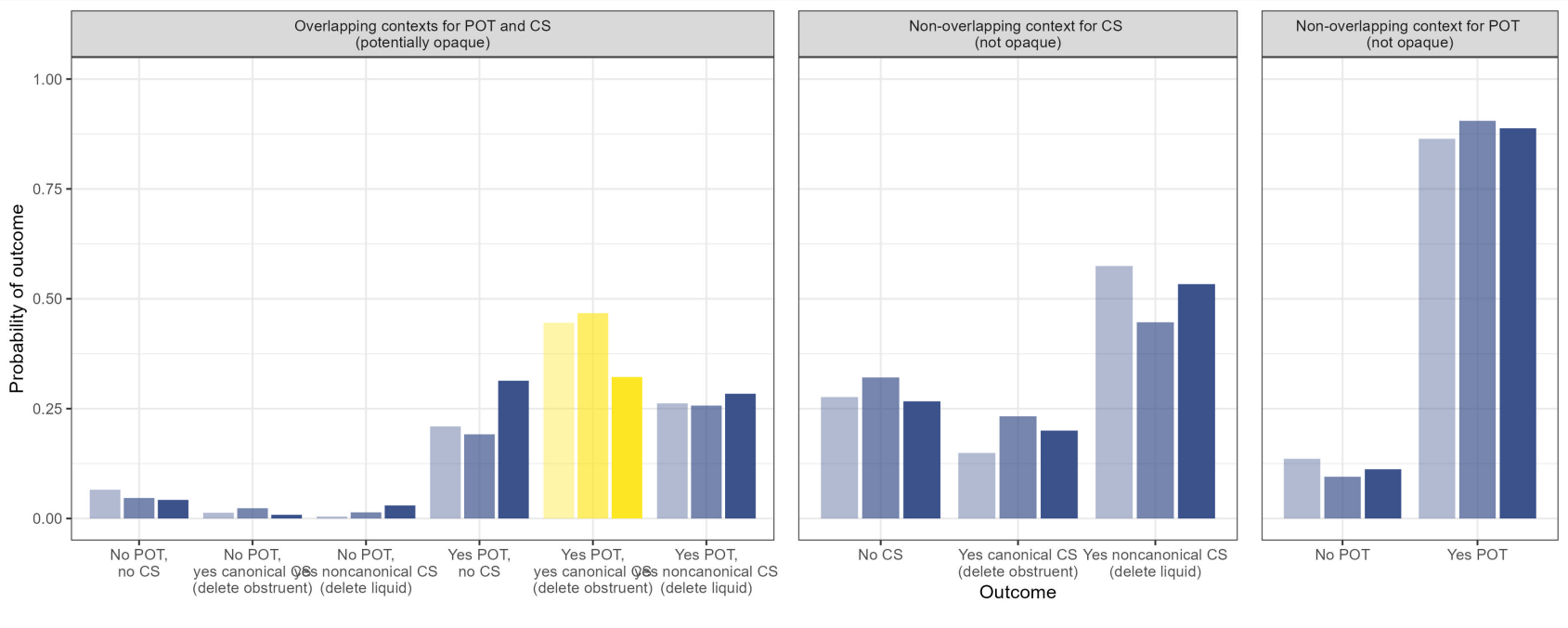
- In non-overlapping cases, similar story to corpus
- No evidence of UR-restructuring (not shown)

Experiment: Results



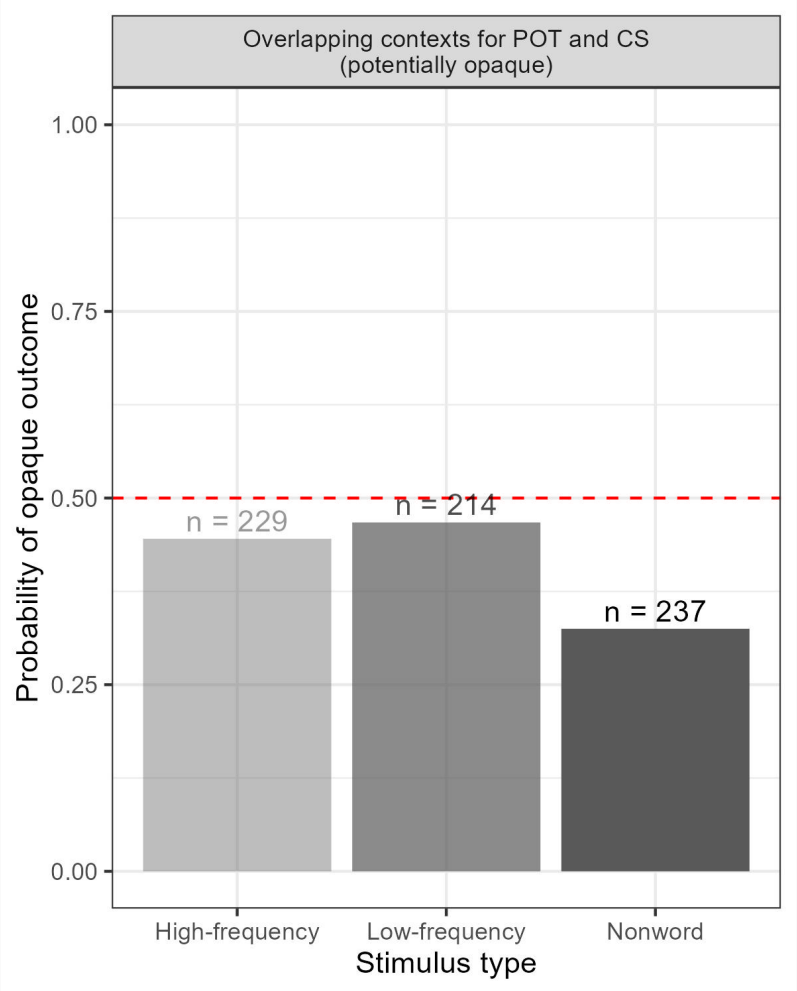
- Frequency-mediated application of opacity-creating CS, but not liquid-targeting CS

Experiment: Results



- In non-opaque contexts, less clear frequency-mediated story.

Experiment: Results



- The proportion of overlap cases that are opaque (that is, both obstruent-targeting CS and POT applying) is appreciably lower in novel words.

Interim summary

- Broad correspondence between corpus and experimental data
- Interaction between stem frequency and the probability of an opaque outcome $\rightarrow$ high-frequency words are more opaque, lower-frequency ones are more transparent.

Timeline and logistics

Material online

Training data:

- Adult corpus counts
- Infant corpus counts
- Dictionary list of all -lC verbs with their web-derived stem frequencies
- 15 participants' experimental data (test), 3 participants' experimental data (validation)
- Demographic info on all 23 speakers

Test data:

- 5 unseen participants' experimental data

Test metric $\rightarrow$ higher likelihood of test data given their model.

Timeline

Currently timeline:

- March 1, 2023: Data is released to participants
- March 14, 2023: Baseline systems released to participants
- April 14, 2023: Test data is available for participants
- April 21, 2023: Final Submissions are due.
- April 25, 2023: Results announced to participants
- May 8, 2023: System papers due for review
- May 22, 2023: Reviews back to participants
- May 31, 2023: CR deadline; task paper due from organizers.

Baselines

- Non-neural (meant to be easy to beat) $\rightarrow$ predict average of tokens in corpora, mean of all corpus for novel words // tokens not in corpus
- Neural (meant to be harder) $\rightarrow$ character-level transformer network

Thanks for participating!

Thanks also to:

Funding: MIT Computational Psycholinguistics Lab

Feedback: Roger Levy, Ryan Cotterell, Kat Vylomova

Data tagging: Sehyun Kim, Jaee Shin

Contact:

canaanbreiss1@gmail.com // cbreiss.com

References

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

Jun, S.-A. (1998). The accentual phrase in the korean prosodic hierarchy. *Phonology*, 15(2):189–226.

Kim, S. (2003). *Phyocwun Palum Silthay Cosa II* [*A Survey of Standard Pronunciation II*]. National Institute of Korean Language, Seoul.

Kim-Renaud, Y.-K. (1974). *Korean Consonantal Phonology*. PhD thesis, University of Hawaii.

Ko, E.-S., Jo, J., On, K.-W., and Zhang, B.-T. (2020). Introducing the ko corpus of korean mother–child interaction. *Frontiers in Psychology*, 11:3698.

References ii

Mayer, C. (2021). *Issues in Uyghur backness harmony: Corpus, experimental, and computational studies*. University of California, Los Angeles.

National Institute of Korean Language (2021). NIKL Korean Dialogue Corpus (audio) 2020(v.1.3).

Sanders, R. N. (2003). *Opacity and sound change in the Polish lexicon*. University of California, Santa Cruz.

Sohn, H.-M. (1999). *The Korean Language*. Cambridge University Press, Cambridge, UK.

White, J. C. (2013). *Bias in phonological learning: Evidence from saltation*. PhD thesis, UCLA.