A Database of Native Mandarin Speakers' Lexical and Sublexical Frequency Judgments

Seth Wiener

Carnegie Mellon University

Language processing is intimately tuned to input frequency

- L1 listeners use statistics of language to improve perception, production and overcome variability
  (Kleinschmidt & Jaeger, 2015; Cutler, 2012)

- L2 learners have less experience with the language but can track statistics
  (Wanrooij et al., 2013; Ellis, 2002)

Exposure to L2 statistical information may help bootstrap early acquisition.

1. Sublexical frequency affects word recognition
   (McMurray et al., 2008, 2009; Wiener & Ito, 2015)

2. Lexical frequency scaffolds sentence processing
   (Bybee & Hopper, 2001)

GOAL: to provide a reliable source of Mandarin lexical and sublexical frequency information for L2 pedagogy and experimental stimuli design

Existing resources may not reflect native speakers' usage

- Frequency dictionaries:
  《现代汉语频率词典》(1986) 《现代汉语常用词词频词典》(1990)
- Corpora:
  Language Corpus System of Modern Chinese Study
  Academia Sinica Balanced Corpus of Modern Chinese

Limitations: outdated, no electronic version, not always free, not representative of spoken language, limited to characters only, not at the needed levels, may not reflect native intuition…

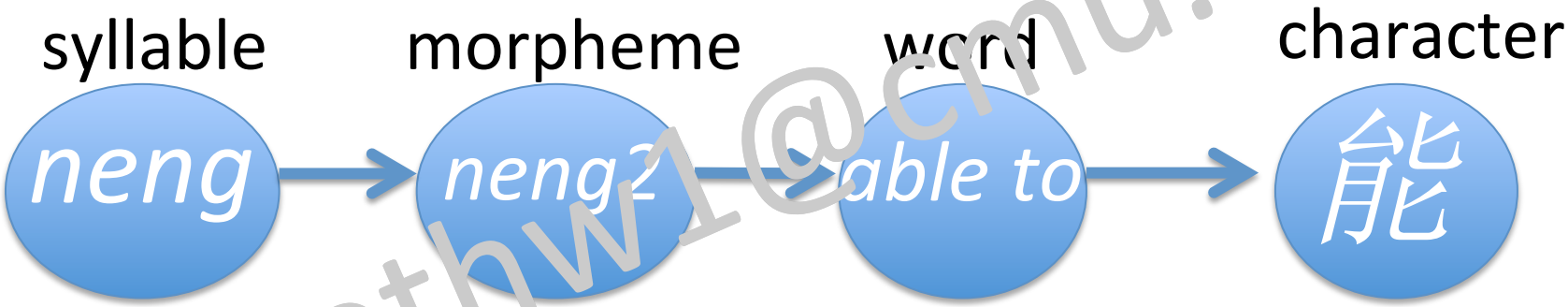
Theoretical framework

400

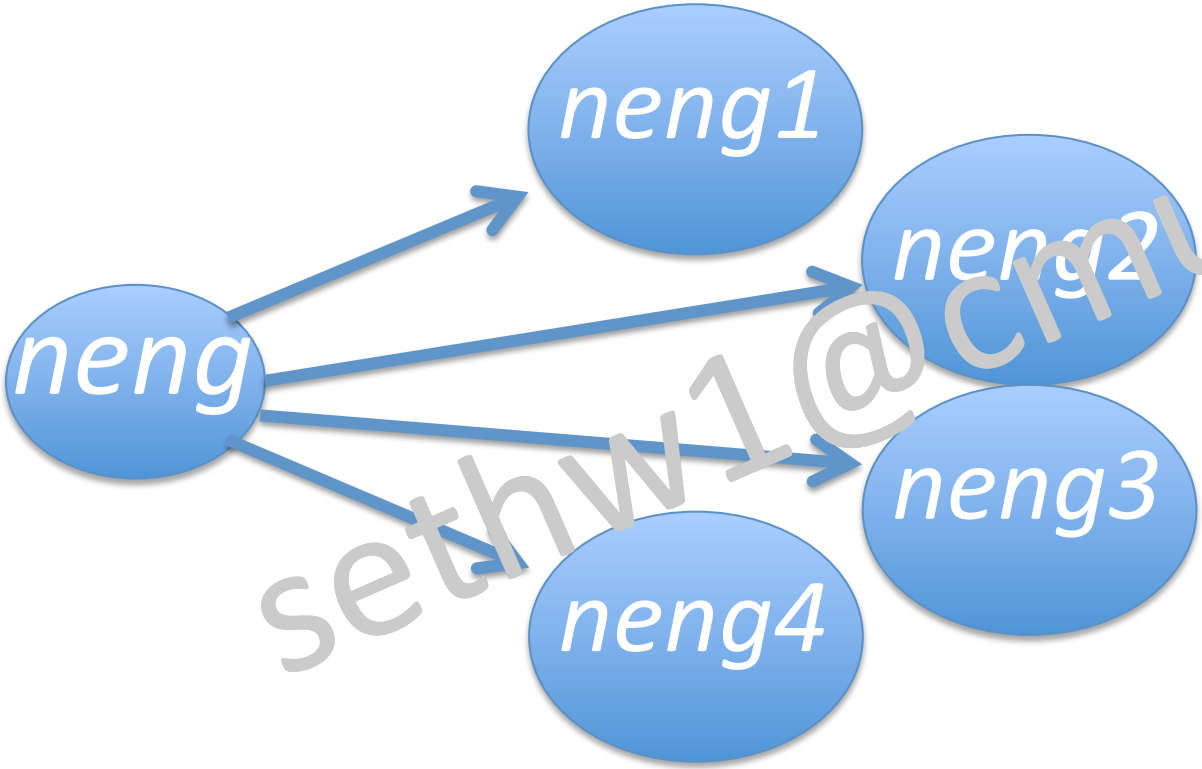
The Mandarin syllable serves as the critical or proximate unit in perception and production

(Chen et al., 2002, 2003; Sereno & Lee, 2014; Verdonschot et al., 2003)

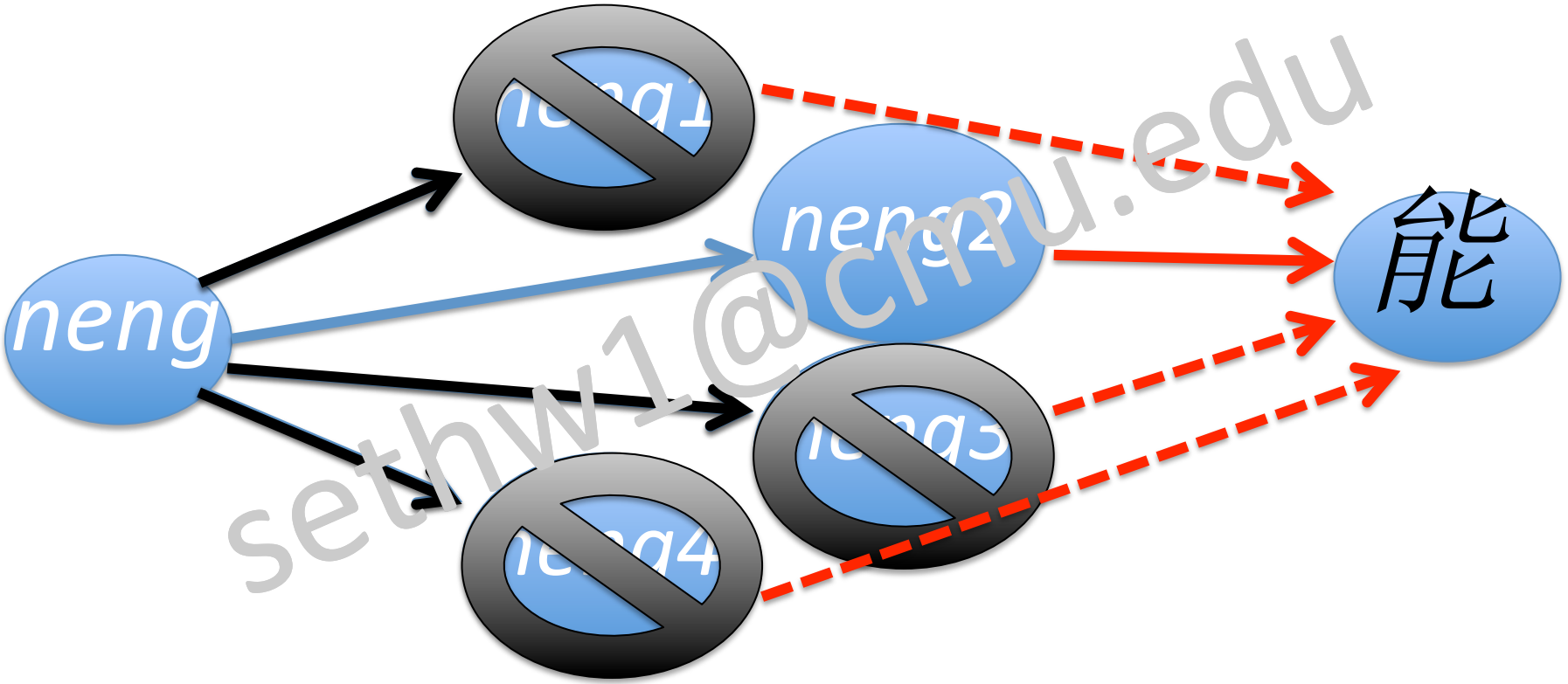
Consistent 1:1:1:1 mapping

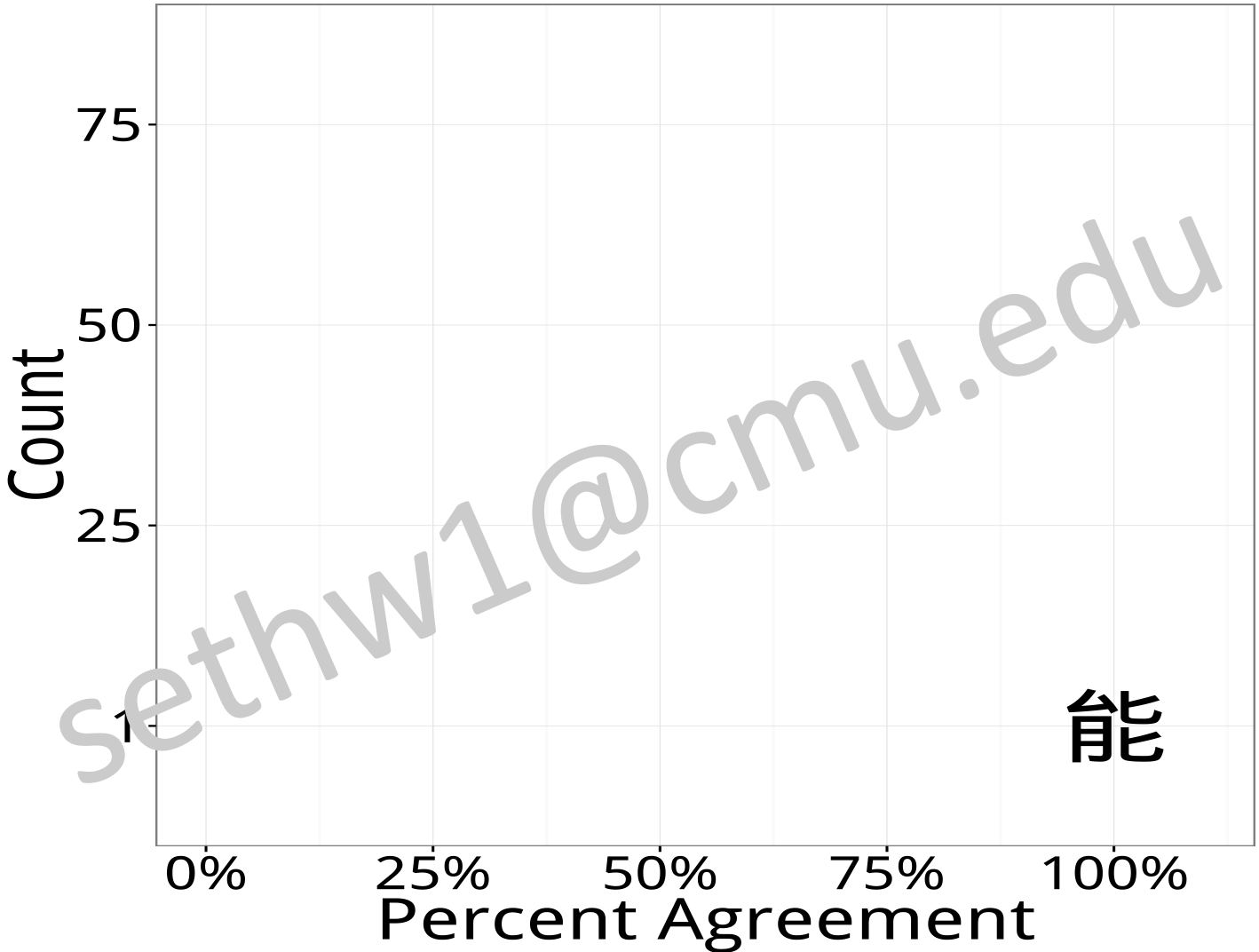


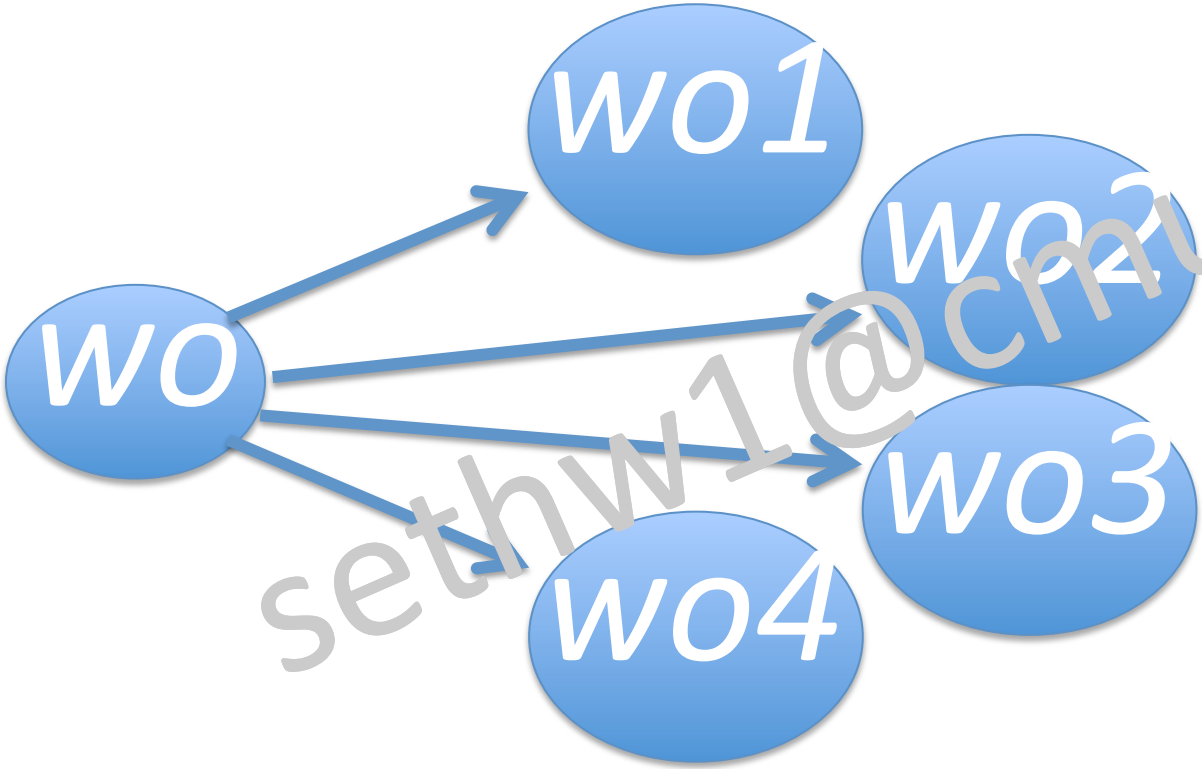
Syllable-conditioned tonal probabilities



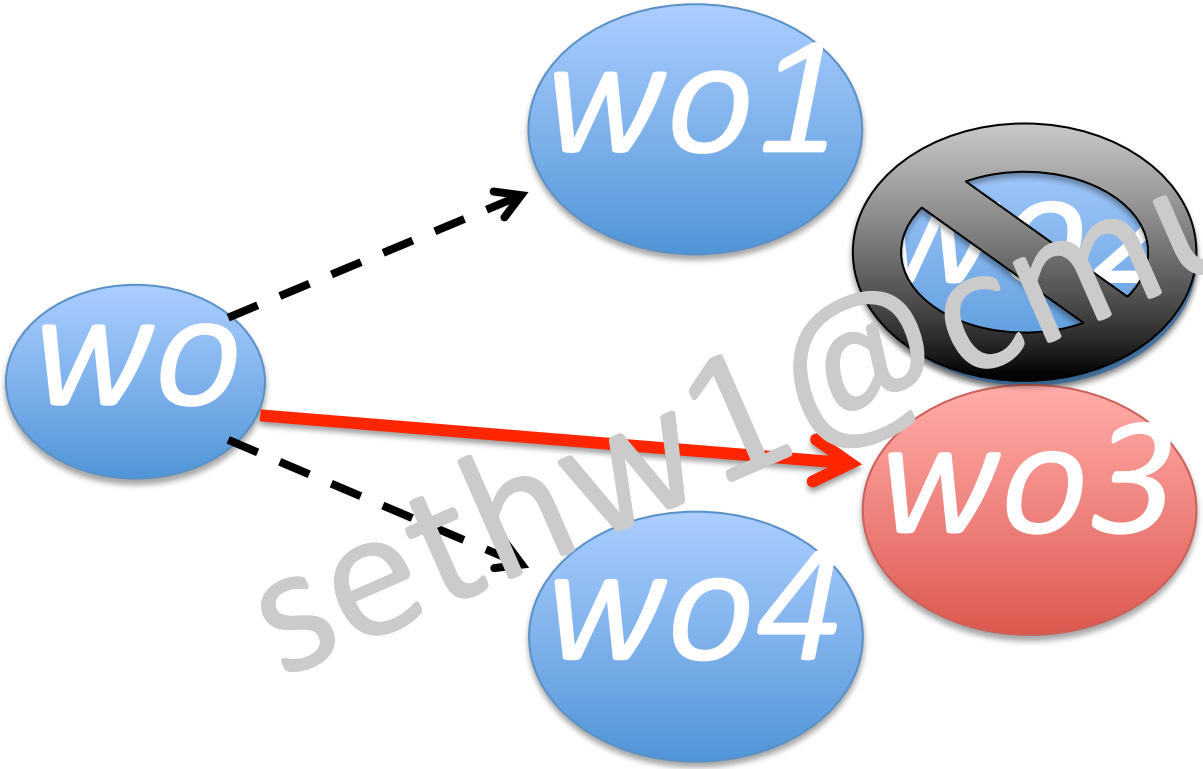
Tone won't affect recognition



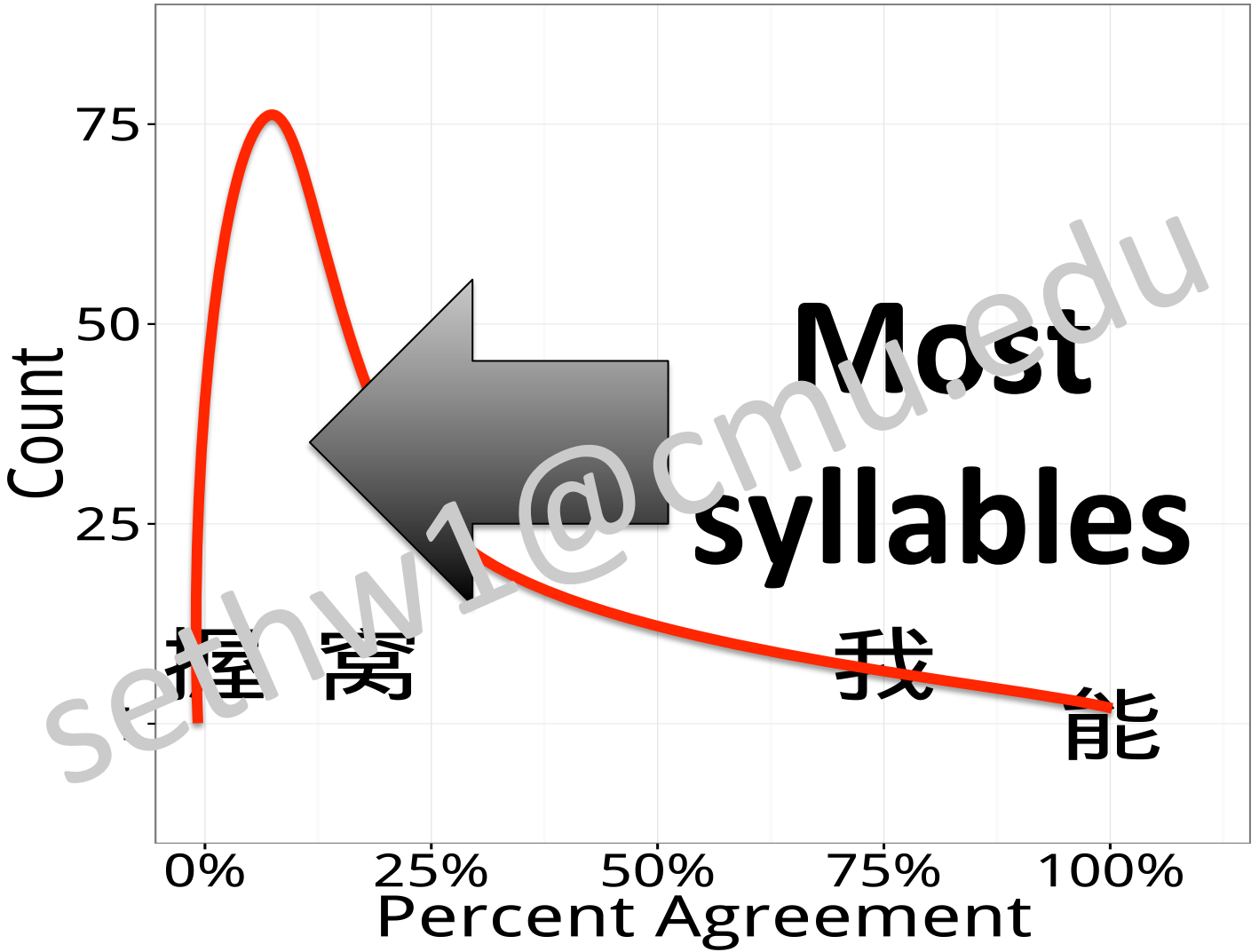




A range of probabilities exists



wo3 (most)
wo1
wo4 (least)
wo2 (non-existent)



Project Goals

① Collect native speaker judgments and intuitions at lexical and sublexical levels

② Confirm that methodology is valid

③ Compare results to published corpora and L2 textbooks

④ Use results to inform L2 pedagogy and experimental stimuli design

Crowdsourced frequency database

- Secure online survey
- Accessible with any web browser
- Monitoring tools for time on task

- All participants finished university in China
- Spoke Mandarin as L1 (dominant language)

Data from 227 native speakers

Syllable-based lexical frequencies

- Presented with individual randomized syllable in pinyin form
- Asked to type first word (using characters) that comes to mind given that syllable (ignoring tone)

wo  我

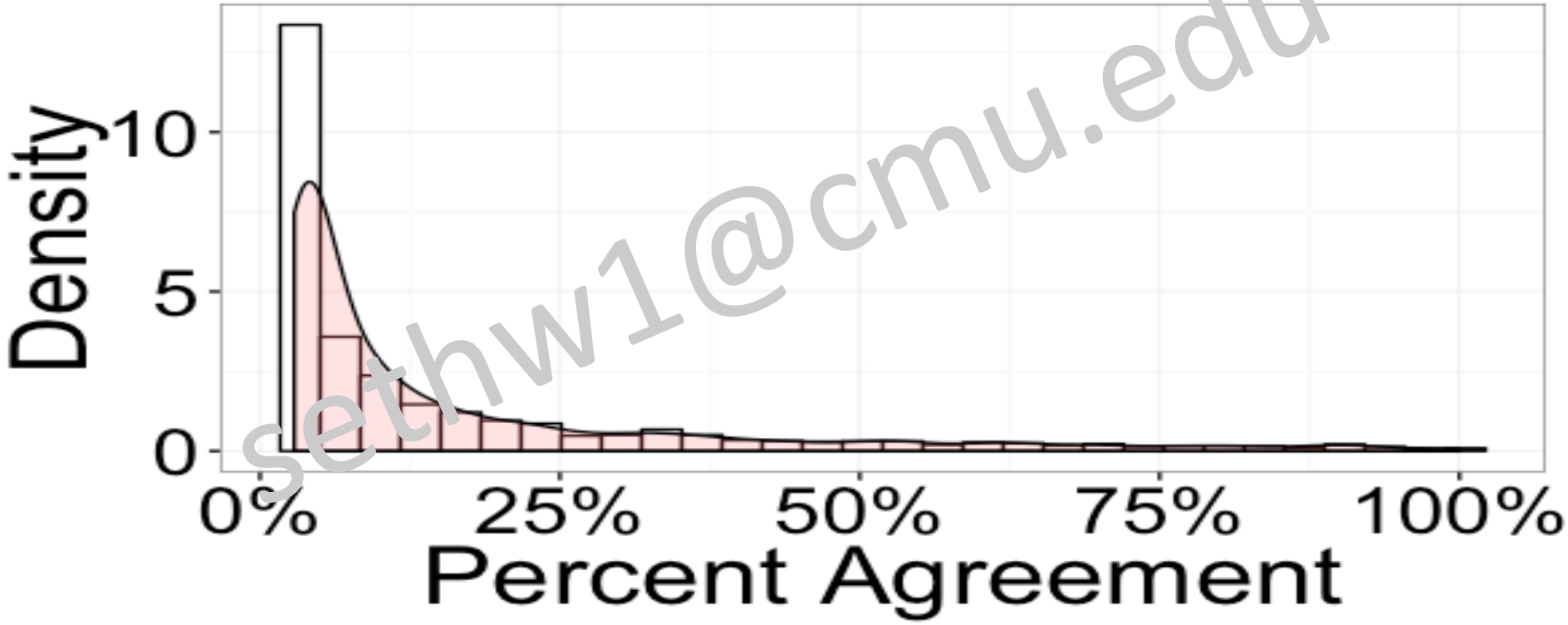
Sublexical syllable+tone frequencies

- Presented with (different) randomized syllable and its five tones (included neutral tone)
- Asked to rank order of likelihood in speech from most likely to least likely or non-existent

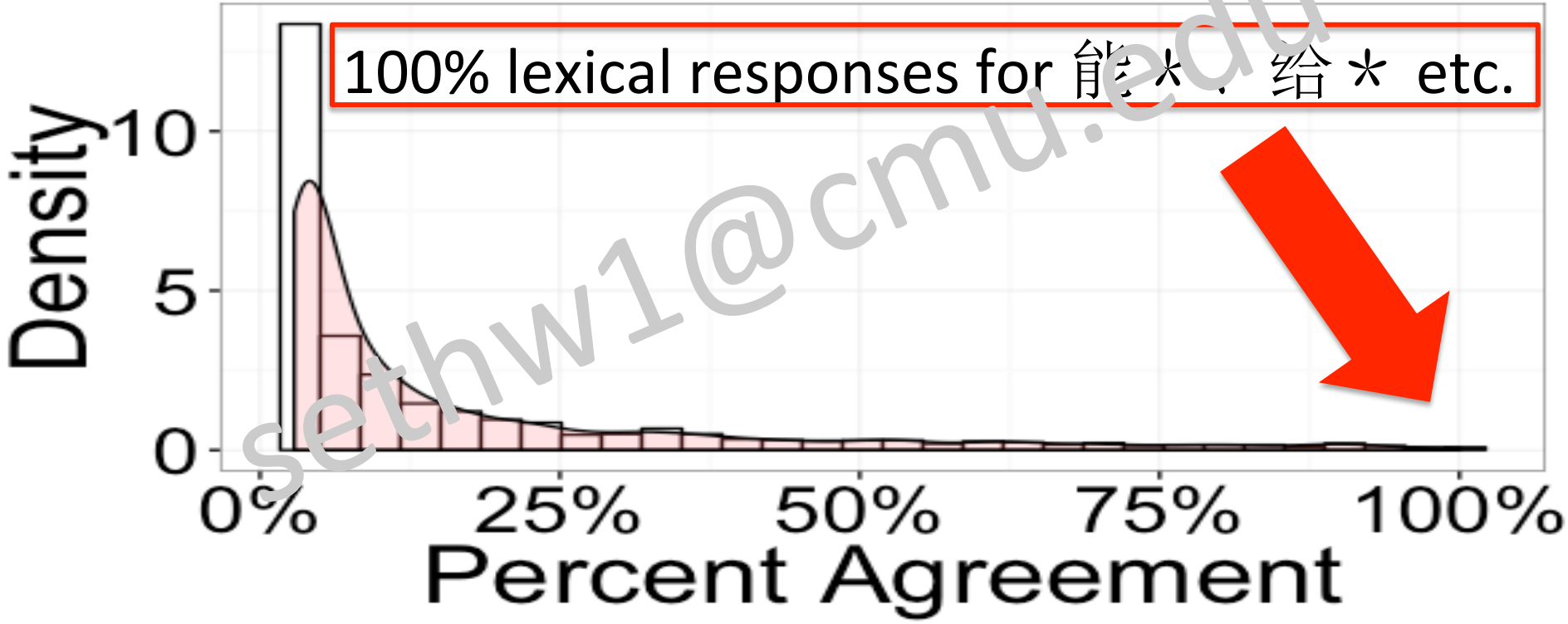
wo1, wo2, wo3, wo4, wo5

wo3
wo1
wo4
wo5
wo2

Crowdsourced results match theoretical predictions



Control syllables confirm methodological validity

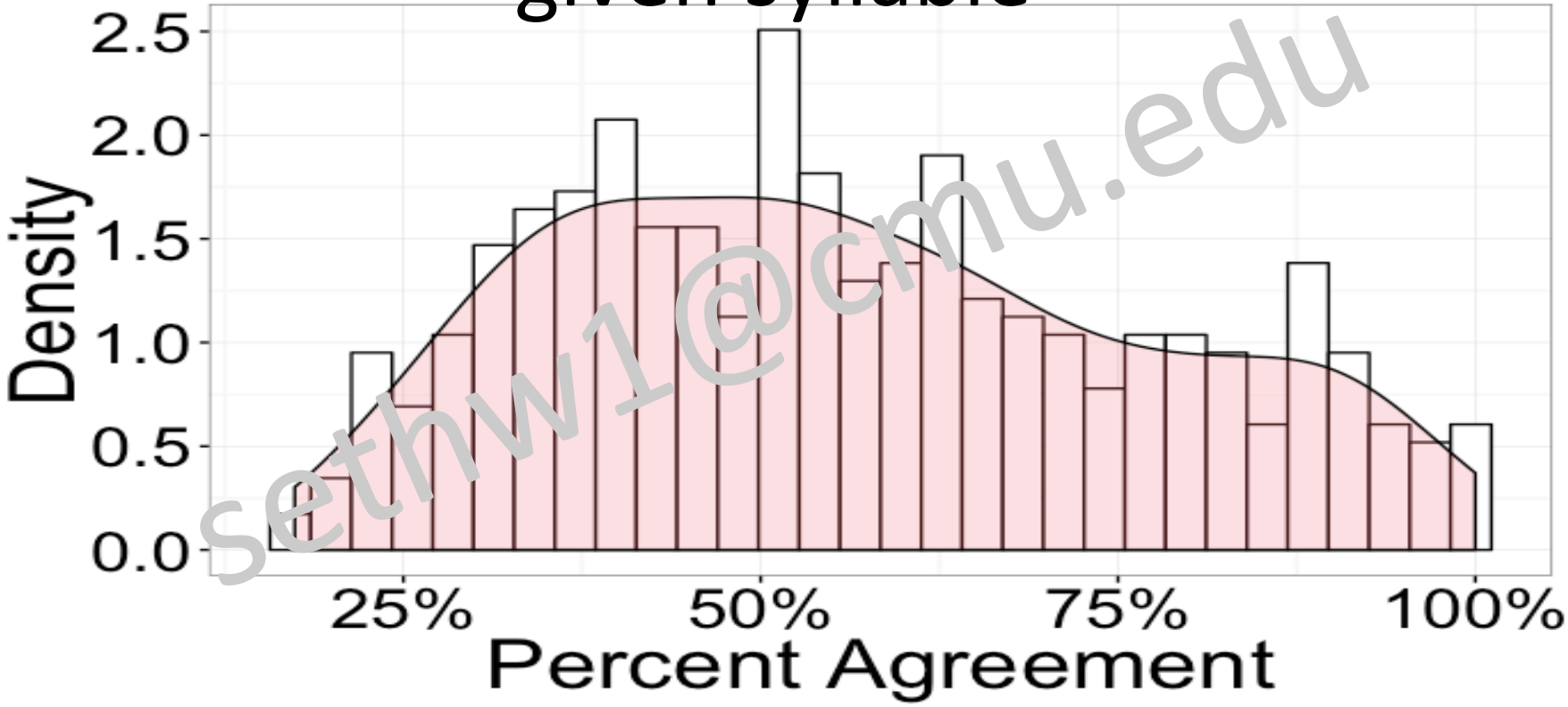


Lack of consensus indicates areas of interest

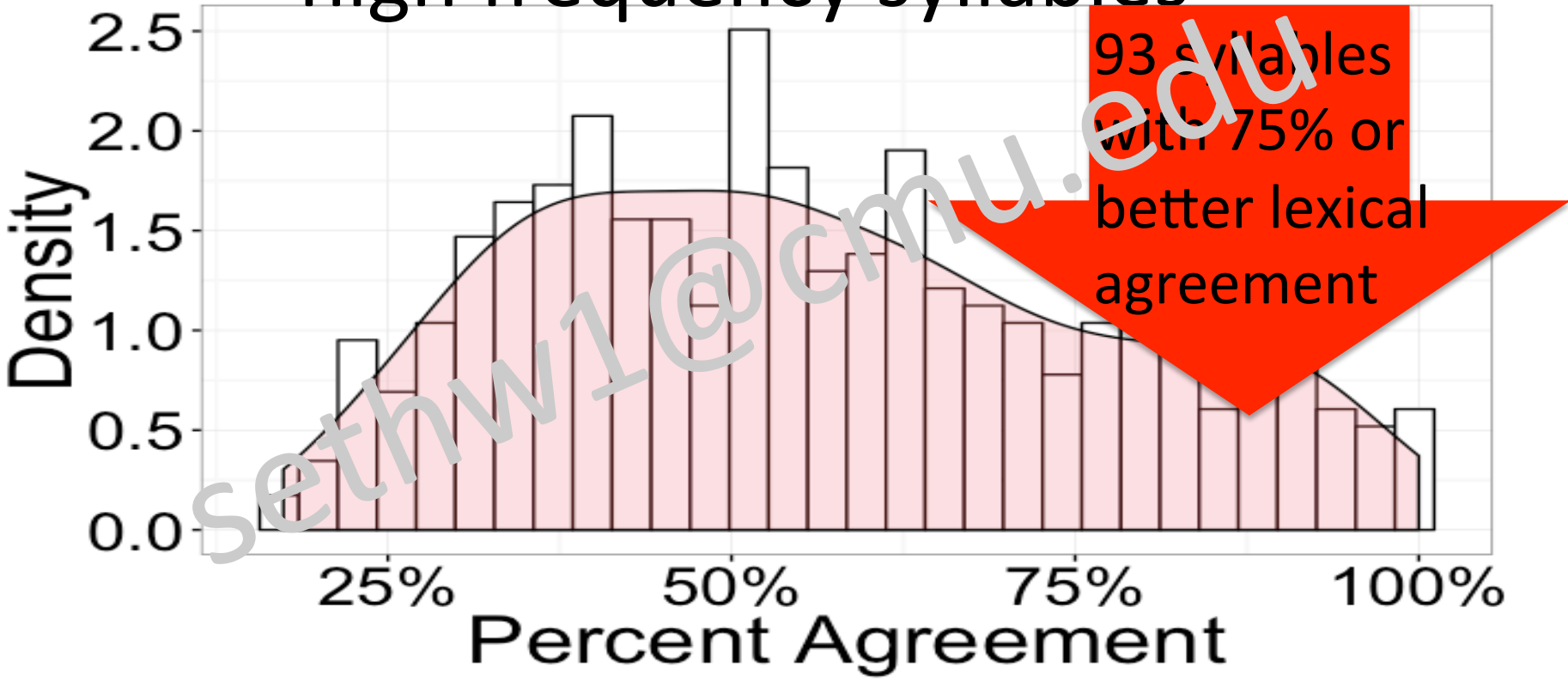
55% response rate;
ma1 most probable

Will drawing learners' attention to statistical distribution aid acquisition?

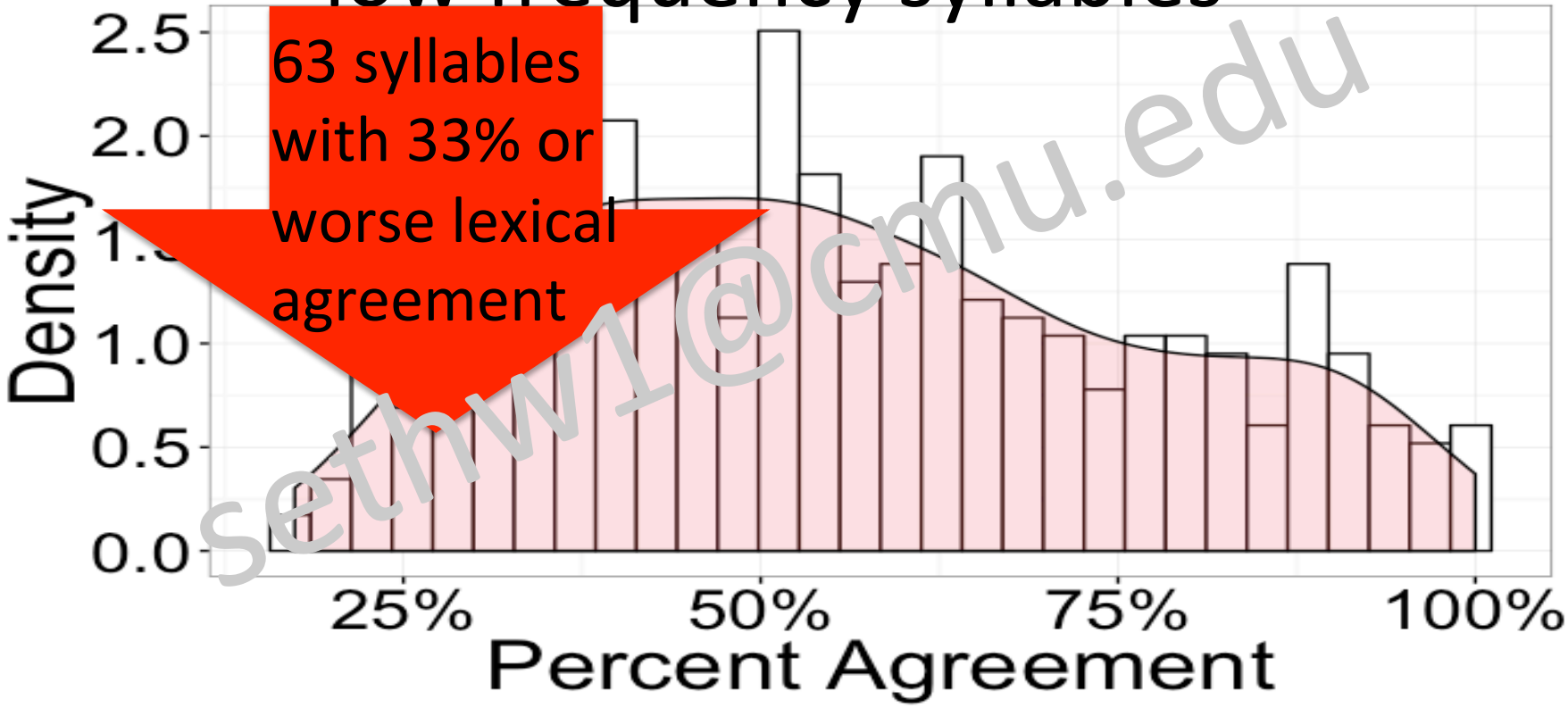
Most probable lexical response given syllable

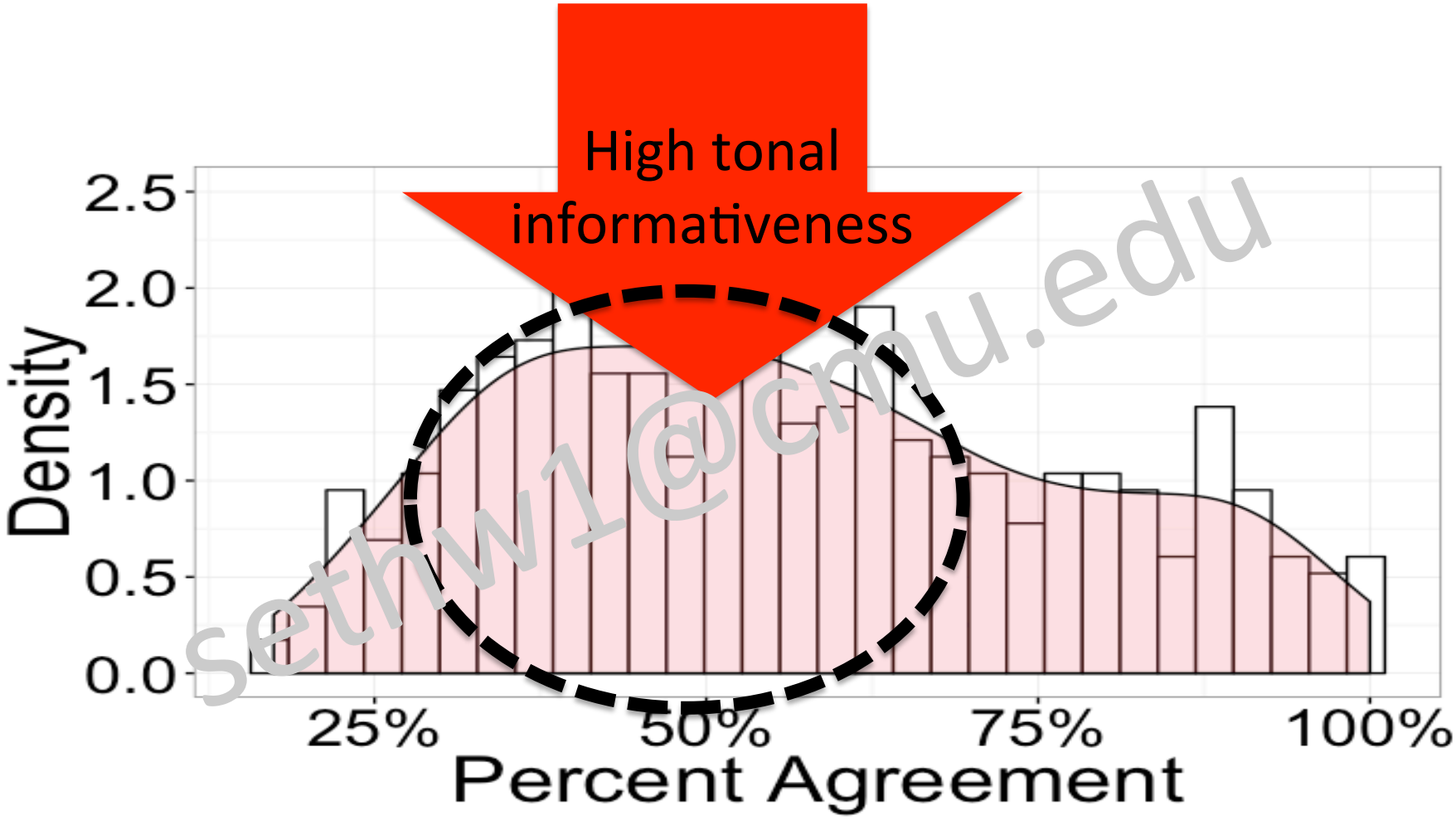


High agreement for high frequency syllables



Low agreement for low frequency syllables





Frequent and homophonous syllables pose problems to learners

What role does explicit awareness of homophony play in acquisition?

鱼、予、余、俞

雨、与、语、宇

预、遇

Next steps

① Continue to collect responses and expand to multisyllabic (and lexical bundle) judgments

② Test the role of explicit awareness as perception and pronunciation heuristic in classroom

③ Refine problem areas for L2 leaners

Contact information

- For more information and demo version contact:

sethw1@cmu.edu