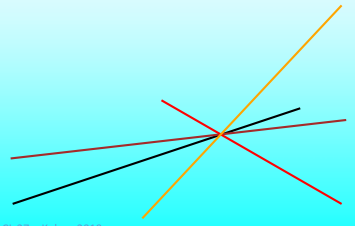
# Variation in syllable decomposition across Sinitic languages



James Myers
National Chung Cheng University, Taiwan
Tsung-Ying Chen
National Tsing Hua University, Taiwan



## Thanks!


- Grants from the Ministry of Education of the Republic of China (MOST 103-2410-H-194-119-MY3, MOST 106-2410-H-194-055-MY3)
- Jane Tsay (蔡素娟)
- Lab CEOs (蘇思綺、羅佳雯) and other assistants (林玟綺、張佑竹、陳奎燁、陳姵珊、劉美君、潘曉音、顏郁珊、蘇思綺)
- And of course the experimental participants




## Overview

- Syllable decomposition
- Language experience and language processing
- Three analyses in four Sinitic languages

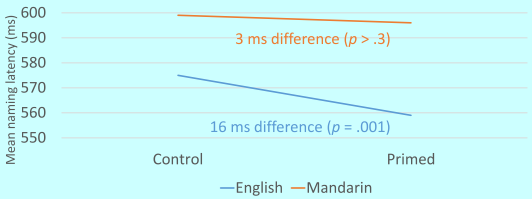


## Syllable decomposition

- English speakers seem to process (produce or recognize) the spoken word **cat** like this:
  - First /k/ … /æ/ … /t/
  - Then /kæt/
- Mandarin speakers seem to process (produce or recognize) the spoken word **貓** like this:
  - First /mɑu/
  - Then (optionally) /m/ … /ɑu/ (or even /m/ … /ɑ/ … /u/)
- These claims are experimentally testable



## Mandarin vs. English onset priming



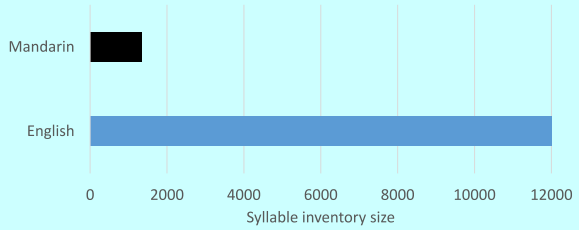

O'Seaghdha *et al.* (2010): form preparation priming in word production

Mandarin: Exp 3, monosyllabic targets, written prompts; English: Exp 6: monosyllabic targets, picture prompts

(But see critical discussion in Yu *et al.*, 2015, and Myers, 2016)



## Mandarin vs. English syllable inventories




Mandarin: Tsai (2000); English: Levelt *et al.* (1999)



## Hypothesis

- Smaller syllable inventory = Process syllables as wholes
  - Syllables are the most natural phonological unit…? (e.g., see review on child language in Treiman *et al.*, 1996)
- Larger syllable inventory = Decompose into phonemes
  - Otherwise there are too many syllables to access from memory…?

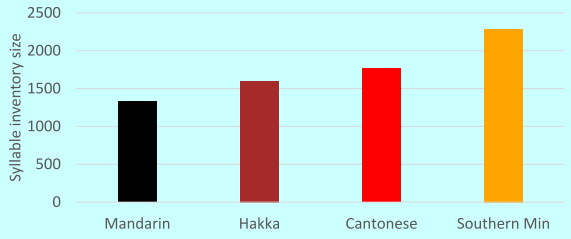


## How many syllables matter?

- Mandarin and English differ enormously in syllable inventory size
- What about within the Sinitic language family?
- We already happened to have some experimental data on
  - Mandarin
  - Hakka (Sixian [四縣])
  - Cantonese
  - Southern Min
- All syllable inventories are quite small (under 2,500, including tone)
- None have any orthographic bias in favor of phonemes
  - Some (Mandarin, Cantonese) are written in a syllable-based orthography
  - Others (Hakka, Southern Min) are virtually never written at all



## Even Sinitic languages vary in syllables




Counts include tone. Mandarin: Tsai (2000); Hakka: MOE (2006); Cantonese: LSHK (2012); S. Min: Ruan *et al.* (2012)

## Collecting the data

https://Worldlikeness.org (Chen & Myers, 2017)





## Wordlikeness judgments

- Native speakers hear nonlexical syllables and decide if each is / isn't "like" the target language (e.g., 像不像國語 for [Taiwan] Mandarin)

| Language | Participants | Items | Notes |
|---|---|---|---|
| Mandarin | 21 | 81 | Recorded with/without S. Min accent |
| Hakka | 22 | 172 | |
| Cantonese | 20 | 153 | Some participants were tested remotely |
| Southern Min | 21 | 101 | Recorded with/without S. Min accent |

Syllables were randomly generated from each language's phonemes, obeying overall syllable structure



## Nonlexical Mandarin syllables

- 81 test items (plus 8 practice items)

ɕan$^{11}$ ɕan$^{55}$ ɕyŋ$^{51}$ ɕyŋ$^{55}$ ɕuai$^{35}$ ɕuaŋ$^{35}$ ɕyn$^{11}$ fiai$^{11}$ fuaŋ$^{51}$ fuo$^{55}$
fyɔŋ$^{55}$ kɻ$^{11}$ keɪ$^{35}$ khaɪ$^{35}$ khuɔŋ$^{11}$ khyæn$^{51}$ kiæn$^{11}$ kiæn$^{51}$ kiai$^{11}$
kiaŋ$^{51}$ kin$^{35}$ kio$^{55}$ kuai$^{35}$ lɣ$^{35}$ lia$^{35}$ luaŋ$^{55}$ lui$^{55}$ miæn$^{55}$ mu$^{55}$
muaɪ$^{53}$ muaŋ$^{55}$ muɔŋ$^{35}$ niu$^{55}$ nuai$^{51}$ nuan$^{55}$ phɣ$^{11}$ phɣ$^{35}$ phiaŋ$^{35}$
phuaŋ$^{35}$ phuɔŋ$^{11}$ phuɔŋ$^{35}$ phuɔŋ$^{51}$ phyɛ$^{55}$ piau$^{35}$ pio$^{35}$ pɣ$^{11}$ ʂɛ$^{55}$
siæn$^{11}$ ʂyæn$^{55}$ tɻ$^{35}$ tɕhɻ$^{35}$ tɕheɪ$^{11}$ tɕheɪ$^{35}$ tɕhua$^{55}$ tɕhuei$^{51}$
tɕhun$^{11}$ tɕo$^{51}$ tha$^{35}$ thioʊ$^{55}$ thuaŋ$^{55}$ tia$^{51}$ tiai$^{11}$ tiaŋ$^{11}$ tʂɐɪ$^{55}$ tʂən$^{35}$
tshɐɪ$^{11}$ tshia$^{51}$ tshuai$^{35}$ tʂhyn$^{55}$ tsiɔŋ$^{51}$ tʂyn$^{35}$ tuaɪ$^{53}$ tyn$^{51}$ xia$^{35}$
xiã$^{51}$ xio$^{35}$ xo$^{11}$ xyn$^{11}$ ʐɻ$^{55}$ ʐɐɪ$^{11}$ ʐin$^{35}$

Tone 3 pronounced as [11]



## Nonlexical Hakka syllables

- 172 test items (plus 4 practice items and 8 vowel-initial items)

bɲ$^{11}$ ɕan$^{24}$ ɕan$^{55}$ ɕiɑm$^{31}$ ɕiɑm$^{55}$ ɕim$^{31}$ ɕin$^{11}$ ɕioi$^{55}$ ɕoi$^{11}$ ɕon$^{24}$ ɕuaŋ$^{31}$ ɕun$^{55}$ faŋ$^{11}$
fem$^{11}$ fiam$^{31}$ fim$^{11}$ fim$^{55}$ fĩn$^{11}$ fin$^{31}$ fion$^{24}$ fion$^{55}$ fŋ$^{11}$ fuai$^{55}$ fuan$^{55}$ hĩ$^{11}$ hio$^{31}$
hio$^{55}$ hua$^{31}$ huan$^{24}$ huaŋ$^{31}$ hui$^{11}$ kai$^{24}$ kan$^{31}$ kʰĩ$^{31}$ kʰĩ$^{55}$ kʰĩm$^{11}$ kʰĩn$^{31}$ kʰio$^{24}$
kʰɲ$^{31}$ kʰu$^{24}$ kʰuaŋ$^{24}$ kĩ$^{24}$ kiun$^{11}$ kuan$^{31}$ lan$^{31}$ leu$^{24}$ lia$^{24}$ lio$^{55}$ liun$^{24}$ liun$^{31}$ loi$^{55}$
luai$^{11}$ lun$^{31}$ mie$^{11}$ mion$^{24}$ miun$^{55}$ mua$^{24}$ niaŋ$^{11}$ nie$^{24}$ niem$^{55}$ nĩm$^{31}$ nioi$^{55}$ niuŋ$^{31}$
noi$^{31}$ noi$^{55}$ nuaŋ$^{24}$ nun$^{24}$ ŋem$^{11}$ ŋĩ$^{55}$ ŋĩn$^{55}$ ŋon$^{55}$ ŋua$^{55}$ ŋuan$^{55}$ ŋun$^{31}$ ŋun$^{31}$
pʰem$^{24}$ pʰieu$^{31}$ pʰim$^{55}$ pʰio$^{11}$ pʰiu$^{11}$ pʰiu$^{31}$ pʰiun$^{24}$ pʰua$^{55}$ pʰuai$^{11}$ pʰuaŋ$^{24}$ pĩ$^{11}$
pieu$^{11}$ pio$^{24}$ pion$^{11}$ pua$^{11}$ san$^{11}$ siau$^{55}$ siem$^{55}$ sien$^{31}$ sĩn$^{31}$ sio$^{55}$ sion$^{55}$ sion$^{55}$
suai$^{31}$ suen$^{31}$ tɕa$^{11}$ tɕam$^{55}$ tɕam$^{55}$ tɕan$^{55}$ tɕan$^{55}$ tɕan$^{55}$ tɕau$^{55}$ tɕʰai$^{11}$ tɕʰam$^{24}$
tɕʰan$^{31}$ tɕʰeu$^{11}$ tɕʰeu$^{31}$ tɕʰĩ$^{55}$ tɕʰie$^{11}$ tɕʰie$^{55}$ tɕʰieu$^{24}$ tɕʰio$^{11}$ tɕʰiun$^{11}$ tɕʰo$^{24}$ tɕiam$^{11}$
tɕin$^{11}$ tɕin$^{11}$ tɕiuŋ$^{55}$ tɕon$^{31}$ tɕun$^{31}$ tʰa$^{11}$ tʰe$^{24}$ tʰiaŋ$^{24}$ tʰĩn$^{31}$ tʰioŋ$^{31}$ tʰiuŋ$^{55}$ tʰua$^{31}$
tʰuaŋ$^{31}$ tʰuen$^{31}$ tiem$^{55}$ tĩm$^{55}$ tion$^{24}$ tiu$^{55}$ tŋ$^{31}$ tŋ$^{31}$ tŋ$^{11}$ tsaŋ$^{11}$ tse$^{11}$ tsʰan$^{24}$
tsʰiam$^{55}$ tsʰiam$^{55}$ tsʰiaŋ$^{24}$ tsʰiau$^{11}$ tsʰin$^{31}$ tsʰiu$^{31}$ tsʰuan$^{31}$ tsʰuan$^{31}$ tsʰuan$^{31}$ tsia$^{24}$
tsiaŋ$^{11}$ tsiau$^{55}$ tsie$^{31}$ tsion$^{31}$ tsiun$^{31}$ tsiuŋ$^{11}$ tsiuŋ$^{55}$ tsoi$^{31}$ tsua$^{55}$ tua$^{55}$ vem$^{31}$ vĩ$^{55}$
viam$^{24}$ vie$^{31}$ vien$^{55}$ vieu$^{11}$ vioŋ$^{11}$ viuŋ$^{31}$



## Nonlexical Cantonese syllables

- 153 test items (plus 4 practice items and 12 vowel-initial items)

dɲ$^{11}$ fɐŋ$^{33}$ fɛ:m$^{45}$ fɛ:ŋ$^{55}$ fɛ:ŋ$^{11}$ fɛ:ŋ$^{22}$ fi:u$^{22}$ fɲ$^{22}$ fœ:ŋ$^{45}$ fœ:ŋ$^{11}$ fɐn$^{11}$ fɐy$^{11}$
hɛ:$^{11}$ hi:$^{23}$ hi:n$^{22}$ hu:n$^{11}$ ja:i$^{11}$ ja:n$^{22}$ ja:ŋ$^{22}$ ja:u$^{11}$ jei$^{55}$ jɐŋ$^{11}$ jɔ:i$^{45}$ jœ:$^{23}$ jɐn$^{45}$
jɐy$^{33}$ ka:ŋ$^{22}$ kɔ:i$^{23}$ kʰi:m$^{55}$ kʰœ:$^{23}$ kʰu:n$^{33}$ kʰu:n$^{11}$ kʰʊŋ$^{45}$ kʰy:$^{55}$ kʰy:n$^{22}$ ki:u$^{23}$
ku:$^{23}$ kʷa:u$^{23}$ kʷɔ:i$^{23}$ kʷɔ:n$^{55}$ kʷɔ:n$^{23}$ kʷɛ:$^{45}$ kʰʷa:m$^{55}$ kʰʷa:m$^{11}$ kʰʷa:u$^{45}$ kʰʷeu$^{55}$
kʰʷɔ:i$^{22}$ kʰʷɔ:n$^{33}$ kʰʷɔ:n$^{11}$ kʰʷɔ:ŋ$^{23}$ kʰʷɛ:$^{33}$ kʰʷi:n$^{22}$ kʰʷi:u$^{23}$ kʰʷɪŋ$^{33}$ kʰʷœ:$^{23}$
kʰʷou$^{22}$ kʰʷu:i$^{33}$ kʰʷu:n$^{11}$ kʰʷy:$^{11}$ kʰʷy:n$^{45}$ kʷi:m$^{22}$ kʷɐy$^{55}$ la:i$^{23}$ lɲ$^{45}$ lœ:ŋ$^{33}$ lu:n$^{55}$
ma:m$^{22}$ ma:n$^{55}$ mɔ:i$^{11}$ mi:u$^{33}$ mœ:$^{45}$ mɐy$^{45}$ mu:n$^{33}$ na:ŋ$^{55}$ na:ŋ$^{23}$ nɔ:i$^{33}$ nɛ:$^{45}$
nɛ:$^{11}$ ni:$^{45}$ ni:u$^{55}$ ŋa:i$^{23}$ ŋa:m$^{22}$ ŋa:ŋ$^{23}$ ŋɐŋ$^{11}$ ŋɐu$^{22}$ ŋi:m$^{22}$ ŋœ:ŋ$^{55}$ ŋɐn$^{45}$ ŋu:$^{55}$
ŋu:i$^{11}$ ŋy:$^{11}$ pa:m$^{33}$ pɛ:$^{22}$ pʰɐn$^{45}$ pʰɔ:i$^{23}$ pʰɛ:ŋ$^{55}$ pʰi:$^{22}$ pʰɐy$^{11}$ pi:$^{33}$ pi:$^{23}$ pɲ$^{45}$
pu:n$^{11}$ sa:$^{23}$ sa:i$^{11}$ sɛ:m$^{45}$ sɲ$^{11}$ sœ:$^{23}$ ta:u$^{22}$ tem$^{23}$ tʰa:$^{11}$ tʰa:i$^{11}$ tʰa:i$^{23}$ tʰa:ŋ$^{45}$
tʰeu$^{23}$ tʰɔ:i$^{33}$ tʰɛ:$^{11}$ tʰœ:ŋ$^{45}$ tʰu:n$^{11}$ ti:$^{11}$ ti:n$^{11}$ tœ:$^{23}$ tɐy$^{45}$ tsa:i$^{45}$ tsɐŋ$^{23}$ tsɔ:n$^{33}$
tsei$^{22}$ tsʰɐn$^{22}$ tsʰɔ:i$^{23}$ tsʰɔ:n$^{22}$ tsʰei$^{33}$ tsʰu:$^{11}$ tsʰy:$^{22}$ tsu:$^{33}$ tsu:i$^{55}$ tsu:i$^{23}$ tsu:n$^{11}$
tu:$^{22}$ tu:i$^{45}$ ty:$^{45}$ wa:m$^{55}$ wa:ŋ$^{45}$ wɐm$^{33}$ wɐu$^{33}$ wɔ:n$^{11}$ wei$^{33}$ wɛ:ŋ$^{33}$ wi:u$^{55}$
wi:u$^{11}$ wi:u$^{23}$ wœ:ŋ$^{33}$ wœ:ŋ$^{23}$ wɔŋ$^{22}$ wy:$^{55}$



## Nonlexical Southern Min syllables

- 101 test items (plus 8 practice items and 1 vowel-initial item)

bam$^{11}$ bɔm$^{11}$ biŋ$^{11}$ buan$^{35}$ ɕan$^{11}$ ɕan$^{55}$ ɕyŋ$^{51}$ ɕyŋ$^{55}$ ɕuai$^{35}$ ɕuaŋ$^{35}$
ga$^{53}$ gɔm$^{11}$ gyŋ$^{53}$ gia$^{11}$ giũ$^{11}$ gu$^{55}$ kɔm$^{35}$ kʰã$^{35}$ kʰai$^{35}$ kʰia$^{53}$ kʰiũ$^{11}$
kʰiũ$^{53}$ kʰoɔ̃$^{55}$ kʰuɔŋ$^{11}$ kʰuĩ$^{35}$ kiæn$^{11}$ kiæn$^{51}$ kiai$^{11}$ kiaŋ$^{51}$ kin$^{35}$ kio$^{55}$
km$^{11}$ kuai$^{35}$ lãu$^{55}$ lɣ$^{35}$ lia$^{35}$ loɔ̃$^{35}$ luaŋ$^{55}$ luĩ$^{11}$ lui$^{55}$ miæn$^{55}$ mio$^{35}$
miũ$^{35}$ mu$^{55}$ muaɪ$^{53}$ muãi$^{55}$ muaŋ$^{55}$ muɔŋ$^{35}$ nam$^{55}$ nam$^{55}$ nio$^{11}$
niɔ̃$^{55}$ niu$^{55}$ nuai$^{51}$ nuan$^{55}$ ŋãi$^{53}$ ŋiã$^{55}$ ŋio$^{35}$ ŋiŋ$^{55}$ ŋue$^{11}$ pʰam$^{11}$
pʰɣ$^{11}$ pʰɣ$^{35}$ pʰiaŋ$^{35}$ pʰioŋ$^{55}$ pʰoɔ̃$^{11}$ pʰuaŋ$^{35}$ pʰuɔŋ$^{11}$ pʰuɔŋ$^{35}$ pʰuɔŋ$^{51}$
piau$^{35}$ pio$^{35}$ siæn$^{11}$ tɕʰua$^{55}$ tɕʰuei$^{51}$ tɕʰun$^{11}$ tɔm$^{11}$ tɕo$^{51}$ tʰa$^{35}$
tʰãi$^{55}$ tʰim$^{53}$ tʰioɔ$^{55}$ tʰuaŋ$^{55}$ tia$^{51}$ tiai$^{11}$ tiaŋ$^{11}$ tsʰia$^{51}$ tsʰio$^{35}$ tsʰiũ$^{35}$
tsʰm$^{55}$ tsʰuã$^{35}$ tsʰuai$^{35}$ tsʰuĩ$^{53}$ tsiãu$^{53}$ tsiãu$^{55}$ tsioŋ$^{51}$ tuaɪ$^{53}$ xia$^{35}$
xia$^{51}$ xio$^{35}$ xɱ$^{11}$ xo$^{11}$



## Three analyses

- Phonotactic probability
- Onset naturalness
- Onset priming

- All analyses use response choice as the binary dependent variable, analyzed with mixed-effects logistic regression
  (maximal models taking participants, participant effect sizes, and items into account: Barr *et al.*, 2013)

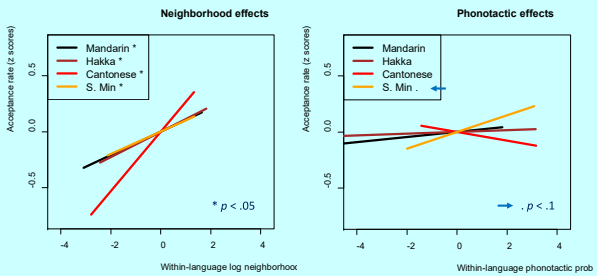


## Phonotactic probability

- **Phonotactic probability** is the overall transition probability between phonemes within an item, based on the lexicon of real words
  /kæt/: mean(P(#k)/P(#), P(kæ)/P(k), P(æt)/P(æ) , P(t#)/P(t))
  - Depends on decomposition of syllables into phonemes
- **Neighborhood density** is the overall similarity of an item with real words
  /kæt/: /hæt/, /kʌt/, /kæp/, /skæt/, /æt/, …
  - Affects processing even if syllables are treated as wholes
  - Distinct from phonotactic probability, but correlated with it
- In English speakers, both affect wordlikeness (Bailey & Hahn, 2001)
- In Cantonese speakers, the effect of phonotactic probability on wordlikeness seems to be much weaker (Kirby & Yu, 2007)



## Phonotactic probability: Results



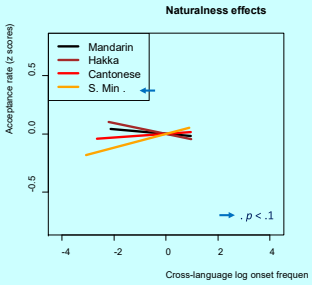

Plots show linear models, but they and the logistic-regression *p* values take both variables into account
No interactions with (log) syllable inventory size ($p > .6$)

## Onset naturalness

- Phonemes also vary in their frequency across languages
  - The more languages that have a given phoneme, the more "natural" it may be
- For example, in the present study we quantified onset naturalness by counting languages in the PHOIBLE database (Moran *et al.*, 2014)
- Naturalness affects English wordlikeness judgments (Hayes & White, 2013)
- It also affects Mandarin wordlikeness judgments (Myers, 2015)
  - But the stimuli were written (BPMF), showing onsets overtly
  - And even weak effects may be detectable in the huge sample (110 participants judging 3,274 items)



## Onset naturalness: Results




- The larger the (log) syllable inventory, the stronger this effect (language x naturalness: $p < .1$)

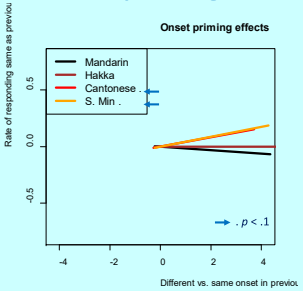


## Onset priming

- Both of the above analyses depend on getting the independent variables quantified correctly (phonotactic probability, naturalness)
- So our third analysis uses only experiment-internal information
- Is a participant's response to an item primed by the response to the previous item if it shares the same onset consonant?



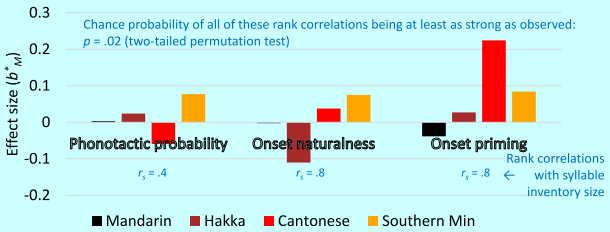
## Onset priming: Results




- The larger the (log) syllable inventory, the stronger this effect (language x priming: $p < .05$)



## Cross-language variation in effect size




Effect size measure from Menard (2004) (different from earlier linear plots). Rank correlations = Spearman correlations



## Conclusions

- A larger syllable inventory prompts speakers to decompose syllables into phonemes
- This pattern can be seen even in our small experiments on a small sample of languages with only small differences in their small syllable inventory sizes
- More stringent tests will require a much larger cross-language sample
- Fortunately, multi-language experimentation is easier than ever with online tools like Worldlikeness



## References (1/2)


Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory & Language, 44*, 569-591.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Chen, T.-Y., & Myers, J. (2017). Worldlikeness: A Web-based tool for typological psycholinguistic research. *University of Pennsylvania Working Papers in Linguistics, 23*(1), Article 4, 20-30.

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry, 44*(1), 45–75.

Kirby, J. P., & Yu, A. C. L. (2007). Lexical and phonotactic effects on wordlikeness judgments in Cantonese. *Proceedings of the International Congress of Phonetic Sciences, 16*, 1389-1392.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(01), 1-38.

LSHK [Linguistic Society of Hong Kong] (2002). *Guide to LSHK Cantonese romanization of Chinese characters* (2nd ed.). Hong Kong, China: Linguistic Society of Hong Kong.

Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *The American Statistician, 58*(3), 218-223.




## References (2/2)


MOE [Ministry of Education, R.O.C.] (2006). Dictionary of frequently used words in Taiwan Hakka. https://hakka.dict.edu.tw//hakkadict/index.htm

Moran, S., McCloy, D., & Wright, R. (Eds.) (2014). PHOIBLE Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://phoible.org/

Myers, J. (2015). Markedness and lexical typicality in Mandarin acceptability judgments. *Language & Linguistics, 16*(6), 791-818.

Myers, J. (2016). Meta-megastudies. *The Mental Lexicon, 11*(3), 329-349.

O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition, 115*, 282-302.

Ruan, J.-C., Hsu, C.-W., Myers, J., & Tsay, J. S. (2012). Development and testing of transcription software for a Southern Min spoken corpus. *International Journal of Computational Linguistics and Chinese Language Processing, 17*(1), 1-26. http://lngproc.ccu.edu.tw/SouthernMinCorpus/

Treiman, R., & Zukowski, A. (1996). Children's sensitivity to syllables, onsets, rimes, and phonemes. *Journal of Experimental Child Psychology, 61*(3), 193-215.

Tsai, C.-H. (2000). Mandarin syllable frequency counts for Chinese characters. http://technology.chtsai.org/syllable/.

Yu, M., Mo, C., Li, Y., & Mo, L. (2015). Distinct representations of syllables and phonemes in Chinese production: Evidence from fMRI adaptation. *Neuropsychologia, 77*, 253-259.