Bilingual Word Familiarity in Cantonese and English

Lauretta Cheng[1] (lspcheng@umich.edu) , Khia A. Johnson[2] (khia.johnson@ubc.ca), & Molly Babel[2] (molly.babel@ubc.ca)

[1]Linguistics, University of Michigan, Ann Arbor, Michigan, United States, [2]Linguistics, University of British Columbia, Vancouver, British Columbia, Canada

Abstract number: 1pSCb16

Introduction

- **Word frequency** (i.e., objective frequency) is often taken as a proxy for **word familiarity** (i.e., subjective frequency) [6].
- Word frequency is typically calculated from a corpus, but is that corpus's frequency representative of language experience for different types of bilinguals?
- For psycholinguistics and phonetics experiments, controlling for frequency can be an integral aspect of the experimental design.

Research Questions

1. How do speaker-listeners with different language backgrounds vary in word familiarity ratings for Cantonese and English words?
2. How do word familiarity ratings for each of the participant groups compare to typical metrics and proxies for familiarity?

Methods

Participants There are four different participant groups.

- Cantonese-English bilinguals who grew up in a/an:
  1. Cantonese-dominant location (Native; $n = 11$)
  2. English-dominant location (Heritage; $n = 33$)
- Early English speakers (no Chinese experience) who grew up in a/an:
  3. English-dominant location in North America (North American; $n = 23$)
  4. non-English-dominant location (International; $n = 6$)

Apparatus Online experiment implemented with jsPsych [2].

Familiarity Task The methods for the task:

- Presented audio stimuli of Cantonese (228 words, 9 nonwords) and English (152 words, 10 nonwords) in separate language blocks.
- Cantonese items consisted of words with (near-)minimal pair target syllables. English items were (near-)minimal pairs.
- Participants rated the familiarity of items on scale from 1 (very familiar) to 5 (somewhat familiar) to 9 (very unfamiliar).

Language Background Questionnaire Focused on knowledge, use and family background, especially for English, Cantonese, and other Chinese languages.

Corpora/Sources

- **Cantonese MacArthur-Bates CDI.** Proportion of Hong Kong children producing word in Cantonese at 30 months [7]
- **Hong Kong Cantonese Corpus.** Conversations and radio programs in Hong Kong Cantonese [4]
- **Heritage Language Documentation Corpus.** Sociolinguistic interviews with 3 generations in Toronto Cantonese [5]
- **English MacArthur-Bates CDI.** Proportion of American children producing word in English at 30 months [3].
- **SUBTLEX-US.** American English subtitles from films and TV series [1].

Word familiarity results

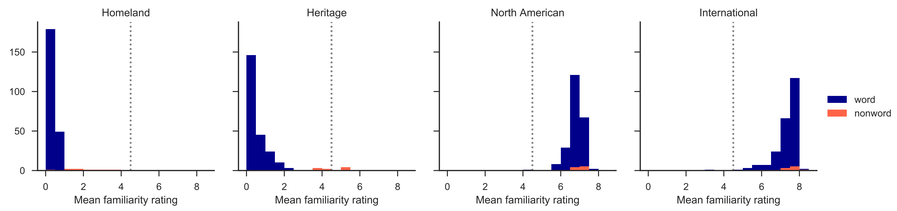

Figure 1: Cantonese mean word familiarity ratings

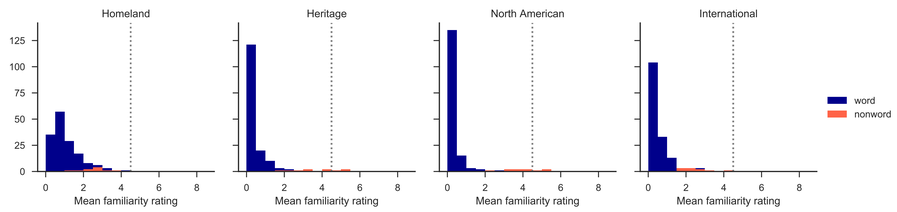

Figure 2: English mean word familiarity ratings

Comparisons: Cantonese

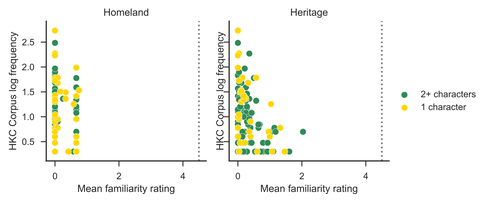

Figure 3: Mean word familiarity ratings for Cantonese are compared against the log frequency values from the HKC Corpus (Native $r = -0.03$, Heritage $r = -0.32$).

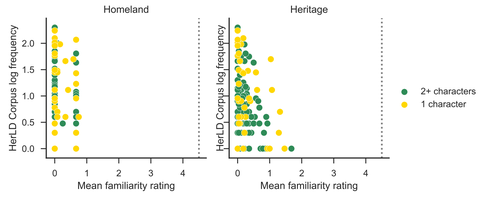

Figure 4: Mean word familiarity ratings for Cantonese are compared against the log word frequency values estimated from the HerLD Corpus (Native $r = -0.03$, Heritage $r = -0.30$).

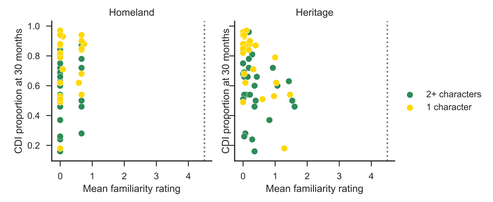

Figure 5: Mean word familiarity ratings for Cantonese are compared against the Cantonese CDI proportions at 30 months (Native $r = 0.09$, Heritage $r = -0.40$).

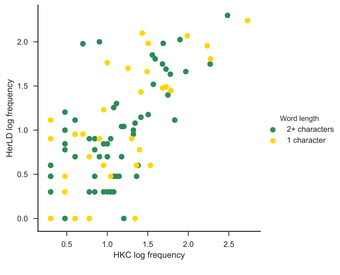

Figure 6: Frequency in the HKC Corpus and HerLD Corpus sociolinguistic interviews is strongly correlated ($r = 0.71$).

Comparisons: English

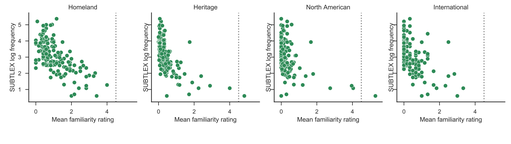

Figure 7: Mean word familiarity ratings for English are compared against the log frequency value from the SUBTLEX-US corpus (Native $r = -0.65$, Heritage $r = -0.59$, North American $r = -0.44$, International $r = -0.54$).

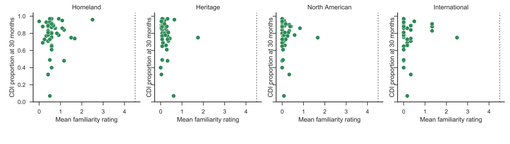

Figure 8: Mean word familiarity ratings for English are compared against the English CDI proportions at 30 months (Native $r = 0.09$, Heritage $r = -0.16$, North American $r = 0.03$, International $r = 0.003$).

Discussion & Conclusion

- Words were highly familiar in both languages, yet groups show slightly different patterns of familiarity ratings. (Recall that our sample sizes vary dramatically between groups.)
  - These different ratings demonstrate that different language experiences lead to unique impressions of word familiarity.
  - This suggests that these groups should be treated as different populations of speakers.
- The HKC and HerLD corpora are well correlated, but both have gaps in word coverage, which likely reflects differences in topics/genre and how they were collected.
- Heritage speakers (moderately) correlate with the corpus metrics, but Homeland speakers do not.
  - Words in this study were selected to be familiar, though they appear to be slightly less familiar for Heritage compared to Homeland groups.
  - If more unfamiliar/uncommon words were included, we would expect to see a stronger correlation for both groups. However, a concern with Heritage speaker populations is low frequency words may be identified as nonwords.
- The strongest correlations are in the English data with SUBTLEX.
  - SUBTLEX is by far the largest corpus we included, and thus arguably more representative of listeners' knowledge.
  - Our English word list also included more low familiarity items compared to our Cantonese list.
- Overall, Cantonese frequency from the available corpora does not correlate all that well with familiarity ratings for these bilingual groups, and thus may not be be representative of the different groups' language experiences.

Take Home Point

For a specific population under study (e.g. language/dialect, bilinguals, children), it is important to consider how representative a "general native speaker corpus" is, and whether the corpus has sufficient coverage. In cases where corpus usage is not appropriate, pretesting stimuli with familiarity ratings may be a safer alternative prior to use in a phonetics or psycholinguistics experiment.

References

[1] M. Brysbaert, B. New, and E. Keuleers.
Adding part-of-speech information to the SUBTLEX-US word frequencies.
Behavior Research Methods, 44(4):991–997, mar 2012.

[2] J. R. de Leeuw.
jsPsych: A JavaScript library for creating behavioral experiments in a web browser.
Behavior Research Methods, 47(1):1–12, 2015.

[3] M. C. Frank, M. Braginsky, D. Yurovsky, and V. A. Marchman.
Wordbank: An open repository for developmental vocabulary data.
Journal of Child Language, 44(03):677–694, may 2016.

[4] K. K. Luke and M. L. Wong.
The hong kong cantonese corpus: Design and uses.
Journal of Chinese Linguistics, 25(2015):309–330, 2015.

[5] N. Nagy.
A multilingual corpus to explore variation in language contact situations.
Rassegna Italiana di Linguistica Applicata, 43(1-2):65–84, 2011.

[6] R. Rapp.
On the relationship between word frequency and word familiarity.
In B. Fisseni, H.-C. Schmitz, B. SchrÃüder, and P. Wagner, editors, *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. BeitrÃdge zur GLDV-Tagung 2005 in Bonn.*, pages 249–263. Peter Lang, Frankfurt, 2005.

[7] T. Tardif, P. Fletcher, W. Liang, and N. Kacirotti.
Early vocabulary development in mandarin (putonghua) and cantonese.
Journal of Child Language, 36(05):1115, may 2009.

Acknowledgements

Thank you to members of the Speech In Context lab for their contributions to this work, especially Stephanie Chung. This work has been supported in part by a SSHRC grant F16-04616 awarded to Babel. We thank the creators of the HLVC Corpus for sharing their data with us. The HLVC Corpus is funded by a research grant from the Social Sciences and Humanities Research Council of Canada 410-2009-2330 (2009-2012) and is developed by Naomi Nagy, Alexei Kochetov, Yoonjung Kang, and James Walker.