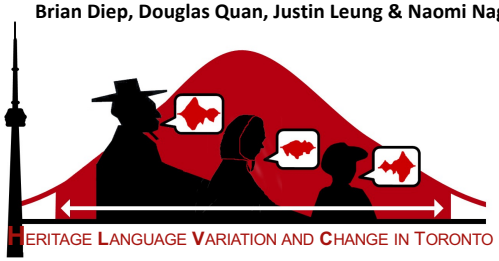
Variation in Transcribing Heritage Cantonese

Brian Diep, Douglas Quan, Justin Leung & Naomi Nagy



HTTPS://NGN.ARTSCI.UTORONTO.CA/HLVC

justinr.leung@mail.utoronto.ca

Workshop on Cantonese, Hong Kong Baptist University, 24 June 2023

Social Sciences and Humanities Research Council of Canada | Conseil de recherches en sciences humaines du Canada

UNIVERSITY OF TORONTO

1

What is the HLVC Project?

- Large-scale project investigating variation and change in Toronto's heritage languages
  - To **document and describe heritage languages** spoken by immigrants and 2 generations of their descendants
    - To compare to a Homeland "baseline" & local English (also variable)
    - to understand changes in progress
  - To **create a corpus** available for research on a variety of topics
  - To **push variationist research beyond its monolingually-oriented core (and its majority language focus)**
  - To **promote HL vitality** through research, training, and "knowledge mobilization" in and out of the classroom

(Nagy & Meyerhoff 2008, Nagy 2017)

2

Cantonese in 2 multilingual cities

http://www.servcorp.com.hk/media/9459/-twoifc-building-3

https://evolvetours.com/wp-content/uploads/2015/04/CN-Tower.jpg

In Hong Kong:

- **86%** of population uses CAN
- 83% are able to speak English
- **5%** have "very good" English

Census and Statistics Dept. (2014:4)

In Toronto (since the 1970s):

- **4.5%** report CAN as mother tongue (267,000 MT speakers)
- 96% are able to speak English
- **77%** (mostly) use English "at home"

Statistics Canada (2016)

Variables previously examined

Dependent variables	Homeland change?	Heritage change?
(I): /i/-/ɪ/ split (Tse 2019)	Yes	Yes
(Y): /y/-/u/ merger (Tse 2019)	No	Yes
(E): /ɛ/ split by coda (Tse 2019)	No	No
(O): /ɔ/ split by coda (Tse 2019)	Yes	No
(VOT) (Tan & Nagy 2017)	No	No
(N): /n/ → /l/ (Nagy in press)	Yes	No
(CL): specializing to singular nouns (Nagy & Lo 2019)	No	Yes
(PRODROP): less pro-drop (Nagy et al. 2011, Nagy in press)	No	No
(MOTION): satellite-/verb-framing (Leung 2022)	No	Yes

Cantonese interviews in the HLVC Corpus

- Audio recordings of sociolinguistic interviews conducted in Cantonese
- Two types of transcripts
  - Jyutping: phonetics/phonology (cf. Nagy et al. in press; Tan & Nagy 2017; Tse 2016, 2019, 2022)
  - Chinese characters: lexicon, morphosyntax (cf. Nagy & Lo 2019; Leung 2022)

5

Traditional methods of interview transcription

- Manual transcription in **jyutping**
  - Easier for heritage-speaking student volunteers and RAs with limited knowledge of characters
- Manual transcription in **Chinese characters**
  - Helps with disambiguating homophones

6

Transcription status

		Jyutping + Characters	Jyutping only	Total
Heritage (Toronto)	Gen 1	11	7	18
	Gen 2	10	15	25
	Gen 3	2	4	6
Homeland (Hong Kong)		11	3	14
Total		34	29	63

The Problem

- Many interview transcripts are jyutping only
- Inconsistencies in transcripts
  - Orthographic variants
  - Mergers in RA/volunteers
- What method(s) can we use to systematically generate additional character transcriptions, leveraging the fact that we have a partially annotated dataset?

Methods

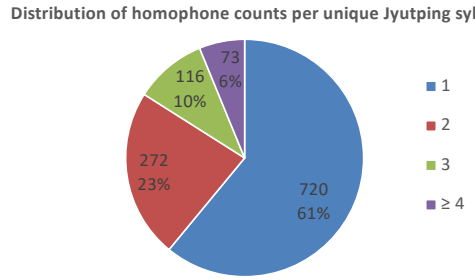
1) Corpus-based imputation

2) Machine-transliteration (Diep 2022)

9

Corpus-based imputation

- Some fully annotated interviews → Use existing annotations to predict the most likely character for any given Jyutping syllable
- Similar themes in interviews based on the same interview script → many recurring words across interviews



10

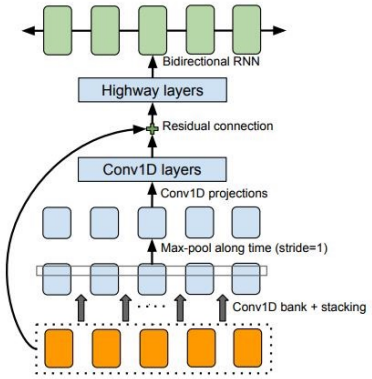
Corpus-based imputation

- For any jyutping string that is missing character transcriptions, we look at the fully annotated corpus and find all the annotated characters that share the same jyutping
- We directly impute the most likely character that shares the jyutping
- We can also store all other possible characters for easier manual correction by a research assistant
- In the future, this could be further improved by looking at bigrams, etc.
  - but this is moving toward our second method

11

Machine-Transliteration

- With advances in natural language processing, applications of machine learning (ML) have become more accessible and widespread in research
  - Still rare for Cantonese
- We adapt an ML model developed for transliteration of Mandarin pinyin to characters for the equivalent task with jyutping



Architecture of a transliterator model (Wang et al. 2017)

12

Machine-Transliteration Data

We train on publicly-available corpora of Cantonese with both character and jyutping transcriptions, such as:

- Hong Kong Cantonese Corpus (Luke & Wong 2015)
- Child Heritage Chinese Corpus (Mai & Yip 2017)
- Guthrie Bilingual Corpus (Guthrie 1983)
- HKU-70 Corpus (Weizman & Fletcher 2000)
- Lee-Wong-Leung Corpus (Lee et al. 1991–94)
- Leo Corpus (Mai & Yip 2022)
- Paidologos Corpus: Cantonese (Edwards & Beckerman 2008)
- Yip-Matthews Bilingual Corpus (Yip & Matthews 2007)

Over 4 million Chinese words from over 1,232 interviews

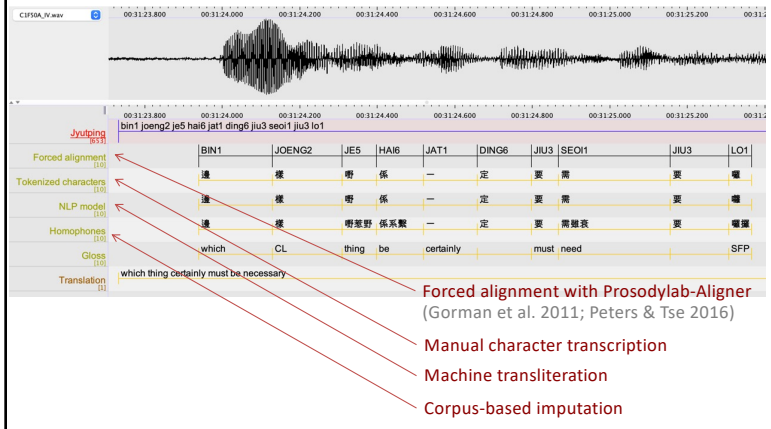
13

Machine-Transliteration

After training on the aforementioned corpora, we apply the model to the HLVC corpus data to convert jyutping annotations to characters.

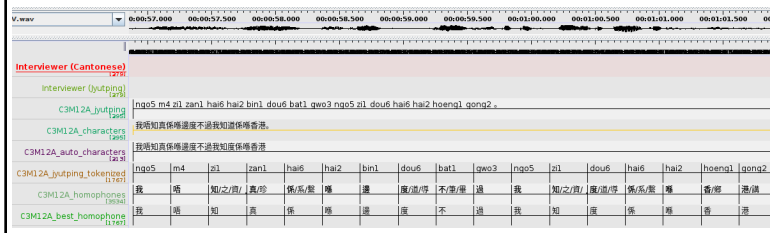
14

Example of the two methods



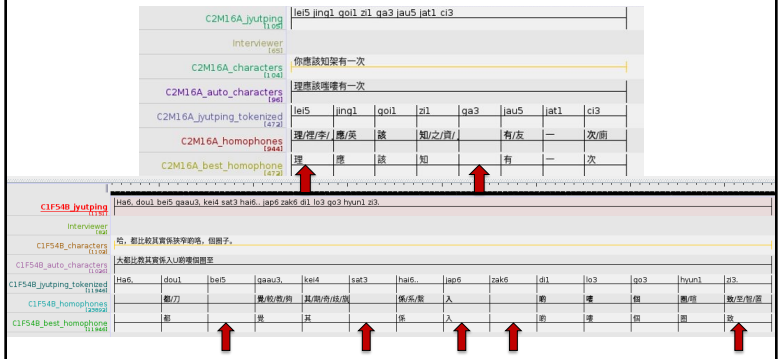
15

Methods Comparison



16

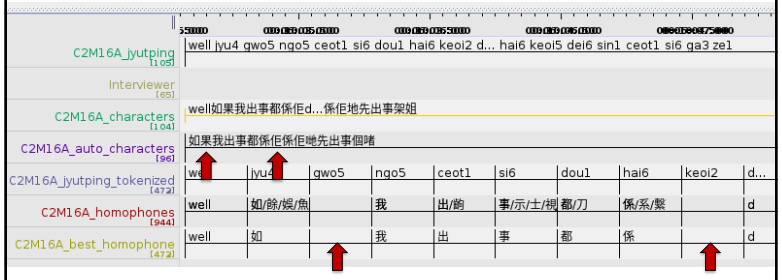
Methods Comparison



Both methods show poor results when the original jyutping transcription is “incorrect”:

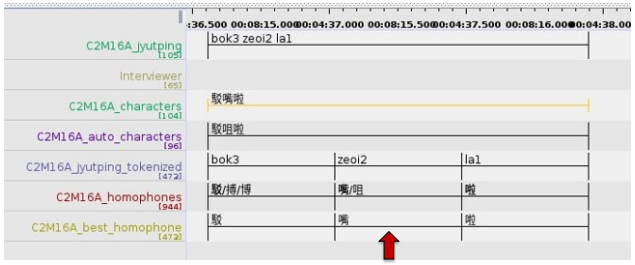
- Genuine transcriber errors: haap3 zaak3 → jap6 zak6, zi2 → zi3, gaa3 → ga3
- Mergers in speaker: nei5 → lei5
- Possible mergers in transcriber: bei2 → bei5, sat6 → sat3

Methods Comparison



- Machine transliteration sometimes has a slight advantage when dealing with problems mentioned in the previous slide
  - gwo2 → gwo5
  - keoi5 → keoi2

Methods Comparison



Corpus-based imputation shows better results when predicting simpler/common words

CER for Transliteration Methods

We report similar average CER (character error rate) for each transliteration method across homeland and heritage speakers

	Corpus-based Imputation	Machine Transliteration
Homeland	0.30	0.21
Heritage	0.31	0.26

Discussion

- **Corpus construction** can be expedited using **(semi-)automated processes**
  - Naïve approach is ~70% accurate
  - ML approach boosts accuracy by 5–10%
  - Comparison of approaches can help with proofreading
- **Promoting HL vitality** through **"knowledge mobilization"**
  - Speakers who are less- or non-literate in the variety can contribute to the documentation of their own language variety

21

Thanks to the HLVC Cantonese RAs

Abigail Chan・Ariel Chan・Karen Chan・Karina Cheung・Kate Cheung・Mira Chow・Mira Chung・Naomi Cui・Joyce Fok・Kei-Fung Ho・Teresa Kwok・Cookie Lau・Vina Law・Ewen Lee・Wilma Lee・Samuel Lo・Tiffany Luk・Yannie Mork・Jonathan Ng・Rita Pang・Valerie Pang・Andrew Peters・Samantha Pong・Ian Quan・Bonita Sham・Janice Shum・Mario So Gao・Haili Su・Katherine Sung・Angel Tam・Ziwen Tan・Josephine Tong・Sarah Truong・Holman Tse・Elaine Wang・Victoria Wang・Ka-man Wong・Joyce Woo・Junrui Wu・Olivia Yu・Christopher Zhu・Minyi Zhu

多謝

22

References

Census and Statistics Department. (2014). Hong Kong monthly digest of statistics: Use of language in Hong Kong in 2012. http://www.statistics.gov.hk/pub/B71406FB2014XXXXB0100.pdf. Accessed April 15, 2017.

Diep, B. (2022). *Towards a neural model for jyutping to character transliteration* [Course paper for STA497: Readings in Statistics].

Edwards, J., & M. E. Beckman. 2008. Methodological question in studying consonant acquisition. *Clinical Linguistics and Phonetics, 22*(12), 939–958.

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics, 39*(3), 192–193.

Guthrie, L. F. (1983). Learning to use a new language: Language functions and use by first grade Chinese-Americans. ARC Associates.

Lee, T. H.T., C. H. Wong, S. Leung, P. Man, A. Cheung, K. Szeto, & C. S. P. Wong. (1991–94). The development of grammatical competence in Cantonese-speaking children. Report of RGC earmarked grant.

Leung, J. R. (2022). Variation in path encoding in motion events in Toronto Heritage Cantonese. *University of Pennsylvania Working Papers in Linguistics, 28*(2), Article 10.

Luke, K. K., & Wong, M. L. Y. (2015). The Hong Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics Monograph Series, 25*, 312–333.

Mai, Z., & V. Yip. (2017). Acquiring Chinese as a heritage language in English-speaking countries and the Child Heritage Chinese Corpus. International Conference on Bilingualism: Language and Heritage, Chinese University of Hong Kong, Dec. 18.

Mai Z., & V. Yip. (2022). Caretaker input and trilingual development of Mandarin, Cantonese and English in early childhood (1;6-2;11). *International Journal of Bilingual Education and Bilingualism, 25*(9), 3389–3403.

Nagy, N. (2017). Heritage language speakers in the university classroom, doing research. In P. Trifonas & T. Aravossitas, eds. *International handbook on research and practice in heritage language education*. Springer.

Nagy, N. (in press). *Heritage languages: Extending variationist approaches*. Cambridge University Press.

23

References, cont.

Nagy, N., N. Aghdasi, D. Denis, & A. Motut. (2011). Pro-drop in Heritage Languages: A cross-linguistic study of contact-induced change. *Penn Working Papers in Linguistics* 17.2.

Nagy, N. & S. Lo. (2019). *Classifier use in Heritage and Hong Kong Cantonese. Asia-Pacific Language Variation & Change 5*(1), 84–108.

Nagy, N., & M. Meyerhoff. (2008). Introduction: Social lives in language. In M. Meyerhoff & N. Nagy, eds. *Social lives in language – Sociolinguistics and multilingual speech communities: Celebrating the work of Gillian Sankoff*. John Benjamins.

Nagy, N., H. Tse & J. Stanford. (in press). Have Cantonese tones merged in spontaneous speech? In R. Rao, ed., *The phonetics and phonology of heritage languages*. Cambridge University Press.

Peters, A. & Tse, H. (2016). Evaluating the efficacy of Prosody-lab Aligner for a study of vowel variation in Cantonese. Workshop on Innovations in Cantonese Linguistics, Ohio State University, March 12.

Statistics Canada. (2017). *Focus on Geography Series, 2016 Census* https://www12.statcan.gc.ca/census-recensement/2016/as-sa/fogs-spg/Facts-CMA-Eng.cfm?TOPIC=5&LANG=Eng&GK=CMA&GC=535

Tan, Z. T. & N. Nagy. (2017). VOT in heritage and Hong Kong Cantonese. Canadian Linguistic Association, Toronto, May 29.

Tse, H. (2016). Variation and change in Toronto Heritage Cantonese: An analysis of two monophthongs across two generations. *Asia Pacific Language Variation, 2*(2), 124–156.

Tse, H. (2019). Vowel shifts in Cantonese?: Toronto vs. Hong Kong. *Asia-Pacific Language Variation, 5*(1), 67–83.

Tse, H. (2022). What can Cantonese heritage speakers tell us about age of acquisition, linguistic dominance, and sociophonetic variation? In R. Bayley, D. Preston, & X. Li (Eds.), *Variation in second and heritage languages: Crosslinguistic perspectives*.

Wang, Y. et al. (2017). Tacotron: Towards end-to-end speech synthesis. *Proc. Interspeech* 2017, 4005–4010.

Weizman, Z. O. & P. Fletcher. (2000). A comparative study of language development: English and Cantonese pre-schoolers in Hong Kong. Committee on Research and Conference Grants, University of Hong Kong.

Yip, V., & S. Matthews. (2007). The bilingual child: Early development and language contact. Cambridge University Press.

24