Exploring automated formant analysis for comparative variationist study of Heritage Cantonese and English

CVC VIII
Change and Variation in Canada
Changement et Variation au Canada
Strathy Language Unit, Queen's University • 31 mai – 1 juin, 2014

Naomi Cui¹, Minyi Zhu¹, Vina Law¹
Holman Tse² & Naomi Nagy¹
¹University of Toronto
²University of Pittsburgh

Heritage Language Variation and Change in Toronto
http://projects.chass.utoronto.ca/NGN/HLVC
What is the HLVC Project?

• Large-scale project investigating language use and change in heritage (non-official) languages spoken in Toronto.

• Goals
  – To document and describe heritage languages spoken by immigrants and 2 generations of their descendants
  – To create a corpus available for research on language change
  – To push variationist research beyond its monolingually-oriented core by focusing on heritage language use among multilingual speakers
  – To develop a framework for research on heritage languages and contact
A Sample of Previous HLVC Work

<table>
<thead>
<tr>
<th></th>
<th>Cantonese</th>
<th>Faetar</th>
<th>Italian</th>
<th>Korean</th>
<th>Russian</th>
<th>Ukrainian</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOT</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ø-subject</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Borrowing</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vowels</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* This presentation
Vowels

• Very well researched in sociolinguistics, but very little work on vowel variation and change in languages other than English.

• Large body of research has made possible the development of new technologies/techniques to make vowel analysis easier
  – Example: FAVE (Rosenfelder et al 2011)
Goals of Current Project

• To determine the extent to which the vowel systems of Cantonese and English may be mutually influencing each other in Toronto

• To extend the use of automated forced alignment and formant extraction as tools for the sociolinguistic study of contact-induced change in Heritage Cantonese.
  – Prosodylab-Aligner (Gorman et al 2011) to be adopted
Methodological Problems

• Large amount of data in HLVC Corpus (~40 speakers/language)
  – Manual formant measurements take a lot of time.
• FAVE designed to work only on English
• Could Prosodylab-Aligner be a viable alternative?
The language quilt

By Catherine Farley and Damian Listar/TORONTO STAR

English is still, by far, the first language across Greater Toronto. But strip away that blanket of dominance and a colourful patchwork emerges, showing where newcomers from around the world chose to settle. The map shows the most prevalent mother tongue after English in more than 1,000 neighbourhoods across the GTA, as revealed by a Star analysis of 2006 census data.

Some discoveries:
- English is the second language in 47 of the GTA’s 1,076 census tracts.
- English is in third place in 7 tracts in Agincourt, on Toronto’s northern edge.
- In 57 tracts, 70% or more of the population has a non-English mother tongue.
- The preponderance of English as a mother tongue is 90% or more in 42 census tracts.
- In 200 tracts, more than 30 distinct mother tongues are spoken by 15 or more people.

The mother tongue mix:
- 56% of the GTA’s 5.4 million residents count English as their mother tongue.

Top 10 mother tongues (after English):

1. Italian 3.5%
2. Chinese 3.2%
3. Cantonese 3.1%
4. Punjabi 2.5%
5. Portuguese 2%
6. Spanish 2%
7. Tagalog 1.9%
8. Urdu 1.8%
9. Tamil 1.7%
10. Polish 1.6%

From irregular stitching, an ongoing work of art.

Where English is the second language... and even the third:

- Brampton
- Woodbridge
- Markham/Anglicourt

Behind the mosaic:

- The Great Lakes region has 4,300 speakers, one of the GTA’s most densely populated areas.
- Many people place second behind English in 20 census tracts in central Toronto.
- Etobicoke, with 4,900 speakers, is one of the smallest GTA communities, but it ranks 8th in bilingualism.
- Dundas, east of the city, and midtown areas are notably low in bilingualism.
- Peel, Halton and Waterloo region have 6,034 speakers.
- Peel, Halton, Waterloo region have 6,034 speakers.

Building the map of second languages:

- The map was compiled by Catherine Farley and Damian Listar from data collected by Statistics Canada.
- The map is intended to provide a broad overview of the linguistic diversity in the GTA.

## Contrasting demographics

<table>
<thead>
<tr>
<th>Language</th>
<th>MT speakers (2011 Census)</th>
<th>Ethnic Origin (2006 Census)</th>
<th>Est. in TO</th>
<th>Speakers come from</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cantonese</td>
<td>170,000+</td>
<td>537,000</td>
<td>1951</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Italian</td>
<td>166,000</td>
<td>466,000</td>
<td>1908</td>
<td>Calabria</td>
</tr>
<tr>
<td>Russian</td>
<td>78,000</td>
<td>58,505</td>
<td>1916</td>
<td>St. Petersburg, Moscow</td>
</tr>
<tr>
<td>Korean</td>
<td>51,000</td>
<td>55,000</td>
<td>1967</td>
<td>Seoul</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>26,000</td>
<td>122,000</td>
<td>1913</td>
<td>Lviv</td>
</tr>
<tr>
<td>Faetar</td>
<td>&lt;100?</td>
<td>300?</td>
<td>1950</td>
<td>Faeto, Celle di St. Vito (Apulia Italy)</td>
</tr>
</tbody>
</table>

www40.statcan.ca/l01/cst01/demo12c-eng.htm;
The year indicates the founding of the first church in Toronto for each group, indicative of community establishment.
Cantonese vs. English Vowel Space

Allophonic lowering of /i/ before velars (Yue-Hashimoto 1972)
- si1, /si1/, 詩, ‘poem’
- sik1, [sɪk1], 識, ‘to know’

Similar Canadian English Vowels
- see, /si/
- sick, /sɪk/

Images from Wikipedia
Expected outcome

Heritage Language / Culture

1st

English/Canadian

2nd
Data

- Two sets of hour-long sociolinguistic interviews from 2 generations of speakers identified as Hong Kong Chinese and who claim Cantonese as a heritage language
  - Not from the same speakers, however.

Interviews in English from the Contact in the City Corpus (CinC) (Hoffman and Walker 2010)

Interviews in Cantonese from the HLVC Corpus (Nagy 2009, 2011)

“My parents came to Toronto in 1972.”

“Ngo5 fu6 mou5 yat1 gau2 cat1 yi6 lin4 lei4 dou3 do1 leon1 do1.”
# Speaker Sample

<table>
<thead>
<tr>
<th>Generation</th>
<th>Sex</th>
<th>CANTONESE</th>
<th>ENGLISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Ages: 42-82)</td>
<td>Male</td>
<td>C1M62A C1M59A</td>
<td>TO.035 TO.038</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>C1F78A C1F54A C1F82A</td>
<td>TO.030 TO.037 TO.039</td>
</tr>
<tr>
<td>2 (Ages: 16-44)</td>
<td>Male</td>
<td>C2M44A</td>
<td>TO.029</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>C2F16A C2F21B</td>
<td>TO.031 TO.056</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>N=8</td>
<td>N=8</td>
</tr>
</tbody>
</table>
Methods - English Data

1. Sentence-level time alignment (manual) using ELAN

2. Word- and phoneme-level time alignment (automated) with FAVE
   - [http://fave.ling.upenn.edu](http://fave.ling.upenn.edu)
Forced Alignment & Vowel Extraction (FAVE)

An online suite for automatic vowel analysis
Prosodylab-Aligner (Gorman 2011 et al)

- A Python script used to perform text to audio speech alignment
- Supports training on arbitrary data
  - With any input from language X, can be trained to deal with acoustic data from language X
- Requirements
  - At least a total of one hour of audio (.wav file in chunks OK)
  - Matching .lab files (.txt files readable by Prosodylab-Aligner) for each .wav file
  - A customized dictionary
Methods – Cantonese Data

1. Interviews transcribed by native speakers of Cantonese using Jyutping Romanization in ELAN
   – Manual sentence-level alignment

2. To create input readable by Prosodylab-Aligner, PRAAT script used to create smaller .wav files with matching .txt files for each annotation.
Translation: “Because at that time, China was at war.”

Translation: “And then the Communist Party came, and then ...”
Training and Evaluation

• .wav files and matching .lab files put in a Training directory

• Prosodylab-aligner uses Training directory and dictionary to build a training model

• Prosodylab-aligner uses training model to evaluate the same files in the same directory

Custom dictionary in the format of The CMU Pronouncing Dictionary
Textgrid Output of Prosodylab-Aligner
Another PRAAT script: formant extraction

- Formant information extracted from Prosodylab-Aligner generated Textgrids and matching .wav files using PRAAT script
- Output: Tab-delimited .txt file
Vowel Normalization

The Vowel Normalization and Plotting Suite

- http://ncslaap.lib.ncsu.edu/tools/norm/norm1.php
- Labov ANAE (Vowel Extrinsic) method used
Prepping for R-Brul

- Tab-delimited .txt file generated by NORM with normalized values for vowel formants
- New columns added for variables
- Ready for statistical analysis with R-brul (Johnson)
Variables of Interest

• External Factors
  – Generation
  – Gender
  – Age

• Internal Factors
  – Following Segment
  – Tone
Cantonese Vowel Charts

Hong Kong Homeland CAN

Toronto CAN (8 speakers), Labov ANAE (speaker extrinsic)

Image from Wikipedia
Toronto Anglo ENG vs CAN ENG

Based on means from Roeder 2012, Boberg 2008, Roeder & Jarmasz 2010

Based on means of 7 speakers
# F1 and F2 Means for /i/ in open syllables

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gen 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>439</td>
<td>2044</td>
<td>3207</td>
</tr>
<tr>
<td>2</td>
<td>423</td>
<td>2106</td>
<td>857</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>435</td>
<td>2057</td>
<td>4064</td>
</tr>
</tbody>
</table>

- Gen 2 has higher and more fronted /i/  
  *p < 0.05

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gen 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>454</td>
<td>2096</td>
<td>1545</td>
</tr>
<tr>
<td>2</td>
<td>434</td>
<td>2324</td>
<td>2370</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>441</td>
<td>2234</td>
<td>3925</td>
</tr>
</tbody>
</table>

- Gen 2 has higher and more fronted /i/  
  **p < 0.01

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Toronto Anglo English</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>474</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>2011</td>
<td></td>
</tr>
</tbody>
</table>

- Anglo English has the lowest /i/.
Discussion of Results

• Evidence of generational change clear with same general developmental trend in both languages.
  – Raising and fronting of /i/ for Gen 2 in both CAN and CAN ENG
• Relative position of /i/ and /ɪ/ are different in CAN and ENG.
• Lack of /u/ fronting in CAN observed, but some fronting in CAN ENG
• How these changes result from contact with English (if that is the case) appear to be quite complex
  – further research required to better understand how.
• Note
  – Tone not considered as a factor
  – Variation and change in other vowels not considered
  – No homeland data available
Discussion of Methodology

• Without human intervention, automatically extracted data creates reasonable vowel plots
• A promising avenue for future research on vowel variation and change in heritage languages
• But need to check and compare results with manual formant extraction
Future Work

• Assessing accuracy of automated alignment and formant extraction by attempting to replicate results using manual methods

• Expanding to more vowels and more speakers
  – 8 speakers for this analysis, ~ 40 CAN speakers in Corpus
  – Comparing homeland data

• Expanding to other heritage languages
  – Italian, Faetar, Russian, Ukrainian, Korean
**HLVC RAs:**
Cameron Abma
Vanessa Bertone
Ulyana Bila
Rosanna Calla
Minji Cha
Karen Chan
Joanna Chociej
Sheila Chung
Tiffany Chung
Courtney Clinton
Radu Craioveanu
Marco Covi
Derek Denis
Tonia Djogovic
Joyce Fok
Paolo Frasca
Matt Gardner

Rick Grimm
Dongkeun Han
Natalia Harhaj
Taisa Hewka
Melania Hrycyna
Michael Iannozzi
Diana Kim
Janyce Kim
Iryna Kulyk
Mariana Kuzela
Ann Kwon
Alex La Gamba
Carmela La Rosa
Natalia Lapinskaya
Kris Lee
Nikki Lee
Olga Levitski
Arash Lotfi

Paulina Lyskawa
Rosa Mastri
Timea Molnár
Jamie Oh
Maria Parascandolo
Rita Pang
Andrew Peters
Tiina Rebane
Hoyeon Rim
Will Sawkiw
Maksym Shkvorets
Vera Richetti Smith
Anna Shalaginova
Konstantin Shapoval
Yi Qing Sim
Mario So Gao
Awet Tekeste
Josephine Tong

Sarah Truong
Dylan Uscher
Ka-man Wong
Olivia Yu
Minyi Zhu

**Collaborators:**
Yoonjung Kang
Alexei Kochetov
James Walker

**Funding:**
SSHRC, University of Toronto,
Shevchenko Foundation

[HTTP://PROJECTS.CHASS.UTORONTO.CA/NGN/HLVC](HTTP://PROJECTS.CHASS.UTORONTO.CA/NGN/HLVC)
References

• Lobanov
• Yue Hashimoto, Oi-kan 1972. Phonology of Cantonese. Cambridge University Press.
Heritage Language Variation and Change in Toronto

HTTP://PROJECTS.CHASS.UTORONTO.CA/ngn/HLVC

Email: naomi.nagy@utoronto.ca

For today’s slides: email: hbt3@pitt.edu