Goals

- training forum in which to develop protocols for sharable data that conform to the spirit of NSF policy (for sharable archived data)
- describe and problematize how we indicate use of multiple languages within one conversation and efforts to maintain consistency across protocols from different languages/communities, commenting on efforts to make these transcripts useful for inquiries developed subsequent to transcription
- appropriate metadata for language choice (at speaker level)
- specific coding conventions for language choice (at word/phrase level, while transcribing)
What is the HLVC Project?

- Large-scale project investigating Variation and Change in Toronto’s Heritage Languages.

- Project’s goals (Nagy 2011)
  - To document and describe heritage languages (HL) spoken by immigrants and 2 generations of their descendants
  - 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) (cf. Nagy & Meyerhoff 2008)

- Descriptive and theoretical goals:
  - develop generalizations about the types of variable features, structures or rules that are borrowed earlier and more often
  - Use consistent methods across languages and variables

If the IVs are in the HLs, why are we in this workshop?

- HLVC goal is to describe heritage languages and we do everything possible to elicit data in those varieties (cf. Labov 1984 on interview methods)

- But, in a multilingual metropolis people regularly use >1 language, including in interviews

- So we need to annotate language choices for 3 reasons:
  - Exclude “English” – however we define that – from HL analysis
  - Many students & scholars are interested in using the data to study code-switching
  - Code-switching rate may be an important independent variable (cf. Torres & Travis 2011)
Contrasting demographics
Toronto, 2011 Census

<table>
<thead>
<tr>
<th>Language</th>
<th>Speakers</th>
<th>Ethnic Origin</th>
<th>Est. in TO</th>
<th>Speakers from</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cantonese</td>
<td>170,000+</td>
<td>594,735</td>
<td>1951</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Italian</td>
<td>166,000</td>
<td>475,090</td>
<td>1908</td>
<td>Calabria</td>
</tr>
<tr>
<td>Russian</td>
<td>78,000</td>
<td>118,090</td>
<td>1916</td>
<td>St. Petersburg, Moscow</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>26,000</td>
<td>130,355</td>
<td>1913</td>
<td>Lviv</td>
</tr>
<tr>
<td>Polish</td>
<td>75,275</td>
<td>214,460</td>
<td>1911</td>
<td>Western Poland</td>
</tr>
<tr>
<td>Korean</td>
<td>51,000</td>
<td>64,755</td>
<td>1967</td>
<td>Seoul</td>
</tr>
<tr>
<td>Faetar</td>
<td>&lt;300?</td>
<td>800</td>
<td>1950</td>
<td>Faeto &amp; Celle (Apulia)</td>
</tr>
</tbody>
</table>

www40.statcan.ca/l01/cst01/demo12c-eng.htm; www12.statcan.gc.ca/

Nagy & Łyskawa / LSA 2016
Data collection methods for naturalistic speech

1. Sociolinguistic interview
2. Ethnic Orientation Questionnaire
3. Picture Description Task
   All conducted and recorded by native speakers in the heritage language

Different languages; different protocols

- Focus on representation and annotation of English in transcripts of conversations in Heritage Languages (Cantonese, Faetar, Italian/Calabrese, Korean, Polish, Russian, and Ukrainian)

1. Review methods that differ by language team
2. Show some implications
3. Discuss best option(s) for standardizing
Metadata (about speakers)

- HLVC Interview catalog contains (some) notes regarding switches to English

Examples:
- RUS & UKR: nothing noted in catalog (but easy to count in interview .eafs)
- KOR, as the result of a year-long study (Chung 2010), has a "code-switch" column: yes/some/no
- POL, as the result of 2 year-long studies (Łyskawa 2015, Łyskawa et al.), has notes on code-switching: lots/Ø
- ITA has few notes: For over 40 speakers, we see 2:
  - “Very chatty, lots of code switching!” (I2F53A)
  - “Does not speak much Italian at all, words are mostly cued in by interviewer, partial transcription as a consequence” (I3M15A)
- CAN:
  - “Clear, Lots of English Phrases” (C2M21B)
  - good sound, lots of English...One-word answers” (C3F12A)
  - “Speaks lots of English” (C3F18A and C2M14A)

Ukrainian- the most straightforward

- transcribe English words with capital letters
- “If a word exists in both language, then I will listen closely to the phonology and transcribe it accordingly.
- If they pronounce an English word with a Ukrainian accent then I will transcribe it in Ukrainian, but I will make a note in the notes tier.” [MH]
UKR example in ELAN
(translation added)

UKR examples

- **UH YEAH tił’ky tak SUBCONSCIOUSLY vin vin je dužhe spravedlyvyj UM v kozhnomu sensi teper**
  
  *Uh yeah just subconsciously he he is very fair um in every sense now.*  
  
  [U3M41A_IV.eaf, 31:30]

- **vin nazyvajet’sja ATTILIO ja joho nazyvaju ARISTOTLE**
  
  *He is called Attilio, I call him Aristotle.*  
  
  [U2F60A_IV.eaf, 48:38]

- **chasamy my jidemo do Fljorydy na MARCH BREAK i todi my USUALLY idemo do des’ na lito SO**
  
  *Sometimes we go to Florida for March Break and then we usually go to somewhere for the summer so*  
  
  [U3F13A_IV.eaf, 9:37]

Regular expression searchable: [A-Z] & Notes tier
RUS examples

- Aga, ja prepodavala francuzskij v [ENG: UofT], jeto bylo vsjo [ENG: part-time].
  Yes, I taught French at UofT, it was all part-time. [R1F55B_IV_PR.eaf, 2:38]

- Tam oni ochen' mnogo tam [ENG: fundraising] i tam raznyx vesjolyx veshhej.
  There they do a lot of fundraising and various fun things.
  [R2F12A_IV_PR.eaf, 0:24]

- Regular expression searchable: “[ENG:"

RUS protocol

- 3-letter language tag “ENG” (or another language) introduces any non-Russian word/phrase, which is bracketed

- “Whether we use English spelling or transliterate the utterance depends largely on how the speaker says it, whether they use English-like or Russian-like pronunciation.” [NL]

- Proper nouns like “UofT” are written in English. Russian words (sometimes) exist for the same concepts.”

- English words with Russian morphemes are transcribed as Russian

English words with Russian morphemes

- *da, vam poslajsat' kolbasku ili pisikom, da.*

  *Yes, would you like your kielbasa sliced or in one piece, yes.*
  
  \[R1M56A_IV_PR.eaf, 0:29:23\]

  - poslajsat
  - pisikom
  - po+slice+at’
  - piece+ik+o

  - “to slice”
  - “in one piece”

- Not regular expression searchable

---

Cantonese

- **Current transcription system: use Jyutping (jyut6 ping3)** romanization
  - every Cantonese word has a number indicating tone as the final character
  - But there are also tone markings on some English words
  - Mandarin borrowings aren’t distinguished

- **Now adding:** transcribing characters (粵語字)
  - Cantonese and English will be more distinct
  - Mandarin borrowings will still not be searchable [SL]

---

Nagy & Łyskawa / LSA 2016
Examples from Cantonese

- seng4 jat6, ngo5 dei6 seng4 jat6 heoi3 pet store go2 zan4 si4 le3 keoi5 hai6
  When we go to the pet store she always goes -- [C2F16A_IV.eaf 29:29]

- "Usually" like mou2 gam3 je6 la1
  usually, like, not very late [C2F16A_IV.eaf 6:55]

- zing3 fu5 le1 dau6 jau5 jat1 di1 giu3 zou6 housing scheme bei5 ni1 di1
  The gov't has something known as housing scheme to provide-- [C1M61A_IV.eaf, 4:20]

- English words without tone are regular expression searchable: [a-zA-Z]s

Integrated English borrowing in CAN

Nagy & Łyskawa / LSA 2016
FAE examples

■ ANNOTATED BORROWING:
  ■ In toska in kiamuntə la i lamponi
  ■ in Tuscan they call the "the raspberries" (ITALIAN) [F1M75A&family_IV_part1.eaf, 29:26]

■ UNANNOTATED BORROWING:
  andʌj vanantə frut: ɛ vɛdʒ:ɛtabl
  where they sell fruits and vegetables [F1F70A_IV.eaf, 2:01.925]

Faetar

■ Probably not regular expression searchable
  ■ “English” is (sometimes) annotated in separate tier or (“ITALIAN”) follows transcription
KOR example

- code-switching (a-theoretical cover term for all kinds of lexical mixing) is marked in transcriptions by use of Roman rather than Hangul characters
- Regular expression searchable: [a-zA-Z]


<table>
<thead>
<tr>
<th>Type of integration into source language</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(Poplack 1980:584)</strong></td>
</tr>
<tr>
<td>Type</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>
# Code-switching vs. Borrowing

## Type of integration into source language

*(Poplack 1980:584)*

<table>
<thead>
<tr>
<th>Type</th>
<th>Phonological</th>
<th>Morphological</th>
<th>Syntactic</th>
<th>Code-Switching?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>No; Borrowing</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Type 1)**

저는 북 [buk] 들 많이읽어요

*I-TOPIC  book-PLUR a lot read-POL*

“I read a lot of books.” [K2F22A]

*book has Korean, not English phonology [buk].

Korean plural morpheme “들” is incorporated with *book.*

Korean syntax (SOV) is used.

→ *book* is a borrowing and not code-switching.

---

**Type 2)**

아빠는 movies 좋아해요

*Dad-TOP  movies like-POL*

“[My] dad likes movies.” [K2M25A]

*Movies has English phonology [muviz].

The plural ‘s’ is English morphology.

Korean syntax: SOV

Thus, phonology and morphology are not integrated into Korean → CS
**Code-switching vs. Borrowing**

<table>
<thead>
<tr>
<th>Type</th>
<th>Phonological</th>
<th>Morphological</th>
<th>Syntactic</th>
<th>Code-Switching?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>No; Borrowing</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Type 3)**

Seventy-three **니까** thirty-six years
Seventy-three **so** thirty-six years
“[The year] ’73, so 36 years” [K1M70A]
“thirty-six years” has Korean phonology”: [tarti].

Sheila Chung 2010 (LIN 497 paper)

---

**Code-switching vs. Borrowing**

<table>
<thead>
<tr>
<th>Type</th>
<th>Phonological</th>
<th>Morphological</th>
<th>Syntactic</th>
<th>Code-Switching?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>No; Borrowing</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Type 4)**

저한테는 I’m hoping they’ll learn it
*Me for-TOPIC I’m hoping they’ll learn it*
“For me, I’m hoping they’ll learn it” [K2M24A]

No integration into Korean → CS

Sheila Chung 2010 (LIN 497 paper)
Where in the typology do we mark speech as “English”?

Type of integration into source language (Poplack 1980:584)

<table>
<thead>
<tr>
<th>Type</th>
<th>Phonological</th>
<th>Morphological</th>
<th>Syntactic</th>
<th>Code-Switching?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>No; Borrowing</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- Where we place the threshold determines how much/where a speaker uses each language.

Threshold (of integration) position determines how much a speaker uses each language
Mean Rate of Code-and Ethnic Orientation

![Graph showing Mean Rate of Code-and Ethnic Orientation](Image)

Sheila Chung 2010 (LIN 497 paper)

Recommendations (for the HLVC proofreading phase)

- General principles for transcribing a corpus for multiple uses:
  1. Find a way that’s fast to do basic mark-up of “everything.”
  2. Let people investigating specific issues do further mark-up.

- Anything English-y should be marked.

- Mark-up could be on a separate tier or bracketed & flagged. Which is better?
  - If on separate tier, then time-aligned (slower to produce; faster to analyze).
  - For KOR (and sometimes CAN) it’s a different orthography so no further annotation is needed.

- Proper nouns need to be marked.
  - Hyphenate proper nouns reliably.
  - Use capitalization only for proper nouns.

Nagy & Łyskawa / LSA 2016
Many tasks require tagging language choice

- Coding sociolinguistic variables
- Measuring phonetic variation
- “Quick” measures of “proficiency”
  - Speech rate – exclude English switches?
  - Vocab size – how many words are English?
  - Code-switching rate
  - Note: We don’t necessarily want these measures to “work,” i.e., correlate to sociolinguistic variation or to EOQ, but there is tension between the methods of var. sociolinguistics, endangered lg. documentation & SLA
- Automation, such as forced alignment

Connecting Ethnic Orientation

- Ethnic Orientation (EO) is assumed to correlate to many linguistic variables including code-switching rates and types.
- Our studies have produced mixed results.
- We begin by quantifying the responses to each item in the Ethnic Orientation Questionnaire on a scale:
  - 0 = English / Canada oriented
  - 1 = mixed
  - 2 = heritage language / Homeland oriented
  - These can be examined in isolation, or totalled, or averaged, or analyzed by Principle Components...
Correlations w/ code-switching rate for Heritage Polish

<table>
<thead>
<tr>
<th>Significant correlations with Code-switching Rate</th>
<th>Non-significant correlations with Code-switching Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual EO (Q A1)</td>
<td>Homeland contact</td>
</tr>
<tr>
<td>Language Use ave.</td>
<td>Parents’ lg. use ave.</td>
</tr>
<tr>
<td>Language Choice ave.</td>
<td>Partners’ lg. use ave.</td>
</tr>
<tr>
<td>Overall EO score</td>
<td>Cultural practices</td>
</tr>
<tr>
<td>Case mismatch</td>
<td>Discrimination experiences</td>
</tr>
<tr>
<td>Devoicing</td>
<td>Age</td>
</tr>
<tr>
<td></td>
<td>Generation</td>
</tr>
</tbody>
</table>

Correlations differ by language

Nagy & Łyskawa / LSA 2016

Nagy, Chociej & Hoffman 2012, Fig. 2
Speech rate and its (non-)correlation w/ social factors
(Italian and Ukrainian, 1,838 sentences)

Predicted and observed correlates to speech rate

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generation</td>
<td>√</td>
</tr>
<tr>
<td>EOQ</td>
<td>√</td>
</tr>
<tr>
<td>Sex</td>
<td>x</td>
</tr>
<tr>
<td>Age</td>
<td>x</td>
</tr>
</tbody>
</table>

Brook & Nagy submitted

Speech rate x Language Use (no effect)
Arrows indicate predicted effect.

Nagy & Łyskawa / LSA 2016
### Summary: HLVC annotation protocols for language choice

<table>
<thead>
<tr>
<th>Language</th>
<th>Method of marking English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ukrainian</td>
<td>capital letters</td>
</tr>
<tr>
<td>Russian, Polish</td>
<td>language tag [ENG: ]</td>
</tr>
<tr>
<td>Cantonese</td>
<td>English has no tones marked (sometimes) now: Cantonese vs. Roman characters (sort of)</td>
</tr>
<tr>
<td>Faetar</td>
<td>note on a different tier</td>
</tr>
<tr>
<td>Korean</td>
<td>Roman characters</td>
</tr>
<tr>
<td>Italian</td>
<td>none</td>
</tr>
</tbody>
</table>

---

**Nagy & Łyskawa / LSA 2016**
### References


Brook, M. & N. Nagy. submitted. Does speech rate indicate proficiency or identity in heritage languages?


---

### References, p. 2


