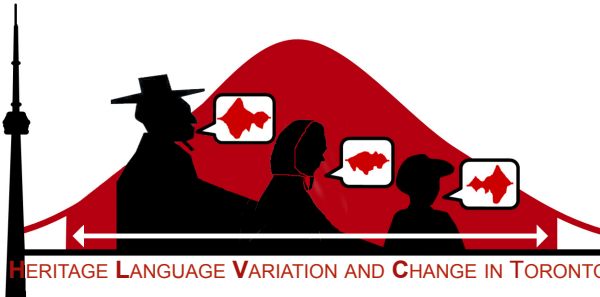



Moving forward with multilingual transcription




Naomi.Nagy@utoronto.ca Paulina Lyskawa@umd.edu



HERITAGE LANGUAGE VARIATION AND CHANGE IN TORONTO

[HTTP://PROJECTS.CHASS.UTORONTO.CA/NGN/HLVC](http://projects.chass.utoronto.ca/ngn/hlvc)

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

Goals

2

- 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?

3

- 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

Nagy & Łyskawa / LSA 2016

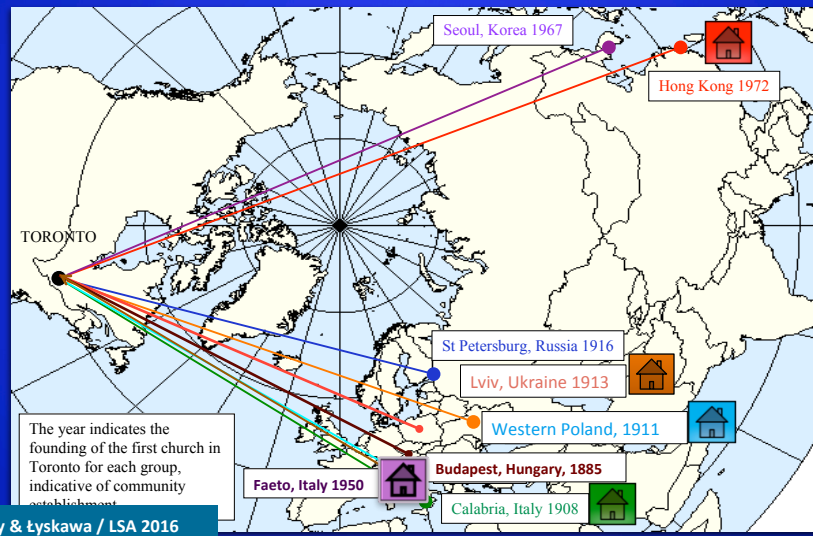
If the IVs are in the HLs, why are we in this workshop?

4

- 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)

Nagy & Łyskawa / LSA 2016

Heritage Language Variation and Change



Contrasting demographics Toronto, 2011 Census

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

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

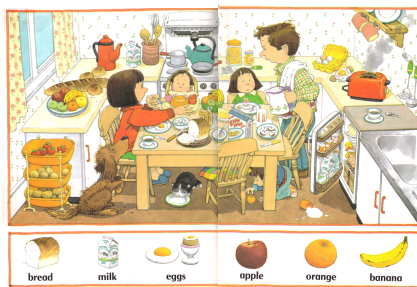
Data collection methods for naturalistic speech 7

1. Sociolinguistic interview
2. Ethnic Orientation Questionnaire
3. Picture Description Task

All conducted and recorded by
native speakers in the heritage language



Nagy & Łyskawa / LSA 2016



Amery & Cartwright 1987

Different languages; different protocols 8

- Focus on representation and annotation of English in transcripts of conversations in Heritage Languages (Cantonese, Faeta, 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

Nagy & Łyskawa / LSA 2016

Metadata (about speakers)

9

- 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. fc.*), has notes on code-switching: *lots/∅*
 - 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)

Nagy & Łyskawa / LSA 2016

Ukrainian- the most straightforward

10

- 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]

Nagy & Łyskawa / LSA 2016

11

UKR example in ELAN

(translation added)

U3M41A

Nr	Annotation
324	che- ja dumajo shho pomohlo bo ja serjozno vs'o brav I MEAN
325	ale ja ne dumaju shho vono by taku velyku riznycu zrobylo jakshho by vin to ne robyv
326	ale ven mene pxav bil'she nizh moju sestru
327	UH YEAH til'ky tak SUBCONSCIOUSLY vin vin je duzhe spravedlyvyj UM v kozhnomu sensi teper
328	ale ja dumaju v tomu chasi vin mav- EXPECTATIONS byly inakshi
329	to tak iak shho do matymatuku i take vo vs'n to hv malo butv moii rechii RIGHT?

0:31:40.160 Selection: 00:31:40.160 - 00:31:48.950 8790

40.000 00:31:41.000 00:31:42.000 00:31:43.000 00:31:44.000 00:31:45.000

U3M41A [539]

ENG TRANS [6]

Uh yeah just subconsciously he he is very fair um in every sense now

Nagy & Łyskawa / LSA 2016

12

UKR examples

- UH YEAH til'ky tak SUBCONSCIOUSLY vin vin je duzhe 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][A-Z] & Notes tier**

Proper names are problematic

Phon. integration influences mark-up

Nagy & Łyskawa / LSA 2016

13

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:”

Nagy & Łyskawa / LSA 2016

14

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
- NB: transcription is transliteratable with Comrie & Corbett’s (2002) system, at <http://www.translit.ru/>

Nagy & Łyskawa / LSA 2016

15

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

po+slice+at'

"to slice"

pisikom

piece+ik+o

"in one piece"

- Not regular expression searchable

Nagy & Łyskawa / LSA 2016

16

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

17

- 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 zau6 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

Nagy & Łyskawa / LSA 2016

Integrated English borrowing in CAN

18



Nagy & Łyskawa / LSA 2016

FAE examples

19

■ ANNOTATED BORROWING:

■ *in toscan i kiamuntə lə i lamponi*

■ *in Tuscan they call the "the raspberries" (ITALIAN)*

[F1M75A&family_IV_part1.eaf, 29:26]

■ UNANNOTATED BORROWING:

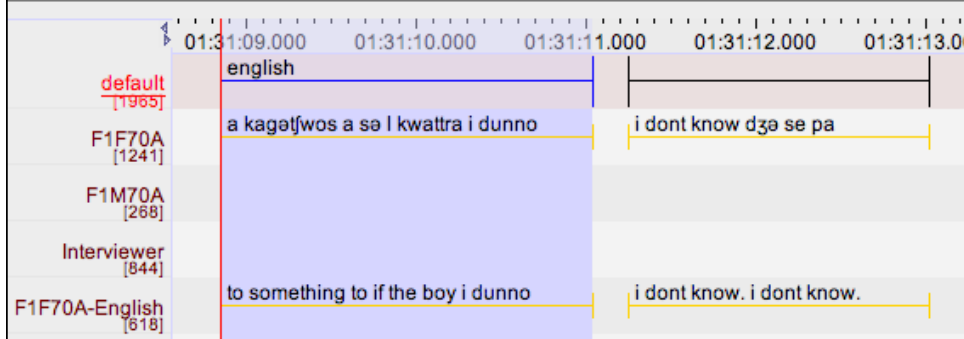
andʌj vənʌntə frut: ɛ vɛdʒ:ɛtabl

where they sell fruits and vegetables [F1F70A_IV.eaf, 2:01.925]

Nagy & Łyskawa / LSA 2016

Faetar

20



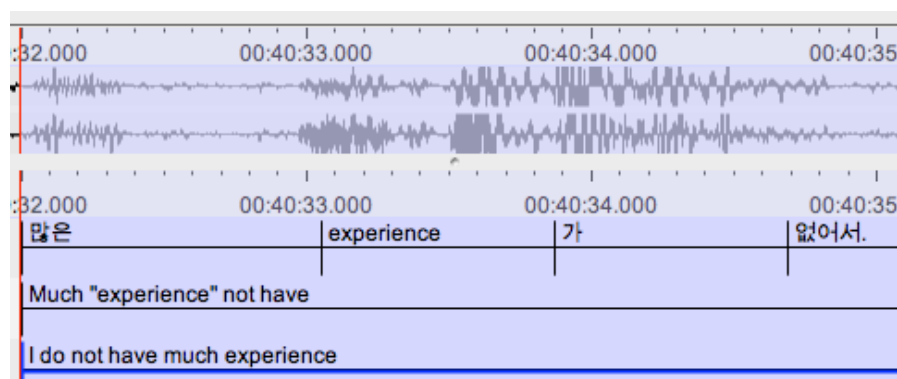
- Probably not regular expression searchable
- “English” is (sometimes) annotated in separate tier or “(ITALIAN)” follows transcription

Nagy & Łyskawa / LSA 2016

KOR example

K2M22A

21



- 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]`



Nagy & Łyskawa / LSA 2016

Chung, S. 2010. **Code-switching as a means of cultural identity among Koreans in Toronto.** TULCON '10 conference, U of Toronto.

22

Type of integration into source language (Poplack 1980:584)				
Type	Phonological	Morphological	Syntactic	Code-Switching?
1	+	+	+	No; Borrowing
2	-	-	+	Yes
3	+	-	-	Yes
4	-	-	-	Yes

Nagy & Łyskawa / LSA 2016

Code-switching vs. Borrowing

23

Type of integration into source language

(Poplack 1980:584)

Type	Phonological	Morphological	Syntactic	Code-Switching?
1	+	+	+	No; Borrowing
2	-	-	+	Yes

Type 1)

저는 북 [buk] 들 많이 읽어요
I-TOPIC book-PLUR a lot read-POL

"I read a lot of books." [K2F22A]

book has Korean, not English phonology [bʊk].

Korean plural morpheme "들" is incorporated with *book*.

Korean syntax (SOV) is used.

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

Nagy & Łyskawa / LSA 2016

Code-switching vs. Borrowing

24

Type of integration into source language

(Poplack 1980:584)

Type	Phonological	Morphological	Syntactic	Code-Switching?
1	+	+	+	No; Borrowing
2	-	-	+	Yes

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

Nagy & Łyskawa / LSA 2016

Sheila Chung 2010 (LIN 497 paper)

Code-switching vs. Borrowing

25

Type of integration into source language

(Poplack 1980:584)

Type	Phonological	Morphological	Syntactic	Code-Switching?
1	+	+	+	No; Borrowing
2	-	-	+	Yes
3	+	-	-	Yes

Type 3)

Seventy-three **thirty-six years**

Seventy-three so **thirty-six years**

"[The year] '73, so 36 years" [K1M70A]

"thirty-six years" has Korean phonology": [tʰɪrtʰi].

Sheila Chung 2010 (LIN 497 paper)

Nagy & Łyskawa / LSA 2016

Code-switching vs. Borrowing

26

Type of integration into source language

(Poplack 1980:584)

Type	Phonological	Morphological	Syntactic	Code-Switching?
1	+	+	+	No; Borrowing
2	-	-	+	Yes
3	+	-	-	Yes
4	-	-	-	Yes

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)

Nagy & Łyskawa / LSA 2016

Where in the typology do we mark speech as “English”?

27

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

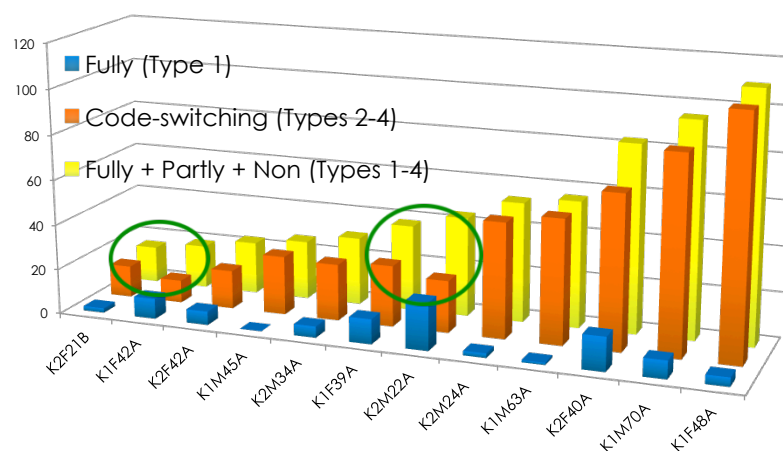
Type	Phonological	Morphological	Syntactic	Code-Switching?
1	+	+	+	No; Borrowing
2	-	-	+	Yes
3	+	-	-	Yes
4	-	-	-	Yes

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

Nagy & Łyskawa / LSA 2016

Threshold (of integration) position determines how much a speaker uses each language

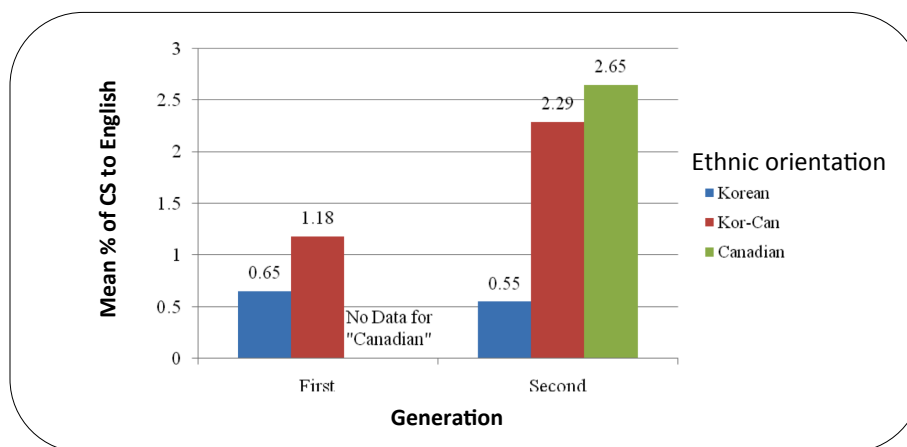
28



Nagy & Łyskawa / LSA 2016

29

Mean Rate of Code-and Ethnic Orientation



Sheila Chung 2010 (LIN 497 paper)

Nagy & Łyskawa / LSA 2016

30

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

31

- 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 sociolx variation or to EOQ, but there is tension between the methods of var. sociolx., endangered lg. documentation & SLA
- Automation, such as forced alignment

Nagy & Łyskawa / LSA 2016

Connecting Ethnic Orientation

32

- 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...
 - (cf. Keefe & Padilla 1987, Nagy, Chocieł & Hoffman 2012 (@ LSA Satellite Workshop))

Nagy & Łyskawa / LSA 2016

Correlations w/ code-switching rate for Heritage Polish

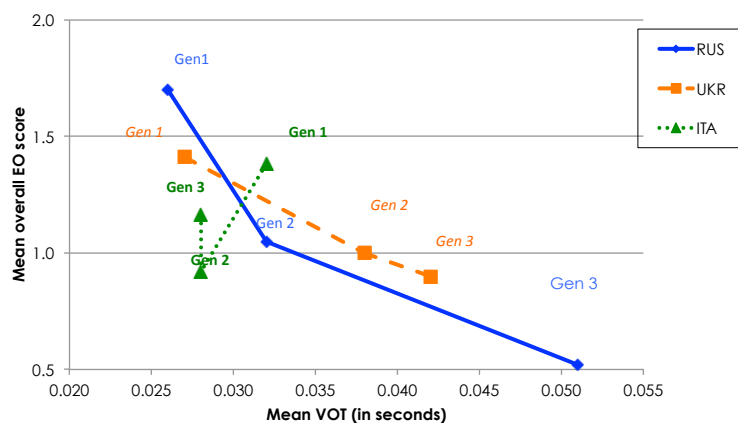
33

Significant correlations with Code-switching Rate		Non-significant correlations with Code-switching Rate
Individual EO (Q A1)	-	Homeland contact
Language Use ave.	-	Parents' lg. use ave.
Language Choice ave.	-	Partners' lg. use ave.
Overall EO score	-	Cultural practices
Case mismatch	+	Discrimination experiences
Devoicing	+	Age
		Generation

Nagy & Łyskawa / LSA 2016

Correlations differ by language

34



Nagy & Łyskawa / LSA 2016

Nagy, Chocieł & Hoffman 2012,
Fig. 2

Speech rate and its (non-)correlation w/ social factors

35

(Italian and Ukrainian, 1,838 sentences)

Predicted and observed correlates to speech rate

	Predicted	Observed
Generation	√	√
EOQ	√	x
Sex	x	√
Age	x	x

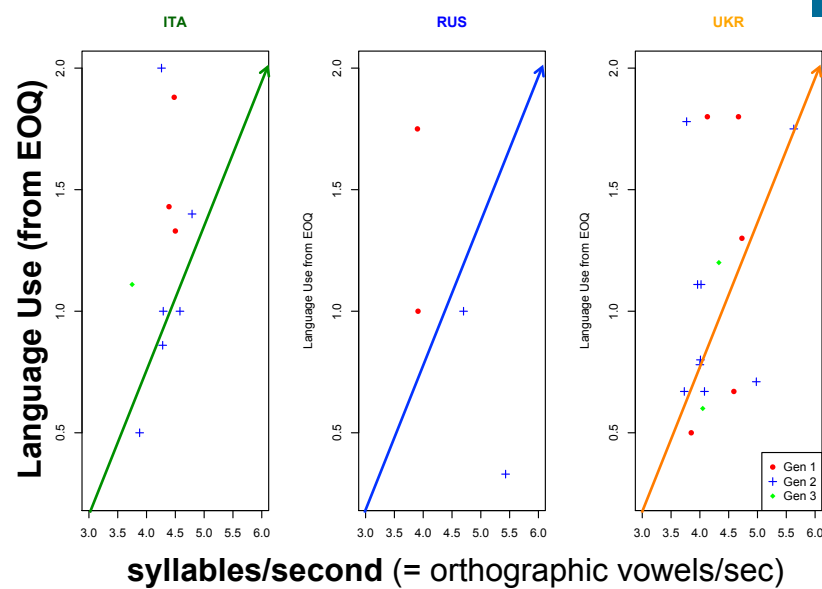
Brook & Nagy submitted

Nagy & Łyskawa / LSA 2016

Speech rate x Language Use (no effect)

Arrows indicate *predicted* effect.

36



Nagy & Łyskawa / LSA 2016

Summary: HLVC annotation protocols for language choice

37

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

Nagy & Łyskawa / LSA 2016

감사합니다 Дзякую Grazie molto Спаси́бо 多謝 gratsia namuorə

38

HLVC RAS:

Cameron Abma	Tonia Djogovic	Vina Law	Will Sawkiw
Vanessa Bertone	Joyce Fok	Kris Lee	Maksym Shkvorets
Ulyana Bila	Paolo Frascà	Nikki Lee	Vera Richetti Smith
Rosanna Calla	Matt Gardner	Olga Levitski	Anna Shalaginova
Minji Cha	Julia Grasso	Samuel Lo	Konstantin Shapoval
Abigail Chan	Rick Grimm	Arash Lotfi	Yi Qing Sim
Ariel Chan	Dongkeun Han	Paulina Łyskawa	Mario So Gao
Karen Chan	Natalia Harhaj	Rosa Mastri	Vlodymyr
Joanna Chocieł	Taisa Hewka	Timea Molnár	Sukhodolskiy
Vivien Chow	Melania Hrycyna	Valeriya Mordvinova	Awet Tekeste
Sheila Chung	Michael Iannozzi	Francesco Muoio	Letizia Tesi
Tiffany Chung	Diana Kim	Jamie Oh	Josephine Tong
Courtney Clinton	Janyce Kim	Maria Parascandolo	Sarah Truong
Radu Craioveanu	Iryna Kulyk	Deepam Patel	Dylan Uscher
Marco Covi	Mariana Kuzela	Rita Pang	Qian Ling Wang
Naomi Cui	Ann Kwon	Andrew Peters	Ka-man Wong
Zahid Daujee	Alex La Gamba	Alessia Plastina	Junrui Wu
Derek Denis	Carmela La Rosa	Tiina Rebane	Olivia Yu
	Natalia Lapinskaya	Hoyeon Rim	Minyi Zhu

Nagy & Łyskawa / LSA 2016

[HTTP://PROJECTS.CHASS.UTORONTO.CA/NGN/HLVC](http://projects.chass.utoronto.ca/ngn/hlvc)

References

39

- Amery, H. & S. Cartwright. 1987. *First 100 Words*. Usborne, London
- Brook, M. & N. Nagy. *submitted*. Does speech rate indicate proficiency or identity in heritage languages?
- Chung, S. 2010. Code-switching as a means of cultural identity among Koreans in Toronto. TULCON '10 & Cornell Undergraduate Linguistics Colloquium 2010.
- Comrie, B. & G. Corbett. 2002. *The Slavonic Languages*. London & New York: Routledge. 827, 832-833.
- Farley, C. & D. Lister. 2007. Greater Toronto's language quilt. *Toronto Star*. Dec. 30, 2007.
- Keefe, S. & A. Padilla. 1987. *Chicano Ethnicity*. Albuquerque: UNM Press.
- Labov, W. 1984. Field methods of the project on linguistic change and variation. *Language in use: Readings in sociolinguistics*, ed. by J. Baugh and J. Sherzer, 28-53.
- Lyskawa, P. 2015. Variation in case marking in Heritage Polish. MA Thesis, Linguistics Department, University of Toronto.
- Lyskawa, P., R. Maddeaux, E. Melara & N. Nagy. *submitted*. Heritage speakers follow all the rules: Language contact and convergence in Polish devoicing. *Heritage Language Journal*.

Nagy & Łyskawa / LSA 2016

References, p. 2

40

- Nagy, N. 2009. Heritage Language Variation and Change. http://individual.utoronto.ca/ngn/research/heritage_lgs.htm.
- Nagy, N. 2011. A multilingual corpus to explore geographic variation. *Rassegna Italiana di Linguistica Applicata* 43.1-2:65-84.
- Nagy, N., J. Chocieł & M. Hoffman. 2014. Analyzing Ethnic Orientation in the quantitative sociolinguistic paradigm. In L. Hall-Lew & M. Yaeger-Dror. Special issue of *Language and Communication: New perspectives on the concept of ethnolect*. 35:9-26.
- Nagy, N. & M. Meyerhoff 2008. The social life of sociolinguistics. In *Social Lives in Language: Sociolinguistics and Multilingual Speech Communities*, M. Meyerhoff & N. Nagy (eds), Amsterdam: John Benjamins. 1-17.
- Poplack, S. 1980. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching. *Linguistics* 8:581-618.
- Statistics Canada. <http://www12.statcan.gc.ca>.
- Torres Cacoullous, R. & C. Travis. 2011. Testing convergence via code-switching: Priming and the structure of variable subject expression. *International Journal of Bilingualism* 15.3: 241-267.

Nagy & Łyskawa / LSA 2016