TITLE: **Analyzing Ethnic Orientation in the Quantitative Sociolinguistic Paradigm**

AUTHORS: Naomi Nagy[a], Joanna Chociej[a], Michol F. Hoffman[b]

[a]Department of Linguistics
University of Toronto
Sidney Smith Hall, 4th floor
100 St. George St.
Toronto, Ontario CANADA M5S 3G3

naomi.nagy@utoronto.ca, joanna.chociej@mail.utoronto.ca

[b] Department of Languages, Literatures & Linguistics
York University
4700 Keele Street
Toronto, Ontario CANADA, M3J 1P3

mhoffman@yorku.ca

CORRESPONDING AUTHOR: Naomi Nagy +1 (416) 978-1767

ABSTRACT

Ethnic Orientation, defined as speakers' sociolinguistic practices and attitudes, does not affect all communities, languages, or linguistic variables equally. We illustrate that the types of differences that emerge depend on methodological decisions, particularly at the analysis stage. We provide examples of inter-community differences including some that emerge differently depending on the method of analysis. This is accomplished by comparison of Heritage Language patterns among groups of Toronto residents: speakers of Heritage Cantonese, Italian, Russian, Ukrainian and Polish; and English patterns in Chinese-descent and Italian-descent Torontonians, comparing across three generations since immigration. We examine the variables pro-drop and VOT in the Heritage Language data. The Canadian Vowel Shift and consonant cluster simplification are examined in English. We show that *no* Ethnic Orientation facets correlate to all types of linguistic variation. The relationships found between linguistic variables and Ethnic Orientation variables suggest Ethnic Orientation as a key factor in modeling variation in Heritage Language communities – their variation should not be attributed solely to subtractive processes like incomplete acquisition or attrition.

KEYWORDS

1. language contact

2. variationist sociolinguistics
3. Heritage Language
4. Canadian English
5. ethnic orientation
6. Toronto

1. Introduction

Many speakers who control two or more languages may use features of one language while speaking the other. We expect that speakers' patterns of language use and attitudes toward each language that they speak (and toward the speakers of these languages) will influence the extent of this type of language mixing. When such contact effects are examined in the majority language (e.g., English in Toronto), the concept of ethnolect, or ethnically-indexed variety, may be called into play. However, when we examine minority language use within each ethnically-specified community, e.g., the use of Heritage Italian in Toronto, the ethnolect concept is not applicable – we are not looking for distinctions within a language that index different ethnicities. It is, however, fruitful to examine contact effects in both directions in tandem in order to best understand the multi-faceted nature of linguistic identity-construction in multilingual communities. Attitudes and patterns of language use are expected to play roles in both of a speaker's varieties. Indeed, Noels (this issue) notes that the very choice of language in which a survey is administered may influence how a speaker responds to a survey. This highlights the importance of considering each linguistic facet of a multilingual speaker in turn: identity may be constructed differently in each language. In this paper, we compare sociolinguistic variation in English and Heritage languages in several Toronto communities.

We refer to attitudes toward a language and its speakers as "Ethnic Orientation," which Noels (this issue) defines as "[…] affective, behavioural and cognitive constructs pertaining to a person's engagement with their ancestral ethnic group […]". In this paper, we compare how Ethnic Orientation measures correlate with linguistic variation in Heritage Languages and English. We show that Ethnic Orientation does not affect all communities, languages, or linguistic variables equally: there is no short, simple set of Ethnic Orientation "facts" that reliably correlate to all types of linguistic variation. We illustrate the types of differences that emerge depending on decisions made by the researchers at each step of analysis. The fact that different analytic methods provide different pictures of Ethnic Orientation effects

highlights the importance of critically evaluating one's methodology throughout the research process. We provide examples of these effects by exploring the role of Ethnic Orientation in variable linguistic behavior through a number of sociolinguistic comparisons, summarized in (1).

(1) **Cross-ethnicity comparisons**

We examine the linguistic behavior of groups of speakers who represent different ethnicities in Toronto. This includes comparisons of Heritage Language patterns among Cantonese, Italian, Russian, Ukrainian and Polish speakers; and English variation among Torontonians of Cantonese Chinese descent and Italian descent.

Cross-generational comparisons

We examine the Ethnic Orientation and linguistic behavior of first generation immigrant speakers versus second (and third) generation speakers.

Cross-linguistic comparisons

We examine the linguistic behavior of speakers in their two languages. This includes comparisons of, for instance, reflections of Ethnic Orientation in Heritage Cantonese to reflections of Ethnic Orientation in the English of the Cantonese community. This also includes comparisons between instantiations of Ethnic Orientation in English in the Cantonese communities versus in English in Toronto more broadly.

Cross-variable comparisons

We contrast effects of Ethnic Orientation on the variables pro-drop and Voice Onset Time in the Heritage Language data. The Canadian Vowel Shift and consonant cluster simplification are examined in English.

This project, investigating different instantiations of Ethnic Orientation via the comparisons described in (1), illustrates that different facets of EO behave differently according to language, community, generation, and linguistic variable within which they are studied. §2 describes the methods used to collect data in order to make these comparisons. We then explore two questions:

- Which aspects of speakers' sociolinguistic behavior, attitudes and perceptions contribute to the construction of ethnic orientation? (§3);
- Which aspects of speakers' ethnic orientation contribute to linguistic variation (§4)?

Before linguistic variation can be properly analyzed, we must resolve which aspects of Ethnic Orientation are independent, robust and relevant, as there are many potentially interacting and overlapping factors to disentangle before they can be introduced into a variationist model. Only then can we explore and develop methods of sociolinguistic analysis that may yield consistent cross-linguistic effects. The specification of generalizable principles about contact-induced language change, particularly with respect to identity and social practices, is at the heart of the research projects in which we are engaged.

One interesting finding is that, in general, factors relating to childhood accounted for more of the variance among speakers in the Ethnic Orientation data than factors relating to the participants' current practices. As Fix (this issue, p. TBD) summarizes, "research in language and dialect acquisition [provides] evidence of the import of exposure to a language or variety before the critical period of early adolescence for its acquisition [… although there] is also evidence that more recent social relations also strongly conditions linguistic behavior." We note, however, that there is not much difference in the effects of childhood vs. adult practices on *linguistic* variation.

A second important point is the different picture that emerges when univariate (correlation) vs. multivariate (regression) methods are applied: when we look at each social factor independently, seeking correlation to *rates of usage* of a particular linguistic variant, no consistent sociolinguistic patterns are revealed. In contrast, multivariate regression analysis does reveal certain sociolinguistic patterns.

Our goal is to understand how Ethnic Orientation operates as a predictor of linguistic variation across languages, communities, generations and variables. To effect such comparisons, a consistent method of organizing our understanding of Ethnic Orientation across languages, communities, and generations is needed. We therefore first operate on each corpus as a whole, in order to see how best to distill the responses to our EO

questionnaire. §3 describes the approach. In §4 we turn to analysis of linguistic variation. At that stage, we model our EO data as described in §3, but using Ethnic Orientation responses only for the speakers who provided the linguistic data for each variable.

2. Data collection methods

Toronto provides an ideal site for this kind of research. It is a multicultural city, where only 54% of residents report English as their (only) mother tongue (Statistics Canada, 2012). We took our data from two ongoing projects that look at the effect of ethnic orientation on the languages spoken by Toronto residents: two distinct samples of the same population. The first sample is from the Contact in the City project (Hoffman, 2010, Hoffman and Walker, 2010) and the second from the Heritage Language Variation and Change (HLVC) project (Nagy, 2009, 2011). Hoffman and Walker's project looks at the effect of ethnic orientation on variation in English. The corpus to date includes recordings of Torontonians of Chinese, Italian, Portuguese, Greek, Punjabi and British-Irish descent. The HLVC project examines linguistic variation in the Heritage varieties of the following languages: Cantonese, Faetar, Italian, Korean, Polish, Russian and Ukrainian speakers who are long-time/life-time residents of Toronto.[1]

We use the term "Heritage Language" in Rothman's (2007, p. 360) sense: "Like all monolingual and childhood bilingual learners, heritage speakers are exposed naturalistically to the heritage language; however, this language is by definition a

[1] To constrain variation in the input variety, the place of origin for each group was limited as follows:

Heritage Language	Place of origin
Cantonese	Hong Kong
Italy	Calabria
Korean	Seoul
Polish	eastern or southern Poland
Russian	Moscow or St. Petersburg
Ukraine	Lviv

nonhegemonic minority language within a majority-language environment"
In this paper, we compare data from Toronto speakers of six Heritage Languages: Cantonese, Korean, Russian, Ukrainian, Polish, and Italian; and English as spoken by residents of Cantonese and Italian ancestry. Participants are self-defined fluent mother tongue speakers of one of these Heritage Languages and belong to a social network inhabited by their interviewer(s). The interviewers were Heritage Language speakers who are part of the community in which they conduct interviews. Most were university students. To the extent possible, the same pair of interviewers interviewed all speakers within one community.

Tables 1 and 2 show the number of speakers in each sample, according to language and type of data (linguistic variables, described in §2.2, or Ethnic Orientation). The Heritage Language speakers (Table 1) range from 16-89 years old. The English speakers (Table 2) range from 17 to 70; the second and third generation speakers, the focus here, are mostly in their 20s and 30s. Throughout this paper, "generation" distinguishes immigrants (first generation) vs. immigrants' children (second generation) vs. immigrants' grandchildren (third generation), not age. It will be seen that effects of Ethnic Orientation correlate with linguistic production in crucially different ways depending on the immigrant generation of the speaker, as noted in, for example, Hall-Lew and Starr (2010) and Del Torto (2010).

The first row of Tables 1 and 2 show how many participants are included in the overall analysis of Ethnic Orientation, which was designed to create a method of distilling the EO responses similarly across groups of speakers and variables, in order to be able to compare the relationships between Ethnic Orientation and linguistic variation. The last two rows of Tables 1 and 2 indicate the number of participants included in analysis of each linguistic variable. At this stage of sociolinguistic analysis, described in §4, we modeled our EO data as described in §3, but using EO responses only for the speakers who provided the linguistic data for each variable.

Table 1: Heritage Language data by community (number of speakers in corpus)[2]

Variable	Cantonese	Italian	Russian	Ukrainian	Polish	Korean
Ethnic Orientation	34	23	10	32	13	15
VOT	--	11	11	12	--	--
pro-drop	16	11	12	--	13	--

Table 2: English language data by community (number of speakers in corpus)

Variable	Cantonese	Italian
Ethnic Orientation	30	26
(t/d)-deletion	16	18
Canadian Vowel Shift	8	8

2.1. Ethnic Orientation data

We investigated the extent to which speakers orient themselves toward their ethnic and linguistic heritage and analyze the relationship between linguistic variation and these practices and attitudes. The sociolinguistic toolbox does not include a widely-used survey of Ethnic Orientation, as can be easily observed by comparing the papers collected in this volume or in *English Today* 26.3. Therefore, we turned to the fields of anthropology and psychology and chose a tool that had been extensively tested in a five-year project that examined several Chicano communities in the United States (Keefe and Padilla, 1987). Like us, Keefe & Padilla (1987, p. 2)

> wanted to determine fairly precise ways of measuring cultural knowledge and ethnic identification, which would describe the ethnic population and its internal variation as well as accurately plot changes over time, especially from generation to

[2] For some speakers, we have linguistic data but not EO data, while for others the reverse is true. Correlations between the linguistic and EO patterns were calculated on the overlapping subset.

> generation [… For this purpose, they constructed] two scales, ***Cultural Awareness*** and ***Ethnic Loyalty***, respectively. Each scale was constructed in a multidimensional fashion, and variation in survey scores within the Mexican American population demonstrate[d] the inaccuracy of stereotypes emphasizing ethnic homogeneity. Nevertheless, certain general trends can be observed in the ethnic group as a whole. The most important trend is the obvious and gradual decline in awareness of Mexican culture from generation to generation, compared to the greater tenacity of ethnic identity.

Also resembling our interests, their investigation was designed to learn:

> What kinds of variation in these patterns exist within the ethnic population? What factors contribute to the separation or assimilation of Chicanos in American life? Why does ethnic persistence and/or change occur?" (*ibid,* 3)

Their project began by defining five cultural spheres (*language familiarity and usage, cultural heritage, ethnic pride and identity, ethnic interaction, interethnic distance* and *perceived discrimination*), each containing features of both ethnic awareness and ethnic loyalty (*ibid,* 47). These were investigated via a 185-part questionnaire administered to participants of five generational groups. They scaled each item in the direction of Mexican cultural awareness and ethnic loyalty, meaning that a high score reflected awareness of or loyalty to the Mexican culture and people" (*ibid,* 47). We also follow this practice. A factor analysis was conducted on these responses to reduce the questionnaire further. Like us, they found that some features are better preserved than others, so [a] multidimensional approach is well-motivated (*ibid,* 7). They also report a lack of correlation between subsets of questions (about acculturation, identification, assimilation) (*ibid,* 23).

The sets of questions regarding social networks and cultural practices reflect sociolinguists' proposals that social networks and patterns of interaction influence linguistic variation (cf. Milroy, 1980, Pierrehumbert, 2006), and Speech Accommodation theory (Giles and Powesland, 1975) and Audience Design (Bell, 1984), the idea that speakers shift speech style to be more like their audiences.

Our version of the questionnaire consists of the 37 questions that Keefe and Padilla found to be most useful in their study and cover most aspects of ethnic identity and practices that have been proposed as relevant to linguistic variation.[3] Questions relate to aspects of ethnic identity such as language use, make up of social network, participation in community activities, attitudes toward cultural heritage, and discrimination.[4] These questions were asked as part of the sociolinguistic interviews from which the linguistic data were extracted. They are open-ended and reflect the self-reports of the individual speakers, in an effort to move away from researcher-imposed categorization and the methodological problems associated with Likert scales (Noels, this issue). In order to enable quantitative analysis of the results of the Ethnic Orientation questionnaire, speakers' responses to individual questions were assigned a score using a three point scale representing a greater orientation toward the English language and mainstream Canadian culture at one end (0 points) and greater involvement with the Heritage Language and heritage culture at the other (2 points).[5] In this way, speakers' responses position them along 37 continua of orientation and practice. The same scoring system was applied to the HLVC and the Contact in the City corpora. We discuss the efficacy of different ways of grouping and interpreting these responses in §3 and the relationship of these measures to linguistic variation in §4. Noels (this issue) notes the utility of quantified responses for comparative and archival purposes, but also the necessary trade-off of a less direct representation of the participants' assessment of their behavior (*ibid*).

2.2. Linguistic variables

The two linguistic variables we explored in the Heritage Language data are pro-drop (Nagy *et al.* 2011; Chociej 2011) and voice onset time (Hrycyna *et al.,* 2011). Pro-drop refers to the variable presence of an overt pronominal subject in a finite clause. English is prescriptively viewed as a non-pro-drop language: every finite clause requires an overt subject. In contrast, the Heritage Languages we examine allow pro-drop to varying degrees, making this a variable in which English contact is one possible source of variation. Italian and Polish are prototypical null subject languages, allowing null subjects in all persons and

[3] One type of practice that is not included is religion, which may have indirect effects through other factors.
[4] The full text of the questionnaire is online at the HLVC Project website (Nagy, 2009).
[5] See §3.3 for a discussion of alternate ways of scoring.

numbers. Cantonese is a radical pro-drop language, allowing null subjects in spite of the lack of verbal morphology indicating person and number. Finally, Russian is a partial pro-drop language, allowing null subjects in certain contexts in oral discourse.[6] Many studies have examined the patterns of variation in situations of contact between English and Heritage Languages spoken in the US, some finding contact effects (*e.g.*, Otheguy *et al.,* 2007; Polinsky 1995) and others not (*e.g.* Torres Cacoullos and Travis 2010). For this analysis, we analyzed approximately 100 tokens for each of 57 HL speakers(N=5,118). For further details on the analysis, and consideration of other conditioning factors, see Nagy *et al.* (2011) and Chociej (2011).

Voice Onset Time (VOT) is a continuous phonetic variable, referring to the duration of the interval between the burst of a stop and the onset of vocal fold vibration. It serves as an important acoustic correlate of voicing and varies considerably across languages (Caramazza & Yeni-Komshian, 1974, Flege, 1991, Fowler et al., 2008, Lisker and Abramson 1964). English has long-lag voiceless stops, with VOT generally longer than 30 ms, while the three Heritage Languages we examine—Italian, Russian and Ukrainian—all have short-lag voiceless stops, mean VOT shorter than 30 ms. We examined word-initial /p/, /t/ and /k/ followed by a stressed /a/ or /o/. We analyzed ~75 tokens for each of 34 HL speakers (N=2,515). For further details on the analysis and relationships with earlier VOT work see Nagy & Kochetov (2013).

The two linguistic features we considered in English are (t/d)-deletion, and the Canadian Vowel Shift (CVS). The first is a stable and well-studied process in which /t/ and /d/ in final position of word-final consonant clusters are variably deleted (*e.g.*, Guy, 1980). Its linguistic and social conditioning is relevant to speech-community membership (Guy, 1980; Tagliamonte and Temple, 2005), second-language acquisition and ethnic identity (Bayley,

[6] While it would be useful to provide pro-drop rates for the non-contact varieties (or source varieties) of the Heritage Languages we investigate, pro-drop rates depend on a number of contextual variables. As rough guidelines, Pustovalova (2011) reports overall null rates of ~50% for Moscow Russian and Rumpf & DiVenanzio (2012) report almost 90% for Southern (but not Calabrian) Italian, both in conversational corpora. Chociej (2011) reports null rates of ~80% for Polish based on data from two life-time residents of Poland. We know of no published studies of pro-drop rates in the other homeland varieties. Suffice it to say that pro-drop rates in all languages considered are quite a bit higher than the ~2% rate that has been reported for Anglo Canadian English (Nagy *et al.,* 2011).

1996; Santa Ana 1992, 1996; Wolfram 1969). The morphological and phonological conditioning of (t/d)-deletion are relevant loci for substrate transfer. 50-100 tokens were extracted per speaker and coded auditorily for the presence or absence of the second consonant in a word-final cluster, along with relevant independent linguistic (e.g., preceding and following segment, morphological category of the token word) and social factors (sex and the Ethnic Orientation factors discussed below). For details regarding analysis see Hoffman and Walker (2010).

The second variable examined in the English data is the Canadian Vowel Shift, in which the front lax vowels BIT and BET are variably shifted to phonetic realizations closer to [ɛ] and [æ], respectively, and the low front vowel BAT is variably retracted toward [a]. First documented by Clarke, Elms, and Youssef (1995), this change in progress has been studied in various locales across Canada (e.g., Boberg, 2004, 2005, 2008; Hagiwara 2006; Hoffman 2010; Roeder and Jarmasz 2008). These studies have identified several linguistic (e.g., following segment) and social factors (as above) conditioning the CVS, which we include in our analysis. We focus here on two vowels, BET and BAT. Starting fifteen minutes into the interview, approximately 100 tokens of each vowel were coded impressionistically as either shifted or non-shifted. As above, for further details on coding and analysis see Hoffman and Walker (2010).

Since all speakers self-identified as fluent speakers of the language in which they were interviewed, we do not make any effort to consider "incomplete acquisition" or L2 effects as explanatory, although such effects may have historically contributed to the construction of some variable patterns.

3.0 Ethnic Orientation quantitative analysis methods

We wish to compare the role of Ethnic Orientation measures in accounting for linguistic variation in Heritage Languages and English. Recall from §2.1 that our Ethnic Orientation data comes from responses to 37 questions from Keefe & Padilla's (1987) study of Chicano Americans. An analysis considering the effect of each of the 37 Ethnic Orientation questions individually seems unwieldy. This section therefore examines possibilities for reducing the number of variables representing Ethnic Orientation by eliminating redundancy and

interactions between those variables. This reduction process must precede introducing the factors into any multivariate model of linguistic variation. While we were originally hoping to find the best measures of Ethnic Orientation, one set that would reliably map onto linguistic variation, we found instead that different decisions made at each step of the analysis process led to different outcomes. We also learned that there is *not* a smaller subset of the Ethnic Orientation Questionnaire that could be used as a starting point: all questions played distinct roles in building the picture of a participant's Ethnic Orientation, and they combined in different ways for different groups of participants.

In the following analyses of Ethnic Orientation variables, we combined all HLVC data in one analysis, grouping together speakers from the various ethnolinguistic backgrounds. We did the same for the English language data (grouping the Italian Torontonians and the Chinese Torontonians), unless otherwise noted. This allows us to learn how the Ethnic Orientation Questionnaire responses reflect EO. In order to investigate the relevance of the Ethnic Orientation questions to the groups of speakers in our samples, we considered several ways of analyzing and grouping the responses, listed in (2), each of which is discussed in turn in the following subsections.

(2) Methods of combining responses to Ethnic Orientation questionnaire questions (at http://projects.chass.utoronto.ca/ngn/pdf/HLVC/short_questionnaire_English.pdf)
1. Average of all 37 questions (§3.1)
2. Subsets of questions, reduced by Principal Components Analyses: organized by Topic (Hoffman and Walker 2010 following Keefe and Padilla, 1987) or by Reference Group (Boyd, Walker and Hoffman, 2011) (§3.2)
3. Two methods of scoring items related to language use (Chociej, 2010, 2011) (§3.3)

After illustrating the *way* Ethnic Orientation is operationalized, we will turn to analyses that consider the relationship between Ethnic Orientation and linguistic variation (§4). In those sociolinguistic analyses, we used the data only for the speakers and language being considered, but we used it in the way illustrated in §3.

3.1 Ethnic Orientation Overall Average

As noted, incorporating the effect of all 37 Ethnic Orientation questions into an analysis would be unwieldy and the quantity of data available does not provide the power to do so in a multivariate analysis. Additionally, some speakers did not answer some questions. Since we seek a comprehensive measure of speakers' linguistic and cultural practices, one straightforward alternative is to assign each speaker a score corresponding to the mean of their responses to all (answered) questions in the Ethnic Orientation questionnaire (*cf.,* Hoffman and Walker, 2010, p. 10). This first method will be compared to the other methods described in the subsections that follow.

3.2 Subsets of questions

As Hoffman and Walker (2010, p. 11) note, ". . . we cannot assume that the responses to the […] questions in the Ethnic Orientation questionnaire are completely independent of each other, nor that each question or set of questions contributes equally to the mean Ethnic Orientation index." Therefore, we assess different groupings and methods of analysis of the Ethnic Orientation questionnaire responses, a process of comparison that is necessary to rule out interacting factors prior to introducing Ethnic Orientation factors into multivariate analysis.

By grouping the Ethnic Orientation questions into subsets, we reduced the problems associated with considering questions individually (e.g., lack of responses, weighting of different questions). We are also able to answer two questions: (1) Is it necessary to ask all questions regarding speakers' linguistic behavior and attitudes toward their ethnicity, or are some of them significantly correlated and therefore redundant? (Our findings suggest the former is true.) and (2) Do some questions reflect more meaningful differences for the Torontonians in our samples than others? (Yes.) We address these questions through Principal Components Analyses.

We used Principal Components Analysis (PCA) for two purposes: to eliminate redundancy and to determine the most meaningful questions in our set. PCA is useful as it reduces the number of independent variables (the 37 individual Ethnic Orientation questions) to a

smaller number of underlying components that account for the most variation. These analyses involved three stages:

1) We first divided the Ethnic Orientation questions into subsets according to the content of the questions.
2) Each subset was independently submitted to a PCA to see which groupings of questions, within each subset, emerged. PCA shows whether the answers to one question predict answers to another (correlation) as well as which questions account for the most variance among the participants.
3) The significant components from the first PCA were used as input in a second PCA. This produced a still smaller number of components that represent the most meaningful variation for each corpus as a whole.

For Step 1, we compared two different methods for grouping the questions into subsets. The first, which we call the **Topic method**, is based on the type of behavior or topic to which each question refers. For example, every question referring to language choice was grouped together (C1, C2, C3, C4 and C5 in the Ethnic Orientation Questionnaire), with the intent of understanding which practices or attitudes best represent Ethnic Orientation (cf. Eckert 2000 on communities of practice). This is based on Keefe and Padilla's (1987) original approach. A detailed account of its application to variationist data is found in Hoffman and Walker (2010). This method creates eight groups of questions.

The other method used in Step 1 is the **Reference Group method**. In contrast to the Topic method, it groups the questions according to the people referred to, with the intent of determining the extent of the role of various types of social networks (cf. Milroy, 1986) in Ethnic Orientation. For example, one group is questions about the participants' family. This includes questions about the family's language use and ethnic orientation (C1, C4, C5, E1, E2, E3, F1, F2 and F3). Another group includes similar questions but focuses on the participants' friends. Boyd *et al.* (2011) apply this method in their sociolinguistic analysis. This method creates seven groups of questions.

The two methods create two distinct ways of grouping the 37 original questions, although there are some questions which are grouped identically (e.g., Questions H1-H5, about Discrimination).

Step 2 applied a Principal Components Analysis to each of the groups we created in Step 1, in order to determine how answers to the different questions are related. From this we learn which answers predict particular answers to other questions in the same group, as well as which questions best distinguish the participants. For an example, in the Heritage Language side of the table in Appendix A, Component 10 indicates that Questions G1, G2 and G3 (about attitudes toward the Heritage culture) group together – how people answered one of these questions correlates well to how they answered the other two. In contrast, Questions H1-H5 do not all group together: while people's experience of discrimination in housing and work are correlated to each other (both are in Component 12), they are not correlated to their perception of General Discrimination against their ethnic group (Component 8). Turning to the English side of the same table, Discrimination at school did not appear at all – responses to that question did not serve to differentiate participants. (This is because very few have experienced discrimination at school, so virtually everyone gave the same answer).

From Step 2 we learned that a number of factors contribute to Ethnic Orientation in both the Heritage Language and English corpora: participants' self-identification, language choice, factors relating to participants' families, their attitudes toward their heritage culture and perceptions of discrimination all emerge as significant predictors of Ethnic Orientation (see Appendices A and B). However, some factors emerged as significant in the Heritage Language corpus but not the English corpus (and vice versa).[7] Additionally, this analysis established a lack of correlation between all listed factors. This meant that we could proceed with the linguistic analysis confident that these factors contribute separately to any observed linguistic variation.

[7] These differences may be due to different demographics of the two samples or to sample size.

In Step 3, the set of components resulting from Step 2 that were determined to be significant (given in Appendices A and B) were all fed, unweighted, into a final PCA. This allowed us to compare the components from the different groups of questions in order to understand, for example, whether one can predict how people respond to a question about language use at work by how they answer a question about perceived discrimination in the workplace. Analytically, this further reduces these still large sets of variables to a smaller number of underlying factors that account for the most variance among speakers. This reduced set is used in the analyses of linguistic variation.

We conducted two PCAs in this step: one for the Topic method of grouping questions and one for the Reference Group method. Table 3 shows the results of PCAs conducted using the Topic method and Table 4 shows the results of PCAs using the Reference Group method. In these tables, we list the significant components accounting for the variation in our speakers' responses to the Ethnic Orientation Questionnaire, alongside the percentage of variance each accounts for. Each component represents factors that cluster together: speakers' responses are similar across these factors. In each analysis, the first component accounts for the most variation, and subsequent components are weaker. The values beside the clustered factors represent their strength: higher values contribute more to the component.[8] The different components are orthogonal – how participants respond to questions in one component is independent of how they answer questions grouped into a different component. These tables also compare components between the two corpora.

Factors in bold appear in the same component in both columns, indicating that these factors are relevant to both the speakers interviewed in English and those interviewed in their Heritage Language, and that they operate in the same fashion. Italicized factors appear in both columns, but in different components, indicating that the factor is, again, relevant in both the English and the Heritage Language data, but that it interacts differently with other factors. The large number of factors that are bold or italic illustrates the

[8] Negative values indicate inverse relationships between the factors within a component. For example, in Table 5's Heritage Language column, Component 2, there is an inverse relationship between parents' ethnic identity and language use and perceptions of discrimination against their ethnic group.

similarity in the set of factors that determine the Ethnic Orientation of members of the two corpora.

Although it seemed, *a priori*, that it would be important to separate individuals' characteristics from social network characteristics, Tables 3 and 4 show that the "Topic" method produces greater consistency between the two corpora than the "Reference Group" method. This suggests that better overall replicability may be found using the original groupings of questions proposed in Keefe & Padilla (1987). For example, greater consistency can be seen in the PCA for Ethnic Orientation, grouping questions by Topic (Table 3), where all the factors in Component 1 for the Heritage Language corpus (Birthplace, Language choice and Language preference) also appeared together in Component 1 in the English corpus. In contrast, no such similarities can be found in the PCA in Table 4, where questions were grouped by Reference Group.

Table 3: Results of PCA for Ethnic Orientation, grouping questions by Topic (Step 3)

Heritage Language Corpus	Var.	English Corpus	Var.
COMPONENT 1	18%*		35%
Birthplace	**.655**	**Birthplace**	**.740**
Language choice	**.640**	**Language choice**	**.920**
Language preference	**.632**	**Language use and preference**	**.912**
		Parents	*.696*
		Partner	.765
		Ethnicity of social network	.608
COMPONENT 2	15%		16%
Parents' Ethnic Orientation and language	*.701*	*Grandparents' age of arrival*	*.725*
Perception of general discrimination	*–.697*		
COMPONENT 3	11%		11%
School and personal discrimination	.778	*Perception of general discrimination*	*.596*
Cultural attitudes	.598		
COMPONENT 4	10%		10%
Experience of economic discrimination	**.880**	**Experience of economic discrimination**	**.779**
COMPONENT 5	8%		
Grandparents language use and age of arrival	*.887*		

* In Tables 3 and 4, percentages for each component indicate how much variance is accounted for by that component. Decimal numbers indicate the strength of the factors contributing to the component: the higher the number, the greater the effect of the social factor within the component. Negative numbers indicate an inverse relationship with the other social factors within a component.

Table 4: Results of PCA for Ethnic Orientation, grouping questions by Reference Group (Step 3)

Heritage Language Corpus	Var.	English Corpus	Var.
COMPONENT 1	17%		36%
Grandparents' language choice and participant's language choice with friends	-.857	*Family's language use* and *parents' identity*	.822
Participant's birthplace and contact with country of origin	.624	Participant's ethnic identification	.722
		Participant's cultural attitudes	.713
		Participant's experience of general discrimination	.656
		Participant's social network ethnicity	.611
COMPONENT 2	13%		20%
Participant's cultural attitudes	.712	*Grandparents' age of arrival*	–.836
Participant's experience of personal discrimination	.636	Partner's identity and language use	.779
		Participant's birthplace and contact with country of origin	.719
		Participant's language use and preference	.704
COMPONENT 3	11%		12%
Participant's social network ethnicity	.751	*Participant's experience of housing discrimination*	.916
Family's language use	.682		
COMPONENT 4	10%		
Participant's experience of housing and job discrimination	.826		
COMPONENT 5	8%		

Parents' identity, language use, age of arrival in Canada	.817
Participant's perception of general discrimination	-.666
COMPONENT 6	8%
Co-workers' ethnicity	*.856*

3.3 Scoring Language Use questions

So far, we have discussed only one way of assigning scores to Ethnic Orientation Questionnaire responses. However, there is another way that responses can be mapped to scores. In this section we compare two models for scoring, which we call "Orientation Continuum" and "Language Mixing." We applied these two models to the subset of the questionnaire reproduced in (3) below.

In the method that has been employed so far, Ethnic Orientation is considered as a continuum, and how strongly a participant identifies with their ethnic group is privileged in calculating the Ethnic Orientation score. This method is based on an expectation that people with a stronger orientation to their heritage culture may maintain a more conservative variety of their Heritage Language, while those more accepting of mainstream culture may more readily innovate, particularly in adopting influences from English. Similarly, in English, speakers with different orientation to their heritage culture may differently make use of sociolinguistic variables. This concept is explored in Becker (this issue), Fix (this issue), and Wong and Hall-Lew (this issue), who all consider potential trade-offs between regional and ethnic indexing.

To operationalize this method, speakers' responses were placed on a scale with greater orientation toward the English language and mainstream Canadian culture at one end, and greater orientation toward the Heritage Language and culture at the other end. We coded responses on a three point scale: if the response indicated greater orientation toward Canadian culture, it was assigned a 0; if the response indicated greater orientation toward the heritage culture, it was coded as 2 points. A mixed response that fit somewhere in the

middle of the scale received 1 point. Thus, speakers fall in one of three positions along a continuum from heritage-culture oriented to mainstream-culture oriented. We refer to this way of scoring responses as the **Orientation Continuum** model.

In contrast, the **Language Mixing** model highlights linguistic behavior by looking at how much participants report being in contexts in which their two languages come into contact. This method was implemented because of the possibility that how much one language affects another in an individual's grammar depends on how often the languages are in active use at the same time. A binary coding of speakers' responses to the questions that deal specifically with which language is used in various contexts contrasts a tendency to keep the two languages separate with a tendency to mix languages. By "mixing languages" we mean that a speaker (reports that she) often uses both English and the Heritage Language in the same situations and/or often combines engagement in their heritage culture and Canadian culture. An example would be a speaker whose main social group typically uses both the Heritage Language and English in interactions. By "separating languages" we mean attitudes and behavior that reflect an effort to keep one's two languages/cultures distinct. An example would be someone who speaks only the Heritage Language at home and only English at work. In this model, if the speaker's response indicated that only one language (either English or their Heritage Language) was used in a given context, the response was assigned 0 points; if the response indicated the two languages were often mixed, it was assigned 1 point.

Comparing the effects of the Orientation Continuum and the Language Mixing models allowed us to determine whether it is a person's orientation or a person's linguistic environment that better correlates to sociolinguistic variation. To contrast these two models, we analyzed responses to the subset of Ethnic Orientation Questionnaire questions that specifically inquire about language use, as well as the response to the first, most general question about ethnic self-identification. These are listed in (3).

(3) Topics considered in the Orientation Continuum and the Language Mixing methods

1. Speaker's ethnic self-identity
2. Language currently used at work and/or school

3. Language currently used at home (with parents or partner)
4. Language currently used with friends
5. Language used at school in childhood
6. Language used at home in childhood (with parents and/or siblings)
7. Language used with friends in childhood
8. Current language preference

Scores to these eight questions were calculated once using the Orientation Continuum model and once using the Language Mixing model. Figure 1 contrasts the mean Language Use scores, calculated by each of these methods, for speakers grouped by language and generation. Note that, in the figures on the right side, which represent the Language Mixing Method, a high score does not mean either more Heritage Language use or more English – just that the speaker reported mixing their two languages. In contrast, a high score in the figures on the left indicates strong orientation to the Heritage Language/culture.

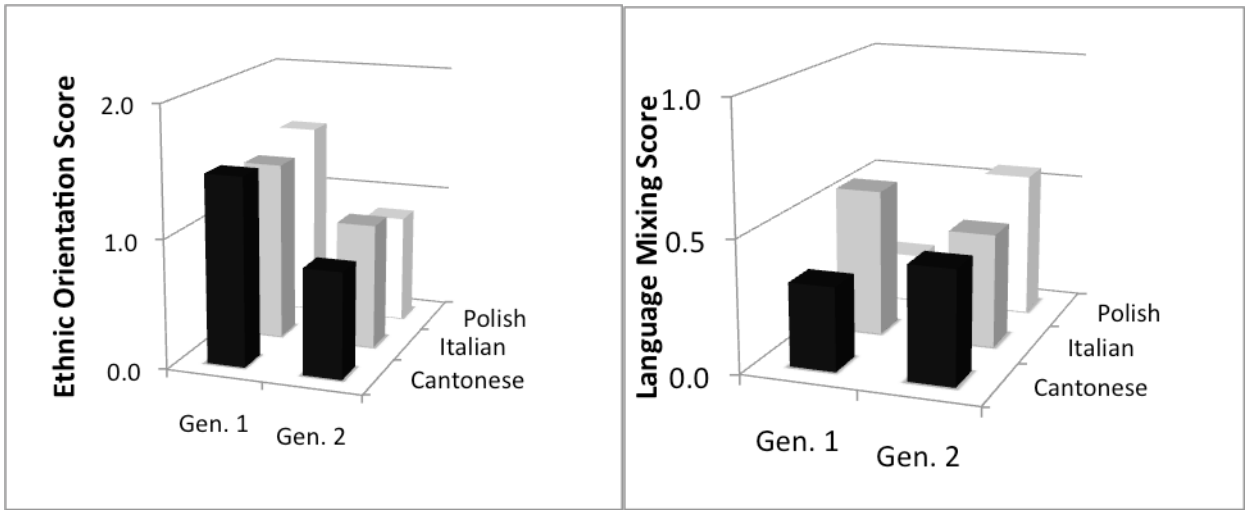
Figure 1: Mean Orientation Continuum scores vs. mean Language Mixing scores

Orientation Continuum Method **Mixing Method**

0 = Canadian/English orientation *0 = No mixing of languages*

2 = Heritage orientation *1 = Much mixing of languages*

Heritage Language Corpus



English Language Corpus

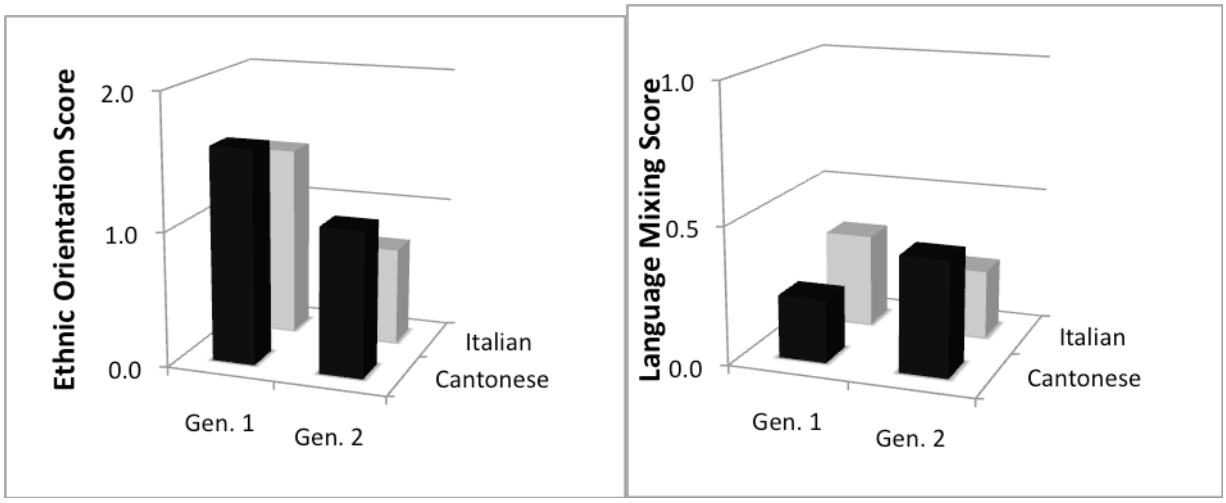


Figure 1 illustrates that the Orientation Continuum method systematically reflects generational differences: it shows a consistent cross-language pattern. In all languages, first generation speakers are more heritage-oriented (have a higher score) than second

generation speakers. This method produces greater consistency and parallelism across the different languages studied. The Language Mixing method results, shown on the right in Figure 1, are less consistent across languages, reinforcing the analytical differences in the two measures. This method gave more "new" information about the speakers, *i.e.*, it does not simply restate the generation the speaker belongs to. In this model, Italian speakers behaved differently from Cantonese or Polish speakers. Differences in behavior between different heritage groups are not unexpected since each group has different histories, experiences, and settlement patterns.

Before introducing these two models into any sociolinguistic analysis, we examined the relationship between the responses to the eight topics listed in (3) above, again with an eye to reducing the number of factors and eliminating redundancies. We performed two PCAs, using only Heritage Language corpus data: one using questionnaire response data scored using the Orientation Continuum method; and one using questionnaire response data scored using the Language Mixing method. Table 5 gives a summary of the Ethnic Orientation questions that made up each component of the Orientation Continuum method and the Language Mixing method. The amount of variance in the average Ethnic Orientation score that each component accounted for is given in the column labeled "Var."

Table 5: Comparison of two methods of coding Language Use Questions, showing the percentage of variance accounted for by the first four components (from PCA), for the Heritage Language Corpus

Orientation Continuum method	**Var.**	**Language Mixing method**	**Var.**
COMPONENT 1	21%		23%
Q7: language used with friends in childhood		Q5: language used at school in childhood	
Q1: speaker's ethnic self-identity		Q7: language used with friends in childhood	
		Q6: language used at home in childhood	
		Q1: speaker's ethnic self-identity	
COMPONENT 2	19%		17%
Q6: language used at home in childhood		Q6: language used at home in childhood	
Q3: language currently used at home		Q8: current language preference	
Q5: language used at school in childhood			
COMPONENT 3	18%		16%
Q4: language currently used with friends		Q3: language currently used at home	
Q8: current language preference		Q2: language currently used at work/school	
COMPONENT 4	14%		16%
Q2: language currently used at work/ school		Q4: language currently used with friends	

Both methods of distilling the scores of questions related to Language Use share one outcome: the components that account for most of the variance reflect aspects of the speakers' childhood, while those that account for less variance reflect speakers' current practices.

The final step in preparing these two ways of measuring language use for insertion into models of linguistic variability was to create categorical factors from these continuous measures. (This allowed us to compare the models to previously conducted logistic regression models.) For this purpose, the following conversions were made. For the Orientation Continuum method, each speaker's score for each component were converted to one of the three categories shown in the left side of Table 6. Conversions for the Language Mixing method are shown on the right. These cut-offs divide the responses into

fairly even sized bins. These categorical variables will be introduced into models of linguistic variables in §4.2.2.

Table 6: Categorical division of component scores, for two methods of scoring Language Use responses

Orientation Continuum component scores			**Language Mixing component scores**		
Definition	*Component*	*Score*	*Definition*	*Component*	*Score*
more English Language oriented	< 1	0	two languages kept distinct in usage	< 0.5	0
	1 - 1.49	1	two languages often mixed.	≥ 0.5	1
more Heritage Language oriented	≥ 1.5	2			

In §3 we have developed a number of ways to convert participants' responses to numerical scores. It is evident that different methods produce different outcomes in terms of degree of similarity across the groups of speakers. The Topic method of grouping Ethnic Orientation questions provides more consistent outcomes than the Reference Group method: here scores are monotonically related to generation. But we cannot say that either is inherently better or more reflective of Ethnic Orientation than the other. Therefore, we next try out each of the methods in analysis of the linguistic variation that we are hoping to account for, in order to determine their explanatory value.

Returning to the questions with which we began the section, we have illustrated that there is *not* a smaller subset of the Ethnic Orientation Questionnaire that could be used as a starting point: all questions play distinct roles in building the picture of a participant's Ethnic Orientation, and they combine in different ways for different groups of participants. The PCA analyses provided groupings of the questions ranked in terms of their ability to reflect meaningful differences for the Torontonians in our samples. We turn in §4 to applying these scores in efforts to account for the linguistic variation displayed by our speakers, both in English and in the Heritage Languages.

4. Effect of Ethnic Orientation on linguistic variables

Having determined several possible ways to reduce the information about Ethnic Orientation that we have for each speaker into a non-redundant subset of factors, we can investigate the role of Ethnic Orientation in variable linguistic behavior. Recall that the processes described above considered the responses from speakers of different Heritage Languages together, in order to develop a method that allows meaningful comparison across these groups. However, for the sociolinguistic analyses below, we applied these methods using only the data relevant to the particular analysis, i.e., the Ethnic Orientation scores of the speakers who produced the data for a particular linguistic variable. Again several methods of analysis are available. We compare two approaches for comparing ethnic orientation scores with linguistic data –correlations and multivariate regression analyses – by applying both to our linguistic variables.

4.1 Correlations

We calculated (Spearman's rho) correlations between Ethnic Orientation scores of individuals and their linguistic patterns: pro-drop and VOT in Heritage Languages, and (t,d) deletion and the Canadian Vowel Shift in English. For various subsets of speakers, correlations were calculated pair-wise for each possible combination of an Ethnic Orientation measure and the linguistic variable's rates.

The most noteworthy outcome is that, of the 168 analyses (14 analysis types x 12 subsets of the data) of the HLVC data, only seven produced significant correlations. Similarly, for the English corpus, of 66 analyses, four correlations were significant. The 11 significant correlations, listed in Table 7, is rather fewer than would be accounted for by chance, and do not appear to form a meaningful pattern. **An analysis method that looks at each social factor one at a time, seeking correlation to *rates of usage* of a particular variant, does not reveal consistent sociolinguistic patterns.**

Table 7: Only significant correlations (out of 168 tests on HLVC corpus data and 66 for the English corpus) [9]

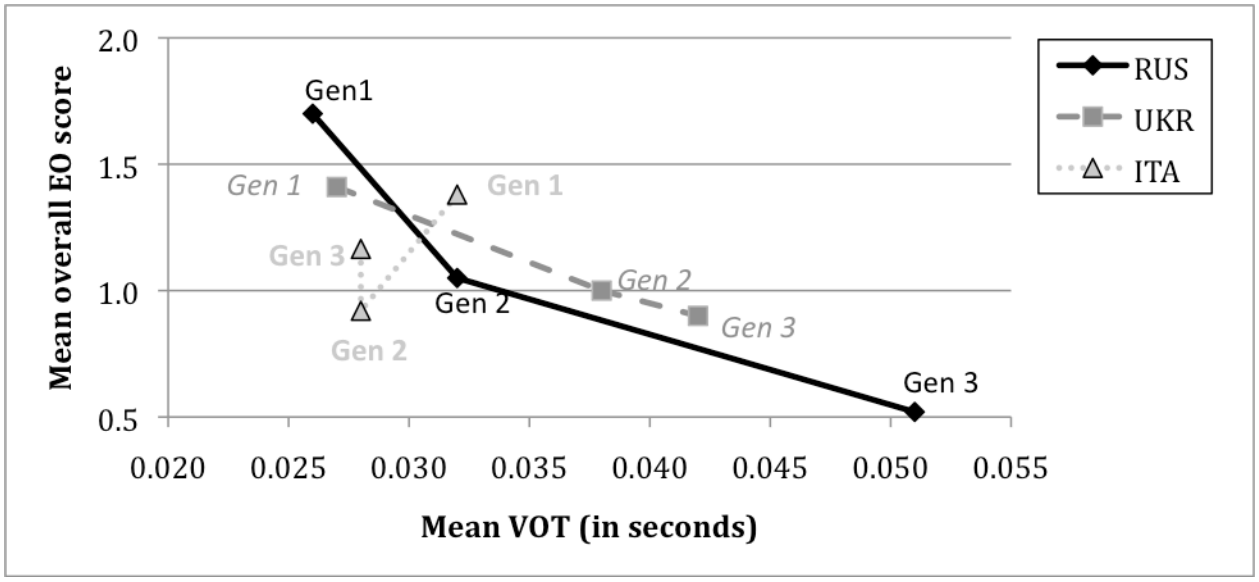
Corpus	Method	Group	Factor/Component	Variable
HLVC	Topic	All speakers	Birthplace; Language Use; Language Choice	VOT
	Topic	Cantonese First Generation Gen 1	Birthplace; Language Use; Language Choice	pro-drop
	Topic	Italian First Generation	Grandparents	VOT
	Reference	All speakers	Ethnicity of Personal Network; Family Language	VOT
	Reference	All speakers	EconDiscrim	pro-drop
	Reference	All Cantonese speakers	EconDiscrim	pro-drop
	Language Use	All Cantonese speakers	Language Mixing	pro-drop
English	Language Use	Italian Second Generation	Language Mixing	CVS (BET)[10]
	Language Use	Italian First Generation	Language Mixing	(t/d)-deletion
	Language Use	All Cantonese	Orientation Continuum	(t/d)-deletion

[9] It should be remembered that the component methods are based on components calculated from the combined Ethnic Orientation data within each corpus, rather than separately for each language or generation. This approach may be useful in comparing orientations across groups and ethnicities. However, for analysis of linguistic variation, it may be more appropriate to conduct separate PCAs for each ethnic group/generation. We plan to pursue this when sufficient data becomes available.

[10] This correlation is very weak.

In this context, it is interesting to note that Nagy and Kochetov (2013), using a slightly different approach (averaging scores for Ethnic Orientation responses within each Topic-method category of questions, rather than subjecting them to PCA for reduction), found that Heritage Russian speakers with higher EO scores, on average, produced shorter VOTs, more like Moscow Russian speakers, while those with lower EO scores had longer, more English-like VOTs. Ukrainian speakers also showed a correlation between shorter VOT and stronger orientation to Ukrainian culture, but we have no homeland VOT measurements to establish with certainty that the change is away from homeland norms. In contrast, the Italian speakers did not show an effect of Ethnic Orientation on VOT. Relatedly, a cross-generational change in progress was evident in Russian and Ukrainian, with VOT scores drifting upward toward English-like durations, but no such change is observable for Italian (see Figure 2). Considering these different outcomes, we speculate that Ethnic Orientation scores may correlate to linguistic patterns only when a linguistic variable is used indexically. That is, in the cases of Russian and Ukrainian, but not Italian, different immigrant generations distinguish themselves by VOT durations and so VOT measurements correlate to Ethnic Orientation scores. In no language do we find cross-generational change for pro-drop, so the lack of correlation between pro-drop rates and Ethnic Orientation here further supports our interpretation.

Figure 2: Average VOT on x-axis (higher values are more English-like) and average Ethnic Orientation score on y-axis (lower values are more English-oriented) for each generation, each Heritage Language (data from Nagy & Kochetov, 2013).



4.2 Regression analyses of the Heritage Language data

In contrast to the results for correlations, which must consider each possible conditioning factor in isolation, multivariate regression analyses fared better in finding relationships between Ethnic Orientation scores and linguistic variation. This approach allows us to consider ethnicity as a factor whose effect is mediated through other factors, such as generation (see Wagner, this issue). In this section we use two methods for performing multivariate regression analyses on the Heritage Language data: we conduct a series of Mixed Effects linear regressions (using the *lmer* package in R, R Core Team, 2012) with VOT as the continuous dependent variable. We use GoldVarb (Sankoff *et al*., 2005) to perform a logistic regression using pro-drop as the categorical variable. In each case, the linguistic and other social factors shown to be relevant in previous work (Hrycyna *et al*., 2011; Nagy *et al.,* 2011; Nagy and Kochetov 2013) are included as independent variables along with Ethnic Orientation scores.

4.2.1. Regression analysis of VOT in the Heritage Languages

We begin with VOT. Using Mixed Effects Models (MEM), we compared different ways of interpreting ethnic orientation data by including in successive analyses different types of independent variables to represent ethnic orientation: the 37 questions individually; components from Topic-based PCAs of the questionnaire responses (see factors listed in Table 3); components from Reference Group-based PCAs of the questionnaire responses (see factors listed in Table 4); and an average of the responses to all questions. Outcomes contrast the effects of the different analytic methods often applied to interpreting linguistic variation.

The MEM is a linear regression method that includes fixed effects (linguistic and social factors) and random effects for individuals and words (to "tease out" the possible effects of non-generalizable differences among individuals or words). The dependent variable (or response variable) was VOT duration, measured in seconds. The independent variables (or factors) included consonant measured (/p/, /t/ or /k/), following vowel (/a/ or /o/), generation and the Ethnic Orientation measure under consideration as fixed effects. Only the Ethnic Orientation measures are continuous variables, the other independent variables are categorical. Linguistic variables and generation were shown to have a significant effect in these data sets in Nagy & Kochetov (2013), in models without Ethnic Orientation scores, and are coded in the same way here, for comparability.

Table 8 shows (in the first row) the lack of a significant effect on VOT of the average of all 37 Ethnic Orientation Questionnaire responses for each speaker. Thus we turned to more complex methods of combining responses to the questionnaire. Components of the two methods of grouping factors (see Table 3: Topic Grouping Method and Table 4: Reference Grouping Method) were each individually added to a Mixed Effect Model containing consonant, following vowel, and generation as fixed effects, and word as a random effect. Speaker was not included as a random effect in the results reported here, as it correlates closely/categorically with the Ethnic Orientation scores in a sample with few speakers. Most components slightly but significantly improved the model, as indicated in Table 8. By this, we mean that the t-value for the added factor was >1.96 (Baayen, 2008, p. 270) and

the AIC, or Akaike number, for the model was smaller than the AIC value of the model with no Ethnic Orientation factors. Because the sample sizes are small and the improvements are moderate (changes in AIC average 3.5 points), we claim significant, but not strong, effects for these factors. For Russian, we did not have enough speakers who were measured for VOT and provided sufficient responses to the Ethnic Orientation Questionnaire to run.

Table 8: Significant Ethnic Orientation components for VOT (√ = significant*,* $p<0.05$*)*

Significant components	Italian (N=368)	Ukrainian (N = 477)
Average of all EO responses		
TOPIC GROUPING METHOD		
1 Birthplace, Language Use, Language Choice	√	√
2 Parents' ethnicity&Language Use; General Discrimination	√	√
3 Culture; Personal Discrimination	√	
4 Economic Discrimination		√
5 Grandparents	√	√
REFERENT GROUPING METHOD		
1 Grandparents&Language with Friends; Birthplace	√	√
2 Culture; Personal Discrimination	√	√
3 Ethnicity of Personal Network; Family Language		√
4 Economic Discrimination		√
5 Parent's Language&Age of Arrival; General Discrimination	√	
6 Ethnicity of Work Network	√	√

In order to check for effects of each question individually, we had to combine the data for the three Heritage Languages to create a large enough sample. When each of the Ethnic Orientation questions was run as an independent variable, in subsequent models, the questions listed with t-test values in Table 9 were identified as significantly predicting VOT. T-test values (for the factor in the output of the Mixed Effects Model) greater than 1.96 are shown, indicating the most promising factors. We provide t-test values to be informative; given the large number of comparisons (22) using the same data set, the t-test values are not high enough to indicate true significance – a repeated measures correction would be necessary. Adding each of those factors to the model improved its fit: the AIC value was lower than in the model with no Ethnic Orientation measures. Blank cells either had t-test values <1.96 or their AIC value was lower than the AIC value for the model without any Ethnic Orientation questions, indicating that that response did not improve the model's ability to predict VOT. "ND" indicates that there were insufficient responses to a given question to test.

Table 9: Significant Ethnic Orientation Questionnaire Questions for VOT, for a combined data set of Italian, Russian and Ukrainian (see questionnaire online at Nagy 2009)

Topic of Question	Ethnic Orientation Question # in model	t-value for this factor
Ethnic Identity	A1	2.048
	A2	2.484
	A3	
	A4	5.995
	A5	
Language	B1	2.707
	B1b	*ND*
	B2	1.073
	B3	
	B4	4.918
	B4b	4.698
	B5	3.01
Language Choice	C1	
	C2	*ND*
	C3	3.391
	C4	*ND*
	C5	*ND*
Cultural Heritage	D1	4.31
	D2	*ND*
Parents/Grandparents	E1	*ND*
	E2 Parents	
	E2 Grandparents	*ND*
	E3 Parents	*ND*
	E3 Grandparents	*ND*
Partner	F1	3.671
	F2	
	F3	3.671
	G1	10.738
	G2	*ND*
	G3	4.856
Discrimination	H1	
	H2	
	H3	*ND*
	H4	6.764
	H5	*ND*

At least one question in each topic of the questionnaire achieves significance, but until we have sufficient data to run such an analysis on the three languages separately, there is not

much interpretable in these results. The best evidence of an effect from Ethnic Orientation scores on VOT comes from the models in which the components of the PCA are introduced as predictors of VOT (Table 8).

4.2.2. Regression analysis of pro-drop in the Heritage Languages

We turn next to pro-drop as the dependent variable in a number of logistic regression analyses using GoldVarb. Each analysis evaluated the contribution of the ethnic orientation data to patterns of null-subject use in Cantonese, Italian and Polish. One goal of these analyses is to compare the two ways of interpreting responses to the Language Use portion of the Ethnic Orientation questionnaire—the Orientation Continuum method and the Language Mixing method (§3.3)—in order to see which method of scoring better matched the linguistic variation.

A number of regression analyses were conducted separately for each language. In each of these analyses, pro-drop was the dependent variable, and linguistic factors which had previously been determined to play a significant role (see citations in §2.2) were included as independent variables. Two analyses were conducted for each group of speakers: one with the Orientation Continuum score as an additional independent variable and one with the Language Mixing score. As a reminder, these are two ways of interpreting responses to questions about Language Use: the Orientation Continuum method gives high scores to participants who express strong Heritage language/culture identity and low scores to those who express strong English language/Canadian identity, while the Language Mixing method differentiates between participants who use both their languages in the same context(s) and those who report one-language-per-setting behavior. The discrete variable versions of these factors are introduced at the end of §3. Table 10 summarizes the significance and the direction of effect that each scoring method produced for each group of participants.

Table 10: Effect of language use scores on pro-drop in Heritage Language Corpus ("√" = significant effect in the predicted direction*)*

Language	**Generation**	**Group favoring more null-subjects**			
		Language Mixing		Orientation Continuum	
Italian	All	*n.s.*		*n.s.*	
	1st	*n.s.*		*n.s.*	
	2nd	√	less mixing		more English oriented
Cantonese	All	√	less mixing		peripheral
	1st	√	less mixing	√	more HL oriented
	2nd	*n.s.*			more English oriented
Polish	All		more mixing	√	more HL oriented
	1st	√	less mixing	√	more HL oriented
				√	more HL oriented
	2nd		more mixing		*(not a significant effect in model)*

Each grouping method is associated with certain logical hypotheses which predict the direction of effect. In Table 10, "√" indicates a significant effect in the predicted direction. In the Ethnic Orientation Continuum method, it is predicted that the more contact a speaker has with English speakers and mainstream Canadian culture, the lower their null-subject rate will be due to the low null-subject rate in English. Trends that do not follow this direction are unexpected. This includes cases where the highest null-subject rate was found with speakers who have extensive contact with English ("more English"), or where speakers who have equal contact with English and their Heritage Language had a higher or lower null-subject rate than both heritage-oriented and English-oriented speakers ("U-shaped"). In the Language Mixing method, it was predicted that speakers who use both English and their Heritage Language in the same situations ("more mixing") will exhibit the lowest null-subject rate. This contrasts with speakers for whom the two grammars never have the opportunity to influence each other because they exclusively use one language or

the other depending on the situation ("less mixing"). In all cases but one (the last cell of the table) where an effect is reported, the Ethnic Orientation factor significantly changed the model (determined by chi-square test on the difference in log-likelihoods of the model with and without that factor). That is, the predictability of pro-drop is improved by including a language use factor, although these factors do not always operate in the predicted direction.

Both methods have advantages and disadvantages. The Orientation Continuum method yielded significant effects for more speaker groups. Notably, the Ethnic Orientation score for second generation Cantonese speakers was significant when the Orientation Continuum method was used, but was not significant when the Language Mixing method was used. However, when considering whether the method yields a significant effect and whether this effect fits in with the logical hypothesis associated with the method, the Language Mixing method produced fewer significant results that contradicted its hypotheses. Nonetheless, the results show considerable differences between speaker sub-groups for both methods. Therefore, we cannot yet make predictions regarding which method will best account for linguistic variation across languages.

Analyses involving more fine-grained subgroups of the Ethnic Orientation questions yielded results that were more consistent across groups. In these analyses, each of the components from our PCAs in §3.3 became an independent variable. Participants' Ethnic Orientation scores became an independent variable in a regression analysis that included pro-drop as the dependent variable and the linguistic factors which had previously been determined to play a significant role (see citations in §2.2). Separate analyses were conducted for each language and each component. Table 11 summarizes the significance and direction of effects: "√" indicates the expected direction of effect (more orientation to the Heritage Language corresponds to more null subjects in the Orientation Continuum method, shown on the left side; less mixing corresponds to more null subjects in the Language Mixing method, shown on the right side); "X" indicates an opposite direction of effect; and "U" indicates a peripheral direction of effect, where mixed orientation behaves in contrast to both Heritage orientation and English orientation. Blank cells indicate no

significant effect of the Ethnic Orientation factor. All reported effects are from models that are significantly improved by adding this Ethnic Orientation factor, according to comparison of log-likelihoods (as for Table 10).

Table 11: Comparison of Orientation Continuum and Language Mixing methods in analyses of pro-drop in Heritage Languages

		Orientation Continuum components				**Language Mixing components**			
Language	**Gen.**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Italian	1st				√				X
	2nd	U	X	u	X		X	√	X
Cantonese	1st			√	√	√			√
	2nd		U	X	X				X
Polish	1st			√			√		√
	2nd	√	U	X	√	X	√	√	

Legend

blank = no significant effect

√ = expected direction of effect

X = opposite of expected direction of effect

U = peripheral (U-shaped) effect

As in the analyses of average Ethnic Orientation scores discussed at the beginning of this section, the Orientation Continuum method yielded a greater number of significant effects than the Language Mixing method: 15 significant effects for the former with 12 for the latter. However, if we only count the significant effects with the expected direction of effect, the two methods are comparable: six for the Orientation Continuum method and seven for the Language Mixing method. Furthermore, an interesting generalization emerged. For the Orientation Continuum method, Component 3 yielded the greatest number of significant effects; for the Language Mixing method, Component 4 yielded the greatest number of significant effects. As Table 5 shows, these two components both reflect current language practices with friends. In both cases, the components are significant for all groups of speakers, albeit sometimes in different directions. **This leads to the testable prediction that patterns of language use among friends, more than any other linguistic aspects of ethnic orientation, will account for patterns of Heritage Language use among other**

ethnic groups in North America. We note that this finding counters Fix's (this issue) finding for African-American English features in a Columbus, Ohio, community.

A comparison of the two generations of speakers further suggests that different aspects of Ethnic Orientation are meaningful in different ways for different generations. Looking again at Components 3 and 4, we see that most first generation speakers behave as expected. With the Orientation Continuum method, highly heritage-oriented first generation speakers employ more null-subjects; with the Language Mixing method, highly language-separating first generation speakers employ more null-subjects. However, the direction of effect is the opposite for second generation speakers. Highly heritage-oriented and highly language-separating second generation speakers employ *fewer* null-subjects. This is not surprising considering the different language-learning dynamic associated with first generation versus second generation Heritage Language speakers. **These results suggest that ethnic orientation measures must be interpreted differently depending on the generation of the speaker.**

5.0 Conclusions

Throughout this paper, we have shown that there is no one-size-fits-all approach to coding and operationalizing ethnic orientation—even in a context where methods for collecting and analyzing data have been held constant and all participants inhabit the same city. Nonetheless, some general conclusions regarding methodological decisions can be drawn:

- Principal Component Analyses show that different Ethnic Orientation questions reveal different (uncorrelated) aspects of speakers' behavior and identity (*cf*. §3.2). Therefore, it is important to ask participants to answer all Ethnic Orientation questions during sociolinguistic interviews.
- Different Ethnic Orientation factors account for variation between speakers of different languages. This is true not only when comparing the Heritage Language corpus with the English corpus (cf. §3.2), but also when comparing different groups of Heritage Language speakers (cf. §3.3).
- Multivariate analyses offer a more nuanced and informative approach to variation than individual correlations. Factors that were not significant in correlation

analyses (§4.1) emerged as significant in multivariate analyses (§4.2), when their effects are considered in tandem with the effects of other variables, notably generation.

- Analyses relating Ethnic Orientation scores and linguistic variables find different effects to be significant for different types of linguistic variation. Compare, for instance, the factors found to be significant for VOT with the factors found to be significant for pro-drop, as reported in §4.2.

Generational differences are critically involved in all relationships between Ethnic Orientation and linguistic variation. For example, generational differences were found for VOT in Nagy & Kochetov (2013) (*cf*. §4.1), and also for pro-drop (§4.2), in terms of how Ethnic Orientation relates to linguistic variation. These suggest that it is useful to apply Ethnic Orientation criteria differently depending on generation. See also Chociej (2011) for clear examples of generational differences in Polish pro-drop effects.

At each stage in the analysis where we were able to contrast the outcomes of two different methods, we got different results. This, in conjunction with the conclusions summarized above, highlights the importance of careful consideration of methodological decisions.

The relationships found between linguistic variables and Ethnic Orientation variables in the Heritage Language data are promising. They suggest that Ethnic Orientation may be a key factor in modeling Heritage Language variation – Heritage Language variation should not be attributed solely to subtractive processes such as incomplete acquisition or attrition. As in English in the same communities (Hoffman, 2010; Hoffman and Walker, 2010), as well as in the communities described in the other papers in this issue, we see that aspects of speakers' attitudes and behaviors are reflected in their linguistic variation. As in these other papers, the varying degrees of strength of connections between ethnicity and linguistic features are highlighted.

The wide range of methods applied in this paper, and others in this issue, illustrate many different types of relationships that can be demonstrated between the multi-faceted concepts of Ethnic Orientation and linguistic variation. We can only report on correlations

between specific measures of Ethnic Orientation and specific measures of a linguistic variable; it would be imprudent to claim direct relationships between ethnicity as a whole and overall language use. By undertaking a project that compares different instantiations of Ethnic Orientation across a range of languages, communities, generations, and linguistic variables, while rigidly controlling the methods applied, we have illustrated that the different facets of Ethnic Orientation behave differently according to language, community, generation, and linguistic variable within which they are studied.

Acknowledgements

We thank SSHRC (SRG 410-2009-2330, SRG 410-2008-2048) for support; our participants; the HVLC RAs, listed at http://projects.chass.utoronto.ca/ngn/HLVC, for recruiting, interviewing, and analysis; the participants and organizers of the LSA 2012 “New perspectives on the concept of ethnolect” workshop, and the reviewers for helpful and encouraging feedback.

References

Baayen, R.H., 2008. Analyzing Linguistic Data. Cambridge University Press, Cambridge. www.ualberta.ca/~baayen/publications/baayenCUPstats.pdf. Accessed 10 June 2013.

Bayley, R., 1996, Competing constraints on variation in the speech of adult Chinese learners of English. In Bayley, R., Preston, D. (eds.), Second Language Acquisition and Linguistic Variation. Amsterdam, John Benjamins, pp. 97–120.

Bell, A.,1984. Language style as audience design. Language in Society 13:145-204.

Boberg, C., 2004. Ethnic patterns in the phonetics of Montreal English. Journal of Sociolinguistics 8, 538–568.

_____, 2005. The Canadian Shift in Montreal. Language Variation and Change 17, 133–154.

_____, 2008. Regional phonetic differentiation in standard Canadian English. Journal of English Linguistics 36, 129–154.

Boyd, S., Walker, J., Hoffman, M., 2011. Sociolinguistic practice among multilingual youth in Sweden and Canada. International Symposium on Bilingualism. Oslo.

Caramazza, A., Yeni-Komshian, G., 1974. Voice onset time in two French dialects. Journal of Phonetics 2, 239-245.

Chociej, J., 2010. Quantifying degree of contact: Determining the factors significant for heritage language speakers. Bilingual Workshop in Theoretical Linguistics, University of Toronto.

_____, 2011. Polish null subjects: English influence on Heritage Polish in Toronto. PhD Generals Paper. University of Toronto ms. http://individual.utoronto.ca/chociej/Files/UToronto-GP1-Chociej.pdf.

Clarke, S., Elms, F., Youssef, A., 1995. The third dialect of English: Some Canadian evidence. Language Variation and Change 7, 209–228.

Del Torto, L., 2010. 'It's so cute how they talk': Stylized Italian English as sociolinguistic maintenance. English Today 3, 55-62.

Eckert, P., 2000. Linguistic variation as social practice. Oxford: Blackwell.

Flege, J., 1991. Age of learning affects the authenticity of voice onset time (VOT) in stop consonants produced in a second language. Journal of the Acoustical Society of America, 89, 395-411.

Fowler, C. A., Sramko, V., Ostry, D. J., Rowland, S. A., & Hallé, P., 2008. Cross language phonetic influences on the speech of French–English bilinguals. Journal of Phonetics, 36(4), 649-663.

Giles, H., Powesland, P., 1975b. Speech Style and Social Evaluation. London and New York: Academic Press.

Guy, G., 1980. Variation in the group and the individual: The case of final stop deletion. In Labov, W. (Ed.), Locating Language in Time and Space. New York, Academic Press, pp. 1–36.

Hagiwara, R., 2006. Vowel production in Winnipeg. Canadian Journal of Linguistics 51, 127–142.

Hall-Lew, L., Starr, R., 2010. Beyond the 2nd generation: English use among Chinese Americans in the San Francisco Bay Area. English Today 3, 12-19.

Hoffman, M., 2010. The role of social factors in the Canadian vowel shift: Evidence from Toronto. American Speech 85 (2), 121-140.

_____ Walker, J., 2010. Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. Language Variation and Change 22, 37-67.

Hrycyna, M., Lapinskaya, N., Kochetov, A., Nagy, N., 2011. VOT drift in 3 generations of Heritage Language speakers in Toronto. Canadian Acoustics 39(3), 166-7.

Keefe, S., Padilla, A., 1987. Chicano Ethnicity. Albuquerque, NM, UNM Press.

Lisker, L., Abramson, A., 1964. Cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 84-422.

Milroy, L., 1980. Language and social networks. Oxford, Blackwell.

Nagy, N., 2009. Heritage Language Variation and Change. http://projects.chass.utoronto.ca/ngn/HLVC.

_____, 2011. A multilingual corpus to explore geographic variation. Rassegna Italiana di Linguistica Applicata 43, 65-84.

_____, Aghdasi, N., Denis, D., Motut, A., 2011. Pro-drop in Heritage Languages: A cross-linguistic study of contact-induced change. Penn Working Papers in Linguistics 17 (2), Article 16. http://repository.upenn.edu/pwpl/vol17/iss2/16/

_____, Kochetov, A., 2013. VOT across the generations: A cross-linguistic study of contact-induced change. In Siemund, P., Gogolin, I., Schulz, M., Davydova, J. (eds.), Multilingualism and Language Contact in Urban Areas: Acquisition - Development - Teaching - Communication. Amsterdam, John Benjamins, pp. 19-38.

Otheguy, R., Zentella, A. C., Livert, D., 2007. Language and dialect contact in Spanish in New York: Toward the Formation of a Speech Community. Language 83, 770–802.

Pierrehumbert, J., 2006. The next toolkit. Journal of Phonetics 34(4), 516-530.

Polinsky, M., 1995. Cross-linguistic parallels in language loss. Southwest Journal of Linguistics 14, 87–123.

Pustovalova, E., 2011. Null subject variation in the Russian spoken language (based on the materials of the Russian National Corpus). National Research University – Higher School of Economics ms.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

Roeder, R. Jarmasz, L., 2008. The lax vowel subsystem in Canadian English revisited. Toronto Working Papers in Linguistics 26, 1–12.

Rothman, J., 2007. Heritage speaker competence differences, language change, and input type: Inflected infinitives in Heritage Brazilian Portuguese. International Journal of Bilingualism 11.4, 359-389.

Rumpf, A., DiVenanzio, L., 2012. Null and overt subjects in Italian and Spanish heritage speakers in Germany. Heritage languages: language contact-change-maintenance and loss in the wave of new migration landscapes Workshop, Wuppertal.

Sankoff, D., Tagliamonte, S., Smith, E., 2005. Goldvarb X: A variable rule application for Macintosh and Windows. Department of Linguistics, University of Toronto, http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm

Santa Ana, O., 1992. Phonetic simplification processes in the English of the barrio: Across-generational sociolinguistic study of the Chicanos of Los Angeles. Ph.D. dissertation, University of Pennsylvania.

_____,1996. Sonority and syllable structure in Chicano English. Language Variation and Change 8, 63–89.

Statistics Canada, 2012. Toronto, Ontario (Code 535) and Ontario (Code 35) (table). Census Profile. 2011 Census. Statistics Canada Catalogue no. 98-316-XWE. Ottawa. Released October 24, 2012. http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/index.cfm?Lang=E. Accessed March 15, 2013.

Tagliamonte, S., Temple, R., 2005. New perspectives on an ol' variable. Language Variation and Change 17, 281–302.

Torres Cacoullos, R., Travis, C., 2010. Variable yo expression in New Mexico: English influence? In Rivera-Mills, S., Villa, D. (eds.), Spanish of the Southwest: A Language in Transition. Madrid, Iberoamericana, pp. 185-206.

Wagner, S. this issue. Linguistic correlates of Irish-American and Italian-American ethnicity in high school and beyond.

Walker, J., Boyd, S., Hoffman, M., in press. Sociolinguistic practice among multilingual youth in Sweden and Canada. In. Nortier, J. Svendsen, B.A. (eds.), Language, Youth and Identity in the 21st Century. Cambridge, Cambridge University Press.

Wolfram, W., 1969. A Sociolinguistic Description of Detroit Negro speech. Washington, DC, Center for Applied Linguistics.

Appendix A

This table shows the results of this first stage of PCA, which are the variables used in the second stage PCA. It uses the Topic method of grouping the Ethnic Orientation Questionnaire questions (see §3.2). Bold font highlights factors which emerged in components for both the Heritage Language and the English samples, illustrating a certain amount of similarity in the behavior of the participants in the two different corpora. In all analyses, the Heritage Language sample and the English sample were analyzed separately, though speakers of different communities/languages were combined within each corpus.

Factors emerging from the first stage PCAs, using the Topic method to group questions

Heritage Language Corpus	**English Corpus**
1. Cultural Heritage: Contact with country of heritage, country of schooling	1. **Self identification** and ethnicity of social network
2. **Language choice**	2. **Workplace network**
3. Language preference	3. Language proficiency and preference
4. **Partner's ethnicity and language use**	4. **Language choice**
5. Past and present social network (school, neighborhood, friends)	5. Birthplace, schooling and contact with Heritage Language country
6. **Self identification** and **workplace network**	6. **Parents' identity** and age of arrival in Canada
7. **Parents' identity** and language use	7. **Grandparents' age of arrival in Canada**
8. **General perception of discrimination to the group**	8. **Partner's ethnicity and language use**
9. Heritage Language proficiency and where it was learned	9. **Attitudes toward culture**
10. **Attitudes toward culture**	10. **Discrimination experienced in housing**
11. Personal discrimination experienced in school and in general	11. **General perceptions of discrimination**

12. **Discrimination experienced in housing** and work

13 Grandparents' language use and **age of arrival in Canada**

Appendix B

This table shows the results of this first stage of PCA, which are the variables used in the second stage PCA. It uses the Reference Group method of grouping the Ethnic Orientation Questionnaire questions (see §3.2). Bold font highlights factors which emerged in components for both the Heritage Language and the English samples, illustrating a certain amount of similarity in the behavior of the participants in the two different corpora. In all analyses, the Heritage Language sample and the English sample were analyzed separately, though speakers of different communities/languages were combined within each corpus.

Factors emerging from the first stage PCAs, using the Reference Group method to group questions

Heritage Language Corpus	**English Corpus**
1. **Grandparents' age of arrival** and language use, friends' language use	1. Family's language use and parents' identity
2. **Participant's birthplace, age of arrival in Canada, contact with heritage country**	2. **Participant's ethnic identity**
3. Participant's language preference	3. **Participant's attitudes toward culture**
4. **Participant's attitudes toward culture**	4. Participant's perception of **general discrimination**
5. Participant's experience of discrimination: in general and in school	5. Participant's social network
6. **Partner identity and language use**	6. **Grandparents' age of arrival**
7. Parents' identity, birthplace and language	7. **Partner's identity and language use**
8. Participant's perception of **general discrimination**	8. **Participant's birthplace, age of arrival in Canada, contact with heritage country**
9. Participant's childhood social network	9. Participant's language use and preference
10. **Participant's ethnic identity**	10. Participant's experience of housing discrimination