Using language models for holistic language variety comparisons*

Joshua McNeill

June 25, 2021

1 Introduction

One of the goals of scientific research is to discover generalizable facts, and work on language variation is no different. Variationists are often interested in describing language varieties and distinguishing between language varieties. This task is typically accomplished by analyzing several salient linguistic variables in fine detail and either generalizing to whole varieties from these several variables. The focus on deeply analyzing a small number of variables has advanced the understanding of language variation in more ways than can be mentioned here. While this paradigm shift from the broad analyses of many variables done dialectology, out of which language variation developed, has been valuable, modern technological advances have made it possible to once again consider analyzing broad swaths of linguistic variables at once. The objective for this study is thus to develop a method for holistically quantifying the distance between language varieties in a way that can account for practically all potential linguistic variables.

Some work on language variation has to already returned to its dialectological roots at least in spirit, and so a look at what has been done recently will help situate the method being proposed here. As such, section 1.1 will discuss some advantages and disadvantages of traditional methods of analyzing language variation to show what can be gained from a more holistic approach, and section 1.2 will review the limits of how holistic such traditional studies tend to get. Section 1.3 will explain a bit of what is done in the field of natural language processing (NLP), and section 1.4 will review how such technological advances have been applied to variationist studies, leading to the inclusion of many more variables than was previously practical.

*Data and code available at https://osf.io/9cjpw/.

1.1 Generalizing from saliency

In traditional variationist work, generalizations about whole varieties are sometimes drawn from a relatively small set of linguistic variables, perhaps as few as five. The assumption underlying that makes this possible is that the variables analyzed are salient and so have social meaning and so are what really distinguish varieties. This connection between saliency and social meaning has indeed been acknowledged in variationist literature (Podesva, 2011, p. 235) and has at least been implicit in variationist theory since its foundations.

Labov (1972) himself introduced the idea of saliency into language variation in proposing the concepts of linguistic variables being either indicators, markers, or stereotypes, the latter two carrying saliency. It was later claimed that most linguistic variables are markers (Bell, 1984), carrying saliency but not being explicitly commented on by speakers when describing one variety versus another. The importance of saliency for classifying types of linguistic variables was also part of the concept of orders of indexicality (Silverstein, 2003), where second and third order indexicals would be somewhat equivalent to markers and stereotypes, respectively, in terms of saliency.

Saliency has not only been a consistent aspect of variationist theorizing, but it has proved useful for analyses. One example is in analyzing how varieties comes to be social constructs, recognized as existing in the minds of speakers, which has been referred to in language variation as enregisterment (Agha, 2003). In essence, the process of enregisterment is based on features increasing in saliency. This was initially applied to Received Pronunciation (Agha, 2003) but has also been applied to the development of Pittsburghese based mostly on an analysis of (aw) monophthongization, though some other features are discussed also such as the term *yinz* 'you (pl.)' (Johnstone et al., 2006). Additionally, the saliency has been used to distinguish varieties spoken by different groups within a single geographic area in that residents will explicitly describe linguistic features that they associate with each group. This was the case in a study of Beijing where two types of Beijingers and one type of non-Beijinger were associated with one linguistic variable each (Zhang, 2005).[1]

The sort of explicit commentary noted in Zhang (2005) is one way to establish whether a linguistic variable is salient or not. To quantify such commentary requires developing indices where perhaps certain types of commentary each count in the index or commentary in general is counts as just one item among other types in the index. In fact, saliency has sometimes even been quantified on purely linguistic grounds (Podesva, 2011, p. 237).[2] Perhaps a more common way of identifying saliency has come through experimentation, however. Perceptual studies that invovle experimentation fall under paradigms such as eye-tracking (e.g., D'Onofrio, 2015) or matched-guise experiments (e.g., Campbell-Kibler, 2009; Delforge, 2012). Findings in such studies include showing that preconceptions of Valley Girls and Californians

[1] Technically, one group, the "smooth operators", was associated with two variables, but each variable involved [ɹ] (Zhang, 2005, pp. 441-444).

[2] Podesva (2011) does of course acknowledge social saliency, as well.

can impact perceptions of the TRAP vowel (D'Onofrio, 2015), that the alveolar variant of (ing) is judged differently depending on the listener (Campbell-Kibler, 2009), and that vowel devoicing in Andean Spanish is less noticeable by listeners when they do not want to notice it, as those who denied having the feature still had it at similar rates to those from elsewhere in whom the feature was regularly perceived (Delforge, 2012).

The point is that saliency does indeed exist, can be measured, and carries import for analyses, particularly if one is interested in social meaning. To the extent that varieties are social constructs, it is also a rather safe assumption to say that salient variables are what distinguish varieties as that which is not socially meaningful could not be a defining feature of a social construct. However, when looking at varieties more as concrete systems that are coherent, consistent, and objectively different from each other, comparing salienct variables alone may miss even a large number of distinctions that a more holistic comparison could capture. Furthermore, being able to compare between varieties both holistically and by looking at only salient variables opens the possibility of finding contrasts that may be enlightening, such as the possibility that two varieties different in many more ways than what is noticeable to speakers, yet the focus in variationist work has mainly been on salient variables alone.

1.2 Looking at clusters of linguistic variables

Language variation studies have of course examined clusters of variables and systems of interrelated variables. The focus in these cases is still linguistic variables that are thought to be salient, though the number of variables analyzed has at times approached a relatively larger scale. This is particularly true when the analysis focuses on "features", which are effectively treated as linguistic variables that have two realizations: present or absent.

Studies of African-American English (AAE)[3], perhaps due simply to there being so many, include both analyses of small clusters of variables but also larger sets of features. In some cases, there is an apparent relationship between the variables, such as in Rickford and McNair-Knox (1994) where the analysis included the zero copula, habitual invariant *be*, the plural *-s* morpheme, the 3rd person present *-s* verbal inflection, and the possessive *-s* morpheme. The first two of these both implicate the verb *to be* whereas the last three all involve homophonous suffixes. Such relationships are not always present, however, as in an examination of Martin Luther King, Jr's language that included the zero copula, (ing), (t/d) deletion, (r), word-final consonant cluster reduction, and prosody (Wolfram et al., 2016). If a structural relationship exists between these linguistic variables, it is not evident at first glance.

When sets of features from AAE are studied, the number of variables grows substantially. In the case of Van Hofwegen and Wolfram (2010), 32 features were tallied in individuals' speech as a way to measure how heavily AAE their

[3]I am including what researchers may refer to as African-American Vernacular English (AAVE) under this term as the distinction is not particularly important for the present study.

speech was. The measure itself, along with the chosen features, was developed in research on speech pathology and education and where it is referred to as the dialect density measure (DDM). While the proposal of an appellation such as DDM suggests something of a novel approach, similar measures have in fact been used elsewhere, as well. In a study looking into the use of AAE features in gay British men on Twitter, 25 AAE features were analyzed according to their frequencies (Ilbury, 2020). Also involving Twitter, over 30 non-standard spellings thought to be associated with AAE were analyzed in order to draw AAE dialect maps in the US (Jones, 2015). While these feature-conceptualized studies include far more than the typical five to seven clustered variables, they still fall well short of including all possible variables in these varieties.

If studies of AAE sometimes include clusters of linguistic variables that appear structurally related, it is not always clear that this structural relationship is what drove the choice of variables. In many other cases, there is little question that the researchers are interested in a cluster of variables because of their structural relationship. This is particularly evident in examinations of vowel systems or large portions of vowel systems, where there is often though not always a chain shift involved. This has been done since the foundational variationist work in New York City (Labov, 2006) and has continued in studies of other areas such as Belfast (Milroy, 1987), King of Prussia (Payne, 1980), Syndey (Horvath & Sankoff, 1987), the Detroit area (Eckert, 2000), and Philadelphia (Eckert & Labov, 2017). Similarly, though without implicating a chain shift, the majority of the pronoun system of Louisiana French as spoken in Houma has been analyzed as a whole (Rottet, 1995) as well as chunks of the verbal morphology system and and copula system in Cajun English (Dubois & Horvath, 2003). Finally, perhaps one of the most explicit treatments of a system of covariation was Guy's (2013) analysis of Brazilian Portuguese as spoken in Rio de Janeiro which had the explicit goal of examining the existence or non-existence of distinctive sociolects by determining if four linguistic variables covaried. While these sort of studies go further towards covering all the variables of particular linguistic systems, there is once again much left to be described if one wishes to compare the distance between varieties in their entireties. Furthermore, not all subsystems of varieties are contained in such a relatively small number of variables. For instance, it is not possible the lexical system of a variety with five, ten, or even one hundred variables.

1.3 Advances in natural language processing

1.4 The impact of NLP on language variation research

1.5 Generalizing from n-gram language models

2 Methods

3 Results

4 Discussion

5 Conclusion

References

Agha, A. (2003). The social life of cultural value. *Language & Communication*, *23*(3–4), 231–273. https://doi.org/10.1016/S0271-5309(03)00012-0

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145–204. https://doi.org/10.1017/s004740450001037x

Campbell-Kibler, K. (2009). The nature of sociolinguistic perception. *Language Variation and Change*, *21*(1), 135–156. https://doi.org/10.1017/S0954394509000052

Delforge, A. M. (2012). 'Nobody wants to sound like a provinciano': The recession of unstressed vowel devoicing in the Spanish of Cusco, Perú. *Journal of Sociolinguistics*, *16*(3), 311–335. https://doi.org/10.1111/j.1467-9841.2012.00538.x

D'Onofrio, A. (2015). Persona-based information shapes linguistic perception: Valley Girls and California vowels. *Journal of Sociolinguistics*, *19*(2), 241–256. https://doi.org/10.1111/josl.12115

Dubois, S., & Horvath, B. M. (2003). Verbal Morphology in Cajun Vernacular English: A Comparison with Other Varieties of Southern English. *Journal of English Linguistics*, *31*(1), 34–59. https://doi.org/10.1177/0075424202250296

Eckert, P. (2000). *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Blackwell Publishers, Inc.

Eckert, P., & Labov, W. (2017). Phonetics, phonology and social meaning. *Journal of Sociolinguistics*, *21*(4), 467–496. https://doi.org/10.1111/josl.12244

Guy, G. R. (2013). The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? *Journal of Pragmatics*, *52*, 63–71. https://doi.org/10.1016/j.pragma.2012.12.019

Horvath, B., & Sankoff, D. (1987). Delimiting the Sydney speech community. *Language in Society*, *16*(2), 179–204. https://doi.org/10.1017/s0047404500012252

Ilbury, C. (2020). "Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics*, *24*(2), 245–264. https://doi.org/10.1111/josl.12366

Johnstone, B., Andrus, J., & Danielson, A. E. (2006). Mobility, Indexicality, and the Enregisterment of "Pittsburghese". *Journal of English Linguistics*, *34*(2), 77–104. https://doi.org/10.1177/0075424206290692

Jones, T. (2015). Toward a Description of African American Vernacular English Dialect Regions Using “Black Twitter”. *American Speech*, *90*(4), 403–440. https://doi.org/10.1215/00031283-3442117

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Labov, W. (2006). *The Social Stratification of English in New York City* (2nd) [origdate = 1966]. Cambridge University Press.

Milroy, L. (1987). *Language and Social Networks* (2nd) [origdate = 1980]. Blackwell.

Payne, A. (1980). Factors controlling the acquisition of the Philadelphia dialect by out-of-state children. In W. Labov (Ed.), *Locating Language in Time and Space*. Academic Press.

Podesva, R. J. (2011). Salience and the Social Meaning of Declarative Contours: Three Case Studies of Gay Professionals. *Journal of English Linguistics*, *39*(3), 233–264. https://doi.org/10.1177/0075424211405161

Rickford, J., & McNair-Knox, F. (1994). Addressee- and Topic-Influenced Style Shift: A Quantitative Sociolinguistic Study. In D. Biber & E. Finegan (Eds.), *Sociolinguistic Perspectives on Register* (pp. 235–276). Oxford University Press.

Rottet, K. J. (1995). *Language shift and language death in the Cajun French-speaking communities of Terrebonne and Lafourche parishes, Louisiana* (PhD). University of Indiana. Bloomington, IN. Retrieved October 18, 2016, from http://search.proquest.com.proxy.bibliotheques.uqam.ca:2048/docview/304210484/abstract/901A763ABD434D22PQ/1

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language & Communication*, *23*(3–4), 193–229. https://doi.org/10.1016/S0271-5309(03)00013-2

Van Hofwegen, J., & Wolfram, W. (2010). Coming of age in African American English: A longitudinal study. *Journal of Sociolinguistics*, *14*(4), 427–455. https://doi.org/10.1111/j.1467-9841.2010.00452.x

Wolfram, W., Myrick, C., Forrest, J., & Fox, M. J. (2016). The Significance of Linguistic Variation in the Speeches of Rev. Dr. Martin Luther King Jr. *American Speech*, *91*(3), 269–300. https://doi.org/10.1215/00031283-3701015

Zhang, Q. (2005). A Chinese yuppie in Beijing: Phonological Variation and the Construction of a New Professional Identity. *Language in Society*, *34*(3), 431–466. https://doi.org/10.1017/s0047404505050153