

Using language models for holistic language variety comparisons*

Joshua McNeill

June 24, 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 come 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 involve 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

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

²Podesva (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 salient 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

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](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 [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9841.2012.00538.x>]. *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 [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/josl.12115>]. *Journal of Sociolinguistics*, 19(2), 241–256. <https://doi.org/10.1111/josl.12115>
- 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>
- Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia University of Pennsylvania Press.
- Podesva, R. J. (2011). Salience and the Social Meaning of Declarative Contours: Three Case Studies of Gay Professionals [Publisher: SAGE Publications Inc]. *Journal of English Linguistics*, 39(3), 233–264. <https://doi.org/10.1177/0075424211405161>
- 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](https://doi.org/10.1016/S0271-5309(03)00013-2)
- Zhang, Q. (2005). A Chinese yuppie in Beijing: Phonological Variation and the Construction of a New Professional Identity [Publisher: Cambridge University Press]. *Language in Society*, 34(3), 431–466. <https://doi.org/10.1017/S0047404505050153>