Disentangling style and priming using Generalized Additive Models

Christopher Ahern,[*] Aaron Ecay,[†] and Meredith Tamminga[‡]

April 14, 2015

Access this handout online: `http://aaronecay.com/papers/style-and-priming-poster-2015/`

1 Introduction

Sequential tokens of a linguistically varying item are rarely independent. Instead, neighboring instances more likely to surface as the same variant. (for early discussions of this phenomenon from a sociolinguistic perspective, see Poplack 1980, 1984; Sankoff and Laberge 1978) There are (at least) two potential causes of this tendency for sameness:

Priming

Priming is a neurally-motivated tendency to recycle linguistic structures that have been recently activated. (The pioneering paper on structural priming in an experimental setting was Bock 1986. For an overview of the literature, see Tamminga 2014, pages 7–21.) Once a certain neural representation has been activated, it is not immediately switched off, but rather its activation gradually decays. In conversation, a residually activated representation of one variant of a variable will be preferentially reactivated, making that variant more likely to reoccur than its competitors.

Style

Style is social process by which speakers situate themselves and their speech in a multidimensional space of identity concepts. It modulates variant choice because speakers engage in 'style-shifting': modulation of variant probability in response to situational factors like interlocutor, stance, topic, or context. In conversation, neighboring tokens of variation are likely to be located in a stylistically-coherent portion of the discourse.

In this work we use Generalized Additive Models (GAMs, Hastie and Tibshirani 1990) to distinguish and quantify these two potential causal factors in a corpus of sociolinguistic interviews.

2 Data and methods

- 18,022 tokens of DH-stopping (*this* ~ *dis*) taken from the 42 interviews in the Philadelphia Neighborhood Corpus (Labov and Rosenfelder 2011)
- Median tokens/speaker = 367; min = 72; max = 752
- Using the `mgcv` package for the R statistical computing language, we fit one GAM per speaker. (R Core Team 2015; Wood 2011)

[*]University of Pennsylvania, `cahern@ling.upenn.edu`
[†]University of York, `aaron.ecay@york.ac.uk`
[‡]University of Pennsylvania, `tamminga@ling.upenn.edu`

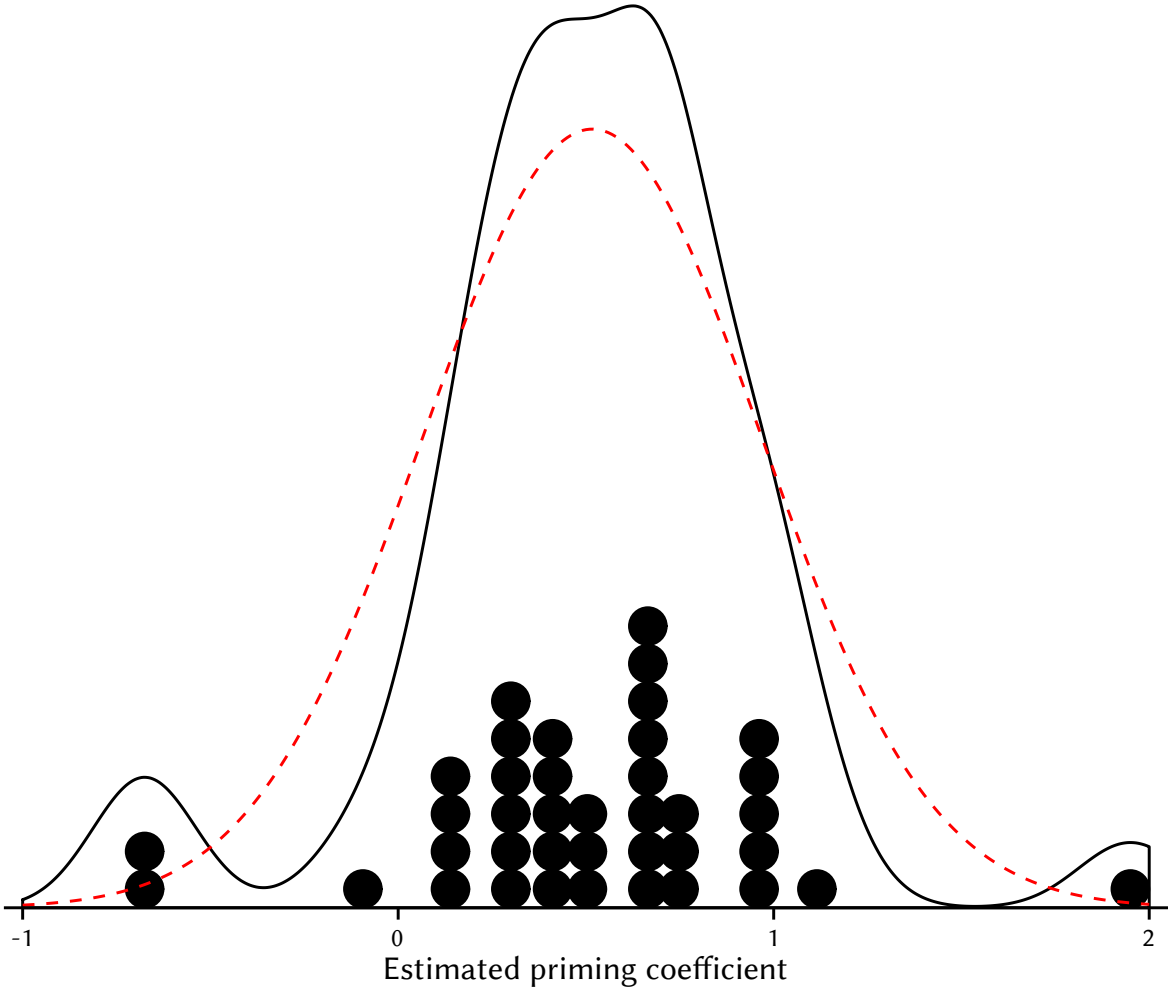

Figure 1: Distribution of priming coefficient in GAMs fit to 39 speakers. The dashed red line is the normal distribution fit by MLE to these speakers.

- Our model was:

$$\text{observation} \sim \underbrace{s(\text{time})}_{\substack{\text{Smooth estimate of} \\ \text{style shifting}}} + \underbrace{\text{previous token}}_{\text{Estimate of priming}}$$

- We use a GAM rather than a traditional logistic regression in order to investigate the hypothesis that different speakers engage in style-shifting to different degrees. That is, we want to let our data "speak for itself."

3 Priming results

- Excluded 3 speakers for whom the model did not converge (priming coefficient ≈ -20)
- Resulting distribution of priming estimates plotted in Figure 1
- Three low outliers with priming values below zero (indicating that the model estimates that these speakers actually engage in anti-priming)
- One high outlier with a priming estimate of 1.9
- Of these
  - One low outlier has a priming estimate which is hardly distinguishable from zero (-0.094)
  - One low outlier has a low N (72)

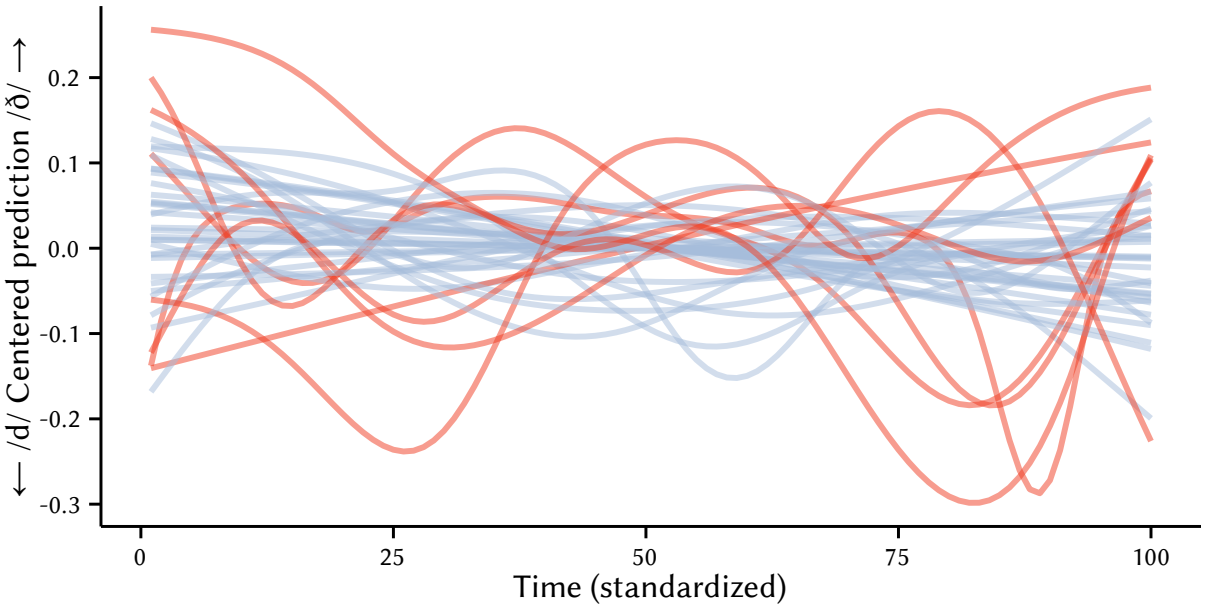

Figure 2: Style splines fit to 39 speakers. Splines in dark red have a range of more than 0.25, whereas those in light blue have less.

  - The high outlier has a low N (78)
  - (One low outlier does not have any obvious problems)
- The remaining 35 priming estimates are normally distributed (Shapiro-Wilk p = 0.30) between 0.1 and 1.1

These results suggest that priming is a universal process. All speakers in our sample for whom our method succeeded in measuring priming participate in the process to some degree. Though individual differences may exist, they do not divide the population into classes, consistent with our hypothesis that priming is an automatic neural phenomenon underlying language production at a deep level.

4 Style results

- Same 39 speakers as in the priming section
- The style splines from each speaker's model are plotted in Figure 2
- Data are normalized with respect to interview length and speaker mean but (crucially) not standard deviation
- Many speakers have simple trajectories estimated for their stylistic behavior: either a flat line or one which slopes down
- A few speakers, highlighted in blue, show style trajectories which both cover large areas of probability space and have a complex functional form
- The number of degrees of freedom that the GAM assigns the style spline are shown in Figure 3
- Bimodal distribution of these values in the population. Most speakers have a linear trajectory (~1 DoF), whereas a minority have a more complicated functional form characterized by a higher number of DoF

These results are compatible with the hypothesis that style-shifting is not an automatic process, but rather one over which speakers have some degree of control. Evidently, different speakers differentially exploit the strategy of variant clustering for stylistic signaling.

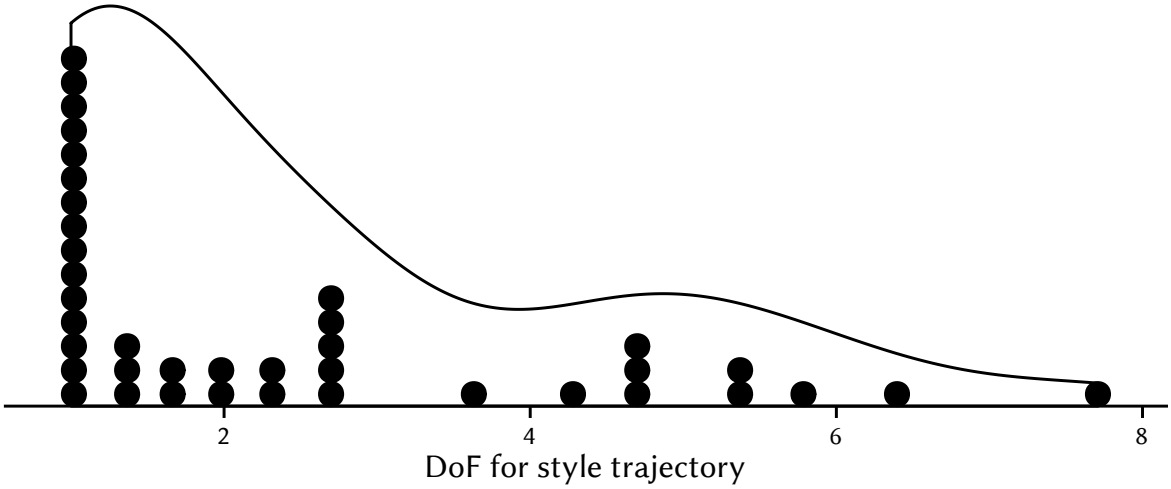

Figure 3: Estimated degrees of freedom for style from GAMs fit to 39 speakers.

5 Conclusions

We have demonstrated that the GAM modeling technique can distinguish between two causal factors which both condition variant selection: priming and style-shifting. The structure that the model assigns to these effects is consistent with their posited sources: a fully automatic, universal process and one over which speakers can exhibit control and variability, respectively.

Take-home message

Speakers vary in their deployment of style-shifting, but priming is a universal, automatic phenomenon which conditions variant selection.

6 Future directions

- **Comparison with top-down approaches to style** In the sociolinguistic literature, a top-down approach to style is common. Labov (2001) exemplifies this approach with his stylistic decision tree. We would like to compare the predictions of such methods to our quite different bottom-up model, to see if they agree on which utterances belong to high/formal and low/informal styles.
- **Investigation of inter-speaker priming differences** Our work has yielded evidence that some speakers are more or less apt to exhibit priming. Do these differences correlate with other neurolinguistic properties?
- **Investigation of inter-speaker style-shifting differences** Our work has also yielded the conclusion that only a minority of speakers style-shift (in an interview setting). Are there systematic correlations between the tendency to style shift and other sociolinguistically relevant factors, such as social network density?
- **More data** It is a truism that more data is always welcome. We are specifically interested in investigating whether these patterns of priming and style shifting are replicated with different types of variable, including phonetic and morphosyntactic variables in addition to the phonological one discussed here.

7 Acknowledgment

We would like to express our gratitude to the speakers who have agreed to be interviewed for the PNC over the past 40 years. Their generosity makes our work possible.

References

Bock, Kathryn (1986). "Syntactic persistence in language production". In: *Cognitive Psychology* 18, pp. 355–387.

Hastie, Trevor and Robert Tibshirani (1990). *Generalized additive models*. CRC Press.

Labov, William (2001). "The anatomy of style-shifting". In: *Style and Sociolinguistic Variation*. Ed. by Penelope Eckert and John R. Rickford. Cambridge University Press. Chap. 5, pp. 85–108.

Labov, William and Ingrid Rosenfelder (2011). *The Philadelphia Neighborhood Corpus of LING 560 studies, 1972-2010*. With support of NSF contract 921643.

Poplack, Shana (1980). "Deletion and disambiguation in Puerto Rican Spanish". In: *Language* 56.2, pp. 371–385.

Poplack, Shana (1984). "Variable concord and sentential plural marking in Puerto Rican Spanish". In: *Hispanic Review* 52.2, pp. 205–222.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Sankoff, D. and S. Laberge (1978). "Statistical dependence among successive occurrences of a variable in discourse". In: *Linguistic variation: models and methods*. Ed. by David Sankoff. Academic Press. Chap. 8, pp. 119–126.

Tamminga, Meredith (2014). "Persistence in the production of linguistic variation". PhD thesis. University of Pennsylvania.

Wood, S.N. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". In: *Journal of the Royal Statistical Society (B)* 73.1, pp. 3–36.