The Gettysburg Corpus: Testing the proposition that all tense /æ/s are created equal

Isaac L. Bleaman and Daniel Duncan, New York University

While sociolinguists have analyzed publicly available written language on social media sites like Twitter (Huang et al. 2015, Eisenstein forthcoming), few have tapped the rich potential of such "big data" in the study of spoken language. This paper demonstrates the promise of one large speech corpus for sociophonetic analysis: hundreds of recitations of Abraham Lincoln's Gettysburg Address. These recordings have many of the benefits of laboratory reading tasks (e.g., a uniform transcript), but due to the sheer number and variety of participants can also be used to uncover patterns of language use correlated with social factors.

To illustrate the corpus's suitability for sociophonetic research, we replicate previous findings regarding the effect of phonetic environment on vowel length. We then apply the corpus to a novel topic: testing whether the phenomenon labeled "/æ/ tensing" is realized uniformly across two different dialects of American English: New York City English and the Northern Cities Shift of the Inland North (Labov et al. 2006). As F1/F2 are relatively unaffected by compression algorithms (Bulgin et al. 2010), variables like /æ/ are representative of the corpus's capabilities.

The recitations are catalogued on www.learntheaddress.org, organized by reciters' state affiliations, and hosted on YouTube. Each video from both Michigan (representing the NCS) and New York (representing NYCE) was scraped from the source code, and the audio of each was extracted as a .wav file and aligned in FAVE (n=402). We measured vowel duration as well as F1/F2 at 10% increments through the vowels in *last, add, task, advanced,* which are tensed in both dialects. We utilize post-hoc measures of the vowel space at large to select a subset of reciters who exhibit traditional features of NCS and NYCE (n=106 NCS, 55 NYCE).

We replicate the finding that vowel duration is longer preceding a voiced consonant than voiceless (Bybee 2001); duration in *add* is significantly longer than in other contexts ($p << .0001$). We also find significant regional differences in the formant trajectory of tense /æ/. Smoothing spline analysis of variation (Gu 2014, Davidson 2006) shows that vowel onset in the NCS group is higher, but with equal fronting, as in the NYCE group. Over the course of the vowel, the NCS group lowers /æ/ to match the NYCE group in F1, but backs the vowel to separate from the NYCE group. These results are consistent with Labov et al. (2006), who suggest that only the NCS has a full diphthong. The differences between the two groups problematize whether "/æ/ tensing," e.g., in pre-nasal environments, can be considered a unified cross-dialectal phenomenon.

These findings confirm the viability of constructing naturalistic speech corpora for sociophonetic research; we particularly show that they can be used to replicate findings made under controlled settings. While phoneticians have rightfully pointed out the complications of relying on YouTube-compressed audio files extracted for phonetic analysis (De Decker and Nycz 2011), we contend that such large corpora can supplement traditional methods when testing hypotheses about socially-conditioned variation.

References

Bulgin, James, Paul De Decker, and Jennifer Nycz. 2010. Reliability of formant measurements from lossy compressed audio. Poster presented at the British Association of Academic Phoneticians Colloquium, University of Westminster.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Davidson, Lisa. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *Journal of the Acoustical Society of America* 120.407–415.

De Decker, Paul, and Jennifer Nycz. 2011. For the record: Which digital media can be used for sociophonetic analysis? *University of Pennsylvania Working Papers in Linguistics* 17.51–59.

Gu, Chong. 2014. Smoothing Spline ANOVA Models: R Package gss. *Journal of Statistical Software* 58.1–25.

Eisenstein, Jacob. Forthcoming. Written dialect variation in online social media. In Charles Boberg, John Nerbonne, and Dominic Watt (eds.), *The handbook of dialectology*. New York: Wiley-Blackwell. http://www.cc.gatech.edu/~jeisenst/papers/dialectology-chapter.pdf

Huang, Yuan, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2015. Understanding U.S. regional variation with Twitter data analysis. *Computers, Environment and Urban Systems*. http://dx.doi.org/10.1016/j.compenvurbsys.2015.12.003

Labov, William, Sharon Ash, and Charles Boberg. 2006. *The atlas of North American English: Phonetics, phonology and sound change*. Berlin: Mouton de Gruyter.