Introducing LingTube: An open-source toolkit for linguistic analysis of YouTube data

YouTube is a vast, yet relatively untapped, source of publicly-available linguistic data (Schneider, 2016). While not without challenges, there are many advantages to using YouTube as a data source, particularly for the computational study of language variation and change: Researchers can collect large amounts of 'naturalistic' speech data representing various contexts (cf., Hall-Lew & Boyd, 2020 on self-recordings), and much of this speech is already captioned, jumpstarting the transcription process. Furthermore, YouTube has the potential to improve access to lesser-studied language varieties or communities. In this vein, this study has two main contributions. First, we introduce LingTube, an open-source suite of tools for automating the downloading and processing of captioned YouTube audio. Second, we present ongoing exploratory work applying these tools to identify potential features of Asian North American (ANA) ethnolinguistic varieties.

In language variation research, data gathered from YouTube videos has been used to study both intra-speaker variation (e.g., Lee, 2017 on style-shifting across contexts) and inter-speaker variation, including larger-scale comparative analysis across regions (Coats, 2020 on articulation rate across US regions) or languages (Kramer, 2021 on cross-linguistic patterns of dependency length minimization). Analysis can be conducted at various levels of linguistic structure, including morphosyntactic, lexical and phonetic. However, working with YouTube data comes with some drawbacks, particularly for phonetic analysis, which LingTube has been developed to address. First, downloading audio and captions are not straightforward. LingTube automates the process of scraping this data from a list of videos, an entire channel, or a set of search results. Second, auto-generated text and/or time-alignment still require hand-correction, and third, unanalyzable speech (e.g., due to background noise or music) must be identified and removed; these tasks remain somewhat time-intensive. Although manual work is still required at this stage, LingTube helps to streamline the process of cleaning, correcting and time-aligning transcripts, as well as identifying usable sections of speech. Ongoing development aims to automate the process of detecting unusable speech. Finally, LingTube also automates the process of creating TextGrids for conducting forced alignment and facilitates hand-correction of forced-aligned segment boundaries.

To demonstrate this novel tool for analysing variation using computational methods, we apply LingTube to study inter-speaker variation across ethnicity and region among ANAs. Although ANAs are not a monolith, anecdotal and perceptual evidence suggest that ANA speakers can often be identified as such by local listeners, hinting at the existence of recognizable ANA varieties (Hanna, 1997; Newman & Wu, 2011; Wong & Babel, 2017). However, this contrasts with thus-far inconclusive evidence for any 'distinctive' ANA ethnolinguistic variety or variants (Reyes & Lo, 2009; Newman & Wu, 2011). Given that few studies have done comparative analyses across regions (Wong & Hall-Lew, 2014) or ethnicity (Bauman, 2016), we are developing a corpus of speech produced by 80 YouTubers, including ANA-identifying speakers and non-ANA comparisons. Although analysis is still ongoing, findings from cluster analyses of vowel space and speech rhythm will, regardless of outcomes, provide new insight into the open question of what, if any, ANA ethnolinguistic features exist as markers of (pan-)ethnic identity.

References

Bauman, C. (2016). Speaking of Sisterhood: A Sociolinguistic Study of an Asian American Sorority [Ph.D. Dissertation]. New York University.

Coats, S. (2020). Articulation Rate in American English in a Corpus of YouTube Videos. *Language and Speech*, *63*(4), 799–831. https://doi.org/10.1177/0023830919894720

Hall-Lew, L., & Boyd, Z. (2020). Sociophonetic perspectives on stylistic diversity in speech research. *Linguistics Vanguard*, *6*(s1). https://doi.org/10.1515/lingvan-2018-0063

Hanna, D. B. (1997). Do I Sound "Asian" to You?: Linguistic Markers of Asian American Identity. *University of Pennsylvania Working Papers in Linguistics*, *4*(2), 15.

Kramer, A. (2021). Dependency Lengths in Speech and Writing: A Cross-Linguistic Comparison via YouDePP, a Pipeline for Scraping and Parsing YouTube Captions. *Proceedings of the Society for Computation in Linguistics*, *4*(1), 359–365.

Lee, S. (2017). Style-Shifting in Vlogging: An Acoustic Analysis of "YouTube Voice." *Lifespans and Styles*, *3*(1), 28–39. https://doi.org/10.2218/ls.v3i1.2017.1826

Newman, M., & Wu, A. (2011). "Do You Sound Asian When You Speak English?" Racial Identification and Voice in Chinese and Korean Americans' English. *American Speech*, *86*(2), 152–178. https://doi.org/10.1215/00031283-1336992

Reyes, A., & Lo, A. (Eds.). (2009). *Beyond Yellow English: Toward a Linguistic Anthropology of Asian Pacific America*. Oxford University Press, USA.

Schneider, E. W. (2016). World Englishes on YouTube: Treasure trove or nightmare? *World Englishes*, 253–282. http://www.jbe.platform.com/content/books/9789027267061-veaw.g57.11sch

Wong, P., & Babel, M. (2017). Perceptual identification of talker ethnicity in Vancouver English. *Journal of Sociolinguistics*, *21*(5), 603–628. https://doi.org/10.1111/josl.12264