Archive of Pittsburgh Language and Speech

Dan Villarreal
Department of Linguistics

Motivation

- Open Science research tools are touted as **democratizing research**, but these tools can themselves pose a **barrier to entry** [1]
- Pitt has a trove of **Pittsburgh speech data** (recordings) that has yielded influential research in **sociolinguistics**, the study of how language intersects with society [e.g., 2]
- This project will transform this data into the Archive of Pittsburgh Language and Speech (APLS)—**the most powerful and accessible sociolinguistics Open Data resource to date**

Context

- Existing Open Data resources in sociolinguistics remove the need for data *collection* but don't address issues with traditional data *preparation* methods
  - Time-consuming, tedious labor
  - Reliant on manual methods (potential for human error)
  - Lacking in reproducibility

Project Description

- APLS will use the linguistic data software LaBB-CAT [3], which makes data prep far easier, faster, and more systematic
  - Synchronizes linguistic annotations to speech, from 30ms speech sounds to hourlong interviews
  - Facilitates searching for complex linguistic phenomena (Fig. 1)
  - Interfaces with R & Python for reproducibility

orthography≈^(sing|ring|king)$ segment≈^(I)$ unisyn syllables≈^(.*IN)$ dictionary-phonemes≈^(.+IN)$

Found 877 results (Total utterance duration: 59:03.912)

Select all results (877) Context: 1 word

CB01interview1.eaf - CB01

1.		a	threatening	type
2.			trusting	and
3.		you	anything	if
4.		of	thing	
5.		very	welcoming	place

University of Pittsburgh

Bringing legacy **Pittsburgh speech data** together with **Open Science**

Project Deliverables

- APLS: a publicly available sociolinguistic data resource consisting of:
  - 271 sound files, representing
  - 44.9 hours of interviews, from
  - 40 Pittsburgh speakers
- Orthographic (text-like) transcription of each sound file, plus numerous layers of linguistic annotations (e.g., phonemes, morphemes, frequency counts, part of speech) and speaker demographic data
  - As of 9/5/2023: 39% of the corpus has been transcribed and uploaded
- Documentation and training materials

Potential Impact

- Making freely available to *all* researchers the type of rich sociolinguistic data resource generally available only to well-resourced linguistics departments
- Advancing data-scientific principles to vastly improve sociolinguistic data methods' efficiency and reproducibility
- Paving the way for advances in our understanding of Pittsburgh speech and sociolinguistics more broadly

References

1. Villarreal, Dan & Lauren Collister. in press. Open Methods: Decolonizing (or not) research methods in linguistics. In *Decolonizing linguistics* , ed. by Anne Charity Hudley, Christine Mallinson, and Mary Bucholtz. Oxford: Oxford University Press.
2. Johnstone, Barbara, Jennifer Andrus & Andrew E. Danielson. 2006. Mobility, indexicality, and the enregisterment of "Pittsburghese." *Journal of English Linguistics* 34(2). 77–104. https://doi.org/10.1177/0075424206290692.
3. Fromont, Robert & Jennifer Hay. 2012. LaBB-CAT: An annotation store. *Proceedings of Australasian Language Technology Association Workshop* 113–117.

Acknowledgments

Thanks to Scott Kiesling, Barbara Johnstone, Robert Fromont, the Computational Sociolinguistics Lab, and all our speakers for making this project possible