LanguageARC:

Using Citizen Science to Augment Sociolinguistic Data Collection and Coding

Christopher Cieri, Jonathan Wright, James Fiumara, Alex Shelmire and Mark Liberman
University of Pennsylvania, Linguistic Data Consortium
Corresponding: {ccieri, jfiumara}@ldc.upenn.edu, www.ldc.upenn.edu

Introduction

- Discipline-specific advances, exploiting improvements in computing, allow quantitative sociolinguistics to analyze more subjects, variables and situations.
- Forced alignment [1, 2], vowel formant extraction [3] and statistical tools [4] allow researchers to undertake investigation and re-analysis until recently impossible [5]. Classifiers for other variables are emerging [6, 7] and search-guided coding [8] alone provides faster access to tokens where no classifiers exist.
- Unfortunately, we continue to suffer from a paucity of recorded observations of language-related behavior and of judgements which cannot be accelerated by purely technical mechanisms.
- Recording new data for under-represented communities and situations, transcription [9] for most languages [10, 11], evaluative judgements of the kind used, e.g., in perceptual dialectology [12] and matched guise tests require human input which is relatively costly and time consuming to obtain.

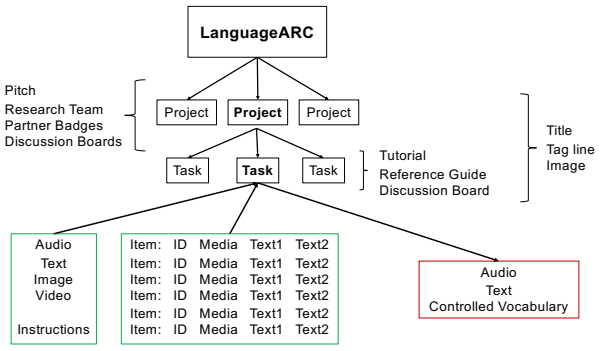
Novel Incentives and Workflows

- NSF sponsored **NIEUW** (*NSF CRI CI-NEW #1730377*) augments existing data sources by offering novel incentives -- opportunities to learn, compete, promote a linguistic variety and make real contributions to science – on a significantly larger scale with less effort.
- Similar approaches proven successful in other fields such as the > 443 million judgments submitted by > 1.9 million Zooniverse [13] volunteers
- Our own language ID game, **NameThatLanguage.org** has collected 382,014 HITs from 31,368 players in < 1 year.

Citizen Linguists

- **LanguageARC** is NIEUW's Citizen Linguist portal designed specifically to collect linguistic data and judgements supporting multiple projects and tasks.
- Underlying toolkit already used in > 100 collection and coding tasks that yielded > 1,000,000 judgements
- Simplified for researcher use
- Tasks short enough to complete during breaks, commuting, etc. and many require nothing more than a smart phone.
- Project Builder allows researchers to deploy new tasks <1 hour, given a design and appropriately formatted data.
- Example Uses: eliciting word, sentence, passage reading; DiaPix and Map Task speech; silent movie narration and judgements such as language, speaker and dialect identification, transcription, grammaticality, and variable coding

Structure



Input Data

ID	Text	Passage Name
1	Rainbow.txt	Rainbow
2	NWS.txt	NWS
3	Stella.txt	Stella
4	Wolf.txt	Wolf
5	Arthur.txt	Arthur
6	Comma.txt	Comma
7	Grandfather.txt	Grandfather

Project Builder

New Project

Name	Title	Subtitle
Pitch	News / Blog URL	Project Image
Forums	Announcements	General Discussion
	Questions	Help & Technical Support
Research Team	Name	Title
	Description	Photo
Partners	Name	URL
	Badge	

New Task

Name	Title	Image
Description	Tutorial	Reference Guide
Assignment Order	In Order	Random
Coverage	Within Contributors	Across Contributors?
Forum	General Discussion	

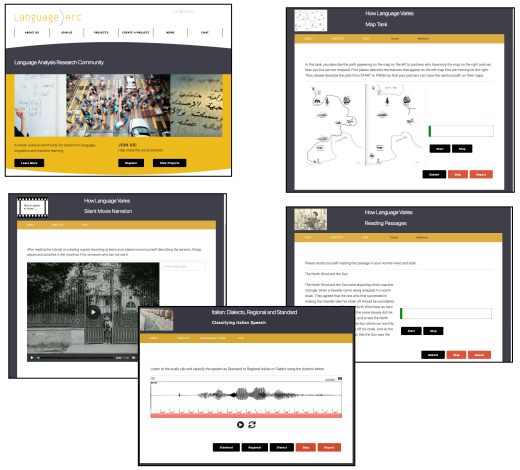
New Dataset

Name	Description	
Manifest File	Data Files	
Randomize Manifest Item Order? Yes No		

Tool

Instructions		
Media Type: Text, Audio, Image, Video, None		Media Column
Language	Limit Language Selection=>	Languages
Item ID		
Item Specific Text	Primary Text/Label	Secondary Text/Label
Response Audio	Level Test? Yes No	Level Meter? Yes No
Response Text Yes No		Text/Label
Judgment Buttons	Multiple Choice	
Allow	Skip? Yes No	Report Bad Item Yes No

Projects/Tasks



Reporting

```
Activity ID Data Set ID User ID Country Code City
       141          90       3           US Philadelphia

Date-Time                  PromptID Judgment Audio Beg End
2019-07-15 07:05:15 +0000         1 Standard         0 2.257
```

Works Cited

[1] J. Yuan, M. Liberman and C. Cieri, "Towards an integrated understanding of speaking rate in conversation," in Proceedings of Interspeech, 2006.
[2] J. Yuan, M. Liberman and C. Cieri, "Towards an integrated understanding of speech overlaps in conversation," in Proceedings of ICPhS XVI, 2007.
[3] K. Evanini, S. Isard and M. Liberman, "Automatic formant extraction for sociolinguistic analysis of large corpora," in Proceedings of Interspeech., 2009.
[4] T. Kendall and E. R. Thomas, "Vowels: vowel manipulation, normalization, and plotting in R.," 2010. [Online]. Available: cran.r-project.org/web/packages/vowels/index.html..
[5] W. Labov, I. Rosenfelder and J. Fruehwald, "100 years of sound change in Philadelphia: linear incrementation, reversal and re-analysis," Language, vol. 89, pp. 30-66., 2013.
[6] J. Yuan and M. Liberman, "Automatic measurement and comparison and vowel nasalization across languages," in Proceedings of ICPhS XVII, 2011.
[7] J. Yuan and M. Liberman, "Automatic detection of 'g-dropping' in American English using forced alignment," in Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, 2011.
[8] C. Cieri, Modeling Phonological Variation in Multidialectal Italy, Philadelphia: University of Pennsylvania, 2005.
[9] S. Reddy and J. N. Stanford, "Toward completely automated vowel extraction: Introducing DARLA," Linguistics Vanguard, vol. 1, 2015.
[10] S. Gonzalez, J. Grama, S. Ananthanarayan, D. Barth and C. Travis, "Opening the door for forced- alignment to minority languages," in NWAV 47, New York, 2018.
[11] P. Meer and J. A. Matute Flores, "Making FAVE ready for New Englishes: Applying and modifying FAVE for semi-automatic acoustic analyses of Trinidadian English vowels," in NWAV 47, New York, 2018.
[12] D. Preston, Handbook of Perceptual Dialectology. Volume 1, Johns Benjamins, 1999.
[13] "Zooniverse," [Online]. Available: https://www.zooniverse.org/. [Accessed 1 June 2019].
[14] Z. Boyd, Z. Elliott, J. Fruehwald and L. Hall-Lew, "An Evaluation of Sociolinguistic Elicitation Methods (2015)," in Proceedings of ICPhS, 2015.

Soft Launch:

https://www.languagearc.org