Using Citizen Linguistics to Empower Indigenous Language Communities

Christopher Cieri, Mark Liberman
University of Pennsylvania

Linguistic Data Consortium

Introduction

- Nearly all Language Technologies require Language Resources, absent for most of the world's languages, especially indigenous languages
- Language Resources = recorded speech, text or transcriptions for written languages, translations and annotations that reveal grammar
- Most publicly available LRs arise from government/NGO sponsored R&D which focuses on languages with greatest Gross Linguistic Product. Even programs for "Under-Resourced Languages" tend to focus on "big" languages
- Funding addresses immediate needs.
- Indigenous Language communities historically at mercy of funders whose attention is elsewhere
- **Approach**: empower Indigenous Language Communities to help themselves by supporting creation of Language Resources and thus Language Technologies

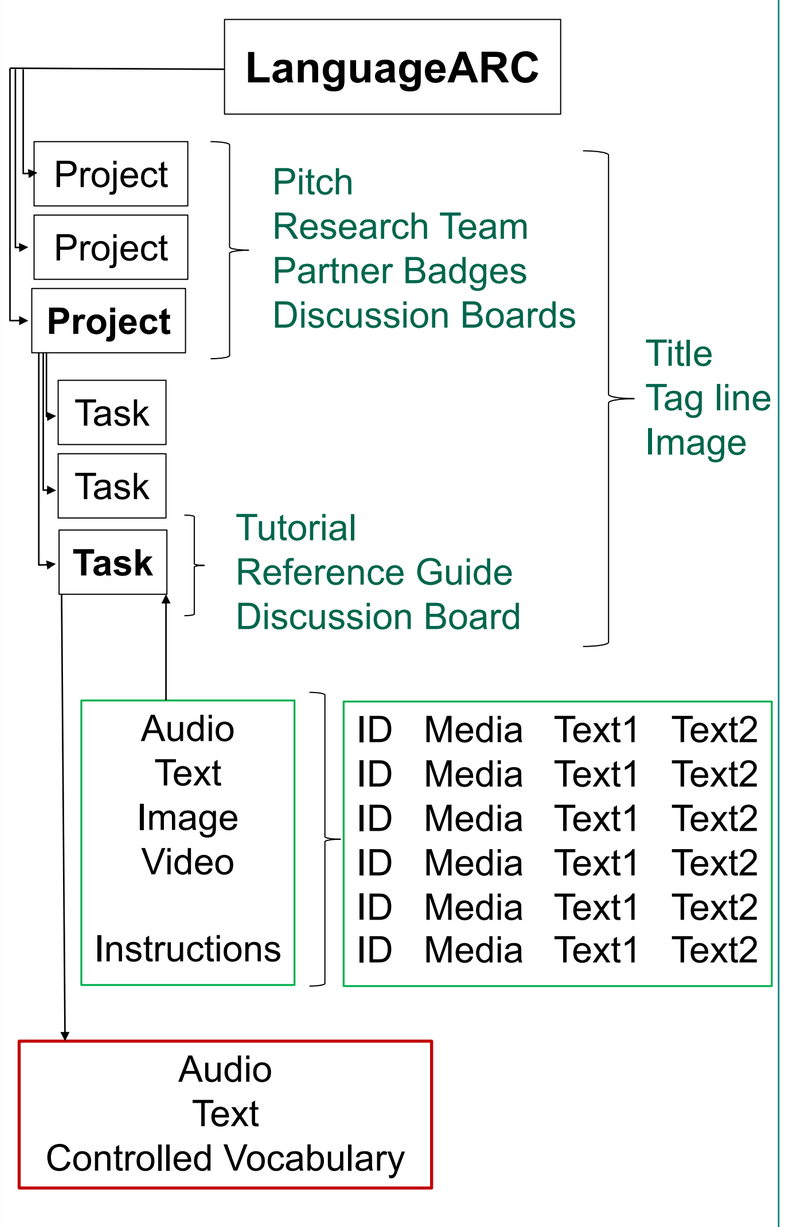
Novel Incentives, Workflows

- NSF sponsored NIEUW (NSF CRI CI-NEW #1730377) augments existing data sources by offering novel incentives: opportunities to learn, compete, promote a linguistic variety, make real contributions to science – larger scale with less effort.
- Similar approaches successful in other fields: >449 million judgments from >1.9 million Zooniverse [13] volunteers

Citizen Linguists

- ***LanguageARC*** is NIEUW's Citizen Linguist portal to collect language data and judgements supporting multiple projects and tasks.
- Underlying toolkit used in >100 tasks, >1,000,000 judgements, simplified for current use
- Short tasks completed in minutes on computer or smart phone, network optional.
- New tasks deployed in <1 hour, given a design & appropriately formatted data.
- Examples: picture description, silent movie narration, DiaPix, Map Task, read / prompted speech; transcription, (grammar elicitation via) translation, judgements of language, dialect, speaker ID, grammaticality, usegae surveys

Structure



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 In Order Random
Coverage Within Across Contributors?
Forum General Discussion

New Dataset

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

Tool

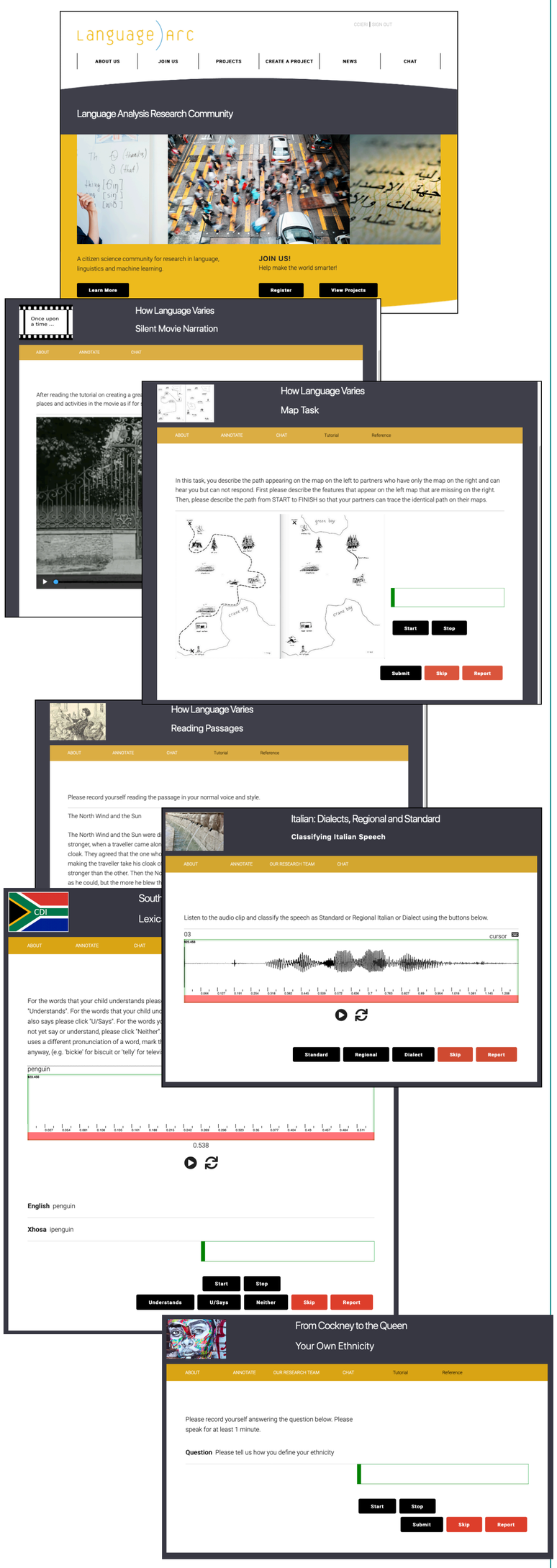
Instructions
Media Type: Text, Audio, Image, Video, None
Media Column
Language: Limit Language, Selection=>Languages

Item ID
Item Specific Text Text/Label1 Text/Label2
Response Audio Level Test? Level Meter?

Response Text ? Text/Label
Judgment Buttons M/C

Allow Skip? Report Bad Item?

Projects/Tasks



Conclusion

LanguageARC simplifies the creation of Language Resources to lower barriers to creating Language Technologies for All.