

NETS 213 : Crowdsourcing and Human Computation - Guest Lecture

Citizen Linguistics (with LanguageARC)

Christopher Cieri

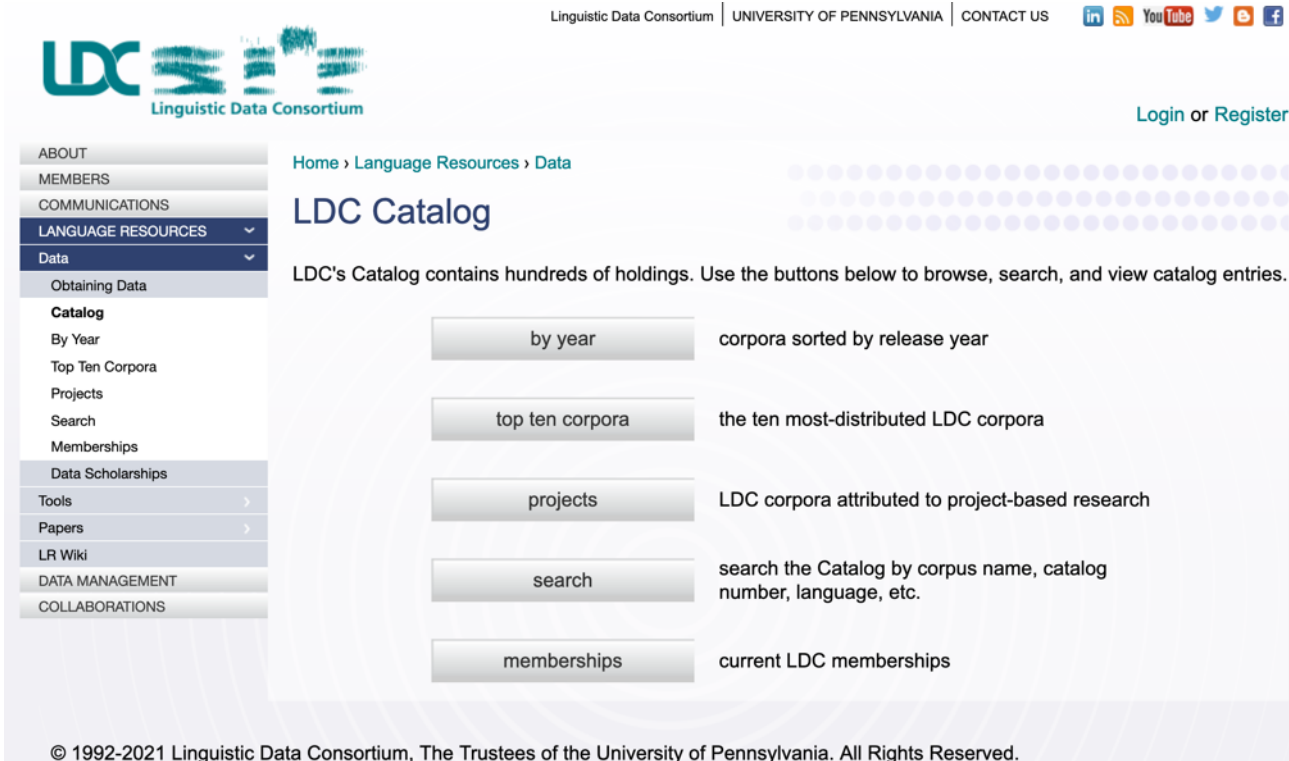
Adjunct Associate Professor in Linguistics, Executive Director, Linguistic Data

University of Pennsylvania

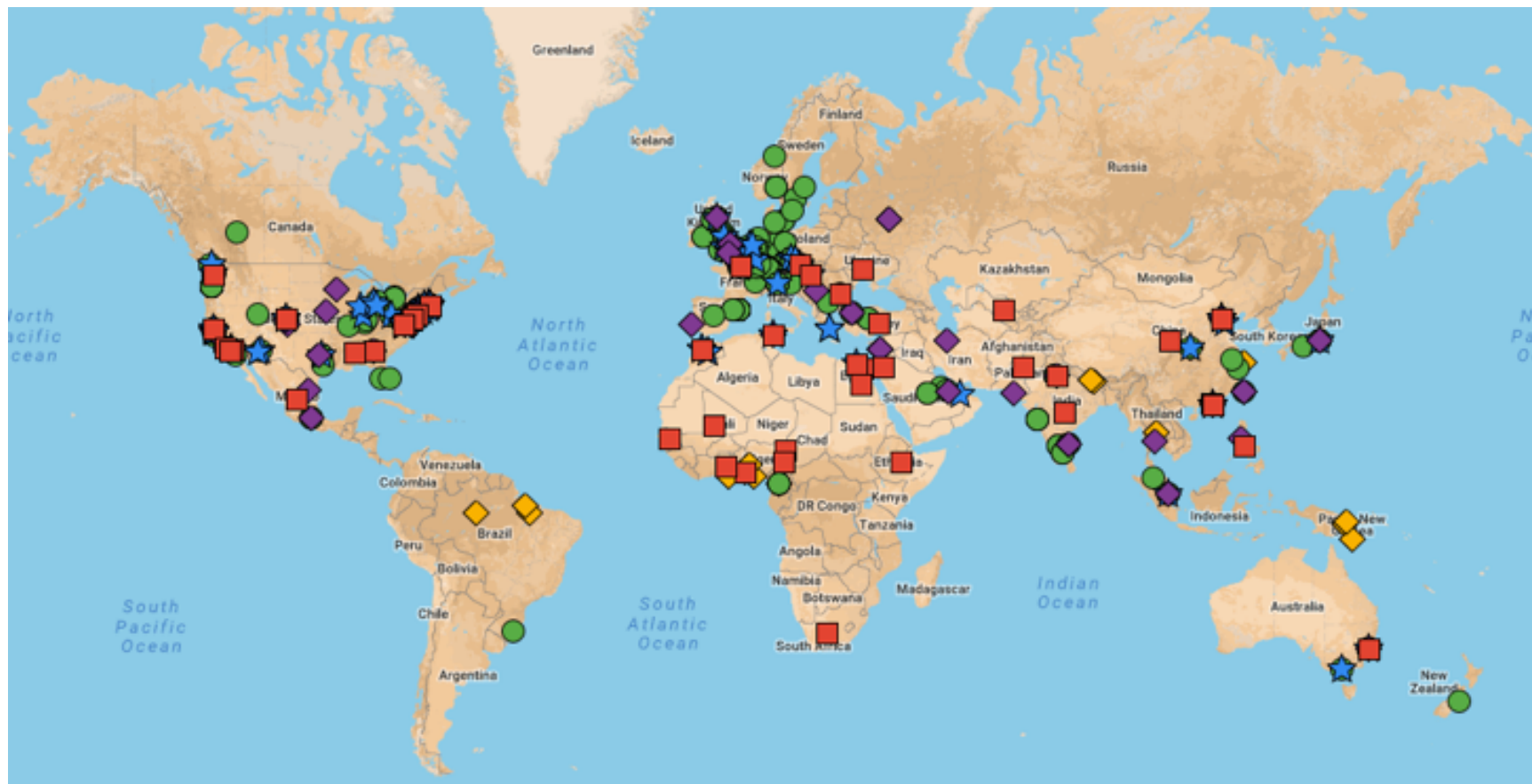
ccieri@ldc.upenn.edu

- ◆ Crowdsourcing is a relatively new term for an old idea made more practical by modern technology
 - *Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.* (Howe 2006)
 - *The practice of obtaining information or services by soliciting input from a large number of people, typically via the internet and often without offering compensation* (Oxford English Dictionary, 2020)
 - A change in source to a large, undefined crowd via open call without compensation.
 - The term and current practice allow for things beyond work on well know platforms like Mechanical Turk.
 - For example, Games with a Purpose (GWAPs) and Citizen Linguistics (that is, the citizen science of language).

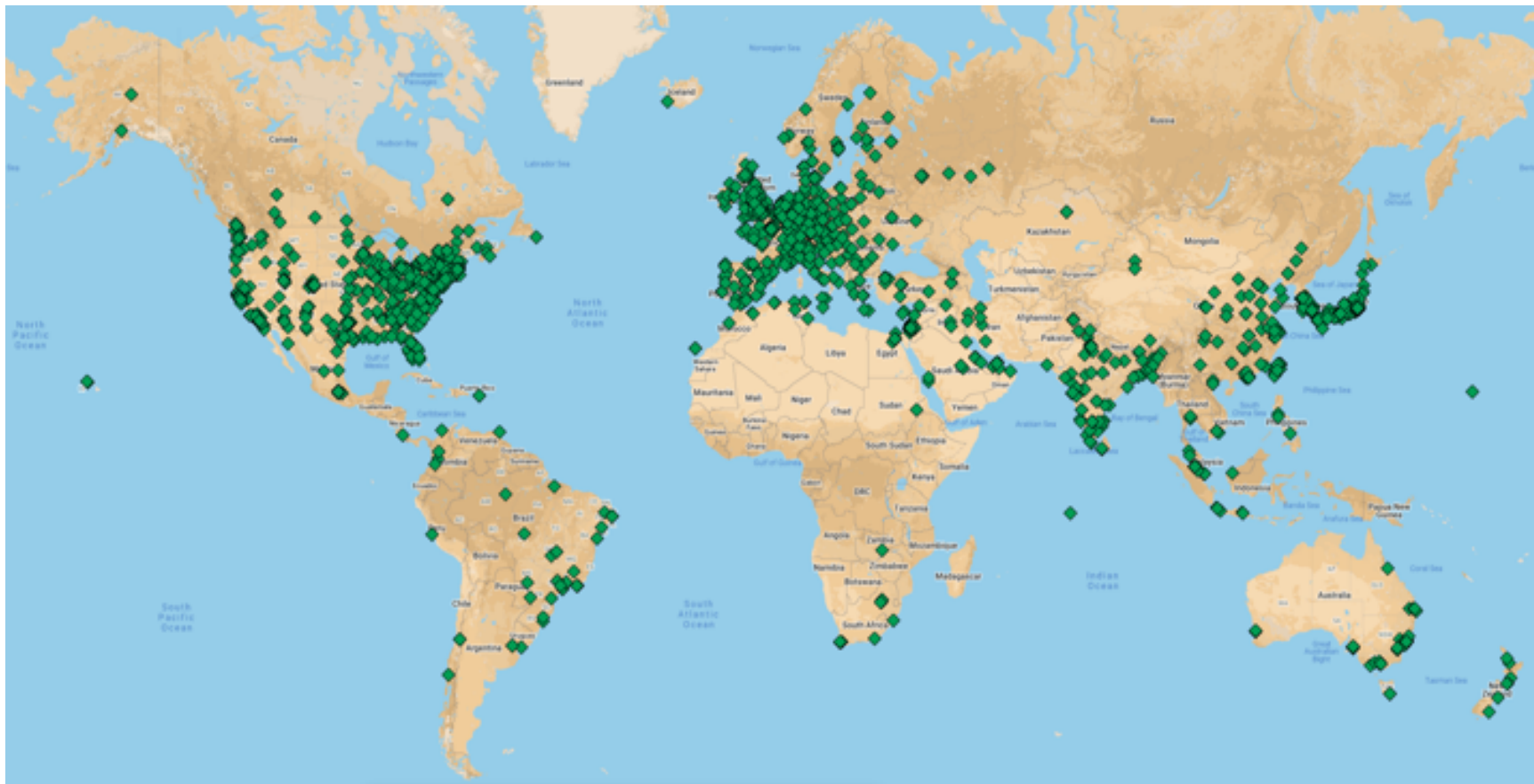
- ◆ I work at LDC. LDC collects, creates and shares data
 - > 175,000 copies of > 3,100 corpus titles in > 90 languages
 - > 40 data & > 100 annotation types (transcription, translation, POS tags)
- ◆ All available to Penn faculty, staff, students @ catalog.ldc.upenn.edu



The screenshot shows the LDC Catalog website. At the top, there is a navigation bar with the LDC logo, the text "Linguistic Data Consortium", and links for "UNIVERSITY OF PENNSYLVANIA", "CONTACT US", and social media icons. Below the navigation bar is a sidebar menu with categories like "ABOUT", "MEMBERS", "COMMUNICATIONS", "LANGUAGE RESOURCES", "Data", "Obtaining Data", "Catalog", "By Year", "Top Ten Corpora", "Projects", "Search", "Memberships", "Data Scholarships", "Tools", "Papers", "LR Wiki", "DATA MANAGEMENT", and "COLLABORATIONS". The main content area is titled "LDC Catalog" and includes a sub-header "Home > Language Resources > Data". Below this, a text box states: "LDC's Catalog contains hundreds of holdings. Use the buttons below to browse, search, and view catalog entries." There are five buttons arranged in a grid: "by year" (corpora sorted by release year), "top ten corpora" (the ten most-distributed LDC corpora), "projects" (LDC corpora attributed to project-based research), "search" (search the Catalog by corpus name, catalog number, language, etc.), and "memberships" (current LDC memberships). At the bottom of the page, a copyright notice reads: "© 1992-2021 Linguistic Data Consortium, The Trustees of the University of Pennsylvania. All Rights Reserved."

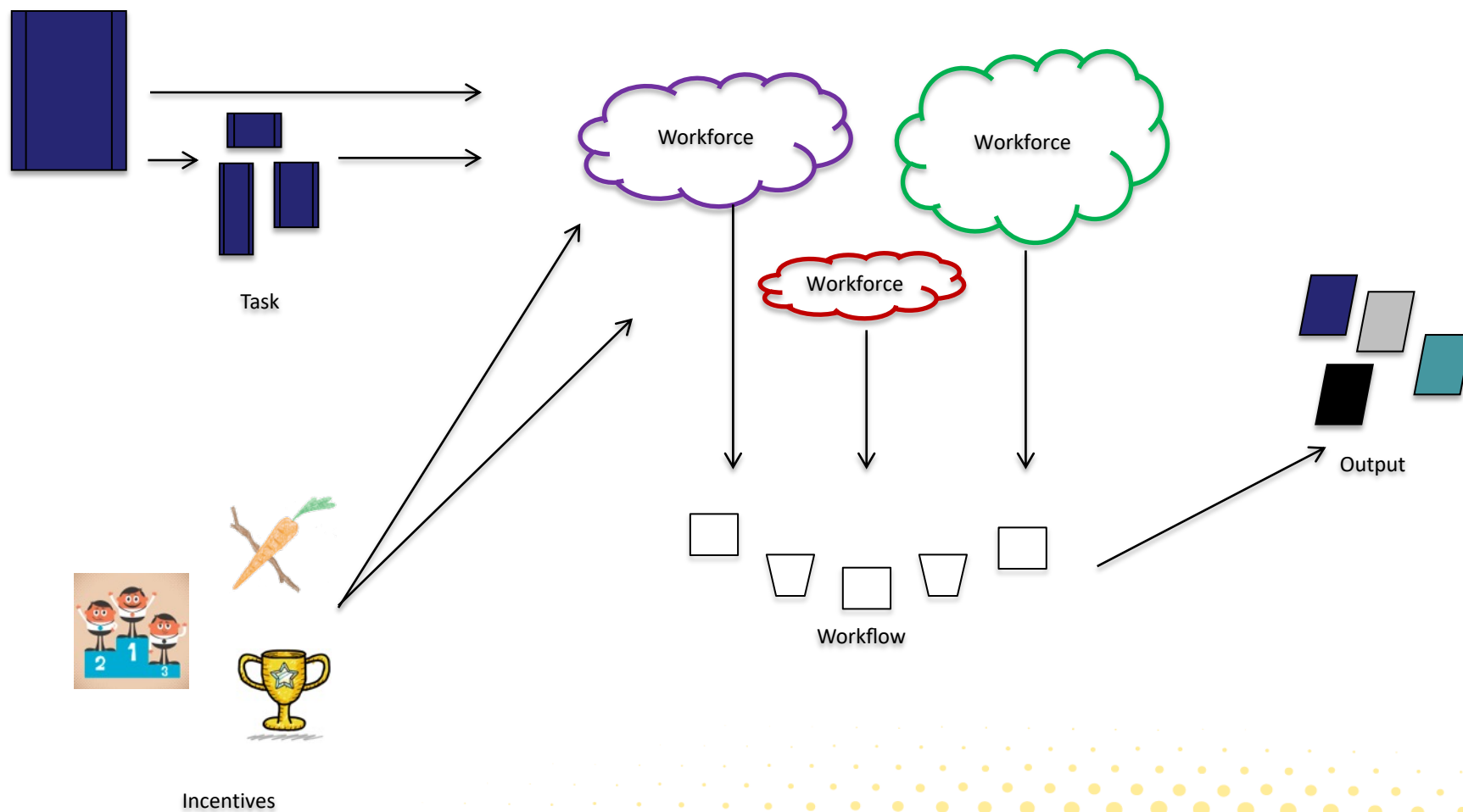


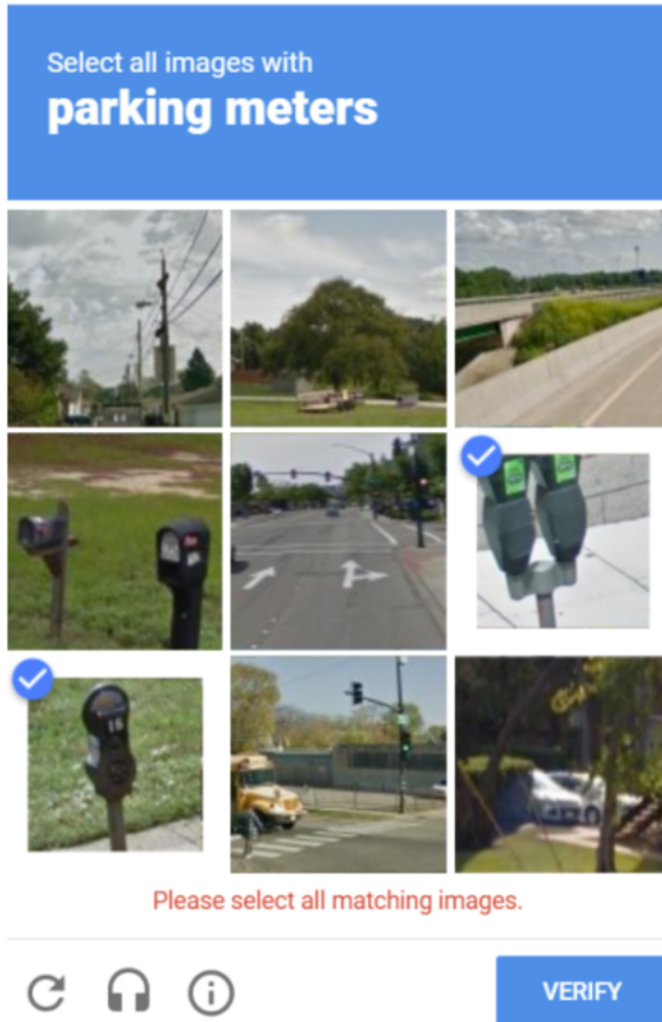
LDC Global Network of select data sources including: ■ = subcontractors and vendors, ● = corpus authors, ◆ = media providers, ◆ = LDC staff collections, ★ = research collaborators. Many markers represent multiple collaborators; many markers partially obscured by others.



Each marker represents a city with between 1 and 101 data users

- ◆ And we're not the only source of Language Data
 - Other international Data Centers
 - ELRA/ELDA
 - SADIaR (RMA)
 - CLDC
 - LDC-IL
 - National-Corpus-like efforts in:
 - Austria, Croatia, Czech Republic, Germany, Great Britain, Hungary, Ireland, Malta, Netherlands, Poland, Romania, Russia, Slovakia, South Tyrol, Switzerland, United States, Wales
 - also many international
 - multisite projects (META-NET, CLARIN)
 - research labs (CNRS-ICL, Essex, QMUL, CMU/LTI)
 - community bake-offs (CONLL, SemEval)
- ◆ But, all together, we're still far from providing the data needed because we use limited resources to tackle a much larger problem.





Task

- apparently web security via Turing test
- but also train AI in image recognition

User Incentive

- access to a site

Workforce

- users eager to access the site

Workflow

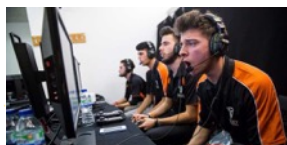
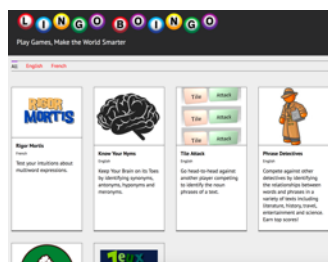
- select target images from larger set
- 9 items, target items change
- System evaluates based on expectations
 - pre-trained on images
 - expects 3 (sometimes 4) responses
- Progress blocked or 2nd pass required if 'incorrect'

Outcome

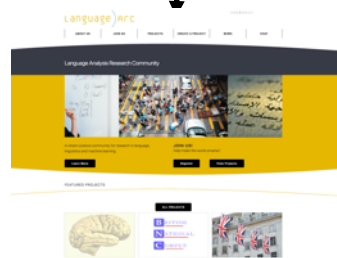
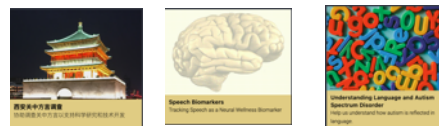
- Some users select mailboxes as parking meters affecting the training data



Games



Citizen Linguists



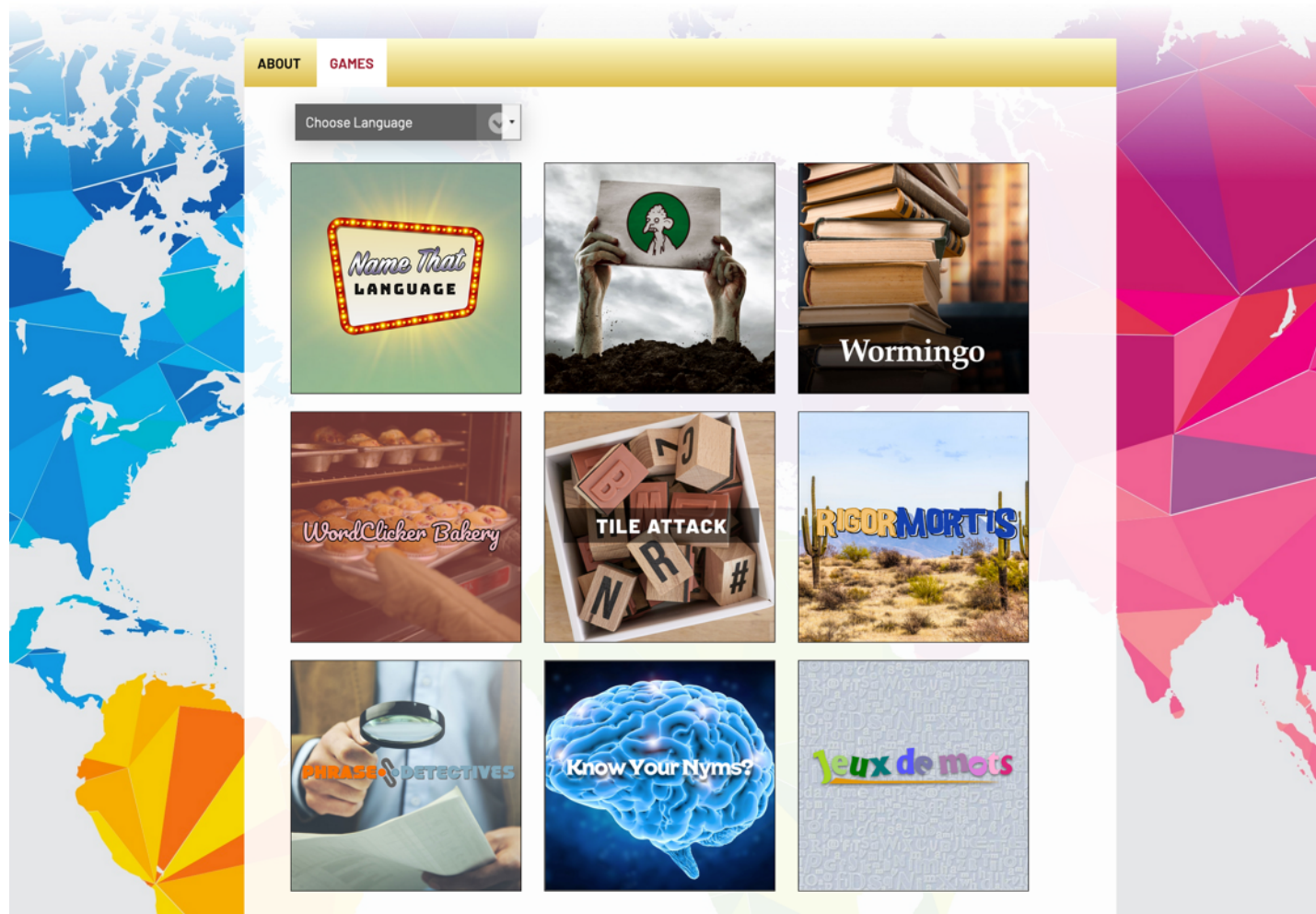
Language Pros

Labov Archive



LINGO **B O I N G O**

Play Games. Make the World Smarter.





Play Games, Make the World Smarter [Learn More](#)

Name That Language!

Score:	0
Round:	1
Lives:	2

1.242



you chose Chinese, Incorrect :(
the answer was Turkish

Chinese

Portuguese

Turkish

New Game

Next

- data type, source
- language selection
- # instances/language
- distractor selection
- # distractors
- game play variants
- scoring
- feedback
- competition

namethatlanguage.org



Language Arc

ABOUT US | JOIN US | PROJECTS | CREATE A PROJECT | NEWS | CHAT

Language Analysis Research Community

A citizen science community for research in language, linguistics and machine learning.

JOIN US!
Help make the world smarter!

Learn More Register View Projects

FEATURED PROJECTS

ALL PROJECTS

BRITISH NATIONAL CORPUS
PERFECTING THE AUDIO BNC

FROM COCKNEY TO THE QUEEN

SOUTHERN AFRICA CDI

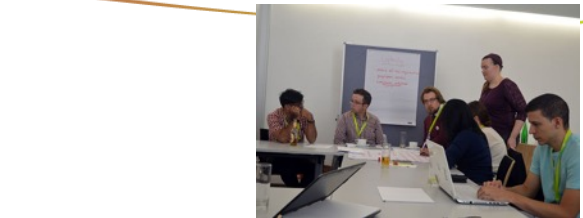
PARTNERS

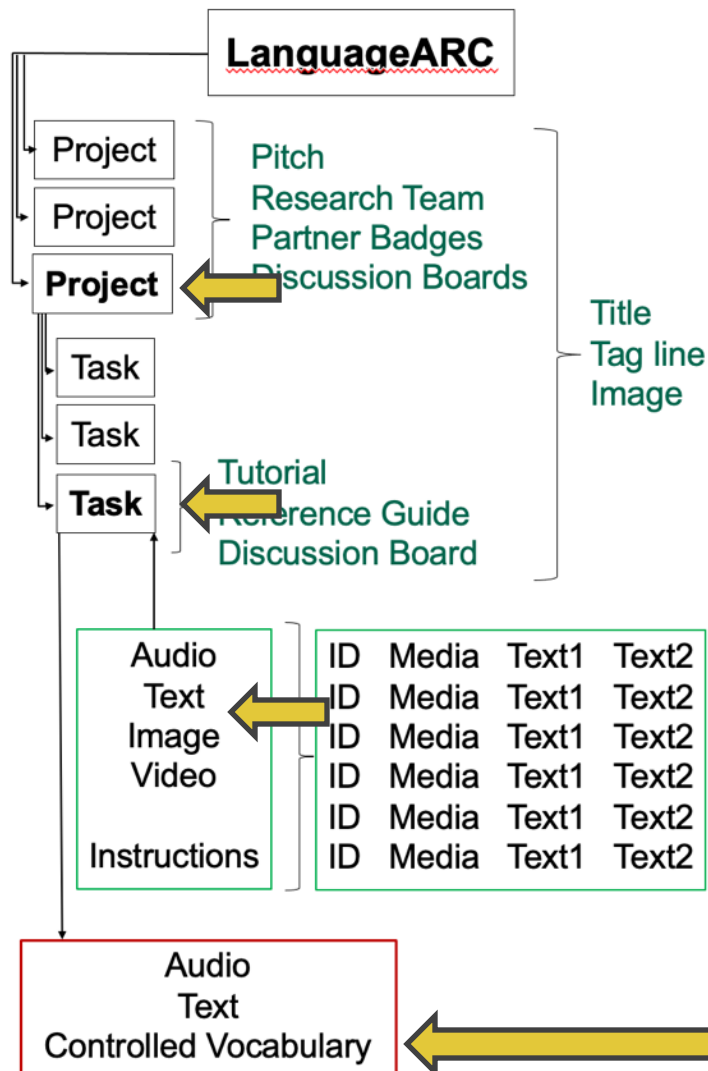
NSF LDC

Portions © 2019 University of Pennsylvania
LanguageArc is funded in part with support from the National Science Foundation under Grant No. 1730377

ABOUT US
FAQ
TERMS OF SERVICE
CONTACT

Facebook LinkedIn





Projects

1 2 3 Next Last

Perfecting the Audio BNC
Contribute to the British National Corpus, an internationally renowned language resource

Cockney to the Queen
Help us understand how people speak across London and Southeast England

Southern Africa CDI
Help document child language development in the languages of South Africa

Tongue Twisters
Record tongue twisters; identify and classify speech errors in others' recordings

FROM COCKNEY TO THE QUEEN
Help us understand how people speak across London and Southeast England

Tasks

ETHNICITY
See if you can guess the ethnicity of the speakers.

Continue

From Cockney to the Queen
Location

ABOUT ANNOTATE OUR RESEARCH TEAM CHAT Tutorial Reference

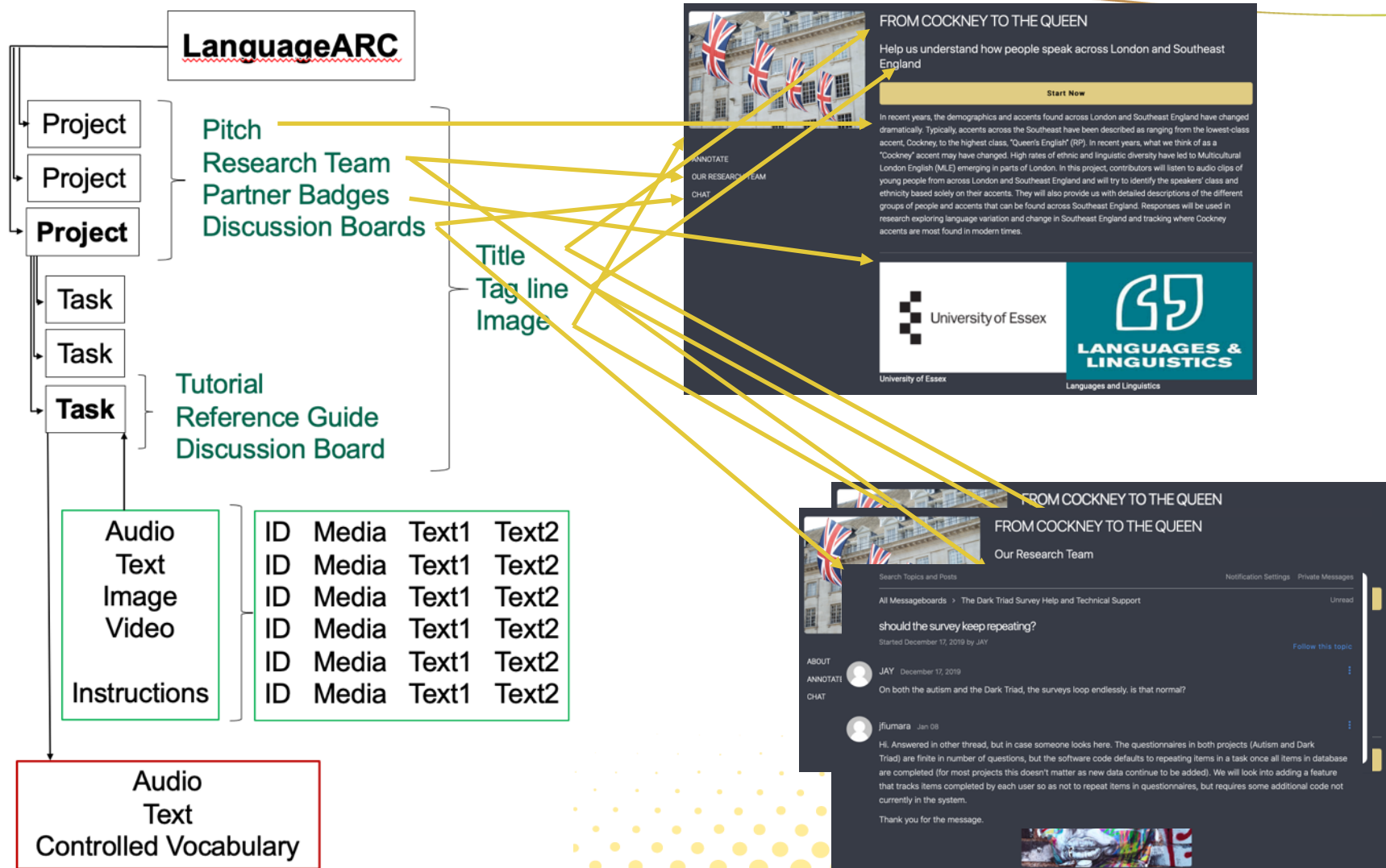
Please listen to the audio clip and then select the home location you believe the speaker is most likely to come from.

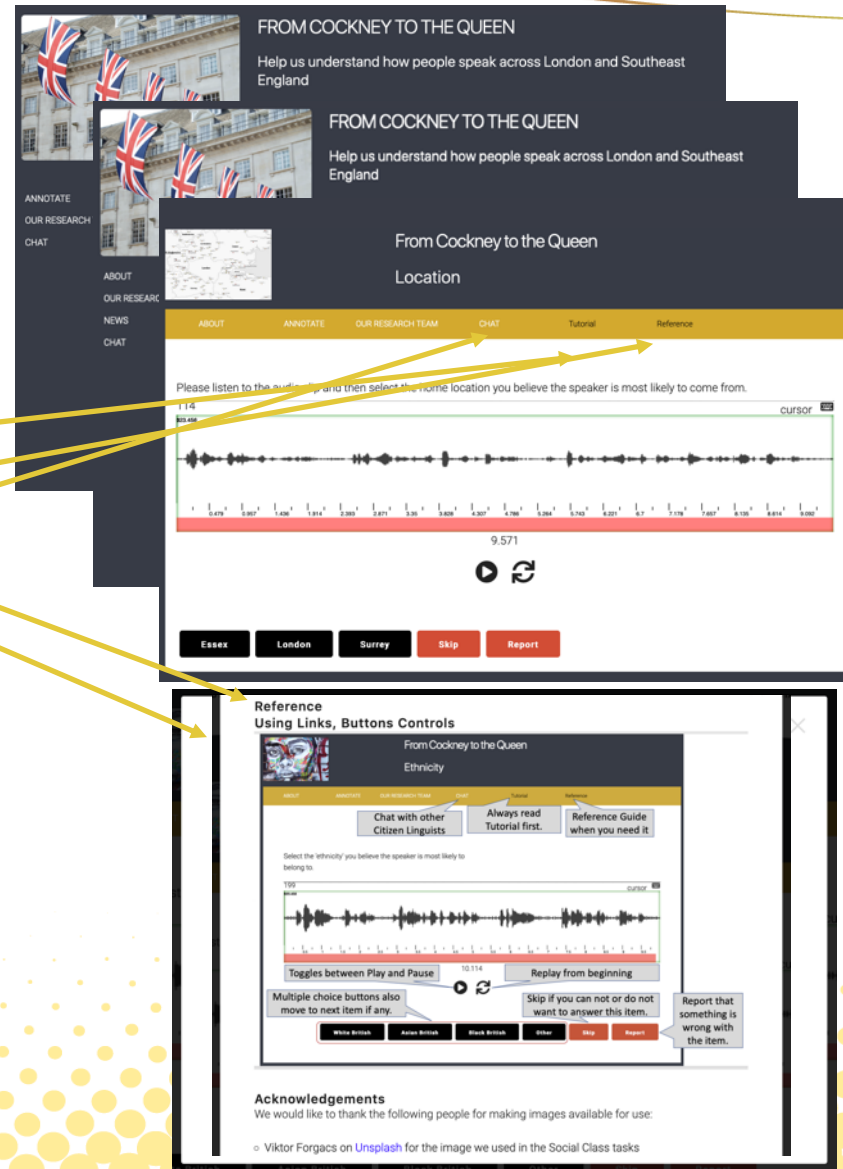
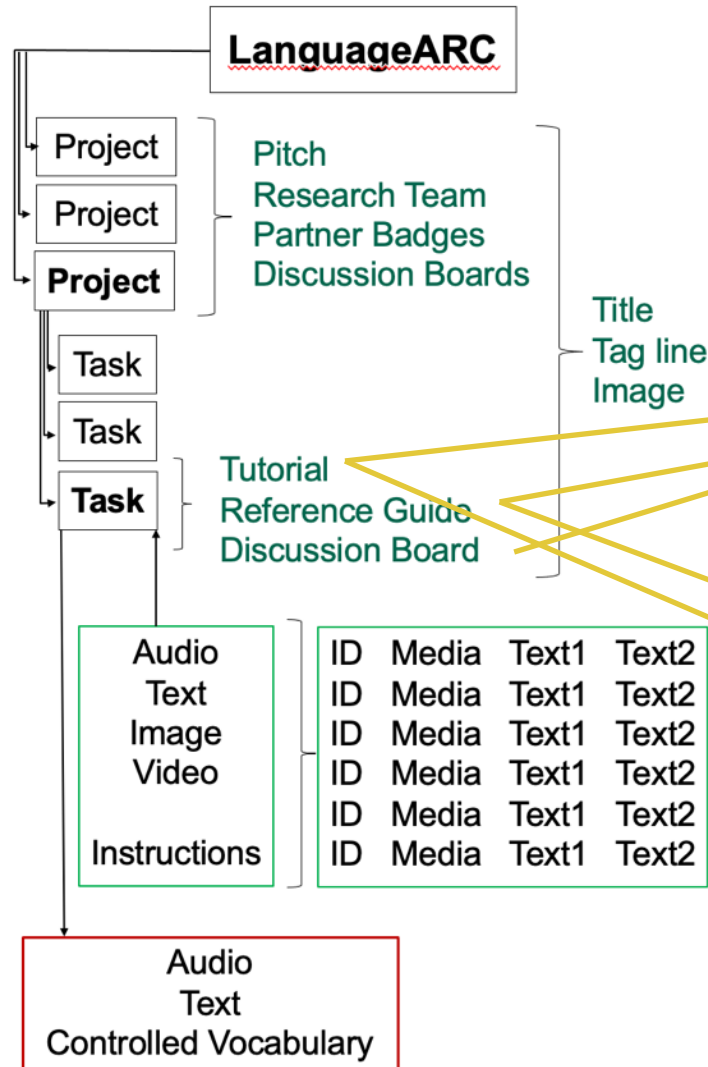
114

0.478 0.507 0.436 0.514 0.385 0.271 0.335 0.828 0.307 0.186 0.284 0.743 0.221 0.7 0.178 0.007 0.105 0.014 0.007

9.571

Essex London Surrey Skip Report





◆ Planning

- Consider end goal, data and annotations required.
- What incentives can you offer?
 - impacts, challenge, learning, community
- Who will be attracted? What (subset of the) work can they do?
- What support can you offer?
 - e.g. tutorial, reference guide, examples, discussion fora, blogs
- What are appropriate (homogeneous) tasks?
- Prepare data converting it into appropriate granularity and format
 - ~ 1 minute per item (HIT)

◆ Complete 4 Forms (<1 hour in total)

- Project
- Task
- Dataset
- Tool

New Project

Name (short internal name, used in menus, globally unique)

Title (displayed on Project Page, globally unique)

Subtitle (displayed on Project Page)

About your project and tasks (accepts [markdown](#))

News / Blog URL (leave blank if none)

Project Image File (displayed on project page)

 Browse... No file selected.

Image Filename (if in project assets)

Project Assets

No assets uploaded. **Upload Assets**

 Browse... No files selected.

Create Project Forums

- ☒ Announcements
- ☒ General Discussion
- ☒ Questions for Research Team
- ☒ Help and Technical Support

Research Team Members

 Add Researcher

Partners

 Add Partner

 Save

Project Builder

Step 1: Create or Select Project

Step 2: Create or Select Task

Step 3: Upload Dataset

Step 4: Create Tool

First you must create or select a project. This project may include just one task, or several.

You must provide a name, title, subtitle, about, and image.

Create Project

OR

Choose Existing Project

New Task

Task Name (short internal name, used in menus, unique within project)

Task Title (unique within project)

Task Description (accepts [markdown](#))

Tutorial (accepts [markdown](#))

Reference Guide (accepts [markdown](#))

Order of items assignment

"In order" - every contributor gets items in the same order "Random" - every contributor gets the same items in a unique, randomized order

☒ In Order ☐ Random

Within or across contributors?

"Within contributors" means each user will eventually be assigned all items, either in order or randomly as selected above "Across contributors" means items will be assigned across users based on order (user 1 might get items 1-10, then user 2 gets 11-20; no user gets the same item unless ITEM_LIMIT is set)

☒ Within contributors ☐ Across contributors

Task Image (optional)

Browse... No file selected.

Image Filename (if already in project assets)

Create Task Forum

☒ General Discussion

Save

New Dataset

Dataset Name

Dataset Description

Manifest File (The manifest is text file with rows separated by newline, the first row is a header, and columns separated by tabs.)

Browse... No file selected.

Data Files (audio, images, videos, etc)

Browse... No files selected.

Randomize manifest items order? This will affect all users of the manifest.

☐ Yes ☒ No

Save

Create Tool from Template

Nothing is saved until the very end when you hit "Save".

Exercise Specific Text (displays within task with each working kit)

Media Type (required) ☒ Text ☐ Audio ☐ Image ☐ Video

Media Content Column (column header in input)

Include language selection? ☒ Yes ☐ No

Limit language selection? ☒ Yes ☐ No

Languages to display

Prompt ID Field (from manifest, used to identify item and results in output) (required)

Include Primary Item Specific Text? ☒ Yes ☐ No

Primary Item Specific Text (column header in input) (required)

Primary Item Specific Text Label

Include Secondary Item Specific Text? ☒ Yes ☐ No

Secondary Item Specific Text (column header in input) (required)

Secondary Item Specific Text Label

Include Response Audio (record response to stimulus)? ☒ Yes ☐ No

Include Level Test? ☐ Yes ☒ No

Include Level Meter? ☐ Yes ☒ No

Include Response Text (translation, transcription, etc)? ☒ Yes ☐ No

Annotation Field Name (saved in db with this name) (required)

Annotation Field Label

Judgment Buttons (one per line). Judgment buttons move to next annotation and are stored in a judgment field. If no buttons are specified, a "Submit" button will be added.

Multiple Choice? The above will not display as buttons, but instead as checkboxes. There will be a Submit button automatically added. ☐ Yes ☒ No

Allow skip? ☐ Yes ☒ No

Allow "report bad item"? ☐ Yes ☒ No

 Save

New Project

Name (short internal name, used in menus, globally unique)

Title (displayed on Project Page, globally unique)

Subtitle (displayed on Project Page)

About your project and tasks (accepts **markdown**)

News / Blog URL (leave blank if none)

Project Image File (displayed on project page)

 No file selected.

Image Filename (if in project assets)

Project Assets

No assets uploaded. **Upload Assets**

 No files selected.

Create Project Forums

- ☒ Announcements
- ☒ General Discussion
- ☒ Questions for Research Team
- ☒ Help and Technical Support

Research Team Members

Partners

- ◆ Mostly just completing a form
- ◆ Allows upload of media into project assets
 - image, audio, video
 - For use in any 'markdown' field
- ◆ Accepts standard markdown
 - familiar from wikis ?
 - headers, bullets, URLs
 - enhanced to inline locally stored media assets
- ◆ Trivial to create fora
 - harder to keep up interactions
- ◆ Optional Research Team
 - may motivate some contributors
 - name, title, description, photo
- ◆ Optional Partner Badges
 - may motivate some (other) contributors
 - name, image, URL

New Task

Task Name (short internal name, used in menus, unique within project)

Task Title (unique within project)

Task Description (accepts markdown)

Tutorial (accepts markdown)

Reference Guide (accepts markdown)

Order of items assignment

"In order" - every contributor gets items in the same order

"Random" - every contributor gets the same items in a unique, randomized order

☒ In Order ☐ Random

Within or across contributors?

"Within contributors" means each user will eventually be assigned all items, either in order or randomly as selected above "Across contributors" means items will be assigned across users based on order (user 1 might get items 1-10, then user 2 gets 11-20); no user gets the same item unless ITEM_LIMIT is set.

☒ Within contributors ☐ Across contributors

Task Image (optional)

Browse... No file selected.

Image Filename (if already in project assets)

Create Task Forum

☒ General Discussion

Save

◆ Items in the data set can be assigned

- in the order of the dataset (manifest)
- in random order, unique for every contributor
- each contributor assigned entire data set
 - maximizes judgements per item (especially of the early items in a large data set)
- data set split across contributors
 - maximizes items judged at least once

◆ Optional task specific forum

The screenshot shows a 'New Dataset' form with the following fields and options:

- Dataset Name:** A text input field.
- Dataset Description:** A text input field.
- Manifest File:** A section with a description: '(The manifest is text file with rows separated by newline, the first row is a header, and columns separated by tabs.)'. It includes a 'Browse...' button and the text 'No file selected.'.
- Data Files (audio, images, videos, etc):** A section with a 'Browse...' button and the text 'No files selected.'.
- Randomize manifest items order?** A section with the text 'This will affect all users of the manifest.' and two radio buttons: 'Yes' and 'No' (which is selected).
- Save:** A black button at the bottom.

Yellow arrows from the text on the right point to the 'Manifest File' section, the 'Data Files' section, and the 'Randomize manifest items order?' section.

◆ Manifest

- Text file, with header
- One item per line, tabs separate columns
- Must have ID
- Media, text1, text2

◆ Media in standard formats

- WAV, TXT, MP4, PNG, JPG, GIF

◆ Dataset can undergo a one-time randomization

Create Tool from Template

Nothing is saved until the very end when you hit "Save".

Exercise Specific Text (displays within task with each working kit)

Media Type (required) ☒ Text ☐ Audio ☐ Image ☐ Video

Media Content Column (column header in input)

Include language selection? ☒ Yes ☐ No

Limit language selection? ☒ Yes ☐ No

Languages to display

Prompt ID Field (from manifest, used to identify item and results in output) (required)

Include Primary Item Specific Text? ☒ Yes ☐ No

Primary Item Specific Text (column header in input) (required)

Primary Item Specific Text Label

Include Secondary Item Specific Text? ☒ Yes ☐ No

Secondary Item Specific Text (column header in input) (required)

Secondary Item Specific Text Label

Include Response Audio (record response to stimulus)? ☒ Yes ☐ No

Include Level Test? ☐ Yes ☒ No

Include Level Meter? ☐ Yes ☒ No

Include Response Text (translation, transcription, etc)? ☒ Yes ☐ No

Annotation Field Name (saved in db with this name) (required)

Annotation Field Label

Judgment Buttons (one per line). Judgment buttons move to next annotation and are stored in a judgment field. If no buttons are specified, a "Submit" button will be added.

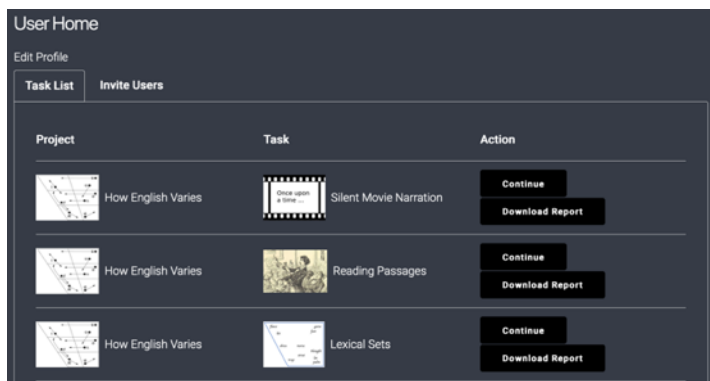
Multiple Choice? The above will not display as buttons, but instead as checkboxes. There will be a Submit button automatically added. ☐ Yes ☒ No

Allow skip? ☐ Yes ☒ No

Allow "report bad item"? ☐ Yes ☒ No

Save

- ◆ Instructions
- ◆ Form reads column headers from manifest
- ◆ Options
 - media: text, audio, image, video
 - language selector powered by SIL list of ~7000 language and ~30,000 language names; can create custom list
 - 1-2 texts that can vary per item
 - audio recorder, level meter
 - text entry field
 - buttons to collect judgements multiple choice or not
 - allow contributors to skip items
 - allow contributors to report bad item
- ◆ Every input field can be named as you like.
- ◆ Tool selects reasonable layout for the combination of widget/controls selected.
- ◆ ~All choices can be edited later.



◆ Dashboard

- access every project you joined
- with appropriate credentials
 - report on all projects & task you manage
 - invite and manage uses
 - manage team roles and team

◆ Report

- project ID, status
- task ID, status
- dataset ID
- user ID
- ~ city, state, country
- date, time
- item and response

◆ Also control status of project

- prototype, private, public (by request)

Project ID	Project Status	Task ID	Task Status	Tool ID	Dataset ID	User ID	Country Code	City	Time	Judgment	Prompt ID
7	Published	24	Published	21	23	6	US	Fayetteville	2019-11-11 03:03:53 +0000		97
7	Published	24	Published	21	23	3	US	Philadelphia	2019-11-11 13:37:43 +0000	skipped	21
7	Published	24	Published	21	23	17	AU	Hobart	2019-12-03 12:41:48 +0000	Essex	131