



NETS 213 : Crowdsourcing and Human Computation - Guest Lecture

Citizen Linguistics (with LanguageARC)

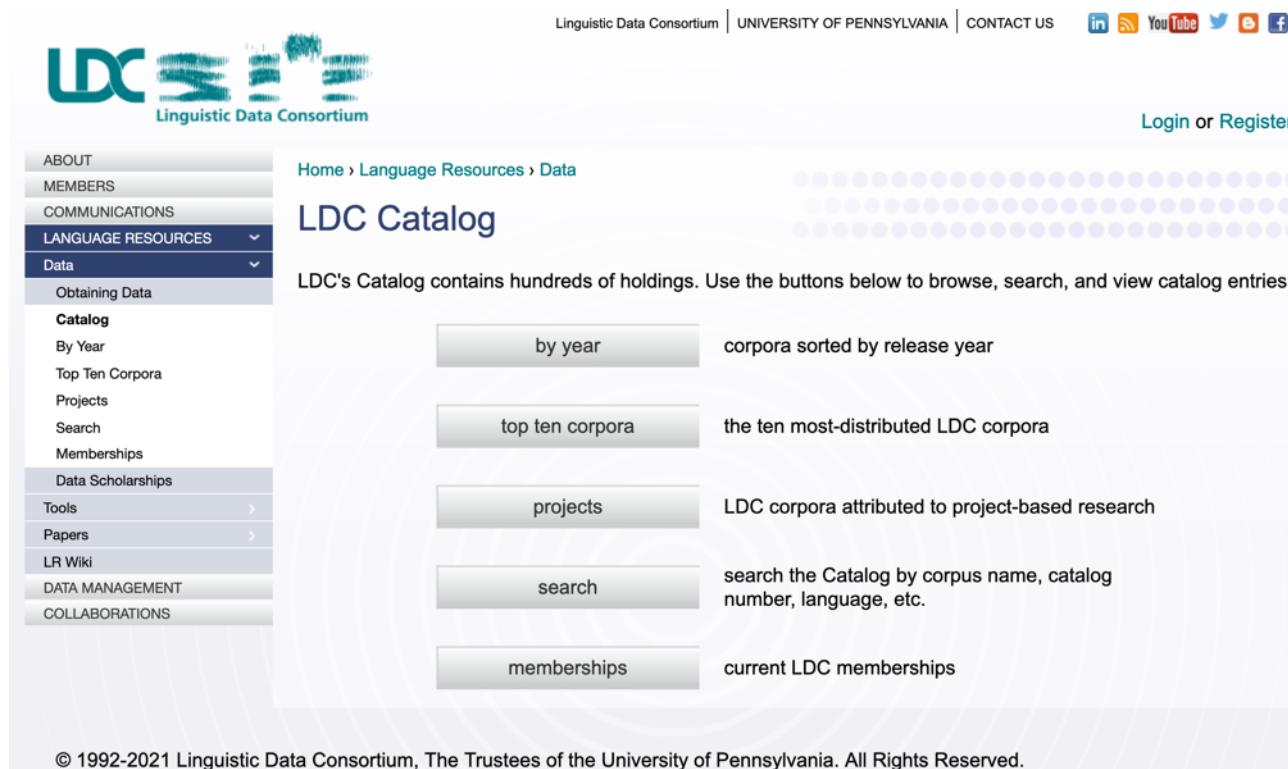
Christopher Cieri

*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 known 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



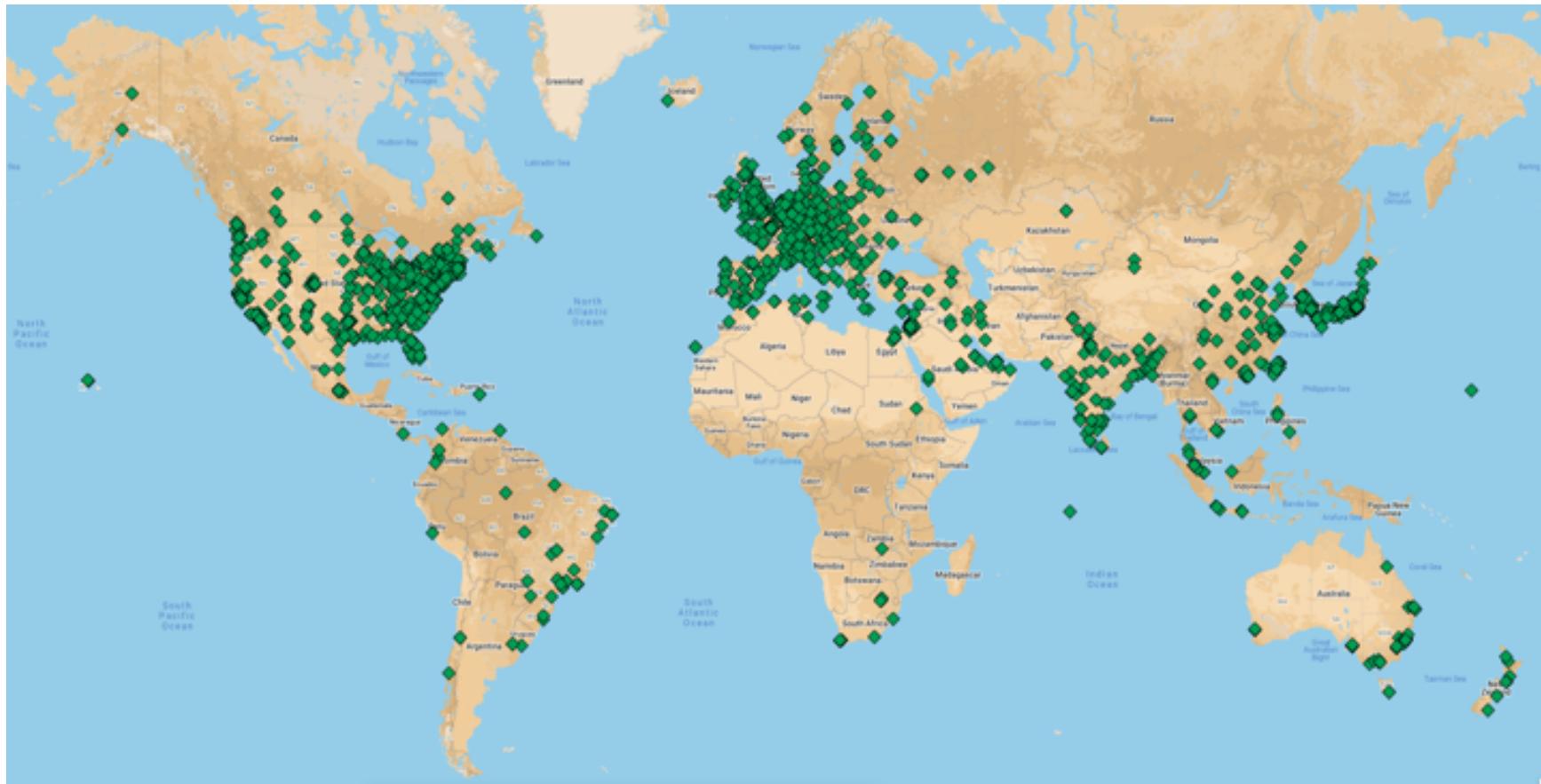
The screenshot shows the LDC Catalog homepage. At the top, there's a navigation bar with links for "Linguistic Data Consortium", "UNIVERSITY OF PENNSYLVANIA", "CONTACT US", and social media icons for LinkedIn, RSS, YouTube, Twitter, and Facebook. On the right, there are "Login or Register" buttons. The main content area has a decorative background of purple dots. On the left, a sidebar menu is open under "LANGUAGE RESOURCES". The menu items are: ABOUT, MEMBERS, COMMUNICATIONS, LANGUAGE RESOURCES (with a dropdown arrow), Data (which is selected and highlighted in blue), and Obtaining Data. Under "Data", there are links for Catalog, By Year, Top Ten Corpora, Projects, Search, and Memberships. Below the sidebar, the page title is "LDC Catalog". A sub-header says "LDC's Catalog contains hundreds of holdings. Use the buttons below to browse, search, and view catalog entries." There are five buttons with corresponding descriptions: "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 Data Contributors



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.

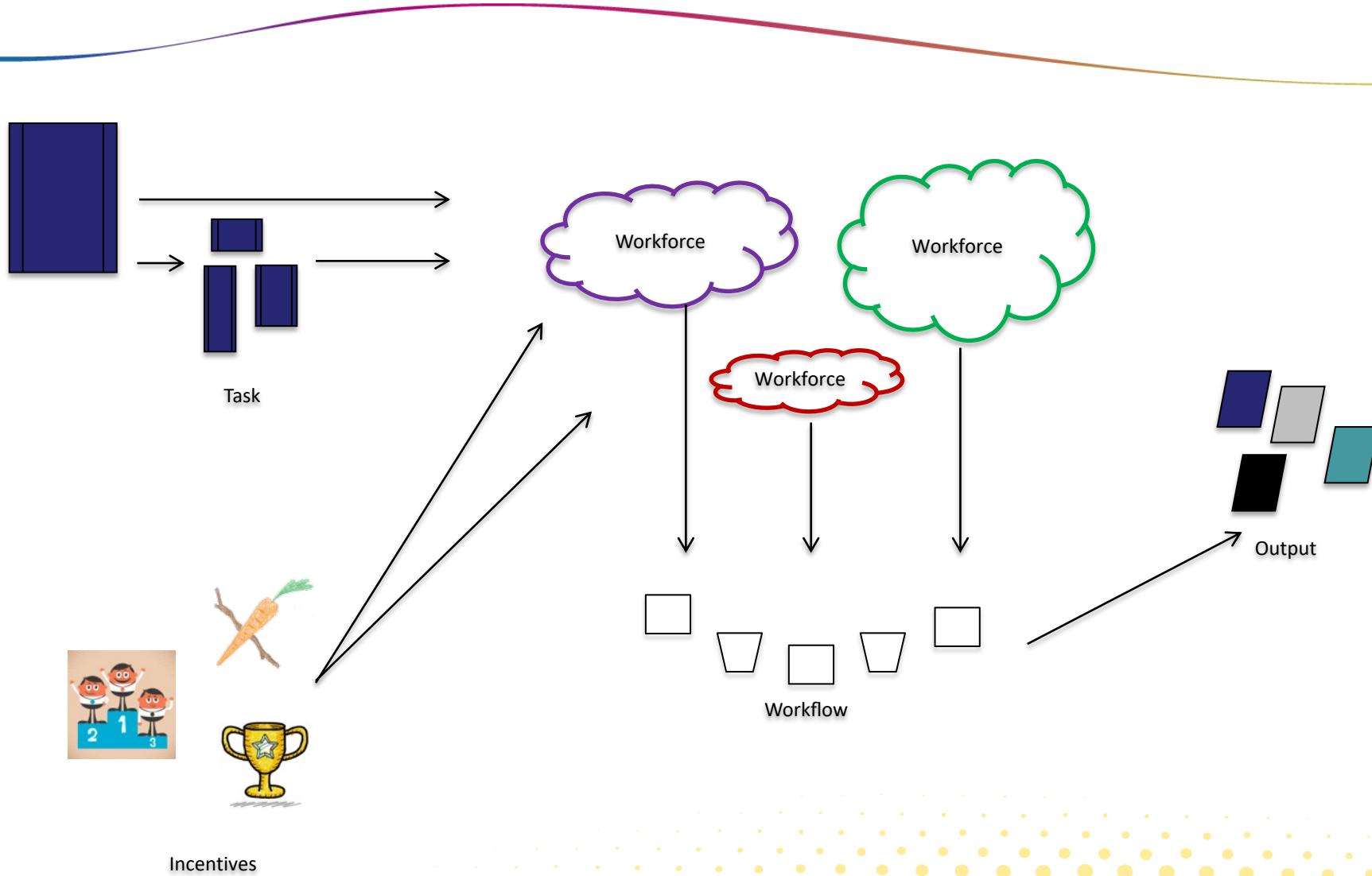
LDC Data Users



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
 - SADILaR (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.

Rethinking Data Creation



When Worlds Collide

Select all images with
parking meters

Please select all matching images.

VERIFY

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

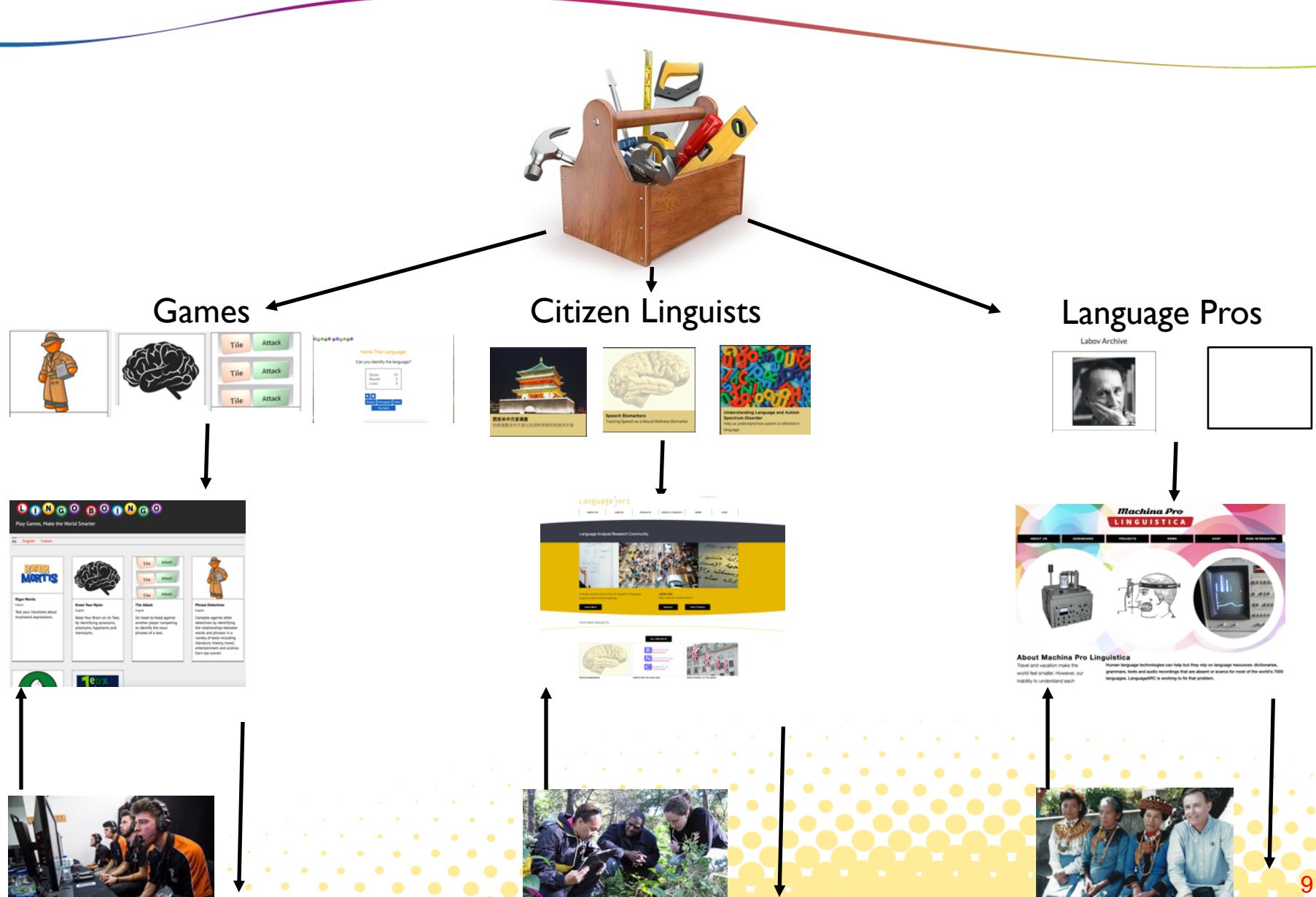
Workflow

- 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

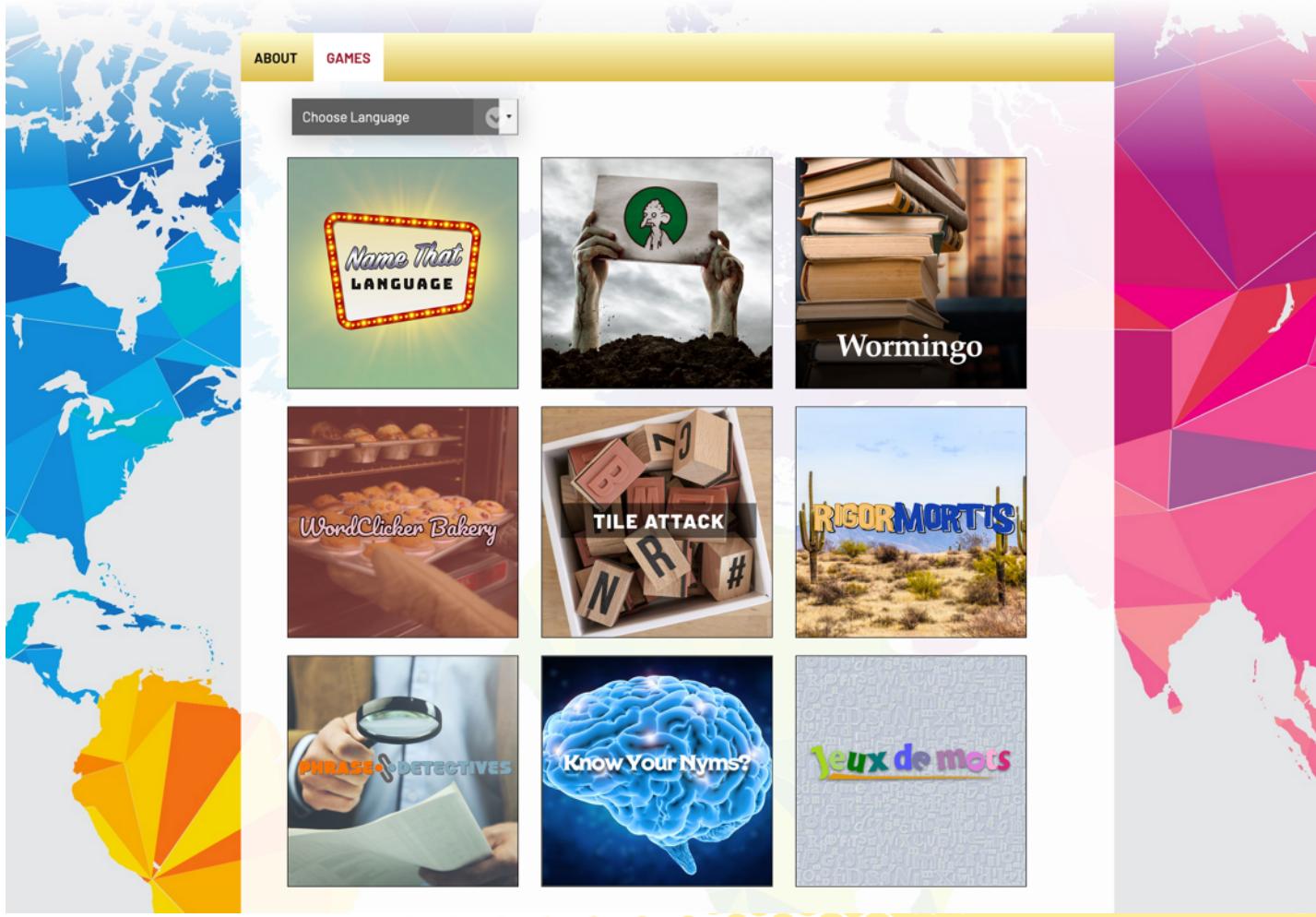
- Some users select mailboxes as parking meters affecting the training data

Approach



LINGO 

Play Games. Make the World Smarter.



The screenshot shows the LINGO website's homepage. At the top, there is a navigation bar with 'ABOUT' and 'GAMES' tabs, and a 'Choose Language' dropdown menu. Below the navigation bar, there are nine game cards arranged in a 3x3 grid:

- Name That LANGUAGE**: A game where players guess languages based on visual cues.
- WordClicker Bakery**: A game where players click on words hidden in a bakery scene.
- TILE ATTACK**: A game involving wooden blocks with letters and symbols.
- Wormingo**: A game featuring a stack of books and a worm.
- RIGOR MORTIS**: A game set in a desert landscape.
- PHRASE & DETECTIVES**: A game involving a magnifying glass and a document.
- Know Your Nyms?**: A game featuring a 3D rendering of a human brain.
- Jeux de mots**: A game set against a background of a crossword puzzle grid.



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



LanguageARC

ABOUT US | JOIN US | PROJECTS | CREATE A PROJECT | NEWS | CHAT | LOGOUT | SIGN OUT

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

B R I T I S H N A T I O N A L C O R P U S | FROM COCKNEY TO THE QUEEN | SOUTHERN AFRICA CDI

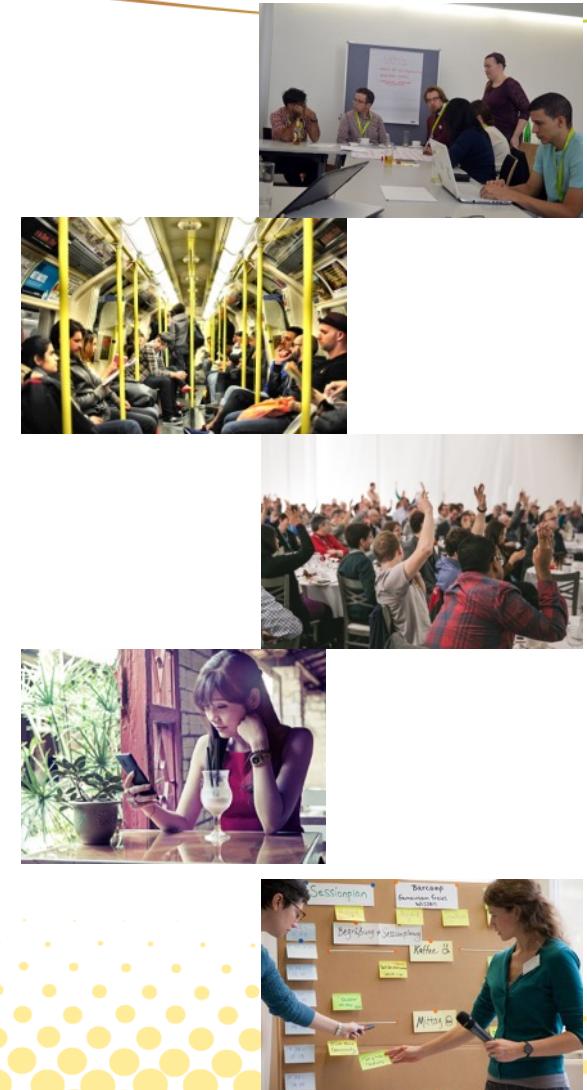
PERFECTING THE AUDIO BNC

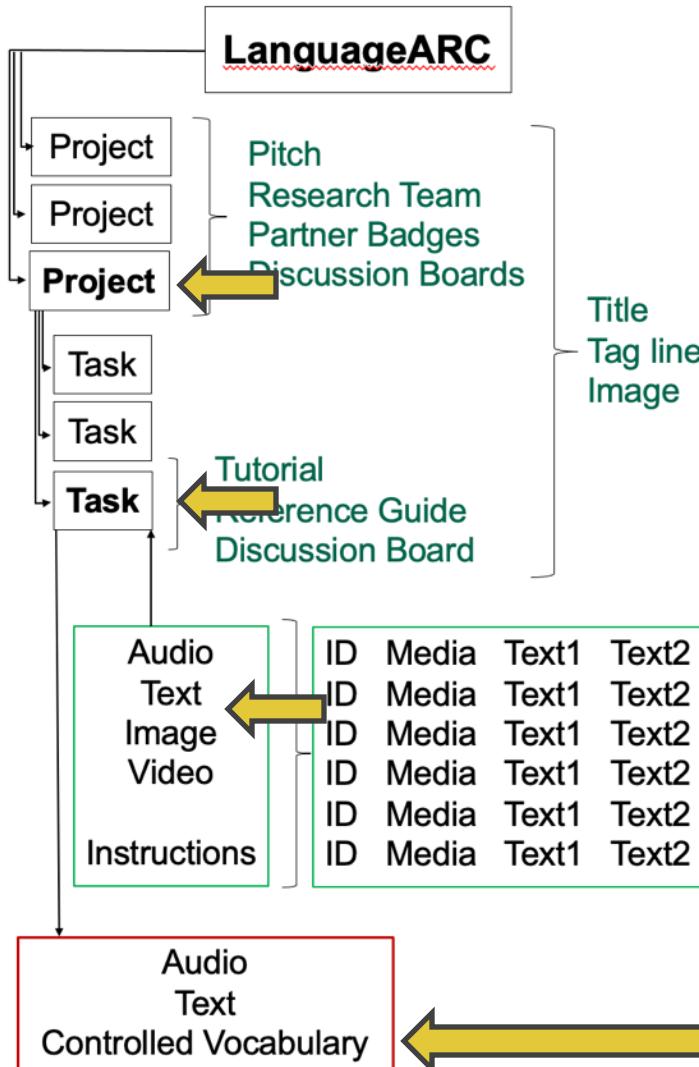
PARTNERS

Portions © 2019 University of Pennsylvania. LanguageARC is based upon work supported by National Science Foundation under Grant No. 1730277.

ABOUT US | FAQ | TERMS OF SERVICE | CONTACT

f in

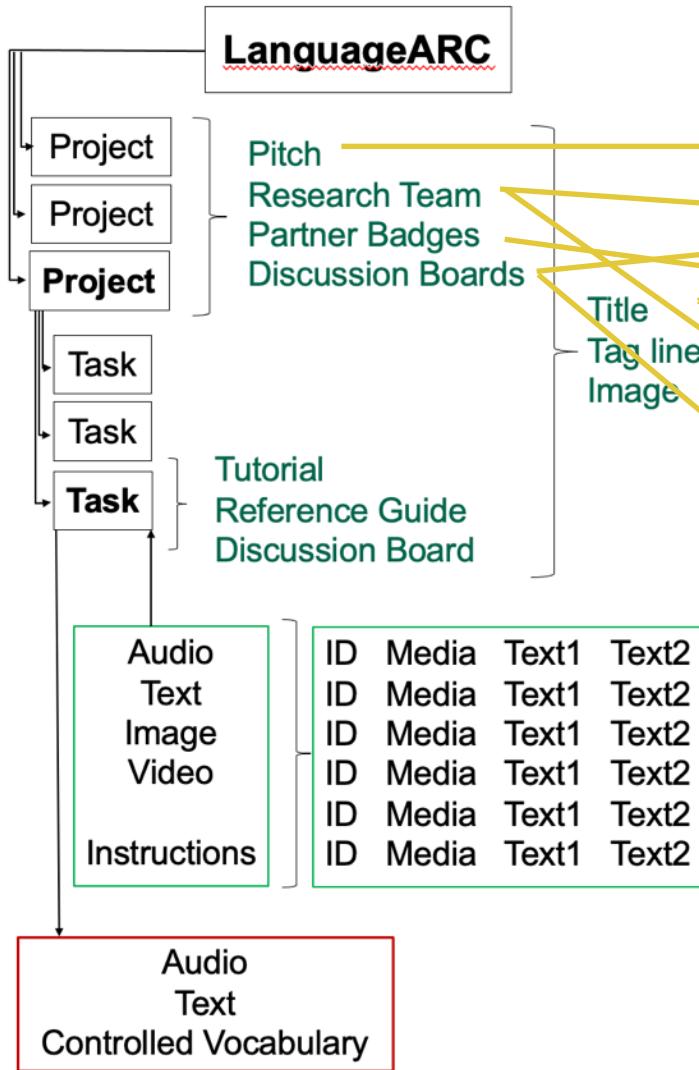




The screenshot shows the LanguageARC task interface. At the top, there's a navigation bar with 'Projects' and page numbers 1, 2, 3, 'Next >', and 'Last >'. Below this are three project cards:

- BRITISH NATIONAL CORPUS**: Perfecting the Audio BNC. Contributing 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.

The 'Cockney to the Queen' card is highlighted. Below it, there's a section titled 'FROM COCKNEY TO THE QUEEN' with the sub-instruction 'Help us understand how people speak across London and Southeast England'. It includes a 'Tasks' section, an 'ETHNICITY' image with the instruction 'See if you can guess the ethnicity of the speakers.', and a 'Location' section with a map and the instruction 'Please listen to the audio clip and then select the home location you believe the speaker is most likely to come from.' At the bottom, there are buttons for 'ABOUT', 'ANNOTATE', 'OUR RESEARCH TEAM', 'CHAT', 'Tutorial', and 'Reference'.



FROM COCKNEY TO THE QUEEN

Help us understand how people speak across London and Southeast England

In recent years, the demographics and accents found across London and Southeast England have changed dramatically. Typically, accents across the Southeast have been described as ranging from the lowest-class accent, Cockney, to the highest class, 'Queen's English' (RP). In recent years, what we think of as a 'Cockney' accent may have changed. High rates of ethnic and linguistic diversity have led to Multicultural London English (MLE) emerging in parts of London. In this project, contributors will listen to audio clips of young people from across London and Southeast England and will try to identify the speakers' class and ethnicity based solely on their accents. They will also provide us with detailed descriptions of the different groups of people and accents that can be found across Southeast England. Responses will be used in research exploring language variation and change in Southeast England and tracking where Cockney accents are most found in modern times.

University of Essex

Languages & Linguistics

should the survey keep repeating?

Started December 17, 2019 by JAY

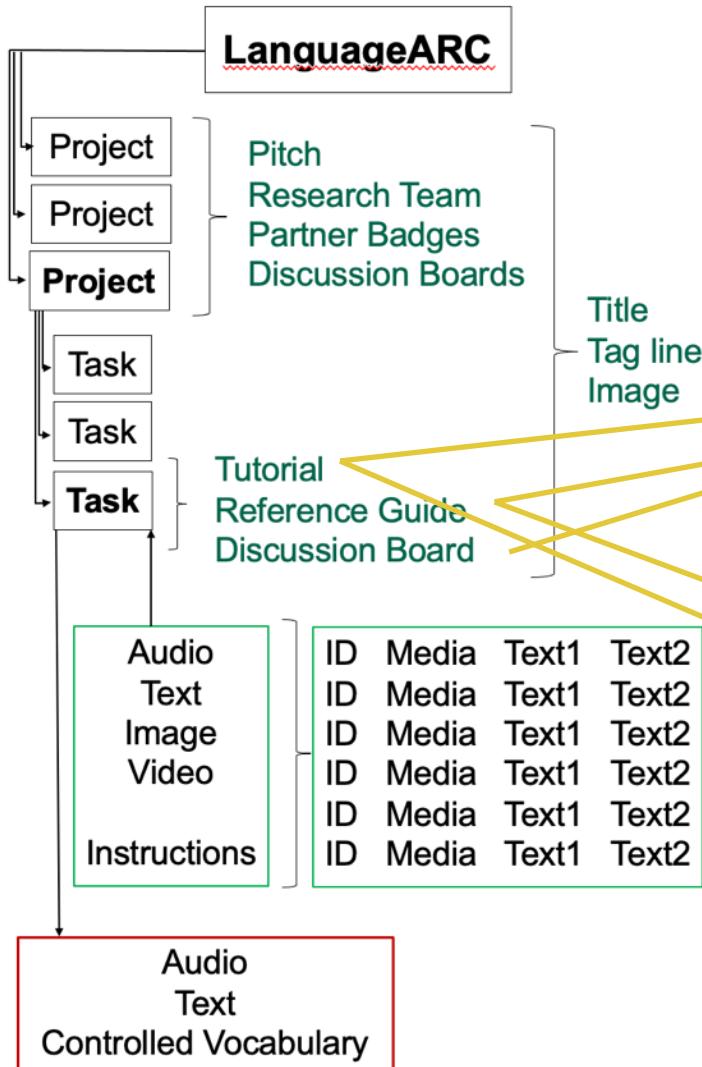
JAY · December 17, 2019

On both the autism and the Dark Triad, the surveys loop endlessly. Is that normal?

Jfumara · Jan 08

Hi. Answered in other thread, but in case someone looks here. The questionnaires in both projects (Autism and Dark Triad) are finite in number of questions, but the software code defaults to repeating items in a task once all items in database are repeated (for most projects this doesn't matter as new data continue to be added). We will look into adding a feature that tracks items completed by each user so as not to repeat items in questionnaires, but requires some additional code not currently in the system.

Thank you for the message.



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

ANNOTE OUR RESEARCH CHAT

ABOUT OUR RESEARCH NEWS CHAT

From Cockney to the Queen Location

ABOUT ANNOTE OUR RESEARCH TEAM CHAT Tutorial Reference

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

cursor

9.571

Essex London Surrey Skip Report

Reference Using Links, Buttons Controls

From Cockney to the Queen Ethnicity

Chat with other Citizen Linguists Always read Tutorial first Reference Guide when you need it

Select the ethnicity you believe the speaker is most likely to belong to.

Toggles between Play and Pause Replay from beginning

Multiple choice buttons also move to next item if any. Skip if you can not or do not want to answer this item. Report that something is wrong with the item.

Acknowledgements

We would like to thank the following people for making images available for use:

© Viktor Forgacs on Unsplash for the image we used in the Social Class tasks

- ◆ 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

Project Builder

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.

No files selected.

Create Project Forums

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

Research Team Members

Partners

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.

OR

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 item 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)

No file selected.

Image Filename (if already in project assets)

Create Task Forum

General Discussion

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.)

No file selected.

Data Files (audio, images, videos, etc)

No files selected.

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

Yes No

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

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](#)

- ◆ 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)

No file selected.

Image Filename (if already in project assets)

Create Task Forum

General Discussion

- ◆ 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

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.)
 No file selected.

Data Files (audio, images, videos, etc)
 No files selected.

Randomize manifest items order? This will affect all users of the manifest.
 Yes No

◆ 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 & Report

User Home

Edit Profile Task List Invite Users

Project	Task	Action
 How English Varies	 Silent Movie Narration	Continue Download Report
 How English Varies	 Reading Passages	Continue Download Report
 How English Varies	 Lexical Sets	Continue Download Report

Project ID	Project Status	Task ID	Task Status	Tool ID	Dataset ID	User ID	Country Code	City	Time	Judgment ID	Prompt
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

◆ Dashboard

- access every project you joined
- with appropriate credentials
 - report on all projects & task you manage
 - invite and manage users
 - 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)