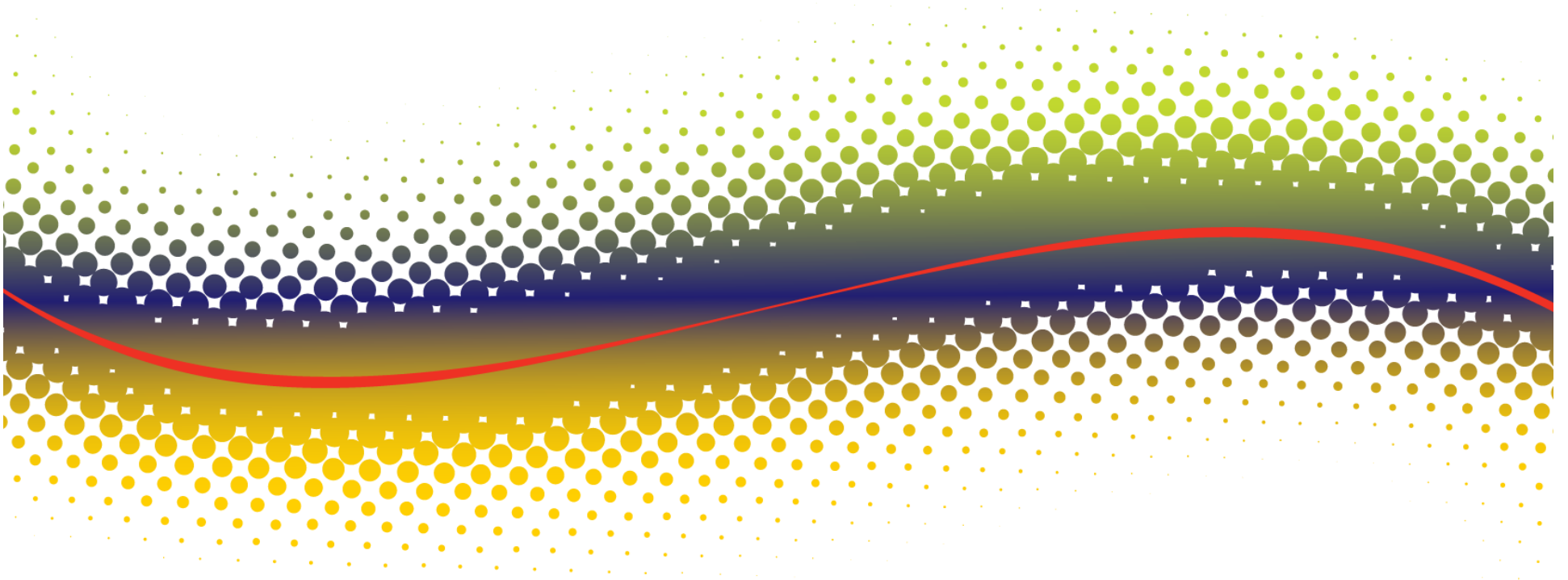




# **NIEUW: Novel Incentives and Engineering Unique Workflows**

Christopher Cieri, James Fiumara

Linguistic Data Consortium



- ◆ ~~1: summary of LDC / DC accomplish~~
- ◆ ~~2: not enough: impediments~~
- ◆ ~~3: lost opportunities~~
- ◆ ~~4: inspiration~~
- ◆ ~~5: features of solution~~
- ◆ 6: outline of infrastructure/portals (3 non-mutually exclusive)
  - ◆ 6a citizen science
  - ◆ 6b gaming
  - ◆ 6c language professionals
- ◆ 7 community: multiple only partially overlapping communities need to be brought together contributors, users, processor each part is at least 2 of these, maybe all three

- ◆ Founded 1992 as archive, publisher of language corpora
  - required to be self-sufficient within 5 years
  - model developed by committee from corporate, academic, government HLT
- ◆ Seeing unmet demand, expanded to
  - collection, annotation, software, best practice, service
- ◆ Benchmarks
  - 120,000 copies, 1860 titles, >3500 orgs, 71 countries, 92 languages
  - >11,000 research papers use LDC data
  - 35 different data collection paradigms
  - >80 different annotation types
  - >dozen 3-5 year multi-site technology programs
  - >110 NIST Technology Evaluations (also SIGHan, SemEval, CoNLL)
  - Data Grants
    - 69 recipients in 26 countries, 110 corpora valued >\$185,000, 64% acceptance rate

# LDC in 1 Slide and a map



corpora=red, media=purple, employees=blue, research collaborators=orange, collection/annotation subs=green

- ◆ Successful at our tasking to date
- ◆ However, despite our work over 24 years and that of:
  - other Data Centers: ELRA, Chinese LDC, LDC for Indian Languages, South African Resource Management Agency
  - programs: METANET, CLARIN
  - National Corpus Efforts: British & American English, Danish, Czech, Slovak, Icelandic, Russian, Turkish, Irish, Welsh

only begun to document world's 7000 languages

- ◆ Scaling LRs beyond current constraints, requires very different thinking
  - Exceeding program constraints time, budget, languages
  - Contributions that don't match current funding
    - William Labov, Shirley Brice Heath, Corky Feigin, Katie Drager, DASS
    - AfrAnaph, Shermin DaSilva, Keelan Evanini, SLAAP, U.AZ CallGrandma
    - Mixer Enthusiasts
  - Contributions that can not be compensated monetarily due to tax law, immigration law, export control

# Cost of Inadequate Coverage

- ◆ Language Resource cost < cost of their absence
  - impedes HLT development
  - impedes research on the language, cultural preservation
  - impedes understanding
  - impedes economic, political, humanitarian efforts
    - in disasters delivering effective relief requires language knowledge
      - collaborators at risk
      - porting HLTs to language comes too late for many, requires LRs
      - media unreliable, provides misinformation



## ◆ People contribute data when asked appropriately

- ◆ Wikipedia, Wiktionary
- ◆ Project Gutenberg
- ◆ LibriVox
- ◆ StoryCorps
- ◆ MOOCs
- ◆ DuoLingo
- ◆ (re-)Captcha
- ◆ Facebook (MySpace), Google+
- ◆ Twitter
- ◆ Linked-In, ResearchGate, Academia.edu
- ◆ Pinterest, Tumblr
- ◆ Mechanical Turk, CrowdFlower
- ◆ Phrase Detectives
- ◆ Zombilingo
- ◆ GWAP
- ◆ Google Image Labeler
- ◆ TrainRobots
- ◆ Fold-It
- ◆ OntoGalaxy
- ◆ The Great Language Game
- ◆ Quizz.us
- ◆ Zooniverse
- ◆ SPICE/RLAT
- ◆ Crowd Curio

## ◆ If we leave collection to commercial market:


- ◆ Where does the data go?
- ◆ Is the data appropriate for our uses?

- ◆ always available, any device, no access barrier, optional vetting
- ◆ connected to social networking, crowd sourcing for recruitment, etc
- ◆ connection to LDC infrastructure for collection
- ◆ self sustaining, seeded with LDC data, self-feeding, selection→prioritization
- ◆ any necessary training and evaluation are automated
- ◆ multiple (novel) incentives:
  - ◆ information, entertainment, self-expression, socializing, competence building (course credit), competition, status, prestige, recognition, payment, discounts (real-world and virtual), access to services (HLT) based on contributions, contributing to cause
- ◆ X-sourcing: task complexity, control, automation vary as appropriate
- ◆ results shared with contributors, research community, publicly
- ◆ general infrastructure (cf. Scribe, LDC WebAnn), multiple instances



# Components of a Solution

- ◆ accommodate users with different: authorization, profiles, skills
- ◆ evaluate skills & contributions, assign microtasks
- ◆ accept and display text, image, audio and video data & metadata
- ◆ virtually segment
  - ◆ text by characters
  - ◆ image by coordinates
  - ◆ audio by time
  - ◆ video by time and coordinates
- ◆ apply annotations to segments or, in tiers, to other annotations
- ◆ encode annotation as
  - ◆ scalar: text, number, true/false
  - ◆ node in taxonomy (controlled vocabulary)
- ◆ design workflow, integrate Human Language Technologies, adapt
- ◆ store data & annotations losslessly & permanently
- ◆ query and report data & annotations (report progress)
- ◆ model 'corpora' for various end-uses


Admin ▾ Home Namespaces Projects Profile Sign out **ccieri**

Task List **Enrollments** Reports

Show  entries Search:

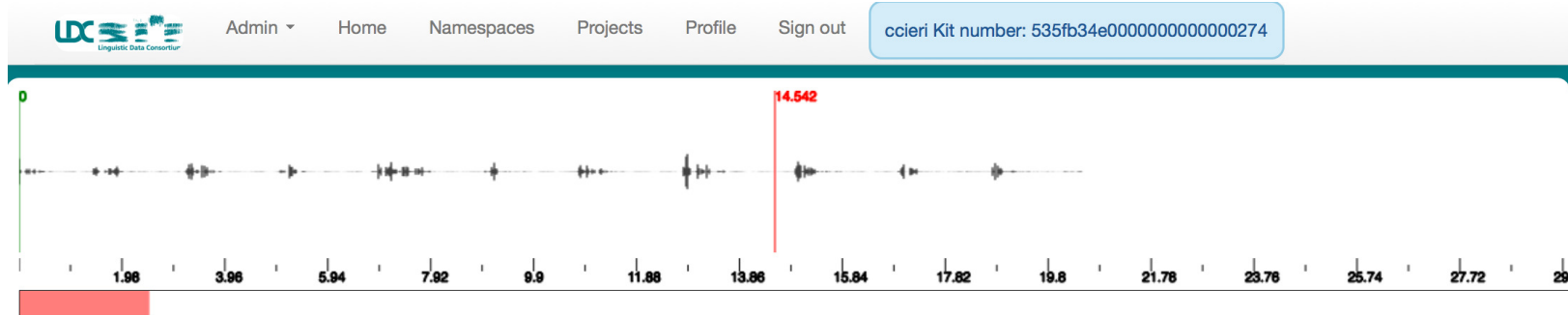
Project	Task	Action	Status
<a href="#">Admin</a>	namespaces	No Available Kits	N/A
<a href="#">Audio</a>	web trans demo	<a href="#">start</a>	N/A
<a href="#">Audio</a>	services	<a href="#">start</a>	N/A
<a href="#">DEFT</a>	ERE demo	<a href="#">continue</a>	in progress (paused)
<a href="#">HighEntropyASR</a>	word list triage	No Available Kits	not in progress
<a href="#">HighEntropyASR</a>	word list audit	<a href="#">continue</a>	in progress
<a href="#">HighEntropyASR</a>	hears reports	<a href="#">start</a>	N/A
<a href="#">VAST</a>	web_trans_test	<a href="#">continue</a>	in progress (paused)

Showing 1 to 8 of 8 entries
First Previous 1 Next Last

**Name:** ccieri  
**Account Type:** lead annotator  
**Created:** 2013-03-14  
**Updated:** 2014-05-29

[Help](#) [Home](#) [LDC](#)





Listen to each word in the kit. If the utterance represents a possible pronunciation, in any American dialect, of any sense of the word, click 'Possible'. If you hear a tiny bit of the preceeding or following word but also hear the target word in a possible pronunciation, choose 'Possible'. Otherwise click 'Impossible'. To begin, you'll need to click the first word. However, after that, clicking the Possible or Impossible buttons will record your decision and play the next word. You can listen to any word a second time if necessary by clicking it.

**nippier**

Possible Impossible

**trashy**

Possible Impossible

**spite**

Possible Impossible

**circumlocutions**

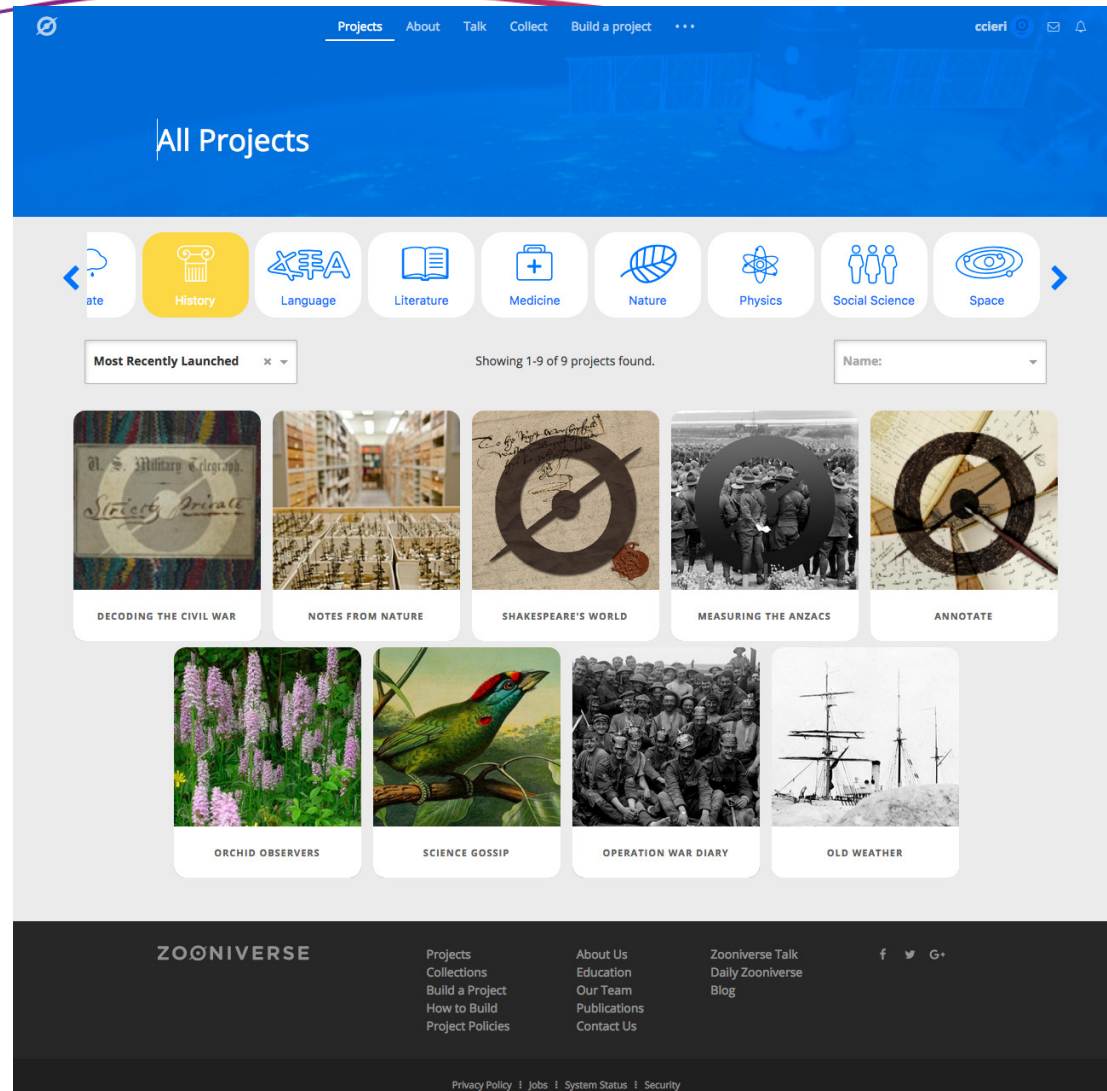
Possible Impossible

**sank**

Possible Impossible

**outages**


Possible Impossible



The screenshot displays the Curio Citizen Science Portal interface. At the top, the 'Curio' logo is on the left, and a 'Log In' button is on the right. Below the header, there are six project cards arranged in a 2x3 grid. Each card features a background image, a title, a brief description, and a 'Learn More' button.

- Thoreau's Field Notes**: How does climate change affect the timing of when buds, flowers and fruits come out?
- Dive4Oceanography**: What can we learn from documenting and studying oceanic diving records?
- Typewriter Girl**: Transcribing Victorian literature can be both fun and exciting!
- CrowdEEG**: Help scientists understand the inner-workings of sleep-staging through EEG classification.
- Ensemble**: Can the untrained ear accurately transcribe music?
- UrbanEars**: Sift through the Sounds of Cities





Science we can do together.

[log in / sign up](#)

[home](#)
[project finder](#)
[events](#)
[our blog](#)

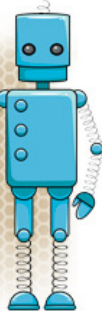
## project finder

pick an activity ▾

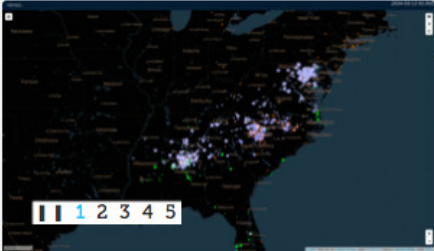
pick a topic ▾

search keywords

[advanced search](#)



## projects of the day



**mPING**

**goal** Collect global observations of weather conditions at the ground

**task** Use app on iOS or Android to submit weather conditions

[Anywhere](#)

### our blog

Guest post by: Egle Marija Ramanauskaite  
Some of you have been keen to hear more news about the project to ...

[read more](#)

### newsletter

Get awesome projects delivered to your inbox each week:





### for scientists

Add a project and we'll connect you to our community of doers!





[premium services](#)

[edit your project or event](#)



### as seen in

### supported in part by

[About](#) | [Contact](#) | [Partners](#) | [Advertise](#) | [API](#) | [Research](#) | [Terms of Use](#) | [Privacy Policy](#)  
 © Copyright 2016 SciStarter.com, a division of Science for Citizens LLC

Great Language Game Play Languages Stats ▾ About



The Great Language Game

Amongst the thousands of languages spoken across the world, here are just eighty. How many can you distinguish between?

Play >



## PHRASE • DETECTIVES

### LOGIN

**USERNAME**

**PASSWORD**

**Login**

Not joined yet?  
[Register here.](#)

### Welcome to Phrase Detectives

Lovers of literature, grammar and language, this is the place where you can work together to improve future generations of technology. By indicating **relationships** between **words and phrases** you will help to create a resource that is rich in linguistic information. Simply register a username and password and you can get started.

[Start here](#)

### Play on Facebook

Play Phrase Detectives on Facebook, which features a new **head-to-head mode** where you play against an expert and **team play** where you score double points if you agree with your friends.

[Play Facebook version](#)

facebook

### 549 docs completed

The most recent was **Alice in Wonderland** (Lewis Carroll) completed by **julie3164** on 20 Mar 2016

[See all](#)

### Detective's Bulletin

**13 Jan 14**

A Facebook group has been set up to discuss some of the more interesting cases of ambiguity found in the Phrase Detectives game. Join the group [Do You Know Your Anaphor From Your Elbow?](#) (Facebook account required).

**19 Apr 12**

Analysis of player motivations will be presented at Collective Intelligence 2012 conference in Boston, MA. It seems the female players are coming out top!

**22 Feb 11**

The long-awaited Phrase Detectives Facebook game has gone live.

[Play the game.](#)

**01 Nov 10**

Here are a few game stats for you number junkies. We have over 1.5 million examples of human language in the database submitted by 3500 players, a collaborative effort of over 2700 hours or 112 days. On average players are quicker to disagree with other players than to agree. Exported data from the game shows that the combined answers of players gives a very high quality result. A huge thank you to all detectives.

### Quick instructions

You must search for **relationships** between **words and phrases** in a piece of text.

**1) NAME THE CULPRIT**

You will be given a word or phrase and you must look for any evidence of it appearing earlier in the text. An example of this would be:

Sherlink Holmes went to the shop. **He** got some tobacco for his pipe.

The word in **orange** refers to "Sherlink Holmes".

### TOP SCORES

**THIS WEEK**  
 cgibbs 569

**THIS MONTH**  
 chox123 229

### LEADERBOARD

WEEK	MONTH	ALLTIME
cgibbs		569
chox123		229
jon		112
magooogy		80
julie3164		46
JMS		33
JRS		27
VB		19
Jemsypie		8
filipk		7
gully		7
turquoise123		7
LucyBlades		5
Sherlink_Holmes		5
jayjay		5
rascal		5
LouiseO'Brien		3
Tobrien		3
KarenLeemc		2
HeIce		1

### MOST RECENT

**Wulfhære of Mercia** (Wikipedia) submitted by **Tobrien**

[Feedback](#)

[Instructions](#)

[FAQ](#)

**SHARE THIS**

## ZOMBI LINGO

### ATTRAPEZ LES TOUTES

JOUE POUR AIDER LES SCIENTIFIQUES !

SUIS MES RÈGLES, IDENTIFIE LES TÊTES ET MANGE-LES.

ATTENTION AUX PIÈGES, ILS SONT NOMBREUX !



SIGNÉ  
Prof. Frankensperrier.

**JOUER**

*Pas de limite pour toi !  
Tu accèdes à toutes les options, bonus cachés !*

**ESSAYER**

*Cette version limitée va te rendre accro !  
Mais tu ne pourras pas sauvegarder !*

**CHALLENGE** MORS TOTAL

points

1 CHOUCOU : 1160 350  
2 METHOSS1 : 840 097

3 NICOZOMBI : 325 295  
4 LYCO : 145 137  
5 MARLIEBO : 45 590


675 inscrits, dernier inscrit : Chomsky, 1 connecté. Chouchou








- ◆ Profile data & Social Networking: post, forward, like, comment
- ◆ Transcription of the Penn Sociolinguistic Archive, etc.
  - w/ SAD, FA, VE, sound classifiers, as class exercise
- ◆ Documentation of my language
  - translation of disaster sitreps, requests for help (cf. LORELEI)
  - Web Video Accessibility: transcription, translation
  - Massively Parallel Visual Dictionaries
  - Virtual Multimedia Language Atlas (Phonemica, You Say Potato)
- ◆ LID annotation as in Great Language Game (w/ Phonexia LID)
- ◆ Scope of conjunction annotation (Lieberman, Kulick)
- ◆ Global TIMIT
  - w/ voice selfies, w/ TTS models (Anderson), w/ STT models (Schultz)
- ◆ My Story, My best \_\_\_\_ ever!, Oral History of my hero
- ◆ ...



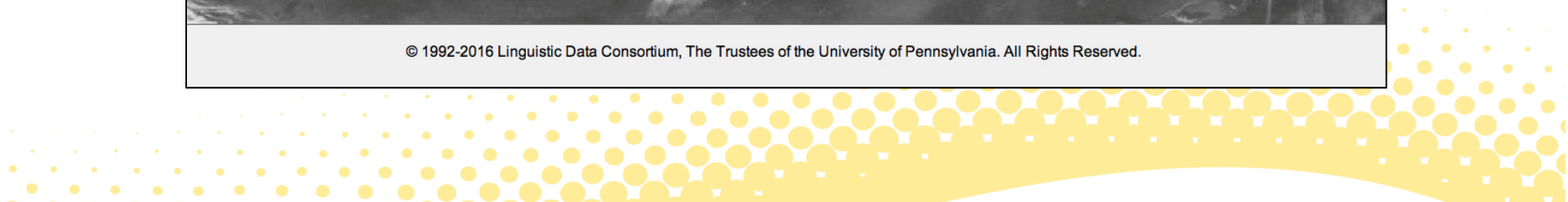
Contact Us About Sign In / [Register](#)






Web-based annotation and transcription projects for language professionals and students



© 1992-2016 Linguistic Data Consortium, The Trustees of the University of Pennsylvania. All Rights Reserved.






Linguistic Data Consortium

[Contact Us](#)
[About](#)
[jfiumara](#)
[Sign Out](#)

## Transcription Task Set-Up

- Assign to Group:

☐ Linguistics 101
☐ Linguistics 101 - Section 2
☐ Sociolinguistics 102

[Create New Group](#)
[Add Students to Group](#)
- Select Activity:

Transcription
- Select the data collection the Group will transcribe or upload your own data

Penn Sociolinguistic Archive  
Feagin Anniston Alabama Archive  
Digital Archive of Southern Speech  
LDC Mixer Corpora  
LDC Callhome English Corpus

Browse
Upload

Are you providing the audio and transcripts?

☐ YES
☐ NO
- How many minutes of speech will Group transcribe for this task?

Note: ~15 minutes of effort for each minute of speech.
+
To evaluate student performance we add 10 minutes to each task.
=

Submit

© 1992-2016 Linguistic Data Consortium, The Trustees of the University of Pennsylvania. All Rights Reserved.

Language ARC

Projects
Contribute
About
Forums

Sign In / Register

**Language Analysis Research Community (ARC)** is an open resource for anyone to participate in language preservation and research.

Learn More
Projects

Click map to find or add your language, or use pulldown menu.

English


[Contact Us](#)
[Privacy Policy](#)
[Terms of Use](#)





[Language ARC](#)
[Projects](#)
[Contribute](#)
[About](#)
[Forums](#)
[Sign In / Register](#)


## PROJECTS


Sort by: 
Search:



Language ARC


Projects



Contribute



About



Forums



Sign In / Register



Multilingual Lexicon Generator

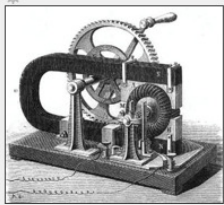

Language ID

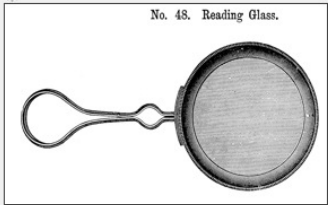

Phrase Detectives



Transcription

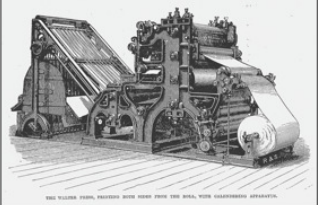

ZombiLingo



Global TIMIT























[Contact Us](#)
[Privacy Policy](#)
[Terms of Use](#)





LANGUAGE MATCH GAME

Drag the tile to match the language spoken in the audio clip.

Spanish  
52

Tagalog  
52

Hausa  
52

Italian  
52

1. \_\_\_\_\_

2. \_\_\_\_\_




3. \_\_\_\_\_


4. \_\_\_\_\_

Submit


LANGUAGE MATCH GAME







Another player identified the language clip below as Albanian.  
Do you agree?

1.  

Albanian

52

Yes

No

I don't know

AMERICA

EUROPE

SOUTH AMERICA

AFRICA

ASIA

AUSTRALIA & OCEANIA



About

Contact Us

Privacy

Terms of Use





## ◆ Community

- no research community devoted to novel incentives & unique workflows for language resource development
- many communities interested in parts of the problem
- you are the avant-garde in those communities
- we want to build the new research community

compute	Input	Processing	Output
<b>LDC</b>	Donate Data	Collaborate	Acquire Data
<b>NIEUW</b>	Contribute Data <ul style="list-style-type: none"> <li>• incentives</li> </ul>	Process <ul style="list-style-type: none"> <li>• workflow</li> </ul>	Use Data <ul style="list-style-type: none"> <li>• evaluation</li> </ul>