



Introducing NIEUW: Novel Incentives and Workflows for Eliciting Linguistic Data

Christopher Cieri, James Fiumara, Mark Liberman, Chris Callison-Burch*, Jonathan Wright

University of Pennsylvania Linguistic Data Consortium and *Department of Computer and Information Science

supported by the National Science Foundation via CISE Research Infrastructure (CRI) grants CRI CI-P 1629923 and CRI CI- NEW 1730377.

- ◆ How many LRs needed to enable today's HLTs for a given language?
 - Krauwer (1998) defined BLARK as a **B**asic **L**anguage **R**esource **K**it
 - Minimal set of resources necessary for research, education, and HLT development: written and spoken corpora, dictionaries, grammars, modules (taggers, parsers, etc), annotation standards & tools, etc
 - would continue to evolve as new technologies emerge
 - Binnenpoorte et al. (2002) implementing BLARK for Dutch-Flemish
 - names 43 technological + data components
 - ≥ 24 data types, still more would be needed to match, e.g. genres
 - Needed: 24 - 43 LRs * ~ 7000 languages $\sim 150,000$ - 300,000 LRs
 - Europe: 288 languages
 - The Americas: 1060 languages
 - The Pacific: 1306 languages
 - Africa: 2143 languages
 - Asia: 2300 languages
- (Ethnologue: <http://www.ethnologue.org>)

- ◆ How many LRs do we have?
 - LDC + ELRA < 2000, LRE Map = 6143
- ◆ And many address the same need
 - e.g. LDC Gigawords: 5 * Arabic, Chinese, English; 3 * French, Spanish
 - ~12 other corpora like English Gigaword
- ◆ And LRs not evenly distributed over languages
 - Cieri et al. (2016) surveyed languages in US & EU Low Resource Language (LRL) programs
 - Bias in LRL selection: National, Provincial > Broader Communication but lacking official status > Vigorous Use with standardization and literatures. None were Threatened, Dying or Extinct
- ◆ However, even languages with many millions of speakers and worldwide economic importance remain under-sourced

- ◆ After 25 years, still building on LRs mostly for the largest languages
 - Still far to go
 - In the meantime, needs go unmet
 - Languages die undocumented.
 - Whole populations continue without benefits HLTs offer: in everyday life, during disasters
- ◆ Our project oriented approach will not allow much beyond this
 - need to justify government investment in the face of competing priorities
- ◆ Is there another, complementary approach?
- ◆ What, besides money, motivates people to contribute to LRs
 - incentives: information, learning, entertainment, self-expression, socializing, working with like minded people, developing skills, demonstrating competence, competition, status, prestige, recognition, local pride, contributing to a greater cause or good, access to (HLT) services
- ◆ But if we offer such incentives, what ‘workforces’ do we attract
 - What are their skills? How can we assess, train, task them to provide data we need?
 - What workflows most effective for these workforces? What post-processing needed?
 - Are there models we can emulate?

- ◆ People contribute massive amounts of time, samples given proper incentives and appropriate workflows

- ◆ 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 the market:

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

Free public domain audiobooks

— Read by volunteers from around the world. —

Read

LibriVox audiobooks are read by volunteers from all over the world. Perhaps you would like to join us?

[VOLUNTEER](#)

Listen

LibriVox audiobooks are free for anyone to listen to, on their computers, iPods or other mobile device, or to burn onto a CD.

[CATALOG](#)

- ◆ LibriVox
 - ◆ “free public domain audiobooks”
 - ◆ recruit, train, organize volunteers who
 - ◆ record readings of literary works out of copyright in US

- ◆ January 1, 2017
 - ◆ 53,430 hours of readings in English
 - ◆ 8,689 hours in 36 other languages
 - ◆ 1 recorded hour requires 2 hours of reading time + 2-4 hours editing > 248,476
- ◆ At industry standard rates, \$500 per finished hour of audio, cost to fund LibriVox would exceed \$31,000,000.
- ◆ Incentives
 - ◆ belief in LibriVox mission
 - ◆ interest in content
 - ◆ promoting specific work, author, genre
 - ◆ challenge of acting without visuals
 - ◆ opportunity to develop and apply skills; spin-offs to Iambik and Audible
 - ◆ work with like-minded collaborators in multiple sub-communities
 - ◆ enlightened self-interest



- ◆ Great Language Game
 - ◆ players hear brief audio clips in ~80 languages
 - ◆ select correct language from multiple choice
 - ◆ 3 lives
 - ◆ correct = 50 points
 - ◆ incorrect = lost life
 - ◆ # choices increases during game
 - ◆ Ethnologue entries for languages missed

- ◆ History
 - ◆ 2013 appeared
 - ◆ 2014 public data release, < 1 year later
 - ◆ >16,000,000 judgments
 - ◆ > all NIST LRE judgments ever
 - ◆ >38,000,000 before stats disappeared
- ◆ Players include: Central & South America, Sub-Saharan Africa, Central Asia, Oceania.
- ◆ Incentives
 - ◆ information
 - ◆ entertainment
 - ◆ competition
 - ◆ status
- ◆ Judgments suboptimal for LRE
 - ◆ too few clips
 - ◆ too many judgements/clip
 - ◆ all answers known

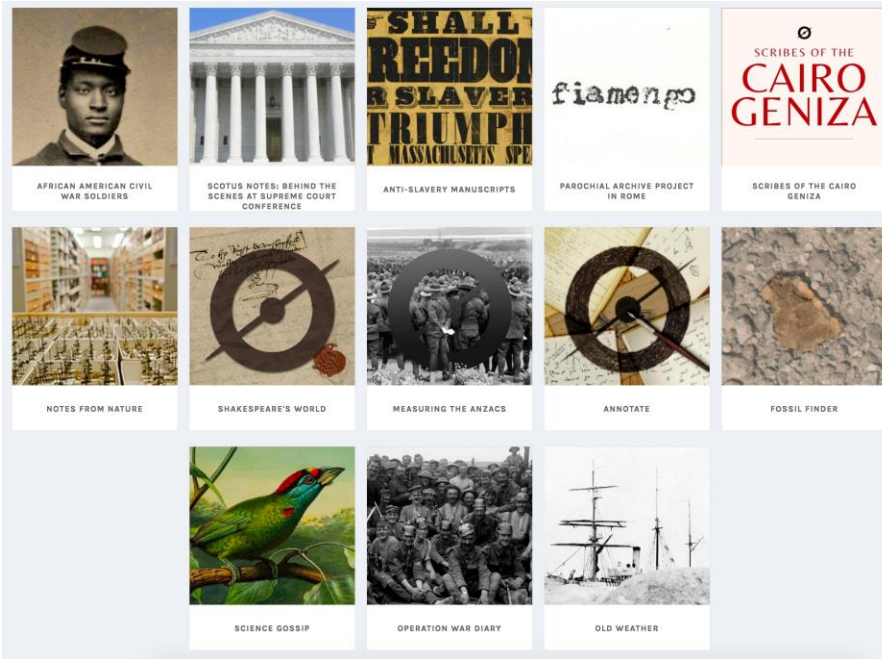
Farewell Listeners!

After nearly 5 years of running the Great Language Game, it's time for me to say farewell and move on to other projects. This means that the game is no longer available to play. If you're interested in the game's history, you can read more about it [here](#).



Thank you to all the players, polyglots and everyday people who have enjoyed the language game and shown support, either with your emails or on Patreon. The languages of the world are a big part of its cultural heritage, it's made me very happy to help so many people enjoy them.

— Lars



- ◆ Zooniverse
 - ◆ citizen science portal
 - ◆ originally for astronomy and biology
 - ◆ expanded to many hard and social sciences

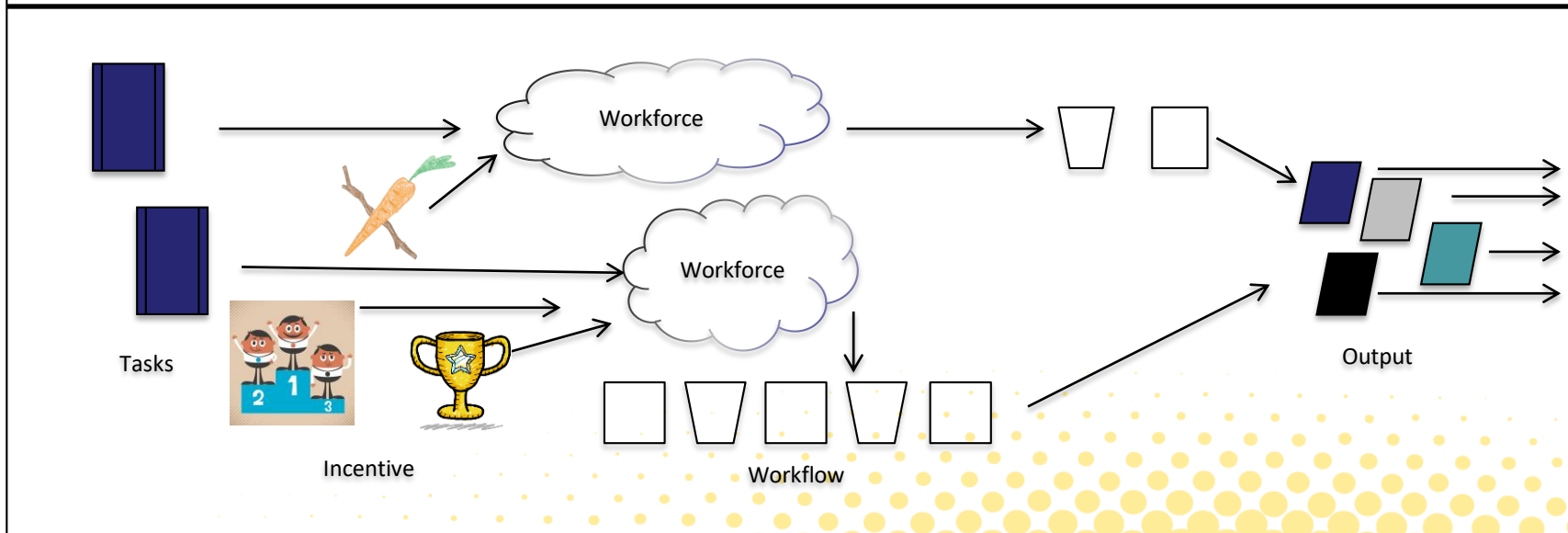
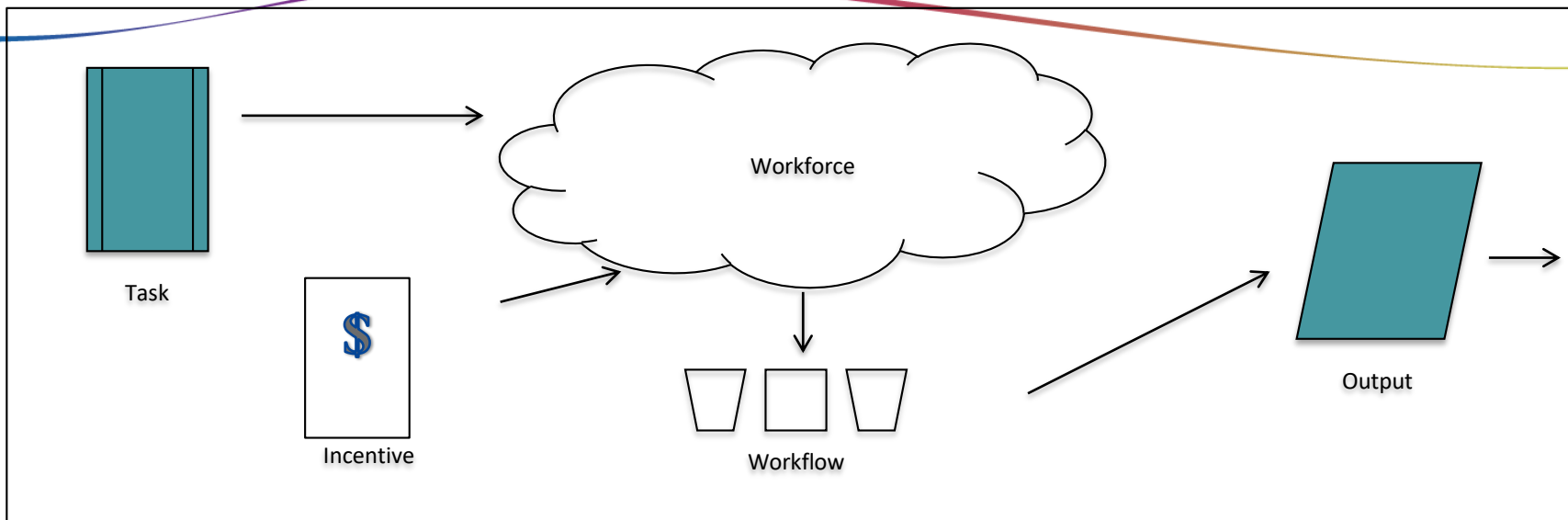
- ◆ Founded in 2009
- ◆ >80 activities, 12 linguistic
- ◆ from Zooniverse and partners
 - ◆ Zooniverse toolkit
 - ◆ some activities use external tools
- ◆ >1,000,000 contributors
- ◆ high throughput
 - ◆ Anzacs: >279,988 facts extracted / 3 months by 7,000 volunteers
- ◆ relatively challenging work
 - ◆ galaxy morphology
 - ◆ animal classification
 - ◆ transcription, entity tagging, KBP
 - ◆ paleography: English, Arabic, Hebrew, ...
- ◆ training through tutorial, reference material, FAQs, discussion fora

The screenshot shows the Zooniverse 'Shakespeare's World' interface. On the left, there is a sidebar with metadata: Author: Unknown, Title: Cookbook [manuscript], Catalogue: Original, and buttons for Snapshot, Transcribe Text (highlighted), Identify Graphic, and Tutorial. A notification states: 'Remember, you're not required to transcribe the whole page! I'm done!'. The main area displays a handwritten manuscript page with a 'Transcribe text' overlay. The overlay includes a text input field, a table of expansion and deletion options, and a 'Save Changes' button.

Expansion	Insertion	Deletion	Unclear	Superscript	Marginalia
w ^h	w ^h , w'	y ^e	y ⁱ	y ^m	yo', y'
Ma ^{te}	Wor ^d	La, La ^p	L, L ^p		
S'	o'	Ex', Exaite	lrm, it	lre	Hc., hono ^{ble}
esq.	es	m ^t	p ^d		
A ^e					

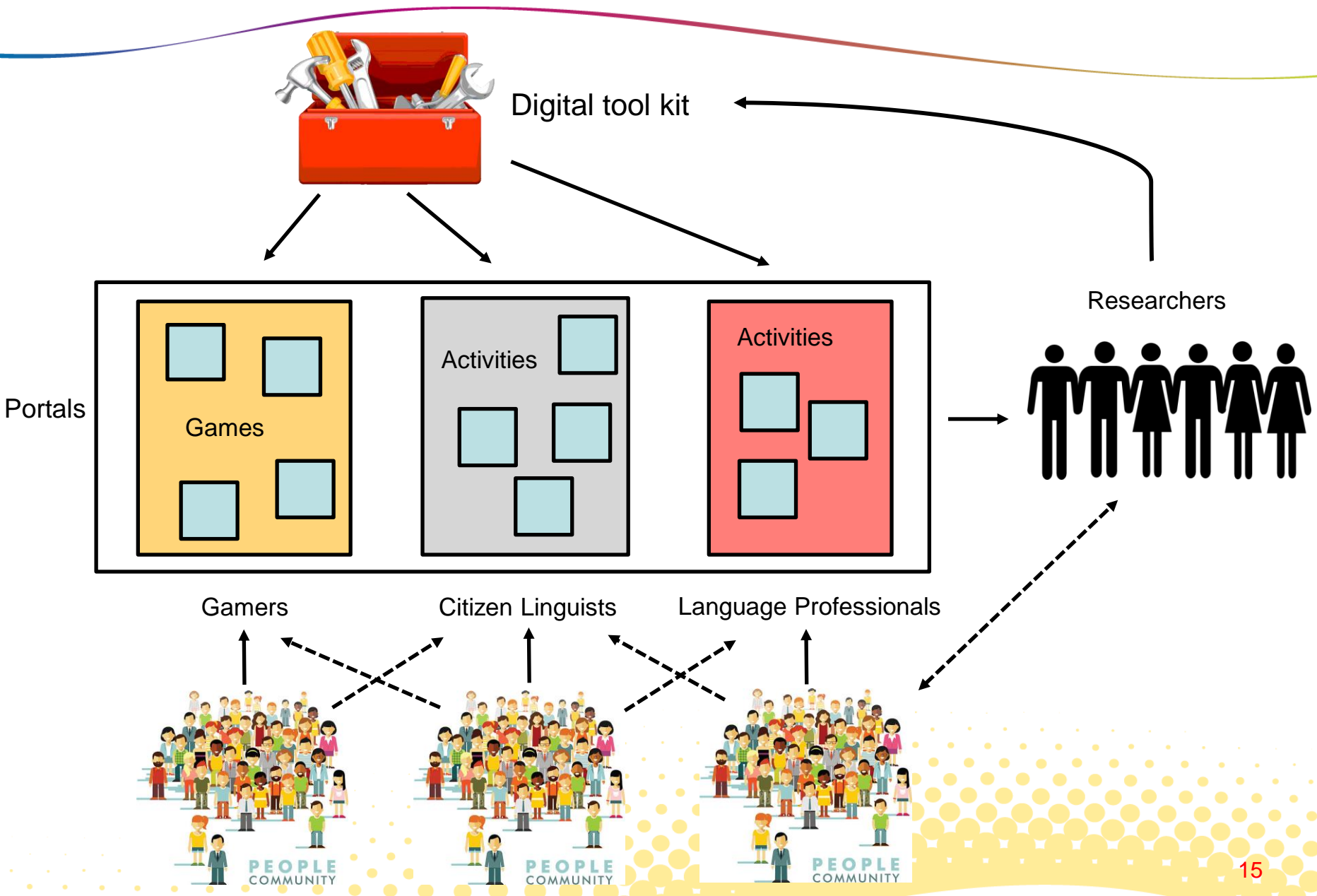
- ◆ Create mechanisms to promote people-power in LR development
- ◆ Offer a range of incentives beyond the 'mere' monetary, organized in rational groupings
- ◆ Support multiple workforces via appropriate workflows.
- ◆ Provide resulting LRs to researchers
- ◆ Where possible & appropriate offer access to resulting HLTs
- ◆ Help people help us, help one another, help themselves.

Incentives Aware LR Development



- ◆ minimizes barriers to entry
 - ◆ always available to anybody, on any device
 - ◆ though some activities require additional skills, data
- ◆ connection to social networking sites for recruitment
- ◆ offers multiple incentives
- ◆ automates worker training and assessment
- ◆ data prioritization replaces data selection
- ◆ X-sourcing: tasking matches workforce, incentives
- ◆ algorithms for task assignment, finding experts, modeling variation
- ◆ integration with current LDC operations to be sustainable

- ◆ NIEUW (Novel Incentives and Workflows) is an NSF funded project to create infrastructure to support linguistic data elicitation
 - NSF CRI-CI-NEW: focused on Community Infrastructure
 1. Toolkit for building collection/annotation exercises
 2. Examples activities as proof of concept
 3. Portals to host groups of related activities
 1. Language Games
 2. Citizen Linguists
 3. Language Professional
 4. User Community



The screenshot shows the 'World Language Games' portal. At the top, the letters 'LINGO B O I N G O' are arranged in a decorative arc. Below this, there are navigation tabs for 'All', 'English', and 'French'. The main content area displays four game cards:

- Jeux de mots** (French): Lexical and semantic games with a purpose in French.
- Phrase Detectives** (English): Compete against other detectives by identifying the relationships between words and phrases in a variety of texts including literature, history, travel, entertainment and science. Earn top scores!
- Tile Attack** (English): Go head-to-head against another player competing to identify the noun phrases of a text.
- Zombilingo** (French): Identify syntactical dependencies, collect brains and eat them! This language game is fun for both fans of grammar and zombies.

At the bottom of the page, it says 'Portions © Copyright 2017 University of Pennsylvania.'

- ◆ Portal - links
- ◆ Authentication
- ◆ Back-end
- ◆ Publicity
 - analysis

L I N G O B I N G O

Name That Language!

Can you identify the language?

Score:	10
Round:	2
Lives:	3

▶ ◻

Russian Portuguese Polish

Play Again

- ◆ Game Play: using bonus round to collect novel annotations
 - # moves, levels, 'lives'
 - reward/penalty for correct/incorrect – amount, fixed/variable, normal/bonus round
 - frequency of bonus rounds
 - skip, or answer before clip complete
 - # languages, # distractors, closed set
- ◆ Data
 - type: broadcast, conversation telephone speech, etc
 - language of clips relative to player's
 - location
 - native language
 - performance
 - language of distractors relative to (all of the above plus):
 - distance from target language
 - phonemically
 - in terms of family tree

Language Professionals

Activity: Group Setup ccieri [Sign Out Profile](#)

Progress: 60%

Welcome to Group Setup.

- Identify your group using up to three levels of labels (e.g. university, course, section)

Level 1 Penn	Level 2 Ling001	Level 3 Section A
-----------------	--------------------	----------------------
- Select the Activity you want the group to undertake and Language


Transcription	English
---------------	---------
- Select the collection the group will **Transcribe** or upload your own.

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


Are you sharing the audio and transcripts?
 Yes No
- How many minutes of speech will each participant **Transcribe**?

Your minutes (NB ~15 minutes effort for each minute speech) + To evaluate student performance we add 10 minutes to each + Based on your collection & sharing choices, add these to benefit community =
- Select processing to perform automatically for each complete transcript (dependent on Language)

Forced Alignment
 Vowel Formant Extraction
 (l) vocalization
 (r) apicalization
 (s) palatalization
- Paste in the email addresses of the participants in your group.
- Click to approve. We will contact the participants. You can track progress [here](#).


Language Professionals


Activity: Transcription
Group: Penn, Ling001, Section A
Collection: Penn Sociolinguistic Archive
File: Rose B.
Progress - Yours: 60%, **Class:** 64%




ccieri [Sign Out Profile](#)


Transcribe each audio clip below using only standard orthography.




Well, during the War, I used to go down to the USO.



They had dances for the soldiers.



And everything was very nice.



They had food and soft drinks and all the guys were polite.

Language ARC

ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER

Language Analysis Research Community

A CITIZEN SCIENCE COMMUNITY FOR RESEARCH IN LANGUAGE, LINGUISTICS, AND MACHINE LEARNING

[Click Here for Projects](#)

Travel and communication makes the world feel smaller. However, our inability to understand each other, even to speak the same language, hinders our efforts to share scientific and medical knowledge, to govern and conduct diplomacy.

Human Language Technologies can help but they rely upon language resources: dictionaries, grammars, texts and audio recordings that are absent or scarce for most of the world's ~7000 languages. Language ARC is working to fix that problem.

You can help. Nearly all humans on the planet speaks, reads, understand or writes at least one language, qualifying them to contribute to this effort. In Language ARC there are activities from the very simple to the challenging that contribute to language resources that are then shared for research, education and technology development purposes.

The activities are designed so that you can contribute something of value in just a few minutes, while on the bus or during lunch break or whenever.

The more you contribute the more we understand about language, they better our research and technologies become, the closer we are to understanding each other.

Take part in cutting edge research in linguistics and language technology

Language ARC


ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER

PROJECTS

SORT BY


SEARCH

Train Your Robot




Help train a system to recognize the languages of the world

Perfecting the BNC




Help us correct time-aligned spoken language transcripts from the British National Corpus.

Voices Across the Globe



Upload your voice to help promote research in language recognition technologies.





Labov Archive



Help transcribe William Labov's sociolinguistic archive of regional American English.


Portions © 2017 - University of Pennsylvania All Rights Reserved
Funded in part by a grant from the National Science Foundation.

About Us Privacy Policy
FAQ Contact

Language ARC

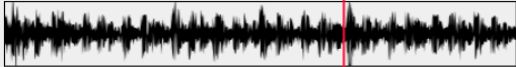
ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER

 Train Your Robot



About Project Guide Tutorial Discussion Board

Select language:

Is the following audio entirely in Spanish?



9.57

YES NO





Don't know Bad audio

NEXT

[Click Here for Training Results!](#)

Portions © 2017 - University of Pennsylvania All Rights Reserved
Funded in part by a grant from the National Science Foundation.

About US Privacy Policy
FAQ Contact

Language ARC

[ABOUT](#)
[DO RESEARCH](#)
[CREATE A PROJECT](#)
[NEWS](#)
[TALK](#)
[SIGN IN / REGISTER](#)

Train Your Robot

[About](#)
[Project Guide](#)
[Tutorial](#)
[Discussion Board](#)

	Your Score	Average Score	Top Score	
March 9, 2018				Select language: Español
March 2, 2018				
February 23, 2018				
February 16, 2018				

[Click Here for Older Results](#)
What Do These Scores Mean?
Continue

Portions © 2017 - University of Pennsylvania All Rights Reserved
Funded in part by a grant from the National Science Foundation.

[About US](#)
[FAQ](#)

[Privacy Policy](#)
[Contact](#)

The screenshot shows a web page for 'Language ARC' with a navigation menu including 'ABOUT', 'DO RESEARCH', 'CREATE A PROJECT', 'NEWS', 'TALK', and 'SIGN IN / REGISTER'. The main content area features the BNC logo (British National Corpus) and the heading 'Perfecting the British National Corpus'. Below this, it says 'Help correct speech to text alignment in the British National Corpus' and provides two buttons: 'Project Guide' and 'Start Now'. A secondary navigation bar includes 'About the British National Corpus', 'Our Research Team', and 'Discussion Board'. The main text under 'About the British National Corpus' explains that the BNC is the first of national corpus efforts to document spoken language, consisting of a 100 million word collection of written and spoken language from the later part of the 20th century. It details the composition of the corpus, including regional and national newspapers, specialist periodicals, academic books, popular fiction, unpublished letters, memoranda, school and university essays, and spoken language collected in various contexts from formal business meetings to radio shows and phone-ins. It also notes that work on building the corpus began in 1991 and was completed in 1994, with no new texts added since then.

Language ARC


ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER

B N C
BRITISH NATIONAL CORPUS

Perfecting the British National Corpus

About Project Guide Tutorial Discussion Board


Do the speech and transcript match perfectly?

 00:00 00:00

[Get help with this task.](#)

Portions © 2017 - University of Pennsylvania All Rights Reserved
Funded in part by a grant from the National Science Foundation.


About US Privacy Policy
FAQ Contact



Language ARC ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER

B N C BRITISH NATIONAL CORPUS Perfecting the British National Corpus About Project Guide Tutorial Discussion Board


Is the transcription of the audio incorrect? (ex., wrong word, added word, missing word)



Yeah. Oh he's all shy this morning.

[Get help with this task.](#)

Portions © 2017 - University of Pennsylvania All Rights Reserved About US Privacy Policy
Funded in part by a grant from the National Science Foundation. FAQ Contact



Language ARC


ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER

B N C
BRITISH NATIONAL CORPUS

Perfecting the British National Corpus

About Project Guide Tutorial Discussion Board

Indicate below how the audio alignment is incorrect.



Additional audio before first transcribed word

First transcribed word cut off

Additional audio after last transcribed word

Last transcribed word cut off


Yeah. Oh he's all shy this morning.

[Get help with this task.](#)

Continue


Portions © 2017 - University of Pennsylvania All Rights Reserved
Funded in part by a grant from the National Science Foundation.

About US Privacy Policy
FAQ Contact



Language ARC

[ABOUT](#)
[DO RESEARCH](#)
[CREATE A PROJECT](#)
[NEWS](#)
[TALK](#)
[SIGN IN / REGISTER](#)








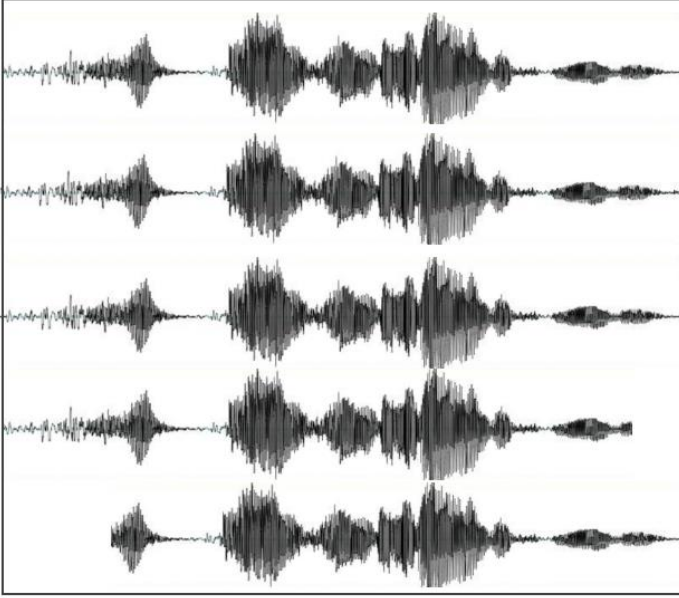
BRITISH NATIONAL CORPUS

Perfecting the British National Corpus

[About](#)
[Project Guide](#)
[Tutorial](#)
[Discussion Board](#)

Alignment Tutorial


00:00 00:00

00:00 00:00

00:00 00:00

00:00 00:00

00:00 00:00



Here the alignment is correct.

Extra audio after the last transcribed word

Extra audio before the first transcribed word

Last transcribed word cut off

First transcribed word cut off

Yeah. Oh he's all shy this morning.


[Back](#)

Portions © 2017 - University of Pennsylvania. All Rights Reserved.

[About US](#)
[Privacy Policy](#)

Language ARC

[ABOUT](#)
[DO RESEARCH](#)
[CREATE A PROJECT](#)
[NEWS](#)
[TALK](#)
[SIGN IN / REGISTER](#)


Perfecting the British National Corpus

[About](#)
[Project Guide](#)
[Tutorial](#)
[Discussion Board](#)

Perfecting the British National Corpus Discussion Board

SEARCH

Announcements
Updates, news, and info.

Participants
Threads
Comments

Project Discussion.
Talk about project with other participants.

Participants
Threads
Comments

Questions for the Researchers
Ask the research team questions.

Participants
Threads
Comments

Technical Issues
Report any technical issues here.

Participants
Threads
Comments

[Recent Comments](#)

Active Commenters:
[jfumara](#), [ccieri](#), [nchomsky](#)

Popular Tags:
[missing_audio](#)
[alignment](#)
[regional_dialect](#)
[wave_form](#)
[british_heritage](#)

Introducing NIEUW, LDC Workshop, Mexico City, May 24-25

30

Language ARC
ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER

Voices Across the Globe
About Project Guide Tutorial Discussion Board Your Profile

Click the Map or use Search

Search

Portions © 2017 - University of Pennsylvania All Rights Reserved

Funded in part by a grant from the National Science Foundation.

About US

FAQ

Privacy Policy


Contact

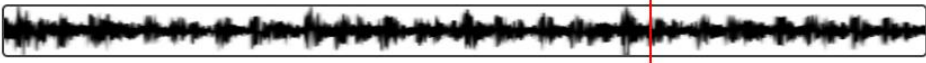
Language ARC ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER





Voices Across the Globe About Project Guide Tutorial Discussion Board Your Profile

[Click here for Map](#) Search

Record Your Voice, Record Your Story

 TIMIT Sentences







   

[Get Help with this Task](#)

Your Other Recordings:
[My Hometown 11/01/16](#)


Portions © 2017 - University of Pennsylvania All Rights Reserved
Funded in part by a grant from the National Science Foundation.

About US Privacy Policy
FAQ Contact

Language ARC

ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER



Multilingual Lexical Elicitation Task

Help build multilingual lexicons by naming objects and entities in your language

[Project Guide](#) [Start Now](#)

About Multilingual Lexical Elicitation Task

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.


Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem.

Portions © 2017 - University of Pennsylvania All Rights Reserved
Funded in part by a grant from the National Science Foundation.


About US
FAQ

Privacy Policy
Contact



Language ARC

ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER

 Multilingual Lexical Elicitation Task

About Project Guide Tutorial Discussion Board

Select language:

Select Subject Type:

ANIMALS


FRUITS & VEGETABLES

Machines and Tools


Portions © 2017 - University of Pennsylvania All Rights Reserved
Funded in part by a grant from the National Science Foundation.

About US
FAQ

Privacy Policy
Contact




Language ARC
ABOUT DO RESEARCH CREATE A PROJECT NEWS TALK SIGN IN / REGISTER



Multilingual Lexical Elicitation Task

About Project Guide Tutorial Discussion Board









Don't forget to sign in or create an account!

Sort by Category:

What do you call this animal?

Scientific name:
 Order: Carnivora
 Family: Canidae
 Genus: Vulpes
 Species: vulpes
 Binomial: Vulpes vulpes

Common name:
 English: Red fox 





Submit

Portions © 2017 - University of Pennsylvania All Rights Reserved

Funded in part by a grant from the National Science Foundation.

About US Privacy Policy

FAQ Contact

- ◆ NSF CRI Planning: June 2016 – April 2018
- ◆ NSF CRI NEW: July 2017 – June 2019 (2020)
- ◆ presentations, discussions
 - NSF, DARPA, IARPA, DOD, CASL, AFRL, ARL
 - MIT-LL, CLARIN, SADiLaR, Carnegie Mellon U.
 - LREC, NSF workshops on Novel Incentives and Workflows, LTC
 - program committee, etc. for LREC, Games4NLP, HCOMP
- ◆ Development Partners
 - University Essex, QMUL, Sorbonne, LORIA, enetCollect
- ◆ Advisory Board
 - Nicoletta Calzolari, Khalid Choukri, Paul Dingman, Maxine Eskenazi, Edith Law, Walter Lasecki, (Matt Lease), Verena Lyding, Lionel Nicolas, Kevin Werbach,
- ◆ New Data Contributors
 - Labov, Drager, D'Arcy, Feigin, Yaeger-Dror
- ◆ NIEUW Infrastructure
 - LDC WebAnn ported to run outside LDC on VMs (currently heroku)
 - LingoBoingo portal live, buzz building
 - LID Game v.1 complete, demoed at Philadelphia Science Festival