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 **Basic Language Resource Kit**
  - 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 ~ 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?
LibriVox

- 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

LibriVox

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

Introducing NIEUW, LDC Workshop, Mexico City, May 24-25
Great Language Game

- 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

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

Introducing NIEUW, LDC Workshop, Mexico City, May 24-25
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
Citizen Science: Zooniverse

- 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

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

- 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.
Introducing NIEUW, LDC Workshop, Mexico City, May 24-25
◆ 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
Games

- Portal - links
- Authentication
- Back-end
- Publicity
  - analysis
Name That Language!

Can you identify the language?

Score: 10
Round: 2
Lives: 3
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

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

---

### Activity: Group Setup

**Progress: 60%**

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penn</td>
<td>Ling001</td>
<td>Section A</td>
</tr>
</tbody>
</table>

#### Welcome to Group Setup

1. 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

2. Select the Activity you want the group to undertake and Language
   - Transcription
   - English

3. 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

4. How many minutes of speech will each participant **Transcribe**?
   - Your minutes (NB ~15 minutes effort for each minute speech) **60**
   - To evaluate student performance we add 10 minutes to each **10**
   - Based on your collection & sharing choices, add these to benefit community **0**
   - **Total**: **70**

5. Select processing to perform automatically for each complete transcript (dependent on Language)
   - [ ] Forced Alignment
   - [ ] Vowel Formant Extraction
   - [ ] (l) vocalization
   - [ ] (g) apicalization
   - [ ] (s) palatalization

6. Paste in the email addresses of the participants in your group.

7. Click to approve. We will contact the participants. You can track progress [here](#). Submit
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.
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

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

Click Here for Projects
Citizen Linguists

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

Language ARC

PROJECTS

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
Train Your Robot

Is the following audio entirely in Spanish?

Select language: Español

YES  NO
Don't know  Bad audio

Click Here for Training Results!
### Train Your Robot

<table>
<thead>
<tr>
<th>Date</th>
<th>Your Score</th>
<th>Average Score</th>
<th>Top Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 9, 2018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 2, 2018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February 23, 2018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February 16, 2018</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Select language: **Español**

Click Here for Older Results

---

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

Help correct speech to text alignment in the British National Corpus

Project Guide  Start Now

About the British National Corpus  Our Research Team  Discussion Board

What is the BNC?

The British National Corpus (BNC) is the first of the national corpus efforts, to document how the language was not only written but spoken at a particular place and time, something we wish we had for earlier forms of English.

The BNC is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. The latest edition is the BNC XML Edition, released in 2007.

The written part of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10%) consists of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

Work on building the corpus began in 1991, and was completed in 1994. No new texts have been added after the completion of the BNC. However, new material continues to be added to the BNC XML Edition.
Perfecting the British National Corpus

Do the speech and transcript match perfectly?

Yeah. Oh he's all shy this morning.

Get help with this task.
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.
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.

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

Perfecting the British National Corpus

Alignment Tutorial

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.
Perfecting the British National Corpus Discussion Board

**Announcements**
Updates, news, and info.

**Project Discussion**
Talk about project with other participants.

**Questions for the Researchers**
Ask the research team questions.

**Technical Issues**
Report any technical issues here.

Recent Comments

Active Commenters:
jfiurnara, ccieri, ochomsky

Popular Tags:
missing, audio, alignment, regional, dialect, wave_form, british_heritage
Introducing NIEUW, LDC Workshop, Mexico City, May 24-25
Citizen Linguists

Record Your Voice, Record Your Story

Click here for Map

Your Other Recordings:
My Hometown 11/01/16

Get Help with this Task
Multilingual Lexical Elicitation Task

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

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.
Multilingual Lexical Elicitation Task

Select language: Español

Select Subject Type:

ANIMALS

FRUITS & VEGETABLES

Machines and Tools
Work to Date

- NSF CRI Planning: June 2016 – April 2018
- 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