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, processors. Each part is at least 2 of these, maybe all three.
LDC in 1 Slide

- 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
Impediments

- 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?
Features of a Solution

- 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\rightarrow 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
<table>
<thead>
<tr>
<th>Project</th>
<th>Task</th>
<th>Action</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admin</td>
<td>namespaces</td>
<td>No Available Kits</td>
<td>N/A</td>
</tr>
<tr>
<td>Audio</td>
<td>web trans demo</td>
<td>start</td>
<td>N/A</td>
</tr>
<tr>
<td>Audio</td>
<td>services</td>
<td>start</td>
<td>N/A</td>
</tr>
<tr>
<td>DEFT</td>
<td>ERE demo</td>
<td>continue</td>
<td>in progress (paused)</td>
</tr>
<tr>
<td>HighEntropyASR</td>
<td>word list triage</td>
<td>No Available Kits</td>
<td>not in progress</td>
</tr>
<tr>
<td>HighEntropyASR</td>
<td>word list audit</td>
<td>continue</td>
<td>in progress</td>
</tr>
<tr>
<td>HighEntropyASR</td>
<td>heasr reports</td>
<td>start</td>
<td>N/A</td>
</tr>
<tr>
<td>VAST</td>
<td>web_trans_test</td>
<td>continue</td>
<td>in progress (paused)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Word</th>
<th>Possible</th>
<th>Impossible</th>
</tr>
</thead>
<tbody>
<tr>
<td>nipper</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trashy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spite</td>
<td></td>
<td></td>
</tr>
<tr>
<td>circumlocutions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>outages</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Citizen Science Portal

Curio

Thoreau's Field Notes
How does climate change affect the timing of when buds, flowers and fruits come out?
Learn More

Dive4Oceanography
What can we learn from documenting and studying oceanic diving records?
Learn More

Typewriter Girl
Transcribing Victorian literature can be both fun and exciting!
Learn More

CrowdEEG
Help scientists understand the inner workings of sleep-staging through EEG classification.
Learn More

Ensemble
Can the untrained ear accurately transcribe music?
Learn More

UrbanEars
Sift through the Sounds of Cities
Learn More
Amongst the thousands of languages spoken across the world, here are just eighty. How many can you distinguish between?
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.

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.

549 docs completed
The most recent was Alice in Wonderland (Lewis Carroll) completed by Julie3164 on 20 Mar 2016

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.

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:

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

The word "orange" refers to "Sherlock Holmes".

Feedback
ZOmBi O Lingo
ATTRAPEZ LES TOUTES

Joue pour aider les scientifiques!

Suis mes règles, identifie les têtes et mange-les.

Attention aux pièges, ils sont nombreux!

Prof. Frankenstein

Suis-je limité pour toi?
Tu accèdes à toutes les options, bonus cachés!

Essayer

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

675 inscrits, dernier inscrit : Chomsky, 1 connecté, Chouchou
Sample Activities

- 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 (Liberman, Kulick)
- Global TIMIT
  - w/ voice selfies, w/ TTS models (Anderson), w/ STT models (Schultz)
- My Story, My best ___ ever!, Oral History of my hero
- ...

Web-based annotation and transcription projects for language professionals and students
Language Analysis Research Community (ARC) is an open resource for anyone to participate in language preservation and research.

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

English
Examples

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

1. [Audio Icon] [Blank]
2. [Audio Icon] [Blank]
3. [Audio Icon] [Blank]
4. [Audio Icon] [Blank]

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

1. [Sound Button] [Mute Button] Albanian

[Yes] [No] [I don't know]
Community

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

<table>
<thead>
<tr>
<th>compute</th>
<th>Input</th>
<th>Processing</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDC</td>
<td>Donate Data</td>
<td>Collaborate</td>
<td>Acquire Data</td>
</tr>
<tr>
<td>NIEUW</td>
<td>Contribute Data • incentives</td>
<td>Process • workflow</td>
<td>Use Data • evaluation</td>
</tr>
</tbody>
</table>