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
{ccieri, jfiumara, myl, jdwright} @ldc.upenn.edu, {ccb} @cis.upenn.edu

Abstract

This paper introduces the NIEUW (Novel Incentives and Workflows) project funded by the United States National Science Foundation and part of the Linguistic Data Consortium's strategy to provide order of magnitude improvement in the scale, cost, variety, linguistic diversity and quality of Language Resources available for education, research and technology development. Notwithstanding decades of effort and progress in collecting and distributing Language Resources, it remains the case that demand still far exceeds supply for all of the approximately 7000 languages in the world, even the most well documented languages with global economic and political influence. The absence of Language Resources, regardless of the language, stifles teaching and technology building, inhibiting the creation of language enabled applications and, as a result, commerce and communication. Project oriented approaches which focus intensive funding and effort on problems of limited scope over short durations can only address part of the problem. The HLT community instead requires approaches that do not rely upon highly constrained resources such as project funding and can be sustained across many languages and many years. In this paper, we describe a new initiative to harness the power of alternative incentives to elicit linguistic data and annotation. We also describe changes to the workflows necessary to collect data from workforces attracted by these incentives.

Keywords: novel incentives, workflows, language resources

1. Introduction & Motivation

The Human Language Technology (HLT) community has benefitted from massive contributions of linguistic data from data centers, governments and groups around the world. Nevertheless, potential still remains largely untapped because the LRs that fuel development fall far short of need whether measured by volume, data type, or language coverage. Data centers regularly receive requests for data sets that they cannot supply even for the dozen languages with the greatest populations and *gross linguistic product*. A 2010 survey (METANET) of the Language Resources (LRs) required to build the HLTs that would support speakers of a given language working in the information age found that none of the language of the European Union, not even English, were fully supplied and warned that "*21 out of 30 European languages could become extinct in the digital world*". Beyond a few exceptional cases: Mandarin, Modern Standard Arabic, Japanese, language spoken primarily outside the EU, suffer even greater LR deficits.

Although the absence of language resources hinders the development of technologies to support international commerce, the problem becomes truly acute during emergent situations such as natural disasters and refugee crises. The International Association of Conference Interpreters warned in 2008: "*Ending a conflict and delivering emergency and humanitarian aid across language barriers represents a major challenge, for which few of the organisations entrusted with operations in the field are well equipped. This problem is compounded by the fact that there is a chronic shortage of interpreters in zones of crisis and war willing to work in the line of fire or in areas of natural disaster.*"

HLTs can make a critical contribution towards disaster relief as we see in the work of Verma et al. (2011) who were able to identify tweets that provided situational awareness with 80% accuracy using purpose built HLTs. However, their system required LRs not currently available for most of the world languages.

Some US government programs have begun to address the problem. DARPA LORELEI is developing technologies to provide situational awareness based on disaster related communications in low resource languages. Despite DARPA's track record for managing such common task projects to produce effective technologies, LORELEI's impressive array of resources will be available for at most a few dozen of the worlds 7000 languages.

Today's approaches to LR creation to support HLT development cannot hope to address world need. This is not only due to the total effort required to create these resources, but also due to the reliance on finite project-oriented funding and collection. Language resource developers and the HLT community must augment these efforts by rethinking the way we collect and annotate linguistic data, the incentives that we offer, the workforces who react to such incentives, the workflows that maximize the efficiency of such workforces and the downstream processing necessary to make the best use of the data and judgments that result from such new approaches.

2. An Incentives-Aware Model of Language Resource Creation

We envision any process used to collect linguistic data as comprised of the interaction of several components: task, incentives, workforce, workflow and processing. Each task has an inherent difficulty which may be mitigated by careful interface design. Task difficulty determines the level of education, skill set and commitment required of the humans who provide the data required to accomplish the task. At the same time the incentives offered attract different contributors who require custom workflows and interfaces in order to maximize their productivity. The tuple of task incentive, workforce and workflow produces an output that may again require customized processing before it can be exploited by a human language technology. Greenfield, Chan and Campbell (2016) make this clear when they write: "*While annotators who have been trained as professional linguists are able to annotate accurately and consistently from dense annotation guidelines, the amateur annotators who serve as workers on crowdsourcing platforms are not similarly motivated to create the best annotations possible*."

Additionally, data created specifically to support HLT research and development have mostly employed the single incentive of monetary compensation. However, this approach is ineffective when there is a lack of funding or when potential data contributors are motivated by factors other than financial gain. Within the NIEUW project, we will consciously engineer incentives, workforces and workflows to produce output for specific purposes.

3. Incentives in Language Resource Development for HLT

Although the use of alternative incentives is not unknown, it is still relatively uncommon within HLT though some examples follow. Campbell (2016) describes multiple data collections that support research into the production of expressive speech. These efforts augmented monetary compensation with sustenance, curiosity, fun, access to recording equipment and data and unusual social opportunities such as human-robot interactions. The resulting data varied in many ways including regionalisms and emotive speech. In Mitsuzawa et al. (2016) consumers provide reviews via the Fuman Kaitori Center initially to communicate some dissatisfaction with a product or service but also to receive spendable credit based upon the size, quality and compliance of the review. The mixed incentives result in variation in the data such as duplications, marginally useful posts, varied spelling of named entities and inaccurate metadata that, the authors believed, must subsequently be corrected by expert annotators.

Crowd workers may be motivated by the quality of the interface design and the desire to maintain a high approval rating in a reputation market as well as any monetary incentives. By improving interface design, Greenfield, Chan and Campbell (2016) elicited higher quality data without having to offer highly compensated work which can tend to attract a mercenary element in some crowd working communities. Phrase Detectives (Poesio et al. 2016) offers the incentives of entertainment, interesting readings, a variable point system, experience levels, leaderboards, socializing and a lottery. The Great Language Game has elicited millions of language labels via the incentives of information, entertainment, competition and status.

Tyson et al. (2016) show that the difference between the corporate mission of About.com and the motivations of content creators leads to an end product that must be post-processed in order to add the proportion of cross document links optimal to sustain the site. The DialRC Center at CMU offers the novel incentives of access to information and the promise of an improved customer experience in real world transit interactions (Eskenazi et al., 2016).

4. Novel Incentives outside HLT

Although the HLT communities stands to benefit significantly from the use of novel incentives, one finds they are employed more frequently and effectively outside the community.

LibriVox[1] organizes volunteers who created audiobooks from out-of-copyright literary works and place them explicitly in the public domain. The initiative has created well more than 10,000 audiobooks recording more than 50,000 hours of speech in the process. Although most books are in English, at least 31 other languages are represented and the non-English material continues to grow. LibriVox recordings could support multiple HLTs including language and speaker recognition, speech-to-text and text-to-speech. Prahallad, Toth and Black (2007) developed text-to-speech systems from a dozen hours of LibriVox audio, compared those systems with more traditional data sources and concluded that "*a voice could be successfully built from large multi-paragraph speech using automatic segmentation tools*."

LibriVox volunteers make enormous contributions because they believe in the LibriVox mission, enjoy reading aloud, want to help maintain the art of storytelling, and enjoy collaborating. A small number eventually receive paid work through audiobook companies.

Other research disciplines collect data and judgments from a crowd attracted by non-monetary incentives. The citizen science site Zooniverse[2] has recruited more than one millions volunteers who contribute to more than 80 different projects by completing tasks such as identifying movement in star fields, classifying animal species, and transcribing museum records. Zooniverse's well-designed interfaces and highly tuned tasking are incentives for participation along with the motivations of learning and discovery, contributing to scientific advancement, interacting with a community who share these goals, and entertainment.

5. NIEUW Directions in Language Research Development

In order to help address the data needs mentioned above, LDC's NIEUW project is building tools to dramatically increase the store of LRs by employing techniques proven to work in multiple scientific disciplines and industry. NIEUW is supported by a 3-year Research Infrastructure grant from the U.S. National Science Foundation

Social media, games with a purpose, and citizen science have shown that human resources are effectively limitless for some activities. By creating an infrastructure that enables the ongoing construction of scalable data collection and annotation activities available to the public via the web and mobile devices and designed with appropriate incentive models, we will enhance LR development well beyond what project-dependent, direct funding alone can accomplish.

We recognize that in order to make the best use of non-traditional labor, we must offer a variety of incentives that are packaged into coherent clusters to appeal to large classes of potential contributors. It appears that several such overlapping communities already exist: 1) language students and professionals such as linguists, transcriptionists and translators who work directly with language data but would benefit from improved tools; 2) citizen scientists who are motivated to contribute to and participate in linguistic research and technology development; 3) game players who seek entertainment and competition. NIEUW is creating web-based portals for each of these communities populated with language collection and annotation activities that appeal to the respective communities through alternative incentives and task design strategies. The size and complexity of the tasks and activities will also be matched to the contributor community. Activities will initially be created by LDC and later by collaborators using a toolkit developed for the

[1] www.librovox.org

[2] www.zooniverse.org

project. By allowing researchers to develop their own activities, the infrastructure not only serves the larger research community but also creates a sustainable and growing resource for data creation not limited to any particular project goal. Unlike collection and annotation strategies created for specific projects and then allowed to lie fallow once the project's needs are met or the funding has been depleted, the NIEUW infrastructure is always available with little impediment to participation.

5.1 Language Professionals

NIEUW will create a portal dedicated to the specific needs, skills, and motivations of language professionals and students. In addition to having interest and expertise in language, professional linguists, language teachers and students can contribute data and judgments of great value to HLT research. Increasingly, language professionals and their students transcribe their own audio data in order to exploit big data approaches that leverage, for example, speech activity detection, forced alignment, and automatic vowel formant extraction. However, uneven availability of supporting infrastructure, data and tools hinders efficient use of time that could be dedicated to actual annotation and learning.

We will address this problem by adapting LDC web-based transcription tools for use within the Language Professionals portal. Teachers and other language professionals will be able to upload audio or select from LDC holdings including the Penn Sociolinguistic Archive comprised of ~6000 recordings made by William Labov and his students over the past 50 years across worldwide English speaking communities. Performance will be evaluated upon small amounts of audio for which gold standard transcriptions exist. The portal will track progress and accuracy which teachers can use to evaluate students. Where appropriate, speech activity detection, forced alignment, and phonetic classifiers will simplify the task and enhance the output which will be made available to researchers in standard formats.

Other planned activities include creating templates for deploying surveys of linguistic typology and eliciting related translations, including re-implementing the Afranaph surveys on the typology of African languages. Questions may include words or sentences requiring translation or questions whose answers are a combination of controlled vocabulary and example sentences.

5.2 Citizen Linguists

Public contributions to scientific research have a long history. Prominent early examples include Edmund Halley soliciting the public to help map solar eclipses in the eighteenth century (Pasachoff, 1999) and the Audubon Society's annual Christmas Bird Count which began in 1900 (Root, 1988). New digital technologies such as the internet, social media, and smart phones have increased the public's ability to engage in scientific research.

Our plans for a Citizen Linguist portal learn from the success of Zooniverse and similar efforts. Although many projects on Zooniverse involve classifying astronomical and zoological images, the success of transcription projects such as Shakespeare's World demonstrates the ability to crowdsource linguistic annotation with proper tasking, tool design, and incentives. Notwithstanding concerns about accuracy, contributions from citizen science can yield high quality results and achieve a throughput that far exceeds that of individual researchers or even small teams of experts.

A portal for citizen linguists attract similar levels of participation as other citizen science communities perhaps more as language is a common experience for nearly every human on the planet. A number of citizen linguist activities will require only native competence. Additionally, connections between language and identity make local pride and cultural preservation powerful incentives. Unified design and branding will encourage a dedicated community of participants. Contributors provide only user name and email during registration though some activities will require, for example demographical and attitudinal metadata which has been shown to correlate with linguistic variation. The portal will mediate access to multiple projects, research results, blogs, news and community fora and provide public recognition for participant contributions. While appealing graphics and simplified tasking and tool design to sustain contributor interest will be important, the portal will also rely on the contributors' desire to participate in language science and technology research, and in linguistic and cultural promotion and preservation.

Like all of the proposed portals, initial activities will be created by LDC with the future ability for researchers to create their own collection and annotation projects using the toolkit. Initial planned activities and tasks include the following.

There are a number of examples of projects that have used crowdsourcing to elicit contributions of recorded speech from participants across the globe. For example, Phonemica[3] preserves Chinese language and culture by collecting contributed stories in 12 different languages with English translations and a map that plots contributor's location. NIEUW will include activities w here citizen linguists record speech samples via computer or mobile device or by telephone via an 800 number.

Another citizen linguist activity will collect neurotypical control data for comparison with clinical populations. A variety of language tasks are among the tools used by clinicians to evaluate patients with Autism Spectrum Disorders or Neurodegenerative Disease. Thousands of hours of recordings are archived, but their value would be enhanced by an increase in matched neurotypical controls. NIEUW will implement versions of elicitation tasks used by the University of Pennsylvania Center for Autism Research as well as open source personality and speech production and perception tests to provide neurotypical controls for comparison to the clinical data.

LDC will create another activity that is a continuation of the GlobalTIMIT project. TIMIT (Garofalo et al., 1993) is the most popular corpus in the history of HLT and LDC has received numerous requests for TIMIT style corpora in other languages. In TIMIT, native speakers read 10 English sentences constructed to maximize the number and combination of sounds. In GlobalTIMIT, naturally occurring sentences replace the somewhat artificial phonetically rich sentences. To date we have applied this method to Thai, Mandarin, Non-Native English and Ga. NIEUW will include an activity with a simple to use

[3] http://phonemica.net

interface to record GlobalTIMIT sentences in many languages.

5.3 Games and Gamified Activities

Human computation and *games with a purpose* (GWAP) target online gamers who “generate useful data” while playing a fun game (Law and von Ahn, 2011). The NIEUW games portal will contain a variety of language games and gamified activities created in house and by collaborators. An early version of this portal, LingoBoingo,[4] is already available and currently includes the language games Phrase Detectives, Tile Attack, Zombilingo, and Jeux des Mots. LDC will have added its own Language ID game to the portal by the time this paper is published.

Inspired by the Great Language Game, our Language I.D. Game will ask players to listen to short audio clips and identify the language spoken. However, our version will improve tracking, language choices, educational potential and, the ability to collect new judgements. Main rounds of the game will present clips where the language is known and ask the player to choose from possible answers with the correct language always an option. These judgments both provide data about language confusability and also serve to demonstrate game player competence in correctly identifying specific languages. Using that knowledge, bonus rounds will be offered where the game player is presented clips of languages for which the correct answer is suspected, but not known. A variety of techniques such as voting algorithms or obtaining multiple judgments for each clip will yield accurate language identification for raw language data.

The game elements of the activity encourage participation. Players score points for every correct answer, but are eliminated after a certain number of incorrect answers. Scoreboards and advanced rounds encourage competition and continued play. As a player increases their score, the scoreboard changes from personal to daily to weekly and so on in order to encourage higher levels of play. Periodic feedback can meet multiple incentives by providing educational information that may also help players improve their score. This information may include summaries of languages provided by Ethnologue, diagrams of language family trees, and information relevant to making game judgements such as phonological properties of a given language.

We are currently advertising the LingoBoingo web portal on a variety of forums from Facebook and Twitter to Linguist List in order to develop brand recognition, solicit and maintain game player participation, and to develop relationships with collaborators who wish to include their own language games on the portal.

6. Infrastructure

Our experience in developing hundreds of language collection and annotation tasks and tools leads us to conclude that a great deal of the technical infrastructure needed to host annotation exercises shared common features whether the exercise appears superficially as a game or a serious working environment. Linguistic data is presented, selected and segmented, segments are classified and labels, annotation records are stored along with information concerning the source media, time and task and contributors. In addition, when multiple contributors make decisions concerning the same segment, it is necessary to determine how to model that variation. For example, should one expose the variation or use a voting algorithm or other method to reduce the variable answers to a single preferred answer.

Given the potential for sharing underlying infrastructure across traditional annotation, citizen science and language games, NIEUW will develop a toolkit for building data collection and annotation exercises inspired by LDC’s WebAnn toolkit which has been used since 2011 by hundreds of LDC contributors to collect tens of millions of linguistic judgements. WebAnn is a single application that presents different GUIs to the user by reusing fixed components while granting control to an annotation task developer who is generally not a software developer. The application has continued to mature in its ability to allow managers to control their work from a redesigned layout manager for tool widgets to a more sophisticated assignment creation feature that tracks the managers' input and reports back on failures.

The NIEUW Toolkit will add support for annotating image and video and improve ease of use for activity designers especially when laying out GUIs, defining workflow, reporting progress and extracting stable corpora. Of particular importance, given the new contributor communities, will be algorithms for task assignment and the integration of variable responses into useable corpora. Planning for the popularity other efforts have experienced, we will also optimize for speed in responding to human actions. The resulting toolkit will accept data contributions that are keyboarded or uploaded and will deploy modules that connect to external data sources including our partners’ user databases, social media accounts, smart phones and tablets, and LDC’s telephone collection platforms. It will be capable of presenting media in whole or part, addressing segments of text by word or character offset, audio and video by time offset and image or video frame by polygons defined in Cartesian space. It will support annotation schemas involving free text, lists, controlled vocabularies or trees for complex taxonomies. Task designers who are not programmers will use tools to lay out GUIs, specify workflows and task assignments and select algorithms for converting annotations that include discrepancies into stable corpora.

The toolkit will also support tutorials, tests and the evaluation of contributor performance versus a gold standard or other contributors. Project managers will have access to tools for monitoring progress and adjusting task assignment. The NIEUW toolkit will include a suite of *gamification* components: progress meters, leaderboards and badges. To support the Citizen Linguist portal, it will include the capacity to add introductory material, blogs and message boards to each activity. The NIEUW version of WebAnn will be released as open source and data collected with this NSF funding will be released at no cost through the LDC.

7. Conclusion

This paper sketched the importance of incentives in language resource development and their impact on workforces, workflows and post-processing for a specific

[4] https://lingoboingo.org

HLT task. The community has experience with various monetary incentives and the necessity to condition found data. However, the HLT community has spent relatively little time trying to consciously engineer incentives and workflows. We described several instances above but believe the field now needs to benchmark its data creation scale and cost against external efforts that have been much more effective. Innovation in language resource creation, employing novel incentives, workforces and workflows is critical if the field is ever to seriously address the demand for HLTs for the world's languages.

8. Acknowledgements

The authors would like to acknowledge the support of the National Science Foundation via CISE Research Infrastructure (CRI) grants CRI CI-P 1629923 and CRI CI-NEW 1730377. We would also like to thank the participants in the workshops on novel incentives held at LREC 2016 and later that same year at LDC. Their contributions also made this paper possible.

9. Bibliographic References

Campbell, Nick. (2016). Herme & Beyond; the Collection of Natural Speech Data. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.

Erard, Michael. (2007). The Wealth of LibriVox: Classic texts, amateur audiobooks, and the grand future of online peer production, Reason 39:1, p. 46.

Eskenazi, Maxine, Sungjin Lee, Tiancheng Zhao, Ting Yao Hu, Alan W Black, Unconventional Approaches to Gathering and Sharing Resources for Spoken Dialog Research. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.

Garofolo, J., L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren and V. Zue. (1993). "TIMIT Acoustic- Phonetic Continuous Speech Corpus, LDC93S1," Linguistic Data Consortium, Philadelphia.

Greenfield, Kara, Kelsey Chan, Joseph P. Campbell, A Fun and Engaging Interface for Crowdsourcing Named Entities. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.

International Association of Conference Interpreters. (2008). Interpreting in Zones of Crisis and War: http://aiic.net/page/2979/interpreting-in-zones-of-crisis-and-war/lang/1

Kominek, John, Alan W Black. (2003). CMU ARCTIC databases for speech synthesis, CMU Technical Report CMU-LTI-03-177, Ver. 0.95, Pittsburgh, PA. Carnegie Mellon University.

Law, E. and L. von Ahn. (2011). Human Computation, San Rafeal, CA: Morgan & Claypool Publishers.

Lewis, M. Paul, Gary F. Simons, Charles D. Fennig (eds.). (2016). Ethnologue: Languages of the World, Nineteenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

METANET. 2010. META-NET White Paper Series: Press Release, http://www.meta- net.eu/whitepapers/press-release-en, accessed March 16, 2016.

Mitsuzawa, Kensuke. Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, Tomoya Mizumoto, FKC Corpus: a Japanese Corpus from New Opinion Survey Service.

Pasachoff, J. M. (1999). "Halley as an eclipse pioneer: his maps and observations of the total solar eclipses of 1715 and 1724," Journal of Astronomical History and Heritage, vol. 2, no. 1, pp. 39-54.

Poesio, Massimo, Jon Chamberlain, Udo Kruschwitz and Chris Madge, Novel Incentives for Phrase Detectives. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.

Prahallad, Kishore, Arthur R Toth, Alan W Black. 2007. Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases, Proceedings of Interspeech, Antwerp, Belgium.

Root, T. (1988). Atlas of Wintering North American Birds: An Analysis of Christmas Bird Count Data, Chicgo, IL: University of Chicago Press.

Székely, Éva, João P. Cabral, Peter Cahill, Julie Carson-Berndsen. 2011. Clustering Expressive Speech Styles in Audiobooks Using Glottal Source Parameters, Interspeech.

Tyson, Na'im, Jonathan Roberts, Jeff Allen, Matt Lipson, Evaluation of Anchor Texts for Automated Link Discovery in Semi-structured Web Documents. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.

Verma, S., S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson, "Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency," in *Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.