Novel Incentives in Language Resource Development

Christopher Cieri
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA
{ccieri} AT ldc.upenn.edu

Abstract

The gap between supply of and demand for Language Resources continues to impede progress in linguistic research and technology development, even in the face of immense international effort to create the requisite data and tools. This deficiency affects all languages in some way, even those with worldwide economic and political influence. Moreover, for most of the world's 7000 linguistic varieties the absence is acute. Current approaches cannot hope to meet the resource demand for even a reasonable subset of the languages currently spoken because they seek to document phenomena of great variability principally using resources, such as national funding, that are highly constrained in terms of amount, duration and scope. This paper describes efforts to augment the traditional incentives of monetary compensation with alternate incentives in order to elicit greater contributions of linguistic data, metadata and annotation. It also touches on the adjustments to workforces, workflows and post-processing needed to collect and exploit data elicited under novel incentives.

Keywords: novel incentives, workflows, language resources

1. Introduction & Motivation

Despite the immense contributions of worldwide data centers, national language corpus projects, government agencies, research groups and individual contributors, the supply of language resources still falls far short of demand. Human Language Technology (HLT) developers experience this shortfall not only in the average but also for every single human language. The METANET (2010) white paper series documents the language resources required to build the technologies needed to *future-proof* European language against *digital extinction*, that is to allow them to participate in an increasingly digital, information driven world. As the reports compare need to existing resources for EU languages, they demonstrate that no language, not even English, enjoys the full range and that "*21 out of 30 European languages could become extinct in the digital world*". What is true for EU languages is at least as much true for the remainder of the world's languages.

Success in the digital domain is only one of many motivations for creating HLTs and the pre-requisite resources. A 2008 report from the International Association of Conference Interpreters warned: "*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.*" Although technologies have the potential to streamline disaster relief, the delay between the onset of the disaster and the integration of the technology continues to thwart relief efforts: "*Effective communication in Haiti was confronted by language barriers and the limited utilization of technology. Media played an important role in communicating about the disaster relief effort to the international community, but their reporting at times included misinformation.*" (Harvard Humanitarian Initiative 2011).

HLTs have a growing role – and will have a critical role in the future – in disaster relief. Varma et al. (2011) showed their potential by using natural language processing techniques to filter tweets, with 80% accuracy, according to whether they provided situational awareness. However, the system required training data to be annotated not only for situational awareness but also for subjectivity, formality, and personal versus impersonal viewpoint. In addition, their automatic processing included a part-of-speech tagger, which cannot be assumed to exist for most low resource languages. Indeed even the tokenizer, list of stop words, unigram and bigram frequencies are absent for many of the world's languages almost certainly some that will figure into future disasters.

A number of US government programs over the past several years have begun to address the need for HLTs and pre-requisite LRs to support disaster relief efforts. In 2011 the National Science Foundation (NSF) provided $2.8M in support to the EPIC (Empowering the Public with Information in Crisis) project at U. Colorado and U.C. Irvine researching technologies to facilitate disaster relief communications. DARPA LORELEI is developing technologies to deal with disaster related communication in low resource languages. However such programs last for just a few years and provide their impressive array of resources for at most a few dozen languages. The 19th edition of the Ethnologue (Lewis, Simons, Fennig 2016) reports the tally of living languages to be 7,097 worldwide most of which lack the resources required by Varma et al.'s system and will not be the focus of LORELEI or any current program.

Finally, the societal need for multilingual technologies and enabling data extends well beyond commerce, defense, and disaster relief. A 2010 article published by the American Psychological Association echoed the growing need for greater translingual capability, which they characterized in terms of interpreters within counseling services.

In summary, current approaches to HLT and LR development will not meet the needs of human language technologies for the world's languages or even an appreciable subset of them. In order to scale significantly beyond current production it will be necessary to revolutionize multiple aspects of LR development including the conceptualization of the tasking, the target workforces and their motivations, the workflows used to acquire data, metadata and judgments and the post-processing necessary to exploit next generation LRs in the development of HLTs.

2. An Incentives-Aware Model of Language Resource Creation

Each process that creates data or annotation used in linguistic research and technology development, whether it does so intentionally or in the service of some other goal, can be seen in terms of several interacting components: the task, the incentives offered, the work-force that the incentives attract, the workflow required to permit to workforce to complete the required task and the output. Different workforces are motivated by different incentives, require different tasking and workflows and produce different outcomes. All of these factors impact the researcher who would use the data as well as the organization that would collect it. Greenfield, Chan and Campbell (2016) provide an example: "*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.*"

The Human Language Technology (HLT) communities are already familiar with found data types such as newswire and broadcast news that are created for purposes unrelated to HLT and rely upon workforces and workflows outside their control. For data types created specifically to support HLT research and development, common incentives include monetary compensation and in smaller scale efforts the potential to use the data for ones own research. However the conscious engineering of incentives, workforces and workflows to optimize output for a specific task is rather limited within the HLT LR production. There are obvious counter-examples. Much of the recent work on crowd-sourcing discusses the impact f factors such as HIT size and complexity, payment rate, and instructions on the quality of the outcome and design sophisticated interfaces to harness the wisdom of the crowd, and reduce cheating. However this valuable research relies principally on the incentive of monetary

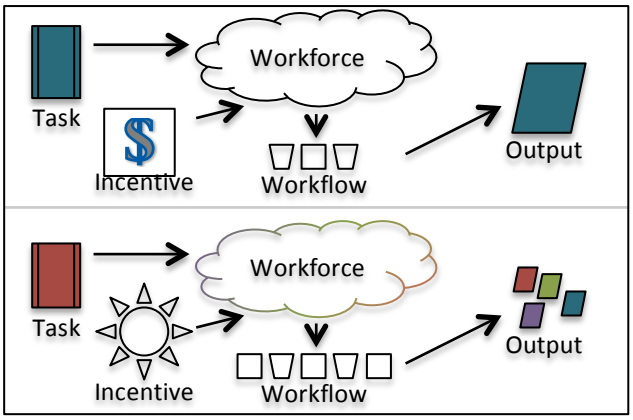

Figure 1: Different incentives attract different workforces that require different tasking and workflows and produce different outputs.

compensation. In much older work multiple LDC studies (Cieri et al. 2006, 2007) have reported on the relative effects of graduated pay scales, completion bonuses and random prizes upon performance in telephone collections. However, again, the incentives were principally pecuniary. In the next section we will review some very recent work within HLT communities in engineering incentives and/or engineering workflows to deal with data created under non-traditional incentives. These include cases of HLT development for industry where the specific combination of workforce, incentive and workflow is the target environment for the technology as well as other cases in which the environment has been engineered to provide data for some other purpose.

3. Incentives in Language Resource Development for HLT

In the sections below we focus on several very recent efforts with the HT communities to make use of novel incentives in data collection and annotation including new workflows and post-processing necessary to use such data in system building.

3.1. Collection

Campbell (2016) reports on a number of data collections intended principally to support the development of systems capable of producing expressive speech. These collection efforts experimented with a variety of incentives and adjusted to the different characteristics of the output. In addition to any monetary compensation, additional motivations included access to the resulting data for research purposes, sustenance, curiosity, fun, ability to keep the recording device used and opportunities for unusual social interactions including apparent conversations with a robot and extended interactions with colleagues outside the lab. The data resulting from these studies naturally varied along many dimensions including the proportions of regional speech, emotion, non-speech vocalizations and contact events. Based on his own experience in both worlds, Campbell also emphasizes the growing divide between academic and industrial HLT research especially in terms of data

volumes. From our perspective in this paper, the motivations of acquiring some product or service can be seen as leading commercial customers to provide vast quantities of 'data' to HLT researchers in industry.

Continuing that theme, Mitsuzawa et al. (2016) describe their efforts to process product and company reviews from the Fuman Kaitori Center. Like many developers in industry they enjoy a reduced train-test mismatch because the data they use to build their systems is quite similar to, or an earlier instantiation of, the data their system will ultimately process. Consumers post their reviews initially to communicate some dissatisfaction with a product or service to the responsible company. A second order incentive is the opportunity to receive points that convert into monetary value, based on the length of the review and the quality of associated metadata. The mixture of incentives naturally yields variation in the data including duplicate, vacuous or offensive posts, variable renderings of named entities and inaccurate metadata necessitating post-processing that is fed by human annotation.

3.2. Annotation

Greenfield, Chan and Campbell (2016) describe their experiments in crowd-sourcing annotation to support information extraction research. They note that at least some of their workforce of Mechanical Turkers seemed to be motivated by the quality of the interface design and the desire to maintain a high approval rating as well as the monetary incentives. By focusing their system improvements on interface design they elicit higher quality data without attracting a mercenary element interested only in highly compensated work.

Poesio et al. (2016) describe *Phrase Detectives*[1], a game-with-a-purpose for collecting anaphora annotation. Players' incentives, in addition to entertainment, are interesting source material, a variable point system, the opportunity to progress through experience levels, leaderboards, the social motivations of teaming with friends in the Facebook version and prizes awarded via a lottery and also according to performance.

The Great Language Game (GLG) asks contributors to listen to short audio clips and indicate what language is spoken. Clips are currently selected apparently at random from 80 languages so that most players are not speakers of most of the target languages. Although created in 2013, The Great Language Game (GLG) has already collected millions of judgments. The developer, Lars Yencken released a corpus of 16 million judgments collected through March 2014 though we estimate that the number collected to date is more than double that amount. GLG employs incentives of information, entertainment, competition and status. Players compete against posted high scores and can brag about their accomplishments in a forum created for contributors. The game displays Ethnologue posts for languages the player has misidentified and players report finding the work fun. In its first year, GLG created a volume of language identification judgments significantly greater than all of the judgments created to support all of the NIST Language Recognition Evaluations since the campaign began in 1996. However, these annotations are not directly useful for LRE. Because the games relies on the ability to tell players when they have gotten an answer correct, each new judgment adds little information about a clip whose language is already known though the many judgments for each clip provide information about confusability.

3.3. Exploitation

Tyson and colleagues (2016) describe their research on automated link discovery among *About.com*[2] texts. Their work shows that, compared to the corporate mission of recirculating users to maximize exposure to advertising, the different motivations of content creators leads them to create fewer links than desired, a problem that the research team is now addressing through a combination of automated techniques and additional human annotation.

Eskenazi et al. (2016) describes a series of dialog system research and development efforts that have employed novel incentives such as automated access to information and the promise of an improved customer experience in real world interactions. The data resulting from the efforts naturally contain challenging levels of noise and variation in speech. Eskenazi and her colleagues at the DialRC Center have extended the notion of novel incentives to apply to the research community as well as the subjects of a study or users of a system. By offering free access to their data and dialog system and by organizing a range of outreach activities, they continue to attract researcher cycles to problems of interest to them. A recurring theme of community organized shared task challenges is that: "*optimization for lab test subjects may not reflect the outcome with real users*".

4. Language Data Collection outside HLT

Despite the obvious benefit to HLT development, initiatives outside the HLT communities have employed novel incentives more frequently in a wider range of contexts and to greater effect. In many cases, the motivation for such collections is quite remote from HLT developers' goals. Furthermore, neither the contributors nor the leaders of the effort may see what they do as language data collection; however we will show here that their outcomes may be extremely beneficial to research in linguistics and language technology both directly and as a model of collections that we may imitate.

4.1. Librivox

LibriVox[3] creates "free public domain audiobooks" by recruiting, training and organizing volunteers who record

[1] https://anawiki.essex.ac.uk/phrasedetectives/

[2] http://www.about.com/

[3] www.librivox.org

themselves reading literary works that are out of copyright in the US. LibriVox readers also declare their recordings to be in the public domain. As of March 25, 2016, the LibriVox catalog listed 10,185 books [4] comprising at least 57,369 hours of read speech. Approximately 86% of all LibriVox recordings are in English. However, there is at least one hour of speech in at least 31 other languages. Figure 1 shows the growing volume of recordings by language, measured in hours of speech as indicated in the LibriVox Catalog.

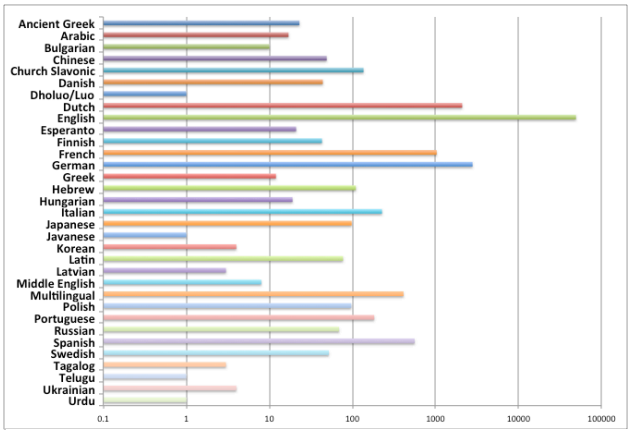

Figure 2: LibriVox Hours Recorded per Language on a log(10) scale

LibriVox recordings are typically careful readings, often of well-known works of literature for which the original written text is also available digitally. Sound quality is variable but generally good with many of the readings made in quiet environments using appropriate recording equipment and undergoing quality control by an independent producer. One or more readers may produce a single work dividing the effort either by chapter or by character. A single reader may also read multiple characters, using different voices and accents when the text seems to call for it. Most readers are amateurs from around the world, including some non-native speakers. Many LibriVox volunteers produce multiple works thus providing multiple samples of their voices over time and under different circumstances.

LibriVox recordings are relevant to a number of HLT fields including language, speaker and speech recognition. However the HLT area that makes the greatest use of LibriVox is probably speech synthesis where the large volumes of long-duration, read speech across a number of genres supplements existing data used to build TTS systems. In one early example of this use, Prahallad, Toth and Black (2007) built statistical parametric speech synthesis systems with male and female voices from a total of ~11.25 hours of LibriVox reading. They compared these to similar systems built from the Arctic Corpus (Kominek, Black 2003), designed specifically to support speech synthesis research, and concluded that "*a voice could be successfully built from large multi-paragraph speech using automatic segmentation tools.*" Braunschweiler, Gales and Buchholz (2010) used lightly supervised, recognition-based alignment to select paragraphs as training material for a speech synthesis system appropriate for reading longer extents of coherent text. Székely et al. (2011) experimented with approaches to clustering utterances in LibriVox readings according to voice quality parameters in order identify utterances associated with different voice characteristics and use them to build systems capable of synthesizing "*speech which is rich in prosody, emotions and voice styles.*" Mamiya et al. (2013) experimented with and evaluated lightly supervised VAD prior to grapheme-based alignment of LibriVox audio to corresponding text in the process of building TTS systems. The VAD system required 50 sentences of the same text to be hand aligned. To evaluate the systems they elicited 90 preference decisions from each of 20 judges who listened to system output. They concluded that the performance of the lightly supervised systems was equivalent to that of their fully supervised system. Proctor and Katsamanis (2011) elicited judgments from 13 listeners concerning the *felicity* of multiple LibriVox readings. Although the judges as a group clearly preferred some and dis-preferred other readers, individual preferences foiled a rigorous classification. Similarly, attempts to correlate preferences with standard prosodic measures failed to create a robust classification of reader felicity.

These studies show both the benefits of using sources like LibriVox in HLT development as well as the pre-processing needed to condition it. To the extent that the processing can be done efficiently sources like LibriVox become critical additions to the set of available LRs for HLT development, data that would be impossible to create using the traditional incentive models in our field. Each hour of recorded LibriVox audio apparently requires 2 hours of reading time and 2 to 4 hours of editing time, meaning that the initiative has elicited at least 229,476 hours of volunteer labor and probably much more. Assuming rates of $500 per finished hour of audio, one would have paid more than $28 million to produce the same material professionally. Volunteers make such enormous contributions for a variety of reasons. Many believe in the LibriVox mission or its connection to the broader open source or free culture movements (Erard 2007). Some enjoy reading aloud, in some cases continuing or expanding an activity they began with friends or family. Others are happy to think they are helping maintain the art of storytelling. Some clearly enjoy collaborating with others of similar interests and having the ability to control the size of their own contributions. A small number of the best readers also receive paid work through Iambik[5], a spin-off audiobook company, or parlay their LibriVox experience into

[4] What the LibriVox Catalog tags as a <book> is typically a single reading of a work which could also be a pamphlet, poem or collection of poems. Additional readings of the same work receive their own Catalog record. Thus there are fewer than 10,185 unique titles.

[5] www.iambik.com

professional narrator with Audible[6], ACX[7] or similar organizations. LibriVox stands not only as a data source but as a model of how initiative may use non-monetary incentives effectively.

4.2. Citizen Science: Zooniverse

Outside HLT, other research disciplines have effectively engineered environments to collect data using non-monetary incentives. Zooniverse is a citizen science portal with many opportunities to contribute to research most of which is in the hard sciences. Tasks include identifying signs of movement in star fields, classifying animal species based on photographs and transcribing museum records for insect specimen collections. The beautiful interfaces are fine grained tasking attract participants and allow them to complete meaningful tasks in minutes. More than 800,000 volunteers have registered, contributed data toward the science of many peer-reviewed publications and even made serendipitous discoveries of astronomical objects.

5. Future Directions for Language Research Development

The initiatives sketches above make it clear that there are numerous opportunities to acquire data from corpora developed under non-monetary incentives and to engineer environments with optimal combinations of incentives and workflows to develop data products for specific tasks. For example, a citizen science-of-language portal could attract equal or greater contributions because while the sciences are only one of many areas of intellectual interest, language is a common experience for nearly every human on the planet. Tasks for citizen linguists could require nothing more than native speaker ability and could scale according to the dedication of the workforce. Finally, for many, language is connected to identity so that local pride, cultural preservation and "putting ones language on the map" become additional incentives. Additionally, games-with-a-purpose, gamified interfaces and even soberer efforts that pay attention to task size and complexity relative to the workforce can increase efficiency and quality.

6. Conclusion

This paper has opened the dialog on incentives in language resource development and how they attract different workforces and require different workflows in order to optimize outcomes for a specific tasking. The HLT community is quite familiar with the impact of various monetary incentives and the effort needed to condition data acquired under non-traditional motivations, for example found data. However efforts to consciously engineer incentives and workflows within HLT have been rather limited. We described several in this paper but also believe the field needs to benchmark its data creation efforts 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 serious address the demand for HLTs for the world's languages.

[6] www.audible.com
[7] www.acx.com

7. Acknowledgements

The author would like to thank the participants in the LREC 2016 Workshop on Novel Incentives for Collecting Data and Annotation from People. Their contributions made this paper possible.

8. References

Braunschweiler, Norbert, M.J.F. Gales, Sabine Buchholz. 2010. Lightly supervised recognition for automatic alignment of large coherent speech recordings, Interspeech, Makuhari, Japan, September 26-30.

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.

Cieri, Christopher, Walt Andrews, Joseph P. Campbell, George Doddington, Jack Godfrey, Shudong Huang, Mark Liberman, Alvin Martin, Hirotaka Nakasone, Mark Przybocki, Kevin Walker. 2006. The Mixer and Transcript Reading Corpora: Resources for Multilingual, Crosschannel Speaker Recognition Research, 5th International Conference on Language Resources and Evaluation, Genoa, May 22-28

Cieri, Christopher, Linda Corson, David Graff, Kevin Walker. 2007. Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora, Interspeech: 10th International Conference on Spoken Language Processing, Antwerp, August 27-31

DeAngelis, Tori. 2010. Found in Translation in Monitor on Psychology, 2010, Vol 41, No. 2, print version: page 52, American Psychological Association, http://www.apa.org/monitor/2010/02/translation.aspx

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.

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.

Harvard Humanitarian Initiative. 2011. Earthquake Relief in Haiti. http://hhi.harvard.edu/sites/default/files/publications/earthquake-relief-in-haiti.pdf

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.
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.
Liberman, Mark, Oral Histories: Linguistic Documentation as Social Media. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.
Mamiya, Yoshitaka, Junichi Yamagishi, Oliver Watts, Robert A.J. Clark, Simon King, and Adriana Stan. 2013. Lightly Supervised GMM VAD to Use Audiobook For Speech Synthesiser. ICASSP.
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
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.
Proctor, Michael, Athanasios Katsamanis. 2011. Prosodic Characterization of Reading Styles using Audiobook Corpora, 162nd Meeting of the ASA, Thursday November 6, San Diego, CA
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.