New Directions for Language Resource Development and Distribution

Christopher Cieri, Denise DiPersio, Mark Liberman, Andrea Mazzucchi, Stephanie Strassel, Jonathan Wright

Linguistic Data Consortium. University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA
E-mail: {ccieri|dipersio|myl|amazzu|strassel|jdwright}@ldc.upenn.edu

Abstract

Despite the growth in the number of linguistic data centers around the world, their accomplishments and expansions and the advances they have help enable, the language resources that exist are a small fraction of those required to meet the goals of Human Language Technologies (HLT) for the world's languages and the promises they offer: broad access to knowledge, direct communication across language boundaries and engagement in a global community. Using the Linguistic Data Consortium as a focus case, this paper sketches the progress of data centers, summarizes recent activities and then turns to several issues that have received inadequate attention and proposes some new approaches to their resolution.

Keywords: data center, language resources, collection, annotation, data distribution, infrastructure

1. Introduction

Since the Linguistic Data Consortium [1] (LDC) was established more than two decades ago, such data centers have played a large and growing role in creating Language Resources (LR) to support linguistic education, research and technology developments. Despite the many thousands of hours of audio and billions of words of text collected and the hundreds of millions of annotation decisions made, the resources created are but a small percentage of those needed. Current focus on narrowly defined program goals with short-timelines and rapid incremental changes in program focus has enabled impressive development of many technologies for a small number of languages. To begin to address the problem of processing a significant percentage of the world's language would require data centers to augment project centered approaches with others that allow ongoing LR growth in an arbitrarily large number of languages.

2. Data Center Progress

Over the past 22 years, LDC activities have grown to include all aspects of creating, distributing and archiving LRs, including quality control, intellectual property rights, human subject protocols; data collection, annotation and lexicon building; tool, specification and best practice development; research on LRs, knowledge transfer via documentation, metadata, consulting and training; support of large multisite programs; workshop organization and service to multiple research communities via review panels and advisory boards.

Despite differences in business models, other data centers have had similar trajectories. For example, the European Language Resource Association [2] (ELRA) established a few years after LDC has also expanded from an archive and publisher of language data into the broader role described above and also coordinates technology evaluation campaigns. Other data centers, Chinese LDC[3], LDC for Indian Languages,[4] South African Resource Management Association [5] follow similar paths, sometimes explicitly modeling LDC or ELRA.

2.1. HLT Program Support

LDC's support of multi-site HLT programs includes: needs assessment for sponsors, developers and evaluators, developing priorities and timelines to translate wish lists into action plans, coordinating LR activities within and across programs and sponsors, integrating HLTs into data production while supporting robust, objective system evaluation, rapidly cataloging, licensing, replicating and distributing LRs among program participants, broadening program impact and encouraging additional research by distributing LRs to external researchers while protecting restricted data such as evaluation sets.

2.2. Data Collection

The ever growing list of data types LDC collects and publishes includes text from news sources, journals, financial and biomedical documents; internet sources including newsgroups, (micro)blogs and discussion fora; text interactions via email, chat and SMS; and printed, handwritten and hybrid documents, for example printed forms completed by hand. LDC also collects audiovisual data from broadcast news and conversation, podcasts, conversational telephone speech, lectures, interviews, meetings, field interviews, read and prompted speech, task oriented speech, role play, speech in noise, web video and even animal vocalizations. LDC also digitizes analog media including interviews in a variety of tape formats.

[1] http://www.ldc.upenn.edu
[2] http://www.elra.info/
[3] http://www.chinip.csdb.cn/
[4] http://www.ldcil.org/
[5] http://rma.nwu.ac.za/

2.3. Annotation

The annotation types in which LDC has developed expertise have also grown rapidly over the past two decades and include: data scouting, data triage and smart data selection; alignment of paired audio streams; auditing for bandwidth, signal quality, language, dialect, program and speaker; quick, quick-rich and careful transcription, audio segmentation and audio-text alignment at story, turn, sentence and word level; orthographic, spelling and phonetic script normalization and transliteration; tagging of phonetic, dialect, sociolinguistic and supralexical features; document zoning, handwriting transcription, OCR QC and tagging of reading order; tokenization and tagging of morphology, part-of-speech and gloss; Treebanking, PropBanking, SemBanking; sense disambiguation, fine and coarse-grained topic relevance annotation; novelty, text entailment, hypothesis generation and inference annotation; annotation of committed belief, sentiment, disfluency, discourse features and hedging; detection and classification of entities, relations, events, time, location and their co-reference in text; knowledge base population; single and multi-document summarization of various lengths from titles to 200 words; query generation and question answering; translation, multiple translation, edit distance, translation post-editing and quality control; alignment of translated text at document, sentence, phrase & word levels; describing the physics of gesture via joint angles and rotations; identification, classification and tracking entities and events in video; assessment of IR, MT, KBP, QA and other system output.

2.4. Projects

LDC supports several large, multisite HLT development and evaluation programs coordinating LR activities to meet all stakeholder needs.

For DARPA DEFT (Deep Exploration and Filtering of Text), LDC produces LRs for three research areas: Relational Analysis, Semantic Filtering and Anomaly Analysis by annotating discussion forums and other informal genres in English, Chinese and Spanish, for a wide variety of tasks. In *Entities, Relations, Events* (ERE), annotators label documents for the entities mentioned, relationships among those, and the events in which they participate, co-referencing multiple mentions of the same entity or event. In *Abstract Meaning Representation* (AMR), annotators produce whole-sentence semantic representations via rooted, labeled graphs. In *Textual Entailment*, annotators judge whether pairs of sentences entail or contradict one another. *Inference* annotation requires enumerating the reasoning steps required to reach the each entailment judgment. Finally, annotators label documents for *Committed Belief*. Future efforts will include modalities like sentiment. With DEFT support, LDC produces resources for the NIST TAC KBP evaluation. In 2014, LDC annotators will create English, Spanish and Chinese queries, annotations and assessments for five evaluation tasks: *Cold Start*, building a knowledge base from scratch; *Entity Linking*, linking entity mentions in text to knowledge base entries and extracting information from unstructured texts about entities (*Slot Filling)*, *Events* and *Sentiment*.

For the DARPA BOLT (Broad Operational Language Translation) program, LDC produces LRs for training and evaluating machine translation (MT) and information retrieval (IR) technologies focusing on informal genres of English, Chinese and Egyptian Arabic. MT resources include large volumes of source text or speech and millions of words of sentence-aligned parallel text. Much of the parallel text is also manually word aligned, Treebanked, PropBanked and annotated for entity and event coreference. LDC annotators also post-edit MT system output to produce HTER scores for system evaluation. For IR, LDC has produced over 250 natural language queries in three languages and manual assessment of IR system output among other LRs. The final phase of BOLT, currently underway, shifts attention from online discussion forums and naturally occurring SMS/chat messages to conversational telephone speech.

To support DARPA RATS' (Robust Automatic Transcription of Speech) goal to process potentially speech-containing signals received over extremely noise, distorted channel, LDC collected and annotated thousands of hours of conversational speech in multiple languages for four tasks: Speech Activity Detection, Language ID, Speaker ID and Keyword Spotting. Rather than collect and annotate noisy data, LDC produced the desired signal by simultaneously rebroadcasting clean, annotated data over 8 independent radio channels configured to introduce various types of signal noise. Post-processing then projected the original audio annotations onto the degraded recordings. In RATS' final phase, focus shifted to smaller data sets supporting search for speech in large, heterogeneous, unpredictable communications containing disruptions, interference, competing transmissions, and non-speech audio artifacts.

In 2014 LDC concluded work on DARPA MADCAT (Multilingual Automatic Document Classification, Analysis and Translation), whose goal was to develop technology for converting foreign text images into English transcripts. LDC collected over 68,000 handwritten pages of Arabic and Chinese data, scanned and annotated them with sentence and token bounding boxes. Over 40,000 pages were also ground truth annotated and a portion of all data was translated into English. The program's final phase focused on resources for structurally complex documents, in particular those containing tables and ledgers with handwritten text.

LDC continues to produce new LRs to support language recognition R&D. Our current collection effort will yield hundreds of samples of narrowband speech extracted from telephone calls and broadcasts in twenty languages and dialects, including confusable varieties. Native speakers audit each recording to verify its audio properties and language. The resulting corpus will support the next NIST Language Recognition Evaluation, expected to take place in 2014 or 2015.

Since kickoff in 2007, the Heterogeneous Audio Visual Internet Collection (HAVIC) program has collected several thousand hours of amateur video, annotated it for multiple features including genre and topic, and provided a synopsis. Annotators also indicate whether videos contain one of the pre-defined HAVIC events (e.g. Making a Sandwich), depicted in the video, audio or embedded text. The HAVIC corpus has been used in the TRECVid Multimedia Event Detection (MED) and Multimedia Event Recounting (MER) Evaluations since 2010.

To support language and speaker recognition work within the VAST (Video Annotation for Speech Technology) program, LDC collects thousands of hours of amateur and broadcast video in over a dozen languages and annotates it for speech activity and language. A portion of the corpus is carefully transcribed and labeled for speaker identity.

2.5. Research

LDC maintains a program of research into linguistic data and related technologies. Earlier work investigated topic detection and tracking (Schultz, Liberman 2000) and linguistic annotation formalisms (Bird, Liberman 2001). More recent work involves big data approaches to sociolinguistics (Cieri et al. 2008) including automated analysis of dialect features (Yuan, Liberman 2011), demographic, situation and attitudinal metadata (Yaeger-Dror, Cieri 2013) and error analysis in Treebanks and parsing (Kulick, Bies, Mott 2011). Very recent work applies novel techniques to phone segmentation and labeling (Yuan et al. 2013), speech activity detection (Ryant, Liberman, Yuan 2013) and tone classification (Ryant, Yuan, Liberman undated).

2.6. Data Publications

Essentially all data LDC produces for sponsored program are published as are contributions from our partners. Since the last LREC, LDC has released[6]:

- Broadcast conversation in Arabic (2013S02, 2013S07) and Chinese (2013S08, 2013S04) with transcripts (2013T04, 2013T17, 2013T20, 2013T08, respectively)
- Conversational telephone speech in Persian (LDC2014S01) with transcripts (LDC2014T01)
- Read and Spontaneous Speech in Arabic contributed by King Saud University (2014S02)
- Deceptive Speech (2013S09) contributed by Columbia University, SRI and University of Colorado, Boulder
- Multichannel Read Speech in a meeting room (2014S03) contributed by the Edinburgh Centre for Speech Technology Research
- USC-SFI MALACH interview speech and transcripts in Czech (2014S04) and English (2012S05)
- Speaker Recognition Corpora including Mixer 6 (2013S03) and Greybeard (2013S05)
- Parallel text in
  - Arabic from newswire (2012T17), broadcast news (2012T14, 2012T18) and web text (2013T01) and in Arabic Dialects (2012T09)
  - Chinese from broadcast news (2014T04), broadcast conversation (2013T11, 2013T16) and scientific text (2013T02, 2012T22, contributed by MITRE)
  - Russian from technical text (2012T23)
  - Multiple languages from 1993-2007 United Nations parallel text (2013T06) produced by Google Research
- Word Aligned (Tagged) Text in
  - Arabic from newswire and web sources (2014T05)
  - Chinese from broadcast (2013T23), newswire (2012T20), web sources (2012T24, 2013T05) and combined genres (2012T16)
- Post-Edited Translation Text with similarity values in Arabic (2013T18) from the STS 2013 shared task
- Timebanks in Catalan (2012T10) and Spanish (2012T12) contributed by Barcelona Media
- Treebanks in
  - Arabic from broadcast news (2012T07)
  - Chinese (2013T21) from U. Colorado and Brandeis
  - English from web text (2012T13)
  - the English Gigaword with syntactic and discourse annotations (2012T21) contributed by the JHU Language Technology Center of Excellence
  - and a new OntoNotes release (2013T19)
- Parallel (Aligned) Treebanks in
  - Arabic from news text (2013T10), broadcast news (2013T14, 2014T03) and web text (2014T08)
  - and the Prague Czech-English Dependency Treebank (2012T08) contributed by Charles University, Prague
- Chinese PropBank (2013T13) from U. Colorado and Brandeis
- Manually Annotated Sub-Corpus (2013T12) of the ANC contributed by Nancy Ide and colleagues
- ARRAU Corpus of Anaphoric Information (2013T22) from the Universities of Essex and Trento
- Domain-Specific Hyponym Relations (2014T07) of English from Xian Jiaotung University
- Page images and transcripts with bounding boxes and reading order (2012T15, 2013T09, 2013T15)
- Corpora supporting NIST OpenMT evaluations (LDC2013T03, LDC2013T07, LDC2014T02)
- Maninkakan Lexicon (2013L01)
- American English Nickname Collection (2012T11)

2.7. Distribution

Since 2012 LDC has published new corpora at an average rate that has grown from 2.5 to 3 per month. As of the time this paper was written, more than 108,000 copies of 1860 titles had been distributed to more than 3500 organizations in 70 countries. Recent developments promise to further improve distribution.

LDC is completing a redevelopment of its business system, the database and associated processes supporting membership and catalog functions. The new version incorporates an updated software architecture and e-commerce principles. Users may register accounts, join LDC and license data online. Transactions include

[6] LDC Catalog numbers, minus the LDC prefix, follow each corpus description in parentheses. See: catalog.ldc.upenn.edu

automated, secure credit card processing, digital signatures for membership and license agreements and automatic email notifications. Organizations have greater flexibility to manage their accounts. Catalog metadata schemata conform more closely to the OLAC[7] standard. LDC will continue to offer the personal support members have come to expect while the business system's new automation will streamline the user experience.

In 2013, LDC began Cloud-based data delivery to support shared tasks, including the REVERB challenge, SemEval and SPMRL, and to meet occasional unusual needs. Data size has varied up to one terabyte. The process is transparent to users who request data from LDC's distribution system and receive it directly or via the Cloud depending upon how the path is configured. Not surprisingly, delivery speed varies with data size and available bandwidth. Cost savings can be significant compared to media duplication, shipment and associated human effort. Our initial efforts provide a basis from which to evaluate broader cloud-based distribution. LDC has also joined Brandeis Vassar and CMU to develop the first US Language Applications Grid that provides access to tools and data as web services. This infrastructure allows users to create pipelines that connect resources and processing on the Grid removing the need to maintain copies of data and technologies locally.

2.8. Outreach

LDC outreach efforts include the monthly newsletter, occasional member surveys and service to multiple research communities via advisory boards (e.g. LINDAT-CLARIN) and funding panels, among other efforts below.

In October 2013 LDC introduced its new website, the first significant change to site structure in over a decade. The enlarged and reorganized site map highlights LDC activities – member services, language resource preparation and distribution, data management, project support, collaborations – and simplifies the user experience while the sites' content management system reduces staff effort.

In September 2012 LDC celebrated its twentieth anniversary with a workshop in Philadelphia that focused on the future of language resources. Invited speakers from around the world discussed new domains, and emerging innovations in data collection and distribution. In addition, the workshop fueled some of the ideas that appear in this paper. We also continue to participate in major conferences as planners, speakers and exhibitors: ACL, ALA, ICASSP, ICA, ICPhS, IEEE, Interspeech, LSA, NWAV, Odyssey and LREC. We also attended the NLP12 meeting, ELRA 18th Anniversary, and DGA Workshop on Multimedia Information Processing as well as the Errare, AARDVARC, DARPA Data Framework and Sustaining Domain Repository workshops.

2.9. Systems Infrastructure

The infrastructure enabling this progress has itself undergone significant change in the past two years, following two broad principles.

2.9.1. Outsourcing Commodity Services

As the first international center for linguistic data, LDC has sometimes sat on the bleeding edge of innovation, needing to implement technologies that were unavailable or inadequate in the open market. However, advanced technologies eventually become commodities so that the capabilities offered in general are adequate to meet the needs of the data center. With this in mind, LDC has outsourced most of its network and telephony and website. Allowing local IT staff to focus on technologies that meet unique Consortium needs. This outsourcing has resulted in greater capacity and business continuity at a cost savings to members. Reducing support demands on local Systems staff has allowed them to devote more time innovating storage and compute solutions, system monitoring and network authorization. LDC continues to locally manage its business system and catalog, the unique telephone system used in call collection.

2.9.2. Generalizing Internal Processes

Among internally managed technologies, traditional approaches are no longer compatible with the very rapid ramp-up and turn-over recently characterizes some of our sponsored programs. For example, some new projects couple demand for massive storage and compute power with timelines that do not permit its acquisition. To address this problem, LDC has developed Storage- and Computation-as-a-Service. Central servers already in place offer virtualization and provisioning. New projects do not acquire new hardware but use slices of existing capability scaled to their need at the time. Costs to projects are currently lower than those in the open market and fees contribute to upgrades and expansions. This approach has also prepared LDC for a smoother transition to commercial products once their prices fall below ours. Within the new storage solution, we also distinguish the different life-cycles of data, annotations and other file providing different storage tiers, backup policies, data protection strategies, availability and costs appropriate to each.

2.10. Software Development

Recent LDC software development has focused on creating a unified framework not only for collection and annotations technologies but also for the very tools developers use to build them. The principles guiding this evolution are:

- generalize infrastructure across projects
- modularize functions via web services
- reduce reliance on OS, directory structure and file I/O
- normalize structure of all annotations
- employ single code base to satisfy (nearly) all needs

[7] http://www.language-archives.org/

The following sections illustrate these principles by discussing specific new developments.

For many years, the goal of LDC software development was meeting the needs of specific projects. While this narrow focus has allowed us to rapidly scale up collection volume and provide custom solutions to program needs, it also reduces opportunities to reuse code and reduce cost by building upon prior knowledge. The LDC People database is a good example of reversing this trend. During the creation of the Greybeard[8] corpus it was necessary to combine all prior human contributor databases in an attempt to find those willing to participate in another study. Having completed a universal contributor database we used it for all subsequent collections, extending schemata appropriately, for example to distinguish contributors in different studies, and developing general enrollment software. The latest version of LDC People includes all contributors: annotators and managers as well as human subjects. Sensitive personally identifying information (SPII) is stored in an entirely separate database from the demographic and performance data associated with contributors so that access to the former can be restricted to just those processing compensation. Finally, all developers access LDC People though a single API.

LDC developers also use web services to simplify access to raw data. Where, in the past, each annotation tool would access text, audio or video directly from disk storage allocated to the task, today a combination of web services and associated databases store the location of each newly collected media file, index the contents of text documents using the Solr engine, track the processing and annotations applied to each and retrieve documents on demand. This approach also allows us to organize media files by type and collection epoch making them easier to store and retrieve.

Beyond the use of web services, LDC developers have also revolutionized tool design in the past few years. Most annotation tools are now web-based and exploit unified task and workflow management and reporting to retrieve and manipulate source material and annotations in a consistent way. Within this framework all annotations are stored in relational database records that track the annotator, time of annotation and atomic annotation decisions. A single WebAnn GUI typically writes many such atomic decisions into the database. The annotation decision, when unpacked, may contain information about extents, annotation types and values but even these conform to a controlled vocabulary. Many changes to the GUI, can now be done by a non-technical task manager using a screen layout tool without requiring changes to database schema. It is also worth noting that additional annotations about the same piece of raw data do not overwrite each other. Each decision is stored and associated with its annotator and time so that subsequent processes can decide how to handle disagreements. This 'transactional' approach to annotations also allows us to roll back the database to a certain point in time when necessary.

Multiple workflows have been pre-defined allowing coordinators to assign tasks to annotators manually or to instruct the tool to assign them at random, or in first-come-first-served order. A specified percentage can be assigned to multiple annotators in a double blind fashion allowing measurement of inter-annotator agreement. Particular assignments, or prohibitions on assignments, can be defined independent of the annotation GUI used. Coordinators monitor progress in a consistent way independent of project. This approach allows project managers and developers to collaborate to rapidly deploy new annotation tools. For the simplest cases, managers can simply lay out widgets as desired. Where new code is truly necessary, developers insert it in a standardized way that reduces the total amount of code created and the need to understand the internals of existing code. The same tool that allows managers to create new annotation tools allows customization of enrollment pages that interact with the LDC People database. A new, centralized database enabling different modes of data collection (CTS, sound booth, SMS, etc.) interoperates with the annotation infrastructure and a new message collection system (Strassel, et al. 2014) that supports both live collection, where the system mediates IM/SMS communication between participants, and a donation method, where archives are uploaded to the server, parsed, and saved to our database. Participants can even review the archive online to redact sensitive information.

A standardized reporting infrastructure allows managers to specify characteristics of reports that programmers extract, adding any custom code necessary. A single cronjob checks every minute for pending reports and executes them serially to reduce computation load.

A new processing pipeline mechanism operates similarly; cronjobs check every minute for data that needs to be processed. Technologies such as Speech Activity Detection have been integrated into this infrastructure. Inputs and outputs are tracked in set implementations to avoid duplication. The system runs constantly.

An extension of the web interface using a JSON API allows clients other than web browsers to access the same functionality. In concert with the LAPPS Grid project (Ide et al., 2014), a JSON-LD implementation improves machine interoperability. In additional we are providing Grid users with many of the functions we already use internally.

Finally, we have developed a waveform widget to support web-based audio segmentation, an extensive library for quality control of corpora prior to release and interfaces to Amazon Mechanical Turk, S3, and local object-based storage.

3. Remaining Challenges

Human Language Technology development has enjoyed rapid gains in performance and coverage over the past decades in large part due to the attention of researchers in numerous related fields and the specific innovations of

[8] http://catalog.ldc.upenn.edu/LDC2013S05

microprocessors and storage, (inter)networking, statistical machine learning, evaluation driven development and common-task research management. Despite these advances, none have yet to yield a promising approach for satisfying LR demand. Even given the concentrated efforts of individual researchers, data centers and funding agencies around the globe more than 99% of the world's approximately 6700 languages remain under-resourced, including languages with many millions of speakers and worldwide economic importance. Even for the best-resourced language, English, we encounter regular requests for LRs that do not yet exist.

One shortcoming of current approaches to LR development and a contributing factor to their overall dearth is the nearly complete reliance on project-oriented collection. Although we derive immense benefit from defining specific technology development goals, developing resources and evaluation strategies and then focusing intense R&D effort to meet those goals, we also miss opportunities for complementary LR development. Within research projects, LR targets are generally moving and timelines are short by design, leading to the need to frequently and rapidly adapt LR development infrastructure and limiting the opportunity to exploit economies of scale. In addition, data collection that is purely driven by the needs of current projects also misses opportunities to collect data that meet future needs, or even current needs for which no funding has been allocated. At LDC, for example, we receive numerous communications from potential data contributors asking to join studies that are closed or offering contributions that do not match current needs. We also receive requests to host collection efforts innovated and previously managed by individual researchers. Most of these efforts address recognized needs even if not funded at the moment of the request. In a world where "*the best data is more data*" and nearly all HLT developers continue to be data-starved, failing to capitalize on the full potential of available and contributed data is counter productive.

4. New Directions

We believe that addressing some of the remaining LR challenges requires a new approach, which combines innovations already proven effective in other projects to form a new collection and annotation platform.

4.1. Ubiquity and Perseverance

Unlike collection and annotation systems that are conceived, developed, implemented and used for specific projects and then switched-off or allowed to lie fallow once project needs are met or funding is depleted, future data collection strategies must be always available to everyone. As in popular social networking sites, participants should be able to register themselves at their own convenience by supplying as little personal information as a contact email, screen name and password. Once authenticated, participants should have access to a wide range of activities, some available to all and some requiring certification or additional demographic information. For example, activities that offer monetary incentives will require whatever information is necessary to transfer funds. By minimizing barriers, this approach maximizes participation and trust.

Available activities might include those familiar to social media users, posting text messages, video, audio and photos, encouraging the posts of others (*liking* them in Facebook parlance), commenting on them and further sharing them (*retweeting* in Twitter parlance). However, such collection platforms would not replace, or even stand entirely apart from, existing and very popular social networking sites. Instead they should include modules that interact with social networking sites to recruit new contributors and to harvest existing data with all necessary permissions as Facebook plug-ins do. WebAnn includes an example of this in its connection to Mechanical Turk. The platform would augment naturally occurring social network interactions with activities that elicit language to support specific research and development activities. Some activities involve interaction with integrate existing technologies for collecting telephone calls, SMS messages and so on. The platform would further differ from typical social networking sites by initiating and encouraging discussion about those linguistic activities providing a natural focus for interaction and attracting contributors committed to similar goals. The LibriVox Forum[9] does this.

An approach based on ubiquity and perseverance requires large volumes of data in many languages to serve as input to the annotation activities. To support annotation on this scale, data centers would need to seed the platform with existing data and add new data as collected. For any given annotation task all appropriate data would be potentially available but prioritized according to current needs. An annotator for such activities would see all segments available for annotation in languages in which the annotator is certified but in the order that best serves current needs.

4.2. Automating Training and Certification

Professional scale collections of linguistic data or judgment typically train potential contributors, use pilot activities to confirm their ability and then assign the target task. Our new direction calls for a training process that scales beyond current methods by recognizing these steps and implementing them in a robust manner. One adds exercises by creating:

1. task definition and data processing routines
2. gold standard corpus and test derived from it
3. 'pitch' to potential contributors
4. online learning material
5. GUI for collecting data and/or annotation
6. scorer comparing contributions against each other and/or gold standard
7. optionally, forum where contributors discuss task, help each other, receive input from a task leader

[9] https://forum.librivox.org/

Potential contributors would review the pitch; those interested would study the learning material and perform pilot annotation based on the gold standard corpus. The scorer would automatically evaluate annotation as each potential contributor finishes the test. Those who perform adequately proceed to production annotation. All contributions, including certification tests and even forum discussions, are available for research use. Many of these components have been used previous LR projects. For the NIST LRE 2009[10] and subsequent evaluations, LDC created a task definition, processing routines, large gold standard corpora and annotation interface. LDC also created material from which the online learning material and pitch could be drawn leaving only the scorer and optionally the forum as new requirements.

The certifications given by these projects could have additional utility. For example, being certified to read, write, speak or understand a specific language may satisfy an employment requirement at organizations that recruit transcriptionists and translators. In addition, many university level linguistic courses require students to learn to transcribe speech for partial course credit.

The components created for a given annotation exercise would become part of a library where they remain available to subsequent efforts. For example the task definition created for language recognition annotation would remain useful in part or entirely even if the target languages changed and the broadcast news transcription scorer could be reused for broadcast conversation.

4.3. Multiple Incentives

The majority of data collection and annotation efforts over the past two decades have relied upon a single incentive type, monetary, to encourage contributions. However, this approach has its limitations especially when funding is in short supply or the target audience is not motivated by purely financial gain. An LDC internal review of some 50 social networking sites has enumerated the following incentives to participation:

1. information
2. entertainment
3. access to services based on contributions
4. sharing intellectual/creative work (self-expression)
5. conveying thoughts & frustrations anonymously
6. payment, discounts (both real-world and virtual)
7. socializing (social networking)
8. competition
9. opportunity to demonstrate competence
10. status, prestige, recognition (levels, high scores)
11. contributing to a greater cause or good

We have seen specific cases of alternative incentives working spectacularly well. The Great Language Game[11] (GLG) asks contributors to listen to short audio clips and indicate what language is spoken. Players are generally not speakers of the target languages. Although created in 2013, GLG has already reported more than 2.6 million games played from which we estimate a bare minimum of 6.4 million decisions. 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.

LibriVox creates "free public domain audiobooks" by recruiting, training and organizing volunteers who record themselves reading literary works. A 2012 survey identified at least 17,500 hours of English readings from which we estimate at least 87,500 hours of volunteer labor at a market value of nearly $9M. Volunteers make such enormous contributions for a variety of reasons. Many believe in the LibriVox mission. Some clearly enjoy collaborating with others of similar interests. A small number of the best readers also receive paid work through Iambik, a spin-off audiobook company.

Within the Games with a Purpose (GWAP) initiative the ESP game, asks players to label images attempting to match labels provided by another, unknown player under time pressure and with scoring. Similar to Google Image Labeler, each of the games sought to improve retrieval of images via services such as Google's image search. After many years of productivity both games are currently unavailable. Google took down image Labeler in 2011. The GWAP developers moved on to other work that same year reporting that some 200,000 players had contributed. Their popularity hints at the power of gamification and other example have emerged: Phrase Detectives[12], Train Robots[13] and OnToGalaxy[14]. Our vision differs from these individual game platforms by providing a sustainable home for language activities, including games, and by building a community of contributors across tasks.

The NSF funded SPICE[15] project at CMU built a web interface that simplified the creation of speech processing components and encouraged contributions from end users without any particular skill in human language technologies. SPICE built partial or complete ASR systems in a number of language for which such technologies were previously absent. SPICE's incentives to contributors included the possibility of downloading a local copy of the ASR engine created through their contributions.

4.4. X-Sourcing

The recent trend toward comprehensive crowd-sourcing has taught LR developers a great deal about data acquisition including lessons that also apply to traditional teams. However, one must recognize that different annotation tasks require different levels of expertise from those of the untrained crowd to the expert. Average native

[10] http://www.itl.nist.gov/iad/mig//tests/lre/2009/
[11] greatlanguagegame.com
[12] https://anawiki.essex.ac.uk/phrasedetectives/
[13] http://www.trainrobots.com/
[14] http://www.kongregate.com/games/phateon/ontogalaxy
[15] http://csl.ira.uka.de/spice/index.php

speakers of the target language, regardless of literacy and level of education are capable of judging whether an audio clip is in their native language, the principle annotation support language recognition technology. Naturally there are difficult cases but those are also difficult for highly educated judges. On the other hand, at least as currently practiced, the best Treebankers, annotators who perform syntactic tagging of text using the phrase-structure formalism, were trained directly by the inventors of the annotation type or by a very small number who were themselves trained by the inventors. Naturally assigning task to under- or over-qualified annotators jeopardizes accuracy and cost respectively. For each annotation type, we thus see three challenges: 1) factoring complex annotations into component tasks according the skills required, 2) identifying the minimal requirements for such tasks 3) effectively providing requisite training, guidance and evaluation. In the Treebanking case sentence segmenting, tokenizing, part-of-speech tagging and syntactic bracketing all require different skill sets. Even within syntactic bracketing, some specific tasks, such as determining the scope of conjunction, require less skill than others. Future collection platforms should acknowledge these challenges by allowing a coordinator to specify the training requirements for a given annotation and then offering tasking only to those certified. It would also allow task prioritization to create a pipeline in which pre-requisite annotation are completed before dependent annotations. Finally, it would include a scheduling engine that prioritized work requiring rare skills over that which requires general skills.

5. Conclusion

Despite the progress sketched above, data centers must revolutionize their approach to collection and annotation if they are to meet R&D needs and create LRs for a wider range of the world's languages. An approach that incorporates the components we have sketched above could scale well beyond current capability. By creating a social networking site open to all, with language related activities and certification available, and participation motivated by multiple incentives, we expect to elicit contributions well beyond what current funding could hope to compensate directly.

Recent technological and managerial innovations at LDC prepare us to develop such a platform. Our vision, begins with an instance of WebAnn, to which social networking activities and interfaces to additional social networking sites have been added. We have already augmented WebAnn with a universal database of more than 12,000 contributors, self-registration, a generalized model of annotation suitable for x-sourcing and tools that simplify GUI creation, workflow management and progress monitoring.

Neverthelesss, there remains considerable work to do. Although it is clear that the initiative we describe here requires multiple incentives, monetary compensation is still the dominant mode. Multiplying incentives requires input both from designers of already successful ventures and from creators of language technologies. Incentives such as those provided by the SPICE program, access to HLTs, are likely to be among the most attractive especially if delivered as web services. In the area of explicit corpus modeling, although we now use a general framework for recording annotations and processes for building corpora from them, we lack a model that describes corpora from the point of view of potential uses, for example HLTs ingesting training data. Finally we need to prove the financial model assuring that the infrastructure is sustainable while reducing LR cost to both creators and end users.

6. Acknowledgements

The Language Application Grid mentioned herein was supported by National Science Foundation grants NSF-ACI 1147944 and NSF-ACI 1147912.

7. References

Bird, S., Liberman, M. (2001) A Formal Framework for Linguistic Annotation, Speech Communication, 33(1,2) pp. 23-60, http://arxiv.org/abs/cs/0010033.

Cieri, C., Strassel, S., Glenn, M., Schwartz, M., Shen, W., Campbell, J. Bridging the Gap between Linguists and Technology Developers: Large-Scale, Sociolinguistic Annotation for Dialect and Speaker Recognition In LREC 2008, Marrakesh, May 28-30.

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., DiPersio, D., Shi, C., Suderman, K., Verhagen, M., Wang, D., and Wright, J. (2014). The Language Application Grid. In LREC 2014, Reykjavik, May 26-31.

Kulick, S., Bies, A., Mott, J., Using Derivation Trees for Treebank Error Detection. ACL-HLT 2011. Portland, Oregon, June 19-24.

Ryant, N., Liberman, M., Yuan, J., Speech Activity Detection on YouTube Using Deep Neural Networks, Interspeech 2013, Lyon, August 25-29.

Ryant, N., Yuan, J., Liberman, M., (undated) Mandarin tone classification without pitch tracking, under review. languagelog.ldc.upenn.edu/myl/ToneWithoutPitch.pd

Schultz, J. M. and Liberman, M., Topic Detection and Tracking using idf-Weighted Cosine Coefficient. In Allan, J. (2000) Topic Detection and Tracking. Dordrecht: Kluwer, pp. 225-239

Wright, J. (2014) RESTful Annotation and Efficient Collaboration. In LREC 2014, Reykjavik, May 26-31.

Yaeger-Dror, M., Cieri, C., Prolegomenon for an Analysis of Dialect Coding Conventions for Data Sharing. In Barysevich, A., D'Arcy, A., Heap, D., eds. (2013) Methods in Dialectology. Lang: pp189-204.

Yuan, J., Liberman, M., Automatic Detection of "g-dropping" in American English Using Forced Alignment. ASRU 2011, Hawaii, December 11-15.

Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V., Wang, W., Automatic phonetic segmentation using boundary models, Interspeech 2013, pp. 2306, 2310.