From ‘Solved Problems’ to New Challenges: A Report on LDC Activities

Christopher Cieri, Mark Liberman, Stephanie Strassel, Denise DiPersio, Jonathan Wright, Andrea Mazzucchi, James Fiumara
Linguistic Data Consortium
3600 Market Street, Philadelphia, PA. 19104 USA
{ccieri, myl, strassel, dipersio, jdwright, amazzu, jfiumara} @ldc.upenn.edu

Abstract

This paper reports on the activities of the Linguistic Data Consortium, the next in a sequence of such data center reports included in each LREC meeting. This report begins by sketching the changing demands for Language Resources driven by the spread of Human Language Technologies throughout the market. One result of the successful deployment of HLT enabled applications is increased demand in ever more languages. This in turn places pressure on data centers to collaborate and form global networks in order to meet the demand for LRs of increasing complexity and linguistic diversity. The report next summarizes the over 100 Language Resources released since the last report, many of which have been contributed by research groups around the world. It also covers advances in Consortium infrastructure that assure the integrity of published data sets and support future collection and annotation. Finally, it discusses recent and current LR creation activities that lead to new LR publications followed by data related research activities particularly in clinical applications.

Keywords: language resources, human language technology, data center

1. Landscape

Each two years between LREC meetings sees significant changes in the landscape of language resources (LRs) and the research and technology development that rely upon them. Perhaps the biggest change is the spread of Human Language Technologies (HLTs) throughout the commercial market. Where HLTs were previously limited to the laboratory and highly constrained external uses, today every smart phone, GPS, web application, and telephone response system can exploit speech-to-text, dialogue management, text-to-speech and translation technologies and many of them do.

Another significant trend is the continuing diversification of languages addressed. While historically many LRs were built for a few dozen languages with the greatest number of speakers and sociopolitical prominence, the past several years have seen increased investment in a wider range of languages often labelled under-resourced. LRs for under-resourced languages tend to form clusters of the resources needed to create speech-to-text technologies or translation technologies thereby reflecting concepts such as the Basic Language Resource Kits (Krauwer, 2003), the REFLEX Language Kits or the earlier CALLHOME corpora.

In addition to the diversification by language, a wider range of research disciplines are embracing the creation use and sharing of LRs and Human Language Technologies. The European CLARIN and PARTHENOS programs now support a vast number of projects in the humanities and social sciences across twenty-one countries with common research infrastructure, especially HLTs, with additional support in the form of training, travel grants, workshops, help-desks and other outreach.

At the same time, clinical groups and the medical field are beginning to recognize the promise of LRs and HLTs not only in the mining and management of vast stores of text and speech data but also in the diagnosis and tracking of disorders and therapies.

These shifts in emphasis should not, and generally are not, seen as proof that our communities have fully solved traditional problems such as multiword expression extraction (Schone & Jurafsky, 2001), language identification from text (da Silva & Lopes, 2006), speech synthesis (Sproat, 2010, p. 206) or speech recognition (Schwartz et al., 2011, p. 399) even in well-studied English. However, in a world with limited resources, they do suggest that sponsors recognize the growing need to address a wider range of new languages and technologies and the growing power of the commercial sector to improve performance in market products.

2. Data Centers as Global Networks

LDC is a consortium of educational, research and technology development groups in the academic, government and commercial sectors joined by a common need for language resources. Not a store but rather a mutual support organization, the Consortium’s members contribute typically in the form of annual fees, though other inputs are possible, and benefit from the many language resources published by the management office hosted at the University of Pennsylvania. The Consortium has always maintained an international focus and its membership has grown to form a global network of LR users and contributors necessary to support the goal of documenting the world’s languages. A snapshot of LDC’s global network of members and contributors appears in *Figure 1*.

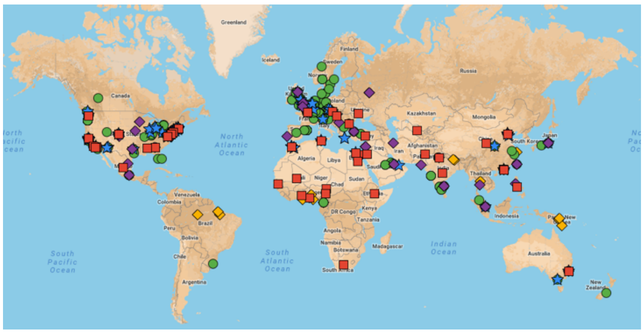

Figure 1: LDC Global Network of select data sources including subcontractors and vendors (red squares), corpus authors (green circles), media providers (purple diamonds), LDC staff collections (gold diamonds), research collaborators (blue stars). Many markers represent multiple collaborators; many markers partially obscured by others.

The business model, established in 1992 by an advisory board of leaders of academic, government and industry research groups, distinguishes the costs of creating a language resource from the costs of distributing copies of that resource to other users. Creation costs are generally covered by the sponsor or customer initially requiring the data set while distribution costs are amortized across recipients. While other models exist, and LDC accommodates sponsors who prefer to vary from this practice by, for example, subsidizing distribution, the original LDC model persists apparently because it is clear, shares the burden of LR cost and is sustainable beyond the duration of a single project or even the career of a specific researcher. As evidence of its continued effectiveness, LDC has distributed more than 140,000 copies of datasets to over 4,000 organizations in more than 80 countries.

3. Publications

The rate of LR publications have continued to increase over the life of the Consortium. During the first decade of its existence, LDC published an average of 23 data sets per year. That average increased to 33 during the second decade and 41 for the past four years. Cumulatively LDC has published 755 data sets and distributed a much larger number of parts and intermediate versions within sponsored programs and closed evaluations. Figure 2 shows the rate of LDC publications since inception including annual, average and cumulative numbers.

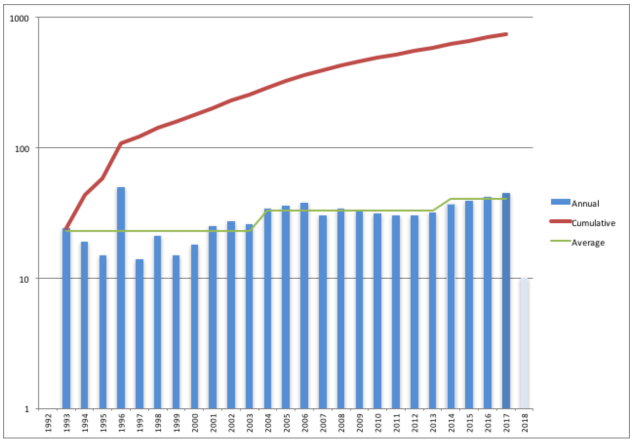

Figure 2: Publication Rate: number of data sets published per year in blue bars, with 10-year average in green and cumulative number of corpora in red. All y-axis values in log scale.

Since 2014, the date of our last report, LDC has released 167 new data sets including 55 corpora that resulted directly from ldc.upenn.edu broadcast conversation with transcripts; Arabic-English and Chinese-English parallel text from broadcast news and broadcast conversation transcripts as well as newswire and web text; word-level alignments of the parallel text and Arabic-English and Chinese-English parallel aligned Treebanks.

Another source of multiple corpora was the IARPA Babel program which produced numerous language packs of which 15 have been released so far via LDC: Assamese, Bengali, Cantonese, Georgian, Haitian, Kurmanji Kurdish, Lao, Pashto, Swahili, Tagalog, Tamil, Tok Pisin, Turkish, Vietnamese and Zulu.

A small sample of additional text publications include: phrase structure Treebanks in Arabic, Chinese, English and Spanish and an Arabic dependency Treebank contributed by the US Army Research Lab; a lexicon of Bambara; informal text (discussion fora, sms, chat) sometimes with translation and word alignment in Arabic and Chinese from the DARPA BOLT program; a range of formal and informal text including translated text annotated for Abstract Meaning Representation; Ancient Chinese text with word-segmentation and part-of-speech tags contributed by Nanjing Normal University; a collection of ~200 annotated samples of student written legal briefs from Georgia Institute of Technology; and a corpus provided by the University of Essex containing >19K words from 40 documents annotated for anaphora by players of the Phrase Detectives "game-with-a-purpose".

Some other examples of speech data include multi-language conversational telephone speech collections in Turkish and three regional clusters: Slavic (Polish, Russian, Ukrainian), South Asian (Bengali, Hindi, Western Panjabi, Tamil, Urdu) and Central Asian (Dari, Persian, Pushto); additional CIEMPISS data from the Universidad Nacional Autónoma de México containing ~18 hours of Mexican Spanish broadcast and transcripts to enable the building of acoustic models for ASR; the KSUEmotions from King Saud University containing five hours of emotional Modern Standard Arabic speech from 23 subjects from Yemen, Saudi Arabia and Syria; Florida Institute of Technology's Noisy TIMIT corpus which reproduces the original TIMIT adding 5-50dB of various noises (white, pink, blue, red, violet and babble); the SRI-FRTIV corpus of ~232 hours of English speech from 34 speakers produced at low, medium and high effort levels in interview, conversation, reading and oration styles to support text-independent speaker verification; and UCLA's collection of audio recording from 9 subjects with time aligned high-speed laryngeal video recordings.

LDC also published a number of data sets to support common task evaluations including the ASpIRE Challenge's data including telephone conversations and speech from far-field microphones in noisy, reverberant rooms; the CHiME Challenge data seeking to develop distant-microphone ASR in real-world environments and the 2010 NIST Speaker Recognition Evaluation Test Set; and the 2007 CoNLL Shared Task data in Arabic, Basque, Catalan, Czech, English, Greek, Hungarian, Italian & Turkish as well as the 2015-2016 Shared Task. The LDC Catalog[1] provides a complete inventory of all publications with descriptions, corpus documentation and samples.

4. LR Creation Projects

Although the Consortium has supported sponsored projects since its founding, LDC's role has expanded from archiving and distributing LRs created by other organizations to collecting and annotating data to create new resources, developing tools and best practices, and

[1] https://catalog.ldc.upenn.edu

increasingly to managing complex data creation teams consisting of multiple partners, subcontractors, vendors and employees with highly specific and complementary roles. Within these teams, LDC crowd-sources or otherwise outsources a growing range of well-defined annotation tasks focusing internal staff on new experimental and high risk collection and annotation paradigms, task definition, quality control and project management.

4.1 DEFT: Deep Exploration and Filtering of Text Program

The DARPA DEFT program which "*aims to address remaining capability gaps related to inference, causal relationships and anomaly detection*" is in its final months and will have ended by the time this paper is published. Over its life-cycle, the program distinguished itself from the earlier practice of extracting isolated information elements from individual sentences in a single language as it evolved toward whole-document, and then whole-corpus (cross-document and cross-lingual) understanding. Over time the program also increasingly focused on extraction of information about events and Sentiment/ Emotive/ Cognitive state (SEC), in addition to entities and relations. To support DEFT, LDC annotated news text and discussion fora for Entities, Relations and Events (ERE), Abstract Meaning Representation (AMR), Textual Entailment and Committed Belief. LRs created for the program are beginning to appear in the LDC Catalog including the DEFT Narrative Text Data which includes ~750,000 words of news text providing the source data for proxy reports based upon single or multiple documents. Proxy reports indicate the date, country and topic of the source text. The body of the report imitates the form of an analyst report.

4.2 Conflicting Accounts of Current Events (CACE)

CACE is new work sponsored by DARPA whose goal is to create an annotated multimedia corpus containing multiple accounts of the same current events in formal and informal data, covering multiple media types and genres including news, blogs, discussion forums, microblogs (Twitter), video, image, speech and other data sources as appropriate. The collected data are centered around a number of topics within a single scenario (e.g. Ukrainian-Russian Relations). Each topic is supported by a topic model that specifies the salient entities and events associated with that topic. A subset of the collected data for each topic is labeled for several features including salient entities as well as salient events and their arguments and attributes (i.e. slots). Additional annotation includes lightweight SEC (Sentiment/ Emotion/ Cognitive State) labeling and/or judgments regarding entailment and contradiction relationships among the various "information elements" that comprise a structured representation of a given event.

4.3 LORELEI

The DARPA Low Resource Languages for Emergent Incidents (LORELEI) program seeks to dramatically advance the state of computational linguistics and human language technology to enable rapid, low-cost development of capabilities for low-resource languages. Acknowledging that even with perfect translation, emergent events such as humanitarian assistance, disaster relief, peacekeeping or infectious disease response generate too much material for analysts to use effectively, LORELEI looks beyond machine translation to provide situational awareness by identifying information elements such as topics, names, events, sentiment and relationships in multilingual sources. LORELEI technology will be applicable to any incident in which a sudden need emerges for assimilation of information about a region of the world where low-resource languages are frequently used in formal and/or informal media. To date technology evaluations have involved Machine Translation, Named Entity Recognition, Entity Discovery and Linking, in which sites are required to link entities to an external knowledge base, and Situation Frame which extracts information from a streaming corpus in the incident language about an emerging situation, to build situational awareness: What's happening, when and where; who is involved; what are the needs and how urgent are they; what are the reactions/responses to the incident among the people involved.

To support the LORELEI Program, LDC is collecting and building a variety of linguistic resources including text in multiple formal and informal genres for several *representative* languages (Hausa, Turkish, Amharic, Arabic, Somali, Farsi, Russian, Spanish, Hungarian, Mandarin, Vietnamese, Yoruba, Tamil, Bengali, Hindi, Indonesian, Tagalog, Thai, Akan (Twi), Swahili, Wolof, Zulu and Uzbek) as well as several low resource *incident* languages whose identity is not disclosed until technology evaluations begin. Large volumes of monolingual and parallel (with English) text are harvested, and smaller volumes of parallel text are acquired via crowdsourcing or traditional translation. Portions of the non-English data that have been translated are labeled for simple named entity, full entity plus coreference, NP chunking, simple semantic annotation lightweight SEC, entity linking and situation frames. Each language includes a 10,000-lemma lexicon and a grammatical sketch, plus basic processing tools including tokenizers, encoding converters, segmenters and entity taggers. Incident languages involve smaller amounts of found data plus (pointers to) formal resources like dictionaries, grammars, gazetteers and primers. The Year 1 surprise language was Uyghur; Year 2 languages were Tigrinya and Oromo.

4.4 Support for NIST Technology Evaluations

In addition to its support for large multi-site technology development programs, LDC also regularly provides LRs to support multiple technology evaluation campaigns organized by the US National Institute of Standards and Technologies.

TAC KBP is the open evaluation series that DEFT performers participate in, along with many non-DEFT sites. As in prior years, LDC has provided data and assessment for all six TAC KBP tracks including the end-to-end Cold Start task, which builds a knowledge base from scratch along with evaluations with EDL, Slot Filling Events and Belief/Sentiment.

NIST has organized an SRE evaluation every year or two since 1996 with evaluations tackling increasingly difficult challenges in channel, room acoustics, language, vocal effort and interaction type. LDC managed data collection in 2015 for the 2016 evaluation under the new CallMyNet protocol. The corpus for the 2016 campaign contains telephone conversations from a total of 220 speakers each of whom completed 10 calls under a variety of noise conditions in one of 4 languages: Tagalog, Cantonese, Mandarin and Cebuano. Calls were manually audited for language, speaker identity and overall quality. LDC has begun collection for the next evaluation, details of which are confidential until the evaluation is complete.

NIST organized the first Language Recognition evaluation in 1996. Since 2003 NIST had held an LRE roughly every other year. LDC has often provided corpora to support LREs including the 2017 evaluation which takes place in September. The latest corpus includes data previously sequestered from the LDC Multi-Language Speech Collection described at a previous LREC (Jones et al., 2016) and containing conversation in two or more linguistic varieties of each of six language clusters: Arabic, Spanish, English, Chinese, Slavic and French. This continues the trend in the NIST LREs to shift from simple language recognition toward the more challenging task of distinguishing highly similar languages and mutually intelligible regional dialects. New data in the 2017 campaign will include speech excerpts extracted from video data.

5. Technical Infrastructure

Although demand for exotic hardware systems has decreased over time, LDC continues to maintain and use its specialized systems for collecting broadcast, telephone calls and messaging as well as its microphone library which addresses a broad range of recording conditions. The past few years have brought increasing demand to miniaturize speech collection systems and deploy them or guide others who deploy them using standard technology in locations around the world.

System infrastructure has evolved to provide ever greater redundancy, data integrity and disaster recovery. Specifically, LDC deploys ZFS as the filesystem of choice to implement continual checks against data corruption. We also distinguish dynamic from static data and internal data from copies stored in the DMZ for public access with appropriate write protection and backup frequency. All storage tiers include offsite storage for disaster recovery.

Our web based annotation infrastructure, WebAnn, first developed in 2011, has since been used to capture tens of millions of annotations. WebAnn is a single application that presents different tools to the user by reusing fixed components, granting much control to the manager of the annotation task. 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. In addition, a new version of the application is under development via the NIEUW project discussed below in section 5.1 that will address key improvements like portability and the ability to run offline.

Our infrastructure has also continued to expand the range of Human Language Technologies we integrate into the data creation pipeline, including those developed at LDC or by partners including Phonexia and Oxford Wave Research. Even well designed command line tools can be difficult to take advantage of in the context of a software group with its own code base and data pipelines. One recent addition has been the incorporation of third party tokenizers into an existing text pipeline which is not trivial as most tokenizers modify the input stream, and our internal processes typically maintain standoff annotation. An even more complex case was taking a recipe for creating a forced aligner for a new language and creating a turn crank procedure.

5.1 NIEUW

For the NIEUW (Novel Incentives and Workflows) project, LDC is building infrastructure and tools to increase the volume and variety of LRs through the use of novel incentives and crowdsourcing. Inspired by the successes of social media, citizen science and games with a purpose, this infrastructure will enable the creation of scalable data collection and annotation activities designed with appropriate incentives and available to the public via the web and mobile devices. We have identified three partially overlapping audiences as the most promising potential contributors: Language Professionals and Students, Citizen Scientists, and Game Players. In order to appeal to these contributor groups and their respective motivations, LDC is creating three portals with varying designs and incentive models. Data collection and annotation activities will be initially created by LDC, however, we will make toolkits available to allow collaborators to create their own activities which can be hosted on the portals.

The infrastructure for NIEUW will build upon LDC's WebAnn described above. By rewriting and enhancing WebAnn, the NIEUW project will create an open source, portable package that supports a wide range of collection and annotation activities with improved ease of use for creating activities, defining workflows, reporting progress, evaluating contributions, and extracting stable data.

6. Research

The Consortium has conducted basic research in addition to, and often integrated with, its data creation activities since inception. Here we distinguish '*research*' activities whose purpose is principally to investigate and report on some linguistic phenomenon but which often produce datasets as a by-product from '*data creation*' activities whose principal goal is to develop language resource which often require research and experimentation in the process. With that distinction in mind, research activities have increased significantly since the last report, particularly in the area of clinical data analysis.

6.1 Exploring the 'Space' of Autism Spectrum Disorders

A collaboration between LDC and the Center for Autism Research (CAR) of the Children's Hospital of Philadelphia began with the goal of identifying linguistic correlates of Autism Spectrum Disorders. CAR had recorded video and audio from many hundreds of diagnostic sessions which included informal conversation and structured linguistic and social activities and for which gold-standard diagnoses were available based on the full range of diagnostic tools. Diagnostic instruments such as the Autism Diagnostic Observation Schedule (ADOS) make reference to multiple linguistic measures that one could imagine extracting from speech and transcripts in an automated fashion for more detailed analysis. Early results showed that relatively straightforward classifiers based on relatively simple linguistic features could predict the diagnoses with good accuracy (Parish-Morris et al., 2017). The corpus used for this research is now undergoing final review before release via LDC.

While encouraging, these results did not truly exercise the power of automated approaches since they relied upon a highly structured diagnostic session requiring several hours of time from both the patient and a highly-trained diagnostician with years of experience in eliciting behavior that distinguished ASD from neuro-typical patients. The real challenge of language technologies would be to achieve the same or better accuracy with shorter informal conversations that could be recorded at home or school via the internet or a telephone call. To support such work CAR is now recording both informal and structured sessions from a stratified sample of ASD and neuro-typical children which will also be released through LDC when complete.

6.2 Linguistic Correlates of Neuro-Degenerative Disorders

Following on promising results with autistic patients, LDC has begun a collaboration with the Fronto-Temporal Disorder laboratory of the Hospital of the University of Pennsylvania. The FTD lab provided digital recordings and transcripts of semi-structured interactions involving 32 patients previously diagnosed with bvFTD, behavioral variant frontotemporal dementia, matched to 17 healthy controls. LDC normalized, QCed and time aligned the transcripts and worked with the FTD lab to develop automatic methods to analyze prosody correlated with clinical evaluations. Specifically, we computed fundamental frequency and log-scale pitch range controlling for individual and sex differences and correlating to neuropsychiatric tests and measures of gray matter atrophy. bvFTD patients had significantly reduced f0 range compared to healthy controls (Nevler et al., 2017) reflecting impaired prosody and supporting the feasibility of automated analysis.

7. Bibliographical References

DARPA (2017) Deep Exploration and Filtering of Text, https://www.darpa.mil/program/deep-exploration-and-filtering-of-text, retrieved October 1 2017.

Jones, Karen Stephanie Strassel, Kevin Walker, David Graff, Jonathan Wright (2016) Multi-language Speech Collection for NIST LRE, Proceedings of LREC.

Krauwer, Steven (2003) The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap, International Workshop on Speech and Computer (SPECOM), October 27-29, Moscow, Russia.

Nevler, Naomi, Sharon, Ash, C. Jester, D. Irwin, Mark Liberman & Murray Grossman (2017) Automatic measurement of prosody in behavioral variant FTD, Neurology: July 19.

Olive, Joseph, Caitlin Christianson and John McCary, eds. (2011) Handbook of Natural Language Processing and Machine Translation. DARPA Global Autonomous Language Exploitation. New York: Springer.

Schone, Patrick and Daniel Jurafsky (2001) Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? Proceedings of Empirical Methods in Natural Language Processing, Pittsburgh, PA.

Schwartz, Richard, Joseph Olive, John McCary and Caitlin Christianson (2011) Machine Translation from Speech in Olive, Joseph, Caitlin Christianson and John McCary, eds. (2011).

Da Silva, Ferreira, J. & Pereira Lopes, G. (2006). Identification of Document Language is Not yet a Completely Solved Problem, in Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation.

Parish-Morris, Julia, Mark Y. Liberman, Christopher Cieri, John D. Herrington, Benjamin E. Yerys, Leila Bateman, Joseph Donaher, Emily Ferguson, Juhi Pandey and Robert T. Schultz (2017) Linguistic camouflage in girls with autism spectrum disorder. Molecular Autism: Brain, Cognition and Behavior 2017 8:48/

Sproat, Richard (2010) Language, Technology, and Society, Oxford, UK: Oxford University Press.