Twenty Years of Language Resource Development and Distribution: A Progress Report on LDC Activities

Christopher Cieri, Marian Reed, Denise DiPersio, Mark Liberman
University of Pennsylvania, Linguistic Data Consortium
3600 Market Street, Suite 810, Philadelphia PA. 19104, USA
E-mail: {ccieri, mreed, dipersio, myl} AT ldc.upenn.edu

Abstract

On the Linguistic Data Consortium's (LDC) 20th anniversary, this paper describes the changes to the language resource landscape over the past two decades, how LDC has adjusted its practice to adapt to them and how the business model continues to grow. Specifically, we will discuss LDC's evolving roles and changes in the sizes and types of LDC language resources (LR) as well as the data they include and the annotations of that data. We will also discuss adaptations of the LDC business model and the sponsored projects it supports.

Keywords: language resource; data center; common task program

1. Introduction

The year 2012 marks the 20th anniversary of the Linguistic Data Consortium. During the past two decades, LDC has observed, adapted to and in some cases even anticipated changes in the needs and methods of the communities it supports. The changes have frequently but not always been in the direction of demands for greater volumes of data in an increasing array of languages with richer annotation and higher accuracy and reliability. Since these desiderata are in conflict with each other and the desire for shorter timeline and greater cost-efficiency, specific projects have made different trade-offs. As some HLT system performance approaches human performance we tend to see increased focus on quality at the expense of volume. However, we also continue to see growth in supply and demand of megascale data sets like the Google n-gram corpora (Franz and Brants 2006). Compare, for example Novotney, Scott and Callison-Burch (2010) arguing for lower cost, higher volume despite lower interannotator agreement with Maamouri, Bies and Kulick (2010) who show parseval scores for Arabic parsers improving with increases in interannotator agreement in the training material.

As sub-disciplines begin to explore data sharing, some researchers engage in work that straddles traditional boundaries (Yaeger-Dror 2002, Clopper & Pisoni 2006) creating a split in their communities between those who can exploit large-scale digital data and those who cannot or will not. The past few years have seen the emergence of initiatives, such as CLARIN (2011) and the numerous projects under its umbrella, that seek to bring large scale computing and data exploitation to the social sciences and humanities.

Finally, the worldwide penetration of computing, social networking and smart phones increases demand for resources in a rapidly growing list of languages.

2. Evolution of LDC Roles

Over the past two decades LDC's role adapted many times to community need. In the early nineties, DARPA communicated a need for a consortium to address the data distribution and archiving needs of HLT developers through a call for proposals. The LDC emerged from submission to this call written by Mark Liberman at the University of Pennsylvania, which became the LDC host institution. LDC's early roles were limited to those of a specialized data publisher and archive guaranteeing widespread, long-term availability. In 1995, after recognizing that the existing U.S. labs doing data collection could not keep up with demand, LDC began collecting conversational telephone speech and broadcast news and doing some transcription. Today, LDC's data collections include news text, (including blogs and newsgroups), biomedical and other text documents (both printed and handwritten), broadcast news and conversation, telephone speech and other spoken interactions including lectures, meetings, interviews, map task and role playing games, read and prompted speech, web video and even animal vocalizations.

By 1998, LDC anticipated another growth in demand and began to undertake annotation projects as well. Cieri who had just joined the staff, was tasked with developing the annotation operation. Early annotation tasks included segmentation of news broadcasts into stories and topic relevance judgments. Today annotation activities include data scouting, selection and triage; various alignments (audio to audio, audio to text, source text to translation); bandwidth and signal quality judgments, identification of language, dialect and speaker, segmentation at program turn, sentence and word boundaries; orthographic and phonetic transcription at different levels of detail, script normalization; annotation of phonetic, dialect,

sociolinguistic feature and supralexical features; document zoning and legibility annotation; tokenization; morphology, part-of-speech, gloss, syntactic, semantic and discourse function tagging, disfluency annotation; relevance judgment; sense disambiguation; readability judgments; entities, relations, events and co-reference tagging; knowledgebase population; time and location tagging; many kinds of summarization; translation, edit distance analysis, post-editing, quality control; analysis of the physics of gesture; video entity and event identification and classification; and pronunciation, morphological, translation and usage dictionaries development.

In 1999, with the arrival of Steven Bird to LDC the organization began develop tools and best practices not just as a means to complete specific project but for general use (Bird, Liberman 1999). The Annotation Graph Toolkit was one of the first products of this effort.

In around 2000, LDC's contribution to common task technology development programs began to grow from simply providing specific corpora to overall language resource coordination across the program. This evolution began during the DARPA TIDES and EARS programs and was fully realized during the DARPA GALE and subsequent programs under Stephanie Strassel's management.

In 2005, LDC began publishing the specifications it developed to guide the creation of new language resources. See for example LDC (2011).

During the GALE program, LDC also realized a long-term goal of integrating HLTs into the LR creation workflow. Specifically, LDC uses the speech-to-text and machine translation systems of multiple GALE performers to intelligently select new and challenging data for annotation (Walker, Caruso, DiPersio 2010). Naturally, such work must proceed with caution to avoid biasing the collection in favor of any specific engine or research group.

This evolution has lead to a situation today in which LDC activities include production, validation archiving and distribution of language resources, management of intellectual property rights and licenses, data collection and annotation and lexicon building, tool, specification and best practice development, documentation and metadata hosting, consulting and training, corpus creation research and academic publication, resource coordination in large multisite programs and service on funding panels, program and oversight committees.

3. Evolution of the Business Model

The LDC business model was designed by an oversight committee composed of commercial, government and non-profit LR users. That group recognized the need to create a business model that would sustain the Consortium's operations independently of government support beyond seed funds and despite rises and falls in the availability of funds for new data development. As a result, LDC provides vast amounts of data at no additional costs to members who support the Consortium typically through membership fees. Specifically, LDC typically publishes around 33 corpora per year. The cost to create any of these corpora is greater that the annual membership fee, in many case one to three orders of magnitude greater. These cost savings are possible because LDC membership as a whole only supports the distribution operation. The cost of creating new LRs is borne by the programs that need such resources.

After a decade of operations, LDC recognized a trend among members that suggested a change to the business model. A significant number of LDC members request all the data sets released. However, the original membership model and fee was based on the expectation that members will acquire a subset of released corpora each year. To adapt to this trend, LDC split the membership into two types: one limiting the number of corpora a member can acquire to 16 annually, the other, with a somewhat higher fee, that includes automatic delivery of two copies of every corpus released.

The centralization of distribution services reduces the boundaries to accessing LR while maintaining uniform licensing within and across research communities. The transparent cost sharing means that funding agencies who cover most or all LR development costs are relieved of the burden of subsequent maintenance and distribution while research users gain access to vast amounts of data. A single membership agreement imposes minimal, standard terms over nearly all of the 520 LRs in the LDC Catalog. Members retain ongoing data access to data via a consistent interface that groups original data with any subsequent patches. This approach encourages sharing, re-annotation and use and the comparison of competing analysis or algorithms over benchmark data.

Benchmarks of LDC's contribution of the research communities it supports are the volume and diversity of LRs distributed, over 84000 copies of more than 1300 titles to 3100 organizations in 70 countries. In addition LDC has identified some 8000 academic papers that reference LDC data after having searched for just 55% of all titles in the Catalog (Ahtaridis, Cieri, DiPersion, 2012).

4. Evolution in Publication Practice

In addition to the changes in the business model described above, LDC has implement several changes in publications practice. The publication operation has always accepted data sets from contributors, reviewed these for accuracy, consistency, adherence to some specification and adequacy of documentation. However, once LDC began creating corpora for using in federally funded common-task programs it became necessary to adapt to this new stream of data. Regarding the division of labor among LDC employees who support projects and those who maintain the publication operation, the former are responsible for collecting, annotating, and preparing a single publication ready copy of each corpus they create while the latter are responsible for an

additional round of quality control, production replication, distribution and archiving

Another specific change in publication practice has been in volume. In 2003, LDC noticed that the backlog of data sets waiting for publication had begun to grow. To reduce the backlog and provide greater value, LDC decided to increase publication productivity. Since 2004, LDC has targeted a range of 30-36 publications per year. Figure 1 shows the number of publications per year as columns and cumulatively as a line adding to 520 in total. Since 2004, LDC has released on average 33 per year representing a 43% increase over the average for all prior years, 23. Productivity has also remained more consistent as seen in the post-2004 standard deviation of 2.68 versus 3.16 for prior years. The broken, horizontal "average" lines emphasize relative consistency since 2004 compared to the previous decade.

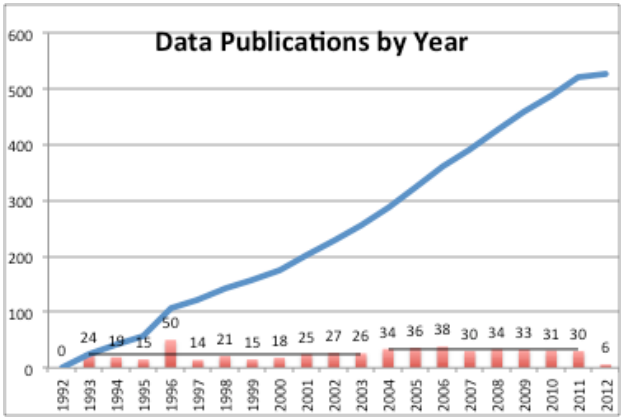

Figure 1: LDC Data Publication by Year

5. Evolution in Outreach

LDC has always followed the practice of finding a way to make data available to credentialed researchers with a bona fide need for data and a genuine inability to pay. In some cases this meant making the researcher aware of an existing affiliation between their organization and LDC. In other cases it has meant correcting a misunderstanding about the fee for a resource, discounting or delaying the fee if necessary or arranging some kind of trade of data or services. In every case, LDC balances the needs of the individual researchers with the needs of the Consortium included assuring its longevity.

In 2010, LDC decided to formalize this procedure to assure that the funds the Consortium invested in supporting worthy but impecunious researchers were distributed wisely. Grants in data are given twice each year to students who demonstrate a need by submitting a two-page proposal summary and a letter of support from their advisors. Because regular members subsidize this activity through their fees, it is important to apply a rigorous procedure. The proposal summary must instill confidence in the research and make clear how and why LDC data will be used. The letter must express the advisor's confidence in the project and assert an inability to pay. Applications are competitive. Incomplete applications are rejected out of hand and not every complete application succeeds. In some cases, LDC staff contact successful applicants to provide additional information about offerings and suggest alternative data sets, in particular when the applicant has not selected data that will allow the work to be properly evaluated. To date, LDC has provided, at no cost to applicants, 8 corpora in 2010, 24 in 2011 and 8 in 2012. Winning applications have run the gamut of language related disciplines from computer science and electrical engineering to oriental studies, second language acquisition and teaching. Had these data been license at normal fees the total would have exceeded 40,000 USD.

From its beginning LDC has always worked to keep the communities it supports informed of its work and research products by attending and presenting its research products at conference and workshops. Since 2007, LDC has expanded its conference presence to include vendor tables, exhibiting at one event in the first year, two in 2008, three in 2009, two in 2010, five 2011 and two at the time of writing in 2012. Still in the process of determining which conferences allow one to reach the broadest audience given the cost, LDC has exhibited at: NWAV (New Ways of Analyzing Variation), Interspeech, ALA (American Library Association), LREC, ACL (Associate for Computation Linguistics), LSA (Linguistic Society of America), ICASSP and NEALLT (Northeast Association for Language Learning Technology).

LDC also conducts surveys occasionally to elicit feedback from it user communities. The first two surveys (2006, 2007) were designed principally to establish a user sentiment baseline. The 2012 survey probed sentiment as well, but it also addressed topics of interest to the Consortium including respondents' "favorite" LRs, LDC's 20th Anniversary (April 2012), mobile technologies, "open data" and social networking.

The 2012 survey was sent by email on January 24, 2012 to 1541 contacts from two groups: those who joined LDC or licensed data from 2009 through 2011; and LDC's primary contact at the surveyed organizations. A reminder email was sent on February 7, and the survey closed on March 2. 99 respondents completed the survey (6.42% of the sample); 169 provided partial responses (10.97%). A 6.42% response rate approximates a 95% +/- 15% degree of accuracy for the sample group[1]. A 10-20% response rate is standard for web-based surveys[2].

Although the results are still under analysis as of this writing, we can note some general points. Respondents indicated overall a high satisfaction level with LDC's language resources (LRs) and services; this confirmed sentiments expressed in 2006 and 2007. Respondents' favorite LRs tended to cluster around "benchmark" data sets, i.e. TIMIT and Penn Treebank, large text sets, such as the multilingual Gigaword corpora, and evaluation data sets, e.g., NIST's Speaker Recognition Evaluation (SRE) releases. When asked to comment on open data, most respondents agreed that the community needs more open data, though

[1] http://www.greatbrook.com/survey_accuracy.pdf
[2] http://constantcontact.custhelp.com/app/answers/detail/a_id/2965/~/predicting-survey-response-rates

responses varied on how this is to be achieved. Respondents also expressed a need for multilingual annotated corpora, particularly in under-resourced languages. The group also noted that while they are becoming more accustomed to using social media networks to gather information on human language technology (HLT) and language-related developments, attending conferences and visiting organizational websites remain their primary information gathering sources.

Several respondents stated that LDC and its services are important -- and in many cases, necessary -- for their work, indicating that LDC's founding principles were sound and continue to be relevant. Nevertheless, some responses reflected a degree of confusion or uncertainty about LDC's membership and data license terms. This suggests the need for continued attention to community education and outreach.

6. Recent and Current Projects

In addition to acting as an archive and specialized publisher of language resources, LDC has undertaken collection, annotation and distribution projects since 1995 with funds coming from the U.S. Departments of Commerce, Defence, Education, Homeland Security, the Interior and Justice as well as the National Science Foundation, other non-profits and several commercial organizations.

LDC tasks in support of multisite HLT programs, most frequently DARPA programs, include: needs assessment, specification, planning, project management and creation of data matrices; intellectual property rights management; human subject coordination; tool development; data scouting, collection and triage; annotation, analysis of inter-annotator agreement and quality assurance; outsourcing; data distribution; sourcing and acquisition of existing data, data sharing across programs and management of reserved data such as evaluation and progress sets. A summary of some projects follows.

TDT (Topic Detection and Tracking) contributed to the development of translingual news understanding systems by creating technologies that could segment news broadcasts into individual stories and detect new stories or stories related to a specific topic. LDC provided audio transcripts, story segmentation and topic relevance annotation.

TIDES (Translingual Information Detection, Extraction and Summarization) further developed news understanding systems by creating translingual technologies to perform detection, extraction and summarization of audio and text. LDC provided audio transcripts, story segmentation and topic relevance annotation.

EARS (Effective Affordable Reusable Speech to Text) developed high quality systems that transformed speech to text in multiple languages and channels. LDC provided multilingual broadcast and conversational telephone speech with time-aligned transcripts and annotations of syntactic-semantic units, disfluency and repair that facilitated the downstream processing of transcripts and the conversion into human readable form.

GALE developed technologies to transcribe, translate and distill speech and text in multiple languages into structured, information. LDC provided data, annotations, tools and specifications to the program.

BOLT (Broad Operational Language Translation) is a new program that will create technology to translate text in multiple foreign languages and all genres, search the translations and media including bilingual spoken and written communication. LDC will provide multi-genre text, translation and multiple layers of annotations to BOLT developers.

RATS (Robust Automatic Transcription of Speech) is building systems that perform speech activity detection, language and speaker detection and keyword spotting in very noisy speech. LDC is supporting RATS by providing clean and noisy speech in multiple languages, segmenting spoken regions, transcribing and annotating for language and speaker, keywords.

MR (Machine Reading) extracted knowledge from natural language text for automated processing with little human intervention. LDC supports MR with annotated data, new annotation guidelines and tools, use cases and system assessment.

MADCAT (Multilingual Automatic Document Classification Analysis and Translation) is building systems segment documents into graphic, printed and handwritten zones, extract metadata, perform OCR and translate the resulting text into English. LDC is providing large-scale language resources based on new collection and annotation of new and existing data that divides pages into handwritten and printed zones, lines of text, words and characters and then provides transcription and reading order for the handwritten words.

In addition to DARPA, the US National Science Foundation was one of the earliest supporters of the Consortium and has typically provided funding for basic research and infrastructure creation projects. The earliest NSF funding supported the creation of LDC-Online, a service that provided Internet based access to LDC corpora. The second version, locally funded provided access to broad data types such as English, Arabic and Chinese news text and English conversations regardless of the corpora in which these data were or would be included. *Talkbank* created communities of practice in a dozen disciplines that work with primary linguistic data and also sponsored the contributed to the distribution of the American National Corpus (Reppen, et. al., 2005) and Santa Barbara Corpus of Spoken American English. NetDC explored harmonization of LDC and ELRA practice by jointly producing the TED corpus of non-native academic speech with transcripts and an Arabic Broadcast News corpus. More recent work for NSF has included biomedical information extraction and phonetic analysis.

With funding from the Department of Education, LDC has created a reading facilitation and evaluation tool and applied it to two morphologically complex but very different languages Arabic and Nahuatl. The tools use morphological analysis to mediate access to digital dictionaries for language learners. Working with Georgetown University Press and US Department of Education support, LDC is now creating Iraqi, Levantine and Moroccan Arabic dialect lexicons that will be shared as print dictionaries and lexical databases.

In addition to contributing to common task programs for numerous sponsors, LDC has contributed data to many technology evaluations organized by NIST. Some of these corpus-building efforts were funded directly by NIST and some by other sponsors for whom NIST conducted the evaluations. Of course, numerous other organizations including MITRE (Mani, et. al., 2009), APPEN (Schlenoff et. al., 2009) and the Universidad Politecnica de Madrid (Ortega-Garcia, et al. (1998) have also provided data to NIST for these and other evaluations. LDC has also provided data for European technology programs including TC-Star and MEDAR and secured for American programs like LCTL (Less Commonly Taught Languages) access to data created in Europe and previously only available via ELDA, for example the EMILLE corpus.

	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11
LRE	✓							✓		✓		✓		✓		✓
SRE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	
BN Re	✓	✓	✓	✓												
CTS Re		✓	✓		✓	✓										
SDR			✓	✓	✓											
TDT			✓	✓	✓	✓	✓	✓	✓							
ACE					✓	✓	✓	✓	✓	✓		✓	✓			
RT							✓	✓	✓	✓	✓	✓		✓		
STD											✓					
GALE Tr											✓	✓	✓	✓	✓	✓
OpenMT						✓	✓	✓	✓	✓	✓		✓	✓		
MetricsMaTr													✓		✓	
MADCAT													✓	✓	✓	✓
TAC KBP														✓	✓	✓
TRECVid SED													✓	✓	✓	✓
TRECVid MED															✓	✓
DUC						✓	✓	✓	✓	✓	✓	✓				

Figure 2: LDC Data in NIST Evaluations

Abbreviation Key: LRE= Language Recognition, SRE= Speaker Recognition, BN Re= Broadcast News Recognition (HUB-4), CTS Re= Conversational Telephone Recognition (HUB-5), SDR= Spoken Document Retrieval, TDT= Topic Detection and Tracking, ACE= Automatic Content Extraction, RT= Rich Transcription, STD= Spoken Term Detection, OpenMT= Open Machine Translation, MetricsMaTr= Metrics for Machine Translation, KBP= KnowledgeBase Population, SED= Surveillance Event Detection, MED= Multimedia Event Detection, DUC= Document Understanding Conference

7. Conclusion and Future Plans

This paper has sketched the evolution in selected LDC activities over its 20 year history that respond to community demands for more and larger language resources in a growing number of languages annotated with greater sophisticated. The current activities stand in stark contrast to the original ones. The Consortium plans to maintain its leadership role and continue to create and distribute LRs while reaching into new languages genres and user communities. We will also continue integrating HLTs into our workflow, increasing research activities and outreach to students, simplifying production through efficiency and outsourcing and to expanding our work in tools and specification development.

8. References

Ahtaridis, Eleftheria, Christopher Cieri, Denise DiPersio (2012) LDC Language Resource Database: Building a Bibliographic Database in LREC 12: Eighth International Conference on Language Resources and Evaluation, Istanbul, May 21-27, 2012.

Bird, Steven, Mark Liberman (1999) A Formal Framework for Linguistic Annotation, Tech Report MS-CIS-99-01, University of Pennsylvania, Department of Computer and Information Science, arXiv:cs/9903003v1.

Clopper, Cynthia, David Pisoni (2006) The Nationwide Speech Project: A new corpus of American English Dialects, Speech Communication v. 48, pp. 633–644.

CLARIN (2011) CLARIN web site http://www.clarin.eu.

Franz, Alex, Thorsten Brants (2006) All Our N-gram are Belong to You, Google Research Blog, Thursday, August 03, 2006 at 8/03/2006 11:26:00 AM, http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html.

John S. Garofolo, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. CD-ROM. Philadelphia: Linguistic Data Consortium, 1993.

LDC (2011) GALE Project Task Specification and Annotation Guidelines web page, http://projects.ldc.upenn.edu/gale/task_specifications/.

LDC (2009) Linguistic Data Consortium Catalog, http://www.ldc.upenn.edu/Catalog.

Maamouri, Mohamed, Ann Bies, Seth Kulick (2010) Upgrading and Enhancing the Penn Arabic Treebank: A GALE Challenge in Olive, Christianson, McCary, 2010.

Mani, Inderjeet, Christy Doran, Dave Harris, Janet Hitzeman, Rob Quimby, Justin Richer, Ben Wellner, Scott Mardis, and Seamus Clancy (2009) SpatialML: Annotation Scheme, Resources and Evaluation, MITRE Technical Reports 09_3827, The MITRE Corporation, The MITRE Corporation.

Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz and Ann Taylor. Treebank-3 LDC99T42. CD-ROM. Philadelphia: Linguistic Data Consortium, 1999.

Novotney, Scott, Chris Callison-Burch, (2010) Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 207–215, Los Angeles, California, June 2010.

Olive, Joseph, Caitlin Christianson, John McCary (2010) Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, New York: Springer.

Ortega-Garcia, J. et al. (1998) AHUMADA: A Large Speech Corpus in Spanish for Speaker Identification and Verification, in Proceedings of ICASSP '98, Vol.

II, pp. 773-776.

Randi Reppen, Nancy Ide, and Keith Suderman (2005) American National Corpus (ANC) Second Release, Linguistic Data Consortium, Philadelphia, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T35

Schlenoff, C. I., Weiss, B. A., Steves, M. P., Sanders, G. A., Proctor, F. M., Virts, A. M. (2009) Evaluating Speech Translation Systems: Applying SCORE to TRANSTAC Technologies, Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Gaithersburg, MD, September 21-23, 2009.

Walker, Kevin, Christopher Caruso, Denise DiPersio (2010) Large Scale Multilingual Broadcast Data Collection to Support Machine Translation and Distillation Technology Development in Olive, Christianson, McCary, 2010.

Yaeger-Dror, Malcah, Lauren Hall-Lew, Sharon Deckert, (2002), It's not or isn't it? Using large corpora to determine the influences on contraction strategies, Language Variation and Change, v. 14, pp. 79–118.