Adapting to Trends in Language Resource Development: A Progress Report on LDC Activities

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

Abstract

This paper describes changing needs among the communities that exploit language resources and recent LDC activities and publications that support those needs by providing greater volumes of data and associated resources in a growing inventory of languages with ever more sophisticated annotation. Specifically, it covers the evolving role of data centers with specific emphasis on the LDC, the publications released by the LDC in the two years since our last report and the sponsored research programs that provide LRs initially to participants in those programs but eventually to the larger HLT research communities and beyond.

1. Introduction

The language resource landscape over the past two years may be characterized by what has changed and what has remained constant. Constant are the growing demands for an ever-increasing body of language resources in a growing number of human languages with increasingly sophisticated annotation to satisfy an ever-expanding list of user communities. Changing are the relative importance of size, complexity, richness of annotation, multilinguality and multimodality in language resources. While developments in computing infrastructure have put greater power at users' fingertips and advances in human language technologies (HLT) have enabled even solitary researchers to produce corpora of great richness, the demand for data centers continues to grow and mutate in type.

In some HLT subdisciplines, the inexorable march of system performance toward milestones defined by human performance has increased demand for quality sometimes even at the expense of volume. Elsewhere however, we see growth in both the supply of and demand for corpora of a size inconceivable just a few years ago, for example the Google n-gram corpora (Franz and Brants 2006).

As new research communities adopt or consider corpus based methods, their most advanced practitioners engage in research that blurs the old traditional boundaries (Yaeger-Dror 2002, Clopper & Pisoni 2006) while others await the existence of adaptive access to existing data and flexible standards.

Finally, the spread of computing around the world increases the variety of languages represented on the Internet and consequently the demand for technologies to process them, demand that is far outstripping the emergence of language resources in these languages.

2. The Role of Data Centers

As the demand for LRs has grown data centers have seen a growing range of opportunities to contribute. Today the largest HLT data centers are simultaneously: specialized publishers, archives, consultants, project managers and creators of tools, specifications and best practices as well as databases.

At least in the case of LDC, the role of the data center has grown organically, generally in response to stated need but occasionally in anticipation of it. LDC's role was initially that of a specialized publisher of LRs with the additional responsibility to archive and protect published resources to guarantee their availability of the long term. Its expansion into data collection and annotation activities in 1995 and 1998 respectively responded to specific needs that were clear at the time. However other expansions of the LDC role including entity annotation and treebanking, tool and infrastructure development in 1999, the overall coordination of data creation activities across multisite sponsored projects in 2000 and the integration of HLTs into the data creation process in 2005 were based upon local initiative or, in the latter case, the recognition that efficiencies enjoyed within other projects would have broader benefit if they could be implemented more centrally with data centers. The creation of a new type of membership, the subscription membership, resulted from the recognition that a significant number of data users request all data sets produced by LDC. The original membership was based on the belief that most members who need only a few corpora each year. It is also equally that ELRA's recent work on technology evaluation and metadata consolidation responds to needs they perceive among HLT researchers (Mapelli, et. al. 2008).

Today, LDC's routine activities include language resource production including quality control, archiving and distribution, intellectual property rights and license management, data collection and annotation and lexicon building, the creation of tools, specifications, best practices, documentation and metadata, knowledge transfer via consulting and training, corpus creation research and academic publication, resource coordination in large multisite programs and serving multiple research communities as funding panelists, workshop participants and oversight committee members.

LDC data collection activities range across news text, blogs, zines, newsgroups, biomedical text and abstracts, printed, handwritten and hybrid documents,

broadcast news, broadcast conversation, conversational telephone speech, lectures, meetings, interviews, read and prompted speech, role playing, web video and animal vocalizations. Annotations and related activities include: data scouting, selection, triage; audio-audio alignment; bandwidth, signal quality, language, dialect, program and speaker ID, time-alignment and segmentation at the turn, sentence and word level; quick and careful transcription; orthographic and phonetic script normalization; phonetic, dialect, sociolinguistic feature and supralexical annotation; document zoning and legibility annotation; tokenization and tagging of morphology, part-of-speech, gloss, syntactic structure and semantic relations; discourse function, disfluency and relevance annotation; sense disambiguation; readability judgments; identification and classification of mentions in text of entities, relations, events and co-reference; knowledgebase population; time and location tagging; summarization of various lengths from 200 words down to titles; translation, multiple translation, edit distance analysis, translation post-editing, translation quality control and alignment of translated text at document, sentence and word levels; analysis of the physics of gesture; identification and classification of entities and events in video; and the development of pronunciation, morphological, translation and usage dictionaries.

LDC support of sponsored, multisite, HLT evaluation programs includes: needs assessment and the subsequent creation of statements of work, budgets, time lines and data matrices; the management of intellectual property rights and human subjects including the development of efficient human subject protocols; corpus specification and validation; custom tool development; data scouting, collection and triage; annotation, inter-annotator agreement studies and quality assurance; outsourcing and remote annotation; data distribution including segmentation into training, development and test sets, sourcing and acquisition of existing data, cross program data sharing and management of limited rights or reserved data including evaluation and progress sets.

The benefits of centralizing such services are many and include broad knowledge and distribution of LRs with uniform licensing both within and across research communities. Costs are shared transparently so that funding agencies who cover most or all LR development costs are relieved of the burden of subsequent maintenance and distribution costs while research users gain access to vast amounts of data – typically 30 corpora per year – for an annual membership fee that is from one to three orders of magnitude less than the cost of any single LR. In addition, the LRs are permanently accessible under simple, standard terms of use. Members' access to data is ongoing and any patches are distributed via the same methods as the original corpus. This encourages reuse and benchmarking of new algorithms. Finally, membership and data licenses cross-subsidize the creation of tools, specifications and papers that are distributed without fee.

3. Current Publications

One area in which LDC adapts to community demands is in its publications. So far in 2008, 2009 and 2010, LDC has added 63 titles to its catalog and produced dozens of corpora for use in evaluation programs that will be released generally after they use in the relevant communities. The publications of the past two years include continuing contributions from and for the research communities in language and speaker recognition, speech to text, information retrieval and extraction, a very large increase in corpora supporting machine translation and a vast number of resources for NLP. Source material has expanded from newswire and broadcast new to include web text and broadcast conversation. Connections to the avant-garde in linguistics and areas studies also continue to grow. A sampling of the publicly available corpora released since our last report follows.

To support Language Recognition, LDC released the 2005 and 2007 NIST Language Recognition Evaluation Test Sets (LDC2008S05, LDC2009S04). For Speaker Recognition, LDC published CHAINS, a corpus of read speech in multiple styles across multiple sessions by 33 English speaker developed at the University College Dublin (LDC2008S09).

LDC released the following to support speech recognition and speaker diarization:

- 33 hours of Czech Broadcast Conversation and transcripts (LDC2009S02, LDC2009T20) from the University of West Bohemia
- MDE annotations, also from West Bohemia, of the 26 hours of transcribed Czech broadcast speech contained in prior LDC corpora; these MDE annotations make corpora more readable and more useful to downstream processing by marking filled pauses, discourse markers, disfluent speech and natural breakpoints in the flow of speech. (LDC2010T02)
- 163 hours of transcribed, telephone speech from 136 native Caribbean Spanish and non-Caribbean Spanish speakers (LDC2010S01, LDC2010T04) from the Fisher collection
- re-tokenized, POS tagged, XML formatted version of the CALLHOME Mandarin Chinese Transcripts provided by Lancaster Univeristy (LDC2008T17)
- corpus of read Brazilian Portuguese Speech provided by the U.S. Military Academy's Department of Foreign Languages and Center for Technology Enhanced Language Learning (LDC2008S04)
- a wideband mobile telephony derivative of the TIMIT Acoustic-Phonetic Continuous Speech Corpus contributed by Patrick Bauer and Tim Fingscheidt (LDC2010S02)

To support machine translation, LDC published:

- NIST MetricsMATR08 Development Data

(LDC2009T05)

- NIST Open Machine Translation 2008 Evaluation Selected Reference and System Translations (LDC2010T01)
- 22.5k words of each of English, Chinese and Arabic data translated into each of the other two languages and annotated for entities and TIMEX2 extents and normalization (LDC2009T11)

and with DARPA GALE funding contributed:

- 550K words of Arabic newswire translated into English (LDC2009T22)
- 323K words of Arabic weblogs and newsgroups text translated into English (LDC2009T03, LDC2009T09)
- 223K characters of Chinese newsgroup text and its translation selected from twenty-one sources and aligned (LDC2010T03)
- 10.7 hours comprising 56,174 words of Arabic broadcast news, transcription and translation (LDC2008T09)
- 553K characters of Chinese blog and newsgroup text and translation into English (LDC2008T06, LDC2009T15)
- 41 hours of Chinese broadcast news transcribed to yield 539K characters and translated into English (LDC2008T08, LDC2008T18)
- 44.4 hours of Chinese broadcast conversation transcribed to yield more than 556K characters and translated into English (LDC2009T02, LDC2009T06)

Researchers in information retrieval and extraction including from biomedical text have access to the following new data

- subset of the English Gigaword corpus, XML tagged for use in the AQUAINT program (LDC2008T25)
- 601K words of PubMed abstracts on molecular genetics and cytochrome P450 inhibition variously tokenized, annotated for part of speech and multiple biomedical named entities and partially annotated for syntactic structure (LDC2008T20, LDC2008T21). The corpora were prepared by LDC for the Institute for Research in Cognitive Science with NSF funding (ITR EIA-0205448) in collaboration with GlaxoSmithKline Pharmaceuticals R&D.
- New York Times Annotated Corpus (LDC2008T19) contributed by Evan Sandhaus containing 1.8 million articles, subsets of which have been manually summarized, entity and topic tagged

Continuing its support for language modelling and responding to demand for greater volumes of text and n-gram corpora, LDC has re-released its popular North American News Text corpus (LDC2008T15), including a version for non-members (LDC2008T16) and published new versions of the Chinese (LDC2009T27), English (LDC2009T13) and Spanish (LDC2009T21) Gigawords as well as three important new contributions from Google:

- word n-grams (n=1,7) and frequencies from 255 billion tokens of Japanese web text (LDC2009T08)
- word n-grams (n=1,5) and frequencies from 100 billion tokens of each of ten European languages: Czech, Dutch, French, German, Italian, Polish, Portuguese, Romanian, Spanish and Swedish (LDC2009T25)
- character n-grams (n=1,5) and frequencies from 883 billion tokens of Chinese text

The Center for Spoken Language Understanding (CSLU) at the Oregon Health and Science University contributed four corpora of read speech from more than ten thousand speakers in total: CSLU: Numbers (LDC2009S01), CSLU: S4X (LDC2009S03), CSLU: Alphadigit (LDC2008S06) and CSLU: ISOLET (LDC2008S07).

Supporting researchers in natural language processing, Academica Sinica contributed proposition bank-style annotations of 500 English biomedical journal abstracts (LDC2009T04) and a new part-of-speech tagged version of the Chinese Gigaword (LDC2009T14). Additional releases include:

- All available annotations of the Switchboard corpus including syntactic structure and disfluencies, phonetic transcripts, dialog acts and prosody plus new annotations for animacy, markables, focus/contrast and animacy all integrated via NITE XML and contributed by Sasha Calhoun, Jean Carletta, Dan Jurafsky, Malvina Nissim, Mari Ostendorf and Annie Zaenen (LDC2009T26)
- Treebank and structural metadata annotation of 144 English telephone conversations, 140K words, created by LDC for the DARPA EARS program (LDC2009T01)
- a revision of the Arabic Treebank, part 3 (LDC2010T08) using a new rigorous specification, parts 1 & 2 using the same specification to be re-released within the year
- Factbank, 208 documents, 77K tokens, of newswire and broadcast news with double-layered annotation of event factuality (LDC2009T23) contributed by Roser Sauri and James Pustejovsky.
- Language Understanding Annotation Corpus, 7K words of English and 2K of Arabic text intensively annotated for committed belief, event and entity co-reference, dialog acts and temporal relations (LDC2009T10) contributed by a multisite team based on meetings at the HLT Center of Excellence at Johns Hopkins University (LDC2009T07)
- new release of Ontonotes containing 1324k of Chinese, 1150k of English and 200k words of Arabic newswire, broadcast news and broadcast conversation from multiple sources annotated for syntactic and predicate argument structure and co-reference created through a collaboration of BBNT, the Universities of Pennsylvania and Colorado and the University of Southern California's Information Sciences Institute (LDC2009T24)
- Brown Laboratory for Linguistic Information Processing (BLLIP) corpora containing Penn Treebank-style parsing of ~24 million sentences from the LDC North American News Text corpora (LDC2008T13, LDC2008T14)

- new version of Chinese PropBank (LDC2008T07) containing predicate-argument annotation of 750,000 words from the Chinese Treebank contributed by Nianwen Xue, Martha Palmer, Meiyu Chang, Zixin Jiang.
- Hindi WordNet, the first for an Indian language, contributed by researchers at the Center for Indian Language Technology, IIT Bombay with 56,928 unique words and 26,208 synsets (LDC2008L02).
- NomBank (LDC2008T23) contributed by Adam Meyers, Ruth Reeves and Catherine Macleod at New York University annotating argument structure for instances of common nouns in 114,576 propositions of the Penn Treebank's Wall Street Journal texts.
- COMNOM v1.0, an enriched version of COMLEX Syntax Lexicon also created at New York University (LDC2008T24)
- Penn Discourse Treebank (LDC2008T05) containing annotations of 40,600 discourse relations in the Wall Street Journal subset of the Penn Treebank
- Czech Academic Corpus (LDC2008T22) created by a multi-site team in the Czech Republic and containing 650,000 words of news and magazine text and broadcast transcripts manually annotated for morphology and syntax plus tools for corpus search and management and the editing of morphological and syntactical annotations
- CoNLL 2008, the shared task data (LDC2009T12) including excerpts from multiple tagged corpora annotated for POS, syntactic dependency, semantic role set, named entities and WordNet super senses
- 10,567 English posts from age-specific chat rooms annotated with chat dialog-act tags and part-of-speech tags contributed by Eric Forsyth, Jane Lin, Craig Martell of the Naval Post-Graduate School (LDC2010T05)

Connecting to scholars in linguistics and area studies, LDC has released:

- An English Dictionary of the Tamil Verb (LDC2009L01) developed by Professor Harold Schiffman and Vasu Renganathan of the University of Pennsylvania containing translations for 6597 English verbs and definitions of 9716 Tamil verbs and available in PDF and XML formats.
- Global Yoruba Lexical Database (LDC2008L03) created by LDC's Yiwola Awoyale containing definitions of 450,000 words from this Niger-Congo language and varieties affected by it including Gullah, Lucumí and Trinidadian.
- Audiovisual Database of Spoken American English (LDC2009V01) developed at Butler University to support research in speech production and recognition and containing seven hours of audiovisual recordings of fourteen American North Midland speakers producing syllables, word lists and sentences used in both academic and clinical settings.

4. Recent and Current Projects

Beyond its role as archive and distributor of language resources, LDC is actively engaged in a number of data creation projects each of which will provide new LRs for general use. A small selection of such projects follows.

The DARPA GALE program develops technologies that interpret large volumes of speech and text in multiple languages to deliver pertinent information in usable form to monolingual English-speaking analysts and decision makers in response to direct or implicit requests. GALE technologies transcribe, translate and distill speech and text in multiple languages and output structured, integrated English text. LDC provides GALE with data, annotations, tools, standards and best practices for system training, development and evaluation.

DARPA Machine Reading makes knowledge in natural language text available for automated processing with little human intervention. Machines learn to read from few examples and read to learn in order to answer questions and perform reasoning tasks. LDC develops and distributes linguistic resources for MR including source data, annotation guidelines, annotated data, use cases, system assessment, annotation, corpus infrastructure and related evaluation resources.

DARPA MADCAT (Multilingual Automatic Document Classification Analysis and Translation) develops systems to automatically provide page segmentation, metadata extraction, OCR and translation in order to convert foreign language text images into English transcripts for use by humans and processes such as summarization and information extraction. LDC is creating publicly available linguistic resources on a scale and richness not previously available including new collection and annotation of new and existing data to address strategic gaps in genre, dialect, image quality found in existing training resources.

NIST TAC-KBP (Text Analysis Conference Knowledge Base Population) is a technology evaluation campaign building upon the ACE and TAC QA initiatives to foster research in automatically mining named-entity information from unstructured text and inserting it into knowledge bases. LDC creates and distributes English source data in multiple genres; annotations, system assessment, tools and specifications for TAC KBP, including the Entity Linking and Slot Filling tasks.

The Rich Transcription evaluation promotes and measures advances in the state-of-the-art of several automatic speech recognition technologies that produce transcriptions more readable by humans and more useful for machines. LDC creates gold standard evaluation data for RT, including transcription and related annotation of speech in various domains including broadcasts, meetings and lectures.

The NIST OpenMT evaluation series advances the state of the art toward fully adequate and fluent translations. LDC creates gold standard training, devtest and evaluation data, performing data collection, processing and selection; manual translation; and system assessment covering multiple genres and language pairs including Arabic, Chinese, and Urdu with English.

NIST's TRECVid Event Detection track promotes

technology for detection of events from a pre-defined set in video. LDC creates gold standard annotation of evaluation data, adjudication of system output, inter-annotator consistency analyses, and related annotation infrastructure including software, tools and guidelines.

NIST's Metrics MATR is a series of research challenge events for MT metrology, promoting the development of innovative, even revolutionary, measures that are intuitively interpretable and highly correlated with human assessments. LDC supports MetricsMATR by conducting various types of human assessment on MT system output.

Phanotics (Phonetic Annotation of Typicality in Conversational Speech) identifies high-level features characteristic of American dialects, for use in speaker and dialect recognition systems. LDC develops annotation strategies and guidelines, provides audio and transcripts and feature annotation.

The US Department of Education has funded LDC to create dictionaries of the Iraqi, Levantine and Moroccan dialects of Arabic in collaboration with Georgetown University Press based on GUP's published dictionaries. The outcomes will be new GUP print and electronic dictionaries and lexical databases for NLP research.

5. Conclusions and Future Plans

This paper has described selected activities and publications from the past two years at the LDC to address the need for greater volumes of data and associated resources in a growing inventory of languages with ever more sophisticated annotation. The plan for the Consortium over the next two years is to maintain a leadership role in language resource creation and distribution, to continue to support distribution operations and to provide increasing support for local initiatives via memberships and data licenses, to extend outreach to new communities (Cieri and Strassel 2009) including those that require specialized corpora, to continue to integrate HLTs into the LR creation pipeline and to generally increase activities devoted to research, to simplify production through efficiency and outsourcing and to expand provision of tools, specifications and training to members.

6. References

Christopher Cieri, Stephanie Strassel (2009), Closer Still to a Robust, All Digital, Empirical, Reproducible Sociolinguistic Methodology, NWAV 38: New Ways of Analyzing Variation, University of Ottawa, Ottawa, Canada, October 22-25, 2009.

Cieri, Christopher, Mark Liberman (2008) 15 Years of Language Resource Creation and Sharing: A Progress Report on LDC Activities, LREC 2008: Sixth International Conference on Language Resources and Evaluation, Marrakesh.

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

Center for Indian Language Technology Solutions, Indian Institute of Technology Bombay, Mumbai (2009) Hindi WordNet Site,
http://www.cfilt.iitb.ac.in/wordnet/webhwn/

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.

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

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

Mapelli, Valérie, Victoria Arranz, Hélène Mazo, Khalid Choukri, (2008) Latest Developments in ELRA's Services, LREC 2008: Sixth International Conference on Language Resources and Evaluation, Marrakesh.

Prasad, Rashmi, et al. (2008) Penn Discourse Treebank Version 2.0, Linguistic Data Consortium, Philadelphia.

Ralph Weischedel, et al. (2009), OntoNotes Release 3.0 Linguistic Data Consortium, Philadelphia.

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.