More Data and Tools for More Languages and Research Areas: A Progress Report on LDC Activities

Christopher Cieri, Mark Liberman

Linguistic Data Consortium, 3600 Market Street, Philadelphia, PA 19104
ccieri,@ldc.upenn.edu, myl@ldc.upenn.edu

Abstract

This presentation reports on recent progress the Linguistic Data Consortium has made in addressing the needs of multiple research communities by collecting, annotating and distributing, simplifying access and developing standards and tools. Specifically, it describes new trends in publication, a sample of recent projects and significant improvements to LDC Online that improve access to LDC data especially for those with limited computing support.

1. Introduction

Considerable change characterizes the language resource landscape of the past few years. While need continues for resources in an ever growing number of languages with increasingly sophisticated annotation, capacity grows at rates and for reasons that are often independent. The result is a partial match of supply and demand that challenges researchers, managers and data centers.

Advances in computing including 4GHz processors, 1Tb drive systems, full motion, DVD quality video, CD quality sound, gigabit networking and media production capabilities give the average researcher the infrastructure needed to collect, annotate and produce small- and medium-scale corpora in particular to address under-represented languages and disciplines.

Improvements in high-end technologies give large data centers the ability to store and process volumes of data at speeds previously inconceivable. Contrasting the situation a few years ago, when the data needs seemed insatiable, it is now possible to produce some kinds of language data more rapidly than technology developers can exploit them. This has been observed in both the DARPA TIDES and EARS programs where some research groups working on the machine translation and speech-to-text tasks reported, for the first time, not being able to train their systems on all of the data provided during the annual cycle. Although the need for more data continues, current production rate are adequate in some areas. For example, the final years of TIDES and EARS as well as their follow on program, GALE, emphasize source variation, richness and quality of annotation and coordination of resource types over volume.

As some technologies approach human performance, it becomes important to both maximize quality – even at the cost of reduced volume – and to understand the natural limits on human performance including inter-annotator agreement.

The adoption by new research communities of digital linguistic resources and the practice of resource sharing increases demand for simple, adaptive access to existing data, opportunities for interdisciplinary work and the need for flexible standards. Similarly, the growing presence of computing in many parts of the world both increases the diversity of languages represented on the Internet, raising the demand for technologies in these languages that in turn requires language resource kits.

2. The Role of the Linguistic Data Consortium

The Linguistic Data Consortium was originally established in 1992 to serve as a distribution point and archive of language resources; this is still the primary function. LDC began collecting and transcribing conversational telephone and broadcast news speech in 1995 and in 1998 added annotation, more broadly conceived, as an important task. In 1999, the development of tools and standards became an area of focus. LDC's mission is to support language-related education, research and technology development by creating and sharing linguistic resources: data, tools and standards. Specific activities, with examples, follow:

- Resource Distribution
- Intellectual Property Rights Management
- Data Collection
    - news text
    - blogs
    - zines
    - newsgroups
    - broadcast news and talk
    - telephone conversation
    - meetings
    - read and prompted speech
- Annotation
    - transcription
    - time-alignment
    - turn and word segmentation
    - morphological
    - part-of-speech
    - gloss
    - syntactic
    - semantic
    - discourse
    - disfluency
    - topic relevance
    - identification and classification of
        - entities
        - relations
        - events
        - co-reference
    - summarization

    - translation and multiple translation
- Lexicon Building
    - pronunciation
    - morphological
    - translation
- Infrastructure Building
    - OLAC: Open Language Archives Community
    - Annotation Graph Toolkit
    - SPHERE Utilities
    - annotation workflow systems
- Tools
    - Transcriber
    - MultiTrans & TableTrans
    - Buckwalter Arabic Morphological Analyzer
    - BITS: Bilingual Internet Text Search
    - Champollion: sentence aligner for parallel text
    - XTrans: multichannel transcription
- Standards and Best Practices
    - Topic Detection and Tracking v1.4
    - Entity Annotation Guidelines v2.5
    - Relation Anotation Guidelines v3.6
    - Simple MDE v6.2
- Consulting and Training
- Hosting and Maintaining research fora
    - Talkbank Workshops
    - LDC Institute

Since its founding, LDC has distributed more than 31,300 copies of 558 Corpora and otherwise shared data with to 2019 organizations in 93 countries. LDC currently adds three corpora to its catalog each month. Membership and licensing fees support this activity completely.

LDC is organized as a consortium, a group of organizations, hosted by the University of Pennsylvania. The management staff in Philadelphia now numbers 43 full-time and up to 65 part-time employees. Yearly memberships are of three types. Online members have access to the subset of data included in LDC Online, described below. Standard members also have access to LDC Online, may request licenses for up to 16 corpora per membership year and receive discounts on licenses of data from previous membership years. Subscription memberships were added in 2005 and now account for 23% of all members. These members have all the rights of Standard Members but automatically receive 2 copies of all corpora on media as they are released. Many corpora are also available for license to non-members. The LDC model permits broad distribution of data with uniform licensing within and across research communities. It also relieves funding agencies of distribution costs while giving members access to vast amount of data. The cost to create any one of the corpora in the LDC catalog is at least as much as the membership fee; in many cases it is one, two or even three orders of magnitude greater. LDC data comes from donations, funded projects at LDC or elsewhere, community initiatives and LDC initiatives. Tools and specifications are distributed without fee.

LDC increasingly serves as data producer and/or distribution center for common task research programs. The benefit these programs provide to human language technologies is evident in many ways including in the resources they produce and share. For example, TIDES, EARS, ACE, and AQUAINT have together produced more than 100 publicly accessible corpora.

3. Recent Publications

The volume and character of LDC publications has changed since the last LREC progress report (Cieri and Liberman 2004). In order to reduce a corpus backlog that had developed in late 2003, LDC increased corpus production from 2 to 3 titles per month, a rate that has held steady since early 2004. During that same time, LDC has added two titles to the catalog to support the development of dialogue systems: the 2000 and 2001 Communicator Dialogue Act Tagged corpora. Researchers in Speaker Recognition now have access to both the second phase of Switchboard Cellular collection and to the 2002 NIST Speaker Recognition Evaluation corpus.

The largest growth both in terms of titles and sheer volume are in corpora for speech recognition research. The Articulation Index corpus contains multiple speakers pronouncing up to 2000 syllables permitted by English phonology including actual and nonsense syllables. LDC has also published corpora of transcribed broadcast news in Czech and English and of transcribed telephone conversations in Mandarin, English and Levantine Arabic, the last of these requiring the invention of an writing system.

The DARPA EARS program encouraged research into the identification of both conversational disfluency and its repair and into the delimiting of sentence-like units in speech. The goal of this research was to enhance automatic transcripts to benefit downstream processing and improve readability, for example through proper capitalization and punctuation. The fruits of this work include the MDE RT-03 and RT-04 Training Speech and Annotations.

Responding to the growing interest in meeting transcript, LDC has published the ICSI, ISL and NIST corpora of meeting speech with transcripts.

An ongoing agreement with the Center for Technology Enhanced Language Learning (CTELL), of the U.S. Military Academy's Department of Foreign Languages has provided several corpora to support speech recognition research. The collaboration, originally suggested by Colonel Steve LaRocca, now of the Army Research Laboratory, has lead to LDC’s distribution of read speech corpora in Arabic and Russian with Croatian and English added since 2004. Additional publications are planned for the coming months.

A new agreement with the Center for Spoken Language Understanding (CSLU) of the Oregon Health & Science University permits LDC to publish all of CSLU’s corpora for non-commercial education, research and technology development. Under the agreement, proposed by Jan van Santen of CSLU, LDC has already added the CSLU Voices and 22 Languages corpora to its catalog and will release the remainder over the coming months.

In addition to the speech corpora, LDC has produced a large number of text corpora. The largest of these are the second editions of the Arabic, Chinese and English

Gigaword corpora targeting an order of magnitude of a billion words or a billion Chinese characters of news text.

The TDT4 corpora contain English, Chinese and Arabic broadcast news audio, their transcripts and news text all annotated for topic relevance. The 2004 HARD corpora contain text, topics and relevance annotations used in the TREC HARD track.

To support information extraction under the ACE and DARPA TIDES programs, LDC has created several corpora that annotate entities, relations, events and co-reference in English, Chinese and Arabic. Since the last progress report, LDC has published the 2003, 2004 and 2005 ACE/TIDES multilingual training corpora as well as the 2004 ACE Time Normalization English data and the BBN Pronoun Coreference and Entity Type corpus.

Researchers in machine translation were well served with two Arabic-English and two Chinese-English parallel text corpora focusing primarily but not exclusively on news. In addition, the Hong Kong Parallel Text corpus contains law codes from the Department of Justice, press releases from the Information Services Department and excerpts from the Official Record of Proceedings of the Legislative Council of the Hong Kong Special Administrative Region. LDC has also published three Chinese multiple translation corpora containing original Chinese news text and sentence aligned translations into English from multiple human sources. A very large Chinese-English name translation list was also published.

Researchers interested in morphological, syntactic or semantic analysis of text received a new version of the Buckwalter Arabic Morphological Analyzer, multiple updates to the Penn Chinese Treebank and LDC Arabic Treebank, the Prague Arabic and Czech-English Dependency Treebanks and Proposition Banks in English and Chinese.

TalkBank, a collaboration joining researchers at Carnegie Mellon University and the University of Pennsylvania, funded by the National Science Foundation (BCS-998009, KDI, SBE, ITR-0324883), fostered fundamental research into human and animal communication. In its final two years, Talkbank focused attention on data creation and produced: the Klex: Finite-State Lexical Transducer for Korean, Morphologically Annotated Korean Text, the Santa Barbara Corpora of Spoken American English parts III and IV, Field Recordings of Vervet Monkey Calls and the FORM1 Kinematic Gesture corpus.

Finally, LDC has also published Moussa Bamba’s Mawukakan Lexicon, Florian Wolf's Discourse Treebank and the second release of the American National Corpus.

4. Recent and Current Projects

The TIDES (Translingual Information Detection Extraction and Summarization) program that concluded in 2005, built underlying technologies for news understanding systems. TIDES sponsored extensive corpus building including Gigaword news text corpora, annotations for topic relevance, identification of entities, relation, events and coreference, parallel, translated and multiply translated text, and summaries all in English, Arabic and Chinese.

The DARPA EARS (Effective, Affordable, Reusable Speech-to-Text), which also concluded in 2005, advanced the state of the art in speech recognition to produce high quality transcripts of broadcast news and conversational telephone speech that could be variably read by humans or processed by downstream applications. EARS data producers including, LDC, BBNT, HKUST and WordWave produced audio and transcripts for 4000 hours of conversational telephone speech in English, 350 hours in Mandarin Chinese and 260 hours in Levantine Colloquial Arabic.

Much of the data in these collections were transcribed using a Quick Transcription (QTr) specification that retains the necessary information for speech recognition systems training but requires only 5-7 hours of human effort to transcribe each hour of speech. LDC also produced new corpora for evaluating speech-to-text systems in English, Arabic and Chinese and provided MDE (Metadata Extraction) annotation to support systems that identify speakers, syntactic/semantic units and disfluencies and their repairs. The output of MDE systems may be used for downstream processing or filtered for display by, for example, grouping utterances according to speaker, using the boundaries of syntactic/semantic units to determine punctuation and capitalization and removing disfluent speech.

DARPA GALE (Global Autonomous Language Exploitation) will build systems that process media in a variety of languages, beginning with English, Mandarin and Arabic, in order to answer questions. The processing will include transcription, translation and distillation of text into structured information. The media will include not only news text, broadcast news and telephone conversations but also broadcast conversation and round tables discussions, news groups and blogs. GALE will produce numerous data resources to support this effort included large volumes of text and transcribed speech that has been translated and aligned at the sentence and sub-sentence level, annotated for syntactic structure and proposition content and distilled via human effort into structured information.

These are just a sample of current projects. Space constraints prevent a fuller description but the following list gives a sense of the diversity of efforts:

- Mixer: collect up to 30 calls from each of 600 subjects in five different languages using at least 4 unique handsets
- Transcript Reading: record each of 100 Mixer subjects reading samples of their own and others’ transcripts via 9 different sensors
- Language Variation and Dialect Identification: record 100 conversations in each of 32 linguistic varieties auditing calls for language
- Automatic Content Extraction: annotate English, Chinese, Arabic text from written and spoken sources for entities, relations, events and co-reference.
- Less Commonly Taught Language (LCTL): create resource kits for LCTLs including monolingual & parallel news text, bilingual lexicons, encoding converters, word & sentence segmenters, POS tagsets and taggers, morphological analyzers and

tagged text, named-entity tagger and tagged text, personal name transliterator and grammatical sketch.

5. Outreach via LDC Online

With interest in shared language data rising among linguists, anthropologists, psychologists and language teachers, comes a new challenge for easily accessible data. By default, LDC data is organized and formatted to accommodate target HLT research communities who prefer to process under computer control. To broaden its utility, data needs to be re-formatted or otherwise presented to accommodate non-technical researchers. In response to this need, LDC has, over the past eighteen months, re-designed and re-built its LDC Online service. Development of the original LDC Online was funded by National Science Foundation as part of a multi-year effort. The second edition was rewritten completely with local funding.

Rather than providing access to LDC corpora as they appear in the Catalog, LDC Online now delivers rapid access to major collections of LDC data, indexed by language and type, regardless of whether the data has previously been shared in a corpus. Exceptions are the somewhat unique Brown Corpus and the American English Spoken Lexicon provided in their original form.

Current news text indices cover 4.4 billion words or Chinese characters including 500 million words of Arabic news text, 1.4 billion Chinese characters of news text and 2.5 billion words of English news text. About 4000 hours of transcribed English conversations have also been indexed adding 26 million words to the total. The news text collections are indexed at the word level and contain searchable metadata for date and source of the story. The conversations contain additional metadata including the project under which they were collected, topic of the conversation and, for each speaker, an anonymous speaker ID, sex, geographic region where raised, age group and level of education.

The search engine underlying LDC Online was written by Mike Schultz while an LDC employee. Mike optimized the engine for speed and completeness even when faced with very large corpora. The engine provides keyword and phrase search in the text and in the metadata fields. Search terms may be combined with Boolean AND, OR and NOT. Wildcards are also permitted and the engine also supports relevancy searching. Search returns are in the form of keyword-in-context or, when intellectual property rights permit, in full text.

LDC Online is available to all LDC members. A large subset, about 10 million words in 10,000 documents, is also available to non-members for non-commercial research at no cost

6. Conclusion and Future Plans

This paper has described selected activities of 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 constituencies including commercial ventures that require specialized corpora, to make better use of technologies that are based upon LDC data 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.

7. References

Cieri, Christopher, Mark Liberman (2004) A Progress Report from the Linguistic Data Consortium: recent activities in resource creation and distribution and the development of tools and standards, LREC 2004: Fourth International Conference on Language Resources and Evaluation.

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, Ralph Weischedel (2004). Automatic Content Extraction (ACE) program - task definitions and performance measures. LREC 2004: Fourth International Conference on Language Resources and Evaluation.

LDC (2006) Linguistic Data Consortium Home Page, http://www.ldc.upenn.edu/.

NIST (2004), The NIST Year 2004 Speaker Recognition Evaluation Plan

http://www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf.

Maamouri, Mohamed, Tim Buckwalter, and Chris Cieri (2004a) Dialectal Arabic Telephone Speech Corpus: Principles, Tool Design, and Transcription Conventions, Paper presented at the NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, Sept. 22-23, 2004.

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, and Wigdan Mekki (2004b) The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus, Paper presented at the NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, Sept. 22-23, 2004.

Maeda, Kazuaki and Stephanie Strassel (2004) Annotation Tools for Large-Scale Corpus Development: Using AGTK at the Linguistic Data Consortium. LREC 2004: Fourth International Conference on Language Resources and Evaluation

Strassel, Stephanie (2004) Linguistic Resources for Effective, Affordable, Reusable Speech-to-Text LREC 2004: Fourth International Conference on Language Resources and Evaluation