A Progress Report from the Linguistic Data Consortium: recent activities in resource creation and distribution and the development of tools and standards

Christopher Cieri and Mark Liberman

University of Pennsylvania, Linguistic Data Consortium
3600 Market Street, Suite 810, Philadelphia, PA. 19104-2653, USA
{ccieri|myl}@ldc.upenn.edu

Abstract

This paper described recent activities of the Linguistic Data Consortium in the collection, annotation and distribution of language data the developments of tools and standards for using that data, the creation of metadata to facilitate the search for linguistic resources.

Introduction

Rapid changes in the landscape of linguistic research and technology development require continuous adaptation from international data centers who would serve the research communities involved. The only constant among these communities is the need for greater volumes of high quality data and tools with which to process the data. Variable are the human languages, data types, annotations, standards and formats needed. This presentation reports on the progress the Linguistic Data Consortium (LDC) has made in distributing existing resources, in collecting and annotating new resources and in developing and sharing standards and tools to address the needs of multiple research communities who are joined by their use of digital linguistic resources.

Resource Distribution

LDC's original, and still primary mission, is to support education, research and technology development by serving as a central distribution point and repository of language resources. LDC's operational model is strongly tied to the notion of a consortium in which members who believe in the work of the organization provide yearly support and receive benefits well in excess of what their membership fees would acquire on the open market. LDC members receive ongoing rights to each database released in the years in which they support the consortium. LDC released 24 data sets in 2002 and 27 in 2003. Membership agreements, differing by organization type, govern the use of LDC data. On rare occasions, corpus specific agreements supercede the membership agreement and further constrain the use of a corpus. Most LDC corpora are also available for licensing to non-members. LDC membership fees have not changed since the Consortium was founded. The annual fee is significantly less than the cost to produce just one corpus.

There is clear evidence that this model provides extraordinary support to research organizations worldwide. Since its founding, just over 12 years ago, LDC has distributed more than 21,200 copies of 288 different corpora to more than 1720 organizations in 89 countries excluding the data sets available for free download from the web pages. Membership and licensing fees completely support this distribution activity.

The percentage of LDC members in each of the commercial, government and non-profits sectors has remained stable since we last reported it at LREC 2002. Figure 1 shows that about three-fourths of LDC members are in the non-profit sectors. Commercial organizations comprise nearly an additional one-fifth of LDC members while government organizations including the research branches governments around the world account for the remainder.

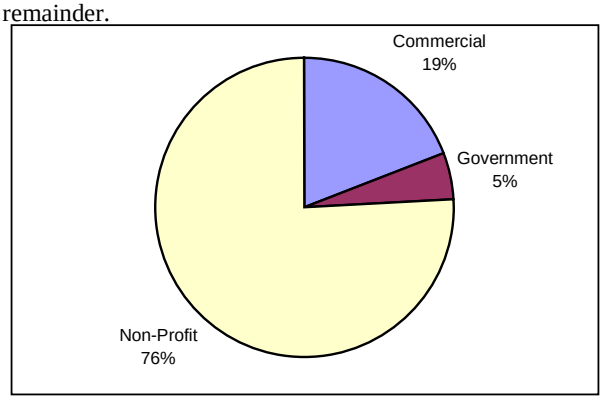

Figure 1: LDC Membership by Organization Type

LDC data users are by no means limited to, or even concentrated in, the United States. Figure 2 shows the geographical distribution of organizations that use LDC data. This map is not limited to Consortium members or even organizations that license data for a fee but also includes those who have requested corpora distributed without fee under the NSF funded Talkbank program and those who have registered to download free data from LDC's web page.

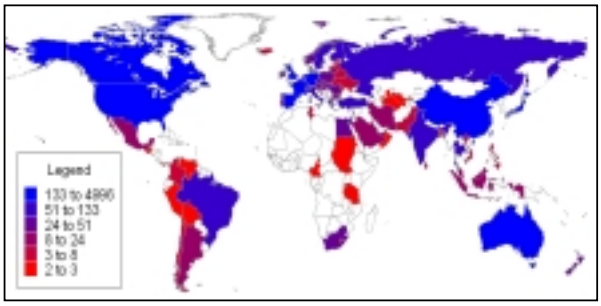

Figure 2: Geographical distribution of LDC Data Users

The activity of the consortium, as measured by the number of new users and by the number of corpus distributions, has continued to grow with a slight increase in the rate of acceleration over the past years. A small but growing part of our work is devoted to experimental corpora, created for a specific research program such as DARPA TIDES or NSF Talkbank, distributed initially to a small research community and subsequently released beyond.

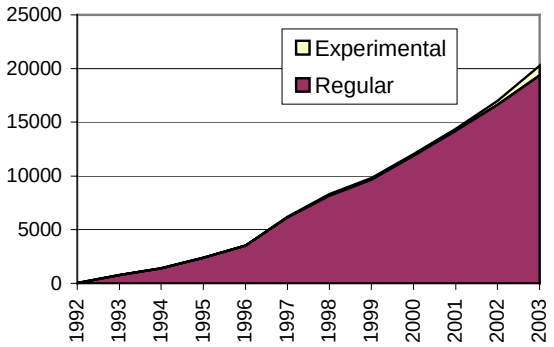

Figure 3: Number of distributions of LDC data sets over time excluding those users download freely from our web site.

Just as the level of publication and distribution activity is increasing so is the variety of data available. The bullets below demonstrate this by giving a very cursory summary of LDC publications over the last two years categorized by the research areas they best address. For those interested in any specific corpus, the LDC Catalog (http://ldc.upenn.edu/Catalog) provides a more thorough description.

- Language Modeling: Gigaword News text Corpora in Arabic, Chinese and English, AQUAINT Corpus of English News Text
- Tagging and Parsing: Arabic Treebank Parts 1 & 2, Korean-English Treebank, Morphologically Annotated Korean Text, Buckwalter Arabic Morphological Analyzer
- Machine Translation: updated Chinese-English Translation Lexicon and Multiple-Translation Corpora in Arabic and Chinese
- Speaker Recognition: Switchboard-2 Phase III Audio, 2001 NIST Speaker Recognition Evaluation
- Speech Recognition:
  - Prompted Speech: West Point Corpora in Arabic and Russian
  - Broadcast News: HUB4 English Speech and Transcripts
  - Meetings: ICSI Meeting Speech & Transcripts
  - Telephone: Voicemail Corpus Part II, HUB5 English, Egyptian Arabic, English, German, Mandarin, Spanish, CallHome style telephone conversation audio, transcripts and lexicon in Egyptian Arabic and Korean
- Dialog Systems: 2002 and 2001 Communicator Corpora
- Information Extraction and Summarization: Message Understanding Conference (6), ACE-2, TIDES Extraction (ACE) 2003 Multilingual Training Data, SummBank 1.0
- Gesture Recognition: FORM2 Kinematic Gesture
- Balanced Text: American National Corpus
- Other
  - Emotional Prosody Speech and Transcripts
  - RST Discourse Treebank
  - Translanguage English Speech & Transcripts
  - Grassfields Bantu Fieldwork: Dschang Lexicon & Tone Paradigms
  - SLX Corpus of Classic Sociolinguistic Interviews
  - Santa Barbara Corpus of Spoken American English Part-II

Resource Creation

LDC began to collect and transcribe conversational telephone and broadcast news speech in 1995; annotation became a major focus in 1998. Since that time, dozens of the large scale corpora that have appeared in the LDC catalog result from these efforts. Two recent examples of the synergy between sponsored, common task research projects and sharable linguistic resources are the DARPA TIDES and EARS programs.

TIDES (Translingual Information Detection Extraction and Summarization) seeks to build the underlying technology for a news understanding system by advancing the state of the art in information detection and extraction, document summarization and translation. Toward this end, TIDES has sponsored extensive corpus building including gigaword news text corpora, annotations for topic relevance and entity identification, parallel text corpora, multiple translation corpora and news summaries in English, Arabic and Chinese. Gigaword corpora contain an order of magnitude of a billion words of news text with consistent markup. Multiple translation corpora are those in which a core set of source language documents are translated by multiple translators with the results being aligned at the sentence level.

The DARPA EARS (Effective, Affordable, Reusable Speech-to-Text) program seeks to advance the state of the art in speech recognition to produce high quality, human readable transcripts of broadcast news and conversational telephone speech. In support of the EARS program, LDC has collected over 2000 hours of conversational telephone speech in English from over 15,000 subjects all living within the United States. Similar, albeit smaller, collections are underway in Arabic and Chinese. The English subjects are balanced by gender, age and the region in which the grew up. Dialect regions were based upon William Labov's Phonological Atlas of the North America which can be found at http://www.ling.upenn.edu/phono_atlas/home.html. Each conversation in this collection has been transcribed using the QTr, or Quick Transcription, specification which retains the crucial information for speech recognition systems training but requires only 5-7 hours of human effort to transcribe each hour of speech. Under the EARS program, LDC has also produced new corpora for evaluating speech-to-text system in English, Arabic and Chinese and to support systems that identify speakers, identify and repair disfluencies and punctuate a text to improve readability.

For both TIDES and EARS, the primary challenge lies not in creating individual resources but rather in coordinating resource creation so that raw linguistic resources meet the needs of multiple technology development areas and so that the allocation of effort matches the estimated benefit to the program. LDC maintains multiple planning documents to assist this process. The first is a data matrix which shows all of the resources the community plans to acquire or create during the upcoming funding cycle. Figure 4 is a snapshot of the EARS Data Matrix for 2004. Readers may find the complete version at http://ldc.upenn.edu/Projects/EARS.

Figure 4: The EARS Data Matrix for 2004

The second planning tool is the Corpus Specification which describes the corpus in detail starting from needs and assumptions and working through raw data collection, annotation and quality control and finishing with distribution formats. Figure 5 is a snapshot of one early LDC corpus specification. Full versions can be found at http://ldc.upenn.edu/Projects.

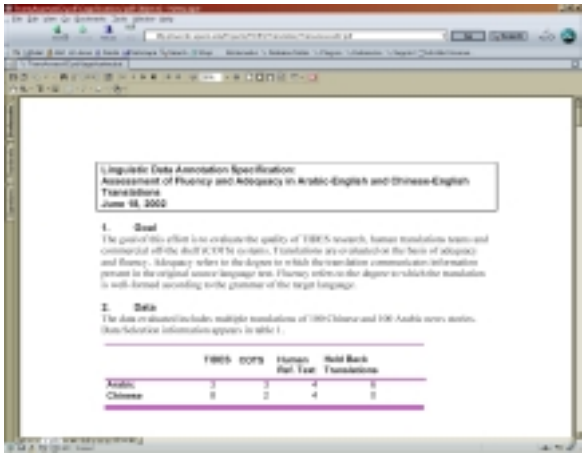

Figure 5: Corpus Specification

The third, and possibly most important, document for coordinating language resources within large multisite programs such as DARPA TIDES and EARS is the Resource Map, a table of resources available to program participants with an indication of which need each resource meets and instructions for retrieving it. These maps may include resources available outside the program but feature a larger number of resources created specifically for the program – resources that will eventually be released more generally. Figure 6 is a snapshot of a piece of the Arabic language resource map for DARPA TIDES. Full versions live at http://ldc.upenn.edu/Projects.

Standards and Tools

Simultaneous with increased activity among the communities that have traditionally used digital language resources, interest has recently grown among new communities such as linguists, anthropologists, psychologists and language teachers. Each new community, with its own history and expectations, presents a challenge to data distribution centers. Not all communities can support full-scale corpus development and many have sought to reuse resources that were created for other purposes. Needs analysis, support of data reuse and re-annotation and the communication of data standards and best practices are primary among goals of the NSF sponsored TalkBank (http://www.talkbank.org/) project. Talkbank has coordinated working groups in multiple disciplines and developed tools such as the Annotation Graph Toolkit and MultiTrans, a multi-channel transcription tool, which cut across traditional boundaries to serve users in multiple research communities. In the past year, Talkbank has also sponsored the distribution of several new corpora including the SLx corpus of sociolinguistic interviews, the FORM corpus of gesture annotated video and the second part of the Santa Barbara Corpus of Spoken American English.

Both LDC and ELRA have joined forces with 21 other archives of language data to build the Open Language Archives Community (OLAC). OLAC has built a union catalog of resources available from its member archives. The catalog is freely available via the Internet at http://www.language-archives.org.

Resource Coordination

Changes in both the supply of and the demand for language resources continue to affect the role of large data centers such as LDC and ELRA within the research communities they serve. The past two years have seen increased demand for extensive batteries of coordinated language resources with sophisticated annotation for several major languages. The DARPA TIDES and EARS programs have already created such resources for use with their programs. LDC has begun to release the resulting corpora for use beyond these programs and all of the data sets will eventually be available for general use.

Data centers such as LDC and ELRA and similar organizations starting up or newly underway in China, Japan, Korean and India will be well placed to address these needs if they can manage to dove-tail new development with existing resources filling known gaps either by creating or assisting in the creation of new data.

The Linguistic Data Consortium is involved in ongoing projects to address all of these issues described above. This paper has described ongoing LDC activity in corpus creation, annotation and distribution as well as efforts to bring together communities of researchers, to identify best practices and develop tools of general use.

Figure 6: A partial Resource Map for DARPA TIDES

References

ACE, 2000, Automatic Content Extraction [www.nist.gov/speech/tests/ace].

Bird, Steven, Kazuaki Maeda, Xiaoyi Ma, Haejoong Lee, 2002, MultiTrans and TableTrans: Annotation Tools Based on the Annotation Graph Toolkit (AGTK), Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

Bird, Steven, Hans Uskoreit, Gary Simons, 2002, The Open Language Archives Community, Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

Cieri, Christopher, David Miller and Kevin Walker, 2002, Research Methodologies Observations and Outcomes in Speech Data Collection, HLT 2002: Proceedings of the Second International Conference on Human Language Technology Research, San Diego, CA, March 24-27, 2002.

Doddington, G. (1999). The 1999 Topic Detection and Tracking (TDT) Task Definition and Evaluation Plan. Available at http://www.nist.gov/TDT.

EARS, 2004, DARPA Program in Effective, Affordable, Reusable Speech-to-Text (EARS) [http://www.arpa.mil/ipto/Programs/ears/index.htm]

LDC, 2004, Linguistic Data Consortium Homepage [http://www.ldc.upenn.edu]

Strassel, Stephanie, Dave Graff, Nii Martey and Christopher Cieri, 2000, Quality Control in Large Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora. In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

TalkBank, 2000, NSF TalkBank Program [www.talkbank.org]

TIDES, 2000, DARPA Program in Translingual Information Detection Extraction and Summarization [http://www.arpa.mil/ipto/Programs/tides/index.htm]

Wayne, Charles, 2000, Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.