# 15 Years of Language Resource Creation and Sharing: 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}@ldc.upenn.edu

### Abstract

This paper, the 5th in a series of biennial progress reports, reviews the activities of the Linguistic Data Consortium with particular emphasis on general trends in the language resource landscape and on changes that distinguish the two years since LDC's last report at LREC from the preceding 8 years. After providing a perspective on the current landscape of language resources, the paper goes on to describe our vision of the role of LDC within the research communities it serves before sketching briefly specific publications and resources creations projects that have been the focus our attention since the last report.

## 1. Introduction

Marking the 10 year anniversary of LREC and the 15 year anniversary of the Linguistic Data Consortium, this paper reviews the activities of the latter with particular emphasis on general trends in the language resource landscape and on changes that distinguish the two years since LDC's last report at LREC from the preceding 8 years.

While it remains true that the language resource landscape is characterized by considerable change, the past two years since our last report at LREC stand out from the previous 8 in a number of ways. We see continuing growth in the need for resources in an ever growing number of languages with increasingly sophisticated annotation to satisfy an ever expanding list of user communities.

Advances in computing continue to endow the average researcher with expanding capacity so that the number of users and indeed creators of language resources continues to grow. Nonetheless the need for large specialist organizations such as LDC, ELRA/MedLTC and Appen to undertake large scale resource creation efforts and to coordinate across smaller efforts has never been greater.

As 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. In contrast to a time when the data needs seemed insatiable, we have observed a gradual shift away from volume-at-any-cost toward an appreciation for higher quality data even if quality comes at the cost of a reduction in scale. We saw the beginning of this trend during the final years of the DARPA TIDES and EARS programs, in 2004-5, 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. The DARPA follow-on to TIDES and EARS GALE, now entering its third year, emphasizes source variation, richness, quality of annotation and coordination of resource types over volume. Similarly in the REFLEX LCTL (Less Commonly Taught Languages) program and the NIST LRE (Language Recognition) evaluation, the focus has been on diversity or resource types and languages as opposed to volume in any specific language or type.

The move toward digital linguistic resources by new research communities and the growing practice of resource sharing has had two effects. On the one hand, communities that have recently adopted the approach of sharing digital resources (perhaps sociolinguistics and language teaching fit in this category) need simple, adaptive access to existing data and flexible standards. On the other hand communities that are extending the ways in which they share data require infrastructure for mapping among alternate representations of the same or relate annotations. 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 established in 1992 to distribute and archive language data. Although this remains its primary function, LDC began collecting and transcribing conversational telephone and broadcast news speech a few years later and has expanded its mission every year or two to include annotation more broadly conceived, the development of tools and standards, the coordination of multi-site efforts of the communication of experience through publications and training. LDC's mission as currently defined is to support language-related education, research and technology development by creating and sharing linguistic resources: data, tools and standards.

The specific activities that support that mission include:

- resource distribution
- intellectual property rights management
- data collection including: news text, blogs, zines, newsgroups, broadcast news and talk, telephone conversation, meetings, interviews, read and prompted speech, printed, handwritten and hybrid documents
- annotation including: quick and careful transcription; time-alignment and segmentation at the turn, sentence and word level; tagging of morphology, part-of-speech, gloss, syntax, semantics, discourse function and disfluency; categorization according to topic relevance; identification and classification of entities, relations, events and their co-reference; summarization of various lengths from 200 words down to titles; translation, multiple translation, translation quality control and the alignment of translated text at the document, sentence and word levels
- lexicon Building including: pronunciation, morphological, translation,
- infrastructure building: for example the Open Language Archives Community (OLAC), the Annotation Graph Toolkit, annotation workflow systems, data scouting systems
- tool creation: including BITS: Bilingual Internet Text Search, Champollion: sentence aligner for parallel text, XTrans: multichannel transcription
- creation and sharing of standards and best practices including those for Topic Detection and Tracking, Entity and Relation Annotation and conducting interviews
- consulting and training
- hosting and maintaining research fora

Since its founding, LDC has distributed 53,580 copies of nearly 800 corpora and otherwise shared data with 2540 organizations in 67 countries. LDC currently adds two or 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 45 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. 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.

In order to keep up with increases in facilities, materials and labor costs, LDC raised its licensing fees in mid 2007 and then its memberships as of January 2008. The amount of these increases was modest, 10% for subscription members and 20% for standard members, considering the 3% average annual increase in the time value of money. The price increases were scaled according the demand each member type places on LDC staff. To support loyal, returning members and those who (re-)join early in the membership year, LDC now discounts membership fee by 5% for any returning member and by 5% for any organization that join in the first two months of the year. The overall effect of these changes is that subscription members who maintain their membership from year to year doing so early in the year will actually see a 1% decrease is costs. Others will experience an increase of vary degrees depending upon membership type, when they join and whether they continue membership from the previous year.

## 3. Current Publications

Following up on our last report, in 2006 and 2007, LDC has added 68 titles to its catalog and produced dozens of corpora for use in evaluation programs that will be released generally after they have been exposed to the relevant communities. A sampling of those corpora includes:

- a corpus of email from the Enron scandal that has been annotated for topic
- Gigaword (billion word) News Text corpora in Arabic, Chinese, English, French and Spanish
- broadcast news in Arabic, Korean
- a large number of contribution from CSLU including their Foreign Accented English, Apple Words and Phrases, Yes/No, Spelled and Spoken Words, Stories, Multilanguage Telephone Speech, Portland and National Cellular Telephone Speech, Names Release, Speaker Recognition, Spoltech Brazilian Portuguese and Voices corpora
- parallel text including Arabic Blogs and their translation into English contributed by the DARPA GALE program
- Hungarian-English parallel text contributed by Dániel Varga, László Németh, Péter Halácsy, András Kornai

- STC-TIMIT contains TIMIT data process through a telephone network contributed by Nicolas Morales
- Urdu speech from the Army Research Labs
- Speech in Korean and Spanish contributed by West Point
- treebanks in Arabic, Chinese, Czech, English and Korean with translations into English of the Arabic and Chinese
- the Penn Discourse Treebank created by Aravind Joshi and his students
- a propbank in Korean
- OntoNotes Release 2.0
- conversational telephone speech in Levantine, Iraqi and Gulf Arabic
- parallel text in Arabic and Chinese including two contribution from ISI
- two broadcast news parallel text corpora created at LDC plus one contributed by MITRE
- video key frames and transcripts created by the TRECVID program
- broadband prompted speech in English and Turkish (contributed by the Middle East Technical University)
- telephone band speech in Russian
- the evaluation data from the NIST 2003 and 2004 Rich Transcription campaigns
- the TimeBank corpus contributed by James Pustejovsky and colleagues
- a new version of the ACE 2005 Multilingual Training Corpus with SpatialML annotations contributed by Inderjeet Mani, Janet Hitzeman, Justin Richer, David Harris

## 4. Recent and Current Projects

Beyond it role as archive and distributor of language resources, LDC is actively engaged in a number of data creation projects. A small selection of such projects follows.

DARPA GALE (Global Autonomous Language Exploitation) is building 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 produces 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.

The Mixer project, Phases 1 through 5, have created corpora of multilingual cross channel conversational telephone speech, transcript reading and face-to-face interviews to support robust speaker recognition technologies. Mixer subjects are recorded with multiple microphones including: a lavalier mics on the subject and interviewer, an Etymotic Link-It micro-array, a podium mic, a PZM mic, a studio mic, a hanging conference room mic, a camcorder, four identical studio mics placed at varying distances from the subject, a microphone array, and a head mounted mic used only for brief telephone calls.

The Language Variation and Dialect Identification project: has recorded more than 100 conversations in each of a dozen linguistic varieties and maintain ongoing collection in another 20 varieties with all calls audited for sound quality and language

The REFLEX Less Commonly Taught Language (LCTL) project has created 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. The languages covered are: Amazigh (Berber), Bengali, Hungarian, Pashto, Punjabi, Kurdish, Tagalog, Tamil, Thai, Tigrigna, Urdu, Uzbek and Yoruba. Colleagues at New Mexico State University produced resource in another half-dozen languages.

## 5. 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.

## 6. References

Cieri, Christopher, Mark Liberman (2006) *More Data and Tools for More Languages and Research Areas: A Progress Report on LDC Activities*, **LREC 2006: Fifth International Conference on Language Resources and Evaluation**, Genoa.

LDC (2007) Linguistic Data Consortium Home Page, http://www.ldc.upenn.edu/.

Mani, Inderjeet, et al. (2008). ACE 2005 English SpatialML Annotations, Linguistic Data Consortium, Philadelphia.

Rashmi Prasad, et al. (2008) Penn Discourse Treebank

Version 2.0, Linguistic Data Consortium, Philadelphia.

Dániel Varga, et al. (2008) Hungarian-English Parallel Text, Version 1.0, Linguistic Data Consortium, Philadelphia.

Ralph Weischedel, et al. (2008), OntoNotes Release 2.0 Linguistic Data Consortium, Philadelphia.