Language Resource Creation and Distribution at the Linguistic Data Consortium: A Progress Report

Christopher Cieri and Mark Liberman

University of Pennsylvania and Linguistic Data Consortium
3615 Market Street, Philadelphia, PA 19104-2608 U.S.A.
{ccieri, myl}@ldc.upenn.edu

Abstract

Changes in the supply of and demand for language resources continues to affect the role of large data centers such as the Linguistic Data Consortium (LDC) and European Language Resource Center (ELRA) within the research communities they serve. The past few years have seen increased demand for: intensively multi-modal resources, larger data sets in high-density languages and new data in low density languages; standards and tools for corpus development and re-useable resources. The next few years will bring demand for extensive batteries of coordinated language resources with sophisticated annotation in several major languages. The DARPA program in Translingual Information Detection Extraction and Summarization (TIDES) has already undertaken such resource development; programs with similarly broad scope addressing other technologies will surely follow. Data centers will be well placed to address these needs if they integrate new resource development with distribution of existing resources to fill known gaps by creating or assisting the creation of new data. LDC has projects ongoing to address all of these issues. This paper will provide an overview of LDC activity in corpus creation, annotation and distribution and describe new efforts bring together communities of researchers, to identify best practices and develop tools of general use.

1. Introduction

The growing importance of languages resources in linguistic education, research and technology development requires international data centers to evolve and accommodate emerging needs. The availability of high quality language resources remains a central issue for the many communities involved in basic research, technology development and education related to language. However, the scope, scale and schedule of language resource development continue to change reflecting, and in some cases anticipating, change in the communities of language resource users. Research communities, in general, need greater volumes of data in a broadening inventory of human languages with ever more sophisticated annotation. Where text corpora of tens of millions of words once seemed luxurious, current projects are targeting one billion word data sets to allow robust language modeling of even uncommon phenomena. Similarly, new approaches to re-training in speech recognition require thousands, not hundreds, of hours of audio data. There now exist technology development programs to create rich arrays of coordinated resources – broadcast news, conversational audio, text, transcriptions, named entity annotations, topical categorization, part-of-speech tags, syntactic bracketing and others – not only for English but for an increasing number of languages.

Advancement in technologies such as speech recognition has increased demand for greater volumes of more challenging data. The DARPA-sponsored LVCSR, SPINE and ROAR evaluation projects have stimulated this demand while raising the bar on the notion of robust speech recognition. Researchers working on automatic transcription of telephone conversation have been tackling the problem of recognition of speech transmitted via cell phone. More recent programs in speech recognition focus on recognition of noisy data in dialogues and multiparty conversations. The Speech in Noisy Environments (SPINE) program having completed its second evaluation cycle has produced several corpora containing more than 350 cooperative task dialogues collected under a variety noise and channel conditions. A number of research sites, anticipating the need of the DARPA ROAR (Robust Omnipresent Automatic Recognition) project have begun to collect and annotate intensively multi-modal, meeting data. Researchers working in information detection, extraction and summarization have similar need for large volumes of data in a variety of languages. Within the DARPA TIDES project, teams of researchers are currently addressing problems in translingual information detection, extraction and summarization in six human languages: Arabic, Chinese and English and, to a lesser extent, Korean, Japanese and Korean. . The TIDES research tasks require broadcast transcripts and news texts to be annotated for named entities, categorized by topic, translated, summarized and processed in a variety of other ways. Research communities are looking increasingly to international data centers like LDC to provide such resources.

Interest in the digital collection and analysis of language data has grown dramatically over the past few years. Linguists, anthropologists, psychologists and language teachers are increasingly relying on publicly accessible language resources. Each new research community brings its own expectations about what constitutes a useful corpus in terms of sampling, collection and annotations. However, not all can support full-scale corpus development from scratch. This raises the issues of data reuse, re-annotation and distribution formats as well as the more basic issue of simply finding what one needs among the growing pool of resources. The NSF sponsored TalkBank (http://www.talkbank.org/) project has coordinated working groups in a number of research communities to address the need for data standards and to develop common tools. The LDC sponsored project in Data and Annotations for Sociolinguists (DASL) is a case study in resource re-annotation and in corpus development coordinated across research communities. The Open Language Archives Community (OLAC) has agreed to produce a union catalog of available resources. Under the Networking Data Centers project, sponsored jointly by the National Science Foundation and the European Union, LDC and ELRA work together to pool knowledge about research needs and to jointly produce and distribute resources.

Changes in both the supply of and the demand for language resources continues 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: intensively multi-modal resources, larger data sets in high-density languages and new data in low density languages; standards and tools for corpus development and re-useable resources. Over the next two years, we anticipate demand for extensive batteries of coordinated language resources with sophisticated annotation for several major languages. The TIDES program has already undertaken such resource development and we anticipate other efforts will follow. Data centers like LDC and ELRA 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. This presentation will provide an overview of ongoing LDC activity in corpus creation, annotation and distribution and describe new efforts bring together communities of researchers, to identify best practices and develop tools of general use.

2. Evolutionary Pressures in Language Resources

Linguistic resources continue to play a crucial role in linguistic education, research and technology development where the role of data distribution centers remains central. Published language resources benefit a wider range of researchers, technology developers and their customers. Standard resources reduce duplication of effort, distribute production costs and encourage new participants in a research community. As communities mature, their published resources may be corrected, improved and re-annotated while they also provide a stable reference point against which to compare new methods and algorithms.

Independent researchers and small research groups now have the desktop capacity to create small- to medium-scale corpora. However, the computational, legal and logistical difficulties of resource creation on a large scale challenge most research organizations whether educational, commercial or governmental. While large corporate research groups routinely engage in medium- to large-scale corpus creation, they often view the resulting resources, created at considerable cost, as a competitive advantage that they are understandably hesitant to share.

Recent, expanded multi-site research programs and the emergence of new training techniques (Lamel, et. al., 2001) impose tougher requirements on size, level of annotation and language coverage. Elevated requirements have widened the gap between the types of resources that may be created by individual researcher or small group and those that require large, specialized data creation centers. The **Gigaword News Text Corpora** provide a compelling example. To support robust language modeling even of uncommon phenomena, LDC is creating corpora of 1,000,000,000 (one-billion) words in each of English, Chinese and Arabic. Simply identifying and acquiring distribution rights for that amount of news text – the equivalent of the entire ten-year archive of the several largest news producers – has required intellectual property negotiations and licensing fees that are probably beyond the reach of most research groups working in isolation.

In this environment, we believe that data centers must not be content with their role as distribution agents but must also research best practices in resource collection and annotation and must contribute to the planning of community wide efforts including sponsored research programs.

3. The Linguistic Data Consortium

In 1992, a cooperative agreement grant between the Defense Advance Research Projects Agency (DARPA) and the University of Pennsylvania founded the Linguistic Data Consortium with the goal of creating a distribution infrastructure to support the sharing of linguistic resources. The University of Pennsylvania, acting as the LDC host, provides the consortium with physical space, legal counsel, technical support including access to high-speed networking and business infrastructure including payroll, purchasing and accounts payable and receivable functions. The University enters into all legal arrangements on behalf of the consortium and the research communities it serves.

At its inception, an external planning committee representing potential academic, commercial and governmental users of LDC resource established the LDC business model including membership fees that have not changed in 10 years. LDC membership is open to researchers around the world. The non-profit and government membership fees are roughly the cost of attending an international conference. The membership fee for a commercial organization is an order of magnitude less than the cost of an average medium-scale corpus. As a matter of policy, no bona fide researcher is denied access to LDC data for genuine inability to pay. LDC encourages resource sharing by offering in-kind memberships in exchange for appropriate, quality resources.

Organizations join the LDC on a yearly basis. The Membership Year parallels the calendar year. Members have ongoing rights to all corpora produced in the years in which they join. Current LDC members also have network access to LDC Online, a service that facilitates browsing and searching of indexed text, speech and lexical corpora.

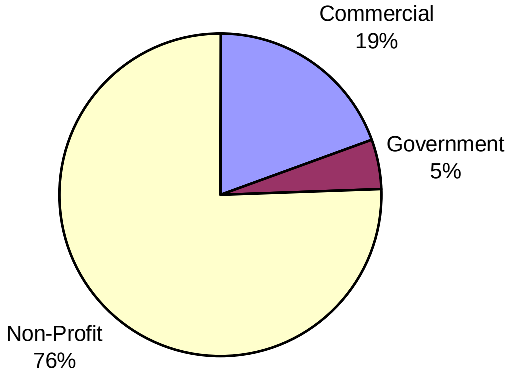

Figure 1: LDC Membership by Type

During its 10 years of operation, nearly 1300 organizations worldwide have used LDC data; more than 380 companies, universities and government research laboratories have joined the consortium; 920 others have licensed corpora as non-members. Of all the organizations that use LDC data, about half are American. Europeans comprise a third of the user base with the remaining groups hailing from Asia, the Middle East, Africa and Australia. By market segment, 76% of LDC members are non-profit organizations; 19% are commercial organization and the remaining 5% are government research labs not only in the United States but also abroad. To date, LDC has distributed more than 15,000 corpora to its members and non-member licensees.

As required by the terms of LDC's founding grant, membership fees and data sales provide funding to support the consortium's ongoing activities in the publication, documentation, maintenance and distribution of databases as well as the negotiation of necessary legal arrangements plus a small amount of new database creation. Often, databases created elsewhere must be extensively transformed to make them suitable for electronic publication; this work is also carried out at the LDC with internal financing.

LDC has maintained fruitful links with research groups around the world who both use LDC data and contribute data for distribution through LDC. LDC has partnered with ELRA on several projects, the most recent of which was the joint publication of the Translanguage English Database (LDC catalog number: LDC2002T03, ISBN: 1-58563-202-3). Hundreds of organizations in Europe are also LDC members and more than one hundred Asian organizations use LDC data.

In response to demands from its constituent research communities, LDC has expanded its role from that of a specialized data publisher to include data collection, corpus creation, and research on the use and structure of language resources. LDC staff has grown accordingly. Twenty-six regular staff members now manage the research, technical, collection, annotation, publication and customer service functions of LDC's Philadelphia office. LDC also maintains a part-time workforce that varies from 15 to 45 staff members depending upon project workload.

4. Data Publication

LDC's primary function is to publish and archive of language resources. Each year LDC publishes between 15 and 25 corpora. Outside organizations contribute about half of these asking LDC to handle final formatting, intellectual property arrangements and distribution. The other half are created by or with the help of LDC typically to support government-sponsored technology evaluation projects. At the time of writing, LDC has published 209 corpora covering about 35 different languages and spanning more than 850 CDs of data. These include 126 speech corpora (60%), 73 text corpora (35%) and 10 lexicons (5%). Some of the most recent include:

- Multiple *Topic Detection and Tracking* corpora including hundreds of hours of Mandarin and English broadcast news plus newswire richly annotated to support story segmentation, topic detection and tracking
- Treebanks in English, Chinese and Czech including the Penn Chinese Treebank and the Prague Dependency Treebank developed by Jan Hajic and his group
- Two discourse annotated corpora including the *Rhetorical Structure Theory Discourse Treebank* containing news text annotated with discourse structure in the framework of Rhetorical Structure Theory and the *CallHome Spanish Dialogue Act Annotation Corpus* containing dialogue act, dialogue game and genre annotations of Spanish telephone conversations among friends and family.
- Multiple *Speech in Noisy Environments* (SPINE) corpora including audio and careful transcriptions of dialogs from a cooperative task conducted in the presence of military noise.
- Multiple corpora of English conversational speech collected under the Switchboard-2 protocol. Many conversations conducted over cellular phones. Some are transcribed.
- The Multiple-Translation Chinese Corpus including Chinese newswire and broadcast news translated into English by multiple human and computer translators.
- Three corpora of Chinese-English parallel text: Hong Kong News, Hong Kong Laws and Hong Kong Hansards.
- Evaluation Material for the HUB-4 1997, 1998 & 1999 Broadcast News Evaluations, HUB-5 1998 Conversational Speech Evaluation, 2000 NIST Speaker Recognition Evaluation, TREC Mandarin and Spanish Evaluation and the MUC 7 Conference
- Broadcast News audio and transcripts in English, Mandarin and Czech.
- News text in English, Chinese, Arabic, Korean and Portuguese
- The Translanguage English Database (TED) Audio and Transcripts released jointly with ELRA

This is just a sampling of LDC publications since the last LREC report. For a complete listing, readers are encouraged to visit the LDC catalog at:

http://www.ldc.upenn.edu/Catalog

5. Data Creation

Over the past several years, LDC has become increasingly involved in the collection and annotation of language resources. Although this was not one of the functions originally envisioned for the consortium, the needs of several research communities for large-scale corpus creation and LDC's success at managing such efforts have combined to make this a productive partnership. The following are data creation projects currently underway or recently completed by LDC staff.

5.1. Telephone Conversation

LDC has managed three types of telephone collection projects: CallHome, CallFriend and Switchboard-2. CallHome supports large vocabulary conversational speech recognition with transcribed collections of audio and lexical resources. LDC has published CallHome corpora in Spanish, Japanese, Mandarin, English, German and Egyptian Arabic. CallFriend supports language identification research with un-transcribed audio in Arabic, Canadian French, English, Farsi, German, Hindi, Japanese, Korean, Mandarin, Russian, Spanish, Tamil and

Vietnamese. Although most of these calls have not yet been transcribed, there is a growing body of transcription for Spanish, Mandarin, Farsi, Korean and Russian to support speech recognition. The Spanish and Mandarin transcripts appear in the LDC catalog.

Switchboard-2 supports speaker identification and speech recognition development with multiple 5-minute conversations from each of several hundred speakers. The subjects do not know each other and are matched in unique pairings and given a topic by the robot operator. Since the last LREC, LDC has collected two new *Switchboard Cellular* corpora. The first one contains 254 speakers participating in multiple conversations. Many of the conversations make use of GSM type cellular telephones on one or both sides. The corpus is gender-balanced. In 2001, LDC released the audio and a separate corpus containing transcripts of 250 of the conversations. The second part targeted 210 speakers each participating in at least 10 conversations. *Switchboard Cellular 2* has not been transcribed and will not be released until 2003.

In 2002 and 2003, LDC plans to begin collecting new large and demographically balanced corpora of English, Chinese and Arabic speakers talking over the telephone with assigned topics. Updates on this project as it evolves with appear on http://www.ldc.upenn.edu/Projects

5.2. Meeting Data

To support robust recognition of conversation recorded during meetings, LDC developed the GroupTalk and GroupMeet protocols. Under the GroupTalk, an interviewer facilitates discussion among a group of 2 to 4 friends or family members who are recorded using unobtrusive microphones. The facilitator introduces a variety of topics with the goal of identifying a few that truly interest the group and engage them in extended discussions. Successful GroupTalk sessions contain relaxed and informal speech. GroupMeet targets planning meetings by identifying groups that meet regularly or were planning to hold a specific meeting and then gaining their consent to record their meeting for research purposes. During GroupMeet sessions, we deploy a variety of microphones. Speech tends to be more formal than in Group Talk.

The LDC' s meeting recording system can record 16 tracks of digital audio. The system features a mixture of wireless and far-field wired microphones. Depending upon the session, either lavalier or head-mounted microphones are used for close-mic'ing of each participant. Room microphones, including a microphone array, PZM, omni-directional and directional microphones are also used. The meeting recording system consists of a digital mixer, a multi-track digital tape recording deck, wireless microphone receivers, a microphone preamplifier, and a multi-channel digital audio computer interface. Meeting sessions are recorded as 16bit/44kHz PCM audio.

To date about 30 hours of meetings involving more than 90 unique speakers have been collected under the GroupTalk and GroupMeet protocols. Some of this material will be included in NIST's Rich Text 2002 Metadata Annotation Experiment. The remainder will be published once it has been thoroughly transcribed and annotated.

5.3. Newswire and Other Text

LDC has acquired large news text corpora in: Arabic, English, French, German, Hindi, Indonesian, Japanese, Khmer, Korean, Mandarin, Persian, Portuguese, Russian, Serbo-Croatian, Spanish, Tamil, Thai, Turkish, Ukrainian and Vietnamese to support language modeling for speech recognition and information retrieval and to build databases to support language teaching. These were collected primarily, though not exclusively, from news agencies. For most of these languages, LDC holds databases of at least 1 million words. In several cases the collections total in the hundreds of millions of words. The texts are normalized by inserting standard SGML markup and by converting the character encoding into either Unicode or the most popular national encoding for the language. LDC has published news text corpora in: Arabic, English, French, German, Japanese, Korean, Mandarin, Portuguese and Spanish. Others will be released in 2002 and subsequent years.

To support the DARPA TIDES program, LDC is produced Gigaword News Text Corpora in English, Chinese and Arabic. These are very large collections of news text, 1,000,000,000 or more words. Gigaword News Text corpora will allow robust modeling of even rare linguist phenomena. Our approach to collecting such large volumes of material is to acquire the entire archive of several large newswire providers.

5.4. Parallel Text

To support research and development in statistical machine translation, LDC has published several parallel text corpora containing original text translated into one or more languages. These corpora are aligned either at the story, paragraph or sentence level. The Bilingual Internet Text Search (BITS) developed at LDC by Xiaoyi Ma and Mark Liberman searches the Internet for parallel text in specific language pairs. While searching for Chinese-English parallel text, Ma found three sets of parallel text produced by the Special Administrative Region of Hong Kong. LDC now distributes these corpora under the names Honk Kong News, Hong Kong Laws and Hong Kong Hansards. The three have proven useful for training statistical machine translation corpora.

During our search for large news archive to satisfy the Gigaword Next Text corpora, we discovered that several major news publishers produce texts that are at least comparable in English, Arabic and German. The BITS component that identifies translation pairs over the web has recently been retooled to find translation pairs in a local archive. As a result, we have identified more then 19,000 translation pairs among the English and Chinese stories

5.5. Lexicons

Under the CallHome project, LDC created lexicons of 40,000 to 100,000 entries including information on the orthography, pronunciation, morpho-syntactic features and word frequency. The LDC catalog contains lexicons for Egyptian Arabic, English, German, Japanese, Mandarin and Spanish and lexicons for Persian, Korean and Russian are under development. In 2002, LDC will release Chinese-English and Arabic English translation lexicons.

5.6. Broadcast News

LDC has been collecting broadcast news since 1996. Early efforts supported broadcast news transcription under the DARPA HUB-4 program. In 1998, LDC began large-scale collection of broadcast news to support the DARPA sponsored Topic Detection and Tracking program. More recently, the DARPA TIDES program has evolved into an umbrella program nurturing developing in information detection, extract, summarization and translation. Each of these technology areas makes use of broadcast news transcripts and news text that LDC collects.

5.7. Annotated Corpora – TDT

During 1998 and 1999, LDC created the TDT-2 and TDT-3 corpora, to support the DARPA research program in Topic Detection and Tracking (Wayne 2000). The Topic Detection and Tracking program seeks to create core technology for a news understanding system capable of processing multi-source, multilingual multi-modal content. The languages are as diverse as Arabic, English and Chinese; the media include broadcast television and radio, newswire and WWW sites. The TDT research tasks include segmenting a stream of news into individual stories, detecting either the first or all stories associated with a new topic, tracking all stories discussing a known topic and linking stories that have a topic in common.

The TDT corpora are collections of broadcast news and newswire with multiple annotations. TDT-2 corpus contains daily samplings over a six-month period from six English and three Chinese sources including over 72,000 stories and 650 hours of recorded audio in English. TDT-3 added two English sources and extended the collection from October through December of 1998. TDT-3 includes more than 65,000 stories and 600 hours of audio.

LDC annotated the TDT corpora for topic relevance by defining 100 topics from January-June 1998 and 120 topics from October-December 1998 and identifying those stories that discusses any of the topics. The TDT corpora as delivered contain the audio of all broadcasts, the newswire text and the text transcripts of the broadcast audio as well as tables showing position of all story boundaries and tables identifying the stories that discuss each topic. The TDT-2 and TDT-3 audio and text corpora have been released. (See Cieri, et. al., 2000 for a more complete description.)

Work is underway at LDC to create the fourth TDT corpus containing 8 Chinese, 8 English and at least 3 Arabic broadcast news and newswire sources collected from October 2001 through January 2001. The corpus is expected to have approximately 61,000 stories and 650 hours of audio. Stories in all three languages will be annotated for relevance to 60 topics selected from that epoch.

5.8. Annotated Corpora – ACE

In the Automatic Content Extraction (ACE) program, LDC is helping to define a new kind of annotation and apply it to a research corpus. ACE seeks to develop technologies to detect and represent entities, relations and events in text. These technologies will support classification, filtering, and selection applications that operate on three kinds of clean and degraded text: newswire, transcripts of broadcast news generated by automatic speech recognition systems and newspaper text generated by optical character recognition systems.

In the early phases, ACE participants worked on Entity Detection and Tracking. To support this task, annotators must identify and classify entities such as persons, organizations, facilities and locations in the source data. For each entity, annotators record its name, classify it as to type and identify all mentions of it in a text. LDC was one of several groups helping to refine the annotation specification and one of three groups annotating the corpora. The data being annotated comes primarily from the TDT corpus and other LDC holdings. LDC plans to release the ACE corpus to its membership after the data has been used in technology evaluations.

5.9. Annotated Corpora – TIDES MT

Within the TIDES machine translation (MT) community, researchers recognized the need to develop a metric for scoring machine translation quality. The IBM MT group (Papineni 2002) developed the “Bleu” metric that scores a translation according to its overlap in n-grams versus an ideal translation. In order to determine whether this metric could predict human assessment, LDC prepared the Multiple-Translation Chinese English corpus. That corpus contained a little more than 100 news stories whose lengths averaged 305 Chinese characters. LDC arranged to have 11 different commercial translation bureaus translate these stories into English on a sentence-by-sentence level. We also ran the data through 6 different commercial MT systems. The output formed development-test and evaluation material for the TIDES MT dry-run evaluation.

In order to provide human assessment of translation quality, LDC recruited a team of 15 assessors who were college educated native speaker of English. This group reviewed each translation of each sentence of each story and rated them on 5-point scales according to their fluency and adequacy. Adequacy was defined as the degree to which a segment communicates all of the information transmitted by a “gold-standard” translation. The gold-standard was simply the most promising of the 11 translations created by humans. Each segment was judged by at least two independent judges. The 17 translations have been published as the Multiple Chinese-English translation corpus. The assessments along with a second Chinese and an Arabic multiple-translation corpus will be published in 2002.

6. Outreach to New Communities

LDC has recently begun to build a relationship with the research community in sociolinguistics. Much of the research in sociolinguistics begins with an empirical and quantitative analysis of patterns in corpora of spoken and written linguistic performance. Until now, the lack of standards and tools has placed barriers in the path of sociolinguists interested in sharing data electronically. However, at the NWAVE (New Ways of Analyzing Variation) conference held in Toronto in October 1999, LDC was invited to talk about its role in data distribution and possible links to the sociolinguistic research community. The LDC project to develop Data and Annotations for Sociolinguists (DASL) resulted from that meeting. DASL encourages data sharing and the re-annotation and reuse of published data as an important complement to first-hand fieldwork.

The quantitative study of linguistic variation in pronunciation, in word choice and in selection of syntactic structures is requires empirical observation and statistical description of linguistic behavior. Collecting and annotating databases is a necessary step in any quantitative linguistic analysis. The current state of computing technology encourages linguists to collect, annotate, analyze and even summarize and present analyses of linguistic behavior wholly within the digital domain. Unfortunately there is a paucity of tools to support such work. The goal of the DASL Project is to conduct a case study involving the analysis of a well-documented sociolinguistic variable as it appears (or does not) in several large-scale speech corpora. In the process of conducting this case study, LDC is developing tools and annotation standards, annotations and data sets that are being made widely available.

DASL investigates linguistic variation in four large digital speech corpora: TIMIT, Switchboard-1, CallHome American English and Hub-4 English Broadcast News, corpora created for a purpose other than the empirical study of linguistic variation. All four were created and have been used primarily to such speech engineering, especially speech recognition, speaker identification and word spotting. Because the data has already been transcribed and segmented, speaker turns can be retrieved separately. Basic demographic information (gender, age, education, region) is available in the four corpora in varying degrees of detail.

Annotators are coding the corpora for examples of linguistic variation. One such example is the so-called process of -t/d deletion whereby a phonemic [t] or [d] is expected but not pronounced in a cluster. Where –t/d deletion occurs "first" is uttered as "firs". The DASL Project begins with the analysis of -t/d deletion in English because it is a well-understood, stable variable common in multiple varieties of English and is easily coded.

DASL annotators use a tool, developed for the project that gives linguists access to the four corpora via the Internet and allows simultaneous annotation at multiple sites. In addition to the empirical study of linguistic variation among the speakers represented, this project addressed methodological issues in the corpus re-use and in team based annotation of linguistic data.

7. Other LDC Projects

It will not be possible in the space and time available, here to discuss all of the projects ongoing at LDC. However, there are several papers in this same volume that provide more complete coverage of some additional LDC activities. Steven Bird will report on his group's development of the Annotation Graph Toolkit and the graphical users interfaces built with it. Stephanie Strassel will report on annotation of time expression. Mike Maxwell will report on morphology learning and Craig Martell will report on the annotation of gesture.

8. Conclusion

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. The past two years have seen increased demand for: intensively multi-modal resources, larger data sets in high-density languages and new data in low density languages; standards and tools for corpus development and re-useable resources. Over the next two years, we anticipate demand for extensive batteries of coordinated language resources with sophisticated annotation for several major languages. Data centers like LDC and ELRA will be well placed to address these needs if they can manage to integrate new resources development with existing resources to fill known gaps. The LDC is involved in multiple efforts to address these issues, and welcomes wider collaborations.

9. 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, Mark Liberman, 2002, A Call for Open Source Lexicons, 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.

Bird, Steven and Mark Liberman, 1999, Linguistic Annotation Page, [www.ldc.upenn.edu/annotation]

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.

Cieri, Christopher, Dave Graff, Mark Liberman, Nii Martey and Stephanie Strassel, 2000, Large Multilingual Broadcast News Corpora for Cooperative Research in Topic Detection and Tracking: The TDT2 and TDT3 Corpus Efforts, In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

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

Graff, David and Steven Bird, 2000, Many Uses, Many Annotations for Large Speech Corpora: Switchboard and TDT as Case Studies, In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

Lamel, Lori, Fabrice Lefevre, Jean-Luc Gauvain and Gilles Adda, 2001, Portability Issues for Speech Recognition Technologies, HLT 2001: Proceedings of the First International Conference on Human Language Technology Research, San Diego, CA, March 18-21, 2001.

LDC, 2000, Linguistic Data Consortium Homepage [http://www.ldc.upenn.edu]

Ma, Xiaoyi and Mark Liberman, 1999, BITS: A Method for Bilingual Text Search over the Web, presented at Machine Translation Summit VII, September 13th, 1999, Kent Ridge Digital Labs, National University of Singapore, [www.ldc.upenn.edu/Papers/MTSVII1999/BITS.ps]

Maxwell, Mike, 2002, Resources for Morphology Learning and Evaluation, Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

Papineni, Kishore, Salim Roukos, Todd, Ward, John Henderson, Florence Reeder, 2002, Corpus-Based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French and Spanish Results, HLT 2002: Proceedings of the Second International Conference on Human Language Technology Research, San Diego, CA, March 24-27, 2002.

Saggion, Horacio, Dragomir Radev, Simone Teufel, Wai Lam and Stephanie Strassel, 2002, Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Multi-lingual Environment, Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

Strassel, Stephanie and Christopher Cieri, 2002, Resource Development for Topic Detection and Tracking Research: The TDT-4 Corpus, Proceedings of the Third International Language Resources and Evaluation Conference, Las Palmas, Spain, May-June 2002.

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.

Stephanie Strassel and Christopher Cieri (1999), Corpus Sociolinguistics: Issues, Data and Tools, Presented at NWAVE-28, York University, Toronto, Ontario October, 1999.
[http://www.ldc.upenn.edu/Papers/NWAVE1999/]

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

TIDES, 2000, DARPA Program in Translingual Information Detection Extraction and Summarization [www.arpa.mil/ito/research/tides]

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.

Wayne, Charles, 1998, Topic Detection & Tracking: A Case Study in Corpus Creation & Evaluation Methodologies, In Proceedings of Language Resources and Evaluation Conference, Granada, Spain, May 1998.