Issues in Corpus Creation and Distribution:

The Evolution of the Linguistic Data Consortium

Christopher Cieri, Mark Liberman

University of Pennsylvania and
Linguistic Data Consortium
Philadelphia, Pennsylvania, USA
{ccieri, myl}@ldc.upenn.edu

Abstract

The Linguistic Data Consortium (LDC) is a non-profit consortium of universities, companies and government research laboratories that supports education, research and technology development in language related disciplines by collecting or creating, distributing and archiving language resources including data and accompanying tools, standards and formats. LDC was founded in 1992 with a grant from the Defense Advanced Research Projects Agency (DARPA) to the University of Pennsylvania as host organization. LDC publication and distribution activities self-support from membership fees and data sales while new data creation is supported primarily by grants from DARPA and the National Science Foundation. Recent developments in the creation and use of language resources demand new roles for international data centers. Since our report at the last Language Resource and Evaluation Conference in Granada in 1998, LDC has observed growth in the demand for language resources along multiple dimensions: larger corpora with more sophisticated annotation in a wider variety of languages are used in an increasing number of language related disciplines. There is also increased demand for reuse of existing corpora. Most significantly, small research groups are taking advantage of advances in microprocessor technology, data storage and internetworking to create their own corpora. This has lead to the birth of new annotation practices whose very variety creates barriers to data sharing. This paper will describe recent LDC efforts to address emerging issues in the creation and distribution of language resources.

1. The Value of Language Resources

Developing realistic models of human language that support research and technology development in language related fields requires masses of linguistic data: preferably hundreds of hours of speech, tens of millions of words of text and lexicons of a hundred-thousand words or more. Although independent researchers and small research groups now have the desktop capacity to create small- to medium-scale corpora, the collection, annotation and distribution of resources on a larger scale presents not only computational difficulties but also legal and logistical difficulties to challenge most research organizations whether they be educational, commercial or governmental. While some large corporate research groups routinely engage in medium- to large-scale corpus creation, these groups typically lack the necessary distribution infrastructure; resources created at considerable cost in those environments are seldom shared outside the immediate group.

Published language resources benefit a broad spectrum of researchers, technology developers and their customers. The presence of community standard resources reduces duplication of effort, distributes production costs and removes a barrier to entry. As research communities mature, published resources are corrected, improved and further annotated. They provide a stable reference point for the comparison of different analytic approaches.

Over the past two decades, the situation for language engineers has evolved from one in which concerns over intellectual property, usage agreements, publication standards and replication costs prevented resource sharing to the current state in which the value of shared resources is widely recognized. Based on the success of the DARPA "common task" methodology and the popularity of early shared databases such as the Brown text corpus and Texas Instruments' TI 46 and TI DIGITS corpora, the LDC was created to foster the development, distribution, archiving and maintenance of language resources in electronic form.

2. The Linguistic Data Consortium

The Linguistic Data Consortium was founded in 1992 with an initial grant from the Defense Advance Research Projects Agency (DARPA) and continuing funding from DARPA and the National Science Foundation (NSF). The University of Pennsylvania serves as the LDC's host institution, providing space, infrastructure and logistical support. LDC staffers are University employees, and Penn enters into all legal arrangements on behalf of the consortium members and the research community at large. From the beginning, LDC has build fruitful links with groups in Europe, Asia and other parts of the world. In Europe, the principal partner has been ELRA but hundreds of European organizations are also LDC members. More than one hundred Asian organizations use LDC data and LDC has maintained links with the emerging GSK group in Japan.

Because progress in several areas of language technology depends upon the accessibility of the consortium's products, LDC is open to researchers around the world. Before LDC was created, an external planning committee set the membership fees that have not changed in 8 years. The membership fee for a university is roughly the cost of a new PC or attendance at an international technical meeting. The membership fee for a commercial organization is roughly the cost of a high-end workstation, certainly less than the cost to create a single small-scale corpus and an order of magnitude less than the cost of an average medium-scale corpus. As a matter of policy, no bona fide researcher is prevented from having access to LDC data by genuine inability to pay.

Organizations join LDC on a yearly basis and gain perpetual rights to all corpora produced in the years in which they join. Current LDC members also have network

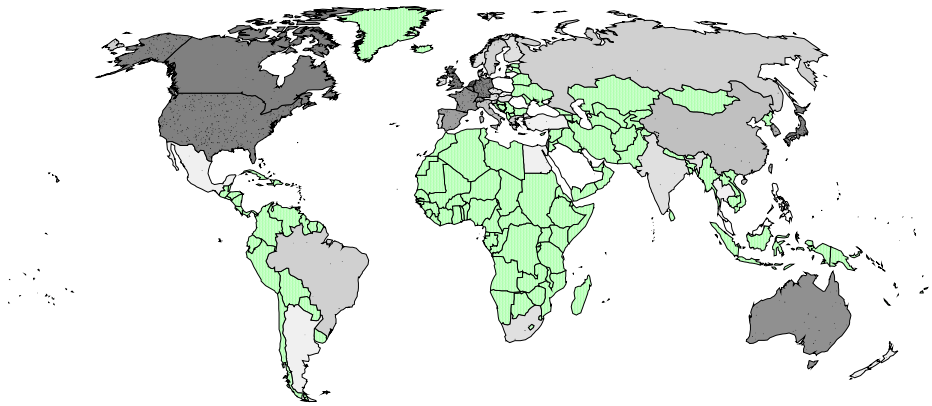

Figure 1: Organizations that use LDC by country. Marks indicate country but not exact location.

access to LDC Online, a service that facilitates browsing and searching of indexed text, speech and lexical corpora.

Since 1992, nearly 1000 organizations worldwide have used LDC data; more than 300 companies, universities and government research laboratories have joined the consortium; almost 700 others have purchased one or more corpora. 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.

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.

Responding 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 regular staffers 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 10 to 35 staffers depending upon project workload.

3. LDC Data Publication

The primary function of the LDC is the publication and archiving of data resources. LDC publishes most corpora on digital media, currently CD-ROM. The sheer volume of some corpora begs for distribution on a denser medium. LDC expects to publish its data on DVD or some future innovation once its market penetration is sufficient to be cost effective. Every corpus, unless prevented by intellectual property agreements, becomes available for network access via LDC Online as well.

3.1. Distribution on Digital Media

Each year LDC publishes between 15 and 25 corpora. About half come from outside organizations that have collected and annotated data on their own but asked LDC to assist with the final formatting, intellectual property arrangements and distribution. The other half are corpora created by or with the help of LDC. The latter typically support government-sponsored technology evaluation projects. About two-thirds of LDC's publications are speech corpora, the remaining third are text corpora and lexicons. At the time of writing, LDC had published 164 corpora including 106 speech corpora, 48 text corpora and 10 lexicons. **Some** of the most **recent** include:

- Topic Detection and Tracking Corpus – richly annotated broadcast news and newswire in English and Mandarin - described further below
- Treebank 3 - the latest update of the landmark hand-parsed corpus of written and conversational English
- Corpus of Spoken American English collected by the University of California, Santa Barbara Center for the Study of Discourse (John W. Du Bois, Director) and representing the American Component of the International Corpus of English (Charles W. Meyer, Director).
- BLLIP 1987-89 WSJ Corpus – a Treebank-style parsing of the three-year, 30 million word, Wall Street Journal archive from the ACL/DCI corpus developed by Eugene Charniak and his group at Brown University
- Speech Under Simulated and Actual Stress (SUSAS) created by the Robust Speech Processing Laboratory at Duke University (Professor John H. L. Hansen, Director)
- Taiwanese Putonghua - 40 transcribed monologues and dialogues in Taiwanese accented Putonghua gathered by San Duanmu at the University of Michigan
- American English Spoken Dictionary – containing recorded pronunciations of 50,000 of the most common English words
- JURIS the database of the Justice Department Retrieval and Inquiry System containing almost 700,000 legal documents from the 1700's through the early 1990's covering Administrative, Case, Statutory and Tax Law, plus Executive Orders, Regulations, and International Agreements, etc.
- Voicemail Corpus – over 1800 messages contributed by IBM volunteers and collected and

transcribed by M. Padmanabhan, G. Ramaswamy, B. Ramabhadran, P. S. Gopalakrishnan and C. Dunn at IBM.

- Broadcast News audio and transcripts in English, Mandarin and Spanish
- News text corpora in English, Mandarin, Japanese, Portuguese and Spanish
- Conversational audio and transcripts in Egyptian Colloquial Arabic, English, German, Mandarin and Spanish
- Pronouncing lexicons in Egyptian Colloquial Arabic, English and German

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:

www.ldc.upenn.edu/Catalog

3.2. Network Access to LDC Data

LDC's most common mode of publication has been to organize data on one or more volumes of computer-readable media in standard or *de facto* standard formats and distribute these upon request. This mode works best for research groups that already know which data sets they require and have the local infrastructure to handle the media and formats and to process the data in large quantities. Other research communities, however, take an exploratory approach, testing hypotheses on small batches of carefully selected data. For these groups and communities, LDC Online provides more useful access.

LDC Online provides network access to LDC's text, audio and lexical resources that are not otherwise restricted by intellectual property arrangements. With LDC Online, users may browse resources linearly, or search text resources by word, lemma, part of speech or any combination of these elements. Statistics such as word frequency are also available. For corpora containing audio data and transcripts, a search against the transcripts also returns a link to the audio. To facilitate network access, LDC transcripts are typically aligned to the audio in small segments (8-10 seconds) and are available in any of the currently popular audio formats.

The American English Spoken Lexicon (AESL) provides an example of audio and lexical data combining with fine-grained indexing to create an Internet resource for linguists, language teachers and others with similar needs. AESL audio files contain pronunciations of each of 50,000 of the commonest English words as counted in LDC's English corpora. AESL is freely available (http://www.ldc.upenn.edu/cgi-bin/aesl/aesl) from LDC's web site where users can either browse the lexicon alphabetically or search for words by spelling or pronunciation.

LDC encourages use of the resources in LDC-Online for research and education. Users or potential users with questions should contact ldc@ldc.upenn.edu. Visitors to our web site who are not LDC members may acquire guest accounts that permit the same kinds of electronic access to a large sample of our data including the Brown text corpus and the TIMIT speech corpus.

4. LDC 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.

4.1. Telephone Conversations

LDC has managed three types of telephone collection projects: CallHome, CallFriend and Switchboard-2.

The CallHome project supports large vocabulary conversational speech recognition by collecting, transcribing and providing lexical resources for a number of languages: Spanish, Japanese, Mandarin, English, German and Egyptian Arabic. In each case, we have recorded over 200 30-minute conversations involving pairs of native speakers, transcribed the best 10 minutes and created lexical entries including pronunciation and, where appropriate, romanization and morphological analysis for each word in the transcripts.

The CallFriend project supports research in language identification for: Arabic, Canadian French, English (from both northern and southern US states), Farsi, German, Hindi, Japanese, Korean, Mandarin (from mainland China and Taiwan), Russian, Spanish (from the Caribbean and South America), Tamil and Vietnamese. In each case, we have collected 100 5-30 minute conversations from native speaker pairs living in the continental United States, Canada, Puerto Rico and the Dominican Republic. Although most of these calls have not yet been transcribed, there is a growing body of transcription for Spanish, Mandarin, Farsi, Korean and Russian. The Spanish and Mandarin transcripts appear in the LDC catalog; LDC will publish the Farsi, Korean and Russian after their use in evaluation projects.

In 1999, LDC began to collect and transcribe Russian and Korean telephone calls. For Russian, LDC collected over 140 telephone conversations among native Russian speakers living in the United States and segmented and transcribed 15 minutes of each of 80 conversations. For Korean, LDC staffers created time-aligned transcriptions for 15 minutes of each of 100 calls originally collected under the 1996 CallFriend project.

The Switchboard-2 corpus supports research and development in speaker identification technology. In each phase we collect, on average, 10 5-minute conversations from each of several hundred American English speakers. The subjects do not know each other and are matched in unique pairings and given a topic by the robot operator. In late 1999, LDC began to collect a small corpus of conversations among cellular phone users. The collection is still underway with a goal to collect 10 conversations from each of 190 participants.

In 2000, LDC plans to continue transcribing its Farsi data and to continue collecting and begin transcribing cellular phone conversations to support both speaker identification and large vocabulary conversational speech recognition over digital, cellular phone channels.

4.2. Parallel Text

To support research and development in statistical machine translation, LDC has published several parallel text corpora in which original content was available in two or more languages. After collecting the material, LDC

staff aligned the corresponding texts either at the story, paragraph or sentence level. The UN Parallel Text Corpus contains United Nations documents in English French and Spanish. The Hansard corpus contains the English and French text of the official records of the proceedings of the Canadian Parliament.

In 1999, LDC programmer Xiaoyi Ma (1999) combined language identification and translation matching technology with Internet indexing software to search the worldwide wide for sources of parallel text. The Bilingual Internet Text Search (BITS) system has been proven to find German-English and Chinese-English bi-text and is extensible to other language pairs.

In the process of the Chinese-English search, Ma learned that the Special Administrative Region of Hong Kong distributes its regulations, news releases and parliamentary proceedings in both English and Chinese. LDC has secured permission to distribute this data to its membership and will publish the three sources in 2000. In conjunction with the TIDES project, described below, LDC has begun to search the Internet for Korean-English parallel text.

4.3. Newswire and Other Text

To date, LDC has acquired large-scale text databases in 20 languages: Arabic, English, French, German, Hindi, Indonesian, Japanese, Khmer, Korean, Mandarin, Persian, Portuguese, Russian, Serbo-Croatian, Spanish, Tamil, Thai, Turkish, Ukrainian and Vietnamese. These were collected primarily, though not exclusively, from news agencies to support language modeling for speech recognition and information retrieval and to build databases to support language teaching. For most of these languages, LDC holds databases of 1 million or more words. In several cases the collections total in the dozens or even 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: English, French, German, Japanese, Mandarin, Portuguese and Spanish. Others will be released in 2000 and subsequent years.

LDC actively encourages the creation of written corpora for languages that currently lack such resources; we stand ready to lend our expertise and assistance to facilitate the development, publication and distribution of such collections, even those that may not be of immediate commercial importance.

4.4. Lexicons

Under the Call Home project, LDC has created lexicons of 40,000 to 100,000 entries including information on the orthography, pronunciation, morpho-syntactic features and frequency of each word in the transcripts. In languages where the structure justifies it, a morphological analysis-synthesis system is included in the form of tables for a finite state transducer. The LDC catalog contains lexicons for Egyptian Arabic, English, German, Japanese, Mandarin and Spanish. As LDC accumulates additional transcripts of conversational or broadcast audio in these languages, we update the lexicons to cover the new vocabulary encountered.

In conjunction with the Persian, Korean and Russian transcription projects described above, LDC has begun to create lexicons for each of those languages.

4.5. Broadcast News

LDC began collecting and transcribing broadcast news in 1996. With the beginning of the DARPA-sponsored Hub-4 speech recognition project, the demand for broadcast news corpora increased. The LDC catalog now includes several broadcast news corpora for Hub-4 with a total of 240 hours of English and 30 hours each of Mandarin and Spanish. The research communities in speech recognition and information retrieval have benefited from LDC's collaboration with the Voice of America.

The Voice of America's charter, codified by U.S. Public Law 94-30, states that "the long-range interests of the United States are served by communicating directly with the people of the world by radio." To achieve these goals, Voice of America broadcasts news, cultural and entertainment programming around the world, around the clock and currently in more than 50 languages: Afan Oromo, Albanian, Amharic, Arabic, Armenian, Azerbaijani, Bangla, Bosnian, Brazilian Portuguese, Bulgarian, Burmese, Chinese, Creole, Croatian, Czech, Dari, English, Estonian, Farsi, French, Georgian, Greek, Hausa, Hindi, Hungarian, Indonesian, Khmer, Kirundi, Korean, Kurdish, Lao, Latvian, Lithuanian, Macedonian, Pashto, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovene, Spanish, Swahili, Thai, Tibetan, Tigrigna, Turkish, Ukrainian, Urdu, Uzbek and Vietnamese. Until recently, the Voice of America's charter had forbidden distribution of their materials within the United States, thus denying not only American researchers but also LDC members worldwide, access to this rich resource. In 1996, however, Congress enacted Public Law 104-269, which authorized Voice of America to cooperate with LDC to make VOA programming available for use in research, education and technology development.

Since 1996, the Linguistic Data Consortium has collected VOA broadcasts in English, Mandarin, Spanish and Czech for use in corpora provided for a number of federally sponsored research projects, and has also published (or will publish) the resulting corpora for general use in research and education. In 1998, using internal funding, LDC established a satellite downlink station and began collecting VOA broadcast audio and (trans)scripts to support speech recognition and topic detection and tracking. DARPA's Hub-4 speech recognition project has used broadcast news in English, Mandarin and Spanish, while DARPA's Topic Detection and Tracking project has benefited from VOA broadcasts in English and Mandarin. VOA Czech broadcasts were used as data for the NSF-sponsored summer workshop "Language Engineering for Students and Professionals Integrating Research and Education" that took place at Johns Hopkins from July 12 to August 20, 1999. VOA broadcasts data will continue to play an important role as LDC collects VOA data for Translingual Information Detection Extraction and Summarization (TIDES) program just now underway.

5. Multi-modal, Multilingual Corpora – TDT and TIDES

During 1998 and 1999, LDC created two of its most ambitious corpora to date, the TDT-2 and TDT-3 corpora, to support the DARPA research program in Topic Detection and Tracking. TDT began in 1997 with a small pilot study. In 1998 and 1999, the program expanded to include new research sites, new languages, new research tasks and enlisted the U.S. National Institute of Standards and Technology (NIST) to evaluate technology and LDC to create the data. Charles Wayne's (2000) presentation at this conference will provide a thorough overview of TDT.

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 could be as diverse as English and Chinese; the media could include broadcast television and radio, newswire, WWW sites, newsgroups, e-mail lists or some future innovation or combination. The TDT research tasks are **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 with multiple annotations. TDT-2 corpus contains daily samplings over a six-month period from two television, two radio and two newswire sources, specifically: ABC's World News Tonight, CNN Headline News, Public Radio International's The World, Voice of America English news radio and newswires from the Associated Press and New York Times services. The TDT-2 Mandarin collection includes daily samples of Voice of America Mandarin radio broadcasts and Xinhua News Service's newswire plus news stories downloaded from the web pages of the Singapore based news agency Zaobao. Over the 180 days of collection, LDC accumulated over 54,000 stories and 634 hours of recorded audio in English. TDT-3 adds two English sources: NBC's Nightly News and MSNBC's News with Brian Williams and extends the collection from October through December of 1998. TDT-3 English includes more than 35,000 stories. The Mandarin data from TDT2 and TDT3 together totals 30,000 stories. In anticipation of future extensions of the TDT-2 corpus, LDC has also collected Voice of America Spanish radio broadcasts and newswire from El Norte's news service during the same epoch. In preparation for the next generation of TDT research, LDC has continued the collection through 1999 adding CCTV Mandarin television broadcasts and Spanish television broadcasts from ECO and Univision. To date, the complete collection starts with six sources in January 1998 and continues, adding sources, so that there are 16 sources from August through December 1999.

LDC has annotated the TDT corpora for topic relevance by:

- selecting and defining 100 topics from January-June 1998 and 60 topics October-December 1998,
- reading each story in the appropriate epoch and
- indicating for each topic-story pair whether the story discusses the topic.

The TDT corpora as delivered contain:

- the audio of all broadcasts,
- the newswire text and
- the text transcripts of the broadcast audio in both
  - reference form and in
  - tokenized form with all metadata removed,
- tables showing position of all story boundaries and
- tables showing the relevance of each story to each topic.

LDC released TDT-2 in 1999 and will release TDT-3 in 2000. The paper by Cieri, et al. (2000), also presented at this conference, will focus on the TDT corpus.

The DARPA Translingual Information Detection Extraction and Summarization (TIDES) program targets technology that will, based on a query in English, locate, translate and summarize relevant documents in multiple languages. The project plans to connect technologies in machine translation, information retrieval and information extraction and summarization to build a complete translingual system. The project also seeks to enable the rapid extension of these capabilities to new languages of interest, even those for which few technical resources exist.

The data requirements for TIDES are significant. To develop the necessary technology, researchers need access to large amounts of text, especially parallel text, in multiple languages with appropriate annotation. To test the generality of TIDES components, technology evaluators will also need medium-scale annotated collections of text in low-density languages, that is, languages for which there are few electronic data resources.

In the context of the TIDES initiative, LDC has undertaken the collection of a unique, large-scale, long-term, intensively multilingual, broadcast news corpus. Since early in 2000, LDC has been collecting up to one hour of broadcast news each day in each language in which VOA broadcasts. The resulting corpus will contain more than 14,000 hours of broadcast news in 53 languages, with overlapping scripts harvested from VOA. It will provide a rich resource that can be organized, subdivided and annotated in many ways for many publications and projects. Some subsets suitable for particular uses will be published as shared data for common-task TIDES projects. LDC will archive the entire collection in its near-line storage facility enabling it to respond to researchers' requests for a specific subsample or projection of the corpus. To the extent that available disk space allows, LDC will eventually make this data available for interactive access via LDC-Online.

The VOA broadcast corpus will benefit several communities of researchers. The presence of audio and text in a multiplicity of languages will support intensively multilingual automatic speech recognition of the kind necessary for the ultimate success of technologies developed under Hub-4 and TDT as well as TIDES. The parallelism in stories based upon original news reporting in English but then translated into multiple other languages by native speakers for rebroadcast also provides a unique resource for machine translation. The scripts themselves provide natural language text suitable for many uses. Finally, the fact that all of these broadcasts focus on the daily reporting of news from a single period in time means that this corpus will provide a unique laboratory for research into translingual information tasks including topic detection, information extraction and story

summarization. LDC expects to release the first data sets based on the TIDES collection in 2001.

6. New Annotations – The TDT Corpora and the ACE Program

With advances in language engineering technology come new needs for annotated data. While LDC has collected and created time-aligned transcripts for telephone speech and broadcast news for several years, the past two years have seen an increased demand for new and more sophisticated types of annotation. To prepare the TDT corpora, LDC identified story boundaries and annotated each of tens of thousands of stories for its relevance to each of hundreds of topics selected from the corpus. This annotation required hiring a sizeable, bilingual staff, training them to annotate topic relevance consistently and developing processes to assure the quality of the final result. The paper by Strassel, Graff, Martey and Cieri (2000), also to be presented at this LREC, will discuss TDT quality control in more detail. In the TDT-3 corpus, LDC staff extended the notion of topic relevance annotation by locating pairs of stories united by a common topic and by using search engines to help identify the first story in the corpus to discuss some topic chosen at random.

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 phase one, ACE participants will work 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 is 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.

7. Reuse of Data, Standards and the ATLAS Project

The drive for efficient use of available funding encourages the reuse of existing data. The re-annotation of the Switchboard I corpus, including its orthographic transcription, partial phonetic transcription, intonational annotation, disfluency annotation, discourse structure annotation, part of speech annotation, syntactic structure annotation, word sense disambiguation is now well-known. Five different sites contributed these annotations over a period of seven years without coordination by any central authority. Some annotations introduced new structure, which was then used by others but each effort imposed various informally defined format changes and most made corrections to the underlying orthographic transcription. Joe Picone's Switchboard Resegmentation project at Mississippi State University will bring together some of these diverse and diverging Switchboard annotations. However, for the future, the field badly needs a framework for "safe" handling of such distributed annotations and corrections.

The reannotations of Switchboard, while probably the most well-known, are certainly not unique. More recently, the broadcast audio from the TDT-2 corpus has been reused in the TREC SDR (Spoken Document Retrieval) task. The TDT-2 text has been re-annotated and reused in the John's Hopkins 1999 summer workshop task on Topic-Based Novelty Detection lead by James Allan. TDT-2 is currently being annotated Entity Detection under the ACE program. The paper by David Graff and Steven Bird (2000) also presented at this conference discusses the reuse of both the Switchboard and TDT corpora. This reuse of data increases the need not only for the standardization of corpus practices and formats but also for their documentation in easily accessible, world-readable forms.

Despite the efforts of groups of researchers around the world, there remain several significant problems in the sharing and re-annotation of language data. Although, the Text Encoding Initiative (TEI) guidelines for SGML mark-up and the NIST SPHERE format for audio files are open standards with tools freely available for dealing with them, our research communities suffer from the same problems that plague other scientific database applications. Dependence upon heterogeneous storage formats, lack of abstract, data definition, incompatible tools and ephemeral citations, are familiar problems in the community of database researchers who are working on general solutions. LDC has benefited from early ties to this community. Elsewhere at this conference, Bird, Buneman and Tan (2000) report on their efforts to develop a query language based on paths through the acyclic, unrooted annotation graphs Bird and Liberman have proposed to structure corpora of time-sequenced linguistic data.

Improved standards will make it easier to create simple, inexpensive (or free) tools for linguistic database creation and use, which can be made widely available to researchers and students. A considerable investment is required to create a good transcription environment (for example), or a good system for searching speech databases based on patterns in transcriptions, or a good system for retraining specialized language models or acoustic models. With improved abstractions for overall database structure, new programming will not be necessary in order to create or access new databases, and such tools can be used more effectively and more widely. Pluralistic sets of inter-operating tools, with development shared widely by the research community through sharing of source code or through clear interface definitions, have already begun to appear, and should have an enormous impact over the coming few years. Although the LDC is not primarily a software development or distribution organization, we do distribute software that we have created for our own use, and we intend to participate more actively in such developments in the future. The presentation by Geoffrois, Barras, Bird and Wu (2000), at this same conference, will describe the Transcriber tool, fruit of the collaboration between LDC and researchers at the French Ministry of Defense, and a good first example

free, useful, extensible software resulting from the creation of open standards.

In addition to collaborating on the development of Transcriber, LDC has joined forces with NIST and MITRE Corporation to promote the idea of corpus standards and shareable annotation components under the ATLAS project (Architecture and Tools for Linguistic Analysis Systems). In 1999, the trio began exploring the feasibility of using the annotation graph formalism proposed by Bird and Liberman (1999) to integrate existing and future language resources. By inserting a layer of abstraction between the physical representation of data on disk and the interfaces used to visualize and analyze it, ATLAS will increase portability of and useful lifetime of language resources. The architecture will facilitate rapid corpus development, data exchange and reuse as well as rapid application prototyping and evaluation through the use of modular software components. The ATLAS architecture will be used in the DARPA TIDES project. The paper by Bird, Day, Garofalo, Henderson, Laprun and Liberman (2000), elsewhere at this conference, will report on ATLAS' progress to date.

8. New Communities - The TalkBank Project

Although the support of pre-competitive research and development in speech and language technology was our first and largest task, LDC has, from the beginning, had productive relationships with linguists, psychologists, clinicians and others interested in the study of language and speech. The widespread availability of new tools for creation and use of language-related data, along with increasingly affordable networked computer power and mass storage, will naturally bring new kinds of researchers into the group of those who prepare, publish and use speech and language databases. Sociolinguists, anthropologists, psychologists, ethnomusicologists, educational researchers, historians and others have already begun this process. Unfortunately, the evolution of clear guidelines for the use of language data lag behind the technological developments that make that use possible. Individual authors and large data centers alike must negotiate individually with each news agency, broadcaster and publisher. And while advances in internetworking bandwidth support the accumulation of small or medium quantities of text data or an extended period of time, they are inadequate for the timely distribution of large quantities of audio or video data. Because the potential advantages of collaboration both to these new entrants and to our current communities are enormous, we feel strongly that the process should be fostered and encouraged.

The TalkBank project funded by the National Science Foundation joins researchers from Carnegie Mellon University and the University of Pennsylvania with an interest in promoting fundamental research on communicative behavior. TalkBank will provide tools to help scholars in multiple research communities acquire, annotate, access and analyze primary linguistic data including audio, video and transcripts, via computer networks.

At the time of writing, TalkBank has established contact with representatives from more than a dozen disciplines focused on human or animal communication. TalkBank tasks currently underway include:

- organizing workshops to solicit ideas about both the core functionality that unites these disciplines and assumptions and modi operandi that make them unique.
- developing a basic architecture and tool set to support annotation of linguistic data in general
- building tools to facilitate data exchange among researchers and research communities
- developing task specific interfaces built from basic components as proof of concept

Readers who are interested in the TalkBank project are encourage to visit www.talkbank.org.

9. Other LDC Projects

For the past several years, the LDC has furnished document collections to a consortium of U.S. Government language schools. LDC is now beginning to make these collections available to all researchers. LDC has also begun collaboration with language teachers and researchers at the University of Chicago to explore ways in which our telephone conversation corpora may support teaching and learning languages, especially via the Internet.

LDC has recently begun to build 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. Strassel and Cieri (1999) report on that work.

LDC seeks to stimulate the growth of linguistic resources throughout the world. Often these resources will be developed primarily through local effort but there be occasions in which a more active LDC role will be justified at least temporarily. The African Language Resource Council (ALRC) is a good example. It is a collaborative venture between the LDC and the Center for African Studies of the University of Pennsylvania. The ALRC, directed by Dr. Yiwola Awoyale of the LDC and of the University of Illorin, will facilitate the creation and the publication of materials for the study of African languages, in particular dictionaries, grammars and texts. These resources, which will be published both in electronic and paper form, should be of considerable value for researchers and students around the world. The first publication resulting from this collaboration, a dictionary of Yoruba is due to be published in 2000.

10. Conclusion

Recent developments in the creation and use of language resources demand new roles for international data centers. International data centers such as LDC and ELRA have been well positioned over the past few years to support the needs of their research communities. However, shifts in both the supply of and the demand for language resources are impacting these organizations. The next few years will see an increased demand for:

intensively multilingual resources covering both high-density and low-density languages; flexible standards and tools for distributed corpus development, maintenance, and reannotation; reuse of existing resources; outreach to new research communities and the extension of existing mechanisms to deliver slices or projections of corpora where bandwidth-intensive media are concerned. The LDC is involved in collaborative efforts to address these issues, and welcomes wider collaborations.

11. References

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

Bird, Steven, David Day, John Garofalo, John Henderson, Christophe Laprun and Mark Liberman, 2000, ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation, In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

Bird, Steven, Peter Buneman and Wang-Chiew Tan, 2000, Towards a Query Language for Annotation Graphs, In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

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

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.

Geoffrois, Edouard, Claude Barras, Steven Bird and Zhibiao Wu, 2000, Tanscribing with Annotation Graphs, In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

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.

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]

Picone, Joe, 2000, Switchboard Resegmentation Page:
[http://www.isip.msstate.edu/projects/switchboard/index.html]

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]

VOA, 2000, Voice of America page, [www.voa.gov]

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.