THE CREATION, DISTRIBUTION AND USE OF LINGUISTIC DATA:

THE CASE OF THE LINGUISTIC DATA CONSORTIUM

Mark Liberman
Christopher Cieri
University of Pennsylvania

ABSTRACT

The Linguistic Data Consortium (LDC) is an open consortium of universities, companies and government research laboratories. It creates and distributes speech and text databases, lexicons and other resources. The University of Pennsylvania is the LDC's host institution. The LDC was founded in 1992 with a grant from the Defense Advanced Research Projects Agency (DARPA). Currently, all LDC publication and distribution activities are self-supporting, while new data creation is partly supported by grant IRI 9528587 from the Information, Robotics and Intelligent Systems division of the National Science Foundation (NSF). The LDC's core mission remains the support of pre-competitive research and development in speech and language technology, but support of other language-related research is also an important focus.

INTRODUCTION

In order to build realistic models of human speech and language, researchers need access to masses of linguistic data: speech, text, lexicons and grammars. The process of creating, maintaining, expanding, modifying and distributing databases of the necessary size presents logistical, legal and computational difficulties that tax the resources of large corporations, and may often exclude universities and smaller commercial research groups.

Publication of linguistic resources benefits the entire research community by effectively sharing production costs, avoiding duplication of effort, and lowering start-up barriers. Published resources also provide the basis for replication of published research results, establishing a stable reference point against which different analyses or algorithms can be compared. Finally, published resources can be corrected, improved and further annotated, to the benefit of the entire community.

A decade ago, most linguistic resources were beyond the reach of most researchers. Concerns over intellectual property rights, the need to negotiate usage agreements, the lack of publication standards and the cost of equipment, media and labor for replication meant that researchers who invested time and money in developing databases for their own use seldom circulated the data beyond their laboratory.

Over the past two decades, a few key exceptions to this tendency showed the value of the publication of linguistic resources. The Brown text corpus has been used by so many researchers that it has become a *de facto* standard for modeling English text. The TI 46 and TI DIGITS corpora, created by Texas Instruments at the beginning of the 80's, and later distributed by the National Institute of Standards and Technology (NIST) in 1982 and in 1986 demonstrated the importance of sharing resources in the evaluation of linguistic technology.

In 1986, the Defence Advanced Research Projects Agency (DARPA) began to use a "Common Task" methodology in its speech research programs, creating a series of shared databases for the development and evaluation of algorithms. This approach led to rapid and sustained progress in speech recognition, and has since been applied to full-text information retrieval, speech understanding and machine-assisted translation.

Based on the success of such models, the LDC was founded in 1992 in order to foster the development, publication, distribution and maintenance of linguistic resources in electronic form. The LDC is an open consortium that can count among its members more than 225 companies, universities and government research laboratories. Individual databases have also been sold or donated to 520 non-member institutions. The University of Pennsylvania serves as the LDC's host institution, providing space and logistical support. LDC staffers are Penn employees, and Penn enters into all LDC-related legal arrangements on behalf of the consortium members and the research community at large.

Since 1995, all consortium activities have been entirely supported by membership fees and data sales, as was required by the terms of the LDC's founding grant. Such activities include publication, documentation, maintenance and distribution of databases, the negotiation of needed legal arrangements, and a small amount of new database creation. Often, databases created elsewhere must be extensively transformed to make them suitable for electronic publication, and this work is also carried out at the LDC with internal financing.

Most new databases are created (at Penn or elsewhere) by specific project funding from various national governments, with the LDC playing the role of publisher. Continued government funding for new database creation, with continued LDC financing of publication and distribution, has been a productive partnership for all concerned.

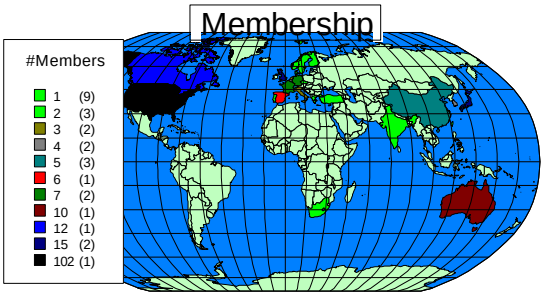
Since progress in several areas of language technology depends upon the accessibility of the consortium's products, LDC membership is open to researchers around the world, with membership fees fixed at reasonable levels. The membership fee for a university

is set at roughly the cost of a new PC or attendance at an international technical meeting. The membership fee for a company is set at roughly the cost of a high-end workstation. These fees were determined by an external planning committee that established the LDC's terms of reference before its founding, and have not been raised since 1992. As a matter of policy, no bona fide researcher is prevented from having access to LDC data by genuine inability to pay.

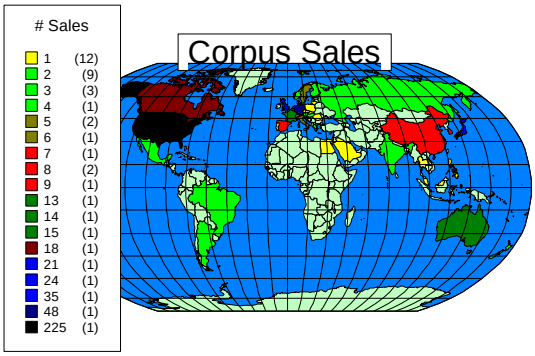
From the beginning, the LDC has had fruitful links with groups in Europe, Asia and other parts of the world. In Europe, the principal partner has been ELRA.

MEMBERSHIP AND PUBLICATIONS

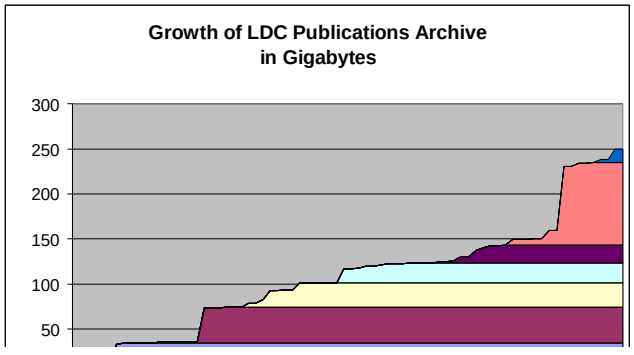
Since its foundation, more than 225 organizations have joined LDC. Among those, 61 are European.



An additional 520 organizations have obtained one or more LDC corpora as non-members. Of these non-member organizations 207 are from Europe.



Since 1993, LDC has published 110 electronic databases, comprising some 360 CD-ROMS and 20 packages to be downloaded over the Internet (this does not include republication or new editions). About half of these databases have been published on behalf of other individuals or organizations. Another 200 gigabytes of data are in the pipeline for publication this year.



ELECTRONIC ACCESS TO THE LDC: LDC ONLINE

All of the text, audio and lexical resources of the LDC that are not otherwise restricted by intellectual property arrangements are available for electronic access over the Internet. These resources are indexed to allow not only linear browsing but also searches by word, lemma, part of speech or any combination of these elements. Statistics such as word frequency are also available.

We strongly encourage use of this on-line resource for research and education. Users or potential users with questions or special needs 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. Other resources, such as Switchboard, will soon be added to this set.

RECENT DATA CREATION AT THE LDC

These are government-sponsored data creation projects currently or recently carried out at the University of Pennsylvania by LDC staff.

Telephone Conversations

The LDC has managed three recent sponsored projects to collect and annotate telephone data: Call Home, Call Friend and Switchboard-2.

Call Home is a multilingual corpus designed to facilitate large vocabulary conversational speech recognition. To date, we have collected data in six languages: Spanish, Japanese, Mandarin, English, German and Egyptian Arabic. For each language, we have recorded and transcribed a minimum of 200 conversations. Participants, who are native speakers of the language of the study, are given one free, 30-minute phone call to the person of their choice provided that person is also a native speaker of the language of the study. The best 10 minutes of each conversation are then selected and transcribed. In the case of Spanish and Mandarin, larger portions of many calls were also transcribed.

The Call Friend corpus facilitates the development of technology in language identification. At least 100 calls are collected in each of the following languages: Arabic, Canadian French, English, Farsi, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. For English, Mandarin and Spanish, we have collected 100 calls in 2 distinct dialects. The conversations in the Call Friend corpus took place in the United States, Canada, Puerto Rico and the Dominican Republic; conversations lasted between 5 and 30 minutes. Although the majority of the calls in the Call Friend

corpus have not been transcribed, 120 30-minute calls in Spanish and as many in Mandarin have been fully transcribed and will be published.

The Switchboard-2 corpus supports the development of speaker identification technology. Each phase of the corpus includes up to 3,700 conversations among up to 700 participants speaking American English, mainly college students, with each phase focusing on a different dialect region. On the average, 10 conversations are collected for each participant. Each conversation lasts 5 minutes and pairs two people who do not know each other discussing a topic suggested by a robot operator. The participant initiating the call must use a different telephone each time. Participants receive the calls at their home, or place of employment, following a schedule they provide. Two phases of Switchboard-2 have been collected to date.

Broadcast News

The LDC began collecting broadcast news in 1995, building, in collaboration with NIST, a small experimental corpus based on broadcasts of KUSC Marketplace. This corpus helped launch "Hub-4" of the DARPA-funded speech recognition project.

In 1996, LDC transcribed 110 hours of various English-language broadcast news sources. This collection was used in the continuation of the Hub-4 project. Like most linguistic resources, it turned out to have uses beyond those it was originally designed for. For example, the STREC (Spoken Text Retrieval Conference) project used it. In 1997, another 100 hours of English broadcast materials were added, as well as smaller corpora (30-40 hours) of broadcast materials in Mandarin and Spanish.

The LDC has also created a database of 200 million words of English collected from commercial transcriptions of broadcast news, soon to be available to LDC members.

At the end of 1996, the United States Congress passed public law 104-269 authorizing the United States Information Agency (USIA) to cooperate with LDC in making USIA broadcast materials available for research and teaching within the United States. The relevant programs of USIA include Voice of America and its 6 sister stations broadcasting in 53 different languages. We have established a satellite downlink station receiving VOA's broadcasts, and we are currently refining processes to allow us to accumulate VOA scripts in electronic form, and associate them with the corresponding broadcast segments. This pilot work is being carried out with LDC internal funding.

Lexicons

For each of the six languages in the Call Home project, we have created a lexicon that combines information on the orthography, pronunciation, morpho-syntactic features and frequency of the words. Each of the lexicons contains between 40,000 and 100,000 words, chosen as a function of their frequency. In languages where the structure justifies it, a morphological analysis-synthesis system is included in the form of tables for an FST (finite state transducer).

As we accumulate new transcribed speech materials in these languages, for example broadcast data, the lexicons are updated to cover the new vocabulary encountered. For most types of spoken material in these languages, current OOV rates are 2% or less.

Texts

The LDC has acquired text databases in more than 15 languages. For most languages, LDC holds databases of 1 million or more words. In several cases the collections total in the dozens or hundreds of millions of words. The texts (primarily from news agencies) are normalized by converting their character encoding into a standard form and inserting a standard form of SGML markup.

Most of LDC's text acquisitions have been funded from projects on language modelling for speech recognition, information retrieval and message understanding and for databases designed in support of language teaching. We actively encourage the creation of written corpora for languages that currently lack such resources, and we are ready to lend our expertise and assistance to facilitate the development, publication and distribution of such collections whether or not they are of immediate commercial importance.

Mixed Media Collection - TDT

During 1998, we are collecting data for the second phase of a sponsored project on Topic Detection and Tracking (TDT-2). This project aims at technology that can monitor news streams across multiple media and in multiple languages. These streams might be newswire, radio, television, internet sites, or some future innovation or combination. The technology should be able to analyze the data streams as needed, performing speech recognition and image analysis for those sources that require it, and assigning stories or story segments to a dynamic set of topics. New topics must be recognized as they arise.

The data for TDT-2 is based on collecting six news feeds for six months. Each week we collect 15.5 hours of broadcast television, 15 hours of broadcast radio and 560 randomly selected stories from each of two newswire services. The broadcast sources are segmented and transcribed. 100 topics are chosen over the six-month period, and each of the roughly 50,000 stories is tagged for relevance to each of these topics.

The (roughly 800 hours of) broadcast audio are also processed by a speech recognition system to produce ASR transcripts at a roughly 50% overall word error rate. These transcripts then form the basis for realistic exercises in segmentation, topic detection, and topic tracking.

OTHER PROJECTS

The LDC has financed the creation of several types of databases that are not part of government

sponsored research projects. Among these are the telephone collection projects: Macrophone, VAHA (Voice Across Hispanic America) and Phonebook.

LDC has established several databases in collaboration with other organizations. For example, UNIPEN is a collections of training and test materials for handwriting recognition, established in collaboration with several universities, industrial researchers and with NIST.

Another collaboration currently underway is the transcription of the TED (Transnational English Database) corpus undertaken jointly with ELRA.

Support for Language teaching

The LDC furnishes collections of multilingual documents to a consortium of U.S. Government language schools. These documents will be published by the LDC in the form of databases available to all researchers. We are in the midst of exploring how the resources of the LDC could be of more general use in teaching and learning languages, especially via the Internet.

OTHER ACTIVITIES OF THE LDC

LDC runs a series of small workshops to examine and resolve practical issues that arise in the course of its activities. Examples include the organization of pronouncing dictionaries, the segmentation of text in Mandarin and Japanese, the transcription of Egyptian Arabic and the semantic annotation of texts.

Finally, the 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, is still under development. The goal of the ALRC is to 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.

CURRENT ISSUES

We would like to draw attention to current issues in two related areas: the need for improved standards and tools, and the opportunity to broaden the research community.

The LDC is involved in collaborative efforts to address these issues, and welcomes wider collaborations.

Standards and tools

Thanks to years of work by researchers around the world, there are reasonable standards for representing many sorts of linguistic data in electronic form. For instance, all LDC text data is released with SGML mark-up, following the guidelines of the Text Encoding Initiative (TEI) where feasible, and all LDC speech data is published with NIST headers. These are open standards, and free tools are widely available for dealing with them.

However, there remain several very significant problems. Most of these are exactly the same as the problems that arise in other database applications, especially scientific databases, and we regard it as important to build bridges to the community of database researchers who are working on general solutions.

First, SGML and TEI mark-up conventions are flexible enough to permit the same material to be structured in a variety of different ways. This can arise because of inheritance from different underlying sources, because of different initial purposes for the particular database in question, or just because of random variation. This flexibility is necessary and good in itself, but it creates difficulties when a program must deal with several different databases at once. The solution is not stricter formal standards – exactly comparable problems of heterogeneity arise in the case of pure relational databases.

Second, we do not have good abstractions for representing the overall structure of speech or text databases as a whole. The standards currently used apply at the level of individual speech files, individual transcriptions, etc., not at the level of large collections of thousands of such objects with a variety of associated tables of side information. The problem of heterogeneity arises even more strongly at this higher level, with the added problem that the structure of individual databases is at best informally defined.

Third, it is increasingly common for layers of annotation to be added over time to existing databases, in a way that causes problems for maintaining connections. For instance, (some or all of) the Switchboard database of conversational speech has been given orthographic transcriptions, disfluency annotation, discourse structure annotation, part-of-speech annotation, syntactic structure annotation, word sense disambiguation, phonetic transcription, and intonational annotation. This work took place at five sites over a period of seven years, and were not planned or coordinated by any central authority. Some of these annotations have introduced new structure, which was then used by others – for instance, the discourse structure annotation used a new phrasal segmentation created by the disfluency annotation. Each of the annotation efforts imposed various informally defined format changes, and most of them also made sporadic to extensive corrections in the underlying orthographic transcription. Some of the annotations were based on the first edition, and other on the second edition, which corrected various errors, including some involving speech files.

In collaboration with Joe Picone at Mississippi State University, we plan to bring all of these diverse and diverging Switchboard annotations together into a consistent structure. For the future, the field badly needs a

framework for "safe" handling of such distributed accretion of annotations and corrections.

A closely related issue is the need to maintain durable citations for (small or large) pieces of such databases. As such databases evolve over time, we need to be able to maintain the validity of citations, whether of isolated examples or of large subcorpora used in training and testing of algorithms.

Finally, 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.

Broadening the research community

The LDC's core mission is to provide linguistic resources in support of pre-competitive research and development in speech and language technology. However, from the beginning, the LDC has also had a productive relationship 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. The potential advantages to both sides – the new entrants and our current community – are enormous, and we feel strongly that the process should be fostered and encouraged. It is especially important, in our opinion, to involve linguists who do primary description

There are also some essentially new types of speech- and language-related databases that are needed by diverse scientific research communities. Examples include video data for sign language and gesture studies, and physiological data (such as articulatory data and data from functional brain imaging studies). In each case the data are expensive and effortful to compile, and in each case there are important benefits to be gained from publication of the basic underlying data, not just the particular statistical summaries or examples that may be extracted for publication by the authors. We believe that the experience of the community of speech and language technologies researchers, as represented by organizations like LDC and ELRA, will be useful to these other communities as they learn how to publish and distribute their basic data.

ONLINE REFERENCES

LDC Homepage: http://www.ldc.upenn.edu

Language Resources primer:
http://www.ldc.upenn.edu/myl/LR_background.html

Report from 1997 Language Resources Workshop:
http://www.ldc.upenn.edu/myl/LR_report.html

A free transcription tool (demonstrated at LREC)
http://www.etca.fr/English/Projects/Transcriber