The Linguistic Data Consortium

Mark Liberman
University of Pennsylvania

John Godfrey
Texas Instruments Speech Research Group

1 Background

There is increasing interest in computer-based linguistic technologies, including speech recognition and understanding, optical and pen-based character recognition, text retrieval and understanding, and machine translation. In each area, we have useful present-day systems and realistic expectations of progress.

However, because human language is so complex and information-rich, computer programs for processing it must be fed enormous amounts of varied linguistic data—speech, text, lexicons, and grammars—to be robust and effective. Such databases are expensive to create and document, with maintenance and distribution adding additional costs. Not even the largest companies can easily afford enough of this data to satisfy their research and development needs. Researchers at smaller companies and in universities risk being frozen out of the process almost entirely.

For pre-competitive research, shared resources also provide benefits that closely-held or proprietary resources do not. Shared resources permit replication of published results, support fair comparison of alternative algorithms or systems, and permit the research community to benefit from corrections and additions provided by individual users.

Until recently, most linguistic resources were not generally available for use by interested researchers. Because of concern for proprietary rights, or because of the additional burdens of electronic publication (which include preparation of a clean and well-documented copy, securing of clear legal rights and drafting of necessary legal agreements, and subsequent support), most of the linguistic databases prepared by individual researchers either have remained within a single laboratory, or have been given to some researchers but refused to others.

A few notable examples over the years have demonstrated the value of shared resources, but until recently, these have been the exceptions rather than the rule. For example, the Brown University text corpus ([KF67]) has been used by many researchers, and was cited in [BDD$^+$92] as a generally-available standard against which statistical language models for English can be tested. The importance of shared data for evaluation of speech technology was shown by the TI-46 ([DS81]) and TI DIGITS ([LD84]) databases, produced at Texas Instruments in the early 1980's, and distributed by the National Institute of Standards and Technologies (NIST) starting in 1982 and 1986 respectively. The U.S. Defense Department's Advanced Research Projects Agency (ARPA) began using a "common task" methodology in its speech research program in 1986, creating a series of shared databases for algorithm development and evaluation. This approach led to rapid progress in speech recognition, and has since been applied to research in message understanding, document retrieval, speech understanding, and machine translation.

Building on these successes, the Linguistic Data Consortium (LDC) was founded in 1992 to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. Based at the University of Pennsylvania, the LDC is a broadly-based consortium that now includes about 65 companies, universities, and government agencies. An initial grant of $5 million from ARPA amplifies the effect of contributions (both of money and of data) from this broad membership base, so that there is guaranteed to be far more data than any member could afford to produce individually. In addition to distributing previously-created databases, and funding or co-funding the development of new ones, the LDC has helped researchers in several countries to publish and distribute databases that would not otherwise have been released.

The operations of the LDC are closely tied to the evolving needs of the research and development community that it supports. Since research opportunities will increasingly depend on access to the consortium's materials, membership fees have been set at affordable levels, and membership is open to research groups around the world. At the same time, a significant fraction of the consortium's budget comes from its membership fees. Government investment in LDC database development is continuing, but the LDC's ability to support its activities through membership fees provides a crucial sign of its value to the community it serves. These fees are now adequate to support the central staff organization, pay database publication costs, and underwrite some database creation as well.

Several parallel or related efforts are underway in Europe and the Far East. Productive relationships are being developed between the LDC and these activities, on the principle of open access across continental boundaries to the raw materials of technological progress. The road to the future of linguistic technology is (so to speak) paved with data, and the LDC will serve as part of an international highway system, providing a common infrastructure necessary for world-wide progress in research and development.

2 Linguistic Data: its Value and its Cost

What do we mean by *linguistic data*? What is it good for? Why do we need to spend money to get it?

The core of the problem is the complexity and richness of human language. There are many languages, each containing many words, which combine into messages in intricately restricted ways. Each word, in turn, corresponds to many sounds, depending on the surrounding words, on the speaker's age and sex and dialect, and on style, nuance and setting. It takes a lot of human experience to learn a language, and it takes a lot of data to "teach" one to a computer.

This is true even if we try to pre-digest the data, and "teach" the computer just a list of rules and exceptions—the list winds up being a very long one. However, experience suggests that it is better to retain as much as possible of the rich structure of actual linguistic experience. Language and speech are usually ambiguous in ways that we never even notice. We talk and listen, read and write as if our speech and text were perfectly clear, oblivious to the intricate web of uncertainty that plagues a computer program trying to imitate our behavior. The most effective way to reduce this uncertainty—and the bizarre, inhuman errors it produces—is to furnish the computer with a great deal of information about what ordinary human language is usually like.

This process, which applies at every level of linguistic analysis, is easiest to explain in the case of words in text. For instance, both *last* and *lost* are adjectives in English, and thus could modify the noun *year*, but *last year* occurs in news-wire text more than 300 times per million words, while *lost year*, although perfectly well-formed and even sensible, is vanishingly unlikely. What is usually *lost* is *ground*, *souls*, *productivity*, or *wages*, while *ground*, if not *lost*, is likely to be *high*.

These stereotypical connections are amusing, but there is a serious point: an optical character recognition (OCR) system, unsure whether a certain letter is *o* or *a*, can very safely bet on *a* if the context is *l_st year*, but on *o* if the context is *l_st souls*. A more complete set of such expectations about local word sequences can greatly reduce the effective uncertainty of a letter in English text.

The standard ASCII code allows 2^8 or 256 possibilities for each letter; in this case, we say our *perplexity* (in a technical sense) is 256. Allowing for the differing frequencies of the different characters in ordinary text reduces our perplexity to about 2^5, or 32. In 1951, Claude Shannon estimated the true perplexity of English text at about $2^{1.25}$, or 2.4 possibilities per character ([Sha51]), based on an analysis of human guessing. Mathematical models based on predictions from the frequencies of three-word sequences *known as trigrams* are currently achieving perplexities fairly close to Shannon's estimate, around $2^{1.75}$, or about 3.4 possibilities per character ([BDD$^+$92]).

The effect of this reduction in uncertainty is to make a recognition task—such as OCR—many times easier, with a correspondingly large improvement in system performance. The same technique—improving performance by reducing uncertainty about word sequences—plays a crucial role in most speech recognition applications, and can also be used to store text data in a minimum amount of space, or to transmit it in a minimum amount of time. This is one of the simplest examples of the value of linguistic data in improving the performance of linguistic technologies.

A model of this kind needs tens or even hundreds of millions of words of text to derive useful estimates of the likelihoods of various word sequences, and its performance will continue to improve as its training set grows to include billions of words. To put these numbers in perspective, consider that a typical novel contains about a hundred thousand words, so that we are talking about the equivalent of hundreds or even thousands of novels. It is not easy, even today, to obtain this much text in computer-readable form.

In addition, different sorts of text have different statistical properties—a model trained on the the Wall Street Journal will not do a very good job on a radiologist's dictation, a computer repair manual, or a pilot's requests for weather updates. The sequences input to a pen-based or speech-based interactive system when a user is entering a business letter will be quite different from those that a desk calculator or a spreadsheet sees. This variation according to style, topic and application means that different applications benefit from models based on appropriately different data—thus there is a need for large amounts of material in a variety of styles on a variety of topics—and for research on how best to adapt such models to a new domain with as little new data as possible.

This same topic-dependent variation can be used to good advantage in full-text information retrieval, since words and phrases that occur unusually often in a document tell us a lot about its content. Thus the ten most unexpectedly-frequent words in a book entitled *College: the Undergraduate Experience* are **undergraduate, faculty, campus, student, college, academic, curriculum, freshman, classroom, professor**; we are not surprised to learn that **quilt, pie, barn, farm, mamma, chuck, quilting, tractor, deacon, schmaltz** are the ten most unexpectedly-frequent words in a novel with a rural setting, or that **dividend, portfolio, fund, bond, investment, yield, maturity, invest, volatility, liquidity** characterize *Investing for Safety's Sake*, while *The Art of Loving* yields **motherly, separateness, love, fatherly, paradoxical, brotherly, faith, unselfishness, erotic, oneself**.

Of course, there is much more to text structure than just counts of words or word sequences. For instance, in analyzing the way that words go together into sentences, we can take account of the typical connection between verbs and their subjects, objects, instruments, and so on. Thus if we ask (based on a few million words of Associated Press news-wire text) what verbs have a special affinity for the noun *telephone* as their object, the top of the list is **sit by, disconnect, answer, hang up, tap, pick up, be by**. Such "affinity measures" among words in phrases can be used to help resolve the otherwise-ubiquitous ambiguities in analysis of text structure, so that *he sat for an hour by the hall telephone* is suitably differentiated from *he sat for a portrait by the school* *photographer.*

Text, diverse and ambiguous as it is, is simple and straightforward compared to the universe of speech. Here the comfortable simplicity of the alphabet is replaced by continually-varying, limitlessly-varied sounds, modulated by processes belonging to physics, physiology, and sociology alike. We have a long way to go to reach entirely adequate models of human speech, but the foundation of all progress so far has been the careful fitting of appropriate models to large amounts of data. Through this process, the model is "trained" to incorporate as much of the language's sound patterns as the model's structure and the amount and quality of data permit. For the process to work well, the training data must reflect the expected properties of the task, so that words (or the sounds that make them up) must be pronounced by enough different kinds of people in enough different kinds of messages to sample the real variability of the final application.

Properly designed, such models can then be used to recognize, synthesize or encode speech, and their performance can be evaluated quantitatively on more data of the same sort they were trained on. Improvements in performance can come in three ways: better models, better fitting techniques, or more data. Usually, experiments with new models and new fitting techniques also require new data to be carried out properly. Thus to a large extent the pace of progress in speech technology, especially in the area of speech recognition, has been determined by the rate at which new speech data has become available.

Common sense and experience alike specify the benefits of data-driven research in linguistic technology: it gives us the basis for modeling some of the rich and intricate patterns of human speech and language, by brute force if no better way is devised; it permits quantitative evaluation of alternative approaches; and it focuses our attention on problems that matter to performance, instead of on problems that intrigue us for their own sake.

3 The Future of Linguistic Technology

In discussions about future applications of linguistic technology, it sometimes seems as if the aim is a wholesale replacement of the keyboards and mice of current interactive computers by means of speech or handwritten input, and perhaps the replacement of display screens by means of voice output as well. Although there are doubtless many circumstances in which spoken conversation with a computer is the right solution, and many other circumstances in which pen-based input is the right approach, there are still keyboards and (especially) screens in any future that we can see clearly today. Understanding the role of linguistic technology in the future of our society requires consideration of some larger issues.

We humans spend much of our lives speaking and listening, reading and writing. Computers, which are more and more central to our society, are already mediating an increasing proportion of our spoken and written communication– in the telephone switching and transmission system, in electronic mail, in word processing and electronic publishing, in full-text information retrieval and computer bulletin boards, and so on.

Because information storage and processing is getting cheaper by a factor of a thousand or so every decade, we know that computer technology will continue to penetrate and reshape society for some time to come, as today's expensive laboratory curiosities become tomorrow's electronic commodities.

These trends create an enormous economic and social opportunity for natural language and speech technology. Where computers are already involved in creating, transmitting, storing, searching, or reproducing speech and text, we have the chance, at little marginal cost, to add new features that improve the quality of the process or that increase the productivity of the human labor involved. Simple examples of this kind include the use of spelling correctors in word processing; the use of speech technology to reduce the workload of telephone attendants, by screening calls with voice recognition or providing information through voice synthesis; and the use of machine-aided translation (MAT) programs, which make human translators more productive by providing a rough draft to be corrected. In other cases, linguistic technology improves on other solutions–for instance, hands-free voice control of industrial inspection stations is often faster and more accurate than alternative methods.

It is easy to project future extensions of such linguistic technology, such as the human-like voice communications of computers in *2001* and *Star Trek*; but predicting the details of life even twenty years from now raises questions about the development and social acceptance of many complex technologies. What will be the roles of input/output methods such as voice, video, keyboard, mouse, pen-based systems, direct interpretation of gesture and gaze? To what extent will information flow as full text, as fielded records, as sound, as images? How will computer networking, telephone systems and cable TV divide up the world of information transmission? In what directions will wireless telecommunications develop? We don't know for sure—we probably don't even know how to pose the questions in the right way.

Thus speculating today about the interactive computer of 2012 may be rather like a 1970 discussion of the keypunch technology of 1990. We do know a few things about the world of 2012, however: people will definitely still like to talk, and computers will probably still be getting cheaper. Therefore, human communication will still be heavily based on speech and text, and computers will be involved in increasingly sophisticated ways. This guarantees that any basic improvements in linguistic technology will find important social and economic roles.

3.1 LDC Databases: Past, Present and Future

3.2 Current LDC Databases

The databases now distributed by the LDC are described briefly below.

3.2.1 TI Digits Corpus

Many commercial applications depend on accurate recognition of digits. For example, to build a machine that can listen to a person reading a credit-card number over the phone and know which account to bill, you could use a large set of examples of people reading credit-card numbers. In 1984, Texas Instruments (TI) completed and documented the collection of a large set of spoken connected digits. The collection was undertaken to fuel and guide TI's own development effort in digit recognition technology.

Later, the National Institute of Standards and Technology (NIST) was asked to distribute the corpus, so that others could take advantage of the data, and so that meaningful comparisons between different systems could be made. In 1991, the corpus appeared on CD-ROM, which avoids the labor needed to deal with the dozens of bulky magnetic tapes previously required.

Precise cost records were not kept, but TI estimates that collection of this corpus required about $300-400,000 of TI's internal research and development funds. The corpus consists of data from 326 speakers (men, women, boys, and girls) reading sequences of one to seven connected digits. The data set is more fully documented in [LD84].

The corpus has been used extensively since its appearance for development and evaluation of connected digits in English. Connected digit recognition is still an important area of investigation with many potential commercial applications, and the TI digits have become the standard benchmark. So many researchers have benefitted from this data that a number of companies (e.g., AT&T, NYNEX, TI) have sponsored several even more ambitious efforts to collect larger and more diverse samples of spoken digits.

3.2.2 TIMIT Acoustic Phonetic Corpus

Researchers would like to build speech recognizers that are generally applicable, and not dependent on specific conditions, such as the vocabulary for a specific task. However, English speech is so rich in acoustic and phonetic variability that any new vocabulary is likely to contain sequences of sounds that were not previously observed, even in very large sets of speech data. This is because the commonest sound sequences will occur many times, while only some of the rare sequences will happen to occur, with new rarities cropping up at intervals as new vocabulary is added. The TIMIT corpus was designed to address this problem by containing examples covering the entire sound system of English.

In addition, this corpus samples the diversity of American English dialects, containing read speech from 630 speakers from all regions of the country. Along with the speech data, the corpus includes time-aligned orthographic and phonetic transcriptions.

The work represents the collaboration of several sites under sponsorship of ARPA: chiefly, MIT, SRI, and TI. Precise costs are difficult to estimate, but must be more than $500,000. The data have been used to support research on speech acoustics, to initialize statistical modeling techniques for speech recognition, and to evaluate the performance of phoneme recognition systems.

A new version of TIMIT, recorded over telephone lines, has recently been released by NYNEX. Known as NTIMIT, it is also available from the LDC.

3.2.3 DARPA Resource Management Database

TIMIT provides broad coverage of the sounds of English, but by the same token, its distribution of sounds and words is not at all typical of actual applications. Any particular application will have its own characteristic patterns of sounds and of words, and good models of these characteristic distributions will improve performance.

The DARPA Resource Management corpus was designed to be used in developing and evaluating speech recognizers, in a particular "representative" application. The sentences in the database are questions or commands addressed to a naval database and interactive graphics program, and include examples such as: "What is the Poughkeepsie's ASW rating," and "Are there any CAT-2 CASREPS for Winamac?"

This corpus was jointly designed and implemented by members of the ARPA speech recognition community. The data includes a dialectally diverse set of about 100 native speakers of English reading over 21,000 sentences. The vocabulary is specified and consists of about 1000 words. In addition, the data set includes the orthographic transcriptions of the approximately 2900 unique sentences. 70% of the talkers are male, and the rest are female. Further documentation of this corpus appears in Price et al. (1988), Proc. IEEE ICASSP, Vol. 1, pp. 651-654.

The corpus was designed during the first part of 1986, and recording took place between October 1986 and March 1987. In the Proceedings of IEEE ICAASP, 1988 (the major technical society meeting where speech recognition results are reported) all of the papers reporting results on large-vocabulary continuous word recognition in English made use of the Resource Management database. This is especially significant since abstracts for this meeting were due only a few months after the release of the corpus, attesting to the speech recognition community's hunger for common corpora for development and evaluation.

The DARPA RM corpus has also served to spur and to document a steady improvement in the performance of speech recognition systems, with error rates decreasing by a factor of 4 to 5 from October 1987 through February 1991. The data are available on CD-ROM through NTIS as well as from the LDC. Funding for the project (more than 2 person years of effort) came from DARPA. The sites chiefly responsible for design, collection, formatting and distribution include BBN, CMU, NIST, SRI, and TI.

3.2.4 ACL Data Collection Initiative

In early 1989, the Association for Computational Linguistics set up an ad hoc committee called the Data Collection Initiative (hence ACL/DCI), to oversee the acquisition and preparation of a large linguistic corpus to be made available for scientific research at cost and without royalties. All materials submitted for inclusion in the collection remain the exclusive property of the copyright holders (if any) for all other purposes. Each applicant for data from the ACL/DCI is required to sign an agreement not to redistribute the data or make any direct commercial use or it; however, commercial application of "analytical materials" derived from the text, such as statistical tables or grammar rules, is explicitly permitted.

The ACL/DCI has gathered several hundred million words of text, one dictionary, and a bit of speech data. As of September, 1991, about 40 individuals and research groups had gotten portions of the ACL/DCI's holdings, mostly by cartridge tape.

Four hundred copies of the first ACL/DCI CD-ROM were produced in September, 1991. It contains about 310 MB of Wall Street Journal text, about 180 MB of scientific abstracts, the full text of the 1979 edition of the Collins English Dictionary, and a preliminary sample of tagged and parsed text from the Penn Treebank project. The first printing of this CD-ROM is now almost entirely sold out.

For the first couple of years of its operation, the ACL/DCI was run entirely by volunteer labor, using borrowed computer resources. Although no cash changed hands, this in effect meant that companies such as AT&T, IBM, Xerox, and Bellcore were donating valuable employee time and computer resources. In the spring of 1991, the General Electric company donated $20,000, which was used for dedicated computer resources, especially disk drives, and AT&T Bell Labs lent some other disks. In the summer of 1991, the ACL/DCI got an NSF Software Capitalization grant (IRI 91-13530), which provides for additional computer equipment and support for data preparation and distribution. Dragon Systems Inc. paid for the manufacture of the first CD-ROM.

3.2.5 Penn Treebank

The Penn Treebank Project began in 1989 at the University of Pennsylvania under the direction of Mitch Marcus. It is intended to provide a very large corpus of grammatically-analyzed sentences. By the release of the ACL/DCI CD-ROM in mid-1991, about four million words of tagged text had been produced, and about half a million words of text provided with parse trees.

Funding for this project is now being provided by the LDC, which released the first Penn Treebank CD-ROM in January 1993. This CD-ROM includes the Brown Corpus, which has been re-tagged in order to provide consistency with the rest of the Treebank. The Brown Corpus has also been hand parsed. About 1.6 million words of Dow Jones newswire text have been hand parsed, and about 2.6 million words have been tagged. A variety of other texts have also been analyzed, largely in response to the expressed needs of researchers and system developers in speech and natural language technology.

A corpus of this size would not have been possible without automated tools for assisting in the marking of parts of speech, and in the grammatical analyses. The corpus has already been used in much research on algorithms for part-of-speech labelling and parsing, as a database for research in human parsing, and as a source of evidence in lexicographic and linguistic research.

The Treebank project is now working towards a three-million-word database defining predicate-argument structures for English text. Seed money for this project was provided by a grant from General Electric. Continuing support has been provided over a period of four years by ARPA, AFOSR, and ARO grants, and by the LDC. The present level of effort is equivalent to about 2.5 full-time employees.

3.2.6 ATIS

This corpus was ground-breaking in several ways. It provides speech and text in an interactive service application, the ATIS (Air Travel Information Systems) domain. Unlike the previously-described speech corpora, it is based on spontaneous speech, collected from people interacting with an actual system or a simulated system to solve a realistic problem, and recorded in a normal office environment (as opposed to a sound-isolated recording booth). This was a giant step towards providing realistic, spontaneous speech data, representative of a plausible application. Such speech includes false starts, hesitations, um's and ah's, coughs, and other things that are common in applications, but which are excluded from the read speech collected in previous shared corpora.

The ATIS task involves interaction between a human speaker and a computer, who plays the role of a travel agent providing the information needed to plan a trip by air. The standard evaluation of performance compares the answers given by the computer to standard answers drawn from a shared subset of the Official Airline Guide, and therefore the ATIS databases must include a rating of utterances according to whether they are answerable with or without dialog context, and the standard answers for those utterance that are deemed to be answerable. Experiments have also examined performance measures such as the time to complete a given task.

Collection and publication of the ATIS materials began in 1990 and is still continuing. Four ATIS collections, comprising 12 CD-ROMs, are now available from the LDC, and another release is planned for 1994.

3.2.7 SWITCHBOARD

SWITCHBOARD is a large corpus of conversational speech, produced by many talkers and recorded over long distance telephone lines. It was funded by the U.S. Government, collected at Texas Instruments, produced on CD-ROM by NIST, and released by the LDC earlier this year.

The entire corpus consists of 2,430 conversations, averaging about six minutes in length, by 523 speakers from around the United States. This amounts of about 240 hours of speech, and about three million spoken words. Detailed orthographic transcriptions are time-aligned with with the speech, and formal conventions are used to show speakers' turns, simultaneous talking, interrupted sentences, partial words, and other conversational phenomena.

The SWITCHBOARD conversations were collected under computer control, without intervention of a human operator or supervisory. A special 800 number ensured that all-digital lines were used for all conversations from the point of insertion to the computer doing the recording. The two sides of each conversation were recorded in separate 8 kHz mu-law data files.

A relational database provides information about the speakers, calls and topics. It provides speakers' age, sex, education, and other registration data, and the topic and speakers for each call. There were 70 suggested topics, with the topic for each call matched in advance to the interests indicated by the speakers involved in that call.

The entire SWITCHBOARD corpus occupies 26 CD-ROMs. The single-CD Credit Card corpus contains a sample of SWITCHBOARD calls on the topic of "credit cards." Overall, SWITCHBOARD cost more than a million dollars to produce.

3.2.8 Map Task Corpus

The Map Task corpus includes 128 two-person conversations, with 64 talkers (equally balanced between males and females) each taking part in four conversations. This corpus was collected and transcribed by the Human Communication Research Centre at Edinburgh University, with funding provided by the British Economic and Social Research Council. The LDC funded the production of Map Task CD-ROMs, and the corpus is available both from the LDC and from Edinburgh.

In the Map Task conversations, each participant has a map that is not visible to the other. On one of the maps, a route is traced. The point of the conversation is to transfer the route information to the other map, with speech the only method of information exchange.

The two maps have the same outline, and about a dozen labelled features. Not all features are on both maps, although most of them are. The overall experimental structure controls the relationship of the two maps in a detailed way, so as to produce conversations in which various sorts of words, phrases and situations are likely to arise.

The total corpus is about 18 hours of speech, about 150,000 spoken words. The speech was recorded on DAT recorders in a sound-treated booth with high-quality microphones, using one channel per speaker, and then transferred digitally to a computer and downsampled to 20 kHz. Detailed orthographic transcriptions are provided. The speakers are students from Glasgow.

3.2.9 TIPSTER

The TIPSTER corpus was funded by ARPA in order to foster research in document retrieval and document understanding.

THe TIPSTER project has two related parts, each with its own database. The first (and much the largest) database is a document retrieval test corpus, amounting at present to more than three billion bytes of English text, from a variety of sources including magazines, newspapers, newswire services, scientific abstracts, the Federal Register, and patents. The documents are presented in a consistent SGML format, and each document is given a unique ID. This collection was built at NIST in cooperation with UPenn.

In order to evaluate algorithms for document retrieval, the collection provides 150 "topics" in the form of user need statements, which were formulated by real users of document retrieval technology. Each such statement is about a page long, and retrieves an average of 300 documents from the overall collection. For each topic, a list of relevant documents in the corpus is provided. Given the size of the corpus, this list is generally incomplete, having been constructed by having professional relevance assessors judge a set of candidate documents produced by a variety of retrieval methods.

Within the TIPSTER project, this corpus is referred to as the "detection" corpus, since an algorithm's task is to detect the documents that are relevant to a query. Three TIPSTER "detection" disks have been released by the LDC, containing about one gigabyte of text each. A fourth disk is will come out next year. This test collection has formed the basis for the TREC (Text Retrieval Conference) meetings, which are open to all groups who register to try their algorithms on the task.

The second part of the TIPSTER project is known as "extraction," and involves filling out a database record with information extracted from a document of the type that might be retrieved in the detection phase. This task has sometimes been called "message understanding," and the TIPSTER extraction database has been used in the most recent of the series of MUC (Message Understanding Conference) meetings.

The TIPSTER extraction collection includes several thousand documents dealing with microelectronics and with joint ventures, with corresponding database structures for each document. This collection is about three orders of magnitude smaller than the detection collection. It has not yet been released by the LDC, but will be in the next year.

3.3 Future Plans

Several other LDC databases are in final stages of preparation and are nearly ready for release. They include four releases of Continuous Speech Recognition (CSR) corpora, which together will include 54 disks of read journalistic text; and the first release United Nations Parallel Text corpus, which will include about 600 megabytes of text in parallel English, Spanish and French, presenting UN documents of public record in these languages dating from 1988 to the present.

Space does not permit a full explanation of this material, nor of several corpora that the LDC will publish shortly on behalf of other groups. These include the CELEX lexical databases in English and German, produced by the CELEX consortium in the Netherlands, and the initial CD-ROM from the European Corpus Inititative.

Lack of space does not permit us to do more than mention several LDC initiatives currently in initial stages of data collection. One of these is COMLEX, a Common Lexicon for English that will include detailed syntactic information, a corpus-based cross-indexing of syntactic frames with WordNet semantic categories, and a large English pronuncing dictionary. Another is POLYPHONE, a multilingual telephone speech database that will be collected in parallel by groups in several countries, of which the LDC is one.

A longer-range plan aims at parallel speech/text/lexicon combinations suitable for exploring large vocabulary speech recognition technology in up to twenty languages.

4 Aims and Structure of the LDC

The problems of collecting, processing and annotating the needed quantities of linguistic data are too large for any one company. In any case, it is inefficient to duplicate the large up-front investment in foundations of research whose commercial implications will not blossom fully for ten years or so (even though the first applications have emerged already). Furthermore, university researchers and small companies, traditionally among the most important sources of technical innovation, may be frozen out entirely.

Thus a consortium is the right way to harness both economies of scale and individual creativity. We get economies of scale by avoiding duplication of effort on fundamentals, by doing the job in a way that serves the broadest range of needs, and by re-using data in a variety of technical areas. We improve the opportunities for individual creativity by making fundamental resources and tools available to researchers in academia and smaller companies.

Such an organization is also well adapted to recent international developments in the structure of language technology research. Over the past half-dozen years, we have seen the development of a new management paradigm for pre-competitive development of language technology. This new paradigm has taken somewhat different forms in different countries. In Japan, new task-oriented laboratories have been founded, notably the ATR Interpreting Telephony Research Laboratories in Kyoto, at which researchers from many companies work together on a common project. The German Verbmobil project is defining a common speech-to-speech translation task. Its goal of combining speech and language technologies in a negotiation dialogue brings together some of the most prominent universities and industrial labs in Germany (and potentially around the world).

In the USA, the ARPA Human Language Technology (HLT) program has made systematic and effective use of a "common task" method. This approach begins each project by specifying a task, defining a formal, quantitative evaluation metric, and developing a large common database for training and testing purposes. Then each participant pursues solutions in an individual way, and all participants meet periodically to compare methods and results (including evaluation scores). Used since 1987, this technique has resulted in rapid performance improvements in several areas. For example, speech recognition word error rate has been cut in half every two years for the past six years. Similarly, the performance of (text) message understanding and retrieval systems, measured in terms of metrics such as precision and recall, has also improved at a rate between 20-50% per year. Common tasks have also been an effective method to engender cooperation and the productive exchange of ideas and techniques.

Such common-task techniques, used in different ways in the three countries cited, are effective because

- they focus attention on the essential problems,
- they foster development of necessary common infrastructure,
- they lead to rapid dissemination of successful methods,
- they give researchers a concrete basis for discussion and exploration,
- they force solutions to system integration issues,
- they recognize and reward technical merit, and
- they provide funders with reassurance that the field is making progress (last but by no means least!)

The LDC originated in the U.S. experience, and is especially well adapted to the needs of that research community. An explict goal, however, has been to provide U.S.-sponsored databases to foreign researchers, and to help negotiate arrangements for general access by U.S. researchers to resources produced elsewhere. As a result, the LDC has become a genuinely international organization, with about a third of its members coming from outside the U.S. Through such international cooperation, we have already seen significant examples of international research cooperation via the common task method, arising in a bottom-up way through the actions of individual researchers and laboratories. In the future, such interactions may become more official, with agreements between government agencies or sponsorship by international scientific and technical organizations.

4.1 The LDC Staff and Board

Most of the LDC's work is done under contract by members of the community of researchers that it serves. Organizational and administrative functions are provided by a central staff, located at the University of Pennsylvania, which has been selected as the LDC Host Institution. Oversight is provided by the LDC Board, which consists of the director, up to three individuals appointed by DARPA to represent the community of researchers, and one representative appointed by each LDC Senior Member.

The founding director of the LDC is Mark Liberman. On-going LDC activities are supervised by Executive Director John J. ("Jack") Godfrey, whose services have been donated for two years by LDC Senior Member Texas Instruments. LDC staff members include Adminstrative Assistant Judith Storniolo (replacing Elizabeth Hodas), Programmer/Analyst David Graff, and Research Coordinator Rebecca Finch.

Victor Zue of MIT and David Pallet of NIST are ARPA-appointed representatives on the LDC Board; Senior Member NYNEX Science and Technology is represented by Judith Spitz; Senior Member Texas Instruments is represented by Raja Rajasekaran. George Doddington of ARPA and Charles Wayne of the U.S. Defense Department are advisors to the Board.

The LDC aims to ensure that no researchers are excluded by virtue of genuine inability to pay, while at the same time raising sufficient funds through membership fees to support on-going work. Therefore the LDC Board has set yearly membership fee at $2,000 in the case of not-for-profit entities, and $20,000 in the case of for-profit entities. Cases in which these fees pose insuperable obstacles will be considered and resolved on an individual basis by the Board.

Senior Members lend significant financial support to the LDC's operations by contributing $200,000 per year. In return, a seat on the LDC Board gives Senior Members considerable influence on the direction, scope and pace of data collection, and thus permits them to leverage their contribution through other LDC income.

Further information about the LDC is available by email from ldc@unagi.cis.upenn.edu.

References

[BDD+92] P. Brown, S. Della Pietra, V. Della Pietra, J. Lai, and R. Mercer. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18.1, 1992.

[DS81] George Doddington and Tom Schalk. Speech recognition: Turning theory to practice. *IEEE Spectrum*, 18.9, 1981.

[KF67] H. Kučera and W. Francis. *Computational analysis of present-day American English.* Brown University Press: Providence, R.I., 1967.

[LD84] R. Gary Leonard and George Doddington. A speaker-independent connected digit database. In *IEEE ICASSP*, volume 3, 1984.

[Sha51] Claude Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30, 1951.