Chapter 8

LINGUISTIC RESOURCES, DEVELOPMENT, AND EVALUATION OF TEXT AND SPEECH SYSTEMS

Christopher Cieri

University of Pennsylvania, Department of Linguistics, Linguistic Data Consortium Philadelphia, USA

ccieri@ldc.upenn.edu

Abstract Over the past several decades, research and development of human language technology has been driven or hindered by the availability of data and a number of organizations have arisen to address the demand for greater volumes of linguistic data in a wider variety of languages with more sophisticated annotation and better quality. A great deal of the linguistic data available today results from common task technology evaluation programs that, at least as implemented in the United States, typically involve objective measures of system performance on a benchmark corpus that are compared with human performance over the same data. Data centres play an important role by distributing and archiving, sometimes collecting and annotating, and even by coordinating the efforts of other organizations in the creation of linguistic data. Data planning depends upon the purpose of the project, the linguistic resources needed, the internal and external limitations on acquiring them, availability of data, bandwidth and distribution requirements, available funding, the limits on human annotation, the timeline, the details of the processing pipeline including the ability to parallelize, or the need to serialize steps. Language resource creation includes planning, creation of a specification, collection, segmentation, annotation, quality assurance, preparation for use, distribution, adjudication, refinement, and extension. In preparation for publication, shared corpora are generally associated with metadata and documented to indicate the authors and annotators of the data, the volume and types of raw material included, the percent annotated, the annotation specification, and the quality control measures adopted. This chapter sketches issues involved in identifying and evaluating existing language resources and in planning, creating, validating, and distributing new language resources, especially those used for developing human language technologies with specific examples taken from the collection and annotation of conversational telephone speech and the adjudication of corpora created to support information retrieval.

L. Dybkjær et al. (eds.), Evaluation of Text and Speech Systems, 221–261.

Keywords Language resources; Data; Data centres; Common task evaluation; Specification; Collection; Segmentation; Annotation; Intellectual property rights; Informed consent; Conversational telephone speech; Human subject behaviour; Quality assurance; Inter-annotator agreement; Adjudication; Distribution.

1 Introduction

The material for this chapter comes from lecture notes for a 2002 ELSNET Summer School with the goal of sketching the issues involved in identifying and evaluating existing language resource and in planning, creating, validating, and distributing new language resources especially those used for developing human language technologies. The workshop discussed these issues in the particular context of the common task technology development and evaluation programs that characterized human language technology research in the United States in the decade prior to the writing of this chapter. In preparing this chapter for publication, issues with momentary relevance for the ELSNET student but no general relevance were removed and facts, figures, and arguments were updated and generalized. The sections that follow begin with a description of the current linguistic resources landscape including the impact of common task programs and the role of data centres. After defining some terms, the discussion moves to planning resources for technology development including both technical and legal issues. After a brief discussion of how to find linguistic resources in the current context, the second half of the chapter details the issues involved in building language resources with emphasis on data collection.

2 The Linguistic Resource Landscape

Over the past several decades, research and development of human language technology has been driven or hindered by the availability of data.

> Modern speech and language processing is heavily based on common resources: raw speech and text corpora, annotated corpora and treebanks, standard tagsets for labelling pronunciation, part-of-speech parses, word-sense, and dialogue-level phenomena. (Jurafsky and Martin, 2000)

This dependence upon data is due in part to the shift toward probabilistic approaches and machine learning.

> By the last five years of the millennium it was clear that the field was vastly changing. First, probabilistic and data-driven models had become quite standard throughout natural language processing. Algorithms for parsing, part-of-speech tagging, reference resolution and discourse processing all began to incorporate probabilities and employ evaluation methodologies borrowed from speech recognition and information retrieval. (Jurafsky and Martin, 2000)

Although research continues on making the best use of limited data in statistical tasks, such as are common in speech recognition and natural language processing, we will argue that the need for linguistic resources in human language technologies is inevitable whether the research is statistical or rule governed. There is ample evidence that research communities and commercial developers of language technologies agree. COCOSDA (http://www.cocosda.org/), the International Committee for the Coordination and Standardization of Speech Databases and Assessment Techniques, established to promote cooperation in spoken language processing, emphasizes resources in its mission statement:

> COCOSDA supports the development of spoken language resources and speech technology evaluation. For the former, COCOSDA promotes the development of distinctive types of spoken language data corpora for the purpose of building and/or evaluating current or future spoken language technology.

Although the past 15 years has seen the birth of more than a dozen organizations that create or distribute language data, demand continues to outpace supply. ELSNET (http://www.elsnet.org/), the European Network of Excellence in Human Language Technologies, began receiving funding from the European Commission in 1991 to advance human language technologies by offering "an environment that allows for optimal exploitation of the available human and intellectual resources". The Linguistic Data Consortium (LDC (http://www.ldc.upenn.edu)) was founded in 1992 to support language-related education, research and technology development by sharing linguistic resources. The European Language Resources Association (ELRA (http://www.elra.info/)) was established as a non-profit organization in 1995 to make language resources available for language engineering. The Bavarian Archive for Speech Signals (BAS (http://www.phonetik.uni-muenchen.de/Bas/)) was founded in 1995 to distribute databases of spoken German to the speech science and engineering communities and has since added 17 corpora to its catalogue. Between 1991 and 1994, a consortium led by Oxford University Press built the 100 million word British National Corpus (http://www.natcorp.ox.ac.uk/). Planning for the American National Corpus (http://americannationalcorpus.org/) began in 2001 with the first release becoming available in 2003 and the second release in 2005.

Many teaching and research groups have contributed valuable language resources. The Center for Spoken Language Understanding (CSLU (http://cslu.cse.ogi.edu/)) at the Oregon Graduate Institute of Science and Technology seeks to teach and conduct basic research and technology development and to help other teachers, researchers, and businesses build and use spoken language technology. They have created 20 different corpora since 1992. The Johns Hopkins Center for Language and Speech Processing (CLSP (http://www.clsp.jhu.edu)), established in 1992, promotes research and education in language and speech technologies and develops one or more databases each year, in particular as a product of its annual summer workshops. The Institute for Signal and Information Processing (ISIP (http://www.isip.msstate.edu/)) of the Mississippi State University was founded in 1994 to develop public domain speech recognition software and has contributed several important data resources including JEIDA [ISBN: 1-58563-093-4, ISBN: 1-58563-099-3], a corpus of southern accented speech and the resegmentation of the Switchboard [ISBN: 1-58563-121-3] corpus.

The current landscape for language resources is characterized by individual researchers, small and large research groups, and data centres all striving to create data and yet failing to keep pace with the demand for greater volumes of data in a wider variety of languages with more sophisticated annotation and better quality. As of the writing of this chapter, "large scale" data collection might be defined as billions of words of text, tens of thousands of hours of broadcast speech, thousands of hours of conversational telephone speech, and hundreds of hours of meeting recordings. For part of speech, entity, and syntactic tagging projects, "large scale" is currently defined as a million or more words of tagged text.

When we speak of a wider variety of languages, we mean that several research communities working in speech and text engineering have begun to move beyond the dozen most commercially viable languages that have often been the subject of intensive resource development and toward those languages that are sometimes called *low density* not for the number of native speakers but rather for the scarcity of publicly available resources. In some circles, the term "*low density languages*" has been replaced by the term "*less commonly taught languages*" but even the latter must be understood to mean less commonly taught outside the countries where they are national or regional languages of importance. In the current landscape, English, especially American English is the language for which exist the greatest number and variety of language resources of the greatest volume. There also exist impressive lists of resources for many of the languages of Europe, including Eastern Europe, though not always to the same degree. Beyond the largest languages of North America and Europe, Mandarin Chinese, Modern Standard Arabic, and Japanese are now well represented. Beyond that, however, there are few languages with adequate resources to support the comprehensive development of language technologies. Recent years have seen some attention focused on languages such as Bengali, Cebuano, Hindi, Punjabi, Tagalog, Tamil, Tigrinya, Urdu, Uzbek, and Yoruba with large populations of native speakers, writing systems, a press, and an Internet presence but with very little in the way of publicly available language resources to support technology development.

With each passing year, new research communities embrace an approach to empirical analysis that is both computer-based and collaborative. At the same time, research communities that have traditionally used digital data now demand data in new orders of magnitude. We can see evidence for the former in the experiences of the TalkBank Project (http://www.talkbank.org), an interdisciplinary research project funded by a 5-year grant from the U.S. National Science Foundation (BCS- 998009, KDI, SBE) to Carnegie Mellon University and the University of Pennsylvania. The project's goal was to foster fundamental research in the study of communication by providing standards and tools for creating, searching, and publishing primary materials via networked computers. The TalkBank principals identified and collaborated with 15 disciplinary groups of which seven received focused attention: Animal Communication, Classroom Discourse, Conversation Analysis, Linguistic Exploration, Gesture, Text, and Discourse. The Talkbank principals have observed growing demand for shared data resources and common tools and formats among all of the Talkbank areas. Ten years ago the list of publicly available digital linguistic resources was dominated by resources for speech recognition, speaker verification, information retrieval, and natural language processing with ACL/DCI [ISBN: 1-58563-000-4], TIMIT [ISBN: 1-58563-019-5], TIDIGITS [ISBN: 1-58563-018-7], ECI [ISBN: 1-58563-033-3], Switchboard [ISBN: 1-58563-121-3], ATIS [ISBN: 1-58563-001-2], YOHO [ISBN: 1-58563-042-X], and Penn Treebank [ISBN: 1-58563-163-9] perhaps the most well-known data-sets, of that era. Today, due in part to the efforts of the Talkbank Project, the list of publicly available data-sets is much more varied and addresses the needs of communities that were not well represented in 1995. The Talkbank Corpora include: the FORM1 and FORM2 Kinematic Gesture corpora [ISBN: 1-58563-299-6, 1-58563-269-4], Grassfields Bantu Fieldwork: Dschang Lexicon and Dschang and Ngomba Tone Paradigms [ISBN: 1-58563-255-4, 1-58563-254-6, 1-58563-216-3], the SLx Corpus of Classic Sociolinguistic Interviews [ISBN: 1-58563-273-2], and the Santa Barbara Corpus of Spoken American English, Parts 2, 3, and 4 [ISBN: 1-58563-272-4, 1-58563-308-9, 158563-348-8] and Field Recordings of Vervet Monkey Calls [ISBN: 1-58563-312-7]. The research community working on the quantitative analysis of linguistic variation, which has been devoted to empirical methods since William Labov's seminal work *The Social Stratification of English in New York City* (Labov, 1966), has not traditionally published data-sets. This practice began to change with Gregory Guy's (1999) workshop on publicly accessible data and has borne fruit with the publication of the SLx corpus [ISBN: 1-58563-273-2] of classical sociolinguistic interviews collected by Labov and his students and transcribed, time-aligned, and annotated for examples of sociolinguistic variation by LDC.

While new communities join the ranks of those that work with digital language corpora, communities that worked with digital language data 10 years ago have continued to demand greater volumes of data. The resources created by the DARPA TIDES (http://www.darpa.mil/IPTO/Programs/tides) (Translingual Information Detection, Extraction and Summarization) and EARS (http://www.darpa.mil/IPTO/Programs/ears) (Effective Affordable Reusable Speech-to-Text) research communities offer an example. These include Gigaword News Text Corpora in English, Chinese, and Arabic [ISBN: 1-58563-271-6, 1-58563-230-9, 1-58563-260-0 respectively] with roughly a billion words in each, parallel and translated text corpora now measured in the hundreds of millions of words and the Fisher English corpora (Cieri et al., 2004) that now consist of approximately 4,000 hours of conversational telephone speech of which about two-thirds are publicly available [ISBN: 1-58563-313-5, 1-58563-314-3, 1-58563-335-6, 1-58563-336-4] at the time of writing.

Every speech and language researcher is not only a potential user but also a potential creator of linguistic resources thanks to desktop computing that has advanced to support both collection and annotation of text, audio, and video in amounts measured in the hundreds of gigabytes. Unfortunately, data creation and annotation require skills that are not adequately taught in the typical graduate programs in Electrical Engineering, Computer Science, Linguistics, or Computational Linguistics. Medium to large-scaledata collection and annotation further require specific technological infrastructure and management. International language resource centres such as the LDC and increasingly the ELRA maintain stable infrastructure and manage medium to large-scale language resource efforts. Together they have published more than 600 data-sets of which more than half have been donated by other organizations or individuals.

Simultaneous with the demand for increases in volume and language variety have been demands for more challenging data with annotations of greater sophistication. Speech recognition has seen the progression in its scope from a small vocabulary of words and short phrases to read speech, broadcast news, telephone conversation, speech in noisy environments and, most recently, speech during meetings. Treebanks have been re-annotated to create Proposition Banks and Charles University has produced Dependency Treebanks with tectogrammatical annotation, that is annotation at the level of meaning abstracted from the variations in linguistic form that appear on other linguistic levels. Treebanks have also moved from the realm of text to conversational speech, including languages such as Levantine Colloquial Arabic that lack the standardizing effects of a broadly accepted writing system. Part of speech tagging has evolved to include morphological analysis and gloss in the case of the LDC Arabic Treebank [ISBN: 1-58563-261-9, 1-58563-282-1, 1-58563-341-0, 1-58563-343-7]. The Automatic Content Extraction (ACE (http://www.nist.gov/speech/tests/ace/)) community has developed, from what was simple named entity tagging, a new specification for text tagging that includes entities, relations, events, and coreference. Video has been tagged for text recognition, entity recognition and tracking, and the physics of gesture.

2.1 Common Task Research and Technology Evaluation Programs

A great deal of the linguistic data available today have been created as a result of *common task* technology evaluation programs. Mariani (2002) sketches the history of speech and language technology evaluation in the United States giving the origin:

> Evaluation as a theme was introduced after the first DARPA programme on Speech Understanding Systems (SUS), which lasted from 1971 to 1976. The main conclusion of SUS was - that it was impossible to compare systems which were developed on different tasks, with different languages of various levels of difficulty.

the point at which common task evaluation became a regular theme:

> Evaluation was subsequently included as a theme in the following DARPA programme which started in 1984, but work was not initiated until 1987. The evaluation campaigns were open to non-US laboratories in 1992, and Philips Speech Processing (Germany), Cambridge University Engineering Department (UK) and LIMSI-CNRS (France) participated in the evaluation on that year, with excellent results.

and a mention of what is probably the most inclusive and international evaluation program:

> The Text Retrieval Evaluation Conference (TREC) programme started in 1992. It was opened to the international community from the very beginning and more than 120 groups have participated in this programme since. Both spoken and written language processing were addressed in the evaluation-based programmes.

The common task evaluation program is an implementation of a research management paradigm that has proven itself over the past decades. Under this model, multiple organizations work together to solve research and development problems while benefiting from shared infrastructure that may include task definitions, evaluation metrics and procedures, data and software components. Common task programs often involve direct sponsorship of participants but that is not always the case. Every year dozens of organizations participate in common task evaluation programs organized by the United States National Institute of Standards and Technologies (NIST (http://www.nist.gov)) without direct sponsorship finding that the opportunity to collaborate and have their technology evaluated objectively is a benefit worth the effort required by the evaluation. Common task programs may have one or more tasks required of all participants, one or more optional tasks and even some tasks that are site-specific. Examples of common tasks include automatic speech recognition of read speech, broadcast news, conversational telephone speech and meetings, identification of the languages spoken in a corpus of telephone calls, speaker identification from telephone calls, translation of news stories, identification of all stories in a corpus that discuss a topic, extraction and categorization of entities, relations and events in a corpus, the compression of one or more news stories into a headline or into summaries of varying length, and the development of two-way speech-to-speech translation.

Common task evaluation programs as practiced in the United states typically involve objective measures of system performance on a benchmark corpus that are compared with human performance over the same data. Examples of such metrics include word error or word accuracy rate in which system generated transcripts are compared to human transcripts of the same speech and points are deducted for every reference word missing from, added to, or replaced in the system transcript. Another, more controversial, metric is the Bleu score (Papinieni et al., 2002) in which translation systems are graded on the overlap between the word n-grams in their output and those in a set of independent human translations. The number of reference translations and the length of the n-grams can vary.

Mariani (2002), writes: *"The [European] projects are based on concept of co-operation among consortia, not on competition."* US common task programs have sometimes been criticized for being too competitive. Where common task evaluations measure participant performance directly via stable data and evaluation metrics, competition is inevitable. However, the concept of "competition with cooperation" is prominent in many US programs. Sites share data, discoveries, and software and join to form mini-consortia. Furthermore, in most common task evaluation projects data, evaluation metrics, and research results are published. Meetings are completely open or else include international observers. Research sites are also free to publish their own results at international conferences. It is important to note that different program managers have configured their programs to vary competitiveness. Naturally, if the goal is to identify the site with the highest performing system in order to award a contract for further development or creation of a production system, then research sites will be inclined toward competition and away from cooperation. Compare this with the DARPA EARS program in which annual performance goals were considered very challenging but in which the program required just one site to meet each goal with a system that could be composed of components from other sites. The effect of this approach was that many groups participated in multi-site teams and some participated in more than one such team. Intense cooperation among international research teams continues in the DARPA GALE (http://www.darpa.mil/IPTO/Programs/gale) program among others.

Task definitions originate with the program manager who seeks to accelerate research and development on pre-commercial technologies in order to respond to a government need. The program manager, researchers, resource and evaluation providers all refine the task definitions before they are formalized in an evaluation specification. The community also identifies resource needs and develops a schedule typically at a kick-off meeting. Infrastructure groups create resources and implement evaluation methods, negotiating with the community on any modifications. Evaluation is generally the responsibility of an organization that is independent of sponsors and all sites. In the United States, NIST, part of the Department of Commerce, is the most common evaluation group.

Shared resources lower the barrier of entry to all program participants and reduce the duplication of effort. One or more of the research sites may supply data or an independent organization may be contracted to create data specifically for the program. The LDC hosted at the University of Pennsylvania has been archiving and distributing language resources for common task evaluation programs since 1993 and has been creating them since 1995. In Europe, ELRA fulfils a similar function.

Most US common task programs distinguish two or three kinds of data. *Evaluation Data*, is carefully constructed specifically for the measurement of system performance. At evaluation time, research sites receive raw data and are required to process it and produce output compliant with the evaluation specification. The evaluation group then compares system outputs to human outputs produced according to the same, or else compatible, specifications. In many cases the human outputs are created ahead of time for all benchmark data and held in reserve until system outputs have been submitted. However, human annotators sometimes adjudicate sites' results either as a replacement for or as a complement to up-front annotation. Technology developers are typically unaware of the composition or time epoch of the evaluation data. The difference between evaluation data and *Training Data* may be varied to focus attention on the technology's generality or alternatively on its suitability to a specific task. In other words, technology developers build their rules or statistical models upon training data that may be matched or intentionally mismatched to the evaluation data. The size of the evaluation corpus will depend upon the technology and conditions being evaluated. However evaluation sets are generally sized to be the minimum that will provide robust, statistically significant technology evaluation. Funding for evaluation data is generally reserved before any is allocated to other kinds of data. However, the creation of evaluation data may occur after all other data is created. This is because many evaluation programs seek to take evaluation data from an epoch that is separate and preferably later than the epochs of other kinds of data. Doing so gives an opportunity to evaluate how technologies will fare when dealing with the new vocabulary that inevitably arises over time. Although it is possible to take evaluation data from a later time epoch and still create it early in an evaluation cycle, the desire to have the entire data-set be as fresh as possible, thus making it interesting for purposes of demonstrating technologies, means that data collection is often ongoing during an evaluation cycle and that evaluation data is often created after the other types, indeed just in time for its use in an evaluation.

Some but not all common task programs create a third kind, *Development/Test Data*, generally similar to evaluation data differing only in that development/test data is provided directly to sites for their own internal evaluation of the generality of their technologies. In multi-year programs, previous years' evaluation data is frequently reused as development test data.

In common task evaluation programs, such as those organized by DARPA and NIST, all three data types as well as the specifications for creating data and for evaluating systems, sites' system descriptions, and NIST's reports of results are published on an annual basis. In some cases, evaluation corpora are held in reserve until they can be replaced by newer evaluation corpora. This allows NIST to evaluate the systems of research sites who seek to enter the program mid-year.

Mariani (2002) provides a European perspective on common task evaluation programs. On the difference between US and European sponsored research, he writes: "*Simply stated, the US focuses on fewer but larger-size projects, whereas European funding is spread thinner over a larger number of projects and players.*" Further contrasting the availability of infrastructure for evaluation, he writes:

> The main actors in this framework are: the National Institute for Standards and Technology (NIST) as the organiser - defining the calendar, the protocols, the metrics, organising workshops and meetings; the Linguistic Data Consortium (LDC) as the Language Resources provider; several technology developers, both from the public and industrial sectors. The tasks addressed in this framework were made more and more difficult with time.

and then: "*... there is no infrastructure for evaluation in the EU, and the definition of the measure of success is still open.*" Seeing this as a major obstacle to progress, Mariani writes:

> The question arises, therefore, whether it is acceptable that European technology development and applications are conducted in Europe, with a dependence on technology assessment in the US, due to a lack of proper evaluation infrastructure in Europe. [...] "As for EU-US co-operation in Human Language Technologies, especially in Standards, Resources and Evaluation, this appears to be working well at present. The consensus is that it is well worth investing future effort in this direction for all concerned."

Over the past 5 years, the EU has sought to correct the situation Mariani mentions, in part due to his own efforts. Quite recently, ELRA has begun to evaluate technology in EU sponsored programs.

Mariani concludes with a set of common challenges for European and American technology development:

> Multilingualism is a major challenge on both sides of the Atlantic, for very different reasons. In Europe, there is a need to address all languages of EU citizens, for cultural and political reasons. In the USA, they feel that they have a strong strategic disadvantage: everyone understands English but they don't understand other languages. Therefore they cannot get the information from abroad!

From a contemporaneous American perspective, we report on the results of a breakout group on innovation and infrastructure held during the 2000 NIST Transcription Workshop. Participants acknowledged the benefits of shared task definitions, data, and evaluation metrics as reference points for comparison and noted that they function as a driving force for research agendas and tend to encourage solutions that tune to one task. Some researchers noted that they spent considerable time duplicating others' approaches in order to maintain competitive scores rather than focusing on innovation. The fundamental challenges they identified at the time, understanding signal characteristics, localization, speaker variability, lack of data, lack of sharing of tools and components, improvement of diagnostics beyond word error rate, and information transfer across linguistic levels, have since become the focus of intensive research. DARPA EARS devoted considerable energy into creating a large corpus of conversational telephone speech that represents the differences in regional accent, as well as age and sex that characterize American speech. The EARS community also developed a stable benchmark corpus and rules for its use that supported the measurement of progress from year to tear. The Novel Approaches working group in EARS developed a new set of features used in acoustic decoding of speech. The research management innovations continue in the DARPA GALE program where large, multi-site teams collaborate intensively to reduce the pressure on each participant to reproduce technological innovations developed by all other participants.

Researchers also sought to lower barriers to enter into the community in order to increase the size of the gene pool. They suggested a reinforcement of the idea of hub and spoke design, whereby a small number of required evaluation conditions made it possible to evaluate all systems on a consistent basis while optional tasks allowed researchers to pursue their own areas of inquiry. Similarly a mixture of large and small tasks would allow teams of different size to focus their efforts appropriately. Finally they lauded events such as the training workshops coordinated by the Johns Hopkins University and Mississippi State as ways to bring new researchers into the community. They recommended that future programs focus on tool and component sharing. Researchers generally agreed that well annotated, stable data, formal evaluation specification, and the knowledge transfer that takes place at workshops sponsored by the evaluation community were crucial to progress.

2.2 The Role of Data Centres

Data Centres play an important role in enabling education, research, and technology development. Within the United States, several multi-site, common task research programs have collaborated with the LDC to meet their data needs. LDC began in 1992 with the goal of serving as an archive and distribution point of corpora for technology development and evaluation. Over time the mission of the LDC has expanded either in response to or in anticipation of growing needs. In 1995, LDC began its first data collection projects when it became clear that there were not enough other labs to meet the growing demand. By 1998, it was clear that demand would continue to grow and data collection and annotation became a central focus for LDC. That same year, LDC also began to focus on infrastructure and tool development to support data collection and annotation. At the time of writing, LDC has grown to include 43 full-time employees and a transient staff of part-time annotators that has been as large as 65. 2019 unique organizations in 89 countries have used LDC data. To date, LDC has released 31,269 copies of 558 titles including more than 2500 copies of more than 160 titles within common task programs. The data produced for common task programs may be held in reserve to support evaluation before it is eventually released generally.

LDC is an open consortium that unites researchers in the non-profit, commercial, and government sectors with a common interest in language research, teaching, and technology development. The basic model is that organizations join the consortium on a yearly basis paying a membership fee that supports consortium operations. In return they receive rights to no-cost copies of all data released during the years in which they were members. Membership is open to any organization. Rights are ongoing and can be exercised at any time. For example, 1993 members may still request data under their membership for that year. The membership fees have never increased since they were set in 1992 by a board of overseers that included participants from the government, commercial, and non-profit sectors. Although preferable, it is not strictly necessary for an organization to become an LDC member. Many corpora may be licensed to non-members. To support new entrants into human language research communities, LDC also allows current members to acquire data from previous membership years at reduced rates.

As mentioned above, LDC serves as both a centralized distribution point and an archive. Every corpus ever released is still available. In some cases, newer versions with additional data or bug fixes replace old ones. However, where

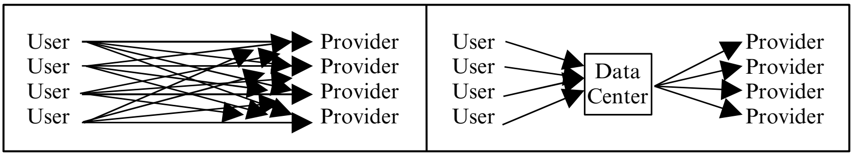

Figure 1. Data Centres serve as intellectual property intermediaries, reducing the number of user-provider negotiations necessary.

corpora have served as evaluation benchmarks they are preserved in exactly the same form in which they were originally released. This stability in the data is necessary to allow comparison of system performance over time. Where errors are discovered in benchmark corpora they are documented without being repaired.

LDC also acts as an intellectual property intermediary thereby reducing the amount of negotiation involved in resource sharing. Figure 1 demonstrates the benefit of such an arrangement. In order for a group of users to acquire rights to data directly from each of a group of providers, the number of agreements would be the product of the number of users and providers. Data centres acting as intermediaries provide a level of abstraction whereby each user signs one user agreement and each provider signs one provider agreement with terms that have been coordinated. The total number of agreements needed is just the sum, not the product, of the number of users and providers. More importantly, researchers gain consistent rights regardless of their negotiation skills and providers reduce the effort they spend to support research. This is especially important for commercial providers because they receive little or no direct revenue from supporting research.

There are additional advantages to the centralization of data resources. To the extent that resources are centralized, potential users have a smaller number of places they must search. LDC corpora are identified by catalogue number and ISBN, and their authors and titles are given within the catalogue. Centralized corpora can also be better standardized both in terms of structure and format and in terms of content and quality. Corpora published through the LDC are checked for quality and released in a consistent form, to the extent allowed by the variation in their target audiences.

Data centres consolidate resources from a multitude of disciplines in a single location. The LDC Catalogue, for example, contains resources that were developed to support research in: speech recognition under a variety of circumstances (including broadcast news, conversation telephone speech, and meetings), speech synthesis, language and acoustic modelling, information retrieval, information extraction, summarization, natural language processing, machine translation and speech-to-speech translation, and dialogue systems.

ELRA's catalogue has similar diversity. LDC also serves as a repository of benchmark corpora used in NIST technology evaluations so that new algorithms and innovative approaches may be tested and compared directly against state of the art systems whose scores on these same database have been published.

In its role as data centre, LDC strives to address the needs of its research communities. In part this has meant distributing and in many cases creating corpora in a greater variety of languages with more sophisticated annotation for use in an expanding number of disciplines. It has also meant creating and distributing data collection and annotation tools and corpus standards, and integrating data creation and technology development. Specifically, over the past few years LDC has increased collaboration with the beneficiaries of its data in order to use their technologies and technology evaluation tools to improve corpus creation. In the commonest cases, LDC corpora are often dually annotated and then scored for consistency using the same tools that score system performance against a benchmark corpus.

A corpus contains data selected and prepared for a specific purpose. However, it is sometimes possible to create a corpus that supports more than one kind of research. For example, in December of 2004, LDC received a request for conversational telephone speech that was dense in named entities and therefore useful for projects focusing on information extraction, such as the Automatic Content Extraction (Doddington et al., 2004) program. Having met the goals of a phase of the Fisher English (Cieri et al., 2004) collection of conversational telephone speech supporting speech-to-text technology evaluation, LDC was able to change the topics assigned to subjects in order to highlight people and places in the news. Topics were formulated to remind subjects of the important persons and places associated with each event. The results were conversations that were certainly useful for the original goals but were also much richer in entities and events, and therefore able to serve the ACE project as well.

Over the past dozen years, several other data centres have begun to perform functions similar to LDC with models that are similar but not identical. The ELRA also serves as an archive and repository of language data. Under their model, members receive discounts on licensing fees. Where LDC and NIST have specialized to focus on data and technology evaluation respectively, ELRA has recently begun to handle both functions for European programs.

3 Background on Linguistic Data and Annotation

A *corpus* is any body of raw data selected, sampled, formatted, and annotated for a specific purpose. The terms *language data*, *linguistic data*, and *raw* *data*, here refer to recorded observations of any performance or experiment involving a spoken, written, or signed language or communicative system. *Annotation* is any process of adding value to raw data through the application of human judgement. For example, an audio recording of a telephone conversation is raw data. A transcript of that same conversation encodes subtle human judgement about what was said, and in some cases intended, and is thus annotation.

Annotation may be accomplished with direct human effort or mediated by some technology based upon rules or statistical observation. Morphological analysis, in particular generating one or more analyses of a surface form, is frequently rule-based while part of speech-tagging, particularly selecting the most probable analysis from among several possibilities, is frequently based upon statistical models. Other examples of annotation include transcription, segmentation, part-of-speech tagging, parsing, entity, relation, event and co-reference tagging, sense disambiguation, topic categorization, summarization, and translation.

Segmentation, a very specific type of annotation, involves dividing some larger unit of observation into smaller pieces to facilitate future annotation, search, or analysis. For some purposes it may be necessary to segment by actually cutting a recording, for example, a large audio or video file, into pieces and storing them separately. However, segmentation can generally be accomplished by storing the time stamps that mark the beginning and end of significant events in a separate annotation file in order to preserve the integrity of the original recording. Recordings of conversations might be segmented at speaker changes, breaths, or pauses. News broadcasts might be segmented at speaker changes or story boundaries.

In-line annotations are embedded within raw language data. Though very common, especially in text, in-line annotations can be problematic for several reasons. Although there are formats that permit raw data and annotations of different modes to be mixed in the same file, this practice may compromise the integrity of the raw data. In many cases the formats that allow mixed modes to be embedded in the same file are proprietary and reduce the generality of access. Multi-tiered annotation, in-line, can make data unreadable. Finally, the distribution rights for raw text and annotation are frequently very different. Reliance on in-line annotation reduces the ability to distribute by multiplying the restrictions that may apply to the raw data and its annotation.

Stand-off annotations are separate from the signal and refer to it or portions of it via time codes, byte offsets or word offsets, for example. Stand-off annotation is widely recognized as the best option for those who work with audio and video data. Stand-off annotation is equally effective though less common when working with text. In the simplest cases, where the signal is text and there is a single layer of annotation, stand-off can be slightly more complicated to parse than in-line. However, even in this case there are advantages of stand-off annotation, which can be associated equally well with text, audio, or video without changing the raw data in any way and can be used to maintain separate layers of annotations.

It is important to reiterate that a corpus contains data selected for a specific purpose. Selection is an important part of corpus creation. The match or mismatch between the data in a corpus and the technology or application for which it is used may have a large impact on the success of the venture. To give some simple examples, a speech recognition system trained on broadcast news will not fare as well when tested on conversational telephone speech as would the same engine trained on matching data. Unfortunately, many research communities lack the specific data they need to conduct research and technology development and must settle for, or choose to settle for as a cost-saving measure, data that was developed for some other purpose. The Switchboard [ISBN: 1-58563-121-3] corpus for example, was originally developed to support research in speaker identification and topic spotting but has been re-annotated to support research in everything from speech recognition to natural language processing and parsing. The Topic Detection and Tracking 2 (TDT-2) [ISBN: 1-58563-183-3] corpus has similarly been re-purposed for speech recognition and spoken document retrieval.

It should be clear by now that *time* is an important dimension in spoken linguistic data. Time is evident in recordings of live linguistic performance such as conversations or monologues. In written language, *sequence* replaces time as a significant dimension. Text need not be written or even read in the order in which the final version appears. Writers are free to edit and reorder their writings and readers are free to progress through a text in some non-linear fashion. Nonetheless, the sequence of words on a page or in a file represents the author's deliberate intent and thus forms a dimension upon which subsequent analysis is based. Of course, not all linguistic data need be ordered along a chronological or even sequential dimension. In lexicons, for example, the order of entries means something entirely different than the order of written words in a text or spoken words in a conversation.

Speech and gesture are necessarily captured as the communicative performance takes place with time playing a major role in the analysis of these modes. Time is generally represented as the offset from the starting time of the event recording not as absolute time. However, in some cases the absolute starting time of the recording of the event is present either in metadata or encoded in the file name of the recording or both. This is desirable. Now that desktop computing makes it possible for individual researchers to collect and annotate small to medium sized corpora, the number of collection projects has grown creating opportunities to study phenomena across corpora. With time playing such a crucial role in the introduction of new vocabulary, especially named entities, the encoding of absolute time in data collection offers the possibility of placing recordings from multiple sources on a single timeline.

The recording of written language on the other hand generally does not preserve the order in which the components, words or characters, of the communication were produced. This may be because the recording technology, stone tablets, papyrus, or sheets of paper, lacked the ability to represent time or because, as is the case with modern word processing technology, time is ignored. In the recorded version, written communications order language according to the author's desires, which may be very different from the order in which it was produced. One exception is handwriting which, like gesture, may be recorded via motion capture devices such that timing information is available. For handwriting a combination of stylus and writing surface translate handwriting into time-sequenced movement data. In the case of gesture, motion may be captured, for example, by a combination of transmitters placed on a subject's joints and receivers placed at the corners of a three-dimensional bounding box or may be interpolated from two-dimensional video. Although the grammars of spoken, written, and gestured language differ significantly from one another and although there is a common association between written language and text and between spoken language and audio, one should not conclude that all text encodes written language or that all audio encodes speech. One may read written material aloud and transcribe spoken material proving that the mode of recording language does not guarantee the type of language recorded.

4 Data Planning for Technology Development and Evaluation

Data planning for technology or application development and evaluation depends upon a number of factors including the purpose of the project, the linguistic resources needed, the internal and external limitations on acquiring them, availability of data, bandwidth and distribution requirements, available funding, the limits on human annotation, the timeline, the details of the processing pipeline including the ability to parallelize or the need to serialize steps.

4.1 Technical Issues

Planning for corpus creation involves matching the types and volume of raw data and the complexity and coverage of their annotation to the needs of the human language technology, such as speech-to-text, and sometimes to an application using that technology, such as a voice mail transcription system. Basic speech-to-text technologies require audio recordings of speech with time-aligned annotation. Acoustic modelling requires a close alignment of the speech signal with symbolic labels. Were it not for phonetic variation and differences between orthography and phonetic reality, a single layer of time-aligned orthographic transcription, which would thus also be phonetic, would be sufficient. However, because writing systems often differ from phonetic reality, transcriptions need to be mediated either through a second layer of annotation or through a pronouncing lexicon. The lexicon generally contains an entry for each surface form showing its alternate pronunciations, possibly its morphological analysis in morphologically complex languages, its part of speech and preferably its frequency in different domains. The former approach involves creating two or more tiers of transcription in which the tiers are aligned to each other and the audio. One tier provides an orthographic transcription and the other a phonetic transcription. It should be noted that the two-tiered transcription is a kind of contextualized version of the pronouncing dictionary.

Another difference to consider in comparing corpora is the variability potentially present in the raw data. Language varies according to region. Regional varieties of a language are generally called dialects. However, the reader is cautioned that the meaning of the term changes with the situation. Linguists use the term dialect to refer to mutually intelligible varieties of the same language. When varieties become mutually unintelligible they are considered different languages. However, there is great variation in usage; varieties sometimes labelled dialects, such as the dialects of Chinese, evince less mutual intelligibility than varieties traditionally considered different languages, such as Swedish and Norwegian. The situation is further complicated in dialect continua, chains of regional dialects in which adjacent pairs are mutually intelligible while varieties separated by greater space are not. Although there has been some work done on cross-dialectal training of human language technologies, this work is still in its early stages so that it remains important to match the dialects of training data to dialects targeted by the technology or application. Variation in spoken and written language may also be conditioned by social factors, time, register, domain, and mode. Vocabularies may be general, technical, literary, or conversational. Speaking and writing are different modes of communications marked by different grammar.

We can elucidate the process of analyzing needs for technology or application development through an example, a system designed to gather information for purposes of evaluating the merits of accusations of fraud in the stock market or other illegal business practice. Such a system might process multiple sources of information in order to help investigators find correlations between events and trading or business activities. The sources of information might be news text and press releases present on the World Wide Web or available via subscription feeds, news broadcasts and cable news programmes; wire taps and corporate email archives of companies under surveillance or investigation; video recordings of their meetings and of depositions of their employees, partners, and customers, financial analyses, transaction reports and filings. The system would help answer a number of questions using this data. For example, within the telephone conversations, one would like to know who speaks, what they say both in summary and in detail, whether they mention the company under investigation or its employees regardless of whether that mention used a fully specified name or a nickname. In meeting video, one would also like to analyze gesture, gaze, and body language in order to help determine the power relations among the participants and their disposition toward each other and toward the topics discussed. These needs suggest a number of technologies already under development including speaker recognition, summarization, information extraction, speech recognition, video and gesture analysis, and information each of which have their own data requirements.

4.2 Intellectual Property Rights and Informed Consent

The acquisition of intellectual property rights and the informed consent of human subjects are important parts of the data planning process and the responsibility of the collection team. Where a technology development effort benefits from existing data, these issues have generally been handled though their impact may show in the cost of licensing the corpus. A complete review of the legal issues goes beyond the scope of this work. Here we will simply discuss representative issues taking examples from the current situation in the United States noting that laws and practice differ from country to country.

Within the United States, the creator of an original work is owner of that work and the only one who has the right to copy and distribute. The creator can assign copyright by contract and employers generally arrange to acquire copyright for the work done by their employees. To support research and teaching, the principle of *fair use* permits copying for those purposes. US law provides the parameters with which fair use is evaluated but does not actually define the space. The parameters include: the use of the material whether for commercial purposes or for education and research, the size of the material used relative to the entire body of work from which it is extracted, the degree to which the data is transformed before use, and the probable impact the use will have on the owners' ability to derive income from the material. The interpretation of fair use is left to the discretion of courts. In practice, organizations typically define safe harbours in which they believe they can operate with reasonable assurance of avoiding charges of copyright violation. Given the uncertainty surrounding copyright law, the LDC normally acquires explicit rights to distribute the data in its corpora.

Much of the data used in human language technology development comes from news publishers and broadcasters. That data is intrinsically interesting and has a broad vocabulary due to the variety of topics covered in the news. Other benefits of news text and broadcast news are that the data is broadly available, that they can be licensed for research use without unusual limitations and that licensing costs are reasonable given creation costs. Some sources are willing to share their data for research use without cost. For example, some governments, including the US government, consider their publications to be in the public domain. Other sources have offered data centres deep discounts off their normal licensing fees. In negotiating rights for language data, LDC generally seeks non-exclusive, perpetual, worldwide, royalty free license to distribute to both LDC members and non-members who sign an agreement limiting their use of the data to linguistic education, research, and technology development without limitation as to quantity. These conditions are motivated by the use of corpora as benchmarks for evaluating new technologies. Distribution restrictions that limit distribution either by number of copies, time, or region limit the usefulness of the data for technology evaluation.

For conversational and meeting speech, the primary issue is *informed consent*. Within the United States, research that involved human subject must proceed according to a collection protocol approved by an Institutional Review Board (IRB). These boards, apparently designed to regulate clinical medical trials, review collection protocols for their risk versus the benefit presumed to result from the research. For Human Language Technology (HLT) research and development, the risks are generally no greater than those subjects encounter in their everyday lives. In most data collection efforts, subjects are simply asked to talk or write or to make judgements about language. One area of concern, however, is the risk to anonymity. Collection efforts maintain the anonymity of human subjects by separating the identifying information used to contact and compensate subjects from the actual recordings of speech or decisions. For many linguistic research projects the benefits to human subjects are also minimal compared, for example, to the benefits of having access to experimental pharmaceuticals in clinical trials. However, this minimal benefit to the individual is acceptable given the minimal risk and the potential benefit to society resulting from the creation or improvement of technologies that become part of our everyday lives.

5 Finding Resources

One of the first decisions one must make is whether to use existing resources or else build them to specification. Found resources are generally less expensive and may have the virtue of being previously used, discussed, and improved. However they may not necessarily be ideal for the target use. Resources built to specification are optimized for the target use but with added cost and time. Finding digital linguistic resources is somewhat more difficult than finding books due to the distributed nature of their publication and the lack of a single catalogue. In order to find resources published directly by their creators, one must know the creators in advance, learn about them via published papers, request advice from experts using networked discussion lists, or perform Internet searches by resource type and language. Resources that are distributed via data centres, such as the LDC or the ELRA are somewhat easier to find due to their centralization and due to the catalogues the centres maintain.

At the time of writing, the ELRA (http://catalog.elda.org) catalogue allowed full text search with the user entering one or more search terms and specifying whether those terms were to be matched exactly, in conjunction or in disjunction. The search engine responds with a hit list containing brief descriptions of matching corpora and links to fuller descriptions. Available information includes pricing information and, for selected corpora, additional documentation, samples, a validation report, a description of the design of the database, and a list of known bugs. Pricing information is distinguished along three dimensions, whether the organization is an ELRA member or not, whether the organization is commercial or not, and whether the use is commercial or not.

The LDC catalogue (http://www.ldc.upenn.edu/Catalog), at the time of writing, can be browsed by data type, data source, and release year and can be searched using full text search and fielded records in any combination. The fields: catalogue number, corpus name, authors, and corpus description can be searched with keywords. The fields for languages, data types, associated research projects and recommended applications have controlled vocabularies that are selected from a pick list. The user can specify whether to conjoin multiple fields with Boolean AND or OR. The search engine responds with a hit list containing pointers to the full catalogue entries for each item. Catalogue entries include the corpus name, authors, the catalogue number and ISBN number, release date, data type, data source, associated programs, recommended applications, languages and ISO language codes, distribution media, licensing information, links to online documentation, and samples. Once licenses are properly executed, LDC data are distributed on media, CD, DVD, or hard drive or via HTTP transfer depending upon the size of the corpus.

The Open Language Archives Community (OLAC (http://www.language-archives.org/)) indexes 33 different collections of language data, including the holdings of the LDC and ELRA, as a union catalogue of language resources. OLAC separates the function of hosting data from the functions of hosting, indexing and searching metadata. Participating archives export their metadata to search providers who index and maintain search engines. For example the advanced search functions hosted by the LinguistList (http://linguistlist.org/olac/) allow keyword searching in the title, creator/contributor and corpus description fields and searching with controlled vocabulary among the language, type, and discourse type fields. Queries against the LinguistList search engine return hit lists with links to fuller catalogue records. The contents of these fuller descriptions vary from one data provider to another. The OLAC metadata language accommodates most of the metadata types needed by its constituent data providers and is extensible. Metadata types include creator, contributor, publisher, title, coverage, date, description, format including encoding and markup, identifier, language, relation to other resources, rights, source, subject, functionality, and linguistic type.

6 Building Resources

The actual steps in building language resources include planning, creation of a specification, collection, segmentation, annotation, quality assurance, preparation for use, distribution, adjudication, refinement, and extension. We have already discussed planning, including planning for the acquisition of distribution rights and consents. The sections that follow cover the other steps with particular emphasis on collection.

6.1 Specification

During the course of a corpus creation project, a multitude of decisions are made and implemented. Unless the project has a very short life cycle or the principals have exceedingly good memories, some decisions will be forgotten, reviewed, and possibly revised though not always with the effect of improving the effort. A corpus specification describing the overall use of the corpus, the raw data used as input, the collection and annotation processes including dependencies among parts of the process, the output formats, and assumptions about all of the above, can help stabilize and coordinate such effort. The specification contributes to planning, training of collection and annotation staff, and documentation of the final products. It also reminds internal staff, sponsors, and potential users of the decisions made.

6.2 Collection

In preparation for the development or evaluation of language technologies, one may collect data representing spoken, written, or gestured communicative modes that may be captured as audio, text, or video.

6.2.1 Collection parameters. Before beginning, one must determine the parameters of collection. Here we will discuss two such parameters *sampling resolution* and *quantization*. Sampling resolution is the frequency with which an analogue signal is sampled to produce a digital artefact. The sampling of two-dimensional graphics, for example, is measured in dots per inch. Video is generally sampled at roughly 30 frames per second. Speech is sampled as it is digitized at rates that tend to range from 8 to 48 thousand cycles per second or kilohertz (kHz). The range of frequencies involved in spoken language (0–8 kHz) and the frequencies the human ear can detect (0–11kHz), as well as the need to double sampling frequencies in order to avoid aliasing, have figured historically in the selection of sampling rates for digital technologies and formats. The most common sampling rates are: 8, 11, 16, 22, 44, and 48 kHz. Quantization refers to the range of values any single sample may have. For example, common quantizations for two-dimensional images range from two bits, representing just black and white dots, to 32 bits representing more than 4 billion colours or shades of grey. Speech is typically quantized in 8, 16, 20, or 24 bits. The greater dynamic range offered by 20 and 24 bit quantization reduce the probability of reaching the sample peak (clipping) when increasing microphone gain or when dealing with an audio signal that has especially great or especially variable amplitude.

Deciding upon a sampling rate and quantization often involves compromise. The principle of full information capture (Chapman and Kenney, 1996) states that sampling and quantization should be fine enough to capture the smallest detail considered significant. In image digitization this might involve reviewing an image with a jeweller's loupe to identify the smallest detail to be preserved, measuring that unit and then setting resolution to assure capture. In the domain of spoken language, full information capture might be interpreted as recording the highest frequencies used in human language in which case audio sampling rates of 16 kHz would be adequate. Another approach sets capture parameters at the limits of the biological system. In the case of audio data for human listening, 22 kHz sampling reflects this thinking. Current needs also play a role in setting collection parameters. A collection designed to provide training data for a technology that expects data at a certain sampling rate may reasonably decide to collect at just that rate. This approach optimizes for short-term gain and may prove problematic if current needs prove less demanding than future needs. In LDC's experience, data have a protracted job cycle being re-annotated and reused far beyond original intent. Other constraints include available time and funding to conduct the data collection and the capacity of available technologies at the time. Following Moore's Law, we expect the capability of computer technology to increase and its cost to decrease. As a result, constraints based upon technical capacity and cost tend to loosen over time. A nearly ideal situation exists when affordable technology is capable of collecting data that not only meets current needs but satisfies the principle of full information capture and exceeds the ability of the biological system. We have reached that state with respect to collection of digital text and audio. A billion words of text in a language that averages six characters per word encoded as two bytes per characters would require 12 gigabytes of storage if uncompressed. Even inexpensive notebook computers generally have that

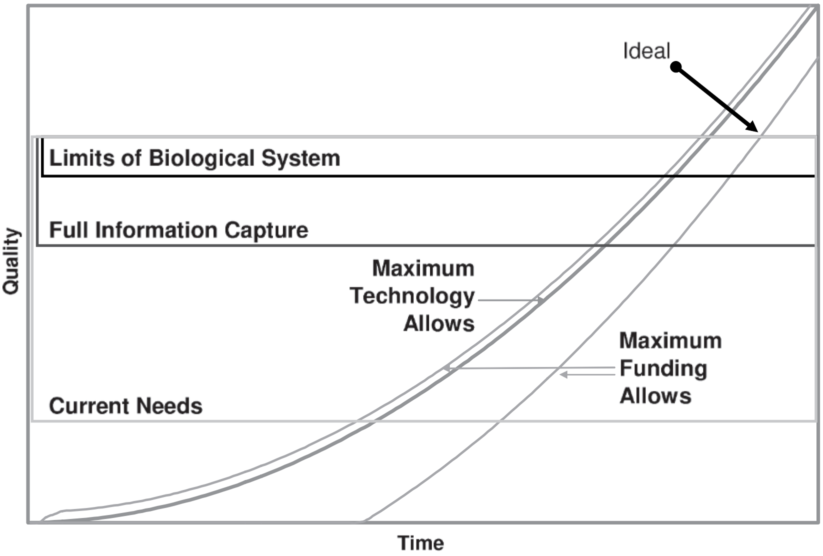

Figure 2. Factors influencing data collection parameters.

much storage to spare. Similarly, an hour of single channel audio sampled at 22 kHz with 16 bit quantization requires about 150 MB (megabytes) of storage per hour of recording. Although current notebook computers lack the capacity to store, say, 1,000 hours of such data, desktop computers and external drives for notebook computers can easily store this much data at costs that range from one-half to US$2 per gigabyte. The situation for digital video is not quite so far along. Large collections of high quality video still require storage solutions that, while possible, are not standard on average desktop computers. Figure 2 presents these factors graphically. Note that the principle of full information capture need not require lower quality than is imposed by the limits of the biological system though the graph presents them in that relationship.

6.2.2 Text. The volume of text corpora is typically measured in bytes or words. Because the number of bytes per word varies from language to language even when a general purpose encoding such as Unicode is used, byte counts can only be compared where the texts are in the same language and encoding. The number of characters per word varies by language, while the number of bytes per character varies by language and encoding. In languages such as Chinese, where words are not generally space separated, and where the conceptualization of word does not benefit from centuries of space separated writing, segmentation, and thus word count vary with the counter. Furthermore it is important to note whether word counts include tags that can comprise a considerable percentage of the tokens in the text especially in those harvested from web pages. Consider the following *New York Times* article excerpted below. The complete document contains 467 space separated tokens of which 35 are tags and 30 are non-text tokens. That leaves just 402 text tokens or 86% in a news story that appears to have very little extraneous mark-up.

```
<DOC>
<DOCNO> NYT20000101.0002 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 2000-01-01 00:21 </DATE_TIME>
<HEADER>A3886 &Cx1f; taf-zu a BC-NYTIMES-ISSUE-NUMBER-
01-01 0415 </HEADER>
<BODY>
<SLUG> BC-NYTIMES-ISSUE-NUMBER-NYT </SLUG>
<HEADLINE>A CORRECTION: WELCOME TO 51,254 </HEADLINE>
(lh)
c.1999 N.Y. Times News Service
<TEXT>
<P>NEW YORK _ For those who believe that in the good old days _
before calculators, before computers _ people were better at
mental arithmetic, The New York Times offers a sobering New
Year’s message: Not necessarily.</P>
<P>On Feb. 6, 1898, it seems, someone preparing the next day’s
front page tried to add 1 to the issue number in the upper left
corner (14,499) and came up with 15,000. Apparently no one
noticed, because the 500-issue error persisted until Friday
(No. 51,753). Saturday The Times turns back the clock to
correct the sequence: this issue is No. 51,254.</P>
<P>Thus an article on March 14, 1995, celebrating the arrival
of No. 50,000 was 500 days premature. It should have appeared
on July 26, 1996.</P>
<P> As for the other number on the front page _ the volume,
in Roman numerals _ it remains CXLIX. It will change to CL on
Sept. 18, when The Times enters its 150th year.</P>
</TEXT>
</BODY>
<TRAILER>NYT-01-01-00 0021EST &QL; </TRAILER>
</DOC>
```

Standard desktop computing is more than adequate to support small- and medium-sized text collections. The specialized tools, such as leased lines and dedicated modems, previously used to distribute newswire are rapidly being replaced by distribution via the Internet using ftp, http, and rss protocols. If large-scale text collection remains challenging, the cause is not the network bandwidth or storage capacity of desktop computers but rather the long-term commitment required. The largest news text corpora now exceed one billion words; for example, the second edition of the English Gigaword News Text [ISBN: 1-58563-350-X] corpus contains 2.3 billion words of text selected from both daily distributions and archives of several major news providers covering a 12-year period.

Parallel Text, that is text and its translation into another language, is generally measured in words of *source language* text. Source refers to the language in which the text was originally written while *target* refers to the language into which it is translated. In some cases it will be difficult to distinguish the source and target languages because that information is not provided with the text. The form of the text itself may provide hints. For example, in parallel text involving English, non-native use of the determiner "the" will suggest the text was translated, imperfectly, into English. In parallel text produced by multilingual publishers such as the United Nations and many international news sources, original text may have been written in any of the several languages. The largest parallel text corpora contain tens of millions of words of text in the source language. Examples include the Hong Kong Parallel Text [ISBN: 1-58563-290-2], Arabic English Parallel News, Part 1 [ISBN: 1-58563-310-0], and UN Parallel Text Complete [ISBN: 1-58563-038-1] corpora.

6.2.3 Speech. In studies of speech conducted over the past decade, audio recordings of spoken language have included isolated words, short phrases selected at random or selected to be phonetically rich or balanced, read speech, task-oriented dialogues among humans or between humans and machines, broadcast news, conversations, and meetings. Some collections, particularly those involving broadcast news and meetings have included video. Although not strictly necessary for all applications, audio collections are frequently accompanied by time-aligned transcriptions, which may be orthographic, phonetic or some hybrid.

Corpora that support speech-to-text systems typically include audio recordings of speech in the language and preferably in an acoustic environment and genre that match the target application. A voice mail transcription system would need to be able to handle telephone speech characterized by reduced bandwidth and multiple encodings and decodings, compressions and decompressions of the signal as it passes over landlines, cellular networks, or even the Internet. In addition there is variability at the handsets of both talkers, which may be traditional handsets with carbon button microphones, head mounted headphone/microphone combinations, or speaker phones and which may involve additional retransmission of the signal if the handset is cordless. In just the few examples given above, distance from mouth to microphone may vary from a few centimeters to a few meters. Because of the severe differences in spectral properties, telephone speech, broadcast speech, and broadband speech, are generally treated separately.

The increasing integration of speech-to-text with other technologies has increased the requirements for raw data and annotations. Where video accompanies audio, researchers have investigated the fusion of lip shape recognition with speech recognition to improve accuracy. Similarly, optical character recognition of text appearing in broadcast video, face and object recognition, and tracking and related technologies enrich the information that can be extracted from broadcast. However, the development of these technologies requires the collection of other types of data, as well as new annotations.

In contrast, data for text-to-speech systems generally include carefully produced speech, recorded in a quiet environment with high quality microphones. The speech is commonly read from prompts and may include words, short phrases and sentences, and nonsense syllables selected to give the best possible coverage of the language's phonotactics. Sentences may be written to be phonetically rich at the risk of sounding unnatural. Alternatively, sentences of actual text may be selected for their coverage of the consonant and vowels combinations, prosodic features and the like.

Speech corpora are measured in hours of recording. Hours of actual speech will be somewhat less because stretches of silence do occur in conversational corpora, and stretches of music and commercials occur in broadcast. Speech corpora may also be measured in the number of words contained in their transcripts. Of course, the number of words per minute of speech varies by language, speaking style, speaker, and format. In the CallHome English [ISBN: 1-58563-112-4] transcript, of which a small piece is presented below, 32% of the tokens are something other than words.

```
825.89 828.31 A: Oh. How's he doing?
827.81 828.48 B: And he's had
829.40 835.67 B: Well he's working for an American firm
over here, and he's doing very very well. %um, and his wife
just had their fourth child.
835.80 836.63 A: Wow.
836.34 838.15 B: A little boy. &Benjamin. yeah.
838.51 839.22 A: Wow.
838.82 842.10 B: %um, about, %uh, t- two weeks ago maybe.
842.12 842.64 A: %huh.
842.33 842.85 B: Tops.
843.76 846.20 B: And I got a card from, you remember &Amy
&XXX?
846.10 846.60 A: yeah.
847.25 849.27 B: yeah. Well she just had a baby a couple of
849.62 850.44 B: (( ))
850.63 851.50 A: heard that.
851.38 852.65 B: Well this is this is number two.
```

The largest corpora of conversational telephone speech, for example, the Fisher English corpus parts 1 and 2 [1-58563-313-5, 1-58563-335-6] are now measured in the thousands of hours.

High-end systems for collecting conversational telephone speech may consist of a server augmented by the same kind of telephony hardware and software used to manage interactive phone systems. These systems can both initiate and accept calls. With custom-written software they may be programmed to acquire informed consent, authenticate and pair subjects, describe the collection protocol, guide subjects through a session, ask questions, accept answers as speech or as selections from multiple choices with the help of push button tones, and record speech either separating or mixing the audio from the participants. Lower end solutions exist. However, those that collect speech at the phone of one of the participants reduce the mobility of the effort by forcing one participant to always be at the collection point or else to carry the collection system with her. Such systems add unwanted variability to the quality of the collection by recording one side of the conversation at the source and the other after transmission through a telephone network. Some may mix the near and far sides of the conversation into a single stream loosing the ability to separate them subsequently.

In contrast, low-end systems for collecting broadcast speech may be quite effective. Digital video recorders have, at the time of writing, become popular on the consumer market due to the broad availability of digital video and large capacity hard drives. Although many consumer-oriented systems prevent access to the video signal, there are now dozens of video tuner boards, digital video cameras-and software interfaces that allow digital video to be streamed directly to computer disk and accessed independently. For large-scale collection, a more customized solution may be desirable. For example, LDC's current broadcast collection system includes satellite dishes and receivers to capture both proprietary signals and standard C and Ku band signals, as well as wideband and shortwave antennae and cable television. Each of the dishes and antennae are connected to one or more receivers whose output is routed through an audio- video matrix to a number of processors. Servers stream digital audio and video to disk or else digitize analogue signal and then store it. Closed caption decoders extract captions from the video signal and write them to disk as a rough time-aligned transcript of some broadcast sources. Audio is also streamed to systems running best-of-breed commercial speech to text software in order to provide time-aligned transcripts. Video, audio- and transcripts are stored on spinning disk with coordinated file names and are written to tape as a back-up. The entire operation is automated by a master computer that maintains a database of programs to record, the times and channels on which they appear, instructions for tuning dishes and receivers, and indications of whether closed captioning is to be extracted or automatic transcripts created. This system automates broadcast collection so that it runs over nights, weekends, and holidays without human intervention. However, it is important to note that such automation requires ex post facto human auditing since programs may be pre-empted and programming schedules may change.

6.2.4 Communicative interactions in data supporting speech technology development. Speech data from broadcast news, telephone conversations, and multiparty meetings vary significantly along a number of other dimensions. Figure 3 lists several interactions and indicates how they affect the difficulty each presents to human annotators and speech systems.

The three speech types differ with respect to the degree of variability present in the physical environment and in the audio capture equipment used. In broadcast news, most speech takes place in the studio environment where high quality equipment and quiet prevail. When news broadcasts contain audio of correspondent reports, telephone calls, or satellite transmissions, humans notice the difference in quality and the performance of speech-to-text systems degrade. Recognition of telephone speech suffers from the greater variability present in the physical environment in which the speakers find themselves. Broadcast news and telephone conversations also differ with respect to both movement in situ and change of location as factors affecting the ability to recognize speech. Broadcast news personalities tend to sit in a single place and minimize movements that would create noise. Conversational speech lacks this level of discipline. Not only may participants generate additional noise through movements but they may also change their location relative to the data capture devices either by moving a phone away from their mouths, by walking out of range of a wireless phone base or, in the meeting environment, by

	Broadcast	**Telephone**	**Meetings**
Variable Environment		🟨	🟥
Variable Capture		🟨	🟥
Movement		🟨	🟥
Change of Location		🟨	🟥
Multimodality	🟥		🟥
Informality		🟥	🟥
Impromptu Speech		🟥	🟥
Overlapped Speech			🟥
External Apparatus	🟨		🟥
Multiple Speakers	🟨		🟥
Information Handicap	🟨		🟥
Observer's Paradox			🟥
Readable Transcript	🟨	🟨	🟥

Increasing Challenge =>		🟨	🟥

Figure 3. Comparison of human interactions underlying three speech data types.

walking alternately toward and away from room microphones. Broadcast news does present a greater challenge than telephone conversation in its multimodal signal. The modern television broadcast may contain not only the video of the on-air personality but also background images, closed captioning, sidebar text, and the horizontally scrolling text, or “crawl”, that CNN, for example, conspicuously employs. Integrating these sources of information is an open research problem for information management technologies. Broadcast news speech, relatively formal and well-rehearsed, contains a narrower variety of linguistic styles, and fewer disfluencies and rapid speech phenomena than conversational speech. In telephone conversations the number of speakers is usually small and fixed while in broadcast news there may be studio guests, call-ins, and man-on the-street interviews. *Information handicap* refers to the paucity of information the annotator or recognition system has relative to the participant in a communicative interaction. During telephone conversation, a recognition system has as much signal data as the interlocutors. However in meetings and broadcast television, the facial expressions, maps, visual aids, etc. that help to disambiguate the audio for participants are lacking in the audio signal generally provided to systems. The *Observer’s Paradox* states that in order to understand human communication one must study it even though the very act of observation affects the phenomena under study. Broadcasters know they are being watched for their news content and their register is formal as a result. Observation by speech researchers has no additional impact in this case. Among LDC telephone collections, there is both evidence that participants believe they should monitor their speech and evidence that they sometimes forget to do so. The effect of observation has the potential to be the most profound in meetings where special rooms may be required and where microphones may be in plain sight.

Case Study: Collection of Conversational Telephone Speech

This section summarizes observations from LDC’s experience covering five phases of Switchboard collections (Cieri et al., 2003), four Fisher collections and two Mixer collections (Cieri et al., 2006). All three types of collection recruit large numbers of subjects to complete conversations on assigned topics with other subjects in the study. In Switchboard studies, subjects are encouraged to participate in up to 10 six-minute telephone conversations. Because, Switchboard has been used primarily to support speaker identification technology development during a time when the research focused on low level acoustic features, not all of the Switchboard data has been transcribed. On the other hand, special care was taken to verify speaker identity. The behaviour of the robot operators that enable the studies differs somewhat in each case. The Switchboard robot operator, waits for an incoming call from one of the subjects at which time it initiates outbound calls, using a single line, to a series of participants until one accepts the call. The Fisher robot operator takes control of the call flow by initiating calls simultaneously to a number of subjects pairing them as soon as it has two on hold who agree to participate. In Mixer studies, the robot operator operates similarly but the goals of the collection, speaker recognition in a multilingual, multi-channel environment, led to changes in the types of subjects recruited, in collection parameters such as the increase in the number of calls a subject could make and in equipment aspects, such as the microphones and handsets used.

Conversational telephone speech proceeds in two phases, recruitment and collection. The importance of recruitment is sometimes overlooked. Without participants, there is no data. Experience from previous studies has shown that several factors in the recruitment process can have profound effects on the collection's outcome. These include the time of year in which the collection takes place. For example, in Switchboard Cellular Phase I (1999–2000), the requirement that participants make a minimum number of calls from locations outdoors led to the study beginning 3 months earlier than planned simply to avoid the winter months. It was also observed that restricting the hours during which participants can make calls raised the probability that they would actually reach another available participant by concentrating call activity into a small number of hours. In most cases, recruitment must begin several weeks prior to the beginning of collection. When recruitment occurs too far in advance, participant interest wanes prematurely. When collection begins too soon, the lack of a critical mass of available participants may frustrate callers. The best recruitment efforts, however, are only as good as the technology that supports them. Recruitment requires a reliable database of participants and a user-friendly interface to support the recruitment team. Subject data generally includes: name, gender, age, education, occupation, location born and raised, and where appropriate, ethnicity. For purpose of payment and participant care, contact information and identifying numbers, such as social security numbers in the United States are also crucial. Generally speaking, this data is collected during the initial discussion between the participant and the recruitment staff. Indeed, that may also be the only time the recruiters speak directly with a participant.

LDC generally advertises via print media and electronic announcements. Potential participants contact the LDC via phone or e-mail, or by completing electronic forms whence they learn: (1) that speech will be recorded for research and educational purposes, (2) that personal information will be kept confidential, not be released with the data, (3) when the study begins and ends and how to participate, (4) how, how much, and when they will be compensated.

Registered participants receive detailed written instructions that reiterate everything discussed in person and sketched on the project's web pages. In telephone studies, the instructions include a person identification number (PIN) and the series of prompts that participants will hear. Some conversational studies (Switchboard, Mixer, Fisher) require a critical mass of recruits before they can begin. In other protocols (CallHome) a participant can begin immediately after registering. In either case, participant compliance is closely monitored to ensure a successful study. If a study does not proceed according to plan, adjusting study parameters including the number of recruits, their demographics, and their compensation may prove helpful.

Collection systems must be accurate, reliable, economical, and capable of delivering real world data. For broadcast, telephone, and meeting speech, LDC has developed robust systems that leverage off-the-shelf hardware. The telephone system consists of customized software, telephony hardware, and a project database and can record multiple simultaneous conversations with no need for operator intervention. The database contains demographic information and call activity statistics for each participant and supports all recruitment, collection and reporting. Call activity is logged each time a participant tries to make a call, or receives one.

The LDC's meeting recording system can record 16 tracks of digital audio from a mixture of wireless and far-field wired microphones. Lavalier or head-mounted microphones are used for close recording. Room microphones, including a microphone array and PZM, omnidirectional and directional microphones are also used. The meeting recording system consists of a digital mixer, a multi-track digital tape recording deck, wireless microphone receivers, a microphone preamplifier, and a multi-channel digital audio computer interface. Meeting sessions are recorded as 16 bit/44 kHz PCM audio.

6.2.5 Human subject behaviour. The goal of Switchboard Cellular Phase I, was to collect 10 six-minute calls from 190 GSM cellphone users balanced by gender. The most successful recruiting effort involved employees of a local GSM provider, in which 293 participants were recruited. Unfortunately calls to many of the registered phones went unanswered during times the subjects had agreed to receive calls. This proved to be a result of participants' habit of turning off their cellphones when not using them. To counter this problem and to generally improve customer care, LDC initiated multiple participant call-backs and mailings and a participant lottery for those who completed the study. Although this study had a high rate of success in terms of subjects who completed the required number of calls, it was very labour-intensive. Switchboard Cellular Phase II included several adjustments to these challenges.

The goal in Switchboard Cellular Phase II was 10 calls each from 210 participants balanced by gender with no restriction on cellular network. LDC recruited 591 participants and instituted a sliding pay scale that covered subject costs for each call while simultaneously providing strong incentives to

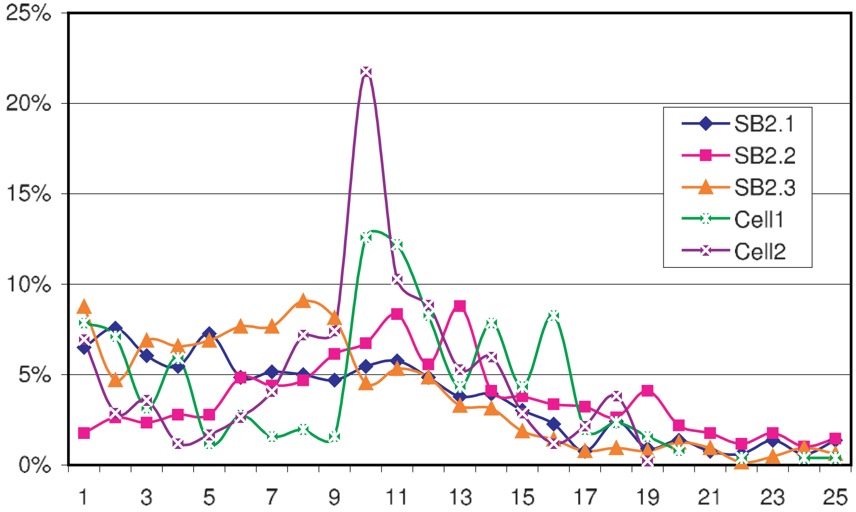

Figure 4. Switchboard call summary. The vertical axis shows the number of participants who made the number of calls on the horizontal axis.

complete the study. As a result of these measures, LDC was able to complete Switchboard Cellular II in about 1 month. Figure 4 shows participant behaviour in each of the Switchboard collections. Switchboard Cellular 2 has the tightest distribution of subjects around the goal of 10 calls. For Switchboard 2.1–2.3, the goal was to collect an **average** of 10 calls per participant. Although these studies eventually met their goals, Figure 4 shows a very diffuse distribution of participant performance. In the Cellular studies, the goal became having a minimum number of subjects who participated in at least 10 calls. The labour-intensive approach adopted in Switchboard Cellular 1 produced a funny distribution of subject performance and was costly in terms of recruiter effort. The approach used in Switchboard Cellular 2 produced a distribution that is very tightly centred around a mode at 10 calls and was in every other way, more efficient.

6.3 Segmentation

As noted previously, *segmentation* is the actual or virtual division of a recording of communicative performance into pieces to facilitate annotation, search or analysis. Similar to records and fields in a structured database or chapters and sections in a book, time-aligned segments allow the user to zoom in on target phenomena, for example, by searching the transcript and then using the time stamps to play the corresponding audio. Segments may correspond to pause or breath groups in conversational speech, speaker turns in conversation or broadcast news and to stories in a news text collection, speaker turns, stories, or sections in an audio or video collection. More recently the segments correspond to “SU”s, extents whose syntactic and semantic properties suggest they be treated as units. The granularity of segments will depend upon intended use and may change over time. Stand-off segmentation is generally more suitable and adaptable to change both because it allows for multiple segmentations of these files and because it preserves the first generation digital artefact.

Naturally, most annotation is dependent upon segmentation. Formalisms and tools may have a constraining effect on annotation possibilities. For example, some assume that each recording of linguistic performance must be uniquely and exhaustively segmented. The strong form of this approach allows for one and only one segment at each time sample in an audio or video file and at each character offset in a text file. In this case, creating a segment that occupies the first second of an hour-long conversation has the side effect of creating a second segment that is 59 minutes and 59 seconds long. The start time of the first segment must be zero, the end time of the last segment must be 60 minutes 0 seconds, and the end time of any segment after the first must be equal to the start time of the following segment. This approach is problematic for speech involving more than one speaker where overlaps occur. A weaker form of the same approach applies the same constraints within an annotation tier but allows for multiple tiers. This approach accommodates overlapping speech but is still awkward for partially annotated files because it forces the portions that are not annotated to be either included in nonce segments, increasing segment count or else attached to annotated segments increasing the duration of segments unnecessarily. The alternative approach is to allow segments to be independent of each other. The strong form of this alternative approach allows multiple tiers of annotation and removes the constraints that govern start and end times of adjacent segments thus allowing some to overlap and allowing gaps between others; segments are no longer forced to abut. The disadvantage of this approach is that an error in which a human annotator fails to include interesting material in a defined segment results in that material being overlooked during annotation by default. Annotation projects that use this approach to segmentation need a quality control pass in which annotators or systems look for annotatable material between defined segments.

6.4 Annotation

As previously noted, we define annotation as any process that adds value to raw data through the application of human judgement. That judgement may be applied either directly by human annotators or automatically with humans specifying the rules that systems implement. Bird and Liberman’s (2001) survey of a multitude of annotation practices showed a common theme. Formally, annotation may be seen as the attachment of one or more category/value pairs to segments of a corpus. Categories indicate the type of annotation. Each type may have one or more values. For example, a two layer orthographic and phonetic transcription of prompted speech, such as we find in the TIMIT corpus (Garofolo et al., 1990), might have two categories, *word* and *phone*. For an utterance of the word "So", annotations would include *word=so* and *phone=s*. The duration of the segments ranges from relatively small to relatively large. There are several TIMIT phone tags per second of speech. On the other end of the spectrum, categorizations of an entire recording may be formalized as metadata or, to keep the formalism consistent as annotations with time spans that equal the length of the recording. Where a category may have only one value, either the category or value label may be excluded. Where a corpus contains a single kind of annotation, bare values without category labels are often given.

Annotation varies with respect to the expertise required and the variability expected. *Expert annotation* requires specific background and advanced skills. The syntactic annotation of Treebanks and the entity, relation, event, and co-reference tagging of the ACE program are examples of expert annotation. In the former case, the successful annotators have generally been college graduates, graduate students, or post-doctoral researchers whose secondary education included a specialization in the syntax of the language under study where that language was also the medium of instruction. Even such highly skilled scholars spend months perfecting their knowledge of the specification before they can be fully productive annotators. Although individual variation plays some role in the quality of expert annotation, one expects inter-annotator agreement to increase with training and collaboration. *Intuitive annotation*, where the goal is to capture the judgement of an average speaker or potential technology user, requires less specific training. Sometimes native speaking ability in the target language is enough. The specification and tools are also generally simpler and one expects more variation among annotators. Translation in non-technical domains is a kind of intuitive annotation where variation may be extreme. In some cases, the annotator may also act as a user model for a given technology. Topic annotation within the Topic Detection and Tracking (TDT) and High Accuracy Retrieval from Documents (HARD) use the annotator as user model.

Linguistic Resources may also be differentiated as to whether they serve a very specific purpose such as the topic and entity tagging for information retrieval and extraction, or provide general knowledge such as part-of-speech tagged text, and translation lexicons.

6.5 Quality Assurance and Inter-Annotator Agreement

Annotation tasks naturally vary according to level of difficulty. The distinction between intuitive and expert annotation sketched above impacts both the amount of inter-annotator agreement and its importance to human language technologies.

The goal of some collection and annotation tasks is to sample the variability that exists in a human population. For example in the first months of DARPA TIDES' sponsorship of machine translation evaluation, a critical task was to develop an objective measure of translation quality knowing that for any source language text there may be many very different translations that are nonetheless equally valid. The solution adopted by the community was to create multiple human translations of the same source language text and, grossly speaking, credit systems that produced translations that were attested in any of the human translations. Human translators were given very few constraints. They were required to produce a direct translation, avoiding summary and exegesis, that was grammatical in the target language and as faithful to the original text and its cultural matrix as possible. Their atom was the sentence of source language text. They were required to produce one or more sentences of translation for each sentence of source. Otherwise, they were left to their own discretion. In this case, it is important to model inter-translator variation. However, any attempt to force the translators into greater conformity risks distorting the model and the resulting evaluation of system performance.

In contrast, the syntactic annotation of text or transcribed speech is an example of expert annotation in which there is generally assumed to be a right answer. Annotators are expected to be highly trained in general syntax, the syntax of the target language and the annotation specification and the quality of their work is expected to increase with training and experience assuming a positive disposition toward the work. Measures of inter-annotator agreement in syntactic annotation are useful in determining how difficult the task is. At the same time, ongoing training, error analysis, and similar measures that increase agreement are valid.

With that background in mind we distinguish four kinds of quality control (QC): precision, recall, discrepancy, and structure. *Precision QC* attempts to find incorrect assignments of an annotation. Annotators review each case in which a data span has been given an annotation and verify that the annotation is appropriate. Unless an annotator misunderstands the specification, mistakes of this kind, false alarms, should be relatively less common than the next type we will discuss. Where annotations are sparse, a greater percentage of annotation may be submitted to precision QC. For example, LDC reviewed 100% of all annotations in the TDT corpora where a small number of news stories will have been marked as relevant to a given topic. *Recall QC* attempts to find failed assignments of an annotation. Annotators review segments where an annotation was not applied to verify that it should not have been. Errors of this kind, misses, result from waning attention and are relatively more common among human annotators. The search for misses may employ computer assistance, for example, a search engine may identify documents with high relevance scores for a given topic for a human annotator to review. *Discrepancy QC* reviews annotations of the same data done by multiple independent annotators. Depending upon the nature of the annotation discrepancies may be used to calculate scores of inter-annotator agreement, to identify cases in which annotators misunderstand the specification, to identify cases in which the specification fails to treat specific phenomena, and to identify cases that require a judgement call. Naturally, some findings may lead to revision of the specification or scoring metric. Others may lead to remedial training of annotators. Generally, LDC performs discrepancy analysis on 5–10% of all annotated data using a double-blind protocol. Finally *Structure QC* uses facts about relations among annotations in order to identify suspect annotations. To take a simple example, in Arabic a prepositional phrase may occur within a noun phrase, as in "the woman from Tunis". However, structurally, the PP tag must actually be subjacent to the N tag not to the NP tag directly. In bracketed notation, this structure is (NP(N(PP…))) and any case of (NP(PP…)) is suspect. Once the rules have been established, Structure QC can generally be done automatically so that one expects 100% of all annotated data to be subject to this level of scrutiny. XML validation and checks of audio file headers are also species of Structure QC.

6.6 Preparation, Distribution, Adjudication

In preparation for release, corpora are generally associated with metadata and documented to indicate the authors and annotators of the data, the volume and types of raw material included, the percent annotated, the annotation specification, and the quality control measures adopted. Although authorship may seem the simplest of these for corpora of linguistic data it is often difficult to identify the author because corpus creation was, and often still is, viewed differently where authorship is concerned than writing academic papers or presenting at conferences. Furthermore, there is no standard for determining what kind of contribution to a corpus counts as authorship. For a corpus of annotated conversational telephone speech, the subjects, transcriptionists, other annotators, their immediate managers, senior managers or principal investigators, financial and contracting personnel, sponsors' technical representative, and their management will all have contributed to the realization of the corpus.

In some cases it may be necessary or just preferable to annotate data in a different format than the one in which it is distributed. To give a simple example, transcripts that include markup for disfluency, non-lexemes, partial words and the like will be more readable for humans if these items are tagged simply and in-line. On the other hand to permit robust processing of the transcripts, these simpler user tags may be converted into a mark-up language, for example, XML. In large-scale projects involving rotating teams of native speakers of different languages who must learn complex annotation specifications and execute them with consistency, LDC gives relatively high priority to simplifying annotation even if that means reformatting prior to release.

Despite multiple passes of quality control, errors still find their way into corpora. Adjudication processes may help reduce such error. Adjudication typically involves reviewing two or more independent annotations of the data to identify and resolve areas of disagreement. The independent annotations may involve humans or machines or both. During the creation of the TDT corpora, adjudication was used to identify human annotator errors in the assignment of documents to topic clusters. This process was implemented when the project abandoned *exhaustive* annotation in favour of *search-guided* annotation. *Exhaustive* annotation in the case of the TDT-2 corpus meant that each of more than 50,000 documents were compared against each of 100 topics yielding more than 5,000,000 decisions. In TDT-4 and TDT-5, exhaustive annotation was replaced by search guided annotation in which a search engine seeded with keywords from topic descriptions and text from on-topic documents, searched the corpus and returned a relevance ranked list of hits. Human annotators then reviewed those hits looking for truly on-topic documents and following consistent rules to decide how many documents to review. Because the success of this method could have been skewed by problems in the search engine, the results were adjudicated in the following way. Once the evaluation had taken place, the results from the tracking systems developed by the several research sites where compared to each other and to LDC's human annotators. Documents were ordered with respect to how many research systems disagreed with the human annotators. Human annotators then proceeded in priority order through the cases where the majority of systems disagreed. Figure 5 shows the results of this adjudication and confirms something that makes sense intuitively. Humans using search-guided annotation to decide whether a news story discusses a specific topic are more likely to miss relevant stories than they are to erroneously judge a story to be relevant. When all seven systems concluded that the LDC human annotator has missed a relevant story, the systems were correct 100% of the time. Otherwise the human, search-guided annotation generally made the right decision in the majority of cases. The human false alarm rate was very low. Even in the few cases in which all systems disagreed with the human judge who thought the story was on-topic, the human was generally correct. Naturally such adjudication can also be used to validate annotation based only on human effort. In some cases, for example, in the TREC cross-language document retrieval track, adjudication is used instead of human annotation of an evaluation corpus. In other words, the evaluation corpus is given

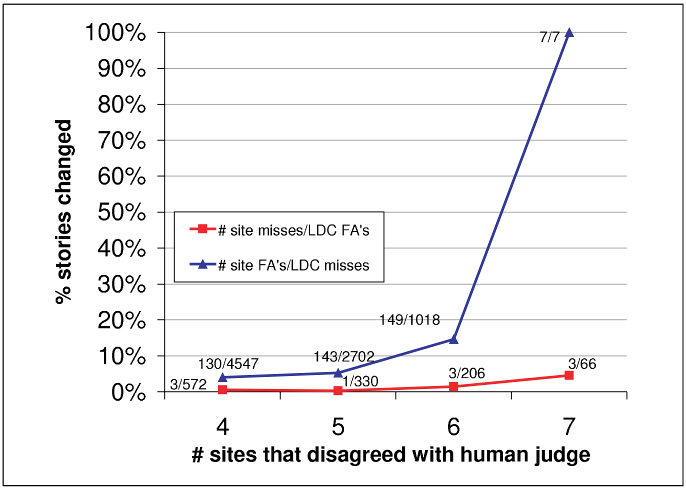

Figure 5. Results of human-system adjudication. In search-guided topic annotation of news, human errors are more commonly misses than false alarms.

to sites without annotation. System results are pooled using a process similar to the above and then adjudicated by humans to produce the final answer key and system scores.

7 Conclusion

This chapter has tried to give a perspective on the creation and sharing of language resources for purposes of technology development and evaluation informed by experience within the rubric of common task technology programs described above. It bears repeatation that there are other perspectives on language resource creation informed by other experiences and moulded by other approaches to research management. A theme that runs throughout the research communities working in linguistic education, research and technology development is the increasing use of language resources. Each year new communities embrace the practice of sharing language resources. The communities that had established that practice a decade or more ago, continue to rely upon shared resources of ever-increasing sophistication, diversity and volume. Technological advancements endow the average desktop with the ability to create and share small and medium-scale resources. Perhaps the greatest challenge currently facing HLT communities is the generalization of technologies developed for a small number of languages. Work in resource sparse languages also termed "low density" or "less commonly taught" offers both the difficulty of truly generalizing technologies to handle human languages of startling diversity as it also offers the rewards of improved communication and access to information leading to improved understanding.

References

Bird, S. and Liberman, M. (2001). A Formal Framework for Linguistic Annotation. *Speech Communication*, 33(1–2):23–60.

Chapman, S. and Kenney, A. R. (1996). Digital Conversion of Research Library Materials: A Case for Full Informational Capture. *D-Lib Magazine*. http://www.dlib.org/dlib/october96/cornell/10chapman.html.

Cieri, C., Andrews, W., Campbell, J. P., Doddington, G., Godfrey, J., Huang, S., Liberman, M., Martin, A., Nakasone, H., Przybocki, M., and Walker, K. (2006). The Mixer and Transcript Reading Corpora: Resources for Multilingual, Crosschannel Speaker Recognition Research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 117–120, Genova, Italy.

Cieri, C., Miller, D., and Walker, K. (2003). From Switchboard to Fisher: Telephone Collection Protocols, their Uses and Yields. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1597–1600, Geneva, Switzerland.

Cieri, C., Miller, D., and Walker, K. (2004). The Fisher Corpus: A Resource for the Next Generations of Speech-to-Text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 69–71, Lisbon, Portugal.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). Automatic Content Extraction (ACE) Program - Task Definitions and Performance Measures. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 837–840, Lisbon, Portugal.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1990). The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. [CD-ROM] US Department of Commerce, Gaithersburg, MD.

Guy, G., editor (1999). *Symposium on Public Access Data Bases, NWAVE 28: The 28th Annual Conference on New Ways of Analyzing Variation*, Toronto, Canada.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, New Jersey.

Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington, DC.

Mariani, J. (2002). Are We Losing Ground to the US? A Contrastive Analysis of US and EU Research Frameworks on Human Language Technologies. http://www.hltcentral.org/page-975.0.shtml.

Papinieni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, USA.