A. Innovative Claims

The essential basis of all language-related technology is linguistic data: speech, texts, and lexicons, along with various annotations, linkages and derived information. The Linguistic Data Consortium (LDC) was founded in 1992, with seed money from DARPA, in order to create, collect, publish and distribute such data.

Creation of linguistic data can be expensive and time-consuming. As demand grows for data in many languages, and for rapid development of capabilities in new languages with little advance notice, both the expense and the time become more and more problematic. The LDC proposes to create large volumes of multilingual data at low unit cost, with three goals: to support the multilingual research and development within the TIDES program; to build up a ”bank account” of multilingual data that can be used for future needs; and to streamline methods for meeting future needs in new languages as quickly and cheaply as possible. In order to make this (and other) data accessible to researchers, we also propose to develop (jointly with NIST, MITRE and other interested parties) a set of standards and tools for creation, modification, visualization, search and retrieval of multilingual linguistic resources of all types. These standards and tools are essential for our own development of multilingual resources on the necessary scale, and they are also essential for efficient transfer of such resources to the research community, as well as for the development of inter-operable software for TIDES research, development and eventual deployment.

We will create large amounts of low-cost multilingual data by taking maximal advantage of existing resources in three areas, in each case continuing and expanding our previous work. First, we will collect a large corpus of VOA broadcasts and scripts, in up to 45 languages, over a coordinated time period, taking advantage of our status as a VOA satellite downlink station, and the law authorizing VOA to cooperate with us in such endeavors. Second, we will produce digital database versions of paper bilingual dictionaries. Some 37 of these have previously been produced by our group; we propose to digitize and obtain distribution rights to TIDES-related dictionaries at a rate of six per year. Third, we will search out bilingual text corpora, on the internet and elsewhere, obtain IPR releases for redistribution, and clean them up and align them. This continues our recent web-based effort that has resulted in three Mandarin/English parallel corpora totalling about 21 million words.

The experience of the past decade shows that as the number of languages increases, and the scope of systems broadens, common standards and tools become increasingly important. Without flexible and general formats for transcriptions, texts, annotations, lexicons and so on, researchers will waste much time modifying tools and data to fit together, and many bugs will be introduced despite their best efforts. We have begun the development of the ATLAS framework, jointly with NIST and MITRE, precisely to address these issues, and propose to complete it under the requested funding.

B. Technical Rationale

The Voice of America Data Collection

The Voice of America's charter, codified by U.S. Public Law 94-30, states that "the long-range interests of the United States are served by communicating directly with the people of the world by radio." To achieve these goals, Voice of America broadcasts news, cultural and entertainment programming around the world and around the clock in dozens of languages. Currently VOA broadcasts in 45 languages: Afan Oromo, Albanian, Amharic, Arabic, Azerbaijani, Bangla, Bosnian, Bulgarian, Burmese, Cantonese, (Haitian) Creole, Croatian, Czech, Dari, Farsi, French, Greek, Hausa, Hindi, Hungarian, Indonesian, Khmer, Kirundi/Kinyarwanda, Korean, Kurdish, Lao, Latvian, Lithuanian, Macedonian, Mandarin, Pashto, Polish, Portuguese, Romanian, Russian, Serbian, Slovakian, Swahili, Thai, Tibetan, Tigrinia, Turkish, Urdu, Uzbek and Vietnamese.

Until recently, the Voice of America's charter has forbidden distribution of their materials within the United States, thus denying American researchers access to this rich resource. In 1996, however, Congress enacted Public Law 104-269, which authorized Voice of America to cooperate with the Linguistic Data Consortium at the University of Pennsylvania to make VOA programming available for use in research, education and technology development. With advice from VOA and technical assistance from VOA-recommended consultants, LDC has installed a satellite downlink station that can access all of the audio channels for all of VOA's worldwide broadcasts, in digital form. LDC and VOA have also established mechanisms for LDC access to all of VOA's internal programming scripts, including a large amount of material that is not posted on the VOA web site. Most of these scripts are now prepared and archived internal to VOA in an accessible electronic form.

Since 1996, the Linguistic Data Consortium has collected VOA broadcasts in English, Mandarin, Spanish and Czech for use in corpora provided for a number of federally sponsored research projects, and has also published (or will publish) the resulting corpora for general use in research and education. DARPA's Hub-4 speech recognition projects have used broadcast news in Mandarin and Spanish, while the DoD-sponsored Topic Detection and Tracking project has benefited from VOA broadcasts in English and Mandarin. VOA Czech broadcasts are currently being used as data for the NSF-sponsored summer workshop "Language Engineering for Students and Professionals Integrating Research and Education" at John's Hopkins from July 12 to August 20, 1999.

The LDC catalogue already includes 12 broadcast news corpora in English, Spanish and Mandarin. Another major broadcast news collection, TDT-2, will be released during the summer of 1999. TDT has been LDC's first bilingual Broadcast News collection, combining some 52 thousand English news stories and 20 thousand Mandarin news stories collected from three newswire, three radio, four television and one worldwide web source over a six-month period in 1998.

Through these projects, LDC has established a viable procedure for collecting VOA material, and the research community has shown that this material is useful in research and development. In the context of the TIDES initiative, it makes sense to harvest a much larger sample of this interesting stream of material, which is not otherwise archived, except (by statuatory requirement) on extremely low-fidelity high-capacity analog tape equipment at VOA headquarters.

Voice of America satellite broadcasts are structured so that English, Spanish and Portuguese programs sent to Latin America appear on special sidebands while the majority of programs are carried on a pair of T1 signals designated T1-1 and T1-2. T1-1 carries broadcasts to Europe and Africa while T1-2 carries broadcasts to Asia.

In order to collect VOA broadcasts, LDC has acquired and installed a satellite downlink station consisting of a 15-foot C-band satellite dish. The dish collects digital audio from Intelsat, a geostationary communication satellite that VOA uses for transmissions to its various broadcast stations around the world. The Intelsat signal is demodulated via a Radyne 2401RX demodulator and split into the two T1 digital signals and three sideband radio signals. The individual T1 signals are demultiplexed into MPEG encoded channels using an Ascon Timeplex Mini-Link/2+ with associated signal processing components (QSP.2 and ILC.3 boards). Individual broadcast channels are decoded using Comstream ABR200s, and collected in digital form on computer disk using Townsend DAT Links attached to a dedicated SunSparcstation. The current configuration allows LDC to collect up to two sideband channels and up to three pairs of broadcast channels from either T1 signal. All of the current equipment was purchased with funds from LDC membership fees and data sales, at a total cost of about $150K.

Although VOA currently broadcasts around the clock in 45 languages, distributing material to its various broadcast stations on 24 satellite channels, a careful analysis of the VOA satellite schedule shows that LDC can collect up to an hour of news per day in each language with capture stations serving just fourteen channels: two for the sidebands serving Latin American; eight for T1-1 serving Europe and Africa; and four for T1-2 serving Asia. The equipment requested in the attached budget will upgrade LDC's task specific hardware to accommodate all the channels necessary for a complete VOA collection.

Table 1 shows the VOA broadcast schedule as of June 1999. The first and second columns show the language of a broadcast and the geographic region targeted. For example, VOA French broadcasts target Francophone Africa, not Europe. The third column categorizes broadcasts as: exclusively news, features (where 80% of the broadcast, measured in minutes, is feature oriented) or mixed news and features. The fourth and fifth columns indicate days and hours of broadcast in universal time unless otherwise noted. This table excludes VOA English and Spanish broadcasts since LDC is already collecting these and already has adequate facilities to continue doing so.

For a simplified picture of this broadcast schedule, we can ignore the distinction between news and features, and assume that all programs run for a complete hour on the hour seven days a week. This yields the schedule in table 2, where the columns contain the hour in UTC, the T1 channel and the languages in which VOA broadcasts during that hour on that channel. This will determine the worst cases of simultaneous recording ever required.

If we were to plan for the worst case scenario, we would need 11 recording stations to cover all languages broadcast on T1-1 at 1800 UTC and 5 stations to cover all languages broadcast on T1-2 at 0000 UTC and at 1300 UTC. The connections between channels and recording stations are not program-settable, but rather determined by physical wiring in the rack. Thus without introducing a programmable switch, or relying on human labor to switch patch cords several times a day, we would seem to require 16 recording stations. However, several of the languages broadcast at the busiest times are broadcast in many other time slots as well. Mandarin Chinese, for example, is broadcast in 11 different time slots. This redundancy in the broadcast schedule will allow us

Afan Oromo	Africa	Mixed	Mo-Fr	1845-1900
Albanian	Europe	Mixed	Mo-Su	0500-0530 1200-1215 1300-1315 1400-1500 1600-1630 1830-1900
Amharic	Africa	Mixed	Mo-Fr	1800-1830
Arabic	Middle East	Mixed	Mo-Su	0400-0500 0730-0830 1900-2100
Arabic	Middle East	News	Mo-Su	1700-1900
Azerbaijani	Middle East	Mixed	Mo-Su	1830-1900
Bangla	Asia	Mixed	Mo-Su	0130-0200 1600-1700
Bosnian	Europe	Mixed	Mo-Fr	1500-1530
Bulgarian	Europe	Mixed	Mo-Su	0300-0330
Burmese	Asia	Mixed	Mo-Su	2330-0000 1130-1200 1200-1230
Cantonese	Asia	Mixed	Mo-Su	1500-1700
Creole	Caribbean	News	Mo-Fr	1230-1300 1730-1800 2200-2230
Creole	Caribbean	News	Sa-Su	1730-1800
Creole	Caribbean	Mixed	Sa-Su	2200-2230
Croatian	Europe	Mixed	Mo-Su	0630-0700 1730-1800 1930-2000 2030-2100 2200-2230
Czech	Europe	News	Mo-Fr	0530-0600 1230-1245 1930-2000
Czech	Europe	Mixed	Mo-Fr	2100-2130
Czech	Europe	News	Sa-Su	1510-1530
Dari	Asia	Mixed	Mo-Su	0200-0230 1515-1600
Farsi	Middle East	Mixed	Mo-Su	0300-0400 0400-0430 1700-1800 1800-1900
French	Africa	News	Sa	1830-1840 1930-1935 2050-2057
French	Africa	News	Su	1830-1900 1930-1935 2050-2057
French	Africa	News	Mo-Fr	0530-0545 0600-0615 1830-2000 2100-2126
Greek	Europe	Mixed	Mo-Su	1645-1700 1900-1930 1900-1920
Hausa	Africa	Mixed	Su	0500-0530 1400-1430 2030-2100
Hausa	Africa	Mixed	Mo-Fr	0500-0530
Hausa	Africa	Mixed	Sa	1400-1430
Hindi	Asia	Mixed	Mo-Su	0030-0100 1600-1700
Hungarian	Europe	News	Mo-Su	1800-1830 2100-2130
Indonesian	Asia	Mixed	Mo-Su	2200-2330 1130-1230
Khmer	Asia	Mixed	Mo-Su	1330-1500 2200-2230
Kirundi/Kinyarwanda	Africa	Mixed	Mo-Fr	0400-0430 0330-0430
Korean	Asia	Mixed	Mo-Su	2130-2200 1300-1400
Kurdish	Middle East	Mixed	Mo-Su	1600-1700
Lao	Asia	Mixed	Mo-Su	1230-1300 1300-1330
Latvian	Europe	News	Mo-Fr	0640-0700 1730-1736
Lithuanian	Europe	News	Mo-Fr	0600-0710 1730-1736
Macedonian	Europe	News	Mo-Fr	1330-1345
Mandarin Chinese	Asia	Mixed	Mo-Su	2200-2300 0000-0300 0700-1500
Pashto	Asia	News	Mo-Su	0130-0200
Pashto	Asia	Mixed	Mo-Su	1430-1515
Polish	Europe	News	Mo-Su	0630-0640
Polish	Europe	News	Mo-Fr	2200-2300
Portuguese	Africa	Mixed	Mo-Fr	0430-0500 1700-1730 1730-1830
Portuguese	Africa	Mixed	Sa-Su	0430-0500 1730-1800
Romanian	Europe	Mixed	Mo-Su	0330-0400
Russian	Europe	Mixed	Mo-Fr	1300-1400 1700-1800
Serbian	Europe	Mixed	Mo-Su	0330-0400 0530-0600 1700-1730 1930-2000 2100-2130 2200-2215
Slovakian	Europe	News	Mo-Su	0500-0515 1800-1830 2045-2100
Swahili	Africa	Mixed	Mo-Su	1630-1730
Thai	Asia	News	Mo-Fr	0605-0625 0630-0645 0830-0900
Thai	Asia	Mixed	Sa	0830-0900
Tibetan	Asia	Mixed	Mo-Su	0000-0100 1400-1500
Tigrinia	Africa	Mixed	Mo-Fr	1830-1845
Turkish	Middle East	Mixed	Mo-Su	1800-1900
Urdu	Asia	Mixed	Mo-Su	0100-0130 1330-1430
Uzbek	Asia	Mixed	Mo-Fr	1500-1530 1530-1545
Vietnamese	Asia	Mixed	Mo-Su	2230-2330 1230-1330 1500-1600

Table 1: Current VOA Broadcast Schedule

0	T1-2	Hindi, Mandarin Chinese, Thai, Tibetan, Urdu
1	T1-2	Bangla, Mandarin Chinese
1	T1-1	Pashto
2	T1-2	Dari, Mandarin Chinese
3	T1-1	Bulgarian, Farsi, Kirundi/Kinyarwanda, Romanian, Serbian
4	T1-1	Arabic, Czech, Farsi, Kirundi/Kinyarwanda, Portuguese
5	T1-1	Albanian, French, Hausa, Serbian, Slovakian
6	T1-1	Croatian, French, Latvian, Lithuanian, Polish
7	T1-1	Arabic
7	T1-2	Mandarin Chinese
8	T1-1	Arabic
8	T1-2	Mandarin Chinese
9	T1-2	Mandarin Chinese
11	T1-2	Burmese, Indonesian, Mandarin Chinese
12	T1-1	Albanian, Creole, Czech
12	T1-2	Burmese, Lao, Mandarin Chinese, Vietnamese
13	T1-1	Albanian, Macedonian, Russian
13	T1-2	Khmer, Korean, Lao, Mandarin Chinese, Urdu
14	T1-1	Albanian, Hausa
14	T1-2	Mandarin Chinese, Pashto, Tibetan
15	T1-1	Bosnian, Czech, Uzbek
15	T1-2	Cantonese, Dari, Vietnamese
16	T1-1	Albanian, Greek, Kurdish, Swahili
16	T1-2	Bangla, Cantonese, Hindi
17	T1-1	Arabic, Creole, Croatian, Farsi, Latvian, Lithuanian, Russian, Serbian
18	T1-1	Afan Oromo, Albanian, Amharic, Arabic, Azerbaijani, Farsi, French, Hungarian, Slovakian, Tigrinia, Turkish
19	T1-1	Arabic, Croatian, Czech, French, Greek, Serbian
20	T1-1	Arabic, Croatian, French, Hausa, Slovakian
21	T1-1	Czech, French, Hungarian, Serbian
21	T1-2	Korean
22	T1-1	Creole, Croatian, Polish, Serbian
22	T1-2	Indonesian, Khmer, Mandarin Chinese, Vietnamese
23	T1-2	Burmese

Table 2: VOA broadcasts by time and T1

to capture automatically up to one hour of programming in each language each day, with nearly complete freedom of choice, using a total of 14 recording channels. We can do nearly as well with 12.

This analysis allows us to propose the collection, over a period of a year, of multiple hours of data each week in every language in which VOA broadcasts, while adding only modest additional hardware costs beyond our current set-up (since per-channel hardware costs are significant). The proposed collection will be roughly 1.6 terabytes in size, so inexpensive near-line storage, in the form of a multi- terabyte tape robot, will be used to permit easy access to the collection without requiring a large investment in mass storage. We will work out procedures with VOA for automatically harvesting scripts and correlating them with broadcast intervals.

We can now do accurate automatic word-level alignment between digital audio and VOA scripts for three languages (English, Mandarin and Spanish). This alignment is accurate, even though significant portions of the audio are not covered by the scripts. To build an aligner for a new language requires only to create a minimal pronouncing dictionary and to train acoustic models. No language model is required, of course, and alignment will work with relatively crude acoustic models and relatively high OOV rates. We therefore expect to be able to create and use aligners for a large number of other DARPA-specified languages over the course of this project. Where no automatic aligner is available, scripts will be associated at the program or story level.

The resulting corpus will contain more than 14,000 hours of broadcast materials in 45 languages,

with overlapping scripts. It will provide a rich resource that can be organized, subdivided and annotated in many ways for many publications and projects. Some subsets suitable for particular uses will be published as shared data for common-task projects. The near-line storage system, along with the LDC's existing facilities for data publication, will enable an educator or researcher who wants a certain amount of material in a particular language, or across a certain set of languages, to get a "custom corpus" created to order. A large sample (of a size constrained only by available disk space) will be made available for interactive access via LDC-Online.

The presence of audio and text in a multiplicity of languages will support intensively multilingual automatic speech recognition of the kind necessary for the success of projects like Hub-4 and TDT (Topic Detection and Tracking) as well as TIDES. The parallelism in stories based upon original news reporting in English but then translated into multiple other languages by native speakers for rebroadcast also provides a unique resource for machine translation. The scripts themselves provide natural language text suitable for building language models. Finally, the fact that all of these broadcasts focus on the daily reporting of news from a single period in time means that this corpus will provide a unique laboratory for research into translingual information tasks including topic detection, information extraction and story summarization. A sample collection scheme is presented in table 3.

The average total number of minutes to be recorded daily is 2,325, or 38.75 hours. This amounts to 14,144 hours per year. With 16-bit linear samples at a 16 KHz sampling rate (as has been used in Broadcast News collections in the past) this amounts to 115.2 MB per hour, or about 1.6 terabytes for a year's collection before compression. Of course, the actual details of the collection would be adjusted to reflect sponsor interest and the detailed facts of VOA's (changing) schedule during the period of collection. However, we feel that a collection of this general size is technically straightforward, and will serve the research community well for years to come.

Parallel text corpora

Several promising directions in multilingual IP depend on the existence of large parallel text corpora. The well-known IBM research on statistical models for machine translation was based on a corpus of about 2.9 million English/French sentence pairs, since published by LDC along with other parallel text. However, follow-ups to that research have been hampered by a lack of available parallel corpora in other languages.

LDC has previously published two large parallel text corpora. The Canadian Hansards corpus (LDC95T20, English/French) includes both the IBM corpus and another disjoint portion of the Canadian Hansard archive of about equal size. The UN Parallel Text corpus (LDC94T4A, English/French/Spanish) includes the UN archives from 1988 to 1993; the English portion is about 48 million words. Both of these corpora were quite difficult to get, from a technical as well as a legal point of view, because the form of the data made access difficult. The Canadian Hansards originated as typographer's tapes in a proprietary low-level markup language, while the U.N. archives existed only as disk cartridges for an obsolete Wang multilingual word processor.

These days, the technical problems (at least for new material) are likely to be easier, since files are likely to be archived in higher-level or at least more widely-available logical and physical formats. One obvious example of easily available parallel material is text on the web – though

Language	Minutes Collected Daily
Afan Oromo	15
Albanian	60
Amharic	60
Arabic	60
Azerbaijani	30
Bangla	60
Bosnian	30
Bulgarian	30
Burmese	30
Cantonese	60
Creole	60
Croatian	60
Czech	60
Dari	60
Farsi	60
French	60
Greek	60
Hausa	60
Hindi	60
Hungarian	60
Indonesian	60
Khmer	60
Kirundi/Kinyarwanda	30
Korean	60
Kurdish	60
Lao	60
Latvian	30
Lithuanian	30
Macedonian	15
Mandarin Chinese	60
Pashto	60
Polish	60
Portuguese	60
Romanian	60
Russian	60
Serbian	60
Slovakian	60
Swahili	60
Thai	60
Tibetan	60
Tigrinia	15
Turkish	60
Urdu	60

Table 3: Proposed VOA sampling: average daily minutes per language

as yet there are few large parallel corpora in any language, and few languages with even modest parallel material available.

More than a year ago, we began searching the web nightly with a special web spider program that harvests sites with significant amounts of parallel text, and works out the detailed page-to-page and phrase-to-phrase alignments. In those cases where researcher interest and our resources

warrant, we plan to obtain permission to redistribute the harvested parallel, publishing the results in an assimilable form. So far, we are following this process though with three moderately large English/Chinese corpora from Hong Kong. These are the Hong Kong legal code (about 8 million words of English), the Hong Kong Hansards (about 6 million words of English), and an ongoing collection of new stories (about 7 million words of English, growing at about 600K words per month). We have permission in hand for these materials, and have released the legal code, with the other two corpora requiring some additional pre-publication preparation. Taken together, these amount to some 21 million words on the English side, which is perhaps a third the size of the IBM Hansard corpus, but still large enough for some interesting experiments. In any case, as far as we know this is the largest Mandarin/English corpus that is likely to become available.

After considerable study, we feel that our approach – in which we harvest the sites and obtain permission to redistribute a (cleaned-up) version – is necessary, at least for some kinds of material. The obvious alternative – publishing lists of URLs along with software to enable researchers to download their own copies – has the advantage of side-stepping certain copyright issues, but it has some serious disadvantages as well.

First, the web is ephemeral. Some of the most interesting sites (for instance the Hong Kong news mentioned above) have revolving content, with no on-line archive. Other sites may change their content erratically, or even disappear. In such cases, our method is the only way to guarantee even medium-term access for the research community as a whole.

Second, it is not clear that distributed downloading solves the legal problems. Many sites assert terms of use like this: "Subscriber" (defined as "each person who establishes or accesses a connection") "may download copyrighted material for Subscriber's personal use only" but "may not modify, . . . , create derivative works, or in any way exploit, any of the content, in whole or in part." The legal force of such unilateral assertions is unclear, but a company eager to avoid legal entanglements will be wary.

Many – probably most – parallel corpora of significant size are not yet web accessible. We know of interesting materials in the possession of various U.N. agencies, the World Bank, other NGOs, and many national and local governments. Depending on the needs and priorities of the TIDES program, we propose to obtain and publish such corpora, as well as focusing our web activities in TIDES-relevant directions.

Dictionary digitization

An especially crucial resource for TIDES will be "transfer lexicons" for IR and MT. Despite our best efforts and those of others, few TIDES languages will have adequate parallel text available to induce such lexicons statistically. Electronic versions of traditional paper bilingual dictionaries are one natural raw material for the development of transfer lexicons.

The Language Analysis Center (a University of Pennsylvania lab that merged with the LDC several years ago) worked for several years on a contract to produce digital database versions of paper dictionaries of various multilingual sorts. The practice was to scan and OCR the dictionary, after training any needed fonts; then to parse the scanned text in accordance with an SGML DTD for the dictionary structure, adding compliant SGML tags; and then to have native speaker editors post-

edit the text to correct OCR errors and introduce any segmentation that could not be inferred from the typography. Some 37 (Monolingual or bilingual) dictionaries involving 23 languages were digitized, from Africaans and Basque to Uzbek and Xhosa. The process was relatively fast and efficient, with each dictionary taking about one to two months of work, and costing about $15K. Unfortunately, IPR permissions for general research access were not part of the picture.

We suggest that access to such bilingual dictionary databases will be an important asset to TIDES research, and that it is also valuable to have in place a smoothly-functioning system for digitizing dictionaries and obtaining redistribution rights. Therefore, we propose to establish a new version of the general procedure previously used, with the addition of IPR negotiations. The same set-up can also be used to create text corpora for languages where electronic text is not available in adequate amounts.

ATLAS: Architecture and Tools for Linguistic Analysis Systems

Background

Annotated corpora have been a essential component of research and development in language-related technologies for some years. As such corpora have proliferated, and have found uses across a rapidly expanding set of languages, disciplines and technologies, the lack of agreed standards has become a critical problem. Of course, standardization of content annotation is necessarily an open-ended task, and is always subject to revision as the underlying domains and tasks change. Yet the standardization of the annotation structures themselves is a goal that could be substantially achieved in a three to five year timeframe. This latter issue is currently the primary roadblock for the creation of general-purpose tools and formats. A widely adopted annotation standard would be an important milestone for research infrastructure in language-related technologies.

The *Linguistic Annotation Page* [`www.ldc.upenn.edu/annotation`] has made researchers aware of the wealth of ongoing activity, of diverse approaches to similar problems and, conversely, of similar approaches to diverse problems. A special issue of the journal *Speech Communication* on "Speech Annotation and Corpus Tools", which is being edited by one of the PIs (Bird), aims to bring together a representative sample of current work and to stimulate dialogue leading to greater consensus about the annotation process.

During the past 12 months, two of the PIs (Bird and Liberman) have demonstrated the existence of a common conceptual core to a wide variety of types of linguistic transcription and annotation, as documented in two research papers: Bird and Liberman (1999a, Bird and Liberman (1999b). This work provides an elegant and expressive algebraic formalism to serve as a universal foundation for linguistic annotation, and work on prototype implementations is already underway. The formalism comprehends all known annotation formats, opening the door to interoperable tools and intertranslatable formats. It will soon be straightforward to apply an existing software tool to a corpus format for which it was not originally designed, permitting users to use their preferred tool to manipulate arbitrary corpora. Likewise, someone will be able to prepare a new corpus in the most convenient format, knowing that it will be straightforward to link it to existing tools as the need arises. Furthermore, it is now possible to create platform-general tools for creating, combining, searching and transforming annotated corpora.

The ATLAS project will provide a general-purpose infrastructure to support the development of annotated corpora and associated analysis tools. The project has the following goals:

1. to implement an annotation graph API as a function library;
2. to develop specialized application programs permitting speech annotations to be created and manipulated;
3. to construct an XML surface representations of annotation structures, to serve as a universal interchange format;
4. to provide import/export functions to connect the annotation structures with existing data formats, and with the XML interchange format;
5. to integrate the application programs into a new version of MITRE's Alembic Workbench;
6. to design query languages and indexing mechanisms to allow annotation structures to be efficiently retrieved and transformed.

The three-level architecture

In the early days of database systems, data manipulation required explicit reference to physical storage in files, and application software had to be custom-built. In the late 1960s, with the development of the so-called "three-level architecture", database functionalities were divided into three levels: physical, logical and external. The Bird/Liberman model applies the same development to databases of annotated speech. Figure 1 depicts the speech annotation version of the three-level architecture.

This model permits users to create and manipulate annotation data in the way that conforms most closely to their own conception of the structure of the underlying data, to the contingencies of the task at hand, and to individual preference. Furthermore, it is possible to change an implementation at the physical level while leaving the higher levels intact – the *data independence principle*. By adopting this model, the volatile nature of formats and the open-ended issues associated with user interfaces no longer present barriers on the road towards standardization. In fact, a large number of tools will be able to comprehend a large number of formats, so tools can interoperate and formats are translatable. Therefore communities that are wedded to particular formats or tools are not left out in the cold.

The annotation model

The formal model is simply stated: annotations are expressed as labeled acyclic digraphs with time references on some of the nodes. Each arc label specifies a property of the extent of signal denoted by the arc. Various constructions on these graphs, such as namespaces defined over the nodes and equivalence classes defined over the arcs, make it possible to express the full range of extant speech annotation formats, and to encompass the full range of conceivable associations between

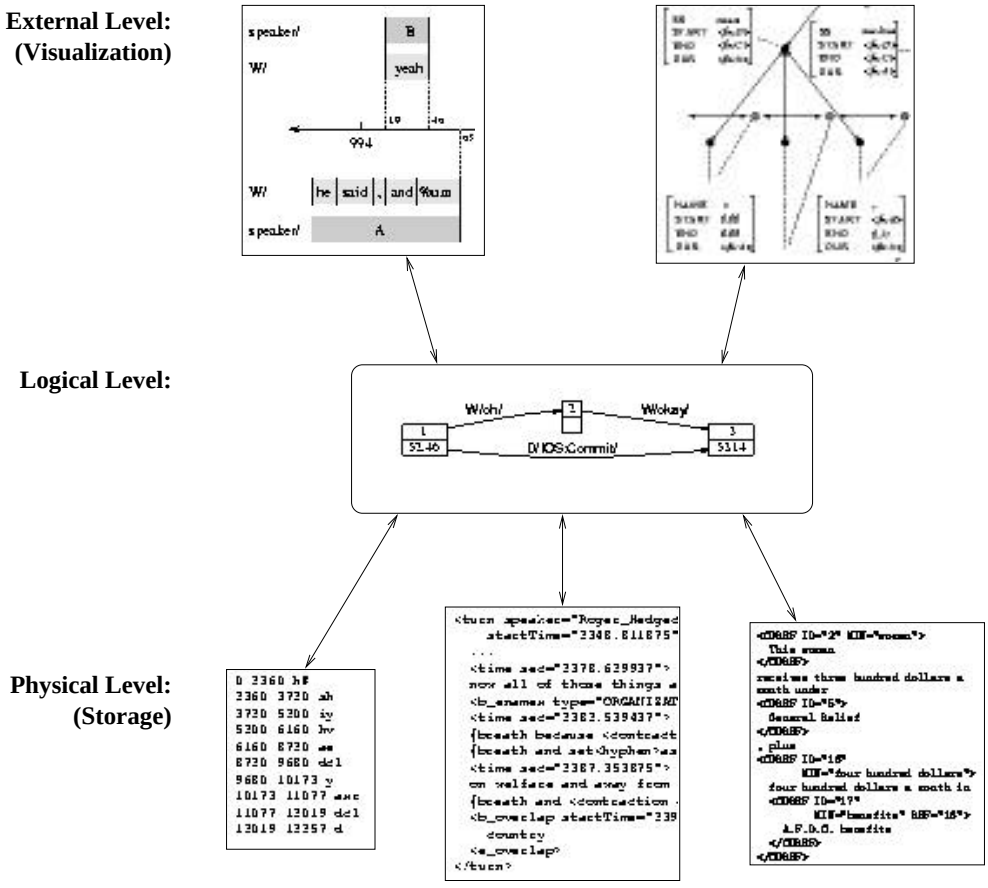

Figure 1: Three-Level Architecture for Speech Annotation

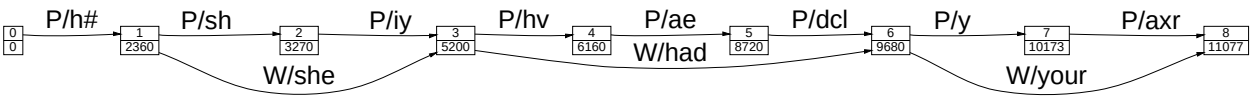

Figure 2: Graph Structure for TIMIT Example

---

annotations and signal files. In this section we present an exposition of the annotation graph model with two examples.

TIMIT was the first annotated speech database to be published, and it has been widely used and also republished in several forms. Here, we just give one example taken from the TIMIT database.

```
train/dr1/fjsp0/sa1.wrd:
2360 5200 she
5200 9680 had
9680 11077 your
11077 16626 dark
16626 22179 suit
22179 24400 in
24400 30161 greasy
30161 36150 wash
36720 41839 water
41839 44680 all
44680 49066 year
```

```
train/dr1/fjsp0/sa1.phn:
0 2360 h#
2360 3720 sh
3720 5200 iy
5200 6160 hv
6160 8720 ae
8720 9680 dcl
9680 10173 y
10173 11077 axr
11077 12019 dcl
12019 12257 d
...
```

The file on the left combines an ordinary string of orthographic words with information about the starting and ending time of each word, measured in audio samples at a sampling rate of 16 kHz. The file on the right contains a corresponding broad phonetic transcription. We can interpret each line: <time1> <time2> <label> as an edge in a directed acyclic graph, where the two times are attributes of nodes and the label is a property of an edge connecting those nodes. The resulting annotation graph for the above fragment is shown in Figure 2.

NIST has recently developed a set of annotation conventions called 'Universal Transcription Format' (UTF). A key design goal for UTF was to provide an SGML-based format that would cover both the LDC broadcast transcriptions and also various LDC-published conversational transcriptions, while also providing for plausible extensions to other sorts of material. A notable aspect of UTF is its treatment of overlapping speaker turns. In the following fragment (from the Hub-4 1997 evaluation set), overlapping stretches of speech are marked with the <b_overlap> (begin overlap) and <e_overlap> (end overlap) tags.

```
<turn speaker="Roger_Hedgecock" spkrtype="male" dialect="native"
    startTime="2348.811875" endTime="2391.606000" mode="spontaneous" fidelity="high">
  ...
  <time sec="2387.353875">
  on welfare and away from real ownership
  {breath and <contraction e_form="[that=>that]['s=>is]">that's a real problem in this
  <b_overlap startTime="2391.115375" endTime="2391.606000">
    country
  <e_overlap>
</turn>
<turn speaker="Gloria_Allred" spkrtype="female" dialect="native"
    startTime="2391.299625" endTime="2439.820312" mode="spontaneous" fidelity="high">
  <b_overlap startTime="2391.299625" endTime="2391.606000">
    well i
```

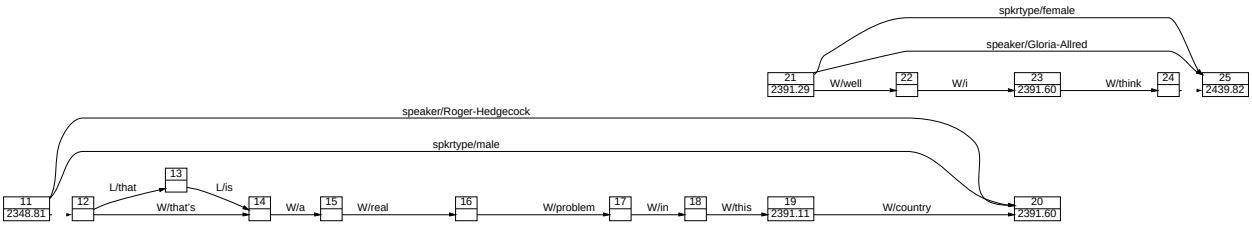

Figure 3: Graph Structure for NIST UTF Example

```
  <e_overlap>
  think the real problem is that %uh these kinds of republican attacks
  <time sec="2395.462500">
  ...
</turn>
```

Observe that there are two speaker turns, where the first speaker's utterance of 'country' overlaps the second speaker's utterance of 'well I'. The structure of overlapping turns can be represented using annotation graphs as shown in Figure 3. Each speaker turn is a separate connected subgraph, disconnected from other speaker turns. This situation neatly reflects the fact that the time courses of utterances by various speakers in conversation are logically asynchronous.

Multiple annotations

Linguistic analysis is always multivocal, in two senses. First, there are many types of entities and relations, on many scales, from acoustic features spanning a hundredth of a second to narrative structures spanning tens of minutes. Second, there are many alternative representations or construals of a given kind of linguistic information. The AG representation offers a way to deal productively with both kinds of multivocality. It provides a framework for relating different categories of linguistic analysis, and at the same time to compare different approaches to a given type of analysis.

As an example, Figure 4 gives a visual representation of an annotation graph which incorporates eight different sorts of annotation of a phrase from the Boston University Radio Corpus. The original BU annotations include four types of information: orthographic transcripts, broad phonetic transcripts, and two kinds of prosodic annotation (tones and break indices), all time-aligned to the digital audio files. We have added four additional annotations: coreference annotation and named entity annotation in the style of MUC-7, syntactic structures in the style of the Penn TreeBank, and an alternative annotation for the F_0 aspects of prosody, known as *Tilt*.

As usual, the various annotations come in a bewildering variety of file formats. These are not entirely trivial to put into registration. For example, the TreeBank terminal string contains both more (e.g. traces) and fewer (e.g. breaths) tokens than the orthographic transcription does, and the connection between the word string and the break indices are mediated only by identity of the associated the floating-point time values. Since these time values are expressed as ASCII strings, it is easy to lose the identity relationship simply by reading in and writing out the values to programs that may make different choices of internal variable type (e.g. float vs. double), or number of decimal digits to print out, etc.

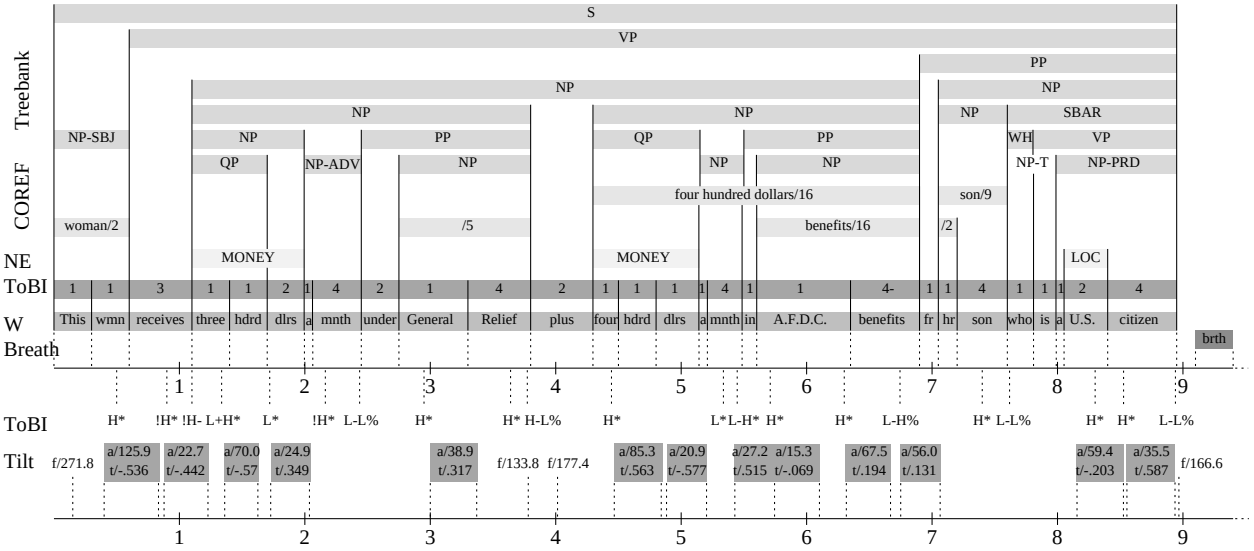

Figure 4: Visualization for BU Example

Problems of this type are normal whenever multiple annotations need to be compared. Solving them is not rocket science, but does take careful work. When annotations with separate histories involve mutually inconsistent corrections, silent omissions of problematic material, or other typical developments, the problems are multiplied. However, once this common framework is established, via translation of all eight "strands" into annotation graph terms, we have the basis for posing queries that cut across the different types of annotation.

Validation

An annotation may need to be submitted to a variety of validation checks, for basic syntax, content and larger-scale structure. First, we need to be able to tokenize and parse an annotation, without having to write new tokenizers and parsers for each new task. We also need to undertake some superficial syntax checking, to make sure that brackets and quotes balance, and so on. Second, various content checks need to be performed. For instance, are purported phonetic segment labels actually members of a designated class of phonetic symbols or strings? Are elements marked as 'non-lexemic vocalizations' drawn from the officially approved list? Do regular words appear in the spell-check dictionary? Do capital letters occur in legal positions? Finally, we need to check for correctness of hierarchies of arcs. Are phonetic segments all inside words, which are all inside phrases, which are all inside conversational turns, which are all inside conversations? We propose to meet these validation needs by developing conversion and creation tools that read and write well-formed graphs.

Standards and tools for dictionaries

The CallHome project involved community sharing of some very simple dictionaries in six languages. The dictionaries included orthographic word forms, pronunciations, unigram frequencies

in various corpora, sometimes morphosyntactic features, and sometimes glosses or semantic class information. Their structure was basically a simple, single relational table. Despite the simplicity of the situation, and our best efforts to produce resources with a clear formal structure, both the producers and the users of the data encountered many technical and conceptual problems.

The TIDES program will be confronted by a far richer ontology of lexical information, including cross-lingual glossaries for CLIR; IR ”thesauruses”; tables for driving stochastic MT systems; database forms of paper bilingual dictionaries; tables for finite-state morphological transducers; connections between lexical distinctions and corpus annotation; and so on. Since there is a premium on handling some 30 languages, and on rapid development of support for new languages, it is vital to avoid the situation where researchers spend weeks trying to match lexicons with tools, or waste weeks because of unforeseen mismatches.

The recent TDT-3 dry run gave a small foretaste of the many kinds of ”impedence mismatches” that are in store for us; different ideas about tokenization (i.e. what the database entries are), about character sets and normalization techniques, about vocabulary choice, about transliteration, and so forth, are going to be rampant – never mind all the simple bugs in assimilating the data in the first place.

It’s not possible to foresee or prevent every possible problem; and we have to be careful about foreclosing research directions by making too-restrictive choices. However, we believe that it is possible to devise an extensible standard for most of the basic kinds of lexical resources that will be involved in TIDES, and to provide some basic tools for creation and access. Although this process is less far advanced than the planning for the transcription and annotation aspects of ATLAS, it is essential for to cover this area as well, since LDC will be creating and/or distributing many multilingual lexicons in association with TIDES and similar projects.

Collaboration with NIST and MITRE

In recent meetings at NIST and MITRE, we have refined the annotation model (e.g. to include full support for inter-arc pointers) and we have identified a preliminary division of labor, based on the 3-level architecture: with MITRE generalizing its Alembic tool as the primary implementation of the “External” level (visualization and user interaction), LDC mainly responsible for the “Logical” level, and NIST mainly responsible for the “Physical” level (file formats and other storage issues).

More specifically:

1. LDC with NIST and MITRE input will define the formalism, create I/O routi nes to a wide variety of existing data formats, and define an XML interchange forma t;

2. LDC will implement the formalism, providing an API for annotation graphs and a library of elementary annotation-graph manipulations, and developing indexing mechanisms and a query language;

3. MITRE with LDC and NIST input will build a new version of the Alembic workbench on top of the API, encompassing both text and speech annotation, and LDC will roadtest the workbench on a variety of annotation tasks;

In the case of standards and tools for dictionaries, our proposal is to begin by doing a rapid survey of existing practices, similar to the survey of annotation practices documented in `http://www.ldc.upenn.edu/annotation`. We will then create a formalism for the logical level of dictionary database representation, and follow that by cooperating with NIST, MITRE and other interested parties to design and implement the external and physical layers.

C. Deliverables

VOA Corpus

For the proposed VOA corpus, the entire 1.6 TB archive of 14,144 hours of digital audio recordings in 45 languages, and all associated scripts, will be completed during the first year. It will be available for delivery to researchers, in whole or (more plausibly) in parts, from the beginning of the second year onward.

Starting in the second year, we will prepare speech-to-text aligners, based on standard speech recognition technology, for six languages per year. For each language, this requires as components a pronouncing dictionary with 90-95% coverage of lexical tokens in scripts, and a set of acoustic models trained from a few hours of scripts hand-aligned at the phrase level with VOA audio. These components, as well as the resulting aligned texts, will be published each year.

These activities will continue through the option years, if the options are exercised.

Parallel and Congruent Corpora

We expect to be able to deliver one or two large parallel text corpora each year (where "large" means tens of millions of words), and four or five small parallel text corpora (where "small" means hundreds of thousands to millions of words).

Precise plans for each year will be based on the language interests of the TIDES projects, the existence of parallel texts, and the attitude of particular IPR holders.

We also expect to be able to deliver large amounts of "congruent" text of the kind that we have been collecting for TDT, that is, journalistic text streams for different languages during the same time periods and covering an overlapping set of topics.

These activities will continue through the option years, if the options are exercised.

Bilingual Dictionary Databases

We will deliver six digitized bilingual dictionaries each year, subject to selection according to TIDES interests and successful negotiation of IPR releases from the copyright holders for research distribution.

Each dictionary will be delivered in the form of a set of fielded records, one per lemma, with the structure encoded as XML tagging.

These activities will continue through the option years, if the options are exercised.

ATLAS

In the first year, in consultation with NIST and MITRE, we will deliver a complete specification of the ATLAS transcription/annotation formalism, the XML interchange format, I/O routines for

all relevant existing data formats, and a library of annotation-graph functions. We will also deliver a survey of current lexical database practices, and a preliminary recommentation for a logical structure for lexical data in ATLAS. We will test various prototype GUIs for multilingual transcription/annotation creation.

In the second year, after consultation with the research community, we will deliver an expanded inventory of annotation-graph manipulations, and an improved set of libraries. We will provide indexing mechanisms and a query language. With NIST, MITRE and others, we will deliver stable systems for creation, access, search and update of multilingual transcriptions and annotations. We will deliver a final version of the lexical formalism, and a preliminary API and library for lexicon creation and manipulation.

In the third year, we will deliver a stable version of the multilingual lexical tools. We will provide improved integration of the transcription/annotation and lexical tools, and will provide (in experimental form) a general software framework for integration of machine assistance for human annotators and lexicographers.

Throughout all three years, all models, formats, libraries and tools will be documented. Scientific results will be reported to the wider community via conference presentations and journal publications.

Proprietary Claims

There are no proprietary claims to the results of the proposed work, except for necessary IPR licensing for copyrighted material.

In particular, all ATLAS specifications will be freely usable by all, and software developed for ATLAS will be distributed under an “open source” model.

D. Statement of Work

Note that because of the nature of this proposal, detailed discussions of the work to be performed can be found in the technical rationale section. We summarize the work to be performed here in quantitative terms, with reference to section B for additional specific information, such as the tables of VOA languages.

VOA Corpus

We will produce a 45-language corpus of speech with associated scripts, sampled from VOA broadcasts accessed via digital satellite downlink during the first year of the project. Most languages will be sampled at an hour per day over a year, with some sampled more sparsely. Over up to five years, we will produce software speech-to-text aligners in up to 24 languages for this corpus, able to produce accurate time alignments at the word level. This software will be based on a basic pronouncing dictionary and a set of acoustic models for each language; given a language model, these would also suffice for a minimal ASR system.

Parallel Text

Over up to five years, we will produce up to 10 large parallel text corpora, and up to 20 small ones, obtained from web sites and from international governmental and non-governmental agencies, subject to TIDES interest, corpus availability and success of IPR negotiations. We will align these corpora at the phrase level.

Dictionary Digitization

Over up to five years, we will produce up to 30 dictionary databases corresponding to published bilingual dictionaries, based on scanning, OCR, entry parsing, and post-editing by native speakers. IPR permissions will be negotiated to permit distribution for research use, both within the TIDES project and beyond it.

ATLAS

We will work with NIST, MITRE and others to produce the standards and tools needed for creation, validation and maintenance of multilingual resources, and for access to those resources by researchers and their systems.

1 E: Schedule of Milestones

F. Technology Transfer

For the VOA corpus, the parallel text corpora, and the digitized dictionaries, we will accomplish technology transfer by distributing the resulting corpora to TIDES researchers, and publishing them via the LDC for the research community at large. They will thus join the more than 150 corpora and lexicons that the LDC has published since it was founded (with seed money from DARPA) in 1992. More than 250 member organizations and 520 non-member organizations have received LDC corpora, including nearly every R&D organization in the world involved in language technologies. Since there have been more than 10,000 transactions in which one lab received one corpus or lexicon, each of the more than 770 participating labs has gotten an average of about 13 items.

In the case of ATLAS standards and tools, the same route of technology transfer also applies, since new LDC resources, and republications of old ones, will use ATLAS standards. This will make resource creation and maintenance more efficient, and it will also make resources easier to assimilate. In addition to this kind of technology transfer, ATLAS standards and tools will be freely available to the research community for their own use in creating additional resources, with the tools being distributed under an “open source” model.

G. Comparison with other approaches

VOA Collection

No other non-governmental organization in the U.S., as far as we know, is in a position to collect a multilingual speech/text database on the scale of the proposed VOA collection, especially with respect to the number and diversity of languages involved. This is a unique opportunity, and one that may expire after 2001, since the law authorizing VOA cooperation lapses then. Renewal would be dependent on new congressional action. While there is every reason to hope for this, it cannot be viewed as automatic – the current law was the third attempt, with the first attempt failing because it was part of a larger package that was blocked by political disagreements over a foreign policy issue, and the second attempt failing because it was inadvertently omitted during an all-night session to resolve House and Senate versions of a bill, even though it had originally been present in both versions.

Parallel Text

In the case of parallel text acquisition, the primary alternative is the STRAND proposal by Philip Resnik, in which what is distributed is not parallel text itself, but rather a set of URLs at which parallel text can be found, and "site sucker" software that will allow other researchers to download the data for themselves.

While the STRAND model is definitely worth pursuing, we feel that it has certain disadvantages. First, the web is ephemeral. Some of the most interesting sites (for instance many journalistic sites) have revolving content, with a limited on-line archive or no archive at all. Other sites may change their content erratically, or even disappear. In such cases, STRAND will not work, and our method is the only way to guarantee even medium-term access for the research community as a whole.

Second, it is not clear that STRAND-style distributed downloading really solves the legal problems. Many sites (again, especially journalistic ones) assert terms of use that forbid users to "modify, create derivative works, or in any way exploit any of the content," Transfer of access to others (even within the same lab) is also typically forbidden. In general, providers want to limit usage to individuals viewing pages in their individual web browsers, and to prohibit any uses that go beyond this, even where traditional notions of "fair use" would allow them. While the legal force of such unilateral assertion of terms is unclear, a company eager to avoid legal entanglements will be wary.

Therefore, we feel that the STRAND model is by no means a panacea, and that our approach is often preferable, especially for large shared parallel corpora where reliable access over time is an important goal.

Digitized dictionaries

We do not know anyone else who has attempted digitization of paper dictionaries on the scale in question, or who is proposing to do it now. Like the VOA corpus, this seems to be a unique opportunity.

ATLAS

ATLAS will adopt the three-level model for linguistic annotation, which is an important innovation for language language technology arising from the Bird/Liberman annotation formalism. The three-level model was introduced to the database research community some 30 years ago, and

has become a foundational conception in database practice. Without this conception, superficial differences in file format or visualization technique may loom very large, making equivalent content look very different; at the same time, quite different logical structures may look misleadingly similar. Previous work has primarily focused on issues like whether SGML is used to express document structure, which we believe to be an issue about the physical or storage level, rather than the logical level.

The "annotation graph" conception provides an excellent basis for the logical level of linguistic texts, transcriptions and annotations. In addition to expressing the content of current annotation practice relatively directly, it offers good properties for indexing and search, and will permit important advances in incremental annotation, and in distributed creation and storage of annotations.

A number of other general-purpose annotation models have been developed elsewhere, including: Festival's Heterogeneous Relation Graphs, MATE's XML format, the Macquarie Emu system (see [`www.ldc.upenn.edu/annotation`] for pointers). Since these models ignore the "data independence principle" observed by ATLAS , they are exposed to the same problems encountered by the ad hoc database systems of the 1960s. Moreover, they lack the clean algebraic semantics of the ATLAS model, and much less is known about efficient methods for indexing, query and data transformation.

H. Qualifications of Key Personnel

Mark Liberman is a professor at the University of Pennsylvania (1990-), with appointments in Linguistics and in Computer and Information Science. He was a member of technical staff and department head of the Linguistics Research Department at AT&T Bell Laboratories (1975-1990). He is currently the director of the Linguistic Data Consortium (1992-). He is on the editorial boards of Speech Communication, Computer Speech and Language, and The International Journal of Corpus Linguistics. His research interests are in phonetics, phonology, speech technology, and computational linguistics.

8% compensated effort. (Current: co-PI 2 programs, totaling 4% compensated effort during proposal period. Proposed: co-PI 5 programs, totaling 12% compensated effort.)

Steven Bird is associate director of the Linguistic Data Consortium, and holds the position of adjunct associate professor in the department of Linguistics and the department of Computer and Information Science at Penn. He has been a guest editor of *Computational Linguistics* for a special issue on Computational Phonology, and is guest editor of *Speech Communication* for an upcoming special issue on Speech Annotation and Corpus Tools. He is Founding President of the ACL SIG in Computation Phonology, an editorial board member for *Computational Linguistics*, and International Linguistics Advisor for the Summer Institute of Linguistics. He has done extensive field work on Grassfields Bantu languages of Cameroon, and his other recent projects include HyperLex, a powerful lexical search engine that has been applied to lexical databases in English, Dschang (Cameroon), Yoruba (Nigeria), Mawukakan (Cote d'Ivoire), Nahuatl (Mexico), and Tamil (India).

Compensated effort 35%. (Current: no compensated effort on sponsored projects. Proposed: PI 1 program, co-PI 4 programs, totaling 58% compensated effort.)

Christopher Cieri is the executive director of the Linguistic Data Consortium. His previous positions were as head of computer services for the University of Pennsylvania Law School, and as a researcher in the Language Analysis Center. He has been responsible for overall plannign and project management for several large LDC corpus-development efforts, including TDT-2 and TDT-3, CallFriend Russian and Korean, Switchboard Cellular, a newswire collection in 15 languages, and several dictionary projects. His other research interests include sociolinguistics and the dialects of Italy.

Compensated effort 30%. (Current: no compensated effort during period of proposed grant. Proposed: PI 1 program, compensated effort 30% in year 1, 60

David Graff was the LDC's second employee, and has been in charge of technical support since the beginning. He has played a central role in the production, preparation and maintenance of virtually every data collection created or published by the LDC. He designs new corpus structures and specifications in consultation with sponsors and with NIST. He has designed and implemented many of the custom user interfaces for transcription and annotation, and has supervised development of the others. He also is responsible for the planning and layout of computer and media resources to accomodate the capture and and processing of linguistic data in ever-larger quantities from ever-more-varied sources.

Compensated effort: 20% year 1, 10% years 2 and 3. (Current: no compensated effort during

proposal period. Pending: 40% in year 1, 10% in year 2.)

Relevant prior projects

The Linguistic Data Consortium is an open consortium of companies, universities, and government research labs, that creates, collects and distributes speech and text databases, lexicons, and other resources, in support of research and development in human language technologies. The University of Pennsylvania is its host institution. The LDC was founded in 1992, with seed money provided by DARPA. Since then, it has published more than 150 corpora and distributed them to more than 250 member organizations and more than 520 non-members organizations. The total number of tranctions in which some organization received some corpus exceeds 10,000.

As provided in the terms of its initial DARPA grant, the LDC has become self-supporting, in the sense that membership fees and data sales provide all the funding for core consortium activities such as corpus publication, membership relations, and IPR negotiations. An DARPA/NSF cooperative agreement, IRI-9528587, which began in September 1995 and will end in August of 1999, has had two primary goals: first, the establishment of an on-line repository of all LDC data, along with implementation of a consistent search and retrieval system across all of the various databases; and second, the provision of necessary data for DARPA-sponsored research efforts in the language technology area.

Both of these goals have been achieved. LDC-Online has succeeded in putting all LDC speech and text corpora on the internet, where IPR agreements permit, and in providing a consistent web-based search and retrieval interface for this very heterogeneous set of databases.

Using funding from the current DARPA/NSF cooperative agreement, the LDC has managed the collection, transcription and distribution of hundreds of hours of broadcast speech data in three languages for the DARPA “Hub-4” and TDT projects; has combined newswire text and transcriptions with ASR output and closed-caption capture as the textual basis of the TDT corpus; has added topic definitions and topic relevance judgments for TDT; has collected and published hundred of millions of words of broadcast transcriptions for language-modeling purposes; and has worked with NIST and others to create and distribute ”evaluation kits” for speech recognition (in several domains), speaker identification, language identification, message understanding, and on-line handwriting recognition. A total of 97 databases have been published during the period of this agreement (fall 1995–present).

I. Description of Facilities

The Linguistic Data Consortium maintains a large computing facility that will, with the additional specialized equipment requested in this proposal, be sufficient to support the project as outlined. This facility includes 2 Sun Ultra Enterprise 4000 multiprocessor servers with a total of 1.26 Terabytes (1260GB) of disk, 2.25G of RAM, and ten UltraSPARC processors, with a 3.5 Terabyte DLT tape robot for backups and disaster recovery. An additional Sun E450 server with two 450MHz processors and 1G of memory, and an additional 3.5 TB tape robot for nearline-storage, are due to be delivered soon, funded by a Sun AEG grant for support of TDT and other DARPA-related projects. These servers are connected via high-speed fiber optic networking to the Computer and Information Science Department (CIS) and to the LDC's workstations.

In addition to the network servers, LDC maintains more than 60 Sun Sparcstations, more than 20 Windows NT workstations, more than 5 Linux workstations, and dedicated hardware for the collection of language data from newswire services and the broadcast airwaves. To capture Voice of America broadcasts, LDC maintains a satellite downlink station built by Zycom Electronics, the group that configured VOA's own satellite network. This system consists of a C-band satellite dish along with:

1. Research Concepts RC2000 Dual Axis Antenna Controller to adjust and tune the dish
2. Radyne Demodulator 2401 RX
3. Ascon Timeplex Mini-Link/2+ to split digital satellite signal into individual encoded channels
4. Four Comstream ABR200 Audio Processors to decode the individual channels
5. Four Townsend DATLink+ components to stream decoded digital audio onto computer a SCSI bus
6. SunSparcstation to manage the collection process and control the other components

LDC's offices, recently expanded to 9400 square feet, are adequate to house the 20 full-time staff and 25-50 part-time staff who contribute to LDC projects including the proposed.

J: Cost by Task (five pages?)

III: Bibliography

Bird, Steven (1997). A lexical database tool for quantitative phonological research. *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*, pp 33-39, Madrid, Spain.

Bird, Steven & Mark Liberman (1998). Towards a Formal Framework for Linguistic Annotations. *Proceedings of the Fifth International Conference on Spoken Language Processing*, Sydney, Australia.

Bird, Steven & Mark Liberman (1999a). *A Formal Framework for Linguistic Annotation.* Technical Report MS-CIS-99-01, Computer and Information Science, University of Pennsylvania.

Bird, Steven & Mark Liberman (1999b). Annotation graphs as a framework for multidimensional linguistic data analysi s. Towards Standards and Tools for Discourse Tagging; Proceedings of the Workshop. pp 1–10.

Bird, Steven (1999). Multidimensional exploration of online linguistic field data. *Proceedings of the 29th Annual Meeting of the Northeast Linguistics Soci ety*, pp 33–47.

Cieri, Christopher, David Graff, Mark Liberman, Nii Martey and Stephanie Strassel (1998) TDT-2 Text and Speech Corpus, *DARPA Broadcast News Workshop, Washington, DC.*.

Graff, David and Christopher Cieri (1998) Update on Lexical Resources and Projects at the Linguistic Data Consortium, *Ninth Hub-5 Conversational Speech Recognition (LVCSR) Workshop,* Linthicum Heights, Maryland.

Liberman, Mark and Christopher Cieri (1998) The Creation, Distribution and Use of Linguistic Data *Proceedings of the First International Conference on Language Resources and Evaluation,* Granada, Spain.