Project Description

Networking Linguistic Information:
Creating and Disseminating
Knowledge about the World’s Languages

This is a proposal submitted to the NSF KDI initiative by Akin Akinlabi (Rutgers), Steven Bird (Penn), Peter Buneman (Penn), Will Leben (Stanford), and Mark Liberman (Penn)

1 Overview

Advances in mass storage technology now make it possible to collect almost arbitrary amounts of speech, text and other linguistic data in digital form. Advances in computer hardware and software make it possible to annotate this data efficiently with representations of linguistic structure and function, and to search the results flexibly and easily. Advances in networking mean that almost arbitrary amounts of such data can be published at little or no marginal cost.

These developments have revolutionized research practices in speech and language technology over the past decade, and have strongly influenced scientific research on language. More profound changes will follow, as research practices catch up with the new opportunities. A focused interdisciplinary effort at this crucial time will accelerate the process, and help to avoid unnecessary inconsistencies and duplication of effort.

This project aims to foster new modes of fundamental research in linguistics and language technology, by providing standards, tools and practices for creating, searching and publishing primary linguistic materials via networked computers. It brings together linguists, computational linguists and computer scientists. The goal is to facilitate the work of individual linguists, to permit geographically separated groups to collaborate easily, to allow primary data to be shared with the wider research community, to foster research on typology by establishing standards that facilitate comparison across languages, and – underlying all of the above – to explore and exploit new methods for representing and analyzing online linguistic data.

For this effort to succeed, it must be based on an active and extended collaboration between working linguists and specialists in various areas of computer science. In order to get critical mass for collaborative work, we’ve chosen a focus on the tonal languages of sub-Saharan Africa. This choice is based on the theoretical importance of the topic, the large amount of on-going primary description, and the cohesive and cooperative nature of the research community. In order to ensure generality of results, we’ve also planned a wide range of applications to other areas and subdisciplines. Benefits will accrue to all researchers and scholars who work with linguistic data, from speech recognition engineers to anthropologists and historians. The tools and the resulting data will be useful in teaching as well as in research.

All of the primary data created by this project will be published on the web site of the Linguistic Data Consortium (LDC), for general public access. Materials created by other collaborators will also be published there, subject to the permission of the authors. All tools and documentation produced by the project will be freely available to others.

1.1 Uses and Users of Linguistic Databases

There are many sorts of linguistic databases, collected for many different reasons, and used for an even wider variety of purposes. The standards and tools that we are discussing are applicable across the spectrum of database types and database uses.

The community of users in our focus area – linguists working on African tone – will exercise the various types and applications of linguistic databases especially widely. They also have especially strong motivation to make use of new technology of this type. Before explaining this in further detail, it will be helpful to survey some more general background. (In addition to the material presented here, readers may want to refer to the Language Resources Primer [1] prepared for the 1997 Language Resources Workshop, or that workshop’s final report [2].)

Linguistic databases span a multidimensional space of cases, which we can divide up in several ways: the scope
and design of the data collection; the goals of the creators; the nature of the material included; the goals and methods
of the users (which are often not anticipated by the creators). Three examples follow.

In one type of linguistic database, the design unfolds interactively in the course of the creator’s explorations. This
is the pattern typical of traditional “field linguistics,” in which material from elicitation sessions is analyzed and re-
analyzed intensively repeatedly as it is gathered, with tomorrow’s elicitation often based on questions that arise in
analyzing today’s. The resulting field notes are then used during subsequent years of research, and may serve as an
archival resource indefinitely – the field notes of linguists and anthropologists working in the early years of the 20th
century remain an important source of information today. Computerization is an obvious boon to work of this type, as
exemplified by the popular program Shoebox [3], now about 15 years old, which replaces the field linguist’s traditional
shoebox full of file cards.

Another pattern is represented by experimental approaches in which a specified body of carefully-designed material
is collected from a range of subjects, and then subsequently analyzed to evaluate a particular hypothesis, or to build
or test a particular technology. Today, such databases are of course collected and analyzed in digital form. Among
scientists (such as phoneticians or psychologists), they are rarely published and therefore rarely preserved. Among
engineers, it has become common for such databases to be shared and re-used at least within a laboratory or company,
and often to be published more widely. Linguistic databases of this type are the basis of the “common task” method of
research management, which over the past decade has become the norm in government-funded research programs in
speech- and language-related technology.

Finally, there are efforts to gather a “reference corpus” for a particular language. A large and well-documented
recent example is the British National Corpus [4] (BNC). The goal in such cases is to produce a set of linguistic
materials that cover the many forms, styles and uses of a language as widely as possible. The core application is
typically lexicographic, that is, the construction of dictionaries based on a careful study of patterns of use. The BNC
was constructed by a large government/industry/university consortium. Its planning and execution took more than five
years, and involved tens if not hundreds of person-years of work. There is a long and distinguished history of other
humanistic reference corpora, such the Thesaurus Linguae Graecae [5].

There are no hard boundaries among these categories. Accumulations of smaller bodies of data may come in
time to constitute a sort of reference corpus, while selections from large databases may form the basis for a particular
experiment. Further instructive examples follow.

A linguist’s field notes may include extensive examples of many genres (proverbs, conversations, narratives, rituals,
and so forth), and may come to constitute a reference corpus of modest but useful size. There are many extinct
languages for which such material is all the data we will ever have, and many more endangered languages for which
such documentation is urgently needed. Sociolinguists typically base their work on analysis of a set of recorded
interviews, which may over time grow to create another sort of “reference corpus.” In some labs, the residue of decades
of work may comprise literally thousands of hours of recordings, many of which have been transcribed and annotated
to one extent or another. The CHILDES corpus [6], comprising transcriptions of parent-child interactions in many
languages, contributed by many individual researchers, has come to constitute a widely-used “reference corpus” for
language acquisition research. Speech technologists aim to produce training and testing material of broad applicability,
and in effect wind up creating another sort of “reference corpus.” To date, linguistic technology R&D has been the
primary source of published linguistic databases of all sorts (see [7] and [8] for an extensive catalogue).

As large, varied linguistic databases are published, phoneticians or psychologists are increasingly likely to base
experimental investigations on balanced, focused subsets extracted from databases produced for entirely different rea-
sons. Their motivations include the desire to save time and effort, the desire to work on material available to others
for replication, and sometimes a desire to study more ecologically valid forms of linguistic behavior. The process of
choosing a subset for such a study, and making the measurements involved, is usually in itself a non-trivial addition to
the database. This recycling of linguistic databases for new purposes is a normal and expected consequence of pub-
lication. For instance, the Switchboard database [9], which was originally collected in 1991 for speaker identification
research, has since been used as the basis for published studies in speech recognition, word pronunciation, disfluency,
syntax, intonation and discourse structure.

At present, only a tiny fraction of the linguistic databases that are collected are published in any meaningful sense.
This is mostly because publication of such material was both time-consuming and expensive, and because use of such
material by other researchers was also both expensive and technically difficult. In principle, general improvements in
hardware, software and networking have changed this. In principle, linguistic databases can now be created, published,
stored and used without inordinate effort or large expense.

In practice, the implications of these cost-performance changes are only beginning to be felt. The main problem is that adequate tools for creation, publication and use of linguistic data are not widely available. In most cases, each project must create its own set of tools, which hinders publication by researchers who lack the expertise, time or resources to make their data accessible to others. Furthermore, we do not have adequate, generally accepted standards for expressing the structure and content of linguistic databases. Without such standards, general-purpose tools are impossible – though at the same time, without available tools, adequate standards are unlikely to be developed, used and accepted. Just as importantly, there must be a critical mass of users and published material to motivate maintenance of data and access tools over time.

1.2 Goals of our Project

Our project aims to provide the means for efficient entry, publication, search and retrieval of primary linguistic data. The techniques should be accessible and useful to any working linguist, and should make primary description easier, as well as improving the quality and accessibility of the results. Tools for constructing such linguistic databases should check consistency, and help authors maintain consistency across updates. The resulting databases (including those in the process of construction) should support convenient browsing and specific search for examples as well as large-scale statistical analysis. The approach should support cooperative work, both on data creation and on access for research. Both small and large scale collaborations should be supported, including collaborations that are geographically distributed. The resulting databases should be directly publishable without any requirement for further human labor, and should not become inaccessible in the future due to technological change. Such databases should support durable citations that are maintained across corrections, additions and combinations.

By “primary linguistic data” we mean lexicons, texts, annotated audio recordings, derived information such as pitch tracks and spectrograms, “paradigms” (in the extended sense of rational catalogues of relevant cases), demographic and geographical information, and appropriate links among all of these. Tools will include ways to create, organize, modify, search, explore, model and cite such bodies of data. Machine learning algorithms will be used to minimize or eliminate the programming required to create new tools for (computer-aided) morphological and syntactic analysis.

Descriptive text (including grammars, monographs and papers as well as less formal discussion) is an essential part of the picture, both directly as documentation for the primary data, and indirectly as an on-going scientific discourse referring to it. The empirical grounding of such text may often be seen as an annotated restructing of primary data, such as a table containing pointers to selected entries in a lexicon and selected excerpts from a set of audio recordings, or a graph summarizing a specified set of phonetic measurements. It is possible to treat such restructuring as an active process, so that a reader may (re-)assemble the presentation from the primary materials that it is based on, which may in principle be from diverse sources and stored in diverse locations. In the limit, such a process may turn a scientific paper into an object that automatically replicates the experiment it documents each time it is read. Without going so far, such active citations to published databases are still a boon to subsequent research.

All data representation standards developed by this project will of course be open, and all tools will be made freely available to all researchers, including source code wherever possible.

1.3 Components of a Solution

A key problem is how to take advantage of the rapidly evolving technology of networked computers, without having reference databases and published citations become obsolete every few years.

We will adapt or create a set of evolving, interoperating tools, which will couple to a set of stable but flexible data exchange formats. As long as the initial data abstractions remain valid, the tool set can be completely replaced every few years without making old results inaccessible. Entirely new kinds of data models can also be added, as long as older structures are still supported or can be translated.

It is essential that the initial approach to representing linguistic data be carefully constructed with these requirements in mind. Finding the best solutions raises some open research problems, and so a close collaboration with database theorists is necessary. Co-PI Buneman is a database researcher who has worked on analogous problems in bio-informatics and other areas.
At the same time, solutions must be designed in consultation with working linguists, and must be extensively tested by them. Linguistic phenomena are complex, the representational issues are often subtle, and it is easy to leave out crucial features, or to generate needlessly complex solutions to problems that do not arise in practice. In terms of tool design, small changes can turn out to make large differences in accessibility, flexibility and other aspects of usefulness. Therefore, we need a community of linguistic researchers to test and validate the approach though extensive real-world use. We also want to be sure that a critical mass of users will continue to work with the resulting data, to provide motivation for keeping the tool set technically fresh. Finally, in order to provide a test bed for collaborative work, our user population should be engaged in both small-scale (2-3) and larger-scale (10-15) collaborations.

These requirements are best met by focusing initial effort on one subdisciplinary area. We’ve chosen the tonal phonology of Africa, because of the theoretical importance of the topic, the large amount of on-going primary description, and the cohesive and cooperative nature of the research community. Two of the co-PIs are primarily Africanist phonologists (Akinlabi and Leben), while two others have published in this area while being primarily a computational linguist (Bird) and a phonetician (Liberman). More than a dozen other prominent scholars in this area have agreed to cooperate, and we hope that nearly the whole community will participate to some extent within a few years, if the project is successful.

Since African tone connects to essentially every aspect of grammar and language use, we expect that standards and tools adequate for this task will be generally applicable. As one check of generality, we will ensure that all databases published by the Linguistic Data Consortium can be expressed in our formats and accessed via our tools. We will add specific collaborations involving a sample (opportunistic but as representative as possible) of work in other language areas and other subdisciplines. We have arranged more than ten such collaborations. For reasons of space, we will provide only three examples.

Our tools will be used in the course The Study of the Speech Community taught by William Labov and Gillian Sankoff at Penn, in which techniques and theory derived from sociolinguistic studies are used to define neighborhoods, enter the community, analyze social networks, and obtain tape-recorded data from face-to-face interviews.

We will collaborate with Jonathan Amith of Yale, to verify that our tools can provide a useful web-based interface to his analytical dictionary of Nahuatl.

We will work with the Pacific Languages Electronic Resources Project of Manning and Walsh at the University of Sydney, Australia to ensure compatibility of our tools with their databases and their user community.

1.4 Why do it now?

It would not have made sense to begin this project a few years ago. The essential preconditions for a solution are now available: affordable hardware/software platforms with the capacity to create and search large linguistic databases; practical access to relevant basic data standards such as Unicode, XML and CORBA; programming environments that enable rapid creation and widespread use of interactive tools, such as Tk/tcl, python, Java; some software piece parts and data formats good enough to start experimentation with; years of experience in the practical and legal aspects of large-scale electronic publication of linguistic data (LDC); and last but certainly not least, the development of the internet as a tool for communication, collaboration, tool distribution, and data publication.

For some varied examples of the new possibilities, we draw your attention to Steven Bird’s HyperLex system, to Zhibiao Wu’s LDC-Online, to Harrington and Cassidy’s EMU speech database system, and to Paul Boersma’s Praat tool for phoneticians.

Although the background conditions are favorable, and there are many inspiring examples, the current situation with respect to the structure and use of linguistic databases remains very far from ideal. From the point of view of overall structure, an uncharitable observer might refer to today’s linguistic databases as “a babel of inadequate and inconsistent solutions.” The overall structures of current linguistic databases are only informally specified, are sometimes inconsistent in parts even within a single database, and are different (sometimes for good reason but often not) even across similar databases from related sources. Formats for individual pieces of such databases (audio files, transcriptions, etc.) are more standard but still tend to differ from database to database. Most current linguistic databases are supplied without tools for access or searching, so that users must create their own, but in most cases, each new database requires new programming to make it accessible. Some formats used by linguists are now largely inaccessible (e.g. HyperCard stacks), and others are likely to become so. Existing searching and visualization tools
have many gaps and flaws, and often require expensive proprietary subsystems that limit their usage among (generally impecunious) linguists.

For individual projects, the path of least resistance is to adopt the simplest and most accessible partial solution that is locally adequate. This tends to lead to further fractionation. Precisely because the rapidly changing technology makes large-scale use increasingly attractive, now is the most valuable time to try to create some durable general-purpose abstractions.

1.5 Research on African Tone

The Ethnologue database [15] lists more than 1,400 languages in sub-Saharan Africa. Of these, about 70 now have a million or more primary speakers, and at least 400 have 100,000 or more. Because of this sheer diversity, there is probably more primary linguistic description underway in Africa than anywhere else in the world.

However, this is not the only reason that linguistics “stands in profound debt to work on African languages” [16]. In particular, the study of complex tonal phenomena in African languages gave birth to ‘autosegmental phonology’ (e.g. [17, 18]), a theoretical model incorporating multi-linear structures representing parallel, partially-asynchronous speech gestures, as exemplified in the figure below.

![Autosegmental diagram](attachment:autosegmental-diagram.png)

In this Dschang phrase, the top plot is a time function of fundamental frequency (here calculated from laryngograph measurements). The middle plot shows a typical transcription of surface tone categories aligned with phonetic consonant and vowel segments. The bottom plot shows a representation of the corresponding phonological structure, with a (circled) “floating” high tone, and a final low tone that is “downstepped” by its context.

The type of model that this figure exemplifies has been found to have enormous heuristic and analytic value. It has driven theoretical phonology since the mid 1970s, and is a principal focus of the dozen or so phonology textbooks published in the last decade. To see a concrete example of what the tonal system of an African language is like, see the online sketch at [20].

With the recent theoretical shift towards constraint-based approaches in linguistics, data from African languages continues to provide some of the most challenging test cases. This is in part because of the large number of languages in Africa, and the great variety of phenomena of all types that they manifest. However, lexical tone plays a central role in the theoretical relevance of African linguistic phenomena [19]. All but a handful of the 1,400-odd languages sub-Saharan Africa are tonal, and in most cases, tone interacts in complex and varied ways with other aspects of the language’s sound system, word structure and phrase structure.

Tone is arguably the most typologically variable of the phonological features used in human language. Tonal systems differ in numbers of phonemic tone levels, in the type of segments that can bear tone, in the role of syllable-
level tonal glides, in the degree of constraint on word-level tone melodies, in the prosodic domains of tone rules, in the presence and realization of features such as downstep and downdrift, and in whether tone coexists with accent or similar features. Of the world’s tone languages, African languages (and in particular West African languages) exhibit the greatest typological diversity.

Finally, the physical medium of lexical tone—the fundamental frequency of the voice, or F0—is also shared with the expressive dimensions of linguistic performance, and with other non-lexical functions that are reminiscent of the role of intonation in languages like English. It is a general characteristic of speech to simultaneously encode texts and contexts, phrases and feelings. In the case of tone and intonation, this duality is expressed in a simple univariate time function, posing the general problem in a way that is particularly amenable to modeling.

For all these reasons, the study of African tonal systems is both scientifically interesting and representationally demanding. There is no aspect of language structure and use that is not effectively connected to this research enterprise, across a wide range of languages, and there are a wide range of perspectives and goals among the researchers involved. As a result, we get the benefit of subdisciplinary specialization with few of the drawbacks.

A great deal of work, both descriptive and theoretical, has already been done on tone in African languages. For hundreds of languages, we have orthographies and at least minimal grammars and dictionaries. A smaller number of languages are thoroughly described. A great deal of work remains to be done even on the best-described languages, and certainly on the many languages that are almost entirely undescribed. Documentation of dialect variation represents an enormous challenge.

The reasons for attempting this enormous task are both scientific and practical. The diversity and intrinsic interest of African languages has attracted a large international community of scholars. Internationally, there are several hundred linguists working on the languages of sub-Saharan Africa—more than 200 have attended individual U.S. meetings such as ACAL. Other scholars are motivated by the need to use linguistic analysis in sociological, anthropological or historical studies. For Africans, the issues are also practical ones: language is on the critical path to goals in education, health, religion and culture.

In several West African countries (among them Cameroon, Nigeria, Ghana and Cote d’Ivoire), there is a history of several decades of research by local linguists, with whom members of this project have been working (in some cases also for decades). Thus there is an academic linguistic community in these countries that we can draw on to compose our on-site research teams. Members of our project already have several cooperative projects underway in these countries, and have firm commitment from local researchers to cooperate in the relevant aspects of the proposed new work.

Other researchers have agreed to test the project’s ideas in on-going work on languages in other parts of Africa, and to give feedback and advice based on their own research interests and experience. These include Lee Bickmore (SUNY Albany), Nick Clements (Paris), Bruce Connell (Oxford), Laura Downing (UBC), Dafydd Gibbon (Bielefeld), Larry Hyman (Berkeley), Francis Katamba (Lancaster), Scott Myers (Texas), David Odden (OSU), Harry Otelemate (West Indies), and Doug Pulleyblank (UBC), among others. Among them, these researchers provide coverage of representative languages from Central Africa, East Arica and Southern Africa.

There are nearly 100 African languages with a thousand or fewer speakers. Many of these smaller languages must be considered endangered. Although this is not our primary goal, our project will be directly involved in documenting some of these endangered languages, and will facilitate the endeavor to document others.

1.6 The Use and Development of Database Technology

The purpose of database technology should be to provide tools for representation, storage and access of data. Let us briefly review the development of this technology and its relevance to current and proposed linguistic data sets. In reviewing this we concentrate on two closely related topics: the data model and the query language. The data model lays out the kinds of structures that are provided for the representation of data and the query language provides an interface to that data. While query languages are designed to be understandable, they have a more important function of being optimized interfaces to the data. In fact it is common to find queries being generated by applications programs such as user interfaces. The prime example of this is the relational data model in which data is represented in tabular form and SQL is the most widely used query language for this data. It is important to note that databases should be considered as abstract databases. The implementation of the data may be very different than its representation. What
is important is that the implementation should support the interface, that is the query language.

Relational databases have not been widely used for scientific data processing. While the “administrative” data – such demographic information on the participants in an experiment – lends itself to a relational representation, the data itself and the scientific analyses and annotations do not. In many current linguistic databases, the raw data consists of digitized audio signals together with “transcriptions” in the broad sense – analyses expressed as symbol strings representing words, phonetic segments, grammatical structures, and so on. All of these can in principle be encoded in relational form, but such encoding is counter-productive for one cannot express typical retrieval requests, which typically involve pattern matching on these analyses, in a relational query language.

Object-oriented and object-relational systems offer a solution to this problem by providing a much wider variety of data types such as arrays, lists and user-defined types with which to encode data. Also programs can be written in the ambient language, typically C++ or Java, to perform any kind of query for automatic analyses, pattern matching etc. These systems provide a uniform programming environment and the expected database functionality: concurrency control, checkpointing and a degree of optimization in retrieval. Relatively expressive query languages, notably OQL, have been developed for such systems. Queries in OQL can be mixed with user-defined methods. Object-oriented databases and technology have been used with limited success in some areas of scientific data processing [21, 22, 23].

The drawbacks are that restructuring of data can be extremely difficult; that optimization technology is still in its infancy; and, when concurrency control is not an issue, the functionality they provide when compared with, say, an indexed file is not worth the commitment to a specific database management system.

For whatever reason most scientific data is held in data formats, and each scientific field has generated its own formats and standards, which is unfortunate given that there is considerable commonality in the types of data and the analysis techniques.

Linguistic databases are relatively new, and there is little experience with standards for overall linguistic database structure, but there are certain lessons to be learned from other scientific areas. The first is that fixing a format for data archiving and exchange is more important than the choice of a database management system, if one is needed.

The second point is that there should be no need to develop a “home grown” format or protocol. There is a variety of generic standards such as ASN.1, XML, and CORBA. With each of these there is a substantial body of software for parsing and indexing so that the adoption of such a standard provides a certain amount of database functionality.

We noted in the introduction that for some linguistic databases, the “design unfolds in the course of the creators’ explorations.” This case poses the greatest challenge to the database aspect of the project, because all database management systems and most data exchange protocols require a type or schema before they can be implemented, and restructuring – as we noted earlier – can be difficult and dangerous. The question is whether there is anything that can be specified in the initial stages of this project.

Recently a few systems have been developed that allow one to put more in the data than there is in the schema. An extreme case of this is semi-structured data in which one only deals with data [24, 25, 26]. However certain data formats also allow this. ACeDB (a “lightweight” data manager popular with biologists) and XML allow such flexibility. In XML [27] one may declare an element as OPEN, indicating that there may be more fields in the data than are specified in the schema. One may also declare the components of an element optional, so that it is possible to create an instance that bears no relationship to its class.

We believe that the adoption of such a flexible format may be appropriate as a starting point for this project. Moreover query languages have been developed for semistructured data that may be useful in this context. These may prove useful for basic analyses and restructuring of the initially “loose” data structures.

Most of the foregoing remarks would apply to any exploratory scientific data gathering project. Let us look at a few of the issues associated with the special forms of data in this project. In the analysis of speech there is an underlying signal to which all analyses are pinned in the sense that each component, say a phoneme, has an associated temporal region of a signal. Representing this relationship in a way that allows us to relate analyses is an important issue. For example, we may want to check the effect of vowel height on the pitch of the voice, after allowing for tone, phrasal position, speaker, and so on. XML, being based on SGML, allows the hierarchical annotation of a linear structure. However it does not easily permit two independent annotations. For this, some kind of “stand-off” markup strategy is needed (see [52] for suggestions of how to use XML for stand-off markup).

A second issue is that of representing and querying the annotations. Typical linguistic annotations (e.g. phonetic transcriptions, morphological analyses, syntactic structures) can be described as finite state automata, and search pat-
terns can be described similarly. Then a query system can be based on the algebra of finite automata, which offers interesting opportunities for optimization via techniques such as “lazy determinization” or “lazy composition” (see [28] for some examples of these techniques used in speech recognition applications). It is interesting to note that semi-structured languages also allow the use and optimization of regular expression matching, and that this approach is a natural one for the annotations we are considering here.

A final point concerns data integration. It is normal to want to make connections across linguistic databases, for instance to tie words in a text with entries in a dictionary. Here we may be able to make use of relatively standard database technology for efficient queries across databases.

1.7 Problems for Computational Linguistics

Perhaps the most significant and enduring contribution of computational linguistics to the study of phonology and morphology has been the demonstration that linguistic generalisations in these fields require no more than finite-state power. Far from demeaning these application domains, this finding shifts the focus of inquiry to the design of higher-level languages and compilation procedures. A linguistic generalisation that is easily stated in an expressive high-level language might compile into an FSA—the machine language—having many thousands of states. On this view, we can have more than one high-level language, selected according to the task at hand by the linguist user. Indeed, compilation algorithms for the last three major paradigms of phonology (SPE, Autosegmental Phonology, Optimality Theory) have been provided. Therefore it is reasonable to assume a finite-state foundation for the following four strands of research.

The representation of multidimensional information about speech. Speech is a coordinated motor activity, situated in space and time; it is also the resulting pattern of sound; and it is also implicitly the cognitive structure intended by the speaker and perceived by the listener. These articulatory, acoustic and cognitive patterns are each complex, all related, and all relevant to linguistic research. For some purposes, we may primarily be interested in symbolic representations of cognitive structures: features, phonemes, syllables, words, phrases. In other cases, we may want to focus on acoustic patterns—such as bursts, pitch trajectories, resonance frequencies—or articulatory data—such as lip opening or velum height. We need to represent the relative timing of these various types of data, as well as the hierarchical relations of the cognitive structures. Thus a /t/ involves articulatory gestures of the tongue and the glottis, coordinated with surrounding and simultaneous gestures in complex ways; it involves acoustic events such as a release burst and the cessation and resumption of voicing, coordinated with other acoustic phenomena such as the time variation of pitch in voiced regions; it is also part of a syllable, a word, and a hierarchy of phrases. This intricate ballet of simultaneity, sequence and hierarchy is fundamental to phonological structure [29]. However, no entirely satisfactory way has yet been proposed for representing and processing this kind of multidimensional information using finite-state methods (see [31] for one promising direction). We will devise a flexible and expressive scheme that permits the kinds of annotation that linguists need, along with methods for restricting search or display to the aspects of interest in a particular case.

Linguistic knowledge discovery. Each item of a speech corpus may contain complex annotations which are difficult to visualise in totality, even for a single utterance. A corpus containing many thousands of such items will be impenetrable, and it becomes necessary to provide methods for extracting useful information. The rich assortment of techniques offered in the literature on knowledge discovery and data mining cannot be applied off-the-shelf. For example, a user may wish to know what kinds of “assimilation” are found in a corpus for a certain language. With hierarchical annotation, the corpus will contain information about individual sounds and corresponding words (e.g. word level: “ten pin”; segment level: /t E m p l n/, with assimilation of /l/ to [m]). Lexical lookup on the word “ten” will give the segments /t E n/ and reveal the assimilation. Research by speech technologists in the area of “pronunciation modeling” has applied machine-learning techniques to induce rules operating on finite automata to generate variant pronunciations, given a pronouncing dictionary with citation forms, and speech data annotated for word identity and phonetic segment sequence. This work can be easily generalized to new languages, and also to distinguish the various different types of phonological processes involved. For instance, Ellison [32] has shown how the abstract notion of assimilation can be expressed as a finite-state device with underspecified labels and particular constraints between label sets.
Automatic learning of paradigms and generation of lemmatizers. Typically, a corpus will contain multiple instances of a word in different forms. These may be variant pronunciations, inflected forms differing in number, tense, etc., or new forms derived by regular word-formation processes (e.g. “SGMLization” or “desktop”). Tone languages bring these problems into sharp focus, thanks to a pervasive kind of contextual variation known as “tone sandhi.” A hierarchical annotation permits phonetic and/or orthographic transcriptions aligned with information about word identity. From this data, it is straightforward to identify all the contextual forms of a word that occur in a corpus. However, no general-purpose corpus is large enough to include every form of every word. Rather, most words will only be present in a fraction of their possible forms, especially in the case of rich systems of inflection and derivation, where thousands of forms of a word may in principle be possible.

There are many practical reasons to want to automate the process of relating word forms to lexical entries. This process is often called “lemmatization,” based on the usage of “lemma” for the headword of a lexical entry. There are well-known techniques for using finite-state transducers to implement such analyzers in a general way—see e.g. [33, 34, 35, 36, 37] for a review. However, even with the help of available compilers, writing such systems is in effect a programming task, and thus requires specialized training and often prolonged debugging. We believe that it should be possible to use a “programming by example” approach to good advantage in this case. The idea is to recruit the linguist to establish a set of paradigms, in roughly the traditional sense, providing a catalogue of examples of inflection and word-formation processes, organized along their natural dimensions, and furnished with as many specific examples as the corpus affords or the linguist is able to provide. The machine then induces an automaton that can (sometimes incorrectly) “fill in the blanks” for new words or for unseen forms of old words. Over time, mistakes or inconsistencies will be discovered, and the automaton improved by an iterative process of refinement through dialogue with the user.

Theory testing. In complex linguistic systems of the type under discussion, it is often difficult to know whether or not a formal grammatical description is consistent with the facts. There may be millions of relevant facts—for instance, instantiations of a complex verb paradigm with a thousand possible forms for each of five thousand verb stems—and the grammar may engage any particular fact through a complex (though well-defined) chain of inferences. In a real sense, the grammatical description is a kind of program, and verifying that the program gives the right output for all possible inputs is hard to do just by thinking about it. In our approach, it is natural to look for computer assistance. For example, we might test a tone analysis expressed in a version of optimality theory, compiled into finite-state automata using the method defined by Karttunen [38]. Karttunen will collaborate on developing this trial application of his system and testing it on some of our data. Additionally, we will explore methods for “source-level debugging” of an analysis, whereby the system identifies data items that are exceptional under the analysis, and identifies those constraint statements or constraint rankings that are problematic given the lexical and surface forms of particular items.

1.8 Scope and Limitations

Even if our project is entirely successful, there will still be data analysis tasks for which researchers will have to write ad hoc Perl scripts or the equivalent. However, we will push back the boundary of the wilderness, so that an increasing number of familiar tasks can be carried out without new programming. As a result, a larger number of working linguists will create and publish a larger volume of shareable linguistic data in electronic form.

There will also remain types of linguistic data for which researchers will have to define new structures. However, we again will push back the frontier, so that there are commonly-understood formats for more kinds of linguistic data, not only for lower-level objects like speech files or textual transcriptions, but also for the structure of large inter-related collections of such objects. As a result, much published linguistic data will become—and will remain—accessible to the research community at large, without new programming for each new publication accessed or each new researcher accessing it. This access will also be available to students and to the general public.

2 Plan of the Project

Our plan is evolutionary. Initially, we will employ existing standards, tools and practices in creation and dissemination of linguistic materials, and in research based on these materials. Step by step, we will refine and replace the standards,
tools, and practices to meet our evolving needs and to exploit new opportunities afforded by the underlying technologies. We will involve other researchers through visits, networked exchanges, joint projects and yearly workshops.

These workshops will bring together for several days the linguists, database theorists, computational linguists and others working on the project. There will be technical discussions within each group, as well as tutorial explanations across disciplinary boundaries, and extended joint exploration of specific problems and solutions. In addition to key African linguistics collaborators, these workshops will bring in outside collaborators with specific technical expertise, such as Lauri Karttunen with respect to applications of the Xerox finite-state tools, or Henry Thompson with respect to the Edinburgh tools for stand-off markup via XML. We will also invite representatives of other groups involved in similar enterprises, such as the EMU group from Australia or the LACITO group from Paris.

While building gradually on what exists, we have to take a longer view to avoid entering blind alleys. Although our ultimate goals are functionally minimal, they are nevertheless ambitious and difficult. For instance, a “digital library” of primary linguistic information must be universally accessible and remain so, and citations must be universally interpretable and remain so. The step-by-step evolution of useful tools in no way guarantees such results. There many prior examples of unfortunate evolutionary dead ends – for instance, the excellent UCLA HyperCard stacks for Sounds of the World’s Languages [39, 40].

In our opinion, this longer-range planning is mainly a problem in database theory and its application. If the underlying data abstractions are coherent and appropriate to their purpose, and if they are mapped appropriately onto open standards (such as ASN.1, XML and CORBA), then we (and others) will be able to keep our databases accessible over time, and also develop (and discard) tools for creation and use of such databases in relative safety.

However, not all of our needs are met by a straightforward application of existing database techniques. We face several significant and general problems for which database theory does not now offer general solutions. One example is the problem of heterogeneity across databases with different authors. Even in the case of mature and well understood database models, different authors may express the same information in different ways, creating familiar problems for anyone who wants to search across several such heterogeneous databases. Another example is preservation of “citations” under updating of a given database across time. These are areas of on-going research. An active collaboration with database theorists is therefore motivated, since our efforts both require guidance and provide interesting problems and test cases.

### 2.1 Standards, Tools and Publication Techniques

#### 2.1.1 Data Representation and Access

Hundreds of linguistic databases have now been published; see [7] and [8] for some catalogues. There are many proposed and de facto standards for representing such linguistic data. The Handbook of Standards and Resources for Spoken Language Systems [41] is an excellent survey. A general approach and an extensive catalogue of particular proposals for encoding various sorts of “texts” in the broad sense was produced by the Text Encoding Initiative [42] (TEI). Several de facto standards have also emerged from the experience of the Summer Institute of Linguistics [43] (SIL).

Current practice remains pluralistic, despite this extensive practical experience and attention to standards. On the other hand, there are many proven formats on which to base the encoding of individual pieces of linguistic data, such as digital audio recordings, transcriptions, and lexical databases. Although the variety of alternative formats presents problems of its own we will not have to break new ground in order to get started.

In contrast, few precedents exist for a consistent formal framework encompassing the overall structure of a variety of linguistic databases, each which may comprise many (often more than 10,000) individual files interconnected in complex ways. Existing databases with similar content are likely to be structured in quite different ways. For instance, some databases encode information in pathnames, so that in TIMIT [44] the path `timit/train/dr1/fcjf0/sal.wav` means TIMIT corpus, training set, dialect region 1, female speaker, speaker-ID cjf0, sentence text sal, speech waveform file, and the path `timit/test/dr5/mbpm0/sx407.phn` means TIMIT corpus, test set, dialect region 5, male speaker, speaker-ID bpm0, sentence text sx407, phonetic transcription file.

The Switchboard [9] corpus of telephone conversations organizes analogous information in a set of five relational tables, expressing information about callers, conversations, ratings, and topics. For instance, an entry in the “caller”
and tells us (among other things) that speaker 1000 is a female born in 1954, from the South Midland dialect region (which is TIMIT’s dialect region 4). Other databases use other strategies, such as file headers or specific markup conventions in individual files.

Now consider the problem of a user who wants to know how the word “water” is pronounced by speakers from various dialect regions of North America. TIMIT and Switchboard each have hundreds of instances of this word, with exact information about where each instance occurs in which digital audio recording, and also considerable (though slightly different) demographic information about the speakers. However, finding the examples requires a significant amount of poking around in documentation, cross-referenced files, path names and headers, in a way that is completely different for the two databases. If we now introduce a third published database (say the CSR database of read sentences from newswires), another substantively different process will be required. Similar issues arise in getting information out of dictionaries, purely textual corpora and indeed every sort of linguistic database.

Several research groups have addressed this problem, at least implicitly, by creating systems for accessing certain kinds of linguistic databases of diverse origin in a consistent way. We know of some early work of this type at AT&T Bell Labs and at IBM Watson Labs, and no doubt there are others. However, we have found only three open, published research efforts in this area. One is the so-called TIPSTER architecture [45], which is an approach to providing a consistent access interface for text databases. second is LDC Online [46, 47], giving access to nearly all of the speech and text databases available from the Linguistic Data Consortium (LDC). A third example is the EMU [48] speech database system developed at Macquarie University.

All of these systems offer a “meta-model” into which diverse database structures can be cast, along with software tools for access once the recasting has been done. While none of these constitutes the general solution that we need, each provides important components of such a solution. The source code of all three is available to us, giving us a reasonable place to start.

### 2.1.2 Data Creation Tools

At the international level, hundreds of research sites and thousands of researchers are creating and using linguistic databases, employing computational tools in the process. Out of this experience we can identify several starting points for our work. We will begin with packages and libraries where we have access to redistributable source code, or with stable and accessible tools with a well-defined interface. We will support transcription in the broad sense, handling of complex text, lexicon development and relevant signal processing/display. An important factor is that the resulting tools should run on a range of popular and accessible platforms, and should be accessible to a wide range of users, novice or expert, poorly or well-resourced, and supported or unsupported.

By transcription in the broad sense we mean the creation of linked annotations such as basic orthographic transcription, phonetic transcription, structural annotation, and so on: essentially, whatever can be represented with finite directed acyclic labelled graphs, optionally aligned with speech data. The result is what we are calling complex text. For lexicon development, we need to support at least the sorts of functions that Shoebox does, along with linkages to speech and complex-text data. Certain signal processing functions are needed, including generation and display of spectrograms and pitch tracks, and (semi-) automatic alignment of orthographic transcriptions with speech.

### 2.1.3 Phased Plan for Tools and Standards

The three stages sketched below will each occupy roughly a year.

In the first stage, we will pursue five goals in parallel, as sketched below, to establish a initial set of tools, standards and practices to meet the basic needs of our project. During the same time period, these initial tools will be tried out, especially in adapting or further annotating already-collected data.

Consolidation. Existing work on tools and standards will be brought together to be evaluated and adapted. Specifically, we will try out and modify various LDC tools, Transcriber [49], HyperLex [50], EMU [48], the Mississippi State
tools [51], Shoebox [3], and various others. We’ll investigate stand-off markup [52] as applied to several standard linguistic database types, and adapt some existing tools to use it.

**Prototyping.** We’ll develop initial prototype tools for web access of linguistic data, based on merging features of existing tools. We’ll configure laptop PCs to run these prototype tools, both as a standalone application and in client-server configurations.

**Markup.** We’ll try out alternative markup conventions for multi-dimensional speech annotation, providing import and export functions for embedded linear or hierarchical markup in common existing formats.

**Search.** We’ll experiment with finite-state search mechanisms over multi-dimensional annotations, providing general projection functions to input or output various aspects of such markup in a user-friendly visual form.

**Citations.** We’ll experiment with mechanisms for persistent and consistent citations across (at least common kinds of) database revisions.

In the second stage, we’ll aim at better integration of tools, refine and more thoroughly test our markup standards, and create a formal proposal for the key kinds of markup. In particular, we want to choose an approach to multi-dimensional speech annotation from among those tried out in the first stage. Key tools should become robust enough to be used in the field with minimal support.

**Integration.** We don’t aim at a monolithic system, but different tools should interact conveniently, should share user interface conventions (“look and feel”), and so forth. The tool set and its documentation must become usable in the field, in the hands of outsiders to the project, and we will begin testing this.

**Markup.** At this point we will have enough experience to make some decisions and propose basic data structures and data formats for the project.

**GUI improvements.** A key goal is finding a workable GUI for using regular expressions to search multi-dimensional annotation.

**Typological Comparison.** Enough data should have been converted or gathered to begin experiments in typological comparison across databases from different languages. This raises some interesting database issues.

**Knowledge Discovery.** We’ll experiment with a knowledge discovery tool for standard phonological phenomena, as discussed above.

In the third stage, we will provide supplementary tools, while consolidating, documenting and disseminating those already developed.

**Publication.** Our primary focus is on networked electronic publication. However, we should also develop various kinds of support for print publication, such as convenient methods for project our hypermedia documents onto pairs of documents, one being a print document (for a monograph or journal article) and the other being a corresponding online document, including just the multimedia portions missing from the print version.

**Knowledge Discovery.** We’ll develop tools for paradigm-based learning and lemmatizer generation, as described above.

**Theory Testing.** We’ll demonstrate a finite-state implementation of a tone analysis tester, as described above.

**Tutorials.** Interactive tutorials and so on for the tools will be developed, to make it easy for would-be users to download them and get started without instruction.

### 2.2 Plan for Research in Africa

Our Africa research has two main goals: to produce new primary descriptions with the project’s tools, and to use this material in linguistic research. The first step tests the approach from the perspective of linguistic data creation; the second step tests it from the perspective of subsequent use by the research community. We expect this process to uncover many problems, both in design and in implementation, and also to suggest many opportunities for new kinds of work.

The descriptive work needs to be broad enough (in terms of the languages covered and the styles and interests of the linguists involved) to ensure that we design broad enough solutions. The associated linguistic research needs
to be varied enough to ensure that a large enough range of modes of search and retrieval can be supported. Both
the primary description and the associated research need to be successful enough and attractive enough to get other
linguists involved in carrying the ideas forward at the end of the project.

2.2.1 Research sites

We have chosen Cameroon and in Cote d’Ivoire as our initial sites for gathering primary data, with plans for additional
work in the neighboring countries of Nigeria, Benin, Ghana, and Burkina Faso. These choices seem particularly
appropriate, because of the exceptional diversity of the languages available for study, because these locations already
possess an infrastructure that will permit our project to get off the ground quickly, and because our key personnel are
already experienced in doing research in these locations.

Cameroon offers a diverse stock of languages, from Chadic in the north, through Adamawa/Ubangi in the centre,
and to Bantoid in the south. The Bantoid group itself covers a rich diversity of languages, from narrow Bantu in the
south-east of Cameroon to Grassfields, Beboid, Jarawan, Ekoid, Tivoid, Jukunoid, Cross and others in the south-west.
Co-PI Steven Bird has recently returned from two and a half years of research in Cameroon, where he specialised on
Grassfields languages and gained extensive experience in leading team-based research in Africa. For our project he
will direct a research effort in Cameroon in close association with key individuals from all of the institutions presently
engaged in research on Cameroon’s languages.

Cote d’Ivoire, the other initial site, provides access to four major groups of Niger-Congo languages that have figured
most prominently in the linguistic description of African languages—Kwa, Kru, Mande, and Gur. The University of
Cocody is already the host of an ambitious research project in the comparative phonology of fifteen Kwa languages
which is to be completed this year. The director of this project, William Leben, who is a Co-PI of the proposed new
project, spent a total of fourteen months in Cote d’Ivoire working on this project and will benefit from the momentum
and the infrastructure that this project has established.

Nigeria, like Cameroon, is host to languages of the Niger-Congo, Afro-Asiatic and Nilo-Saharan families and
is the most linguistically diverse country in Africa. Co-PI Akinbiyi Akinlabi, a Nigerian national who is associate
professor of Linguistics at Rutgers University, received a significant portion of his linguistic training in Nigeria, taught
at University of Ibadan, and has collaborated on past linguistic research with a number of linguists currently in Nigeria
who we hope to draw on for the proposed project.

These choices give us access to large bodies of prior research in many of the languages we will include in our
sample, and to an extensive network of contacts in the countries involved.

2.2.2 Research plan, year by year

Our research in Africa and development of a database framework in the U.S. will need to proceed in parallel, with re-
results in one dictating the next steps in the other. Also, the different research teams in Africa will need to be coordinated,
to better learn from one another’s activities and to provide coherent input to the framework development effort.

Our earliest efforts in developing a framework will thus be based on work already in progress by the project
participants and their close associates, focusing on a few languages where there is already reliable, well understood
primary data. In later stages, we will extend data collection efforts over a wider-ranging and less well studied set of
languages, and will involve a larger group of collaborators.

Year 1: In the first year, we will focus on applying the initial framework to work that is already underway, and in most
cases to data that has already been collected. This will include Leben’s Ivory Coast data, Bird’s Cameroon data, Aki-
labi’s work on Yoruba, and two current dictionary projects underway at Penn (a Yoruba dictionary by Yiwola Awoyale,
and a dictionary of several Manding languages by Moussa Bamba). We will also begin preparing the personnel at some
of our African sites.

Year 2: During the second year, we will focus on the problems that arise in helping outsiders (whether producers or
consumers of data) to work in the framework, by working with a gradually expanding network of collaborators, both
in the U.S. and outside it. We will also prepare the ground for a larger project in Africa in the third year.

Year 3: In the third year, we will focus seriously on the problem of scaling the work up. We will plan and carry
out some large-scale primary description in Africa, involving a larger number of collaborators working together. We
will work to spread the framework to a broad cross-section of Africanist linguists around the world. We will “seed” the work in a variety of other linguistic subdisciplines, ensuring that the framework is genuinely adaptable across the spectrum of interests and applications.

2.2.3 Project organization and activities

Each team in Africa will be directed by one of the project co-PIs and by a resident director.

For Cameroon the resident director will be Prof Maurice Tadadjeu, head of the linguistics department at the University of Yaoundé I and president of NACALCO, the national association responsible for language development. NACALCO will provide office space and access to staff with local knowledge of all areas of Cameroon. Ngessimo Mutaka, an internationally recognised tone specialist at the University of Yaoundé I will coordinate the research side of the project. The field linguists will be identified by Tadadjeu and Mutaka and will include promising PhD students and recent PhD recipients. SIL Cameroon, which has projects in about 25 languages, will support the activity by producing its descriptive work on tone using the framework and tools we provide. SIL’s international tone consultant, Keith Snider, based in Cameroon, will provide consultant advice to the project.

The project will start with three languages in the first year, expanding in the second and third years. For each language, the team will collect and compile lexical data, tone paradigms and texts from a variety of genres, and record and transcribe the data for a number of speakers. They will prepare their findings in the form of a tone description document, with guidance from a local consultant and/or some of our above-named collaborators under the close supervision of Mutaka.

Steven Bird will visit Cameroon early in the first year to set up the project. The training of field linguists will begin during this visit and will be continued afterwards under the supervision of Mutaka. The rest of the year would be devoted to collecting and refining the field data.

Early in the second year, Bird would visit Cameroon to train an enlarged team in the tools developed in the US. Training would essentially consist of practicing how to convert data from the first year of the project into the new framework, and use the framework to guide further data collection. The remainder of the second year will involve work on five more languages.

At the start of the third year, the project would expand to a dozen or more languages, covering as much genetic diversity as possible. Bird would make an extended visit of up to two months to help put the second year’s results in order, to coordinate the expansion, and to assess the strengths and weaknesses of the framework in supporting linguistic research.

The parallel team in Cote d’Ivoire would operate in an analogous fashion. William Leben will direct the project, with Firmin Ahoua (the local director of the current comparative Kwa language project funded by NSF) serving as the resident director. For two of the fifteen languages of the current project, four members of the current Ivoirien team, consisting of PhD linguistics students and faculty, will analyze the existing data and collect new data to encompass more kinds of phenomena in these languages and consequently to deepen the analysis.

Leben will travel to Abidjan for a week at the beginning of the first year to help initiate the team’s activities. Around the beginning of the second year, Leben would spend up to two months in Abidjan to oversee its expansion. The enhanced team would add to the existing team researchers in the three other language groups of Cote d’Ivoire: Kru, Mande, and Gur. During Leben’s visit, a training session would be organized to train the team in using the tools developed in the US during the first year. As in Cameroon, training would essentially consist of practicing how to convert data from the first year of the project into the new framework.

At the time of the second year visits of Bird or Leben or both, Akinlabi will travel to Nigeria for four weeks to set up a team in Ibadan, modeled after the first year experiences in Cameroon and Cote d’Ivoire. The resident director will be Francis Egbokhare, a phonologist and Senior Lecturer in the Dept. of Linguistics at the University of Ibadan. Akinlabi will also work with one or more Co-PIs in Africa on preparations for additional teams for the third year, in Nigeria, Ghana, Benin, and Burkina Faso as the situation permits.

For the third year, the aim is to have several teams of about four members each, modeled on the ones from the first year in Cameroon and in Cote d’Ivoire. Their purpose would be to expand the typological diversity of the languages further, to test the project’s framework and methodologies against a significant body of new data, and to deepen the level of understanding of these languages, as a further test of the analyses that come from applying the framework to
data.
3 Liberman: Results from Prior Support

NSF IRI-9528587: Improved Speech and Text Data Resources
Project Period and Amount: 1995-8
PI: Mark Liberman

Goals of the Project. The Linguistic Data Consortium (LDC) 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. Since its foundation in 1992, more than 225 organizations have joined LDC, and an additional 520 have obtained one or more LDC databases as non-members. LDC has published more than 110 electronic databases, comprising some 360 CD-ROMs and 20 packages to be downloaded over the internet.

The LDC has become self-supporting, so that membership fees and data sales provide all the funding for core consortium activities. The current NSF cooperative agreement, IRI-9528587, has 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.

Results of the Completed Work. 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 this 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” project; has collected and published hundred of millions of words of broadcast transcriptions for language-modeling purposes; has published the Topic Detection and Tracking (TDT) pilot corpus, and is now working to collect and distribute the large TDT-2 corpus; 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.

Publications resulting from award:

The most important publications are the new databases that the LDC has released since the award began in September of 1995. These are listed below with their month and year of release; detailed documentation is available on the LDC web site http://www.ldc.upenn.edu.

During the period of the award, thousands of publications by researchers around the world have reported on research based entirely or in large part on LDC databases. At major international meetings in the speech and language technology area, such as Eurospeech or ICSLP, roughly half of all papers present research based directly on LDC data. At most of the important DARPA-sponsored workshops in this research area, essentially all of the presentations are based on LDC data. There have also been a number of publications from the LDC documenting the database production and distribution process, but these are of relatively little importance, and are not cited here.

2. CELEX-2 Lexical Databases of English, Dutch and German. Dec-95.
3. MUC VI Message Understanding Corpus. Jan-96.
5. SPANISH NEWS TEXT. Mar-96.
7. CSR-IV (DARPA Continuous Speech Recognition Hub 4) May-96.
8. FFMTIMIT May-96.
10. CSR-IV (DARPA Continuous Speech Recognition Hub 3) Jun-96.
11. DCIEM/HCRC MAP TASK (Sleep Deprivation Study). Aug-96.
16. CALLHOME German Telephone Conversations. Jan-97.
32. CALLHOME Mandarin Chinese Jun-97.
37. JEIDA (JCSD) Japanese Speech Data Channel 0. Aug-97.
38. CALLFRIEND Tamil Telephone Conversations. Aug-97.
41. CALLFRIEND Egyptian Arabic Telephone Conversations. Aug-97.
42. CALLFRIEND Farsi Telephone Conversations. Aug-97.
43. CALLFRIEND Vietnamese Telephone Conversations. Aug-97.
44. CALLFRIEND Hindi Telephone Conversations. Aug-97.
51. CALLFRIEND Mandarin Chinese (Taiwan Dialect) Telephone Conversations. Aug-97.
52. CALLFRIEND German Telephone Conversations. Aug-97.
54. SWITCHBOARD-1 Release 2 Aug-97.
56. SWITCHBOARD Speaker ID Corpus. Aug-97.
60. LLHDB (Lincoln Lab Handset Database). Oct-97.
63. DARPA Hub-5 Mandarin Telephone Transcriptions. Jan-98.
64. DARPA Hub-5 Mandarin Telephone Speech Corpus. Jan-98.
68. CSR-V Hub-4 Development and Evaluation Data. Mar-98.
74. Spanish News Transcripts (Hub-4). Apr-98.
75. Taiwanese Putonghua Speech and Transcript Corpus. Apr-98.
4 Buneman: Results from Prior Support

NSF CCR-92-16122: Collection Types in Programming Languages and Databases

Project Period and Amount: 1995-8
PI: Val Tannen
co-PI: Peter Buneman

Goals, Objectives, and Targeted Activities. The purpose of this contract was to study the efficient representation and querying of collection types in both programming and databases. It started from work on an simple algebra and syntax for complex object languages that uniformly handles a variety of collection types: lists, multisets and sets, as well as record and variant types. This work has been applied to the problem of querying, transforming, and integrating a variety of non-standard databases such as scientific data formats. It also included the investigation of optimization techniques, aggregate operations and the extension of the basic principles of collection types to other types such as arrays and semi-structured data.

Indication of Success. The greatest success of this project has been its contribution to the Kleisli (now K2) system of integration of non-standard database sources. One of the most interesting developments in scientific databases, especially those used in biology, is the diminishing importance of standard database technology. The data associated with the human genome project is complex and is evolving rapidly. Because the data is so volatile and the boundaries of the domain are ill-defined, efforts to build large integrated repositories using standard relational or object-oriented technology have met with mixed success, and there is increasing reliance on tools for integrating a variety of complex heterogeneous data formats that have been designed for the transmission and archiving of data. The problem of providing scientists with tools for understanding, integrating, and analyzing the proliferating data sources at their disposal is one of the great challenges facing database research.

Because “drivers” for most current data sources have been constructed, Kleisli has been used extensively within bio-informatics projects both within and outside Penn. The software is now deployed in a number of pharmaceutical companies. We have applied for patents on optimization techniques for aggregate queries and for query languages for semi-structured data.

Potential Related Projects. Much of the research in this project resulted from the fusion of ideas in programming languages and in databases. The PIs feel that there is more to be gained from this interaction, especially in the study of type systems for object-oriented languages.

Project Impact.

- The project has been used partly to fund two PhD students, both current, one female.
- Some of the material on collection types is used in Penn undergraduate and graduate courses on databases. It is also used in a course in parallel computation.
- The ability to integrate biological databases was of key importance in the foundation of Penn’s Center for Bioinformatics, the first in any US academic institution.
- A number of industrial collaborations are based upon our database integration work, notably with SmithKline Beecham. As mentioned above we have filed for two patents based on optimization work.

Publications resulting from award:


Software created and how being made available to the research community.

Details about the Kleisli prototype are available at http://www.cis.upenn.edu/~db/home.html. This page contains:

- CPL Web Service: allows you to interactively evaluate CPL queries, and to create and execute parameterized queries encoded in HTML forms.
- Form Based Human Genome Map Search: a sample parameterized CPL query encoded in an HTML form.
- CPL manual: description and examples of the language.

Kleisli is also being used in several industrial and research projects outside Penn: 1) within SmithKline Beecham; 2) within Lockheed-Martin; and 3) at the University of Manchester; 4) by Bob Grossmans’ group at University of Illinois - Chicago; and 5) numerous projects associated with the Limsoon Wong’s group in the Bioinformatics Center, Kent Ridge Digital Laboratories, Singapore.
5 Leben: Results from Prior Support

NSF SBR-9514718: Comparative Phonology in Ivoirian Languages
Project Period and Amount: 1996-98
PI: William R. Leben
co-PI: Firmin Ahoua

Goals, Objectives, and Targeted Activities. This project is currently in progress. Its goal is to examine fifteen closely related languages, to investigate the possibility that similar phenomena (vowel harmony, tonal regularities) can extend over different phonological domains in different languages. The purpose is to help to develop of a theory of prosodic domains in phonology that will lead naturally to accurate treatments in all languages despite their differences. The project involves gathering comprehensive phonological, morphological, and syntactic data for a set of fifteen Kwa languages, and comparing both the data and prospective phonological analyses across this set.

Indication of Success. The most encouraging indication of success is that the foreseen obstacles to success have all been overcome. Leben was on site for a bit less than the project’s projected 31-month duration, yet the team has performed well even in his absence. This is indicated by the regularity with which chapters for the two-volume final project report have been sent from Côte d’Ivoire to the U.S. Internet communication between the U.S. and the Ivoirian team has been frequent and useful. An earlier indication of success was the project’s successful training of a set of fifteen Ivoirian graduate linguistics students with native fluency in the languages of the project. With several months of training to supplement the training received in their linguistics courses, they were able to design questionnaires for elicitation that were well coordinated among the languages. They also succeeded in carrying out the actual data collection in the field as planned. A third sign of success is the cooperation received from the faculty of the Department of Linguistics in Côte d’Ivoire in verifying the accuracy of the researchers’ transcriptions of the data.

Potential Related Projects. The project suggests that detailed cross-linguistic studies of very closely related languages will yield interesting differences among them. Of course, the most relevant potential related project is the one proposed in the current document, which seeks to develop standards, tools and practices for creating, searching and publishing primary linguistic materials via networked computers. This project proposes to use as input primary linguistic data including the very data collected and analyzed in the Ivoirian project.

Project Impact.

- The project provided invaluable hands-on training in linguistic data collection and analysis to the students involved. This suggests ways in which future research projects can include a significant training component.
- The project anticipates the publication of a two-volume final report consisting of comparative research on fifteen languages, among them some endangered languages.
- The project has created a research-oriented and result-oriented infrastructure that promises to enable future projects in Côte d’Ivoire to get off the ground quickly.

Publications resulting from award:

2. (in progress) The Comparative Phonology of Kwa Languages in Côte d’Ivoire, two volumes, W. R. Leben and F. Ahoua. Publication by CSLI Publications is anticipated.
6 Dissemination and Institutional Commitments

6.1 Dissemination Plans

For basic dissemination of tools and data, the answer is easy—everything will be on the internet. Beyond this, we will do three things to inform the research community and the public about our work. First, we will hold yearly workshops for the circle of immediate collaborators, who will number about 20 apart from the project principals. Second, in addition to submitting papers for suitable meetings such as ACAL, LSA, ICSP and ICPhS, we will also request special sessions or panel presentations at these meetings. Third, we will prepare interactive network-based course materials, test them in our own teaching, and advertise them widely.

The course materials that we develop will include short interactive demonstrations suitable for use in undergraduate courses or in high schools; special packages of tools and instructions for use in “linguistic field methods” classes; and more advanced interactive demonstrations suitable for use in graduate courses in phonetics, phonology or computational linguistics. The senior personnel on this project teach introductory linguistics, phonology, phonetics, field methods, and/or computational linguistics at three institutions, and are already using approaches of these kinds.

One of the outreach goals will be to make the content of the published materials culturally as well as scientifically interesting. For example, one of the bodies of data to be prepared in the first year of the project is a set of Yoruba “talking drum” recordings, already made but not yet annotated, analyzed or published. In this material, a trained drummer speaks, chants and drums a large set of passages in Yoruba, such as proverbs, praise songs, and verses from Ifa divination. In addition to posing an interesting set of questions about Yoruba prosody, in modeling how the drumming transmits its message by presenting a formally stylized of the normal tonal patterns of the language, this material has intrinsic aesthetic and cultural value. We will also prepare a digital version of a videotaped lecture by the drummer (a professor at the University of Ilorin in Nigeria) on his craft, its history and its cultural context. In all of our data collection, there will be a component of such material: songs, stories, poems, oral histories and so forth.

The materials developed for these courses will be advertised through mailing lists (such as LINGUIST and corpora), and through the Penn African Studies Center web site, which has become perhaps the largest Internet source of information on Africa, with more than a million hits a month.

6.2 Institutional Commitment of Space and Equipment

Office space and basic equipment. Physical space is not much of an issue for this project. All of the personnel involved have offices at their host institutions, as well as suitable desktop and portable computers with network connections. The space that matters for this project is space for data.

Cyberspace (Web Server and Data Archive): The Linguistic Data Consortium host organization (at the University of Pennsylvania) owns one large server (morph.ling.upenn.edu, a SUN Ultra Enterprise 4000 with 2 167 Mhz UltraSPARC processors and 1G of RAM), and half of another (unagi.cis.upenn.edu, a Sun Ultra Enterprise 4000 with eight 167Mhz UltraSPARC processors and 1.25GB of memory). These two servers are connected by a high-bandwidth link, and together spin nearly a terabyte of RAID disk arrays used for LDC-Online and for storage of LDC databases in process of construction. It is planned to add high-density linear tape robots for “near-line” storage of even larger archives, e.g. of video data. Space on this facility is available for internet publication of nearly arbitrary amounts of data from the proposed project, at low marginal cost.

CD-ROM Production: The LDC owns several systems for production of writable CD-ROMs, including two Kodak systems capable of writing up to 75 disks without reloading. These will be available for CD-ROM production associated with this project. As standards for writable DVD-ROMs stabilize, these will be available as well.
7 Yearly Performance Goals

This section sketches goals that are described in more detail within the body of the proposal.

7.1 First year

Development of prototype tool set from existing sources. Exploration of alternative of markup techniques and overall database structure, with provision for persistent citations. Experiments with finite-state search techniques over multidimensional annotations. Encode enough data (old or new), and do enough analysis of it, to stress tools and formats. Use of prototype tools to generate interactive (web-based) educational materials, and first use by students in field methods courses.

Applications of prototype tools and standards to existing data. Establishment of two African research sites, one in Cameroon and the other in Côte d’Ivoire. In both locations, field work and transcription to expand existing sets of primary data, to provide a good first test of the database framework being developed in the U.S. This will be done for two languages in Cameroon and two in Côte d’Ivoire. In both locations, completion of a document reporting the year’s research results, including both data and analysis.

7.2 Second year

Integration of tool set with consistent “look and feel,” robust enough and well enough documented for some field use by outsiders. Workable GUI for finite-state search and retrieval over multidimensional annotations. Formal proposal for markup and overall data format. Data resources adequate to carry out significant typological study. Experiments with phonological knowledge discovery by finite-state methods.

Expansion of the activities at both existing African research sites. In Cameroon, expansion to five languages; in Côte d’Ivoire, to three new language groups. Extended visits by co-PIs Bird, Leben, and Akinlabi to the research sites for training of new researchers in implementing the database framework. Overtures to prospective research centers in new locations, including Nigeria, Benin, and Burkina Faso, which would carry the activities of the existing centers in Cameroon and Côte d’Ivoire into the realm of new languages, with new researchers.

7.3 Third year

Consolidation and wide dissemination of tool set, including interactive tutorials. Support for creation of paper publications. Experiments with paradigm-based learning and finite-state phonological theory tester.

African data collection is scaled up, with significant materials created on more than a dozen languages in Cameroon and a comparable number in Cote d’Ivoire, with preliminary coverage of the Edoid and Lower Cross languages in Nigeria, and creation of some teams in other countries (Ghana, Benin, Burkina Faso) if the situation is favorable.
8 Project Management Plan

There is a natural division of labor among the sites. Database and computational linguistics research will be centered at Penn, along with software development. The linguistic data collections are geographically divided: Ivoirian work with Leben at Stanford, Cameroonian work with Bird at Penn, and Nigerian/Ghanaian work with Akinlabi at Rutgers.

Bird will be spending 4/5 time on the project, and will play a key role in coordinating the various aspects of the project.

Database research. Buneman, Liberman and Bird will meet regularly during the course of the project, along with the two Penn graduate students, to pose problems and evaluate solutions. Other Penn researchers and implementers will be involved as interest and needs dictate.

Software design and implementation. The database component will be discussed in the previously-mentioned database meetings. Other tools will be designed by Bird and Liberman, in consultation with other interested parties. Implementation will be shared by various LDC-based programmers (1/2 time taken together), Bird, Liberman, and the Penn graduate students. Some implementation will overlap with other LDC-centered projects, so that the available human resources will be larger than what is covered under this proposal alone. Primary responsibility for maintenance of project source code will be vested in LDC senior programmer Zhibiao Wu.

Computational linguistics research. Bird and Liberman will meet regularly with the Penn graduate students and with other interested parties. Applications of machine learning to linguistic problems are an area of active research at Penn, so it is anticipated that other students and faculty outside the scope of this grant will become involved. Bird will direct this part of the project.

Linguistic Data Collection. In the early phases of the project, much of the effort will be in transforming existing data into experimental versions of a new framework. This will be carried out by the researchers most familiar with each dataset. As the project goes forward, the emphasis begins to shift to new data creation in Africa. Akinlabi, Leben and Bird have considerable experience managing data collection projects in the areas of Africa where they work. As a result, we can be confident that they will be able to continue and expand this work successfully, even in the face of logistical difficulties that are frequently encountered when doing fieldwork in Africa. The postdoctoral fellow at Stanford will work closely with Leben on the Ivoirian (and related) components of the project.

Outreach. Our goal is to involve ten or more additional researchers as early, active testers of the approach to data creation and use that we are developing. We have also recruited collaborators interested in markup issues, use of Unicode, development of free software for speech research, applications of finite-state techniques, and machine learning. We will establish collaborations through focused visits, either at one of the project sites or at each collaborator’s home base. Yearly workshops, bringing all of the collaborators together for reports and discussion of on-going work, will be crucial to establishing and maintaining these external collaborations. Primary responsibility for these external relationships will lie with Mark Liberman and Steven Bird.

Distributed Collaboration. The geographical distribution of project participants, including the African participants, creates an excellent framework for developing and testing support for collaborative creation and exploration of linguistic databases. Beyond the necessary long-distance collaborations (e.g. from Stanford to Abidjan, or Rutgers to Ibadan, or Penn to Yaoundé), we will strengthen and expand existing research collaborations among the project sites over the course of the project. This will increase the frequency of contact among the project participants as well as providing a testbed for the ability of the project tools to support distributed collaboration.
References Cited


http://www.xrce.xerox.com/research/mltt/fst/articles/jnle-97/rele.html


[38] Lauri Karttunen. The Proper Treatment of Optimality in Computational Phonology.  
http://xxx.lanl.gov/abs/cmp-lg/9804002


[40] Sounds of the World’s Languages HyperCard Stack.  
http://www.humnet.ucla.edu/humnet/linguistics/faciliti/software.htm

http://www.coral.lili.uni-bielefeld.de/EAGLES/eagbook/eagbook


[51] Speech Data Segmentation Tool.  

http://www.ltg.ed.ac.uk/ht/sgmleu97.html