

**The Linguistic Data Repository Project**

**Fall 2000**

**Che Wilcox**

Texas A&M University is hosting a project that will advance the study of the nature and structure of human speech. This project was initiated by a senior honors thesis by Michael Neal Audenaert in April 2000. Since then, six students under the direction of Dr. Lisa Ann Lane and Dr. Dick B. Simmons became involved in transforming the idea into a reality.

The Language Data Repository project will allow the linguistic community to take full advantage of recent advancements in technology. Audenaert explains, “The Language Data Repository (LDR) project is working to develop a software system capable of storing transcripts and recordings of spoken language data and capable of hosting software tools to aid in the analysis of that data.”

The project has been given a good start in the fall of 2000. We accomplished our target goals as well as a fair amount of research in the linguistic field. Development of LDR project is intended to span several years, which will be past the scope of the original creators. In these early stages, it is imperative that we carefully document our work so that we can pass the information on to future teams.

In detail, we will explain the concept - including the overall idea, as well as the need, standardization, and distribution of the project. We will go into depth about the necessity of community participation, and how we propose to expand the project to include as many participants as possible. We will also give information about the linguistics field in general, and we will include information on our research. Finally, we will describe our progress thus far, and the outlook for the future.

## Concept

The concept for the Language Data Repository Project is relatively simple. We propose to develop a software system that will manipulate all types of linguistic data in many different databases. We will develop a standard for handling and storing this data, but we will also incorporate methods for communicating with existing databases. The linguistic community is aware of the need for such a project, and several institutions are in the process of developing standard file formats. Currently, no one is attempting to collaborate with many separate databases and different database formats. The LDR Project is designed to facilitate the development of the software structure and core elements, not to implement the entire program. We have therefore developed a system that will distribute the work throughout the community.

Audenaert goes into great detail about the overall design concepts and initial software structure in his article, “The Language Data Repository: Project Abstract.” Although the article is not published yet, part of his work is worth reproducing here, because it gives a good picture of the idea behind the LDR Project:

The proposed software architecture will enable multiple researchers to store linguistic data from multiple languages on either local machines or non-local machines that can be accessed via a network by multiple users simultaneously. Such a software system will offer two main improvements over current methods of recording transcripts of linguistic data. First, by utilizing machine-readable storage, it will enable linguists to use computational tools to aid in linguistic analysis, thereby reducing some of the “grunt work” traditionally associated with this aspect of research. In turn, there will be an increase in the ability to quickly and accurately test and evaluate linguistic hypotheses. Second, the LDR system, by providing a common format for the representation of data and by reducing the need to obtain physical access to the data, will enhance linguists' ability to document their research and share their results with a greater number of colleagues than previously possible.

The article gives information about design considerations, design overview, representation of data, the client module, and the persistence framework. It is one of the driving forces behind the LDR Project.

The technology has been available to create a software system such as this for several years. There are basically two factors preventing the development of this system. The first is that the linguistic community has not agreed upon a standard for storing linguistic data. A generic database cannot be developed without knowing exactly what sort of data it needs to contain. Several different formats have been developed, but none have been widely accepted. The second is that no one institution has either the manpower or the authority to create this software system. The LDR Project proposes solutions to both of these problems.

Standardization is very important when developing a system that everyone can use. Steven Bird and Mark Liberman recognize this problem, and discuss it in their May 2, 2000 article, “A Formal Framework for Linguistic Annotation:”

While the utility of existing tools, formats and databases is unquestionable, their sheer variety – and the lack of standards able to mediate among them – is becoming a critical problem. Particular bodies of data are created with particular needs in mind, using formats and tools tailored to those needs, based on the resources and practices of the community involved. Once created, a linguistic database may subsequently be used for a variety of unforeseen purposes, both inside and outside the community that created it. Adapting existing software for creation, update, indexing, search and display of ‘foreign’ databases typically requires extensive re-engineering. Working across a set of databases requires repeated adaptations of this kind.

Eight institutions in the U.S. are in the process of developing database archives. We make number nine. Our approach is different because we do not propose a single database to hold all information at a single place. We are working on a format for storing data that incorporates as many meta-description tags as possible. We are also developing this software in the form of a downloadable application that will be able to query many databases in many different locations with different storage formats. The format for uploading will depend on the database. The classes and tools necessary to communicate with these various databases will be available through either a remote communication object located on a server, or as some type of plug-in for the application. Using this approach leaves the LDR Project in a position to develop the software independent of the imposed standards of a particular database. As we develop a default

standard of our own, we will incorporate every aspect of every database in existence. In essence, we are proposing a universal application.

The linguistic community is working hard to come up with a universal standardized format. A seminar titled, "Linguistic Exploration: Workshop on Web-Based Language Documentation and Description" will be held at the Institute for Research in Cognitive Science (University of Pennsylvania) from 9:00 am December 12 through 3:30 pm December 15, 2000. It is organized by Steven Bird and Gary Simons, and sponsored by IRCS, ISLE, and Talkbank:

This workshop will lay the foundation of an open, web-based infrastructure for collecting, storing and disseminating the primary materials which document and describe human languages, including wordlists, lexicons, annotated signals, interlinear texts, paradigms, field notes, and linguistic descriptions, as well as the metadata which indexes and classifies these materials. The infrastructure will support the modeling, creation, archiving and access of these materials, using centralized repositories of metadata, data, best practice guidelines, and open software tools.

Although the LDR Project is not web-based, we will be presenting at the workshop. We will take home valuable information that will help us incorporate every type of linguistic data available. Once the standards have been set, we will be able to define exactly what a modern database should look like. These standards will define our primary format, but the structure of the LDR Project must also allow other database formats to mesh easily into the system.

The second problem for creating a linguistic software system is that the project is enormous. There are many corpus databases, and many different organizations working independently to create the perfect database and software interface. The LDR Project is designed to disperse ownership among the linguistic community. The original team at Texas A&M University has begun the project and designed the overall structure of the program, but once the core of the system is developed, contributions from many individuals and organizations are necessary to complete the endeavor. This would include portions of code, plug-in tools, computer resources, donations and suggestions. This project requires input from people from many different disciplines. Most people will not be able to make long-term time commitments, but dedicated people with a vested interest would be able to manage certain aspects of the project.

The LDR team has come up with a solution for organizing and maintaining contributions. A central web server has been set up to host the LDR web page at [LDR.2y.net/LDR/index.asp](http://LDR.2y.net/LDR/index.asp). This web page contains information on how to get involved with the project. An individual will be able to view a range of needs, and request to work on a certain aspect of the project. When the web site is complete, it will organize two types of users: managers and contributors. Managers will be able to oversee the whole project, define the architecture, define tasks and integrate solutions into the LDR Project. Contributors will be able to define requirements, create data models and data structures, write software and tools, submit data model extensions, suggest task implementations, and report bugs. More information about the web page is described in the community participation section of this paper.

This project is necessary because there are no institutions taking on such a task. We have the technology to advance the study of linguistics, and using this technology would allow us to better understand human language. The benefits of the LDR Project are innumerable, and we must begin immediately.

## **Community Participation**

In order for the Language Data Repository Project to be a success, we must elicit help from the linguistic community. The project is designed to benefit everyone, and its publicity will only result in momentum for the LDR team. Possibly the most significant tangible item resulting from the work completed in the fall of 2000 is the LDR web site. This is our connection to the linguistic community.

The design of the LDR web site takes its roots from a variety of linguistics related web sites. One of the major influences for the design comes from the “Special Interest Group on Computational Phonology” web site located at [www.cogsci.ed.ac.uk/sigphon/](http://www.cogsci.ed.ac.uk/sigphon/). It is maintained by John Coleman and Steven Bird, and currently has 181 members. Membership on the site is “open to anyone who professes an interest in computational phonology.” We have registered at several related sites, and have joined mailing lists such as LINGUIST List ([linguist@linguistlist.org](mailto:linguist@linguistlist.org)).

In order to appreciate the full extent of the LDR web page, one should visit it at [LDR.2y.net/LDR/index.asp](http://LDR.2y.net/LDR/index.asp). We will give a brief description and explanation of the LDR web

site here as well. Basically, there are sixteen menu options: main, home, who's who, register, concept, updates, development team, development, contact, documents, logout, important news, message board, add news, delete news and moderate message board. If the user has registered as a manager, all of the menu options are available. Registration as a manager requires an administrator password. If the user has registered as a contributor, all menu options are available except for the last three. If the user is not registered, only the first eleven options are available.

#### **Options that are available to anyone visiting the LDR web site:**

- **Main:** Main splash screen with graphics and the ability to log in
- **Home:** Introduction to the LDR Project
- **Who's Who:** List of all members of the web site (name, organization, email, homepage, bio information)
- **Register:** New user registration information (contributor or manager)
- **Updates:** Current status of the project
- **Development Team:** Information on LDR team members at Texas A&M
- **Development:** Research information and general project information
- **Contact:** Contact information
- **Documents:** Research Papers used for reference and documents produced within the LDR team
- **Logout:** Logs the user out of the system

#### **Options that are available to contributors and managers:**

- **Important News:** Information geared towards people making a commitment to the project
- **Message Board:** Allows members to post or review information on the site

#### **Options that are available only to managers:**

- **Add News:** Allows manager to add news items to the web site
- **Delete News:** Allows manager to delete news items from the web site
- **Moderate Message Board:** Allows manager to delete items from message board

We came up with a structure that will allow contributors to submit information to the managers via a database. The flow of information should go something like this: a user requests a feature to be added to the LDR Project; a manager reviews the request; the manager makes a request to all users; a user picks up the task and completes a proposal form; the manager reviews the proposal form and accepts it; the user develops the item.

## **Research**

The research compiled as of October 18, 2000 consisted primarily of links to related linguistic web sites and a few articles found at Evans Library. Through this research, we concluded that ours was the only project that is attempting to collaborate all linguistic corpora into a massive single format that incorporates many existing databases. We found several commercial entities have produced software that will index and sort corpora, and several educational institutes have corpus projects, but none that was similar to the LDR project.

The most significant project that we found relating to the LDR project was the Gutenberg Project. It includes an 80 million word corpus searchable in real-time. This project is quite simple in that it searches for words in over 1350 online texts. It is not a very structured project in that it uses no “pre-generated data indexing.” Thus, it does not produce a means to sort and store specific words collected by someone working out in the field. Some of the important factors that have been described by several sources to take on a project such as ours included things such as tagging, parsing and keying texts in.

Before October 18, we had not produced anything specifically useful to the LDR project that we had not developed already. We decided to focus on specific aspects that would be useful to the software design of the LDR project. We found a true type font that implements the Unicode standard that will be useful to the format of data storage, and looked for information on some of the problems experienced with client server architectures, particularly in the area of data synchronization.

Since October 18, we discovered information on several database formats. For example, we found systems such as TIMIT, PARTITUR, CHILDES, EMU and Project Archivage. All of these formats are referenced on the Linguistic Annotation web page at [morph ldc.upenn.edu/annotation](http://morph ldc.upenn.edu/annotation). These formats are all different, but they all seem to use meta-

description tags to describe linguistic data. Extensive research on these formats will continue in the spring of 2000, but for now we must be aware that they exist as we develop our own default format.

We also found a wealth of information after following Internet links relating to Steven Bird and Gary Simons. The two most important discoveries were that the problems relating to linguistic database standards have already been well documented, and that there are many projects working on linguistic databases.

Gary Simons outlines some of the problems of standardization in his June 12, 2000 article, "Developing an infrastructure for online linguistic archives." He states:

Though we might wish for an ideal world in which all language documentation was archived in a consistent manner in a single archive, it seems clear that the world we live in requires that any institution be able to host an archive of the research it has pursued or sponsored. But developing the infrastructure for such an archive (including delivery mechanism, formatting standards, and supporting software) is a huge task that is beyond the capacity of any of these single institutions to accomplish on its own.

He proposes several necessary steps to solve this problem:

A set of standards for meta-description and data formatting needs to be adopted by the community and managed at a single, centralized site that can be accessed by all archives, linguists, and software developers. Similarly, software that supports these standards needs to be deposited at a single, centralized site so as to prevent individual archives and researchers from having to reinvent it.

Standardized meta-descriptions for every holding of each archive need to be deposited in a single, centralized catalog so that researchers the world over have only a single web site to consult in order to find out what is available in all the linguistic archives on the web. A standardized system for identifying and classifying the world's languages needs to be maintained at a single, centralized site so that resources about the same language are consistently identified as such in every archive that contains documentation for it.

Gary Simons has also compiled a list of institutions that are in the process of developing

database archives. These include:

- The Linguistic Data Consortium at the University of Pennsylvania
- The Linguist List's proposed Digital Library of Endangered and Minority Language Data
- The planned electronic wing of the SIL Language and Culture Archive
- A collaboration between the University of Arizona and a university in Mexico to document languages of Mexico
- The proposed web-based Archive of Indigenous Languages of Latin America (AILLA) at the University of Texas
- The Comparative Bantu On Line Dictionary (CBOLD) at the University of California, Berkeley
- The Center for the Documentation of Endangered Languages (CDEL) at Indiana University
- The Project for the Documentation of the Languages of Mesoamerica (PDLMA) at SUNY Albany

Initially, this new information caused us to reevaluate our objectives. We were unsure if the LDR Project would be able to contribute to the linguistic community as much as we had intended. Fortunately, we concluded that not only does our project meet the needs of the linguistic community, but it also takes the concept of a linguistic database archiving to a higher level. We are right on track, and we should be able to offer not only our software system, but also a good resource of information to the linguistic community.

## **Progress & Outlook**

We have accomplished quite a lot in the fall 2000 semester. We have a firm understanding of the need for the LDR Project, and why we should develop it. The structure of the project and necessary elements in the code have been identified and outlined. We have also begun writing and documenting the code. Several meetings have been held throughout the semester to go over our development process with the team.

For future development, we discussed the need for a specific format for documenting what we are doing. A template of some kind needs to be developed to document class descriptions, attribute descriptions, operations, and the algorithms. We also discussed the need for a predefined format for submitting information to the LDR project. This should be a template that serves as a guideline for submitting this information. The predefined documents are the proposal form, the requirements document, the design document, code templates, and the test document.

The future of the LDR Project is promising. We have a basic structure, and the information we take home from the Linguistic Exploration Workshop will allow us to further define our software system. The LDR team has expressed an interest in continuing the project, and we hope to make a significant impact on the linguistic field very soon.

## **Team Information & Conclusion**

Dr. Lane has overseen the development of the linguistic aspect, and Dr. Simmons has overseen the software development. Our project manager Neal Audenaert developed the idea, and has directed the team through the initial stages of the project. Neal Audenaert, Ryan Saunders, Matt Benton, and Blaine Young have begun to write the code and document some of the information. Travis Reed has contributed ideas and suggestions to the development of the web page. Che Wilcox has researched the subject, developed the web page and documented our progress. We all worked well as a team.

The fall semester of 2000 has seen a good deal of development for the LDR Project. Thanks to the collaboration between the Computer Science department and the English department, we have been able to give the linguistic community the advantage of recent advancements in technology. There has been a need for a system like the Language Data Repository project for quite some time, and we are excited to be a part of its creation.

## References

Audenaert, Neal. "The Language Data Repository: Project Abstract." (April 2000).

Bird, Steven and Coleman, John. "Special Interest Group on Computational Phonology." (Dec. 2000). <[www.cogsci.ed.ac.uk/sigphon/](http://www.cogsci.ed.ac.uk/sigphon/)>

Bird, Steven and Liberman, Mark. "A Formal Framework for Linguistic Annotation." (May 2000). <<http://www ldc.upenn.edu/sb/home/publications.html>>

Bird, Steven and Simons, Gary. "Linguistic Exploration: Workshop on Web-Based Language Documentation and Description." (Dec. 2000).  
<<http://www ldc.upenn.edu/exploration/expl2000/>>

Simons, Gary. "Developing an infrastructure for online linguistic archives." (June 2000).  
<<http://www.talkbank.org/resources/simons.html>>