The Endangered Language Documentation Electronic Resource (ELDER): An online tool for lexicon creation and corpus cross-referencing

by

Anna Kathryn Belew

A Project Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master in Arts
Department of Applied Linguistics
At Boston University

May 2010

Table of Contents

i. Abstract

ii. Acknowledgements

1. Documentary linguistics and best practices
   1.1 E-MELD’s best practices

2. ELDER: Origins
   2.1 Ideology

3. ELDER: Technology
   3.1 Underlying database: MySQL
   3.2 Web development: PHP

4. ELDER: Functions
   4.1 Lexicon
   4.2 Sentences
   4.3 Other functions
   4.4 Future functions

5. ELDER: Implementation
   5.1 Single-user
   5.2 Multi-user

6. Conclusion

Abstract

As linguistic documentation efforts move into the 21st century, there is a growing need for tools to aid the digitization of language data. Traditional paper-based data storage is quickly becoming outmoded, and increasing numbers of linguists are trying to find electronic options for data management. In response to this need, I have created an online database tool for storing, sharing, and organizing language data: the Endangered Language Documentation Electronic Resource (ELDER). ELDER is primarily a tool for lexicon creation, but features capabilities beyond a traditional dictionary: it allows users to embed audio files into dictionary entries, to cross-reference words with their uses in sentences and texts, and to ensure that their data is compatible with the field's best practice standards, among other things. This paper aims to describe the goals, creation, and implementation of ELDER, as well as plans for its future improvement.

Acknowledgements

Without the following people, ELDER would be nothing but a smoke dream in the mind of someone with no master's degree; to them I extend my wholehearted appreciation and thanks.

I offer my sincerest gratitude to my advisor, Cathy O'Connor, for her mentorship, encouragement, and support. Without her I would never have dared apply to the graduate program in linguistics, and without her I would not find myself on the cusp of completing my degree. It is thanks to her that I found my calling in documentary linguistics; her influence has shaped the course of my entire academic career. I hope someday to live up to her example of scholarship, teaching, and just plain human decency.

To my best friend Connor Shaughnessy I cannot express enough thanks. His selfless, tireless, compensation-less work on ELDER borders on saintly; I can think of no one else who would willingly spend a Sunday night programming a database for someone else's thesis. ELDER could not exist without him, and I am fairly sure he deserves an honorary master's degree.

I thank Ariane Ngabeu, my Medumba teacher; her consultancy was invaluable in this project. I am lucky to have known her as a language teacher and as a friend. Mə lƆp tə!

Overwhelming affection and gratitude go to my partner Zac Butcher. Without his support, patience, and pep talks, I could never have finished this degree with my sanity intact.

Enormous thanks to my mom for being the best editor and cheerleader I could ask for. I would also like to thank Andrei Anghelescu, Katie Franich, and Nick Danis for being fantastic collaborators in the BU Medumba working group. Thanks to Devon Shaughnessy for creating ELDER's logo.

Most of all, I thank my dad. He would have thought this was so cool.

1. Documentary linguistics and best practices

At the beginning of the 21st century, there are nearly 7,000 living languages on Earth. By some estimates, at least half are likely to be extinct within the next century.[1] When those thousands of languages die, they will likely leave no trace. Unlike a stone tool or a cave painting, unwritten language is not concrete; it exists only in the minds and tongues of its speakers. When those speakers die, or cease to use that language, it disappears entirely. Material artifacts are often durable enough that we can study them millennia after their creators are gone, but languages leave no such clues. A language, when it blinks out of existence, leaves nothing behind for posterity.

The documentary linguist's role is to create records of these languages, to ensure that something of them is preserved for the human knowledge base. She aims to gather, organize, and disseminate information on languages which might otherwise be lost entirely. Whether or not languages should die, or should be documented at all, is an argument well beyond the scope of this paper; suffice to say that among most linguists, the normative belief is that languages should be documented, and primary data such as audio and video recordings should be compiled and preserved. This field of documentary linguistics is relatively new, but has been gaining support and attention steadily for the past twenty or so years, and for good reason. As languages continue to die ever more rapidly, the need for a widespread and concentrated documentation effort has become more urgent.

Certainly it is a boon that more time and effort than ever before is being directed into language documentation. However, it is imperative that this time and effort be put to work

[1] UNESCO Culture Sector, "Safeguarding Endangered Languages," UNESCO, http://www.unesco.org/culture/en/endangeredlanguages

efficiently. The relatively few documentary linguists in the world, as well as the relatively limited amount of resources available to them, would be put to much better use if they spent their time building on existing knowledge rather than collecting redundant data. If a full record of the pronoun system of Twendi existed, but only as a notebook in a filing cabinet, anyone wishing to document Twendi proforms would have to compile that record all over again. Additionally, the fact that many languages being documented are nearly extinct means that any information lost could be irreplaceable. Cassette tapes could be lost in a fire, videos could be trapped in obsolete data formats, or digital transcriptions could be garbled by the advent of new coding standards, and replacing these resources would be impossible. The increasing availability of electronic resources—e.g. high-volume storage media, the internet, and more affordable digital recording equipment—has allowed documentarians to preserve and organize larger volumes of data than ever before, and to make that data more widely available. However, as with any other form of documentation, there are pitfalls in these technologies that must be avoided. The products of documentation efforts should be as useful as possible, for as long as possible, to as many people as possible; accordingly, a set of guidelines has been established to ensure that documentarians are creating accessible, durable data.

One of the major regulatory bodies in documentary linguistics is the Electronic Metastructure for Endangered Languages Data (E-MELD), a group headed by eminent linguists at five major institutions[2]. E-MELD's self-described purpose is as follows:

> Members of the scientific community are faced with two urgent situations: the number of languages in the world is rapidly diminishing while the number of initiatives to digitize language data is rapidly multiplying. The latter might seem to be an unalloyed good in the face of the former, but there are two ways things may go wrong without adequate

[2] For a list of participating institutions and people, see http://emeld.org/organization/index.cfm

> collaboration among archivists, field linguists, and language engineers. First, a common standard for the digitization of linguistic data may never be agreed upon; and the resulting variation in archiving practices and language representation would seriously inhibit data access, searching, and cross-linguistic comparison. Second, standards may be set without guidance from descriptive linguists, the people who best know the range of structural possibilities in human language. If linguistic archives are to offer the widest possible access to the data and provide it in a maximally useful form, consensus must be reached about certain aspects of archive infrastructure.
>
> The primary goal of **E-MELD** is to promote this consensus.[3]

For the purpose of creating such a consensus, E-MELD has set forth a number of "best practices," or recommendations for how to deal with digital language data. Since ELDER was designed with these standards in mind, it will be useful to our understanding of ELDER to briefly summarize E-MELD's guidelines.

1.1 E-MELD's best practices

Since much endangered language data can never be replaced or augmented, it is imperative that whatever data exists be put to the best use. The goal of E-MELD's suggestions is that they will allow digital language data to be more durable in the long run, easier to find, and easier to use.

The first issue in digitalization of language data is the content of the data itself, namely the terminology used in linguistic description. Much of the application of language documentation data is in cross-linguistic studies; for example, if a scholar is researching aspect across Sino-Tibetan languages, she will need to know what, exactly, a documentarian meant by "perfective." Documentary linguists must be explicit in their terminology, and must ensure that

[3] E-MELD, "Homepage," E-MELD, http://emeld.org/index.cfm

anyone viewing their data understands exactly what their terminology means. In response to this need, the General Ontology for Linguistic Description (GOLD) was devised. GOLD aims to be "a solution to the problem of resolving disparate markup schemes for linguistic data, in particular data from endangered languages[4]"; it is meant to represent the full range of possibilities in human language, while providing a standardized terminology for describing linguistic features. In other words, it hopes to circumvent the problem of not knowing what another researcher meant by "perfective" (linguists may differ in their use of terminology). By providing a widely agreed-upon "metalanguage" for linguistic documentation—a language for talking about language—GOLD allows data from different researchers and different languages to be more easily cross-referenced and compared.

The second major issue for documentation is format. When hundreds or thousands of people engage in documentation projects, they are likely to use very different softwares to compile and organize their data. This creates problems for sharing and posterity; if one linguist has stored all of their data in Microsoft Word, that data will be inaccessible to anyone without a copy of Microsoft Word. Even those who could view Word files might not see the information the way its creator intended: IPA symbols could be garbled due to differences in character encoding, or formatting might be displayed incorrectly. In addition, if Microsoft should ever stop supporting its Word software, that data would be encoded in a format that no one could access, since the code for Word is not available to the public. E-MELD thus recommends that all

[4] General Ontology for Linguistic Description Community, "About GOLD," General Ontology for Linguistic Description, http://linguistics-ontology.org/info/about

language data be encoded in Unicode[5], and stored in non-proprietary formats (or at least formats whose code is publicly available).

The third of E-MELD's best practices is straightforward: data should be easy to find, easy to access, and should remain that way. For researchers to be able to build upon one another's work, they must be able to find and access each other's data. A perfectly-formatted, GOLD-compliant resource is an excellent start, but it is of no use to anyone if no one knows it exists. E-MELD recommends that all linguistic resources be submitted to a linguistic search engine, such as the Open Language Archives Community (OLAC) repository. In addition to making materials discoverable, linguists should also make their materials accessible. In 2010, this generally means putting those materials online, and making them visible to as many people as possible. Not only should materials be discoverable and accessible, they should be stable; if a web address changes or a domain name expires, users will no longer be able to access that resource. Best practice suggests that a copy of language resources be placed in a stable online archive, and that a note of the archival copy's location be submitted to the search engine (e.g., OLAC).

Preservation is an important component of these best practices, since one of the primary goals of documentary linguists is to make a record of the world's languages for posterity. One aspect of preservation is tied to format durability: data should be accessible in ten or ninety years. Data migration—the conversion of data from an obsolete format to a more current one—is an excellent way to ensure long-term survival of a resource. If migration is not possible, materials should be stored in a widely known, widely accessible open format (E-MELD recommends XML for textual materials, for example). In addition to the threat posed by

[5] Unicode is a way for computers to represent and interpret written characters. It ensures that characters, particularly foreign or IPA characters, will look the same regardless of the machine displaying them.

obsolescence, literal preservation of materials is also top priority. Having a resource online is all well and good until a hurricane destroys the server where it is hosted. For this reason, it is wise to have multiple copies of materials stored in multiple physical locations; a very sensible mnemonic is LOCKSS (Lots of Copies Keep Stuff Safe)[6]. To quote Thomas Jefferson: “Let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.”

Finally, E-MELD touches on the sometimes sticky subject of who has rights to language resources, and the resources’ terms of use. The inconsistency of intellectual property laws worldwide, not to mention vastly differing cultural views of intellectual property, make this issue an enormous headache for the documentary linguist. Some speech communities feel that the public should not be able to see data about their languages; others have no objection to making that data public. Some speaker-consultants might become convinced that the linguist is profiting from their language, and some linguists might be territorial about their collected data. However, the complex problems of language data rights are beyond this paper’s reach (for a good general treatment of ethics and intellectual property in linguistic fieldwork, see Dwyer 2006). E-MELD’s best practice guidelines merely suggest that terms of use be well-documented, and that measures be taken to enforce those terms for a limited amount of time— access restrictions put in place indefinitely are unwise, since if the linguist dies without changing them, that information will be inaccessible forever.

[6] LOCKSS is also a data-preservation initiative hosted at Stanford; for more information see their website at http://lockss.stanford.edu/

With these guidelines in mind, we can take a look at the creation of ELDER and the basic motivations behind it.

2. ELDER: Origins

The idea for ELDER was first inspired by the difficulties of group work in a 2009 field methods class. A number of assignments were undertaken as group projects, but it soon became clear that working collaboratively might actually be more difficult than working alone. There was no easy way to decide on a standard method of transcription, or a set of descriptive terminology; we had trouble sharing files across operating systems and software packages; and even when a group of two or three students were able to consolidate their materials, no good method existed to share those materials with other student groups. Data in group work was largely confined to pen-and-paper notes and un-annotated sound files. At best, information would be typed up in Word or Excel and distributed via paper copies. When searching for a particular piece of information, e.g. the third-person plural nominative pronoun, one would have to thumb through stacks of paper hoping that *someone* had typed up a summary of the pronoun system, and that the transcriptions were accurate. By the end of the semester, a great deal of data had been elicited and analyzed; however, any given student only had easy access to a small portion of it. If only a tool existed, I thought, to help researchers catalogue their findings in a standard format, and to search through the data they had collected!

As it turns out, such a tool did exist—in fact, at least two of them. The first I encountered was a FileMaker Pro (FMP) database created by Cathy O'Connor and Amy Rose Deal. The FMP tool allowed users to create a lexicon for the target language, including transcriptions, glosses, and semantic information; similar entries could be made for longer utterances like sentences and texts. It had a number of other wonderful functions, such the ability to cross-reference texts and lexicon, add audio files to lexical entries, and do easy IPA input. However, it still presented problems for group work: namely, the fact that anyone hoping to use it needed a copy of FileMaker Pro. Students were able to use the FMP database if they downloaded a free trial version of the software (valid only for 30 days), or if they all agreed to work on a single computer in a particular computer lab. Here was a prime example of the troubles that come with using a proprietary format to work with language data: without shelling out $300 for the software, prospective users were greatly limited in how they could use this tool. Even though the FileMaker Pro database was head and shoulders above the notebook-and-word-processor format we had been using, it lacked certain features that I would have liked. As a child of the internet age, I am accustomed to the ability to share data easily and instantaneously; why was there no tool which would capture the extraordinary functionality of this FileMaker program, but free it from the bonds of proprietary formatting and offline work?

Continuing my search, I came across the Field Input Environment for Linguistic Data (FIELD)[7]. FIELD, developed under the E-MELD umbrella, is a web-based tool that provides a straightforward, intuitive user interface for inputting lexical data from any language. Like the FileMaker tool, FIELD has a number of excellent features that I was excited to use: it incorporates the GOLD ontology, it provides a tool for analyzing grammatical paradigms, it protects users' data from unauthorized use or editing, and it allows the user to add example sentences for each lexical item. I began inputting some of my data to FIELD, since it was the

[7] FIELD is accessible online at http://emeld.org/tools/fieldinput.cfm

best option available in an online utility for language data. However, also like the FileMaker tool, I found it to lack certain functions which I wanted in a language data utility. I missed the option to work with sentences and texts, as was possible in the FileMaker tool; the ability to add an example sentence was helpful, but not as helpful as being able to cross-reference those sentences with the lexical items in them. I wanted to be able to incorporate audio files into lexical and text entries, since no amount of transcription or analysis is an acceptable substitute for real primary data. And I wanted to add more data about a word than FIELD allowed, e.g. syllable structure and related words. Each of the tools I found had its own advantages and excellent functions, but neither fit all of my needs. So I did what any overambitious young academic would do: I decided to make my own.

2.1 ELDER: Ideals

When I set out to design a tool for linguistic data management, I was driven largely by a pragmatic desire to solve a problem I had encountered. However, as graduate students are wont to do, I also had a number of more ideological parameters in mind. These parameters are consistent with E-MELD's best practices, but in some cases go beyond them.

2.1.1 ELDER should be free and non-proprietary

As discussed above, proprietary software is not ideal for use in language documentation. Such formats limit who can use data, and how long it will be usable. Beyond best practices, though, I believed that language data should be freely available to anyone hoping to use it for scholarly purposes. I had in mind a wider audience: not only linguists, but the speech community as well. Dr. O'Connor's FMP database was first used to catalogue her data on Northern Pomo, a now-extinct Hokan language once spoken in northern California. I first used the FileMaker tool to work with data on KiNande, a Benue-Congo language spoken by roughly 900,000 people in the Democratic Republic of the Congo[8]. Though the two languages were vastly different in their robustness (one extinct and one quite vigorous), they both could have benefitted greatly from data being made available to their communities. In the case of Northern Pomo, an electronic lexicon and text compilation could aid revival efforts, or at least further description; in the case of KiNande, an electronic lexicon could be used to create native-language educational materials for schoolchildren, or to aid in the production of KiNande literature and text translations. However, providing said data to the communities would be rendered very difficult if proprietary software was needed. It was clear that asking speech communities to purchase copies of a third-party software in order to access data about their own languages could be impractical. I hoped to create a tool which would allow people worldwide to contribute, organize, and access language data free of cost.

2.1.2 ELDER should be widely accessible

Another major goal, in keeping with E-MELD's best practices, was that ELDER should be accessible by the maximum number of users, regardless of operating system, location, or nationality; I hoped to create a tool which could be accessed by phonologists in Boston, anthropologists in Germany, and speech communities in central Africa with equal ease. As I had determined that a web-based program was the best way to accomplish this, it was important to

[8] Ethnologue, "Ethnologue report for language code: nnb," Ethnologue, http://www.ethnologue.com/show_language.asp?code=nnb

ensure that the tool would function in a wide range of web browsers: in the US alone, there are at least five major consumer browsers, and many more on the international market. A tool that only functioned properly in, say, Microsoft Internet Explorer (a common failing of many websites in the past decade) would be of little use to much of the world. Finally, while most linguists at Boston University use Apple computers, the majority of the world does not; this tool needed to be functional across operating systems.

2.1.3 ELDER should benefit a diverse range of users

As discussed above, I set out to create a tool that would be useful not only to my fellow field methods students, but to speech communities and scholars working in a wide variety of fields. I did not set out to create a tool useful only for phonological analysis, or pedagogy, or semantic typology. My goal was to create a utility that could serve many needs through a single easy-to-use interface. Indeed, user diversity does not refer only to academic interests; I hoped to create a tool that would be useable by those without a strong background in linguistic software, or even computer use. Some extant documentation tools such as SIL's Toolbox[9] have a "learning curve:" users must be able to familiarize themselves with the program's syntax and consistently use that syntax before they can effectively use the program. For some people, especially those lacking confidence in their computer literacy, this might prove so daunting that they opt to just keep their data on paper and magnetic tape. I hoped to create a program with a highly user-friendly interface, simple enough to encourage even the most determined Luddite to try it. After

[9] Toolbox is a work environment for creating XML markup on textual materials; its users must understand XML markup formats and use that format without error in order for the material to be output properly. For more on Toolbox, see http://www.sil.org/computing/toolbox/

all, the more data is digitized and organized, the less we lose to “data graveyards”[10] in office closets and under desks.

2.1.4 ELDER should encourage academic accountability and preserve primary data

The inclusion, organization, and preservation of primary data like audio or video files was of paramount importance when I began to design ELDER. More than once, when working in a field methods class, I would attempt to build on information gathered by another student, only to find that I disagreed with (or couldn’t understand) their transcriptions or analysis. Since the only information that was available to me was the student’s transcription, I was unable to make my own judgments based on actual data, and would have to re-elicit the information all over again. This inefficient use of time could have been ameliorated by having access to clearly-labeled and organized sound files of the actual speech event. The extreme importance of primary data is not limited to circumventing undergraduate mistakes. Even among the most well-respected scholars the potential for faulty analysis is always present. Errors aside, a perfectly sound analysis by today’s standards might be invalidated by new paradigms in the future; if the language in question is extinct by that point, the description *must* not be the only resource available. The ideal documentation tool would one that encourages and facilitates a compilation of *primary* data, and which allows users to adapt that data to a variety of purposes.

2.1.5 ELDER should promote exchange and cooperation among scholars and speakers

[10] An excellent term borrowed from Himmelmann (2006)

At the time of writing, the author has little experience in the world of professional academia. What experience I do have has left me with the vivid impression that, like many scholars, linguists can be a bit territorial about their data and areas of study. This is understandable, certainly; the common measures of academic success (publications, grants awarded, career status at a major university) require some degree of competition. The tendency among some scholars is to "hoard" their data until they believe they have extracted the maximum benefit from it. However, especially in the case of endangered languages, there is such a dearth of time to document the language that it is essential to share what data exists. I hoped to create a tool that would promote the organized sharing of data, and would facilitate a free exchange of ideas and information between scholars and speech communities.

3. ELDER: Technology[11]

Once the core goals of ELDER were determined, the next step was to determine how this tool should be constructed: in what programming language should the database portion be built? What tools should be used to write the web pages with which users would interact? In keeping with best practice and the values discussed in 2.1.1, it made sense that even the tools used for programming ELDER should be open-source and as transparent as possible. The other major priority was accessibility: whatever programming choices we made should allow easy, quick access by the maximum number of users.

[11] At this point I must take a moment to acknowledge the extraordinary work done by Connor Shaughnessy, who is responsible for ELDER's programming and implementation. In this section I draw heavily upon his technical knowledge and familiarity with ELDER's inner workings.

3.1 Underlying database: MySQL

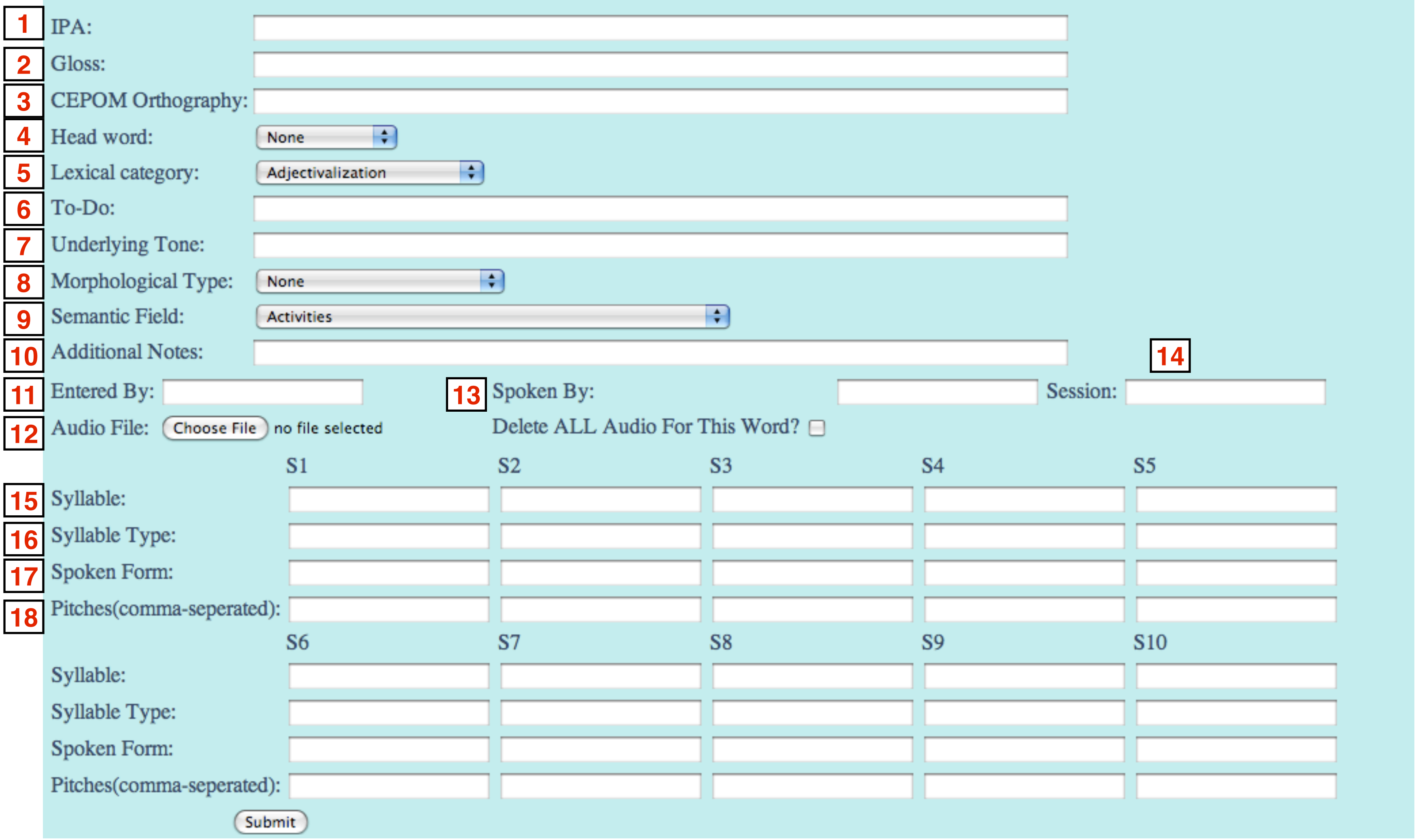
There are numerous softwares for the creation and sharing of databases, including the previously-discussed FileMaker Pro, as well as software like Microsoft Access and Oracle. None of these were suitable solutions for ELDER's purposes, as they are all proprietary products. We decided instead to use MySQL, a freeware application for developing relational database management systems. MySQL uses an implementation of SQL, which is the most widely-used language for modern database development. MySQL is free and open-source, which makes it ideal for a language documentation project; it interacts well with PHP, the language used to create ELDER's "front end" (the web pages users interact with); and it is commonly used and known by web developers worldwide, and is in little danger of becoming an unused, opaque format that data could become "trapped" in. Additionally, it has the capability to export all of its data into XML format, which is easily readable by a multitude of programs and is recommended in E-MELD's best practices[12]. Finally, MySQL was designed especially to be used on web servers and to provide information across networks, making it ideal for web-based tools.

3.2 Web development: PHP

The next question was how users should interact with the SQL database. MySQL is not a comprehensive tool for developing *and* interacting with databases, as is FileMaker Pro; it produces data that is easily accessible by other programs, but incomprehensible to the average person. Since the plan was for ELDER to be a web-based tool, and most of its user interactions to be through web pages, it was necessary to select an appropriate web development language

[12] E-MELD, "What are Best Practices?" E-MELD, http://emeld.org/school/what.html

that could interact with the SQL database. We settled on PHP, one of the most popular web development languages, for a number of reasons: PHP is free, and its source code is available to the public. It contains built-in modules for accessing MySQL databases, and is capable of generating web pages based on content retrieved from the database. One seemingly mundane aspect of a PHP-fronted SQL database is quite useful to ELDER: it does all of its processing on the server side, meaning that instead of users' computers having to search through all the data in the database, the server where ELDER resides performs those searches. This has three major advantages: first, server-side processing heightens security. Users' computers only ever see the *output* from the database, not the workings of the database itself. If anyone ever wanted to disrupt or attack ELDER (for whatever strange reason), it would be much harder to do so. Second, accessibility is enhanced: to use a website made with JavaScript, for example, the user must have a current version of Java installed and functioning on their computer. Some browsers are not Java-compatible, and this limits the range of people who would be able to use ELDER. PHP has no such requirements, and does not demand that the user have any particular software other than a functional web browser. Third, ELDER was intended to be usable by speech communities around the world, and the state of technology in these communities is likely to vary widely. Even those who do have internet access might be using very old computers with very little processing power; these older computers could take a very long time to search through the massive amounts of data ELDER would contain. With server-side processing, the burden of these searches fall instead on the server, which is much better equipped to handle such resource-intensive tasks; accordingly, ELDER will function as quickly and efficiently as possible, no matter the computer used to access it.



IPA:

Gloss:

CEPOM Orthography:

Head word: None

Lexical category: Adjectivalization

To-Do:

Underlying Tone:

Morphological Type: None

Semantic Field: Activities

Additional Notes:

Entered By:

Spoken By:

Session:

Audio File: no file selected

Delete ALL Audio For This Word? □

	S1	S2	S3	S4	S5
Syllable:					
Syllable Type:					
Spoken Form:					
Pitches(comma-seperated):					

	S6	S7	S8	S9	S10
Syllable:					
Syllable Type:					
Spoken Form:					
Pitches(comma-seperated):					

Figure 1: the "Add Word" page in ELDER v.0.5

4. ELDER: Functions

In the following sections, I hope to give a good overview of ELDER's functions, their purposes, and the rationale behind including them. ELDER was designed as a sort of "spiritual successor" to Dr. O'Connor's FileMaker tool, and as such, it integrates many of the features found in that program; it also draws upon ideas garnered from FIELD and from various writings on the practice of language documentation. Most of all, it was designed around my own needs as a documentary linguist and field methods student, and many of its features are the result of my finding something lacking while actually using ELDER. I intend to continue revising and adding features as they become necessary; the hallmark of a good tool is its ability to grow with its users.

4.1 Lexicon

ELDER's primary function as of April 2010 is as a lexical database. There is some debate in the academic community over whether lexicography is really the domain of documentary linguistics; for the purposes of this paper, however, I will assert my belief that the creation of a lexical database (and of traditional dictionaries) can be of great use to linguists and speech communities, and is a worthwhile pursuit by documentary linguists.

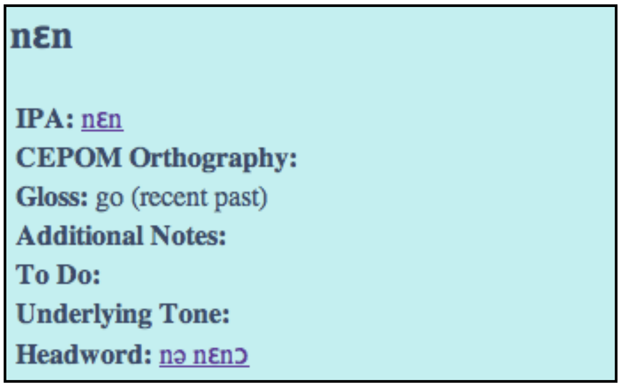
ELDER was designed to do more than a traditional dictionary. Instead of providing only a transcription and definition, it stores a great deal of information about a word and automatically finds all instances of its use within the database. My hope was that many kinds of linguists, as well as non-linguists, would be able to garner useful information from a lexicon stored in ELDER.

The first thing for ELDER to store about a lexical entry is its written form. In this case, the primary identifier for a word is the IPA representation of its phonemic form (see item 1 in figure 1). Obviously, it would make little sense to identify a word by its *phonetic* representation, as this would result in numerous entries being created for the same lexeme in different contexts. That is not to say that the word's phonetic form goes unrepresented, though— its phonetic form in isolation can be represented in the syllable breakdown (17), and a word's phonetic form in the context of a sentence can be represented in the entry for that sentence (see section 4.2). In addition to storing a word's IPA representation, ELDER also has the option to include transcription in the local or working orthography (3)[13]. Where a standardized writing system (or even a regulatory body) does exist for a subject language, it is in the best interest of a documentary linguist to make use of it. If a speech community is already accustomed to seeing its language written in a certain orthography, they will be far more likely to use a tool that makes use of that orthography; inclusion of the local orthography is also a nod of respect to those who have dedicated their time and effort to work on the language.

Like a traditional dictionary, ELDER stores a gloss for each lexeme; this is a fairly self-explanatory feature. It also makes use of a headword system (4). One of the earliest design problems in ELDER was how lexically-related words should be stored: should a noun's plural form and its singular form be given separate entries? Should an inflected form of a verb be considered an entity separate from the infinitive form of that verb? As many traditional dictionaries do, ELDER uses "headwords" (also known as lemmas or catchwords) to resolve this

[13] The current field reads "CEPOM Orthography" due to the fact that ELDER v.0.5 is being used for work on Medumba, a Grassfields language of Cameroon (discussed further in section 5). The Comité pour l'Etude et la Production des Oeuvres Bamiléké Medumba (CEPOM) is the organization responsible for developing and standardizing a Medumba orthography, as well as producing Medumba dictionaries, literature, and grammatical texts. When ELDER is made available for general use, this field will be changed to "Local orthography."

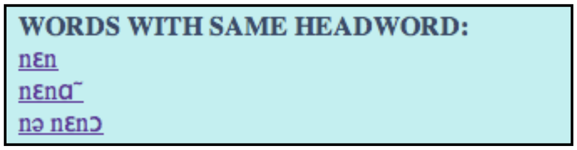
problem. Instead of creating unrelated entries for all forms of a word, or simply lumping all of a word's forms into a single entry, ELDER allows the user to create separate pages for different forms of a word, which are then linked together via a headword. For example, ELDER currently contains an entry for /nə nɛnɔ/, the infinitive form of the Medumba verb "go." It also contains the inflected forms /nɛn/, the recent past form of "go," and /nɛnɔ̃/, the non-recent past form. The latter two entries have /nə nɛnɔ/ as their headword. The entries for these words link directly to the headword, as well as showing the user all other entries with the same headword.



nɛn

IPA: nɛn
CEPOM Orthography:
Gloss: go (recent past)
Additional Notes:
To Do:
Underlying Tone:
Headword: nə nɛnɔ

Figure 2.1: Partial view of entry for /nɛn/, including headword link



WORDS WITH SAME HEADWORD:
nɛn
nɛnɑ̃
nə nɛnɔ

Figure 2.2: Headword portal on the entry page for /nɛn/.

The headword feature not only helps keep track of related entries, but allows the user to view lexemes in a more holistic way. In the context of ongoing research, such as a field methods class or a documentation project, there are bound to be knowledge gaps. Consequently, the researcher(s) may not have a good idea of how the productive morphology works— in other words, all of a word's forms may not yet be predictable from its "dictionary form." If a linguist or team of linguists know that Medumba /nə ʒɯ/ is “to eat”, and that /ʒɯ/ is the recent past form of /nə ʒɯ/, they may still have no idea what other forms /nə ʒɯ/ could take. As more data is added, and more forms of /nə ʒɯ/ are recorded, they will all be linked back to the same headword. By allowing users to view all currently attested forms of a word, ELDER aims to help users discover patterns in words and their morphology.

Headwords are not the only system for relating words to one another. As I began using ELDER to catalogue lexical data, I found that I often wished to express some non-headword relationship between entries. As a result, the “See Also” function was created. “See Also” is a way to create general linking between entries, regardless of the relationship being expressed. There are numerous ways words can relate to one another; when doing documentation work, a researcher may not possess a full understanding of how words are related, be it semantically, morphologically, or phonologically. However, it is valuable to be able to document one’s intuition that words are *somehow* related, and to direct other users’ attention to that fact. As a result, “See Also” is entirely open-ended, and can be used to link any entry to another. A good example of the implementation of “See Also” is in the case of /mɛnkəʙo/, meaning “devil, evil spirit.” This word appears to be divisible into other known morphemes: /mɛn/ “child” and /kəʙo/ “bad, evil,” which in turn appears to be divisible into /kə/ “negative marker” and /ʙo/ “good,” giving /mɛnkəʙo/ a literal meaning of something like “evil child” or “no-good child.” However, in absence of a more complete understanding of the Medumba morphological system, this interpretation is merely conjecture. The “See Also” feature allows such conjecture and suspicions to be noted, and possibly-related entries to be bundled together, without requiring that the user be entirely sure about the precise nature of the relationship— the page for /mɛnkəʙo/ thus links to /mɛn/, /kə/, /ʙo/, and /kəʙo/. In the future, we plan to add a field wherein a user can specify *why* she chose to associate one word with another, so that other users can better understand and work from those linguistic intuitions.

Deciding how to treat lexical categories in ELDER was a non-issue thanks to the GOLD ontology. When entering a word, users are presented with a drop-down menu (5) of all the parts of speech in GOLD. This format ensures that users adhere to GOLD standards when entering data—they have no other option. ELDER also asks for a word's morphological type (bound, free, null, etc.); while not part of GOLD, noting an entry's morphological type is relatively easy (8), and searching by morphological type could potentially be helpful for users. The third drop-down menu for categorizing lexical entries is a list of semantic fields (9). I admit to borrowing this feature from FIELD because I considered it extremely useful and innovative; having a semantically organized lexical database could be useful to language learners as well as to semantically-oriented research. GOLD does not include a semantic ontology of this type, and so the categories used in ELDER are the same as those in FIELD.[14] Users may assign as many semantic fields as they like to a word; this is obviously useful in the case of a word like "chicken (meat)," which could potentially be classified under "Food and Drink," "Animals: Birds," or "Animals: Domesticated Animals."

[14] My sincere thanks to Helen Aristar-Dry and Martha Ratliff for informing me as to the origin of these categories: they are drawn from Shintani and Yang (1990), *The Mun Language of Hainan Island: its classified lexicon.*

Choosing how to represent and store phonetic and phonological information about a word was one of the more complicated facets of designing ELDER. The syllable breakdown has a simple enough purpose: to allow users to examine syllable properties in the target language. "Syllable type" (16) is meant to hold a notation of a syllable's structure, as narrow or broad as the user chooses—currently, users have been instructed only to denote consonants (C), vowels (V), and nasals (N) in the hopes of keeping things consistent and simple. However, users can choose to make their notation as broad or narrow as they like; the "Syllable type" field will accept any alphabetic input in case a user wants to mark liquids (L) or fricatives (F). Similarly, the systems for marking tone are as unrestrictive as possible. The hope was for ELDER to be applicable to the documentation of any language, but its first incarnation was designed primarily for use with Medumba. Medumba has a highly complex tonal system that has eluded the understanding of two semesters of field methods classes, as well as a graduate student workgroup; it was of paramount importance that ELDER handle tone information in a way that made sense for this particular language. Since our understanding of the tone system was (and is still) not complete, it seemed prudent that ELDER store tone data without requiring a great deal of certainty about its theoretical or underlying properties. Most of ELDER's features were designed so that a user could enter as much or as little information as they possessed, since a lot of its anticipated use was in the context of ongoing documentation projects or field methods classes. The solution for handling tone was no different: provide users with the *option* to enter surface-level pitch, as well as the word's underlying tone pattern. Referring back to Figure 1, field 18 is for entering a word's pitch[15] as it manifests in isolation. The use here of a text box

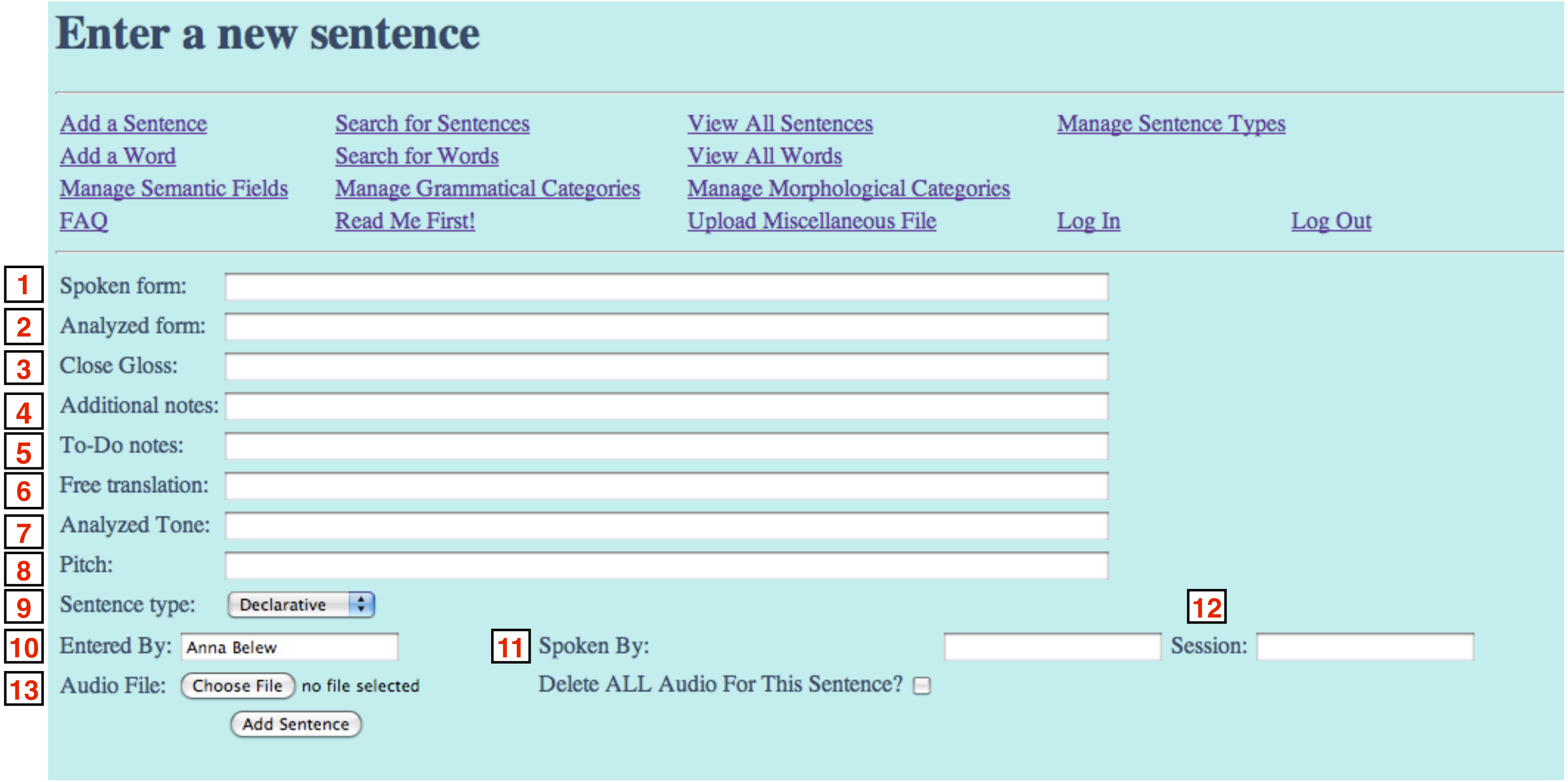
[15] Please forgive the somewhat humorous labeling "pitches;" the web designer is not a linguist, but is in possession of the site's code, and at the time of writing has not had time to correct this typo yet.

rather than a dropdown menu means that it is possible to use any scheme for annotating tone, whether a simple H/L scheme or numbered pitch levels (as we used for Medumba). For example, Medumba /mɛnfi/ “infant” has a high-to-mid falling pitch contour on the first syllable, and a level high pitch on the second. The notation system favored by our working group simply uses consecutive numbers to denote contour pitch, and so the first syllable would be recorded as 53 and the second as 5. This would not work for many other tone languages, though; ELDER will therefore accept almost anything as tone notation, meaning that users can mark pitch in whatever fashion is suitable for the target language (including not at all). Similarly, field (7), “Underlying Tone,” will accept any string of characters as underlying tone notation. While the current users are using a simple H/L (with downstepping marked by exclamation points), ELDER will let users describe analyzed tone in any way they see fit, depending on the descriptive or theoretical paradigm.

Fields (6) and (10), “To-do” and “Additional Notes,” are meant to enhance ELDER’s usefulness as a workspace for ongoing documentation projects. When working with newly-elicited material, I often have unresolved questions, or simply doubt my own interpretation of the data; I find it helpful to take note of these problems and single them out for later work. Having a field specifically for jotting down questions, uncertainties, or future plans helps users better plan out their time. It may also allow other users to see what work remains to be done, or to single out problems which still need addressing. “Additional notes” is just that— a field for including miscellaneous information, e.g. “consultant says only older people use this word,” or “this word refers only to a specific species of ant.” ELDER also stores certain pieces of metadata about a word, namely which speaker provided this word (field 13), during what elicitation session (14), and who entered the data (11). Specifying the speaker is useful in cases when one is working with more than one consultant; if consultants differ in their pronunciation, interpretation, or usage of a word, it may be important to note who said what. Similarly, knowing *who* entered a certain piece of information is crucial in group work or classwork circumstances— it also helps protect data and assign user permissions, which will be discussed further in section 4.3. Finally, the user can make note of the elicitation session where the word was collected. While this may not seem particularly useful in itself, it does something more than just make note of the session's date. Current literature on documentary linguistics encourages the preservation of as much information as possible, including the circumstances in which data was collected. This makes perfect sense: the context in which a word is elicited may affect a consultant's responses, or may contain information not otherwise noted (e.g. discourse-level phenomena). By this logic, I decided to keep and organize all of the raw (unedited) recordings from my elicitation sessions, and to provide them as part of ELDER's database. I devised a simple file-naming scheme for the session tapes[16] and put the filename in the "Session" field. The audio files themselves are then uploaded to ELDER as "miscellaneous files" (more on those in section 4.3), and can be viewed or downloaded by other users.

The final piece of information ELDER stores about a word entry is an audio file (12) of the word being spoken. Ideally, this recording portrays a consultant speaking the word in isolation, though some of the files currently in the database feature a consultant saying a word in the context of a sentence or phrase. ELDER can store files of any audio format for a word's

[16] Session files are named with the creator's name, a six-digit date, and a letter indicating which portion of that session the file contains. For example, if someone named Billy recorded a one-hour elicitation on March 12, 2009, and that session was broken into four 15-minute recordings to minimize file size, he would name them Billy031209a, Billy031209b, etc. While not elegant, this naming system is easy to remember and convey to new users.



Enter a new sentence

Add a Sentence | Search for Sentences | View All Sentences | Manage Sentence Types
Add a Word | Search for Words | View All Words
Manage Semantic Fields | Manage Grammatical Categories | Manage Morphological Categories
FAQ | Read Me First! | Upload Miscellaneous File | Log In | Log Out

Spoken form:
Analyzed form:
Close Gloss:
Additional notes:
To-Do notes:
Free translation:
Analyzed Tone:
Pitch:
Sentence type: Declarative
Entered By: Anna Belew   Spoken By: Session:
Audio File: Choose File no file selected   Delete ALL Audio For This Sentence? □
Add Sentence

Figure 3: "Add sentence" page of ELDER v.0.5

recording, although we anticipate MP3 and WAV will be the most commonly used, and encourage users to upload high-quality recordings in those formats. We also suggest that all audio files uploaded to ELDER have appropriate metadata added to them[17], in case the files are ever separated from their associated database entries: at a minimum, audio files should have metadata containing the creator's name, the consultant's name, the name of the subject language, and the date of recording. Once an audio file is uploaded for a word, other users will see it linked from the word's entry page, and can play the file in their browser or download a copy of it. Other databases currently in use lack the ability to share audio on the web. The ability to easily associate and upload an audio file for words and sentences is one of ELDER's hallmark features, and will hopefully increase organization and sharing of primary audio data among its users.

Once data has been added for a lexical item, other users can view it (but not edit it); the lexical item also becomes part of the "word bank" ELDER searches through when attempting to auto-gloss a sentence, as discussed in the following section.

4.2 Sentences[18]

Documentation of a language entails more than just documentation of a lexicon; an ideal documentation project would contain every conceivable type of communicative event. While not quite able to do *that*, ELDER can currently work with sentences and phrases in addition to lexical items. The primary aim of the sentence functions is to relate the lexicon to actual

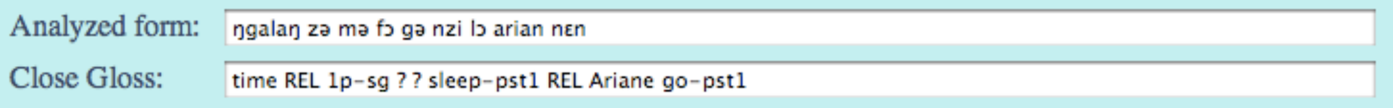
[17] Currently, there is no easy way to add metadata to WAV files; it is possible only through specialized (and rather arcane) software suites. As a result, we are currently encouraging users to contribute high-quality MP3 files with appended metadata if they are able to do so.

[18] For the purposes of this section, I will use the word "sentence" to refer to all sentences, phrases, and clauses stored in ELDER's "sentence" function.

instances of use, but also to preserve and catalogue said instances. Figure 3 above shows the "add sentence" page. As with individual words, sentences can be represented both phonemically and phonetically. Field (1), "spoken form," stores a narrow phonetic transcription of the sentence or phrase, including contractions, repetitions, etc. Field (2), "analyzed form," stores (the user's interpretation of) what lexical items compose the sentence. For example, the Medumba relativizer /zə/ frequently occurs directly before the third-person singular pronoun /a/, as in the sentence /a kɯ zə a sa lɔʔ ɔ/ ("What destroyed the village?"). In normal speech, this sentence would be pronounced /a kɯ **za** sa lɔʔ ɔ/, with /zə/ and /a/ being contracted into /za/. ELDER thus has the former sentence in the "analyzed form" field, and the latter in the "spoken form" field.

A sentence's close gloss, or morphogloss, is stored in field (3). Most language data is presented in a multi-line format following the Leipzig Glossing Rules;[19] ELDER uses roughly the same format (differences are described below). After entering the analyzed form of a sentence, users can enter a close gloss, which will be displayed in a standard interlinear gloss format. In order to correctly display word-by-word alignment (Glossing Rule no. 1), ELDER "reads" the sentence in Analyzed Form, then creates a table which has as many columns as there are words in the sentence (the program interprets spaces as word breaks). The close gloss is then placed in the table as a row below the analyzed form, so that the sentence is aligned to its gloss. For example, the input in figure 4.1 ("While I was sleeping, Ariane went") would be displayed by ELDER as seen in figure 4.2.

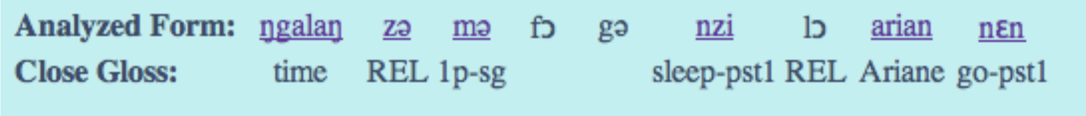
[19] Max Planck Institute for Evolutionary Anthropology Department of Linguistics, "Leipzig Glossing Rules," Max Planck Institute for Evolutionary Anthropology, http://www.eva.mpg.de/lingua/resources/glossing-rules.php



Analyzed form: ŋgalaŋ zə mə fɔ gə nzi lɔ arian nɛn

Close Gloss: time REL 1p-sg ? ? sleep-pst1 REL Ariane go-pst1

Figure 4.1: Inputting a sentence's close gloss

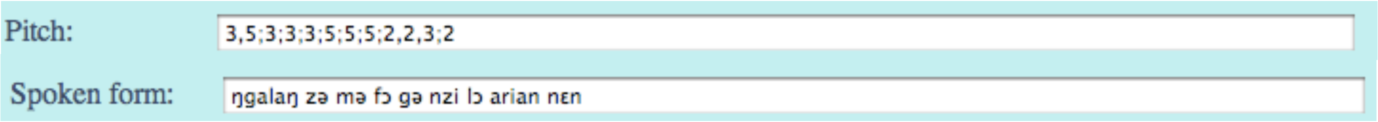


Analyzed Form:	ŋgalaŋ	zə	mə	fɔ	gə	nzi	lɔ	arian	nɛn
Close Gloss:	time	REL	1p-sg			sleep-pst1	REL	Ariane	go-pst1

Figure 4.2: Viewing a sentence's close gloss

Note that certain items above are glossed with question marks; as is often the case during documentation work, some words are still unknowns. In these cases, the user can simply put a question mark in lieu of a gloss, and ELDER will show a blank space under that word.

ELDER also provides word-by-word alignment for pitch— a particularly useful feature for languages with complex tone systems not easily represented by diacritics. The current implementation of ELDER for Medumba uses numeric pitch transcription, as discussed in section 4.1. We wanted to make it easy to enter pitch for a sentence which would then display in correct alignment with the text, so we gave pitch a row in a table (like the table that aligns the close gloss with the sentence). Instead of being aligned with the "analyzed form," pitch is aligned with a sentence's "spoken form." We needed to create a system for users to tell ELDER how pitch should be aligned. The simplest way was just to assign a character to mean "word break" (in our case, a semicolon) and another character to mean "syllable break" (a comma). When entering pitch for a sentence, the user writes out the numeric pitch in field (8) with these symbols inserted at the appropriate places: see Figure 5.1 for an example.

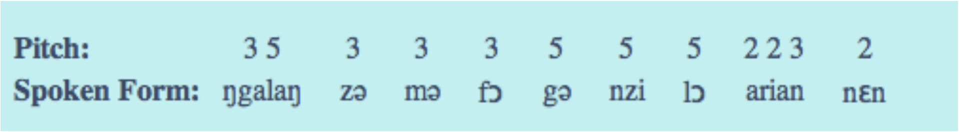


Pitch: 3,5;3;3;3;5;5;5;2,2,3;2

Spoken form: ŋgalaŋ zə mə fɔ gə nzi lɔ arian nɛn

Figure 5.1: Entering sentence pitch

ELDER then reads the pitch input and, upon displaying the sentence, spaces the pitch markings with the words:



Pitch:	3 5	3	3	3	5	5	5	2 2 3	2
Spoken Form:	ŋgalaŋ	zə	mə	fɔ	gə	nzi	lɔ	arian	nɛn

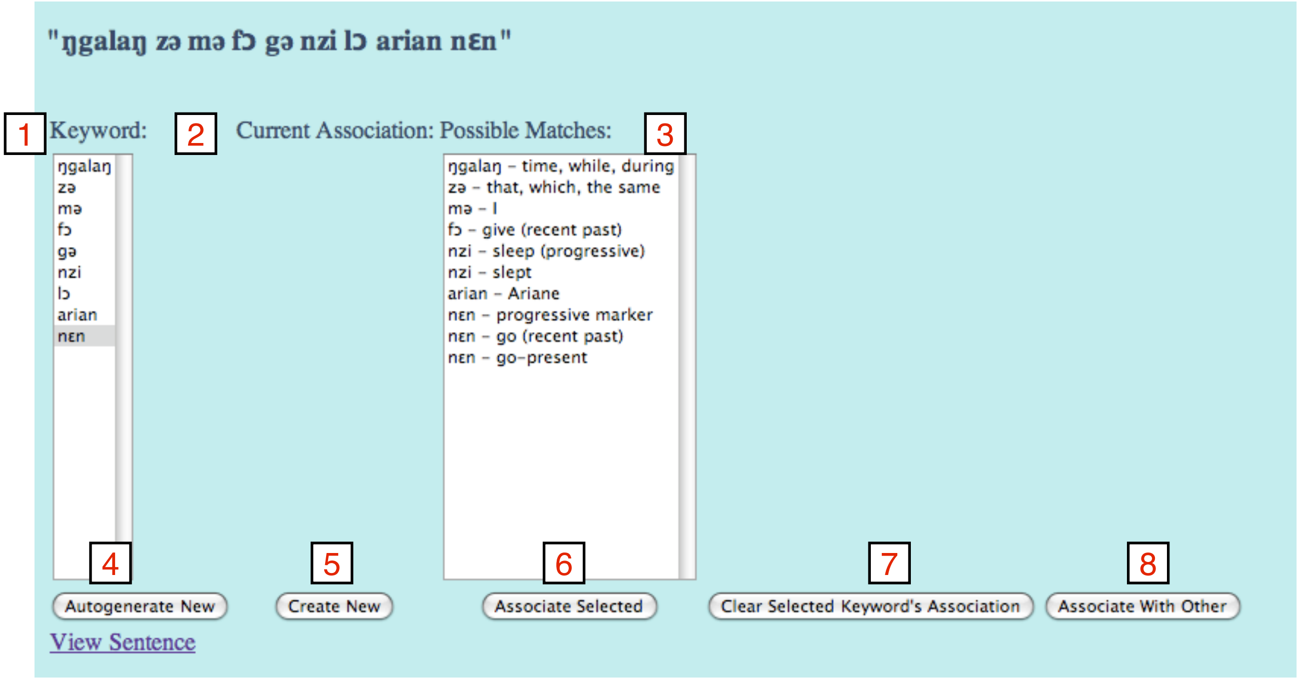
Figure 5.2: Viewing sentence pitch

"Analyzed tone" (7) serves the same purpose for sentences as for words; it is intended to let users record what they think are the underlying tones (within whatever tonological paradigm they choose to use). This is particularly useful for Medumba, which possesses a number of "null" tonemes— morphemes which only manifest in an utterance's tone. These are still poorly understood by our working group, but guesses about them can be recorded in this field. "Entered by"[20] (10), "Spoken by" (11), "Session" (12), and audio files (13) function the same for sentences as for words, as discussed above.

One of ELDER's major features comes into play after a user enters a sentence: the automatic cross-referencing of sentences and lexical items. Once a user clicks the "add

[20] Note that my name is already in the "entered by" field—when a user is logged in, ELDER automatically puts their name into this field. More on user accounts and logins in section 4.3.

sentence” button, a screen appears wherein she can select which lexical items to associate with the sentence, as seen in Figure 6:

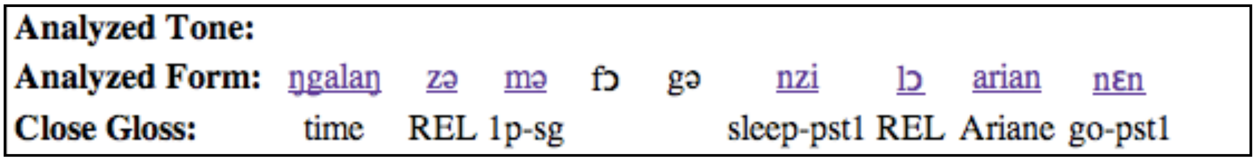

"ŋgalaŋ zə mə fɔ gə nzi lɔ arian nɛn"

ŋgalaŋ
zə
mə
fɔ
gə
nzi
lɔ
arian
nɛn

ŋgalaŋ – time, while, during
zə – that, which, the same
mə – I
fɔ – give (recent past)
nzi – sleep (progressive)
nzi – slept
arian – Ariane
nɛn – progressive marker
nɛn – go (recent past)
nɛn – go-present

Figure 6: Interface for associating sentence keywords with lexical items

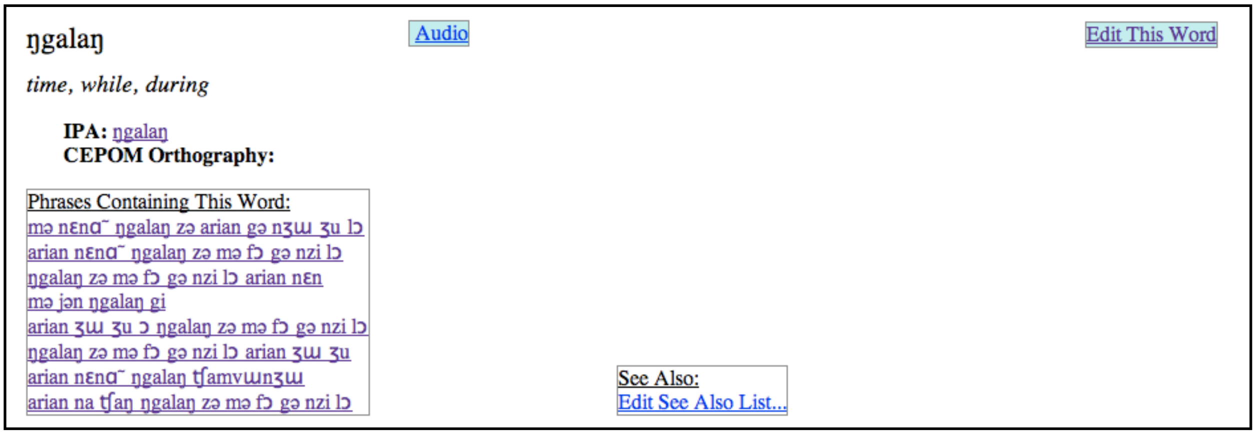
Field (1), “keywords,” shows all of the words in the sentence (meaning everything separated by a space, as ELDER considers spaces to be word breaks). To link words to their lexical entries, the user has several options. If a word is already part of the lexicon, it will show up in the “possible matches” box (3); the user simply selects the word in the keyword list, selects the appropriate word from the match list, and clicks “associate selected” (6). The association will then show up in the “current associations” column (2). If a word is already in the lexicon, but is not showing up in the possible matches field (this may happen if the word is transcribed inconsistently, for example), the user can manually link a word to any other word in the database by clicking “associate with other” (8). This will open a new window allowing the user to search the database, select a word, and associate it with the keyword in the sentence. If a keyword is not yet in the database, there are three options for handling it. “Autogenerate new” (4) will create a new, blank entry for the highlighted keyword, but will not show the user a screen for entering information about it. ELDER specially tags these autogenerated pages; users can browse all the entries created this way to keep track of which words need further work. This feature is ideal when a user is not sure about a word’s meaning, but would like to create a lexicon entry anyway. Having all instances of a “mystery word” linked from its entry page could help contextualize its function or meaning. If a word is not in the database yet, but a user *does* know what it is, she can use the “create new” button (5). This will bring up a new “add a word” window, with the word already entered into the “IPA transcription” field. The user can enter as much information as she has, submit the word, and close the window— this will take her back to the word association page, where the newly created entry will be listed under “current associations.” Once a user has created all the associations she cares to, she can click the “view sentence” link to go to the sentence’s finished display page, where the associations will be displayed as hyperlinks:



Analyzed Tone:									
Analyzed Form:	ŋgalaŋ	zə	mə	fɔ	gə	nzi	lɔ	arian	nɛn
Close Gloss:	time	REL	1p-sg			sleep-pst1	REL	Ariane	go-pst1

Figure 7: Partial view of sentence display page with hyperlinked keywords

Clicking any hyperlinked word will take the user to that word’s entry. The association goes both ways: a word’s entry page contains a linked list of sentences containing that word, as dictated by what sentences the word has been associated with.



ŋgalaŋ

time, while, during

IPA: ŋgalaŋ
CEPOM Orthography:

Phrases Containing This Word:
mə nɛnɑ˜ ŋgalaŋ zə arian gə nʒɯ ʒu lɔ
arian nɛnɑ˜ ŋgalaŋ zə mə fɔ gə nzi lɔ
ŋgalaŋ zə mə fɔ gə nzi lɔ arian nɛn
mə jən ŋgalaŋ gi
arian ʒɯ ʒu ɔ ŋgalaŋ zə mə fɔ gə nzi lɔ
ŋgalaŋ zə mə fɔ gə nzi lɔ arian ʒɯ ʒu
arian nɛnɑ˜ ŋgalaŋ ʧamvɯnʒɯ
arian na ʧaŋ ŋgalaŋ zə mə fɔ gə nzi lɔ

Figure 8: Partial view of lexical entry page with hyperlinked list of sentences

4.3 Other functions

ELDER currently has two features other than its lexicon and sentence database. First is the ability for users to upload “miscellaneous” files, i.e. any type of file in any format. The purpose of this feature is to allow users to share papers, spreadsheets, videos, etc. about the subject language, no matter what format they might be in. Anyone can upload files, and anyone can download them. ELDER also has a user account system, which is particularly critical in a database open to the general public; no one would want to store their data in a place where it could be altered or deleted indiscriminately, or attributed to the wrong person. Protecting data privileges through a user login system allows contributors to take credit (and responsibility) for their work. The user system in ELDER v.0.5 is slightly primitive, but functional: when any word or sentence is added to the database with a name in the “Entered by” box, ELDER makes a record of that name and adds it to a list of potential users. Once her name is added to that list, a user can create a password for their account.

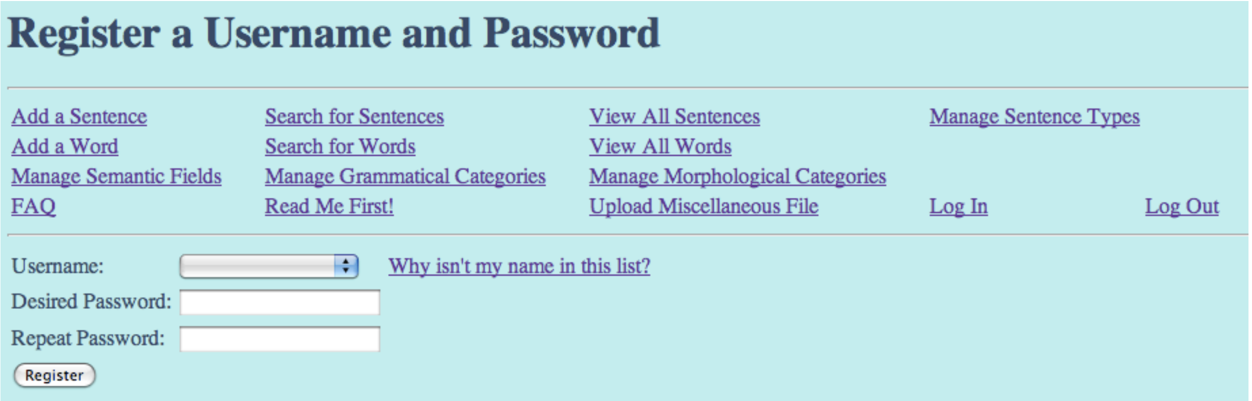


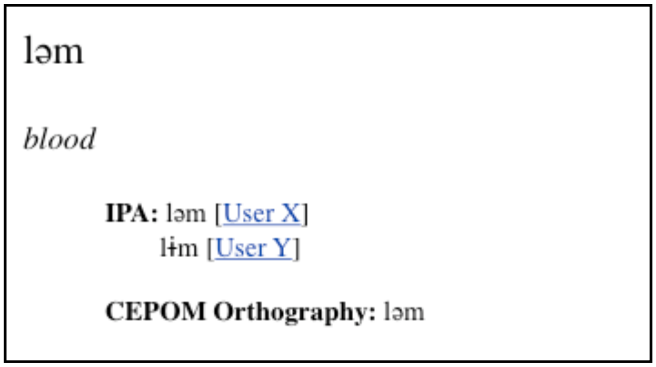
Figure 9: User registration page

After a password is chosen, the user can log in. When logged-in users add a word or sentence, ELDER automatically puts their name in the "entered by" field. Any word or sentence submitted with a name in "entered by" can only be edited by that user; conversely, if a user adds an entry with a blank "entered by" field, anyone can edit or delete that entry. We plan to improve the user account function, among other things, before releasing ELDER for widespread use.

4.4 Future functions

Rome was not built in a day, and ELDER is by no means completed yet. There are a good many features that we would still like to add or improve. One small upgrade will be the addition of a multi-language glossing option (for example, Medumba entries could be glossed in both English and French) for languages in areas with more than one lingua franca in use. The ability to add GOLD standard morphosyntactic and morphosemantic features will be implemented fairly soon as well. On the technical side, ELDER needs to be moved to a hosting service (at the moment it is stored on a personal computer running server software). Once ELDER is properly hosted, we plan to improve the user account system by making logins more secure, registration easier, and permissions assignable by the user. This latter means that we would give an individual user the ability to dictate the permissions on her data: if a linguist is working with a team, she may want to allow other team members to edit her data. The new system would let users decide who can access or modify their entries. We also plan to add a user profile feature; users will be able to provide their contact information (email address, website URL, etc.), and will have the option to make that contact information visible to other users.

The improved user system will be central to ELDER's largest planned overhaul: the implementation of a wiki-style editing system. Unlike Wikipedia, ELDER will not allow users to modify other users' data; however, they will be able to add to it. Imagine that User X adds the Medumba word "blood" as /ləm/. User Y disagrees, believing the word should be transcribed / lɨm/. The proposed wiki feature will allow User Y to append his own transcription to the word's entry, and will display it as an alternate opinion. The display page for /ləm/ would then look something like this:



ləm

blood

IPA: ləm [User X]
lɨm [User Y]

CEPOM Orthography: ləm

Figure 10: Proposed word display with multi-user contributions

Note that the transcription in the header is the one provided by the entry's creator; ELDER would give precedence to the original data provider while making note of other contributions. Search functions would search through *all* contributed information: if User Z is looking for "blood" after the above exchange between User X and User Y, ELDER will direct Z to the correct page whether she searches for /ləm/ or /lɨm/. Note also that the displayed user names would be linked to those users' profiles. If a differing piece of information were added to an entry, ELDER would notify the entry's original creator; users could also contact each other via the contact information in their profiles (if they chose to make that information public). Contributors could choose to accept these proposed revisions, or not; the goal is simply to allow every user's opinion to be heard, and to foster communication. In that vein, we also plan to add a bulletin board system for ELDER users to discuss pertinent topics.

The last major upgrade currently planned for ELDER is to implement multi-language capability. At the moment, ELDER is only storing data for Medumba; once it is hosted on the web and available to the public, we will add the ability to work with multiple languages (done by teaching ELDER to make blank copies of itself each time work on a new language begins). Our goal is to continue improving ELDER as we learn more about the needs of its users. We will be soliciting feedback and suggestions on an ongoing basis, and look forward to making this an even better tool.

5. Implementation

In addition to creating ELDER, I have also used it to compile and annotate a small Medumba lexicon. At the time of writing, I have elicited and input roughly 300 lexical items and

200 sentences; students in Dr. O’Connor’s field methods class have contributed some 40 words and 35 sentences. Below is an overview of how ELDER was first put to use, including problems encountered and their solutions. Note that I aim here not to provide a description of Medumba itself, but simply to describe how the lexicon was compiled and entered into ELDER.

5.1 Single-user implementation

Since the spring of 2009, I have had the good fortune to work with Ariane Ngabeu, a Boston University Ph.D. student from Ndé division, Cameroon. She has served as a Medumba consultant for two semesters of Dr. O’Connor’s field methods classes, and has given a great deal of her time to helping me compile the Medumba database. Her experience as a French teacher and her excellent linguistic intuitions make her an ideal consultant, and I am grateful to have been able to work with her.

Before ELDER was in working order, I began collecting lexical items for the database. The task of creating a lexicon for an entire language is daunting, to say the least. I decided that a good starting point would be a standard 200-item Swadesh list. I made recordings of Ms. Ngabeu saying each of the words in isolation, and in some cases, elicited sentences for context. After the Swadesh list was completed, I needed another system for deciding what words to collect. Since I hoped to include CEPOM’s orthographic representation of as many words as possible, I next began compiling an elicitation list from CEPOM’s 1991 Medumba-French dictionary, *Nə̀t*ɑ̀ *Mə̀dʉmb*ɑ̀. One of the challenges of working with Medumba is that many of the publications dealing with the language are in French; luckily, my own knowledge of French is sufficient to read papers and use it as a metalanguage for basic vocabulary elicitation (though most of my elicitations with Ms. Ngabeu took place in English). When eliciting from the CEPOM dictionary, I provided the word’s English gloss *and* its French gloss to avoid any semantic “cross-contamination;” I hoped to avoid eliciting the wrong word due to a semantic difference between the English and French words. Interestingly, despite the care taken to provide CEPOM’s original French gloss, Ms. Ngabeu sometimes provided different words than those in the dictionary; for example, the dictionary lists “sourd” (“deaf”) as “mbu’ntoŋ,”[21] while Ms. Ngabeu provided the word /kəʒuʔ/ (literally “not hear”). In some cases, she did not recognize the word provided by the dictionary, while in others she recognized the word but admitted she had forgotten it. Whether this difference in lexical knowledge is indicative of language shift over time (the dictionary was written nearly two decades ago) or simple language attrition from disuse is unclear, but could certainly be grounds for an interesting investigation.

After eliciting the lexical items, I began transcribing, glossing, and editing the sound files for them. As discussed in section 4.1, I retained the unedited recordings to add to the database; I used the audio editing program Audacity to chop up these long files into recordings of individual words or phrases. In addition to the information I had elicited specifically for ELDER, I mined data from my elicitation recordings made during the 2009 field methods class; this allowed me to include the maximum amount of information while taking up a minimum of Ms. Ngabeu’s time. In the early stages of work, I compiled all this information in a Microsoft Excel spreadsheet since ELDER was not yet ready for use (working with language data in a spreadsheet, and then in ELDER, confirmed my suspicions that the latter is much more efficient than the former!).

[21] CEPOM orthography follows mostly French orthographic conventions, with IPA characters and diacritics added; for the purposes of the CEPOM data presented in this paper, I will specify that apostrophes represent glottal stops, grave accents mark underlying low tone, and acute accents mark an underlying high tone.

Once ELDER was functional enough, I moved all of the data into it from the Excel spreadsheet[22] and began using ELDER in earnest.

My own experiences using ELDER sparked the creation of some of its features. For example, the “See Also” feature originated when I wanted to note a possible connection between the verb /nə fɯ/, “to cheat,” and the word /nfɯmjak/, “blind,” which the consultant had described as meaning “fake eyes, pretend eyes.” I had a hunch that they might be morphologically related, and wanted to note it. It would be quite cumbersome to put the entire URL of the /nə fɯ/ entry in the “additional notes” section, and I did not want to instruct viewers to go through the trouble of doing a search for /nə fɯ/—and thus “See Also” was born.

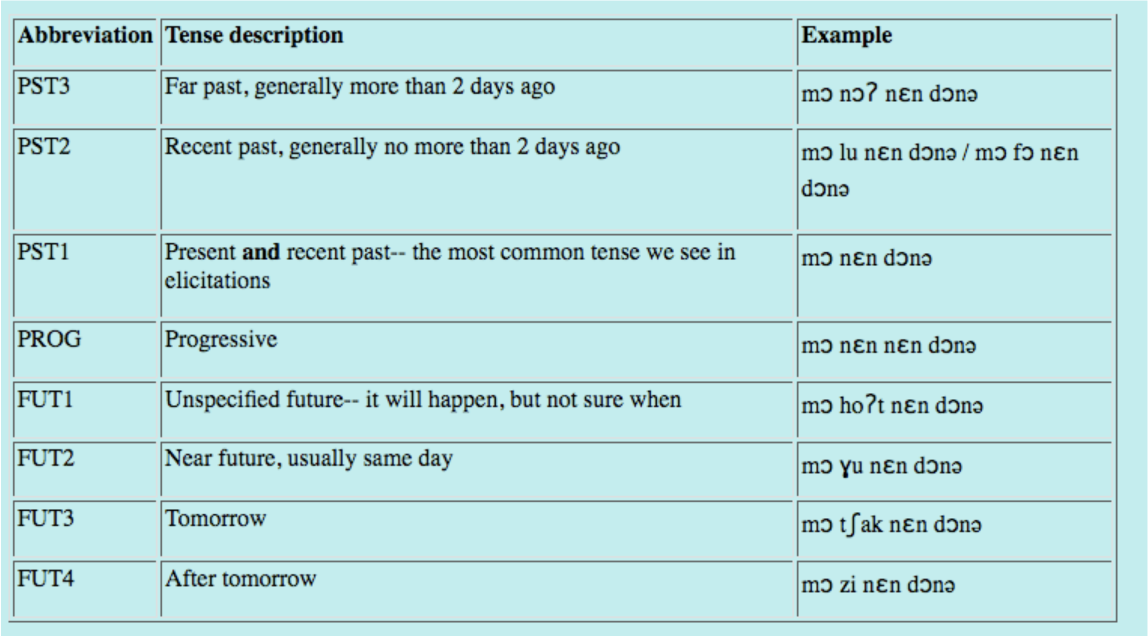
5.2 Multi-user implementation

By the end of March 2010, ELDER was functioning well enough that I felt comfortable making it available to Dr. O’Connor’s field methods class. This was the first true test of ELDER’s usability: it had been designed partly for use in field methods classes, and now was the time to prove its mettle. I created a slideshow explaining ELDER’s purpose and functions, and presented it to the class along with a brief demonstration of the tool. Students were encouraged to enter data into ELDER for extra credit—an arrangement beneficial for them and for me, since they would act as my first round of proverbial guinea pigs.

Before students started to use ELDER, I realized I would need to make some type of instructional document available to them; it was hardly fair to expect them to remember

[22] While it is possible to automatically import data into ELDER from an XML file, I chose to do it manually; I wanted to test the working environment as much as possible and make sure things functioned properly with a wide variety of user input.

everything I said during the demonstration. I began compiling an FAQ to address what I expected to be the most common questions about ELDER. First, I knew that there would need to be a guide to properly formatting audio files. The FAQ instructs users on how to format audio files (MP3 or WAV with a minimum sampling rate of 44.1 kHz), how to add metadata to audio files (using popular media players like iTunes or Windows Media Player, for example), and how to name session files. Also necessary was a guide to inputting sentence-level pitch (recall from section 4.2 the somewhat forbidding semicolons-and-commas format) and a reminder only to mark nasals (N), consonants (C), and vowels (V) when describing syllable structure. Beyond these simpler technical guidelines lay a larger problem: there was no standardized phonemic inventory for the class to use. Our field methods classes had not yet agreed on a basic phonological outline of Medumba, and none had been published to our knowledge; while we had established a few conventions about certain vowels, there was no firm consensus about most segments. For example, some students would transcribe “blood” as /ləm/, others as /lɨm/, and yet others as /lɯm/; one student might write “dog” as /mʙɯ/, while another wrote /mbʉ/. This had the potential to cause significant problems. Most of ELDER’s cross-referencing ability depends on a word’s phonemic form being transcribed consistently across instances. If different students were writing words in different ways, none of their data would end up being linked to other data, and that would be entirely counter to ELDER’s purpose! In the hopes of avoiding these inconsistencies, I consulted with the class and we settled on a phonemic inventory to be used for the purposes of ELDER data; vowel and consonant charts were included in the FAQ for easy reference. Another potential consistency problem was the abbreviations used in the close glosses. The Leipzig Glossing Rules contain a list of standardized grammatical abbreviations for close glosses, and these are included in the FAQ. However, the Leipzig abbreviations only have a single abbreviation each for future tense and past tense; Medumba has four future tenses and three past tenses, based on the distance from the present to the event. To address this, I added a system for numbering the tenses, seen in Figure 11:



Abbreviation	Tense description	Example
PST3	Far past, generally more than 2 days ago	mɔ nɔʔ nɛn dɔnə
PST2	Recent past, generally no more than 2 days ago	mɔ lu nɛn dɔnə / mɔ fɔ nɛn dɔnə
PST1	Present **and** recent past-- the most common tense we see in elicitations	mɔ nɛn dɔnə
PROG	Progressive	mɔ nɛn nɛn dɔnə
FUT1	Unspecified future-- it will happen, but not sure when	mɔ hoʔt nɛn dɔnə
FUT2	Near future, usually same day	mɔ ɣu nɛn dɔnə
FUT3	Tomorrow	mɔ tʃak nɛn dɔnə
FUT4	After tomorrow	mɔ zi nɛn dɔnə

Figure 11: Guide to glossing Medumba tense

With the FAQ in place, and guidelines for standardization provided, I could only hope that the students would find ELDER as productive and easy-to-use as it was designed to be.

ELDER's trial run met with moderate success, to my relief. Over half of the students in the class opted to enter data into ELDER; a few minor errors were prevalent, but the data was largely good. One common mistake was that many students confused "pitch" with "analyzed tone," and input a numeric pitch representation in the analyzed tone field both for lexical items and for sentences. They also tended to put a description of their elicitation topic (e.g., "Verbs of Cognition") in the "Session" field, rather than the file name of the session recording (and accordingly did not upload their session recordings). Another (fairly rare) problem was quite the opposite of what I had expected: instead of using varied and inconsistent abbreviations when glossing their sentences, a few of them seemed to have no idea how to do a close gloss. For example, one student entered a sentence with the free translation “She believes the dog is fat,” and the close gloss “She accepts that the dog is fat,” with no visible correspondence between the Medumba sentence and the English “close gloss.” The only particularly serious mistake they made with any frequency was in re-adding words that were already in the database, using different transcriptions. Nearly a dozen entries were redundant, and some of them were even transcribed exactly the same as the existing entry. The FAQ instructs users (quite emphatically) to search for a word’s gloss before inputting it to minimize duplicates; whether their failure to do this was due to insufficient instruction on my part or simple obstinacy on theirs, I may never know. Despite these errors, I was encouraged by the students’ use of ELDER. They managed to input largely well-formatted data with minimal instruction in the use of the tool; the mistakes they did make could easily be countered by providing a more intensive training session before they began to use ELDER, and by ensuring that they have a good grip on the basics of language documentation (namely glossing procedures). I plan also to create an online walkthrough demonstrating ELDER’s functions, pointing out common mistakes, and reminding users of good practices; I hope that with clear, accessible instructions, users will be able to make the best possible use of ELDER as a resource.

At the time of writing, there are three professors of linguistics considering using ELDER as a tool in their next field methods courses. This is a thrilling prospect, and I can only hope that the number will continue to climb when and if people find ELDER a useful tool.

6. Conclusion

With languages dying at unprecedentedly rapid rates, the task of documentary linguists is more urgent than ever. There is too little time, and too few people, to fully document every language likely to die in the next century. It is imperative that those people willing to do this work have the best possible tools available to them. I hope ELDER will serve to help documentation efforts make the most out of the time and resources available.

The creation of ELDER has truly been a labor of love. Every line of code and every word entered in it represents a single hope: that through dedicated, conscientious, collaborative effort, we will be able to preserve some fragment of the languages being lost every day. Should so many languages slip unrecorded into oblivion, the sum of human knowledge will be bitterly impoverished; I hope that future generations will have at least glimpses of the rich linguistic diversity once extant on Earth.

Bibliography

Duranti, Alessandro. *Linguistic Anthropology*. New York: Cambridge University Press, 1997.

E-MELD. “Homepage.” E-MELD. http://emeld.org/index.cfm (accessed March 19, 2010).

E-MELD. “What are Best Practices?” E-MELD. http://emeld.org/school/what.html (accessed April 16, 2010).

Ethnologue. “Ethnologue report for language code: nnb,” Ethnologue. http://www.ethnologue.com/show_language.asp?code=nnb (accessed April 13, 2010).

Evans, Nicholas. *Dying Words: Endangered Languages and What They Have to Tell Us*. Oxford: Wiley-Blackwell, 2010.

General Ontology for Linguistic Description Community. “About GOLD.” General Ontology for Linguistic Description. http://linguistics-ontology.org/info/about (accessed March 18, 2010).

Gippert, Jost, Nikolaus Himmelmann, and Ulrike Mosel, eds. *Essentials of Language Documentation*. Berlin: Walter de Gruyter, 2006.

Hartmann, R. R. K., ed. *Lexicography: Principles and Practices*. New York: Academic Press, 1983.

Himmelmann, Nikolaus. "Documentary and descriptive linguistics." *Linguistics* 36 (1998): 161-195.

Krauss, Michael. “The World’s Languages in Crisis.” *Language* Vol. 68, No. 1 (1992): 4-10.

Linguistic Society of America Committee on Endangered Languages and their Preservation. “Mission Statement.” Linguistic Society of America. http://www.lsadc.org/info/lsa-comm-endanger-more.cfm (accessed April 7, 2010).

Max Planck Institute for Evolutionary Anthropology Department of Linguistics. “Leipzig Glossing Rules.” Max Planck Institute for Evolutionary Anthropology. http://www.eva.mpg.de/lingua/resources/glossing-rules.php (accessed April 23, 2010).

Tchakoute, Paul Tɑnyi Njanja, Michel Tchayi, Pierre-Mopelt de Bafetbah Mbetbo, Francois Nkwilang, and Njamen Ngatchou, eds. *Nə̀tà’ Mèdʉmb*ɑ̀. Yaoundé: CEPOM, 1991.

UNESCO Culture Sector. “Safeguarding Endangered Languages.” UNESCO. http://www.unesco.org/culture/en/endangeredlanguages (accessed March 10, 2010)

United Nations. “Declaration on the Rights of Indigenous Peoples.” United Nations. http://www.un.org/esa/socdev/unpfii/en/drip.html (accessed March 29, 2010).

Woodbury, Anthony C. “Defining documentary linguistics.” *Papers in Language Documentation and Description,* Volume 1 (2004).