

# NEH Application Cover Sheet

## Digital Humanities Start-up Grants

### PROJECT DIRECTOR

---

Min Chen  
Associate Professor  
32 Campus Drive SS413  
Missoula, MT 59812-4104  
UNITED STATES

**E-mail:** min.chen@umontana.edu  
**Phone(W):** 406-243-2886  
**Phone(H):**  
**Fax:**

**Field of Expertise:** Technical: Computer Science

### INSTITUTION

---

University of Montana  
Missoula, MT UNITED STATES

### APPLICATION INFORMATION

---

**Title:** *Multimedia Search Prototype for Endangered Language Documentation and Analysis*

**Grant Period:** From 5/2014 to 10/2015

**Field of Project:** Social Science: Linguistic Anthropology

**Description of Project:** This interdisciplinary project seeks a Level II Start-Up grant to research, implement and evaluate an advanced multimedia search prototype for language documentation, analysis, and application. This project contributes to documentary linguistics especially in endangered languages by 1) providing an extended text-based search on metadata (transcription, translation, and annotation) to automatically find structurally and semantically related words and 2) providing intelligent processing and retrieval on audio files to efficiently and directly locate targeted audio segments. It will be publicly available and widely applicable via web browsers. Each major component will be made language/platform independent so users have the flexibility to use one or more search functions in their own applications. These computational supports enhance effectiveness of analytical research in morphology, semantics and sociolinguistics as well as other fields such as applied linguistics and ethnomusicology.

### BUDGET

---

<b>Outright Request</b>	\$59,999.00	<b>Cost Sharing</b>	
<b>Matching Request</b>		<b>Total Budget</b>	\$59,999.00
<b>Total NEH</b>	\$59,999.00		

### GRANT ADMINISTRATOR

---

Jeff Conley  
Sponsored Programs Specialist  
32 Campus Drive 4104  
Missoula, MT 59812-4104  
UNITED STATES

**E-mail:** jeffrey.conley@umontana.edu  
**Phone(W):** 406-243-4954  
**Fax:**

## Table of Contents

1. Table of contents	p.1
2. List of participants	p.2
3. Abstract and statements of innovation and humanities significance	p.3
4. Narrative	p.4 – p.9
5. Budget (Budget form and Budget narrative)	9 pages
6. Biographies	p.19
7. Data management plan	p.20– p.21
8. Letters of commitment and support	
9. Appendices	

## **2. List of Participants**

Min Chen (PI)

Associate Professor, Dept. of Computer Science, University of Montana

Mizuki Miyashita (Co-PI)

Associate Professor, Linguistics Program, Dept. of Anthropology, University of Montana

Andrea Berez (Consultant/Product evaluator)

Assistant Professor, Dept. of Linguistics, University of Hawaii

Joyce McDonough (Consultant/Product evaluator)

Associate Professor, Dept. of Linguistics, University of Rochester

Naomi Palosaari (Consultant/ Product evaluator)

Project Manager, Linguist List, Eastern Michigan University

### 3. Abstract and Statements of Innovation and Humanities Significance

This interdisciplinary project seeks a Level II Start-Up grant to research, develop and evaluate an advanced *multimedia search prototype* for language documentation, analysis, and application. This project contributes to **documentary linguistics** especially in **endangered languages** by 1) providing an extended text-based search on *metadata* (transcription, translation, and annotation) to automatically find structurally and semantically related words and 2) providing intelligent processing and retrieval on *audio* files to efficiently and directly locate targeted audio segments. It will be publicly available and widely applicable via web browsers. Each major component will be made language/platform independent so users have the flexibility to use one or more search functions in their own applications. These computational supports enhance effectiveness of analytical research in morphology, syntax, semantics and sociolinguistics as well as other fields such as applied linguistics, ethnomusicology, and language revitalization. Its performance will be evaluated using audio recordings and their metadata in various endangered languages.

#### Statement of Innovation

The project provides multimedia search functions that are lacking and urgently needed in documentary linguistics. It reuses and extends functions provided by existing tools/systems from other fields to serve essential tasks in humanities. It will be free and open source software, deployed in a collaborative, distributed, network-centric environment for public use, interpretation, double-checking and extension. By using the web service technique, the developed system is easily extendable to become a *one-stop platform* for language documentation and analysis.

#### Statement of Humanities Significance

Endangered languages represent a vast repository of human knowledge on the natural world and cultural traditions that is irreplaceable, and must be documented as extensively as possible before they die. *Audio processing/retrieval* quickens database creation in phonetics-phonology; *text-based search* accelerates the annotation process and provides an effective method for discovering information from annotation. Our project also enhances other research activities, such as study of mainstream languages, applied linguistics, ethnomusicology, language revitalization, etc.

## Multimedia Search Prototype for Endangered Language Documentation and Analysis

### 1. Enhancing the humanities through innovation

In recent years, many descriptive linguists studying endangered languages have been showing their interest in *documentary linguistics* by recording, analyzing and preserving languages which contribute to diverse interest groups including speaker communities (Austin 2010). Challenges for documentary linguistics include *interdisciplinarity*, *audio-documentation* and *metadata development*. Language documentation requires a multidisciplinary expertise including linguistics, anthropology, history, musicology, psychology, applied linguistics, computer science (CS), etc. Although it has been mentioned that interdisciplinarity is difficult to reach because of the difference in collaborators' theoretical and practical concerns, this project shows the importance and feasibility of interdisciplinarity. In recording languages during fieldwork, advancements in digital technology have led researchers to obtain massive amounts of digital files and associated metadata. Consequently, data processing and searching has become overwhelming. Fast and direct access to interested segments in the recordings is important for researching especially on (but not limited to) understudied endangered languages. However, computational support on *audio processing and retrieval* is seriously limited. In addition, *metadata* are data about the data that ensures its context, meaning and use for appropriate determination, and are categorized into 'thick data' (e.g., transcription, translation and annotation) and 'thin data' (cataloging). However, computational support has mainly been on cataloging (Nathan and Austin 2004). Our project provides support for processing and searching *multimedia data including audio and metadata* to enhance effectiveness of analytical research activities.

This interdisciplinary project seeks a Level II Start-Up grant to research, develop and evaluate an advanced *multimedia search prototype* for endangered language documentation, analysis, and application. Its **innovation** lies in 1) its extended text-based search on *metadata* with a focus on 'thick data' to automatically find structurally and semantically related words and 2) intelligent processing and retrieval on *audio* recordings to efficiently and directly locate targeted audio segments; both are lacking and are urgently needed in documentary linguistics. It will be made into free and open source software for public use, double-checking and extension. By using the web service technique, the prototype is easily extendable to become a *one-stop platform* for language documentation and analysis. Its performance will be evaluated by various endangered languages including Native American languages.

The proposed project has significance to humanities. Currently, about 90% of the world languages are endangered and many are quickly vanishing (Crystal 1999). Language Endangerment is considered one of the most urgent problems in humanities (Rogers and Campbell 2011). Endangered languages represent a vast repository of human knowledge about the natural world and cultural traditions, and this knowledge is irreplaceable. Language documentation must be done as extensively as possible before they fall silent. The project's results will be valuable to linguistics scholars and students. With *audio processing and retrieval*, database creation in phonetics and phonology will be quickened. The *text-based search* accelerates the annotation process especially for young researchers who are not yet familiar with the language being researched, and it provides an effective search strategy in discovering information from annotated materials. For instance, the search prototype helps find morphologically and semantically related words serving research in morphology, syntax, semantics and sociolinguistics. Our supports for documentary linguistics also enhance other linguistics activities. In descriptive linguistics in Blackfoot, for example, the new searching environment helps find words and morphemes that have not yet been listed in the dictionary (Frantz and Russell 1995). Linguists in mainstream languages can use the system since documentary linguistics is not necessarily about endangered languages per se (Austin 2010). Furthermore, researchers in other fields such as applied linguistics or ethnomusicology and activists in language revitalization can make a use of this project to energize their activities.

### 2. Environmental scan

Currently there are several tools or systems for linguistics studies: FLEx and Toolbox for annotation, CLAN for field and corpus linguistics, Transcriber for audio file annotation, ELAN for multi-level annotation and interlinear analysis, and Praat for acoustic analysis, etc. These tools help scholars generate

metadata for their research. What these tools often lack is a capable search function to help effectively find *information* from *data*. Most of them use an “exact match” scheme. For example, given a search term “examine,” all entries containing this word are returned (e.g., “**examine** the effects”, “want to **examine**” etc.). However, entries containing morphologically related words (e.g., “**examining**,” “**examination**”) or semantically related ones (e.g., “test,” “check”) are not found. This limitation is daunting for endangered language research where prior knowledge of the languages is relatively minimal because annotation results may be different from researcher to researcher. Moreover, the same word may be transcribed with variants in the same course of recording due to natural differences among speakers. Thus it is difficult to find the best search term to cover such research-significant variants, and therefore, an intelligent *text-based search* mechanism is in great need. In terms of *searching on raw audio files*, these tools provided limited, if any, support. Commonly, researchers search for sample files of pronunciations by 1) listening to entire recordings to locate target segments or 2) creating text annotation on recordings to access targeted segments through their time-alignments with annotations. However, both processes are time consuming and infeasible in endangered language research due to its urgency and constraint of resources. Our project provides an advanced multimedia search prototype to address these issues. Literature shows other fields such as information retrieval have researched on intelligent text-based searching for years and developed some algorithms and software systems as results. Among them, the porter stemming algorithm (Porter 1980) and Latent Semantic Indexing (LSI) mechanism (Deerwester et al. 1990) are well-acknowledged and proven effective (Liu & Xu 2007). However, they were not designed for documentary linguistics and need to be properly revised and extended before adopted by our project. Meanwhile, audio searching is a relatively new field where techniques ranging from computational linguistics, artificial intelligence, and databases have made contributions but more research is required, especially for documentary linguistics.

In addition, most existing linguistics tools are stand-alone, PC-based applications. With the advance in networking, researchers across fields increasingly suggest to move software components towards collaborative, distributed, network-centric environments (Zou & Kontogiannis 2000). The Max Planck Institute for Psycholinguistics (MPI) discussed the importance of providing a simple-to-use web-based framework for linguistics work and developed several web-based tools (e.g., Annex, IMDI browser and LEXUS tool). There are other web-based systems such as language archives (AILLA, DOBES, ELAR, etc.) that play recordings and show annotations. All these are good initiatives and demonstrate the feasibility and advantages of deploying complex linguistics applications in a web-based environment. Therefore, our multimedia search prototype will be made publicly available via a web browser. In addition, each major component will be independent to language, platform (i.e., hardware, operating systems), and web browser so users have the flexibility to use one or more search functions in their own applications. As shown in various fields ranging from business management (Zou and Kontogiannis 2000) to scientific applications (Balis, Bubak, & Wegiel 2008), this can be achieved by making commonly useful components into web services.

### 3. History and duration of the project

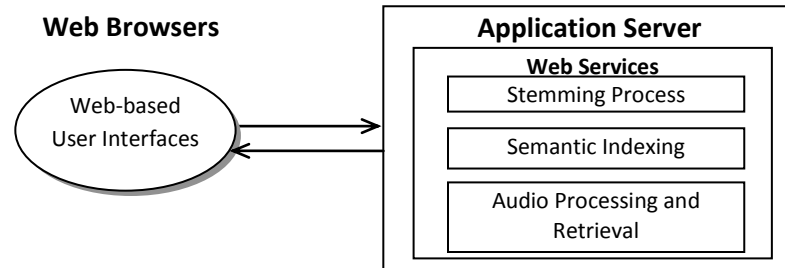
In terms of the **text-based search** component, the PI has received funding from the National Cancer Institute and the Office of Naval Research through subcontracts, which resulted in working programs that can be partially re-used, extended and revised for linguistics applications. As for **audio processing and retrieval**, part of the research was successfully completed with support from the NEH Digital Humanities Startup Level I (2009-2011), which resulted in a prototype system to output a list of audio clips containing one target sound and several forms of dissemination: a paper in the CS proceedings, a presentation at the Conference on the Endangered Languages and Cultures of Native America, and a journal article under review (*Literary and Linguistic Computing*). This prototype system can be improved and extended for this project. In the past year, the Co-PI has transcribed, translated and annotated more Blackfoot recordings supported by the Documenting Endangered Languages fellowship from NEH (2011-2012), and recorded Blackfoot word pronunciation in isolation in 2011. These data together with Navajo and Dene Sùline provided by our consultants, as well as additional data in the open language archive

ELAR will be used to test and improve our algorithm. For **web-based development**, the PI was a development team leader (2004-2007) and a consultant (2007-2009) to a web-based project with a total budget of \$4.3M, which has led to multiple publications as well as a working system (see biographies). This experience and the system architecture are beneficial in making our multimedia search prototype web-based, and major components publicly available via web services.

After the period of this Level II funding, the project is anticipated to continue. The **maintenance** plan is designed to preserve the project developed during the awarded project period with minor updates to the software system and data being produced. We expect financial needs to be minimal, and institutional support, such as office space, laboratories, equipment, etc., is guaranteed. The PIs have previously received and expect to obtain more institutional small grants. The **improvement** plan includes adding more useful functions (such as interactions and integrations with popular linguistics tools FLE<sub>x</sub> and Praat, version control and progress checking for collaborative work) that requires larger funding for research, system extension/improvement, fieldwork, consultation, and hiring graduate student assistants. We plan to seek more funding opportunities such as the Digital Humanities Implementation Grant (NEH), the Digging Into Data Challenge (AHRC/ESRC/IMLS/JISC/NEH/NSF/NWO/SSHRC), the Linguistics Grant (NSF) and/or the Documenting Endangered Languages Grant (NEH/NSF).

#### 4. Work plan

One main goal is to make the search prototype publicly available and widely applicable in linguistics while maintaining minimal project cost. Our plan is to reuse and extend functions provided by existing tools/systems from other fields as well as our preliminary research without needing to re-write from scratch, expose them as web services, and integrate them properly in a web-based platform. In brief, a *web service* is a network-accessible program that uses a standardized messaging protocol (Newcomer E. 2002) to communicate with other programs. It is an XML application mapped to programs, objects, databases, or comprehensive business functions (Aho et al. 2009). The overall system's architecture is shown in Figure 1, and its advantages are outlined below.



**Figure 1.** Overall architecture of the multimedia search prototype

- **Availability and flexibility:** A web service can be published, found and used through the web. Users have the flexibility to use individual search function via its web service in their own applications or to use them through our web applications.
- **Platform, language, and browser independence:** Our search prototype is applicable to different “environments” because the basic web services platform is XML, which is language and platform independent, and our web application uses techniques like HTML/HTTP and JavaScript, which are supported by all major browsers. Therefore users can integrate it easily into their work.
- **Interoperation:** Web services help solve the interoperability problem among various components by following common standards. This characteristic is especially useful for expanding our project to be a one-stop computational platform for linguistics study by adding many other useful components.

#### 4.1 Text-based search

**Stemming processing:** We propose to implement a stemming process as an option for users to enable or disable during the search based on their preference and needs. A *stemming process* is a computational procedure that identifies word variants and reduces them to a single morphological root (Lennon et al.

1981) (e.g., “**examine**” “**examining**” and “**examination**” with the same root “**examin**”). The key problem is to specify the stemming rules. Literature shows that the porter stemming algorithm (Porter 1980) is one of the most well acknowledged algorithms, which has been designed to work with English translations and has been extended to French and Italian (Wexler et al. 1997), Dutch (Kraaij and Pohlmann 1996), and others. Its open source implementations are available in almost every popular programming language, such as Java, C#, and Perl. We will first adopt the algorithm for English in our prototype to test the feasibility of our system architecture and leave its extension on other common languages as one of our future plans. It mainly benefits the search on the translation layer for endangered language research. To better support searching on other layers, such as transcriptions written in the orthographies of indigenous languages or in IPA, we plan to construct additional system interfaces for users to define their own stemming rules. Users can thereafter choose the user-defined rules on their targeted data sets, which adds flexibility to our search prototype and allows researchers to try and test the rules. To integrate this algorithm into our project and to make it work coherently with other important components, we will use the method discussed in Section 4.3.

**Semantic Indexing:** Due to different morphological patterns from mainstream languages, it may not be realistic to define rules for stemming processing using existing techniques. We, therefore, will also adapt a more intelligent algorithm, Latent Semantic Indexing (LSI) (Deerwester et al. 1990), to help search on all annotation layers by discovering semantically related terms. LSI is reported to be used in Google (Liu & Xu 2007). For example, when we search “~examine” in Google, with a tilde as a prefix, which activates LSI, files containing “test” or “check” will also appear in the search results; however, without the tilde, these search results would not be returned. Many studies (Bradford 2009; Zhang et al. 2011) have proven LSI performs well **even without knowledge of the meaning** of documents or words, which is especially useful when prior knowledge of the endangered languages may be minimal. The LSI manipulates the idea that words in documents are related via concepts, and the more words co-occurred across a set of documents, the more these words are semantically related. This idea can be used as an intelligent mechanism (a data mining approach) to help discover or verify semantic relationships among words. It is exciting because we may explore semantic organization within under-studied languages to find distinguished semantic relationships that differ from English. For example, in Blackfoot, *song* may be in closer relation to *war* than *entertainment*. The integration of an LSI algorithm in the proposed system is similar to the integration of a stemming algorithm as its open source implementations are available in many programming languages such as C and C++.

#### ***4.2 Audio processing and retrieval***

We will develop web services to support audio processing and retrieval via a “Query-by-Example” (QBE) mechanism. Users are allowed to submit an audio sample as a search request and the service will search for similar recording segments in the data storage and return them to the users. Part of the algorithm has been studied and implemented into a prototype system in our preliminary work. In this proposal, it will be further improved and extended, then embedded as a web service in our computational platform. The discussion on software implementation will be omitted. Instead, we will focus on the research details and extensions we plan to do.

**Noise filtering:** Noise introduced during the course of recording is likely to affect the accuracy of computer analysis. Since noise has a much higher frequency level in comparison to speech (0-7 kHz), a low-pass noise filter will be integrated into the component and provided as an option for users.

#### **Syntactic analysis:**

- **Data parsing:** Since audio files are often long, they need to be parsed into manageable units for computational processing and analysis. In our preliminary work, audio files were processed at frame level (512 samples; a total of 32ms) based on the literature (Chen et al. 2006). In this project, it will be further extended to include word/phrase level, a more natural concept in morphology and syntax, by using computer-based audio processing developed during our other grant. By default, short pauses between words/phrases will be used by the program as the unit boundaries. In addition, scholars may record words in isolation using self-defined unit boundaries for before and after the pronunciation of a



word (e.g., making tapping noise on the desk or using a built-in markup function in recording devices). We will also develop an interface for users to upload the audio sample of their self-defined boundaries for the wider research community.

- **Feature extraction:** Similar to the traditional database, each basic audio unit is characterized by audio features extracted for acoustic analysis. In our preliminary work, twenty-eight features were extracted from each frame following the common practice, which helped produce reasonably good results in speech analysis as also reported in the literature (Kim et al. 2007). To extend the work for word/phrase level units, a hierarchical structure will be developed where each word/phrase contains a set of frames, and is represented by the statistical combination of the features of its frames. More features such as the ones proposed in Ayadi et al. (2011) will also be studied and tested for further improvement.

**Data mining and retrieval:** To retrieve audio data similar to the search request, the simplest approach is to directly compare audio features between the user sample and the stored units in the system. This approach in fact works well with coherent data, but speakers' speech styles (e.g., volume and speed) can cause variations in features. To handle the heterogeneity, we have preliminarily developed an adaptive-data-mining-based mechanism to detect audio segments containing the velar fricative sound [x] in Blackfoot. Given a set of training data (i.e., the unit is marked as “yes” if containing [x] or “no” if not), the algorithm helps build a statistical model to represent the target sound pattern, which can then be applied to an extended data set (i.e., recordings without annotation) for target sound detection. Our empirical studies have shown that the framework could achieve far better accuracy than other general data mining approaches, and it is expected to perform even better with the planned addition of more features and training data. In this work, this approach will be extended in three directions. First, **more training models** will be built for other sounds such as voiceless stops [p], [t] and [k]. Second, for search requests with sounds different from the ones that the system has built models for, a user feedback mechanism will be applied followed by the data mining process. The idea is to return sample results using pure audio feature comparison. The ones that users mark as correct will be used as training data and the statistical model will then be built and applied on all the stored data for retrieval. This approach still requires some manual efforts and the model takes time to build. However, every time a new model is built, it will be stored for future usage, so in the long run, more and more search requests are expected to be covered by these saved models. Third, a database will be developed for better data management and faster retrieval. We plan to use MySQL, the most popular open source database management system to manage the data.

#### ***4.3 Develop Web Services and Web User Interfaces***

To implement the web service for each search component, we will adopt and extend the methods proposed by Canfora et al. (2008) and Balis, Bubak, & Wegiel (2008). It only requires an understanding of the component interface (its static structure and dynamic user-system interactions) to produce a wrapper as the Web Service. Specifically, we will develop an interactive web UI to map the interface structure and a request/response web service interface to handle interactions. FSA, a model of the human-computer interaction catchable by Web UI and interpretable by the wrapper, is used for specifying user-system interactions, which can then be represented using state diagrams. Developing programs based on state diagrams is a common and well-established practice in software engineering.

With the Web UI and the wrapper, the user's input and action on the webpage results in an HTTP POST request that will initiate an interaction with the application server. The application server then talks to the tools through the wrapper (i.e., web service). Next, an updated version of the window content is created and sent back to the browser.

#### ***4.4 System evaluation***

To evaluate the system functionality and user-friendliness, we will use ELAN files (.EAF) and three types of recordings: (i) conversations, (ii) narratives, and (iii) word-pronunciations in Blackfoot, Navajo and Dene Sųline, as well as the open language archive ELAR. First, ELAN files are used for the **text-based search**. We will submit different text data (e.g., part of word, word, phrase) and compare the results with and without the stemming processing and semantic indexing options. Second, for **audio processing and retrieval**, we will submit a query to the system using sample sounds and check the list of segments

retrieved from the recordings as query results. For both query types, retrieval accuracy will be calculated. Third, we will evaluate the **web UI and web services** generated for the retrieval components. A team of linguistics experts and native speakers (see Section 5) will use the web browser to access the application and submit various searches to determine whether the results are meaningful, the system interface is intuitive to operate, and the retrieval speed is reasonable. The team will also follow our manual to download the web services to be used in their own computers and check whether they can be conveniently used with their local documents and programs. In addition, the PIs plan to attend conferences for dissemination and contact interested researchers and individuals from related fields after our presentations to try the system. An evaluation form will be developed to get their feedback.

Table 1 shows the tentative schedule. Each sub-task will also include the associated experimental evaluations, report generations, system prototyping, etc. System integration, evaluation, deployment, and improvement are required to make sure all components are working properly with each other. Then the system as a whole will be evaluated, deployed, and further improved with collected feedback. Note there are time overlaps among tasks because certain steps can be carried out concurrently and a spiral software development model (Boehm, 1987) is used with iterations (e.g., the algorithms may be revised, validated and improved along with the research progress).

• Table 1: Tasks and Schedules of the Project

Task	Sub-task	Staff	Time
Develop web service & web UI	<i>Build overall system architecture</i>	Chen, student	May 2014 – Aug. 2014
	<i>Develop web UIs for data retrieval component</i>	Chen, Miyashita, student, & consultant	Jun. 2015 – Aug. 2015
Data retrieval	<i>Extension of stemming and develop its web service</i>	Chen, student, & Miyashita	Jun. 2014 – Dec. 2014
	<i>Research semantic indexing and develop its web service</i>	Chen, student, & Miyashita	Jul. 2014 – Jan. 2015
	<i>Extension and research on audio query by example and develop its web service</i>	Chen, student, & Miyashita	Dec. 2014 - Aug. 2015
System evaluation	<i>System integration, evaluation, deployment, and improvement</i>	Chen, Miyashita, student, & consultant	Jan. 2015 - Oct. 2015

## 5. Staff

The PI, **Dr. Min Chen**, will research audio processing and will conduct multimedia data search and prototype development. The Co-PI, **Dr. Mizuki Miyashita**, will monitor the project progress and ensure the project will stay on track with the documentary linguistics. **Drs. Joyce McDonough** (University of Rochester), **Andrea Berez** (University of Hawaii), **Naomi Palosaari** (Eastern Michigan University), and two native speakers will evaluate our project and provide consultation. **Graduate students** in Linguistics and CS will assist the PIs.

## 6. Final product and dissemination

We intend to disseminate our final project in recognized linguistics and CS conferences (e.g., Society for the Study of the Indigenous Languages of the Americas Annual Meeting, ACM International Conference on Multimedia) and prestigious journals (e.g., Journal of Language Documentation & Conservation, and IEEE Transactions on Multimedia.) All search components are implemented into web services, which are publicly available and are not tied to any particular operating system or programming language, so users can access any of them free of charge and include it in their own applications without compatibility issues. In addition, the search prototype is web-based so it can be accessed using web browsers. Source code, detailed readme documents and user manuals will be published on the website for potential interpretation, double-checking, extension, or use of the system. Linguistic data will be available for access and sharing as soon as they are reasonably ready.

## References

Please see the Appendices for the reference list



NATIONAL ENDOWMENT FOR THE  
**Humanities**

# Budget Form

OMB No 3136-0134  
Expires 7/31/2015

Applicant Institution: *University of Montana*

Project Director: *Min Chen*

Project Grant Period: *05/01/2014 through 10/31/2015*

[click for Budget Instructions](#)

	Computational Details/Notes	(notes)	Year 1	(notes)	Year 2	(notes)	Year 3	Project Total
			05/01/2014- 04/30/2015		05/01/2014- 10/31/2015			
<b>1. Salaries &amp; Wages</b>								
PI Min Chen	Academic year salary: \$88,727 (yr1), \$91,389 (yr2)	10 days	\$4,670	10 days	\$4,810	%		\$9,480
Co-PI Mizuki Miyashita	Academic year salary: \$62,552 (yr1), \$64,428 (yr2)	9 days	\$2,963	9 days	\$3,052	%		\$6,015
CS RA student		%	\$8,130	%	\$4,500	%		\$12,630
RA Linguistics		%	\$1,080	%	\$1,080	%		\$2,160
		%		%		%		\$0
		%		%		%		\$0
<b>2. Fringe Benefits</b>								
PI Min Chen	25% of funded portion of salary		\$1,168		\$1,203			\$2,371
Co-PI Mizuki Miyashita	25% of funded portion of salary		\$741		\$763			\$1,504
CS RA student	4.5% of funded portion of salary		\$366		\$203			\$569
RA Linguistics	4.5% of funded portion of salary		\$49		\$49			\$98
<b>3. Consultant Fees</b>								
Linguistics experts and native speakers	Evaluation and consultation		\$500		\$500			\$1,000

Professional editor	Editing service				\$300			\$300
<b>4. Travel</b>								
PI Min Chen	2-day trip for a planning meeting at the NEH offices (Fly from Missoula to Washington, DC; Airfare: \$800, p/d: \$300)		\$1,100					\$1,100
	Conference				\$1,300			\$1,300
Co-PI Mizuki Miyashita	Conference				\$1,300			\$1,300
	Trip to field for evaluation and consultation		\$1,000					\$1,000
Consultant	Travel for evaluation and consultation				\$1,200			\$1,200
<b>5. Supplies &amp; Materials</b>								
Software license	Audio editing, audio/video processing, and server software		\$375					\$375
<b>6. Services</b>								
								\$0
<b>7. Other Costs</b>								
								\$0
<b>8. Total Direct Costs</b>	<b>Per Year</b>		<b>\$22,142</b>		<b>\$20,260</b>		<b>\$0</b>	<b>\$42,402</b>
<b>9. Total Indirect Costs</b>								

41.5% of TDC effective 7/1/2012 per The University of Montana's federally negotiated Indirect Cost Rate Agreement with US Department of Health and Human Services dated July 30, 2009.								
	<b>Per Year</b>		\$9,189		\$8,408		\$0	<b>\$17,597</b>
<b>10. Total Project Costs</b>	(Direct and Indirect costs for entire project)							<b>\$59,999</b>
<b>11. Project Funding</b>	<b>a. Requested from NEH</b>							
							Outright:	\$59,999
							Federal Matching Funds:	\$0
							<b>TOTAL REQUESTED FROM NEH:</b>	<b>\$59,999</b>
	<b>b. Cost Sharing</b>							
							Applicant's Contributions:	\$0
							Third-Party Contributions:	\$0
							Project Income:	\$0
							Other Federal Agencies:	\$0
							<b>TOTAL COST SHARING:</b>	<b>\$0</b>
<b>12. Total Project Funding</b>								<b>\$59,999</b>

Total Project Costs must be equal to Total Project Funding ----> ( \$59,999 = \$59,999 ?)  
 Third-Party Contributions must be

## 1. Salary & Wages

### PI Salary

Annual base salary starts at \$88,727 in year 1 and is escalated 3% in year 2.

### Co-PI Salary

Annual base salary starts at \$62,552 in year 1 and is escalated 3% in year 2.

### Research Assistants Salary

One Computer Science student is budgeted in years 1 and 2 with pay rate at \$15 per hour.

One Linguistics student is budgeted in years 1 and 2 with pay rate at \$12 per hour.

## 2. Fringe & Benefits

Academic Year Faculty – Fringe benefits are computed at 25% of salaries.

Student Assistant – Fringe benefits are computed at 4.5% of salaries.

## 3. Consultant Fees

### Linguistics experts and native speakers

Dr. McDonough, Dr. Berez, and Dr. Fox as well as two native speakers will evaluate the product and provide consultation in years 1 & 2. Consultation fees are budgeted for \$500 each year.

### Professional editor

\$300 is allocated for editing services for abstracts, papers, and other dissemination information.

## 4. Travel

### PI

- 2-day trip for a planning meeting at the NEH offices (Fly from Missoula to Washington, DC; Airfare: \$800, p/d: \$300)

- 3-day research conference in year 2 for project dissemination (Fly from Missoula to conference site, possibly in US; Airfare: \$800, p/d \$360, registration fee: \$140)

### Co-PI

- 3-day research conference in year 2 for project dissemination (Fly from Missoula to conference site, possibly in US; Airfare: \$800, p/d \$360, registration fee: \$140)

- \$1,000 is budgeted for evaluation in the field with linguistics experts and native speakers.

### Linguistics expert

\$1,200 is allocated for linguistics expert's travels to the University of Montana for evaluation and consultation in year 2.

## 5. Supplies and Materials

- Audio editing software and license estimated at \$175 in year 1.

- Software estimated at \$200 in year 1 for audio/video processing and server software.

**9. Indirect Costs**

- Indirect Costs are calculated at 41.5% of TDC effective 7/1/2012 per The University of Montana's federally negotiated Indirect Cost Rate Agreement.

**Total amount of request: \$ 59,999.**

**COLLEGES AND UNIVERSITIES RATE AGREEMENT**

EIN #:

DATE: July 30, 2009

INSTITUTION:  
University of Montana, The  
Research Administration  
University Hall 116

FILING REF.: The preceding  
Agreement was dated  
September 23, 2003

Missoula

MT

59812-4104

The rates approved in this agreement are for use on grants, contracts and other agreements with the Federal Government, subject to the conditions in Section III.

**SECTION I: FACILITIES AND ADMINISTRATIVE COST RATES\***

RATE TYPES: FIXED FINAL PROV. (PROVISIONAL) PRED. (PREDETERMINED)

TYPE	EFFECTIVE PERIOD		RATE (%)	LOCATIONS	APPLICABLE TO
	FROM	TO			
PRED.	07/01/09	06/30/10	41.5	On-Campus	Organized Res. (1)
PRED.	07/01/10	06/30/12	42.0	On-Campus	Organized Res. (1)
PRED.	07/01/12	06/30/13	41.5	On-Campus	Organized Res. (1)
PRED.	07/01/09	06/30/13	56.0	On-Campus	Instruction
PRED.	07/01/09	06/30/13	29.0	On-Campus	Other Spon Act
PRED.	07/01/09	06/30/13	26.0	Off-Campus	All Programs
PROV.	07/01/13	UNTIL AMENDED	Use same rates and conditions as those cited for fiscal year ending June 30, 2013.		

(1) Includes Forestry Research which was indicated as a separate rate in the previous rate agreement.

**\*BASE:**

Modified total direct costs, consisting of all salaries and wages, fringe benefits, materials, supplies, services, travel and subgrants and subcontracts up to the first \$25,000 of each subgrant or subcontract (regardless of the period covered by the subgrant or subcontract). Modified total direct costs shall exclude equipment, capital expenditures, charges for patient care, tuition remission, rental costs of off-site facilities, scholarships, and fellowships as well as the portion of each subgrant and subcontract in excess of \$25,000.

(1)

U41066



INSTITUTION:  
University of Montana, The  
Research Administration

AGREEMENT DATE: July 30, 2009

---

**SECTION II: SPECIAL REMARKS**

---

**TREATMENT OF FRINGE BENEFITS:**

This organization charges the actual cost of each fringe benefit direct to Federal projects. However, it uses a fringe benefit rate which is applied to salaries and wages in budgeting fringe benefit costs under project proposals. The fringe benefits listed below are treated as direct costs.

**DEFINITION OF OFF-CAMPUS:** A project is considered off-campus if the activity is conducted at locations other than in University owned or operated facilities and indirect costs associated with physical plant and library are not considered applicable to the project.

**TREATMENT OF PAID ABSENCES**

Holiday and other paid absences (excluding annual and sick leave) are included in salaries and wages and are charged to Federal projects as part of the normal charge for salaries and wages. Separate charges for the cost of these absences are not made. A separate charge is made to Federal projects for annual and sick leave accruals. Charges for salaries and wages must exclude those paid to employees for periods when they are on annual or sick leave.

**DEFINITION OF EQUIPMENT**

Equipment is defined as tangible nonexpendable personal property having a useful life of more than one year and an acquisition cost of \$5,000 or more per unit.

The following fringe benefits are treated as direct costs:

FICA, UNEMPLOYMENT COMPENSATION, HEALTH INSURANCE, ACCIDENT INSURANCE, AND RETIREMENT.

INSTITUTION:  
University of Montana, The  
Research Administration

AGREEMENT DATE: July 30, 2009

SECTION III: GENERAL

A. LIMITATIONS:

The rates in this Agreement are subject to any statutory or administrative limitations and apply to a given grant, contract or other agreement only to the extent that funds are available. Acceptance of the rates is subject to the following conditions: (1) Only costs incurred by the organization were included in its facilities and administrative cost pools as finally accepted; such costs are legal obligations of the organization and are allowable under the governing cost principles; (2) The same costs that have been treated as facilities and administrative costs are not claimed as direct costs; (3) Similar types of costs have been accorded consistent accounting treatment; and (4) The information provided by the organization which was used to establish the rates is not later found to be materially incomplete or inaccurate by the Federal Government. In such situations the rate(s) would be subject to renegotiation at the discretion of the Federal Government.

B. ACCOUNTING CHANGES:

This Agreement is based on the accounting system purported by the organization to be in effect during the Agreement period. Changes to the method of accounting for costs which affect the amount of reimbursement resulting from the use of this Agreement require prior approval of the authorized representative of the cognisant agency. Such changes include, but are not limited to, changes in the charging of a particular type of cost from facilities and administrative to direct. Failure to obtain approval may result in cost disallowances.

C. FIXED RATES:

If a fixed rate is in this Agreement, it is based on an estimate of the costs for the period covered by the rate. When the actual costs for this period are determined, an adjustment will be made to a rate of a future year(s) to compensate for the difference between the costs used to establish the fixed rate and actual costs.


D. USE BY OTHER FEDERAL AGENCIES:

The rates in this Agreement were approved in accordance with the authority in Office of Management and Budget Circular A-21 Circular, and should be applied to grants, contracts and other agreements covered by this Circular, subject to any limitations in A above. The organization may provide copies of the Agreement to other Federal Agencies to give them early notification of the Agreement.

BY THE INSTITUTION:

University of Montana, The  
Research Administration

(INSTITUTION)



(SIGNATURE)

Daniel J. Dwyer

(NAME)

Vice President for Research

(TITLE)

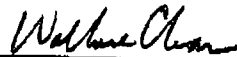
August 5, 2009

(DATE)

ON BEHALF OF THE FEDERAL GOVERNMENT:

DEPARTMENT OF HEALTH AND HUMAN SERVICES

(AGENCY)



(SIGNATURE)

Wallace Chan

(NAME)

DIRECTOR, DIVISION OF COST ALLOCATION

(TITLE)

July 30, 2009

(DATE) 1066

HHS REPRESENTATIVE: Ernest L. Willard

Telephone: (415) 437-7820

**UNIVERSITY OF MONTANA - MISSOULA  
FACILITIES AND ADMINISTRATIVE COST RATES  
FOR THE PERIOD JULY 1, 2009 THROUGH JUNE 30, 2013**

**EXHIBIT A  
PAGE 1 OF 1**

	ORGANIZED RESEARCH					
	JULY 1, 2009 THROUGH JUNE 30, 2010		JULY 1, 2010 THROUGH JUNE 30, 2012			JULY 1, 2012 THROUGH JUNE 30, 2013
	ON-CAMPUS	OFF-CAMPUS	ON-CAMPUS	OFF-CAMPUS	ON-CAMPUS	OFF-CAMPUS
BUILDING DEPRECIATION	1.60%		1.70%		1.60%	
EQUIPMENT DEPREC	0.70%		0.70%		0.70%	
INTEREST	0.80%		0.80%		0.80%	
OPERATIONS & MAINT	11.50%		11.80%		11.50%	
LIBRARY	0.90%		1.00%		0.90%	
GENERAL ADMIN	6.20%		6.20%		6.20%	
DEPT ADMIN	14.70%		14.70%		14.70%	
SPON PROJ ADMIN	3.10%		3.10%		3.10%	
ADMIN COMPONENTS	26.00%	26.00%	26.00%	26.00%	26.00%	26.00%
<b>TOTAL</b>	<b>41.50%</b>	<b>26.00%</b>	<b>42.00%</b>	<b>26.00%</b>	<b>41.50%</b>	<b>26.00%</b>

	INSTRUCTION		OTHER SPONSORED ACT		
	JULY 1, 2009 THROUGH JUNE 30, 2013		JULY 1, 2009 THROUGH JUNE 30, 2013		
	ON-CAMPUS	OFF-CAMPUS	ON-CAMPUS	OFF-CAMPUS	OFF-CAMPUS
BUILDING DEPRECIATION	2.50%		0.20%		
EQUIPMENT DEPREC	0.90%		0.10%		
INTEREST	0.90%		0.20%		
OPERATIONS & MAINT	13.20%		2.40%		
LIBRARY	12.50%		0.10%		
GENERAL ADMIN	3.10%		6.20%		
DEPT ADMIN	16.60%		14.70%		
SPON PROJ ADMIN	1.20%		3.10%		
STUDENT SERV ADMIN	6.20%				
ADMIN COMPONENTS	26.10%	26.00%	26.00%	26.00%	26.00%
<b>TOTAL</b>	<b>56.00%</b>	<b>26.00%</b>	<b>29.00%</b>	<b>26.00%</b>	<b>26.00%</b>

ADMINISTRATIVE COMPONENTS ARE CAPPED AT 26.0% IN ACCORDANCE WITH OMB A-21, DATED JULY 28, 1993.

CONCUR:

(SIGNATURE)

**Vice President for Research & Development**

TITLE

**July 29, 2009**

DATE

## 6. Biographies

**Dr. Min Chen** is an Associate Professor in the Department of Computer Science, University of Montana. Her research interests include distributed multimedia database systems and data mining as well as their application on real life problems and interdisciplinary projects. She has received funding from National Cancer Institute (2009-2011) and the Office of Naval Research (2008-2010) for text data search and analysis, and National Endowment for the Humanities (2009-2011) for Blackfoot audio processing and retrieval. In addition, she was a Computer Science development team leader (2004-2007) and a consultant (2007- 2009) to a web-based hurricane insurance projection project (total budget of \$4.3M from Florida Office of Insurance Regulation). The research and projects have led to more than 30 publications in top journals and conference proceedings and/or working systems, which help provide a solid foundation for this proposed work.

**Dr. Mizuki Miyashita** is an Associate Professor of Linguistics at the University of Montana. She earned Ph.D. in Linguistics at the University of Arizona in 2002. She specializes in sound pattern analysis of Native American languages, specifically, Tohono O'odham (Uto-Aztecan, spoken in Arizona and Mexico) and Blackfoot (Algonquian, spoken in Montana and Alberta, Canada). Current focus of her research is documentary linguistic in Blackfoot, and she is working closely with the Blackfoot native speakers and community members in Montana and Alberta, Canada. As part of her research, she has recorded Blackfoot songs and speech with funding provided by the Phillips Fund of the American Philosophical Society (2007) and the Jacobs Research Fund grant hosted by the Whatcom Museum (2008). She was awarded funding from Humanities Montana to conduct research on the use of Blackfoot conversations in linguistics analysis (2009-2011), National Endowment for the Humanities (2009-2011) to develop Blackfoot audio processing and retrieval, and Documenting Endangered Language fellowship (NEH) to transcribe narratives and conversations and provide interlinear analysis using ELAN. She also founded Blackfoot Language Group to create and disseminate materials for linguistic research and language teaching in Blackfoot ([www.umt.edu/ling/BLG/blg.html](http://www.umt.edu/ling/BLG/blg.html)). Through this activity, she has been engaging in continuous Blackfoot language research and student training in documentary linguistics.

## 7. Data Management Plan

### Expected data

Types of data obtained and generated during the proposed project are: linguistic data, software and documentation.

Linguistic data consists of sound files and ELAN files. The sound files are raw recordings of conversations (with multiple native speakers), narratives (one native speaker), and word-pronunciation sessions (one native speaker and one non-native researcher). ELAN files are generated by using the Eudico Linguistic Annotation program (ELAN). All these linguistics data will be open to anyone who is interested. However, it is possible for natural conversation data to contain private information (e.g., the speaker's income and home address). In such a case, only transcriptions will be made available in which private information is altered, and sound files will be available by contacting the PIs (under the condition of permission from the conversation participants).

Software includes a project website, a source code package, and web services. The website is developed to host project related data/information and to link to the search prototype. The source code package consists of all the programs developed including, but not limited to, the web user interface, search components, and the connections among them. The web services for the search components are included in the source code package. In addition, they are published separately as individual units so users are not forced to download the whole source code package in order to access any of the web services. All software is free and open to the public.

Documentation refers to detailed research reports, software development records, and a user manual. They provide detailed project information and a guide for potential interpreting, double-checking, extension, or use of the proposed project and research products. Documentation is also free and open to the public.

Before all these data are shared with others, they will be managed by the PIs and will be stored in multiple on-site backups, including in the servers owned by the PIs and external hard drives in the Linguistics lab and the Department of Computer Science (CS) at the University of Montana (UM). A copy of the data will also be located in the research media server managed by IT at the UM.

### Period of data retention

All linguistics data will be distributed as soon as they are available. For software, we expect to retain them during the development cycle and will make them available for invited evaluators for system evaluation and improvement. All software will be publicly distributed once they are given a green light by the evaluators and no later than the end of the project. In terms of documentation, research reports and software development records will be released gradually following the project progress, and the user manual will be made available when the software is distributed.

### Data formats and dissemination

All three types of data (*linguistics, software, documentation*) are electronic in both format and content. Specifically, linguistics data include sound files in .wav format and ELAN data in.eaf format. For the software data, the search prototype will be developed using programming languages like HTML and JavaScript so it will be in the formats defined by these languages. Documentation will be in .pdf format.

Because there are many different types and levels of linguistics data involved in the project, the policies for public access and sharing may differ depending on the type and content (as discussed in the Expected data section). For software and documentation data, we plan to make the search prototype open source software, issued under the GNU General Public License (<http://www.gnu.org/licenses/gpl.html>). In brief, the public can access and obtain our software and documentation for free, and have freedom to

distribute copies of our software and change the software or use pieces of it in new free programs.

For dissemination, all data types will be available for access and sharing through our project website, which will be developed using techniques supported by all major web browsers and made publicly accessible. The developed system is free and open source software, and the web services, by their nature, can be publicly found and used through the web. We will also archive them at UM's Mansfield library, which are publicly accessible; therefore, interested research community members and individuals can access them. They will also be disseminated through public repositories, such as GitHub (<https://github.com/>) and disciplinary repositories, such as the Computing Research Repository (<http://arxiv.org/corr/home>), free of charge. In addition, we will disseminate the project details and data access information in linguistics and CS conferences and journals as well as in research forums such as DBWORLD (<http://research.cs.wisc.edu/dbworld/>).

### **Data storage and preservation of access**

Since all our data are important to the humanities as a whole, it is ideal for these to be preserved permanently. They will be managed by the PIs (as discussed in the Expected data section), stored in public and disciplinary repositories, and will also be archived at UM's Mansfield Library that is currently in the process of developing a digital data acquisition regulation. It is expected that the materials would be stored and managed via their existing IT system, which includes standard backup and data redundancy methods. As they build their preservation policy and technical infrastructure, these methods will continue to evolve.

The data acquired and preserved in the context of this proposal will be further governed by the University of Montana's policies pertaining to intellectual property, record retention, and data management.

8 September 2013

Letter of support for proposal : **Multimedia Search Prototype for Endangered Language Documentation and Analysis**

With the development of computational power and new technologies, there is an increasing interest and opportunity in building spoken word corpora for under-documented languages and speech communities. However these data are more difficult to annotate and handle than data from more familiar languages. For one, the orthography or spelling systems are not standardized, and annotation and orthographic systems may not be consistent even within a speech community. Second, we also know little about the variability in the speech or the relationship between the symbol for a sound and its realization, information we have for larger and better know languages. For these type tasks, we need more sophisticated computational techniques, data retrieval, and data mining algorithms developed especially for work with under-documented languages and the special problems that arise with this type data.

Projects developing special cutting edge software for use in language documentation have been developed over the past decade at institutions such as the Max Plank in Nijmegen (MPITools) and SOAS (ELAR) in London, significantly increasing the productivity of the field worker, and the shortening of the time from the collection of data to its analysis and presentation. Importantly, the newly developed software enable the data to be presented in a transparent way. However there is a great deal more to do, especially in the area of data-retrieval and handling queries.

This grant will support the collaboration of a computer scientist and an active field linguist working with endangered language communities. The output of this grant will enable the data collected in the field to be better incorporated into larger databases of better studied languages. The incorporation of this data from typologically distinct languages will broaden understanding of the human language component.

I support this proposal and I'm interested in using these techniques in my own field data in Navajo, in northern Dene languages. This data consist of annotated sound files in Praat.

Sincerely,



Joyce McDonough

September 09, 2013

To Whom It May Concern:

I'm writing to strongly support **Dr. Min Chen's** application for your Digital Humanities Start-Up Grants program. Dr. Min Chen is an Associate Professor in the Department of Computer Science at University of Montana. In my opinion, she is exceptionally talented and unusually accomplished in large-scale collaborative research works for a young scholar at her stage. She has also won great respect from many researchers and professors from Hurricane Research Division/NOAA, University of Florida, Florida Institute Technology, University of Miami, and Florida State University because of her excellent work.

I have known Min since she joined my research team for her Ph.D. study in 2002. Dr. Min Chen has made significant contributions to the field of Multimedia Data Mining and Multimedia Database Systems. This can be evidenced by the facts that many researchers contacted her or me to inquire the work she has contributed. She has published **9** journal papers, **5** book chapters, and **22** refereed conference/symposium/workshop papers (at reputed conferences such as ACM Multimedia, IEEE ICME, IEEE ISM, ACM MDM/KDD, ACM MMDB, IEEE MSE, and IEEE IRI). She has served over 10 program committees of international conferences and workshops, co-organized an international workshop, and is serving as an editorial board member of an international journal in her field.

I would like to emphasize Dr. Min Chen's strength in extending her research to real-life applications and in collaborating with researchers across disciplines. I am always impressed by her technical expertise, scientific curiosity, work ethic, and excellent collaboration skills. She was a technical leader of the "Windstorm and Simulation Modeling" project funded by Florida Department of Community Affairs (FDCA) and Federal Emergency Management Agency (FEMA). This project aims to design and develop a 3-D visualization and animation system to simulate the storm surge effects at real-time. Her excellent work in system development and model construction helped build a solid foundation for the success of the project. Dr. Min Chen has been playing a key role in the "Hurricane Loss Projection Model" project, a multi-million dollar project that is funded by the Florida Office of Insurance

School of Computing & Information Sciences  
11200 S.W. 8<sup>th</sup> Street, ECS • Miami, FL 33199 • Tel. (305) 348-2744 • Fax: (305) 348-3549

Florida International University is an Equal Opportunity/Access Employer and Institution • TDD via FRS 1-800-955-8771



Regulation. The goal of the project is to develop and maintain a computer model to assess hurricane risks, and to project annual expected insured residential losses for specific sites, zip codes, counties, and regions in Florida. The model includes meteorological, engineering, GIS, statistical, computer science, and financial and actuarial components. Dr. Min Chen was the technical leader in implementing and integrating different components for this computer model, and cooperates with a team of meteorologists, wind and structural engineers, statisticians, computer scientists, actuaries, and financial experts from FIU, the State University System and elsewhere. From this project, she demonstrated her skill to manage a large-scale project, her strong programming skill, and her capability to work with multidisciplinary research areas. I envision she will continue to make a major impact on this project which will result in great benefits to the residents at Florida and provide an assistance to the insurance industry in the rate making process in this country to improve the U.S. economy.

I believe that Dr. Min Chen is a talented and enthusiastic academician and an excellent researcher. I strongly support her application without any hesitation. Please contact me if I can be of any assistance.

Sincerely,



Shu-Ching Chen, Ph.D.  
Professor and Graduate Program Director  
Associate Director, The Center for Advanced Distributed System Engineering  
Director, Distributed Multimedia Information System Laboratory  
School of Computing and Information Sciences  
Florida International University  
Miami, FL 33199, USA  
305-348-3480 (O)  
305-348-3549 (Fax)  
Email: [chens@cs.fiu.edu](mailto:chens@cs.fiu.edu)  
<http://users.cis.fiu.edu/~chens/>

School of Computing & Information Sciences  
11200 S.W. 8<sup>th</sup> Street, ECS • Miami, FL 33199 • Tel. (305) 348-2744 • Fax: (305) 348-3549

Florida International University is an Equal Opportunity/Access Employer and Institution • TDD via FRS 1-800-955-8771

## 9. Appendices

### Works Cited

- Aho, P., et al. 2009. MDA-Based tool chain for web services development, In: *Proceedings of the 4th Workshop on Emerging Web Services Technology*, pp. 11-18.
- Austin, Peter. 2010. Current issues in language documentation. In Peter Austin (ed.) *Language Documentation and Description*, Vol 7, 12-33. London: SOAS.
- Ayadi, M. E., et al. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition*, vol. 44, no. 3, pp. 572-587.
- Balis, B., Bubak, M., and Wegiel, M. 2008. LGF: A flexible framework for exposing legacy codes as services, *Future Generation Computer Systems*, vol. 24, pp. 711-719.
- Boehm, B. 1987. A spiral model of software development and enhancement. *Software Engineering Project Management*, pp. 128-142.
- Bradford, R. B. 2009. Comparability of LSI and human judgment in text analysis tasks, In: *Proceedings of the 11th WSEAS international conference on Mathematical methods and computational techniques in electrical engineering*, pp. 359-366.
- Canfora, G. et al. 2008. A wrapping approach for migrating legacy system interactive functionalities to Service Oriented Architectures, *Journal of Systems and Software*, vol. 81, pp. 463-480.
- Chen, S.-C., et al. 2006. A Multimodal Data Mining Framework for Soccer Goal Detection Based on Decision Tree Logic, *International Journal of Computer Applications in Technology*, vol. 27, no. 4, pp. 312-323.
- Crystal, David. 1999. *Language Death*. Cambridge University Press.
- Deerwester et al. 1990. Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407.
- Frantz, D. and Norma J. Russell. 1995. *Blackfoot Dictionary of Stems, Roots, and Affixes*. Second Edition. U Toronto.
- Kim, B.-W., et al. 2007. Speech/music discrimination using Mel-cepstrum modulation energy, In: *Proceedings of the 10th international conference on Text, speech and dialogue*, pp. 406-414.
- Kraaij, W. and R.Pohlmann. 1996. Viewing stemming as recall enhancement. In: *Proceedings, 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96) Zurich*, 40-48.
- Lennon, M., et al. 1981. An evaluation of some conflation algorithms. *Journal of Information Science* 3, 177-183.
- Liu, D. and S. Xu. 2007. Challenges of using LSI for concept location. In: *Proceedings of the 45th annual southeast regional conference*, 449 - 454.
- Nathan, D. and Peter Austin. 2004. Reconceiving metadata: language documentation through thick and thin. In Peter K. Austin (ed.) *Language Documentation and Description*, Volume 2, 179-187. London: SOAS.
- Newcomer E. 2002. *Understanding Web Services: XML, WSDL, SOAP, and UDDI*. Addison-Wesley, Boston.
- Porter, M.F. 1980. An algorithm for suffix stripping. *Program*, vol. 14, no. 3, pp. 130-137.
- Rogers, Chris and Lyle Campbell. 2011. Endangered Languages. *Oxford Bibliographies Online*. DOI: 10.1093/OBO/9780199772810-0013.
- Wexler, M., et al. 1997. Multi-language text indexing for Internet retrieval. In: *Proceedings of the 5th RIAO Conference on Computer-Assisted Information Searching on the Internet*.
- Zhang, W., et al. 2011. A comparative study of TF\*IDF, LSI and multi-words for text classification, *Expert Systems with Applications: An International Journal*, vol. 38, no. 3, pp. 2758-2765.
- Zou, Y. and Kontogiannis, K. 2000. Web-Based Specification and Integration of Legacy Services, In: *Proceedings of the 2000 conference of the Centre for Advanced Studies on Collaborative research*, no. 17.

Screenshot of ELAN file

The screenshot displays the ELAN software interface for the file 'Shirlee Friends.eaf'. The main window is divided into several sections:

- Menu Bar:** File, Edit, Annotation, Tier, Type, Search, View, Options, Window, Help.
- Grid:** A table listing annotations with columns for Nr, Annotation, Begin Time, End Time, and Duration. Row 22 is selected.
- Playback Controls:** A control bar with buttons for play, stop, previous, next, and other functions. A selection range is indicated as 00:02:51.255 - 00:02:54.060 (2805).
- Audio Waveform:** A visual representation of the audio signal corresponding to the selected time range.
- Linguistic Analysis Table:** A detailed breakdown of the selected audio segment into linguistic tiers.

Nr	Annotation	Begin Time	End Time	Duration
12	itamsooksii stoopohsow matsiyikkapisaa	00:01:13.550	00:01:18.360	00:00:04.810
13	Ka'itaamitakiw tsaaniwa akaa'toohkoi aakamotsiyi	00:01:24.320	00:01:29.680	00:00:05.360
14	aamoi l'nakka-in- l'naksii...kapiswa	00:01:30.260	00:01:38.720	00:00:08.460
15	imitaawa skai'itaamitakiwa ni'toyi sspopiwi	00:01:41.560	00:01:46.880	00:00:04.320
16	itaisawaahsitakiw matsiyikkapisaa	00:01:46.080	00:01:49.260	00:00:03.180
17	aanis- aanisiiwaiks aamo aamoyi akk- maanakkaawa	00:01:53.860	00:01:59.860	00:00:06.000
18	matsiyikkapisaa itsii- issimistotoyiw pokatsii- poka- poka...kattsisiyo	00:02:12.600	00:02:23.980	00:00:11.380
19	tsaaniwa isska'isawaahsi- ah! nayiyoo	00:02:27.390	00:02:32.230	00:00:04.840
20	limitaawa nitoyi kii sspopiwa	00:02:32.640	00:02:36.870	00:00:03.230
21	Tsaaniwa awaanistsiiwayi maatsokaa'piiwaatsiksi	00:02:41.540	00:02:47.460	00:00:05.920
22	akkaakomotsiyi'o'p kiaakitaissataistotowa	00:02:48.450	00:02:54.060	00:00:05.610

Tier	Text
IU (S)	akkaakomotsiyi'o'p kiaakitaissataistotowa
Word (S)	akkaamotsiyi'o'p      kitaakitaissataistotowa
Morph (S)	akkaa   m      otsiyi   o'p      kit   aak   it   a   iss   sataist   waw
Gloss (S)	be friend   ?      reciprocal   21      2   fut   loc   dur   in fron   make   4
Trans (S)	We are friends, you do not treat them like that (you will make him angry)