**Good**

**Chen**

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

This project goes to the core problem of language documentation efforts: the difficulty in ensuring that extensive sets of primary and derived data (basically digital recordings of endangered languages and associated transcriptions and translations) are accessible to scientific investigation. "Accessible" in this sense does not mean open access but rather relates to "discoverability" of those elements of the documentation materials that might be useful to researchers seeking to use these materials. Specifically Chen proposes to develop (1) a text-search tool to "automatically find structurally and semantically related words"; and a (2) tool for retrieving sections of an audio file that match particular acoustically defined search criteria. The team is eminently qualified (PI, co-PI, and consultants) and the proposal is a model of clear presentation, written with a deep awareness of both the relevant issues and of the tools and archives that are active in endangered language documentation effort. The close collaboration between a computational and field linguist is particularly important and should serve as a model for others. The tools to be developed will be made publically available and, indeed, could become part of the standard toolkit for documentary linguists and descriptive linguists who wish to take advantage of the growing sets of materials (dictionaries, grammars, and, particularly, digital and transcribed corpora).

Letters of interest from the major archives (AILLA, DOBES, ELAR, PARADISEC, UBerkeley, among others) would have been helpful. This is particularly true in regard to DOBES, whose program development team is actively involved in software development that might already address some of the issues brought up in this proposal.

The evaluation of this project will address the two major areas of activity: (1) text processing; (2) acoustic signal processing.

Language documentation efforts often produce substantial amounts of digital recordings that are transcribed in time-coded format (meaning that a particular segment of the sound file, perhaps some 5 to 10 seconds, often correspond to some syntactic or prosodic unit) using tools such as ELAN and Transcriber. The former allows tiers for analysis (e.g., different types of transcription, a gloss and translation line) the latter does not. Neither facilitates parsing and those who wish to transcribe and parse must port their material between several programs. Praat is used mostly for acoustic analysis of short sound files. Certainly a tool that could find certain items in a large sound file and extract them for acoustic analysis in Praat would be very useful assuming that the phonetician is interested in acoustic analysis of running, non-controlled/elicited speech.

The issue that this project addresses is how to allow scholars to search the texts for material relevant to their own research. Probably the most desirable way to do this is to parse the native language text into morphological units (parses) and their meanings (glosses) and search these lines for the relevant material. Toolbox/FLEx do provide some level of support for interlinearization though I am not sure if the search engine can act on different tiers (e.g., for an English example, search for "chair" with PoS 'verb'). Since Toolbox and FLEx are standard tools for documentation it might have been more useful to develop a tool that would search the native language interlinearized texts that these programs can produce.

Instead, the PI proposes two processes for text-based searches: (a) stemming; (b) semantic indexing. For stemming, the PI proposes to create an graphical user interface "for users to define their own stemming rules". One question is whether the writing of even stemming rules would be beyond the capabilities of anyone but a linguist working on the target language. If this is the case, then two caveats emerge. First, if stemming were the responsibility of researchers not necessarily familiar with the target language, then the goal of making documentation materials accessible would be complicated. Second, if stemming were the responsibility of the researcher working on the language, then one might wonder if the stemming builder that the PI will build would be more efficient or useful than the interlinearization functions of Toolbox and FLEx, which try to parse to the lemma, not stem. Another question is the degree to which any sort of "user defined stemming rules" would work on languages with a somewhat complex morphology (let alone Navajo, one of the target languages). Finally, it is not clear what format the input to the text-based search tool will be and whether the tool will be able to extract and save to file relevant material encountered in a search.

The PI recognizes the problems of stemming in morphologically complex languages and suggests using a "more intelligent algorithm, Latent Semantic Indexing" for discovery in native language texts. However, this presents a new problem (at least in the estimating of this reviewer who is not very familiar with LSI, though the concept is clear): the ability of LSI to work on very small corpora. Fifty hours of transcribed/parsed and glossed text is near the upper limits of most documentation efforts. The PI does not address the issue of data mining in small corpora so the statement that "many studies have proven LSI performs well EVEN WITHOUT KNOWLEDGE OF THE MEANING of documents or words" (emphasis in original) must be taken with a grain of salt if the corpus is small.

My knowledge of acoustic signal processing is extremely limited, as is awareness of the tools presently available. The PI proposes two levels of processing and retrieval: (1) word and (2) phonetic or phonemic (it is not clear which is intended). One question is, again, the nature of the input data. If it is already transcribed in a time-coded format, then the question is more of aligning the transcription to the acoustic signal, a process known as forced alignment. This has been done for phonetic research in major languages and the development of the ability to align signal and transcription in an endangered language would be a significant step. However, from the discussion it seems that the PI is talking more about using simply the acoustic signal as input into the search tool, which would find a section of the signal that matches one for which the user is searching. This is more in tune with automated speech recognition (ASR). An initial step seems to be dividing a signal into word/phrases based on the location of silence. Effectively this will take a long sound file and divide it into smaller units based on the location of "short pauses between words/phrases". The methodology for recognizing and marking as a boundary a given length of silence in an acoustic signal already exist as do the tools for then extracting and naming the word/phrase file. The extent to which the pauses mark significant syntactic units is a question for empirical research.

However, the project goes beyond boundary marking (from silence/pauses) to effectively propose what seems to be an incipient ASR model. My understanding is that in this project recognition of any given segment (e.g., velar fricative [x]) is developed as the result of training/machine learning: a search for [x] will be conducted and annotators will mark those searches that are successful gradually building up a database of the acoustic realizations of [x]: "the algorithm helps build a statistical model to represent the target sound pattern". This seems to be different in some respects from what I understand to be ASR procedures, which is to build up a language model and a database of the phonemic representation of possible words.

It is not clear to me whether the PI plans to work with language models or any sort of probabilistic factors in segment recognition. For example, let us assume that in a given language X the occurrence of a nasal consonant means that all stops in the word must be voiced. Taking this into account in a language with voiced and voiceless obstruents, the presence of a nasal consonant will reduce the possibilities of a subsequent obstruent to only the voiced set. Similar probabilistic models could be build for CV sequences, perhaps. It seems that the PI is basing recognition of a segment simply on acoustic features with no "help" from a model or set of representations of possible words.

Overall, my impression of this project is that while the team is excellently prepared and has a clear grasp of the tasks needed to make endangered language data useful to scientific research there are several flaws that should preclude funding at this point:

1. The project is extremely ambitious as it hopes to provide text and acoustic signal data extraction tools, working with three "evaluation" languages (Blackfoot, Navajo and Dene Suline) while also creating a web user interface. Considering the funds available I think it is necessary to limit the scope of this project drastically. A good starting point would be Blackfoot, leveraging not only Miyashita's knowledge but the support available through NSF. As mentioned at the beginning the collaboration between Chen and Miyashita is a strong argument in favor of this proposal and if not funded this cycle I would recommend the more limited goals of a proof of concept proposal utilizing Blackfoot data.

2. The amount and type of available materials should be clearly expressed in the proposal. For example, the Navajo material seems to be annotated sound files in Praat. This would not be useful for LSI application or even, if short word tokens, retrieval of sound segments. It is not even clear how much Blackfoot data is available and what it's structure is , e.g., how much interlinearized texts. Nor is it clear that stemming is the best approach for data retrieval from Blackfoot texts.

I think that the proposal is very good and innovative, however it would benefit from more focused objectives and greater attention to justifying some of the methodological choices made. It may be a project more suited to NSF Linguistics or Computational Science (or both).

**Overview**

Intellectual significance: The PI has correctly identified several key problems in endangered language documentation and has suggested ways in which the primary materials created by such efforts can be effectively mined for research by others.

Impact on research and technology: The use of natural language processing, particularly of acoustic signals, could have a significant impact on endangered language work.

Innovation: The proposal is highly innovative and potentially of great importance.

Proposal development: Several key issues are inadequately addressed (see comments above).

Feasibility: There are limits to the potential success of this project based on a limited corpus of material for LSI and of insufficient attention to acoustic and language modeling for data mining of acoustic signals.

Qualifications: The team seems to be highly qualified for the type of research proposed.