Narrative: A database of the Mutsun language

I. Introduction

The proposed project will encode all of the voluminous archival information about the extinct Mutsun language into two databases (a lexical database and a text database). The Mutsun language, a Native American language (Costanoan/Ohlone), was spoken in central Coastal California, near San Juan Bautista (see Appendix 1). The last native speaker of Mutsun died in 1930, but a tremendous quantity of archival information on the Mutsun language still exists (see Appendix 2). This archival material is largely in the form of microfilmed handwritten fieldnotes from the 19th and early 20th centuries, written by early linguists and anthropologists who worked with native speakers of the language. There is also an active language revitalization movement in the Mutsun community, and many Mutsun people are working to bring their heritage language back. The databases proposed in this project will serve two main populations. The first is linguists, who will be able to use the databases to analyze the language to provide data for theoretical linguistic investigations and comparative studies of Native American languages. The second is the Mutsun community, who are engaged in revitalizing their heritage language as part of a larger effort to reclaim their ancestral culture. (See II.C-E below.)

In its current state, the archival material is of minimal use for linguists because of the difficulty of locating particular information within it. It is of almost no use to the Mutsun community because it was written for professional linguists and because so much linguistic analysis is necessary to use information from any particular sentence. Funding through the NEH would allow for all of the recorded information about the language to be integrated into a well-analyzed database structure, making it useful for linguistic analyses and providing the information necessary to improve the teaching materials being developed for community use. The archival materials represent a part of our national cultural heritage: early linguists and anthropologists collected vast quantities of information about Native American languages which are no longer spoken, and some of these early linguists and anthropologists (e.g. Harrington and Arroyo below) are themselves important historical figures of our country. Their work is part of the cultural heritage of all Americans, as well as of specific Native American communities. This project proposes to make this cultural heritage accessible both to the Mutsun community and to a broader audience.

The larger project of which this proposal forms a part has two components: a community-oriented aspect, focused on language teaching, and a scientific aspect, focused on development of the databases and analysis of theoretical linguistic issues. This proposal is for the scientific aspect of the project, and involves funding of community-oriented work only to the extent that it also benefits the scientific side of the project, for example by obtaining feedback from the Mutsun community about the usability of products of the project. The more community-oriented aspects of the project are almost entirely budgeted as cost sharing. Both types of work are beneficial to both the linguistic and the Mutsun community, as explained in II.E below.

II. Significance

This project represents a type of linguistic research that is just developing, which combines work with archival primary linguistic data and community work. It will benefit both linguists and the Mutsun community, and ultimately, it will also benefit many other Native American communities.

A. Products of the project

The NEH-funded part of the project will produce four products: 1) A searchable electronic lexical database, containing all known information about every Mutsun morpheme, from every archival source. (Details about structure of the database, distribution, etc., appear below for this and the other products.) 2) A searchable electronic database of annotated texts, containing all recorded texts (primarily sentences, plus a small number of longer texts). 3) A hard copy English-Mutsun / Mutsun-English dictionary, generated from the lexical database. 4) A hard copy publication of the full set of annotated texts, generated from the text database. Additionally, the cost-sharing part of the project will produce an improved language textbook and ancillary teaching materials.

B. Estimated potential audience

There are several partially overlapping audiences for the products of this project. Typologists, historical linguists, and linguists who specialize in Native American languages, as well as syntacticians, semanticists, morphologists, and phonologists working on particular topics for which Mutsun provides evidence, form an audience of several hundreds of researchers. A second audience consists of linguists working on language revitalization, maintenance, endangerment, and languages in contact, an audience of several hundred additional researchers. Linguists working in applied linguistics and language teaching would also be interested in the dictionary and the textbook, since teaching a language when no fluent speakers are available presents interesting problems in the field of applied linguistics. This constitutes a very large audience because of the large size of the applied linguistics community.

A distinct community, which will have great interest in the dictionary and database generated by the project, is the Mutsun community itself. There are approximately 700 enrolled Mutsun tribal members (forming 260 households), and approximately 2000 Mutsun people total. Although language classes have been relatively small so far (approximately 45 people total have participated to date), these classes have only just begun, and interest is growing rapidly. The PI and Ms. Luna-Costillas (the community leader) anticipate that when better learning materials are provided as a result of the project and distance-learning methods are incorporated (currently in progress), the number of community members wishing to study the language will skyrocket. We expect that by the end of the project period, at least 400 Mutsun community members will wish to have a copy of the dictionary, annotated texts, and textbook generated by the project. In the long term, this is a potential audience of up to 2000 people. Because of issues of access to computers, it may be that only a subset of this audience is interested in the electronic products of the project. However, many Mutsun community members do have computers, and the electronic products will be designed to be easily usable for all potential audiences.

Members of other Costanoan groups (see Appendix 1) will also have a strong interest in the products of the project. Most other Costanoan languages are not as well documented as Mutsun, so the Mutsun dictionary and lexical database would provide valuable information for comparison with what is known about the other languages. Of the Costanoan languages, Mutsun, Chochenyo and Rumsien have active revitalization programs, and the products of the Mutsun project would be very useful to the Chochenyo and Rumsien programs (Yamane 2001, Blevins and Arellano 2004) for reconstructing words not recorded for their languages.

Finally, other Native American communities, and the linguists working with them, will form another important audience. There are hundreds of Native American communities with no fluent speakers of their heritage languages [50 languages in California alone (Hinton 1994)]. Many of these communities have a strong desire to bring their languages back using archival documentation. At the Breath of Life Language Revitalization workshops at the University of California, Berkeley (see III.A below), which have been attended by over 30 California language groups, the Mutsun project is already serving as an example for others. Funding of the current project would produce great progress in Mutsun language revitalization, and the Mutsun project would then serve as a model for revitalization for other indigenous languages in California, the rest of North America, and the world. Language death and revitalization are worldwide topics.

C. Importance of the project for linguists

For linguists, this project will provide an easily searchable lexical database of all the known information about Mutsun morphemes (i.e. words and affixes), as well as a related text database. The combination of these two databases will make it easy to create a concordance in order to examine every attested use of each morpheme. Because the language is extinct, it is possible to include all the information about the language in the databases and then to know that one has all recorded knowledge about the language available for analysis. There are a large number of questions in theoretical, historical, or typological linguistics which linguists will be able to study using Mutsun once these databases are complete. Several examples follow:

a. The hypothesized Penutian grouping: There is considerable disagreement as to what larger groupings the Costanoan (including Mutsun), Miwok, Yokuts, Wintun, and Maidun languages may form. One hypothesis groups all of these languages together as a Penutian family (with Costanoan and Miwok as the Utian language families and Yokuts, Wintun, and Maidun less closely related to each other), but it is not clear that the Pen and the Utian branches are related. There have been a wide variety of specific hypotheses as to which of these language families are genetically related into what groups. Because there are no living speakers of any Costanoan language, and the archival documentation of Mutsun is so vast, large searchable lexical and text databases of the Mutsun language could provide an important piece of the puzzle for the Penutian question. There has been considerable scholarly interest in the reconstruction of Costanoan and larger groupings including it (e.g. Adams 1985, Beeler 1955, Callaghan 1997, 1998, 2001, 2003; Okrand 1979).

An example of how work on the database can answer questions of historical reconstruction is the recent clarification of the suffix *–way*. Okrand (1977:160) listed this suffix in his grammar, but he only found two attestations of it in the data he had (*misnisway* ‘Sunday’ and *piinaway* ‘therefore’). Since *misnis* is a borrowing from Spanish *misa* ‘mass,’ and *piina* means ‘that,’ Okrand tentatively proposed that *–way* meant ‘on account of.’ Recently, the graduate research assistant (GRA) on the project found three additional attestations of *–way* in the Harrington notes: *Taalaway* ‘summer,’ *tuurisway* ‘winter,’ and *ammaniway* ‘rainy season.’ *Taala, tuuris* and *ammani* were already well attested as meaning ‘heat,’ ‘cold,’ and ‘rain,’ respectively. Thus, *-way* must have meant ‘the time of ____,’ a meaning substantially different from that ventured by Okrand, but still consistent with his data. This suffix fills a gap in historical linguists’ knowledge of the Costanoan languages in general: Callaghan (2003) is unable to reconstruct a time suffix for Proto-Utian, Proto-Costanoan, or Proto-Miwok, although she does reconstruct such a case for Proto-Eastern-Miwok. The Mutsun suffix *–way* fills this gap, suggesting that some sort of temporal case marker may have existed in Proto-Utian and Proto-Costanoan as well.

b. Imperatives: Mutsun has very unusual command forms for its verbs. The form of the imperative suffix depends on both the number of the subject (the addressee) and the person of the object of the imperative verb. The normal singular imperative suffix (for commanding one person to do something) is *–y*, as in *mehey* 'look!'. The normal plural imperative is *–yuT,* as in *meheyuT,* 'you all look!'. However, if the object is the first person (me/us), entirely different imperative suffixes are used, *-t* for a singular addressee and *–m* for plural addressees [e.g. *mehet kannis* 'look at me!' (*kannis* 'me') and *monsem kannis* 'you all tell me!' (Okrand 1977)]. Thus, the number of the subject but the person of the object is marked, while the number of the object is not marked. This pattern is unusual in the languages of the world, and it is also surprising in Mutsun because the non-imperative verbs are not marked for person or number at all. There are further complications involving objects of imperatives: for third person objects (e.g. 'look at him/her/it'), there is a special pronoun for use only in imperatives, *nuk*. However, if a full noun phrase (e.g. 'look at the mountain!') is the object of an imperative, the object suffix is omitted, and the noun is not marked as an object at all. In non-imperative sentences, noun phrase objects do take an object suffix. All of these special forms for objects of imperatives suggest that there is something unusual in how imperatives and objects are viewed in Mutsun. Callaghan (1998) reconstructs the imperative suffixes for the Proto-Utian language family, but there is much more to be done to understand these unusual patterns of objects of imperatives. Heidi Harley, a specialist in morphology, syntax, and semantics in the PI's department, has expressed interest in collaborating on this topic once the full database is available.

c. Inherent reflexives (mediopassives): Mutsun has a suffix *-ni* which makes verb stems into mediopassives (inherent reflexives), giving them a meaning of change of state, intransitivity, etc. For example, *sacki* 'to split something in half (transitive),' when *–ni* is added, becomes *sackini* 'to split (intransitive), to get split.' (All examples are from Okrand 1977 unless otherwise specified, but are converted to the writing system developed by the PI and Ms. Luna-Costillas, who heads the community language revitalization movement.) There are many other such pairs, as well as verbs which occur only with the *–ni* suffix or only without it. What makes this interesting is that Mutsun seems to go further than other languages in its use of this type of morpheme. In addition to verb categories which cross-linguistically are often mediopassives (inherent reflexives, e.g. change of state, non-agentive events, weather verbs, motion verbs), there are a few unexpected verbs marked with *-ni* in Mutsun (e.g. *saawe* 'to sing,' *saaweni* 'to sing,' with unclear difference in meaning). In order to investigate the extent of use of the mediopassive (inherent reflexive) in Mutsun and determine which classes of verbs can use it under what circumstances, a complete searchable database is necessary. Once this database is completed, one can create a list of all occurrences of the morpheme *–ni*. Then a thorough analysis of its contribution to the meaning of the verb and of the conditions on its use could be conducted, adding to our knowledge of the typology of mediopassives (inherent reflexives), as well as to our knowledge of how to use specific forms in Mutsun. Heidi Harley has also expressed interest in collaborating on this study.

d. Stress and vowel length: Stress and distinctive vowel length are often confused in the archival materials on Mutsun. Okrand (1977) concluded that Mutsun had distinctive vowel length, and that the presence of stress was predictable. For example, *raraS* 'finger' and *raaraS* 'molar tooth' show that vowel length is contrastive. Okrand tentatively concluded that primary stress falls on the second syllable if the word has two syllables and the first is light (i.e. consists of only a consonant and a short vowel), and that in all other cases, stress falls on the first syllable. That is, stress falls on the first syllable if the word has three or more syllables (and of course also if it has only one syllable), but it also falls on the first syllable if the word has two syllables and the first is heavy. However, Okrand believes that even the Harrington notes, which provide the most accurate transcriptions (cf. Appendix 2), sometimes confuse long vowels with stress. The stress pattern as Okrand identifies it is typologically very surprising. It is not unusual to find first syllable stress in all words, nor is it unusual to find second syllable stress if the first syllable is light and first syllable stress if it is heavy. However, it would be quite unusual to find sensitivity to syllable weight appearing only in two-syllable words. With the completion of the database, it will be possible to analyze thoroughly how stress and vowel length are recorded in the various archival sources. This would allow the PI to clarify what the stress rule actually was in Mutsun, and provide further evidence about whether stress, vowel length, or both, were distinctive. If the complete set of evidence supports Okrand's generalization about stress, this would provide interesting typological evidence about stress with implications for phonological theory. A phonologist in the PI's department, Michael Hammond, has studied stress systems extensively and has expressed interest in collaboration in this area.

e. Metathesis: Mutsun displays an unusual metathesis process, in which sounds switch places depending on preceding or following environment (Okrand 1977). For example, the plural suffix appears as *–mak* after a stem ending in a consonant, but as *–kma* after a stem ending in a vowel, and the locative suffix appears as *–tak* after a consonant and *–tka* after a vowel. Many verb stems, such as *kapla* 'to embrace,' show similar alternations (e.g. *kapalmu* 'hug each other'). In some cases (such as *–mak/kma*) the order of the two consonants changes, as well as the order of the final vowel and consonant, while in others, only the final vowel and consonant reverse. This small set of data on suffix alternations has attracted attention in the field of formal phonology (Mielke and Hume 2001, Blevins and Garrett in press). In Mutsun, such metatheses are highly productive and regular in the suffixes and some classes of verb stems, but also appear less regularly in other related forms, such as *paakuc* 'ball (noun)' vs. *pakcu* 'to play ball,' or *uTit* 'two' vs. *uTtin* 'Tuesday.' Although the verb stem and suffix alternations are relatively well understood (Okrand, 1977, 1979), the less regular alternations are not. These are particularly important for linguistic theory, because this type of metathesis (making a distinction in part of speech) has been claimed to be rare or unlikely in language in general. Metathesis in general is a relatively rare sound alteration, and it has attracted a great deal of interest in formal phonology. Completion of the database would provide a wealth of information on this unusual metathesis pattern, which is of interest for theoretical phonology as well as for community members.

f. Language attrition: The databases will provide a valuable resource for the study of language attrition (the study of how a language changes as it becomes endangered). The text database will include both the Arroyo data, elicited in approximately 1815 when there were many fluent Mutsun speakers, and the Harrington data, elicited in 1922 and 1929-30 from the last fluent speaker (see Appendix 2). Furthermore, Harrington re-elicited the Arroyo data, so it is possible to compare directly how each sentence was produced at a time when the language was widely spoken and at a time when it was severely endangered. Because of the quantity of data and because of Harrington's re-elicitation of the same material, this is a very unusual resource.

g. Other topics: The material on Mutsun that has been analyzed thus far [particularly the Okrand (1977) work] provides hints at several more topics of interest to theoretical linguistics, but Okrand lacked enough data to analyze these topics thoroughly. There is evidence of an unusual reduplication pattern (repetition of part of a word to make another word). Reduplication is a major topic in formal phonology. There are also a large number of nominalizer suffixes (suffixes which make verbs into nouns), which make very interesting semantic and syntactic distinctions. Completion of the database will allow the PI and other linguists to easily compile all the data on these and many other topics. Without a searchable database of the archival materials, it is impossible to work on these topics, because there is no way to know where in the voluminous archival data occurrences of a particular suffix or word might be located.

D. Importance of the project for the Mutsun community

There is a highly active language revitalization movement underway in the Mutsun community. Community members (particularly Ms. Luna-Costillas) initiated this movement in 1996, and the PI became involved in the project in 1997. The Mutsun people have an extremely strong desire to bring their heritage language back into use. Language is an important part of culture, and the Mutsun people want very much to be able to speak the language of their ancestors. They feel that this is an important part of maintaining their identity within the larger American society. For Native American groups that are not well known, such as the Mutsuns, it is particularly useful to have the language as something to show who they are.

Ms. Luna-Costillas and the PI have already developed a range of teaching materials for use in community language lessons. (See history of the project below.) However, our linguistic understanding of the grammar, pronunciation, word meanings, etc. of Mutsun is incomplete. The PI has often found that an entry in the dictionary or usage in the textbook had to be revised based on knowledge gained from another archival source after the initial writing. This is unfortunate in that community members may struggle to learn a particular form or grammatical construction, only to find that they have to re-learn it differently later. (The community is well aware that the revitalized form of their language will differ from the way their ancestors spoke Mutsun. The PI has discussed this issue with community members at length. They have decided that they would rather use their ancestral language, even though it will change in the use of it, than allow it to remain extinct.)

Completion of the lexical and text databases, with all archival materials entered into them, would greatly improve this situation. The lexical database is used to generate the dictionary directly, so improvements in the information in the database convert readily to better learning materials for the community. With the text database completed, the PI could easily generate concordances listing all occurrences of particular morphemes, and could thus clear up a great many incorrect entries by comparing all the forms of a morpheme. For example, using this sort of method with the currently available data, the PI was recently able to determine that the basic form for the verb ‘to play’ should be *rorSo* rather than the form *rorois* that was recorded by Arroyo. There is an attested form *roroSpu* ‘amuse oneself’ (with reflexive suffix –*pu*) and Arroyo often spells *is* for *S*. Combining this with the knowledge of metathesis in verb stems discussed above, and considering the facts that Mutsun rarely has two consecutive vowels, and that the basic form of verbs always ends in a vowel and nearly always consists of two syllables (Okrand 1977), it is possible to conclude with fair certainty that the basic form must have been *rorSo*, even though no attestations of that form have been found in the materials examined thus far. This sort of analysis has been done for many of the morphemes of Mutsun, but the as yet unanalyzed Harrington data would clarify far more phonological problems as well as discrepancies in meaning. A further example of how analysis of the archival materials is useful to the community is the suffix –*way* described in II.C.a above (cf. Okrand 1977:160). There is no other known Mutsun word for the general concept of time, so the discovery that –*way* means 'time of something' provides a way for the community to express a very useful concept in Mutsun.

Completing this work within a relatively short time frame (the two years of the proposed project) would be of enormous benefit to the community, because it would mean that any Mutsun community member studying the language after Fall of 2007 would not need to relearn corrected forms. If the proposal is not funded, it will take many years for the PI and students working for independent study credits to complete the entry of archival information into the database, during which time the community will continue working with incorrect materials. (Furthermore, having one research assistant and the PI work intensively on the database would ensure much more consistent analysis than having a large number of independent study students working one semester each on the project over many years.) Of course, since there are no living speakers to consult in order to supplement the voluminous but finite archival data, some uncertainties and gaps will always remain. However, once we know that all of the archival information about the language is in the database, we can make decisions about how to teach the language, how to borrow or create new words, etc., and know that a documented answer will not appear in another archival source later.

Finally, it is important to note that a disproportionate number of members of the Mutsun community are living in poverty, as is the case for many Native American groups. Native Americans are underrepresented in higher education, and the Mutsuns are no exception to this. Learning their heritage language engages Mutsun community members in linguistic analysis and gives them greater motivation than traditional school subjects do to work toward a scholarly pursuit. In the long term, the teaching of the heritage language could have a positive impact on the community by increasing young community members' dedication to educational pursuits.

E. Importance at the intersection of linguistic research and community use

Until relatively recently, linguistic fieldworkers usually recorded information about languages with little thought to giving anything back to the community that spoke the language, beyond payment to speakers and the fact that the language was archived. More recently, field linguists have begun to give back to the communities they work with by providing dictionaries or language teaching materials. However, for a great many Native American languages such as Mutsun, there are no living speakers of the language, but there are large quantities of archival materials from earlier linguists documenting the languages. In most cases, these old archival materials are now lying unused in libraries. Very few linguists study these tremendous quantities of data, because of the difficulty of working with these materials. Furthermore, these materials are inaccessible to almost all of the Native American communities from which they were recorded (although see below for successful cases). Thus, this resource about an otherwise lost part of Native Americans' heritage is languishing unused, even though it would be of use both to linguistics and to the communities from which it came. The current project represents a relatively new methodology in linguistic research. The combination of skills of working with old texts, as is usually done in historical linguistics, and direct work with community members, as is usually done in fieldwork, leads to an interesting fusion of linguistic methodologies. This language revitalization work also advances linguistics beyond documentation of endangered languages to the next step, using linguistic knowledge to bring languages back.

This work benefits linguistics graduate students by training them in a novel combination of archival and community methodologies. They also gain an appreciation for the usefulness of linguistics in the larger world. This work benefits the field of linguistics, in that it advances a new type of scholarship and linguistic research. Furthermore, as language revitalization progresses and individuals in the community begin to acquire competence in the previously extinct language, this will lead to a wealth of fascinating topics in the area of language acquisition. Acquisition of a language in the absence of fluent speakers is very different from acquisition either through immersion or through a more typical classroom situation. As fluent speakers of a previously extinct language begin to emerge, there will be the possibility to study how the language changes as it becomes fluently used, and how speakers develop intuitions about structures of the language in the absence of much input data. Although creating fluent speakers might seem like a far-off goal, this has been accomplished on a small scale already in at least one other Native American language, as described in the next subsection, and the PI, the GRA, and the community leader are already semi-fluent in Mutsun now.

F. Relation of the proposed work to other works in the field

The proposed project has similarities with several other types of linguistic work. First, the process of converting archival or published hard copy materials on a language into a searchable electronic format for future linguistic research is being performed for other languages [e.g. Hyman's Comparative Bantu Online Dictionary (http://linguistics.berkeley.edu/lingdept/Current/research/cbold.html), Bird et al. 2002]. In most such cases, however, the older materials to be included in a database are already in a more analyzed form than the raw fieldnotes we propose to analyze. For example, dictionaries available only in hard copy might be scanned and entered into a database. In the case of Mutsun, this has already been done for the only published dictionary [Mason (1916), resulting in the dictionary edited by Warner (1998)]. Other "easier" sources, specifically all the Mutsun words occurring in the Okrand grammar (1977) and the word lists of C. Hart Merriam (1902) have also already been entered into the database. (See Appendix 2 for description of sources.) What remains to be done is to enter the information available in the largely unanalyzed fieldnotes, and in the original Arroyo work. This is therefore a related but different approach as compared to most other electronic dictionary projects.

In a few cases, extensive analysis of the archival fieldnotes of an extinct language has been performed. Costa (1994) has done this for the Miami language, and has since done similar work on Mohican. The Okrand (1977) grammar of Mutsun is another such example. In such work, old written records of now extinct languages are analyzed to determine the grammar and lexical items of the language. Often, the old records are Bible translations or other texts, with little or no linguistic analysis included. In some cases [e.g. the Okrand (1977) work at the time it was done], this analysis is carried out purely for its linguistic interest. Native American languages are of great interest for typological studies, and the additional data on a particular language provided by analyzing archival materials of an extinct language can be very useful for theoretical investigations. In other cases (such as Costa's work on Mohican), such studies are carried out primarily to assist the community in language revitalization efforts, with a secondary benefit for linguistic theory. However, the number of languages for which massive archival documentation remains unanalyzed and out of use is far greater than the number for which archival data has been analyzed either for theoretical or applied purposes.

The Okrand (1977) work is the most closely related to the current project, since it is for the same language. However, at the time of his 1977 work, Okrand had access to only a small portion of the Harrington fieldnotes (he estimated 2-3%), which are the most important source. He did not have access to or chose not to use the smaller corpus of C. Hart Merriam notes and the unpublished Arroyo, and he made minimal use of the published Arroyo (1861, 1862). Since the time of Okrand's work, the Harrington notes have been collected at the Smithsonian, organized, and made available on microfilm. The availability of the microfilm allows us to analyze the entire set of Harrington notes on Mutsun. (Okrand worked with additional Harrington notes after completing his 1977 grammar, and the project will also make use of his unpublished notes from that work.) Another important difference from the Okrand (1977) work is that Okrand produced a Mutsun grammar but not a searchable electronic database or a dictionary or any materials for community use. Okrand's work was done before the language revitalization movement in the Mutsun community began and before searchable databases were common. The hard copy dictionary and annotated text collection to be published through the current proposal are related to a separate type of work, namely dictionaries and text collections on minority languages worldwide. This type of publication, along with grammars [which Okrand (1977) already provides for Mutsun], is widely used by theoretical and typological linguists in many subfields of linguistics (at least phonetics, phonology, morphology, syntax, semantics, and historical linguistics). Such materials provide a significant proportion of the linguistic data for more theoretical or comparative investigations. Because of the many typologically interesting features of Mutsun discussed above, a dictionary, searchable database, and text collection for Mutsun would be of great use. With the exception of Okrand's excellent grammar, the previous published works on Mutsun (Mason 1916, Arroyo 1861, 1862) are inadequate for the purpose of providing data for further linguistic investigations because of inconsistencies and inaccuracies in transcription and the lack of an organized presentation of data and analysis.

There have been a few related cases of language revitalization from archival materials. The most successful case thus far is Myaamia (Miami), which has gone from being extinct to having a small number of nearly fluent speakers and a large number of community members participating in language classes and workshops. Costa's (1994) work analyzing archival materials on Myaamia was an important step in this process, but Baldwin, the first speaker of Myaamia in decades, has taken the project from analysis to successful revitalization (Hinton 2001a, Baldwin 2003). Costa is now working on analysis of archival materials and development of teaching materials for Mohican, which has been extinct for longer than Mutsun, and Mohican community members are holding language classes and beginning to use the language. Such successful cases are extremely encouraging to other communities engaged in language revitalization, such as the Mutsuns. These successes demonstrate that the daunting task of creating fluent speakers for a language that is not currently spoken is indeed possible.

Finally, the current proposal is related to the Harrington Project at the University of California, Davis, headed by Martha Macri. That project is engaged in transcribing all of Harrington's notes (on any language) to make them available for analysis. That project has developed a system for converting Harrington's notes to a standard character set, since Harrington used a great variety of non-standard symbols. Although the UC Davis Harrington Project facilitates the currently proposed project by providing better access to some of the Harrington notes, the Davis project does not involve in-depth analysis of the language, only transcription of Harrington's notes. Furthermore, the Davis project has, as of yet, not been able to transcribe very much of the Mutsun portion of the notes, so it is anticipated that only a small portion of the transcription will be done by the beginning of the proposed NEH project period.

III. History, scope, and duration

A. History

This project has been underway since 1996, when Ms. Luna-Costillas attended the first Breath of Life Language Revitalization workshop at the University of California (Hinton 2001b). This workshop introduced Native Californians whose languages had no living speakers to the archival materials on their languages available at the University, and helped them learn to use these materials. Ms. Luna-Costillas continued working to learn Mutsun on her own after the workshop. The PI became involved in the project in 1997, as the student mentor working with Ms. Luna-Costillas at the second workshop. Since then, the PI and Ms. Luna-Costillas have continued to work together long-distance and through occasional visits.

The PI and Ms. Luna-Costillas together developed a practical orthography for Mutsun. This writing system encodes the sounds of Mutsun unambiguously, but uses no special characters, making it convenient for e-mail and database use. It has been tested with Mutsun community members, and revisions have been made to increase its learnability. All new work in Mutsun and the examples in this proposal are written in this writing system. At the 1997 workshop, the PI and Ms. Luna-Costillas, together with another Mutsun community member, also translated a traditional Mutsun story that had only been recorded in Spanish back into Mutsun.

After the 1997 workshop, the PI scanned in the only published dictionary of Mutsun (Mason 1916), reversed it to create an English-Mutsun version to accompany the Mutsun-English version, and also converted it to the practical orthography. The result (Warner 1998) allowed look-up of words from English as well as from Mutsun, which was formerly the only option. This was a great improvement for community use, but the PI realized while doing this how important it was to compile the information from the other archival sources, because the Mason and Arroyo works are rife with inaccuracies in transcriptions and meanings.

In 1999, the PI and Ms. Luna-Costillas wrote a partial draft of a Mutsun language textbook for community use. From 1998 to 2001, the PI continued work on the dictionary. The PI developed a lexical database to replace the previous list-style dictionary, using the Linguist's Shoebox software (SIL 2000). She entered all the data in the Okrand (1977) dissertation and the Merriam fieldnotes into the database, along with all the data from the Mason dictionary. This database format allowed for addition of example sentences to demonstrate usage, analysis of alternate transcriptions, and correction of errors in the materials that became apparent through electronic sorting. The PI also added morphological information about alternating verb stems, vowel lengthening and shortening, etc. Specific aspects of the database (which is the one still in use now) will be exemplified below. During this time, the PI held a position to which this project was not closely related, so she worked on this project on her own time without compensation. At the same time, Ms. Luna-Costillas continued to develop additional teaching materials for Mutsun, both independently and with the help of the PI and of other Mutsun community members.

In 2001, the PI moved to the University of Arizona Department of Linguistics, which has a strong emphasis on Native American and endangered languages. (The department's strength in this area, as well as its dedication to involving Native Americans in the study and maintenance of their own languages, are evidenced by the department's Masters in Native American Linguistics program, the joint Ph.D. in Anthropology and Linguistics program, and the research of several faculty members. The department provides a very supportive atmosphere for the proposed work.) In 2002, the PI completed a new version of the dictionary, which includes all the information in the Merriam, Okrand, and Mason sources (Warner 2002). This 261-page, professionally formatted dictionary has been in use in the Mutsun community since then. In 2001, Ms. Luna-Costillas began teaching Mutsun language classes within the community, using the draft textbook and additional teaching materials. In 2004, Ms. Luna-Costillas, the PI, the GRA, and additional community members organized a community language workshop for approximately 40 people. Neither Ms. Luna-Costillas nor the PI nor the GRA is a fluent speaker of Mutsun, but they are now semi-fluent, and it is a testament to Ms. Luna-Costillas' dedication and to the desire of the community for their language that she is able to teach successful language courses at this stage of the revitalization process. Ms. Luna-Costillas has never received monetary compensation for her work on this project.

In 2001 the Mutsun community established the Mutsun Language Foundation for the purpose of raising funds to support language and culture revitalization within the community. This foundation raises money through sales of traditional crafts and t-shirts at festivals as well as through grant writing. Such funding allowed Mutsun community members to attend the Breath of Life Language Revitalization workshop in Berkeley in 2002 and 2004, increasing the number of community members with direct experience in language and archival analysis. The foundation also funds workshops within the community. In 2004, a language committee was formed within the community to create a small group of dedicated learner/teachers who can advance the revitalization process more quickly. Since 2002, several community members have also worked as volunteers for the UC Davis Harrington Project, typing portions of Harrington's notes on Mutsun. The proposed project will have access to the notes they have already transcribed.

Since 2002, four graduate students and three undergraduates in the Linguistics Department of the University of Arizona have worked with the PI on the Mutsun database for independent study credits or as volunteers. These students have been very helpful in improving the lexical database, designing the text database, beginning to enter the Harrington notes and original Arroyo materials into the database, and writing teaching materials. The text database, developed with the help of these students, also uses the Linguist's Shoebox software, particularly its automatic parsing function (demonstrated in IV.D below).

In 2002, the PI and Ms. Luna-Costillas began exploring the use of distance-learning software for teaching Mutsun. Because the Mutsun people are scattered across a wide area in California, holding language classes in person is difficult. The fact that the PI is located in Arizona also makes frequent long-distance communication necessary. The University of Arizona provides distance learning software which allows users to post audio and text messages and language lessons, as well as hearing others' speech and responding to it. This part of the project is still at an exploratory stage, but it is anticipated that distance-learning software will greatly enhance community use and learning of the language by reducing geographical limitations. The PI, GRA, and Ms. Luna-Costillas have successfully used this software to record several Mutsun language lessons. This interface is compatible with all major computer platforms and requires only Internet access and a microphone. This interface will be an important tool for allowing scattered Mutsun community members to practice the language on their own schedule.

In 2003, the PI and Ms. Luna-Costillas received a Woodrow Wilson Foundation Public Scholarship Partnership grant to continue their work on revitalization of Mutsun. This grant was active during the 2003-04 academic year, and funded one semester of research assistantship to begin large-scale work on the Harrington notes. It also funded two trips for the PI, Ms. Luna-Costillas, and the GRA to work together in person. Through this grant, the text database has been expanded from less than 100 to more than 5000 entries, most of which are from the Harrington notes. However, this significant amount of work represents only a relatively small portion of the Harrington notes; the enormous task of extracting all the information from this source (estimated at 36,000 pages of microfilm on Mutsun), as well as from the published and unpublished Arroyo materials, will require far more than one semester of a research assistant's time. Funding for the remainder of this work is requested in the current proposal.

B. Scope and duration

It is anticipated that at the requested funding level, work on all the archival materials will be completed within the two-year NEH project period. Products of the project for the NEH grant period are described in II.A above. By the end of the NEH project period, the theoretical linguistic issues for future research will be more clearly defined based on the knowledge gained during the analysis of archival materials. Therefore, still during the NEH project period, it will be possible to submit a competitive grant proposal to NSF to support further linguistic analyses using the finished database. Thus, the PI has plans for the long-term sustainability of the theoretical component of the project. (See IV.G below on technical sustainability.) As for the community-oriented aspect of the project, the Mutsun Language Foundation will continue to raise funds for language and cultural revitalization. The PI and Ms. Luna-Costillas are also personally committed to continuing work on the project indefinitely, regardless of funding status. The community-oriented aspect of the project is truly open-ended, because continued development of improved teaching materials, classes, and workshops will be necessary until the optimistic long-term goal of a self-sustaining community of speakers who acquire Mutsun as their first language is reached.

C. Publications

The work on Mutsun thus far has focused on creating unpublished materials for immediate community use, rather than on academic publishing. Several works, listed here, are now in circulation within the Mutsun community.

Warner, Natasha, ed. 1998. *English-Mutsun Mutsun-English Dictionary, Based on the Work of Felipe Arroyo de la Cuesta and the Dictionary of J. Alden Mason.* Unpublished manuscript. [109 pages.] [Replaced by Warner (2002), listed below.]

Warner, Natasha, and Luna, Quirina. 1999. *Mutsun riicase ursen ennekmin. [A Mutsun language learning book.]* Unpublished manuscript.

Warner, Natasha. 2002. *English-Mutsun Mutsun-English Dictionary.* Unpublished manuscript. [261 pages.]

Luna-Costillas, Quirina, Luna, Monica, Luna, Genevieve, and Warner, Natasha, translators. 2002. *TcutsuSmin moTese yuu tooTese.* [*Green Eggs and Ham*, by Dr. Seuss.] Unpublished manuscript.

IV. Methodology and standards

A. Use of a standard database and standard annotation methods

The project uses the “Linguist’s Shoebox” database software (SIL 2000), which was developed for organizing language fieldwork data and creating an easily usable dictionary. This choice of software imposes a standard system on the structure of the lexical database. Without altering the Shoebox structure, some adaptations have been made for archival data. Because the Shoebox software was designed explicitly for linguistic fieldwork on minority languages, it provides a very useful format. Furthermore, Shoebox data files are simple text files with a clear structure to their fields. These files can be easily searched and edited, allowing one to make various changes and import the result back into Shoebox for further processing. Shoebox also produces a well-formatted dictionary for hard copy distribution or publication.

The PI is aware that the Linguist’s Shoebox software is not being further developed, and that it has been replaced by SIL’s new product LinguaLinks. However, because LinguaLinks lacks some of Shoebox's features for dictionary production and formatting, we are not currently planning to convert to it. In this proposal, funding is requested to hire a programmer to develop several programs to accompany Shoebox. Although the text database in Shoebox has an excellent automatic parsing function that locates morphemes in the related lexical database, the program does not retain information about which morphemes appear where in the texts. One consequence of this is that as new information becomes available and entries in the lexical database are modified, there is no easy way to update every entry of the text database that contains the modified word. An important task for the programmer will be to create a system for updating the two databases consistently in parallel. A further task for the programmer is to develop a searchable but non-editable version of the database for distribution.

The Shoebox database is not in itself a standard format, so once all information has been entered into the Shoebox lexical and text databases, the databases will also be converted to an XML version, following the best practices for annotation of language data being developed by the EMELD (Electronic Metastructure for Endangered Languages Data) project. Terry Langendoen, one of the leaders of the EMELD project, is a faculty member in the same department as the PI. The presence of the EMELD project within the department is a great resource for the Mutsun project. Use of the EMELD standards will keep the databases accessible as technology changes, because of the XML basis of the encoding. Furthermore, since the EMELD annotation provides standard terms for linguistic categories and links these to the terms used by a particular researcher, conversion to the XML version will solve a long-standing problem for the Mutsun dictionary, namely the balance between the needs of the community and the needs of researchers. For community use, it is best to gloss grammatical terms with English words where possible (e.g. “us” instead of “first person plural objective pronoun,” or “go to do something” instead of “andative”). Conversion to an XML version will allow linguists to search for standardized linguistic terminology, while translations more accessible to the community can also be maintained. The EMELD project is currently developing a tool for performing such a conversion, and has recently obtained permission from the PI to begin converting the Mutsun text database from Shoebox to XML as a pilot project. This conversion and related work is being reported at the 2004 EMELD workshop on linguistic databases in Detroit (Basham et al. 2004). Although XML and EMELD provide the advantage of a standardized format, the project will still need the Shoebox format to produce the dictionary and parse the texts. The XML format cannot accommodate this need.

B. Criteria for selection of materials

The goal of this proposal is, within the project period, to include in the lexical database information about every occurrence of every morpheme in any Mutsun archival source and to include every recorded sentence in the text database. That is, all materials will be included. Because the language is extinct, this is a finite, if large, quantity of materials. Absolute completion of the database will be useful for linguistic research, and will be important for community use, because this will allow community members to know when they should move to creating or borrowing new words instead of waiting for further information from the archival data.

C. Content, form, and length of entries, with examples

The content and form of entries in the Shoebox version of the databases is best demonstrated by examples (below and in Appendix 3). The XML version of the databases will differ in the means of annotating the grammatical categories.

Sample lexical database entry (Shoebox version):

```
lx neppe
ps dem
ge this
re this
xv wattin-ka neppe rukkatkatum.
xe I go away from this house.
xv makkese neppe uTTasi.
xe This one cares for us.
oe close by; can modify a noun or stand alone as a pronoun
cf niSSa
ce this (farther)
cf piina
ce that (more distant)
va ne, nane, nina, nemis, nenis, unta, ister, nepper, nepe
ve Ma/Ar
pdl Obj.
pdv neppes, neppese
pde this (object of sentence)
pdl Instr.
pdv neppesum
pde by means of this
pdl Attrib.
pdv neppewas
pde of this
nt variant nee has form ne
np O, Me
ns NW 11/02, LB 10/03, LB 11/03
so 137, 153, 158, 183, 184, 185, 186, 317, 323, 40c, 53c
dt 23/Dec/2003
```

The lx field shows the main form of the lexeme. ps shows the part of speech (here, demonstrative). ge gives the translation(s), and re gives a version used for making the English-Mutsun (reverse) dictionary. xv, xe pairs give example sentences with English translations. oe describes restrictions on meaning or grammar. cf, ce pairs refer the reader to related entries and give a brief gloss for those entries. va, ve pairs list variant transcriptions of the item (converted to the practical orthography, but otherwise maintained in original form), and show which source (Mason=Ma/Ar, Okrand=O/H, Merriam=Me, or H=Harrington) the variant came from. /pdl, /pdv, /pde triples list any irregular grammatical forms of the word (pdv), along with the grammatical category of the irregular form (pdl) and the English translation of it (pde). Note the distinction between an irregular form (e.g. an exceptional object form) and a variant form (an unexplained alternate form that might be a mishearing or a dialectal difference). The nt field is for general notes for the researchers, not to be included in the final products. The np field shows in which sources the main form listed in the lx field is attested (here, Okrand and Merriam both have /neppe/). (For cases in which the PI reconstructs an unattested form as the most likely correct form, as for /rorSo/ 'to play' discussed in II.C, this fact is recorded by listing "C" for "constructed" as the source. Attested forms are listed as variants in such cases.) ns shows who has edited the entry and when. so lists the page numbers of sources in which the form appears, where they are not retrievable from the text database. The dt field shows when the entry was last edited. Additional fields not used for this entry show usage information (e.g. "rare" or "vulgar"), and what language the word is borrowed from if it is not native to Mutsun.

Corresponding Mutsun-English dictionary entry:
neppe *dem.* **this**. **wattin-ka neppe rukkatkatum.** I go away from this house. **makkes neppe uTTasi.** This one cares for us. *Restrict:* close by; can modify a noun or stand alone as a pronoun. *See:* **niSSa** 'this (farther)'; **piina** 'that (more distant).' *Obj.:* **neppes, neppese** 'this (object of sentence).' *Instr.:* **neppesum** 'by means of this.' *Attrib.:* **neppewas** 'of this.' *[Phon:* O, Me*]*.
The corresponding English-Mutsun entry is similar. These samples are drawn from the current database and dictionary, and it is anticipated that some aspects will be changed in the final product, particularly in the formatting of dictionary entries. The information about who has edited the entry when, about variant forms, and about source page numbers is omitted from the hard copy entry to improve readability, but it will be included in all electronic versions, and will thus remain available for research purposes.

Guidelines as to what information is encoded in which fields, and how to interpret the resulting lexical entries, are included in the introduction to the most recent version of the dictionary (Warner 2002). These guidelines to the organization of dictionary entries are written in non-technical language that is accessible to the community, but that still conveys the information a linguist needs. The length of a given entry varies, depending on how much information is available about the particular morpheme. In some cases, there are multiple variant transcriptions, example sentences demonstrating various usages, quite a few English translations, and cross-references to several related forms, along with source information. For less well-documented forms, there may be only a single translation and the part of speech, along with the source information. No lexical entry is likely to be much longer than *neppe* above.

Sample text database entry (Shoebox version):
```
id  77
idA 33
osA ¿Ara inthrisnane rotes?
t   ara  inTis   nane  rotes
m   aru  hinTise neppe roote -s
g   next where?   this  be_at -remote_past_tense
p   adv  Q        dem   v     -suff
f   Next, where was this?
otA ¿De veras te dueles de tus pecados?
nt  orig. trans. should have been on preceding sentence
ns  TG, NW 12/02
```

The id field simply gives a numerical count of records in the database. idA gives the sentence number in the Arroyo (1862) sentence list, and thereby shows that this text comes from the

Arroyo data. (Corresponding fields are used for identifying the reel and frame number for Harrington, Merriam, or unpublished Arroyo, or page number for Okrand.) osA gives the original spelling of the Mutsun as recorded by Arroyo (and corresponding fields give it for other sources). t gives a transliteration into the practical orthography but does not standardize to the main form of the lexeme (e.g. variant forms appear here) and does not separate polymorphemic words into morphemes. m gives the main (not variant) form of each morpheme, and marks morpheme boundaries. g gives glosses for each morpheme, and p gives parts of speech. f gives the free translation determined by the researcher entering the item. otA gives the original translation from Arroyo (and corresponding fields are used for other sources). nt gives any notes (here, the fact that the original translation is clearly a misprint and was intended for the preceding sentence in the source). ns shows who edited the entry and when. The process of determining the correct parse is illustrated for this sentence in subsection D below.
<u>Corresponding annotated text publication entry:</u>

```
Arroyo 33.
¿Ara inthrisnane rotes?
ara  inTis   nane  rotes
aru  hinTise neppe roote -s
next where?  this  be_at -remote_past_tense
adv  Q       dem   v     -suff
Next, where was this?
¿De veras te dueles de tus pecados?
Note: orig. trans. should have been on preceding sentence
```

All of the information in the database is preserved in the hard copy entry, except the identity of the people editing it (which will be maintained in the electronic version for distribution). Because we are just completing the testing phase for the text database, the format for its output is not finalized. However, the line providing the glosses is a standard linguistic format: glosses for each morpheme (g) aligned with the morpheme in the line above the gloss (m). The original spelling used by the linguist who collected the data is included, but a transcription in the practical orthography and a line giving the main form of the morpheme as listed in the lexical database are also included. Forms in the t field which differ from those in the m field appear in the lexical database as variant entries, with a reference to the main form. In the example above, *inTis* (originally spelled *inthris*) is a common Arroyo variant for the word Okrand confirms to be *hinTis*, and *rotes* is an Arroyo variant for *rootes*. (The former results from Arroyo's use of Spanish for transcribing Mutsun, the latter from his neutralization of the vowel length distinction.) Otherwise, little information about particular morphemes is provided with the texts, because the text publication is meant to be used together with the dictionary.

For the original spelling field of the text database, material extracted from the Harrington notes is transcribed using the standard developed by the UC Davis Harrington Project (see II.F above). Okrand's, Arroyo's, and Merriam's transcription conventions can all be converted to the standard symbol set with little difficulty. Transcriptions from all sources are converted to a phonemic representation in the practical orthography for entry into the lexical database, as well as the t and m fields of the text database. This is because the purpose of the lexical database is to encode the most likely correct phonemic form of each morpheme, while the text database maintains information about how each source transcribed the item. The transcription conversions will be described in the documentation of the products.

D. Editorial procedures

The undergraduate student employee will transcribe the archival materials into the original spelling, text, and original translation fields of the text database. That is, the undergraduate student will be keyboarding the material and converting to practical orthography, and will not be running the parser or otherwise analyzing the data. Because most of the archival material is in poor handwriting, having someone type the material will make the effort of other members of the project more efficient. This stage of the work requires training only in the use of microfilm and in the UC Davis Harrington Project's transcription system, and in a few basic aspects of database use. The undergraduate student employee will never edit the lexical database, and his/her work will always be checked when the GRA performs the parse of the sentences. The undergraduate employee will work on the parts of the Harrington notes that community members have not already transcribed in their volunteer work for the UC Davis Harrington project (see II.F), and will also work on the unpublished Arroyo materials and the unpublished Okrand slip files. Most of the materials are not legible enough for OCR. During the spring 2004 semester, an undergraduate independent study student has been doing this type of work quite successfully for the published Arroyo sentence list. When questions arose about conversion from original spelling to practical orthography, they were easily resolved by consultation with the PI or GRA. Since the parsing process requires the GRA to check the undergraduate's work, and every entry encodes the identity of the person(s) who worked on it, sufficient quality control for the undergraduate's work is built into the process.

The GRA will use the automatic parse function of Shoebox to generate an initial parse of each sentence the undergraduate has entered into the database. The GRA will then determine the accuracy of the automatic parse, based on knowledge of linguistics and Mutsun. General knowledge of linguistics is often helpful, because the automatic parser cannot tell which suffixes attach to nouns and which to verbs, for example. It also generates parses that do not match the translation of the sentence. (These problems are due to the incomplete information in the lexical database—if a new variant transcription or new word appears, it cannot be parsed correctly until it is added to the database.) The task of determining the correct parse also involves frequent reference to the existing lexical database, the hard copy dictionary, and Okrand's (1977) grammar. The GRA has already been working on the project for a year through the Woodrow Wilson Foundation grant (2003-04) and through independent study, and she has gained substantial knowledge about Mutsun morphology and parsing. Once the GRA has identified what she believes to be the correct parse of the sentence, she will add any new information to the lexical database and redo the automatic parsing until the correct parse is generated. When the GRA first began this work, she and the PI worked together closely, but she soon became highly skilled in parsing Mutsun and using the databases, and the PI now needs to consult only on the most problematic sentences. Checks by the PI have verified that the GRA is producing accurate parses. Both databases also include a field for recording who has modified an entry when. This system has been very successfully piloted during the period funded by the Woodrow Wilson grant. The knowledge of Mutsun, the database technology, and the archival materials attained by the graduate student in the past year, as well as her prior knowledge of Spanish, will allow rapid and highly reliable progress should NEH fund this project. If funding is not awarded, the graduate student will only be able to work on the project sporadically through independent study, and continuation of her work after she graduates will be contingent on another student with Spanish proficiency being willing to train and volunteer time to the project.

The PI will exercise final editorial control over the project. The PI will work closely with the GRA on analyzing problematic entries. The PI and the GRA will also refer to the original microfilms, so that when the undergraduate student has difficulty, the PI and the GRA will also be involved in reading the original handwriting. The PI and the GRA have found that when they work together, they are able to use their combined knowledge of linguistics, Mutsun, and Spanish, along with the information in the current lexical database, to be sure of the correct parse for nearly all sentences. Since there are no native speakers to consult, one might think that it would not be possible to know whether one had arrived at the correct parse. However, if 1) all of the grammatical words and suffixes in the sentence are known from the existing database, or they appear in several sentences so they can be analyzed, 2) nearly all the other words (i.e. content words) can be related to an existing database entry, 3) the translation of the parse roughly matches with the translation given in the source, and 4) there are no unexpected syntactic anomalies, then one can conclude that one has an accurate parse of the sentence. Whatever content words cannot be matched to an existing entry can then be identified as nouns, verbs, or adverbs based on the syntax of the rest of the sentence and the translation, and these unknown items can be added to the lexical database. Work on this project will begin with the Harrington data, because it is the most reliable and is closer to the phonemic transcription of the practical orthography, so it is unlikely to result in outright errors being added to the database.

As an example of this parsing process, consider the Arroyo sentence given above, with original spelling *¿Ara inthrisnane rotes?* and an incorrect original translation. The translation (paired with the wrong sentence, not a common problem) would make parsing this sentence very difficult, except that the morphemes in it are clear. First, the researcher converts the original spelling to the practical orthography, based on known patterns in Arroyo's spelling system, and enters *ara inTis nane rotes*. (The researcher must realize that *inthrisnane* is actually two words, by looking up similar forms in the existing dictionary if he/she does not already know these morphemes.) When this sentence was input, the form *ara* was already listed in the lexical database, but both as a variant of *aru* 'next' and as a variant of *hara* 'to give.' The form *inTis* was listed as a variant of two entries *hinTise*, one an adverb and one a question word, both meaning 'what(?), why(?), where(?).' The form *nane* was listed as a variant of *neppe* 'this' and also as the main form of a verb meaning 'to skip over.' *rote* was listed only as a variant of *roote* 'to be in a place.' *-s* was listed as the remote past tense suffix, and in separate entries also as a nominalizer and a question particle. The initial parse gave the researcher a choice of *aru* 'next' or *hara* 'to give' for *ara*, and the researcher must realize that since *rootes* is the verb, the sentence would be impossible to parse with an additional verb 'to give' in it, so the adverb *aru* must be correct. (*Aru* is also extremely common at the beginning of Arroyo's sentences.) For *inTis*, since Arroyo writes the sentence with question marks, the question part of speech is chosen. (Whether there is any syntactic difference between the adverb and question categories can be investigated once the database is completed.) On comparing to the lexical entry for the main form *hinTise*, the researchers noticed that all of Arroyo's forms for this entry end in *s* with no final *e*, and Okrand also has a variant form *hinTis* with no final *e*. Therefore, *hinTis* was made the main form, and *hinTise* was demoted to a variant entry. (*hinTise* appears likely to be the objective form of the word, and this information will be added to the database if more evidence is found.) For *nane*, the parser gives a choice of *neppe* 'this' or *nane* 'to skip over.' As with *ara* earlier in the sentence, an extra verb would not make sense, and *rote* can only be a verb, so *neppe* 'this' was chosen. Finally, *rote* is correctly parsed as a variant of *roote.* For *–s*, the nominalizer and question particle *–s* are very uncommon, and since there is already a question word in the sentence, the remote past tense is clearly the right choice. In this case, all the forms of the morphemes were already present in the lexical database because Mason analyzed Arroyo's published sentences and included most of the morphemes in them. When a form which is not yet in the lexical database occurs, the researcher must determine whether it is a new variant form of a known morpheme (based on phonological and semantic similarity), or a new morpheme, and must then enter the information about it into the lexical database.

The data in this project are not subject to obsolescence. Although the lexical database constitutes a dictionary, since there are no living speakers of the language and the current project will include all existing documentation of the language, the information entered by the end of the project will not be superseded by later information. In the long term, if a group of fluent Mutsun speakers develops, it would be appropriate to begin a separate project documenting their vocabulary and changes from the original vocabulary.

E. Media

The searchable but non-editable final version of the databases will be distributed on CD. The programmer for the project will develop software to allow searching of the encrypted database files. CD is a convenient means for distributing a large database, and since most modern computers have CD drives, this medium is accessible to most people with computers. The hard copy of the dictionary and annotated text collection will be published as a book through a publisher that publishes dictionaries of Native American languages. The PI has made an initial contact with the University of Arizona Press, which is a possible publisher.

F. Organization of and access to digital material

The search function the programmer will write for the encrypted electronic database products will be designed with both Mutsun community members and linguists in mind, and both groups will test it and will suggest changes to improve usability. The programmer will also write documentation to go with the CD, but the search function will be designed to minimize the amount of learning necessary. All products will have an introduction explaining all the information in the lexical and text entries. This documentation has already been written for the current dictionary, but it will be revised based on community feedback, and information about the annotated texts will be added. Documentation of the differences between the Shoebox and XML versions will also be included.

G. Storage, maintenance, and protection of data

The original source material will continue to be housed in the libraries it is currently in. The PI will store and maintain the editable versions of the databases on her computer indefinitely, and copies of the editable versions will also be archived with the Amah Mutsun Tribal Band, the University of Arizona library, and the University of California, Berkeley, library. Since Shoebox database files are text files, they are not expected to become obsolete with future changes in hardware, software, or media, but the inclusion of XML versions makes the databases even more robust to future changes in technology. As described above, the programmer will create a version of the finished databases which is encrypted to prevent editing, but which is searchable. Only that version will be widely distributed, as a means of protecting the data.

V. Plan of Work

The project for which this proposal requests funding will begin July 1, 2005. At the beginning of the project, the PI and GRA will train the undergraduate student assistant on Mutsun sound structure, the UC Davis Harrington Project’s transcription system, and how to work with the microfilmed Harrington data and enter the data into the text database. The GRA does not need to be trained, since she has been working on the project for a year. Shortly after the beginning of the project, the undergraduate student will begin entering sentences from the Harrington microfilms, the GRA will begin analyzing this data and updating the lexical database, and the GRA and PI will consult with each other on difficult cases.

This phase will continue for the first year of the project. During this time, the programmer will work on developing tools for use with the data, such as a spellchecker for Mutsun and database management scripts. (The spellchecker will improve the accuracy of data entry into the corrected form line of the text database, as well the Mutsun fields of the lexical database, such as the example sentence field. It will also be useful to the Mutsun community in the long term.) Data entry can commence, however, before such additional programs are completed.

During either the Winter Break of December 2005-January 2006 or the Spring Break of 2006, the PI, GRA, and undergraduate student will make a trip to California to work with Ms. Luna-Costillas and the Mutsun community in person. In the past, such in-person visits have been extremely productive for developing teaching materials and consulting with community members about structure of information and documentation. Furthermore, it is necessary for the PI and her students to work in-person with members of the Mutsun community other than Ms. Luna-Costillas in order to develop a strong relationship of trust with the broader community.

During the second year of the project (July 2006-June 2007), the undergraduate student assistant will complete the work of entering the archival materials, and will progress to assisting with the development of teaching materials and the final preparation of the hard copy dictionary and annotated text collection for publication. (The undergraduate student's role in that work will primarily be in evaluating documentation of the final products, proofreading, and formatting.) During the second year of the project, the GRA and the PI will finish analysis of the archival information and updating of the lexical database. Ms. Luna-Costillas will make a visit to Arizona to work in person with the PI and the students, and to become more familiar with other work on Native American languages at the University of Arizona. Early in the second year of the project, the PI will also submit a partial draft of the dictionary and texts to publishers. This draft will also be distributed through Ms. Luna-Costillas to a small number of Mutsun community members for evaluation and testing, in order to collect feedback about format, usability, and documentation.

By the beginning of May 2007, the PI and the GRA will be producing the final versions for distribution. During the second year, the programmer will work on developing the software for the searchable but non-editable distribution version of the electronic database. During June 2007, all work will be concluded, and the final version will be submitted to the publisher. During May or June 2007, the PI, GRA, and undergraduate student will make a second visit to the Mutsun community in order to work directly with the community on use of the final products.

The programmer's efforts will be concentrated more in the early part of the project, while the efforts of the other project participants will be distributed more evenly through the project period (except that all university-based participants will work more on the project during the summers than during classes). Although Ms. Luna-Costillas will not be working directly on entering data into the database, the time she spends working on the project is crucial. Ms. Luna-Costillas teaches the community language classes, develops additional teaching materials, publicizes the language revitalization project both within and outside of the Mutsun community, and applies for grants for further funding. Some of these activities provide information about how to format, document, and organize products for effective use in the Mutsun community. Some of Ms. Luna-Costillas' activities increase the size of the potential audience for the final products of the project. Some of her activities (e.g. giving presentations on Mutsun language revitalization at a wide variety of events) increase the general public's awareness of the importance of Native American languages as part of our cultural heritage.

When additional students work on the project through independent study, their efforts will be distributed between assisting the GRA with parsing Mutsun sentences, assisting the undergraduate student with entering archival data, developing teaching materials, and helping with editing of the final products. Their assignment to these tasks will depend on the skills they bring to the project and the needs of the project during the semesters they are available.

Schedule:

July 2005	Training of undergraduate student Programmer: develop data entry tools

July 2005- Summer 2006	Undergraduate: entry of archival materials GRA and PI: analysis of data from archival materials Programmer: Develop additional tools (e.g. spellchecker)
Dec./Jan. 2005-06	PI and students visit Mutsun community in California (1 week)
August 2006- January 2007	Undergraduate: finish entry of archival materials, develop teaching materials and help prepare final versions for publication GRA and PI: continue work on archival materials, submit draft to publishers, distribute draft to subset of Mutsun community for testing
Fall 2006	Ms. Luna-Costillas visits Arizona (1 week)
January 2007- June 2007	Undergraduate: Continue developing teaching materials and preparing final version for publication GRA and PI: Finish work on archival materials and continue preparing final version for publication. Programmer: Develop encryption and searching software for CD
May/June 2007	PI and students visit Mutsun community in California (1 week)

VI. Staffing

The project staff consist of the PI, a graduate student research assistant (GRA), an undergraduate student research assistant, Ms. Luna-Costillas (Mutsun community leader), a programmer, and graduate and undergraduate students working on the project for independent study credits rather than wages.

The PI, who is a faculty member of the Department of Linguistics at the University of Arizona, will devote at least one quarter of her research time to the project during both academic years of the project. Since her appointment is for 40% research, this amounts to 4 hours per week of effort during the academic year. During the summers, the PI will devote one full-time month of effort toward the project. The PI's duties on the project include training the students in all aspects of the work, communicating with the programmer, testing the programmer's programs, working with the GRA on correctly parsing the Mutsun texts and updating the databases, communicating with Ms. Luna-Costillas about the needs of the community, writing the documentation of the dictionary and annotated text publication, and editing the final products of the project. In one year, the PI will also teach a graduate seminar on language revitalization from archival materials, which will be closely related to the project.

The PI has excellent qualifications for this work. The PI holds a Ph.D. in linguistics, and began working on this project under the guidance and training of Leanne Hinton (UC Berkeley), a top expert in language revitalization methods. The PI has already been working on the long-term project for seven years. The PI developed the lexical and text databases, and has entered the information from several sources on the language into them already, and generated a dictionary from the database. The PI, together with Ms. Luna-Costillas, wrote the draft textbook in current use, as well as developing the practical orthography for Mutsun. The PI, Ms. Luna-Costillas, and the GRA have also gained some ability to speak, read, and write Mutsun (which is helpful in parsing sentences from the archival materials). Although their ability to use the language is limited, they are the first people in generations to be able to use it at all. Finally, the PI's publication record in her other research area (phonetics) demonstrates that she has the skills to disseminate the results of her research through publications in high quality refereed journals.

The GRA will devote her entire half-time appointment (20 hours per week) to the project during the academic year, and will work full-time on the project during summers. The primary responsibility of the GRA will be to perform the automatic parse of sentences from the archival sources, determine the correct parse, and add information to the lexical database as the texts supply it to achieve the correct parse. Because of the astounding quantity of archival materials and the amount of analysis necessary to determine the correct parse of an unfamiliar sentence, it is anticipated that this will require almost all of the large amount of time budgeted. The GRA will also assist with creating the final products of the project. The GRA will travel with the PI to California for two trips to work in person with the community members.

The GRA, Lynnika Butler, has already been working on the project through the Woodrow Wilson Public Scholarship grant and independent study, and she is committed to continuing on this project. She is eminently qualified for this work. She has a near-native command of Spanish (necessary for reading and translating the Harrington and Arroyo materials), and she has a strong interest in endangered languages and language revitalization. In the past year she has acquired thorough knowledge of Mutsun grammar, some ability to speak Mutsun, and a complete understanding of the project databases. She has been doing outstanding work on the project for a year. She has just finished her second year of graduate school, so she may be available for the entire grant period, but if she graduates during the grant period, another appropriate GRA could be hired and trained. There is at least one other graduate student currently in the Linguistics Department with a good command of Spanish and an interest in language revitalization, and because of the strength of the Linguistics Department at the University of Arizona in Native American languages and language endangerment, the department attracts many students who wish to work in these areas. The current GRA could help to train a subsequent GRA if necessary.

The undergraduate research assistant will also work 20 hours per week during the academic year and full-time during the summers. This student's primary responsibility will be entering basic information from the archival materials to make it easily accessible for the GRA and the PI. Because of the very large quantity of handwritten materials, this time commitment is necessary. In the later stages of the project, this student will shift his/her effort toward work on the language textbook, and will also assist with the final products of the project. This student will also accompany the PI and the GRA on two trips to work in person with the Mutsun community. No particular undergraduate student has been identified for this project, but because of the geographical location of Tucson, quite a few linguistics undergraduate majors speak Spanish well or natively. Many of the linguistics undergraduate majors are dedicated workers with good attention to detail, and these characteristics, along with good command of Spanish, are the main qualifications necessary for this position. If a suitable Linguistics major is not available in the relevant years, it should be possible to find a suitable student from within a related discipline. The number of undergraduates with Spanish competence who have volunteered through independent study evidences the availability of such students.

Ms. Luna-Costillas currently works approximately 20-30 hours per week on the long-term project, and she will continue to do so during the requested project period. She will be responsible for disseminating the materials to the Mutsun community, determining the usability of the materials for other community members, providing feedback to the PI, teaching language classes (which serve to evaluate the usability of the materials as well as to teach the language), and developing additional teaching materials. She will also continue to give presentations on Mutsun language revitalization both within and outside the Mutsun community, and will work with the K-12 school system in areas with many Mutsun residents to integrate information about Mutsun into public school teaching. Ms. Luna-Costillas is perfectly qualified for this work. She has attended the Breath of Life Language Revitalization workshops at UC Berkeley for years and has taken a university-level course in linguistics. She has been working on the project, which she initiated, for approximately eight years, and she is currently teaching community language classes in Mutsun. She has spoken at numerous events on Mutsun language revitalization. She has been a Tribal Councilwoman of the Amah Mutsun Tribal Band, and is now the Language Director of the Mutsun Language Foundation. She is also currently involved in transcribing parts of the Harrington notes through the UC Davis Harrington project.

The programmer for the project will dedicate a total of 400 hours to the project, concentrated at the beginning and end of the project time period. The programmer's responsibilities will be to create improved data entry and data checking software, a Mutsun spell-checker, a program for searching the encrypted final electronic product, and scripts for manipulating the data structure. The programmer will also write documentation for the software aspects of the final products. The programmer, Keith Alcock, has 20+ years' experience programming in a variety of computer languages for several platforms. He also has experience working with language technologies, including spellcheckers and grammar checkers. He has already been working for the long-term project as a volunteer for approximately three years. (The programmer is the PI's spouse. The PI has checked with the University of Arizona about the conflict of interest involved in hiring him, and since he will be hired for only a short time period for a specific project, this will be possible. The time period of the hire will be shorter than the project period, because the programmer will donate half of his effort as cost-sharing. His willingness to work on the project partially as a volunteer and his previous experience with the project, along with his programming qualifications, are the reasons for choosing him as the programmer for the project.)

The independent study students will work on the project 5 hours per week each during the semester sessions. These students will provide additional labor for typing the archival materials, and will work with the PI and the GRA on parsing Mutsun texts. They will also contribute to the development of the Mutsun language textbook and other teaching materials. The particular students who will work on the project through independent study have not yet been determined. However, four graduate students and three undergraduates of the Linguistics Department have done independent study (or volunteered) to work with the PI on this project in the past, and it is anticipated that as the project grows, more students will wish to participate through independent study. (Other projects within the department use this method of partial staffing through independent study quite successfully.) Because of the Linguistics Department's strength in Native American and endangered languages, many students are interested in this project and qualified to work on it. The work with archival materials affords them the opportunity to learn a different methodology than is used with living speakers. Past independent study students on the project have been outstanding students, and have put a great deal of effort into the project and produced very high quality work. Usually, only highly motivated students take the initiative to work as independent study students, and these students make a valid contribution to the project.

VII. Dissemination

As described above, the project will produce a hard copy dictionary, a hard copy annotated text collection, and encrypted but searchable electronic copies of both the lexical and text databases in two formats (Shoebox and XML). The hard copy products will be published through a dictionary publisher, as described in the plan of work above. The electronic versions may be included with the published hard copy, or may be distributed separately, depending on the requirements of the publisher. The cost for the final published product is estimated at $20 per copy, based on the photocopying costs for past self-published versions, adjusted for a longer hard copy product plus a CD. A search for dictionaries of Native American languages on Amazon.com indicates that this is near the middle of the price range.

Both the hard copy and the electronic versions (if separate) will be advertised to linguists through the Linguist List (which has 14,000+ subscribers worldwide), through appropriate topic-specific e-mail lists, e.g. those for endangered and Native American languages, and through book sales at appropriate linguistics conferences, as determined by the publisher. The final products will be disseminated to the Mutsun community through Ms. Luna-Costillas, with help from the tribal council and the Mutsun Language Foundation. These two groups have the ability to reach a large proportion of the Mutsun people directly. The Mutsun Language Foundation will also disseminate the final products at fundraising events, at which they currently sell traditional crafts and t-shirts for fundraising purposes. The products of the grant will be disseminated to other Native American communities through the Breath of Life Workshops at UC Berkeley, through the Endangered Language Fund (based in Connecticut), and through other relevant workshops in other parts of the country. The appropriateness of the chosen product formats for these communities is discussed in section IV.E and IV.F above.

The proposal includes funding to provide approximately 200 free copies of the final products to Mutsun community members who would find it difficult to afford a copy themselves. As with many Native American groups, many of the Mutsuns are living in poverty, and subsidized dissemination of the final products is necessary in order to avoid charging them to have their cultural heritage returned to them. Beyond the free copies, Mutsun community members will be charged a reduced rate for the products of the project. The Mutsun Language Foundation may be able to provide funding to subsidize (or even completely pay for) dissemination to Mutsuns through their fundraising efforts, or it may be possible to slightly increase the price for regular dissemination in order to subsidize dissemination to the community.

Broad dissemination of information about the Mutsun language raises issues of intellectual control and ownership of the language. The Mutsuns are sensitive to these issues, but they also desire to make their language, their culture, and their very existence known to the broader society. They have chosen to make materials about their language and culture widely available, even putting some language material on the World Wide Web. However, to preserve the integrity of the dictionary and text databases, the electronic versions of the final products will only be made available as encrypted files, so that they are searchable, but not editable. Information about the final electronic products will be posted through the Open Language Archives Community (http://www.language-archives.org/), a standard repository for language data. The Mutsun community may choose at that time to make the electronic products available directly on the web in encrypted form through OLAC, or may choose to place only a metadata heading through OLAC with information about who to contact to receive a copy of the electronic files. (This will also depend on arrangements with the publisher of the hard copies.)

The project will provide documentation about the development of software on request. The programmer may also publish information about the development of some of the programs, such as the spellchecker. However, information about the encryption method for the non-editable electronic databases will of course not be widely published.

VIII. References

References to manuscripts generated by the project are given in III.C above rather than here.

Adams, D. 1985. “Internal reconstructioin in Mutsun morphology”. *IJAL,* **51**:329-31.

Arroyo de la Cuesta, F. 1861. *Grammar of the Mutsun Language, Spoken at the mission of San Juan Bautista, Alta California.* (Shea's Library of American Linguistics, 4.) New York: Cramoisy Press. (Reprinted 1970 by AMS Press, New York.)

Arroyo de la Cuesta, F. 1862. *A Vocabulary or Phrasebook of the Mutsun Language, Spoken at the mission of San Juan Bautista, Alta California.* (Shea's Library of American Linguistics, 8.) New York: Cramoisy Press. (Reprinted 1970 by AMS Press, New York.)

Baldwin, D. 2003. *The Myaamia Project Messenger,* v. 2. Newsletter of the Myaamia Project.

Basham, R., Farrar, S.O., Fitzsimons, B., Gonzalez, H., Langendoen, D.T., Lanham, A., Lewis, W.D., and Simons, G.F. 2004. "A model for interoperability: XML documents as a distributed RDF database". Paper presented at the EMELD Workshop on Linguistic Databases, Detroit MI, July 15-18, 2004.

Beeler, M.S. 1955. "Saclan". *IJAL*, **21**:201-9.

Bird, S., Hammond, M., Amarillas, M., Jeffcoat, M., Harley, H., Miyashita, M., Moll, L., Willie, M.A., and Zepeda, O. 2002. "Web-based dictionaries for languages of the South-west U.S.A." *Literary and Linguistic Computing,* **17**:427-38.

Blevins, J., and Arellano, M. 2004. "Chochenyo revitalization: A progress report". Talk given at the SSILA Winter Meeting, Boston, MA.

Blevins, J., and Garrett, A. In press. "The evolution of metathesis" to appear in a volume edited by B. Hayes, R. Kirchner, and D. Steriade (Cambridge: Cambridge University Press).

Callaghan, C. 1997. "Evidence for Yok-Utian". *IJAL,* **63:**18-64.

____. 1998. "The imperative in Proto-Utian". *UCPL*, **131:**160-67.

____. 2001. "More evidence for Yok-Utian: A reanalysis of the Dixon and Kroeber sets". *IJAL,* **67:**313-45.

____. 2003. "Proto-Utian (Miwok-Costanoan) case system". *IJAL,* **69:**49-75.

Costa, D. 1994. *The Miami-Illinois Language.* Ph.D. dissertation, University of California, Berkeley.

Hinton, L. 1994. *Flutes of Fire: Essays on California Indian Languages.* Berkeley: Heyday Books.

Hinton, L. 2001a. "Sleeping languages: Can they be awakened?" In *The Green Book of Language Revitalization in Practice,* ed. L. Hinton and K. Hale, Academic Press, pp. 413-18.

Hinton, L. 2001b. "The use of linguistic archives in language revitalization: The Native California Language Restoration Workshop." In *The Green Book of Language Revitalization in Practice,* ed. L. Hinton and K. Hale, Academic Press, pp. 419-24.

Hinton, L. 2001c. "The Ohlone languages." In *The Green Book of Language Revitalization in Practice,* ed. L. Hinton and K. Hale, Academic Press, pp. 425-28.

Mason, J.A. 1916. "The Mutsun Dialect of Costanoan Based on the Vocabulary of de la Cuesta." *University of California Publications in Archaeology and Ethnology.* 11.399-472.

Mielke, J., & Hume, E. 2001. "Considerations of word recognition for metathesis," in *Surface Syllable Structure and Segment Sequencing,* ed. E. Hume, N., Smith, & J. van de Weijer, Leiden: HIL.

Okrand, M. 1977. *Mutsun Grammar.* Ph.D. dissertation, University of California, Berkeley.

____. 1979. "Metathesis in Costanoan Grammar". *IJAL,* **45:**123-30.

SIL (Summer Institute in Linguistics). 2000. *The Linguist's Shoebox.* http://www.sil.org/ computing/catalog/.

Yamane, L. 2001. "New life for a lost language." In *The Green Book of Language Revitalization in Practice,* ed. L. Hinton and K. Hale, Academic Press, pp. 429-32.