

MAKING A DICTIONARY FOR COMMUNITY USE IN LANGUAGE
REVITALIZATION: THE CASE OF MUTSUN

Natasha Warner

University of Arizona

Lynnika Butler

University of Arizona

Quirina Luna-Costillas

Amah Mutsun Tribal Band

Contact information:

Corresponding author: Natasha Warner

Address: Department of Linguistics
University of Arizona
PO Box 210028
Tucson, AZ 85721-0028
U.S.A.

E-mail: nwarner@u.arizona.edu

Phone: +1-520-626-5591

Fax: +1-520-626-9914

I. Introduction

This article discusses the creation of a dictionary for use in a Native American community that is attempting to revitalize its dormant ancestral language entirely from archival materials.¹ Some issues are the same as those involved in creating a dictionary of an endangered language for use by the community rather than by linguists. Other issues differ because of the lack of any living speakers and reliance on archival source documents. The problems introduced by limited data and the impossibility of collecting new data are the same as those faced in making a dictionary of a dead language.

Mutsun, the language we will discuss, is a Costanoan (Ohlone) language, part of the Yok-Utian family (Callaghan 1997, 2001). It was spoken near San Juan Bautista, California (South of San Francisco), until the last known fluent speaker died in 1930. There are no audio recordings of spoken Mutsun, although there are a few songs. However, there is a large quantity of written documentation of Mutsun, recorded by early linguists, spanning 115 years from early post-contact times to the death of the last fluent speaker. The original source material varies greatly by author in its consistency and reliability of transcription, degree of analysis, topics, and accessibility. Since 1998, we have been working on producing an English–Mutsun Mutsun–English dictionary from the information contained in these sources for use by Mutsun community members in their language revitalization project. The database we use to create this dictionary also provides a useful tool for linguistic research. This article describes the issues we have encountered as part of the project.

In 1996, members of the Mutsun community began working toward learning and revitalizing Mutsun. The optimistic long-term goal is for all interested members of the

community to achieve reasonable fluency in (the revitalized form of) the language, at which point it is likely that some Mutsuns would be raising their children in Mutsun. There are not yet any fluent speakers, but progress is underway (and is described in Warner et al. in preparation-a). The project takes a multi-faceted approach to revitalization, involving archival source analysis, lexicography, development of teaching materials, work toward fluency, teaching of community language classes, and cultural revitalization. This article focuses only on issues of developing the Mutsun dictionary.

The structure of this article is as follows: the following section describes the structure of the lexical database. Section three gives an overview of the primary and secondary source materials available for Mutsun. Section four discusses extracting information from the sources and entering it into the lexical database. Section five covers how to organize and format the dictionary appropriately for the audience. In section six, we turn to the text database used to encode information above the lexical level. Our conclusions about making a community-oriented dictionary in this situation form the final section.

2. The sources

There are three major original sources of data on Mutsun, as well as a few shorter original sources and some analyses of the original sources. The sources differ greatly in how they should be treated for lexicographic use, and in how reliable they are. Costa (1991, 2003) discusses similar issues for Myaamia (Miami), another dormant Native American language. More generally, any lexicography project for a dead language is

likely to face a similar variety of sources, even if the dictionary is not intended to be used for revitalization.

2.1. Arroyo de la Cuesta

The earliest sources are a 'phrase book' and a grammar by Father Felipe Arroyo de la Cuesta, a mission priest, who compiled these works in approximately 1815 at the San Juan Bautista mission (in the area where Mutsun was spoken). The phrase book consists of a list of 2,884 sentences, alphabetized by first letter of the sentence, with Spanish translations. At the time of Arroyo's work, there were many fluent native speakers of Mutsun, which is an advantage of this source. Also, Arroyo's transcriptions are based on Spanish orthography, a far better choice for transcription than English. Finally, the phrase book represents a substantial quantity of data.

However, the Arroyo work has several deficits: 1) Arroyo's transcription conventions are inconsistent and neutralize several distinctions, especially distinctive length. 2) The typesetter for the published version (Arroyo 1861, 1862) introduced a great many errors (Harrington, part IX, reel 17, pp. 598-599). 3) Arroyo often records multiple phonological forms for the same word (e.g. *murtei*, *murteis*, *murtoeis* for "night", and both *akkes* and *awes* for "salt"). 4) The Arroyo data is entirely at the sentence level (with no longer texts), and alphabetizing the sentences destroyed any discourse context. 5) Arroyo's choice of topics was influenced by the mission context, leading to inclusion of a large number of words regarding religion, morality, and immorality.

2.2. *Merriam*

No more major fieldwork was conducted on Mutsun until 1902, when C. Hart Merriam collected a word list of approximately seventy pages, working with one of the last fluent speakers, Barbara Solorsano (Merriam 1902). The positive aspects of Merriam's list are that he gives very specific translations for animals and plants, because he was a naturalist, and that his choice of content was more culturally relevant than Arroyo's. However, his translations are grammatically unreliable, mistaking entire sentences for a single verb or noun, for example. Most importantly, Merriam's English-based transcription system was notoriously inaccurate and inconsistent (Berman 2002). This data is almost entirely at the word level, without context.

2.3. *Harrington*

The final major original source is by far the largest and linguistically most accurate. This is the work of J. P. Harrington (1922, 1929-30). Harrington's Mutsun work consists mostly of unanalyzed field notes taken during his work with the last fluent speaker of Mutsun, Mrs. Ascensión Solorsano, in 1922 and 1929-1930. Based on the Mills guide to the Harrington notes (Mills 1981), we estimate this collection at 36,000 pages of microfilm on Mutsun, a formidable quantity of documentation. Harrington's notes are not densely filled, and include some duplication, material on another language, and cultural rather than linguistic material. Still, this is a very large source.

Harrington is known to have been an excellent phonetician (Callaghan 1975), and his transcriptions are detailed and extremely reliable. His data includes re-elicitation of the Arroyo material, the Merriam material, and some smaller sources, as well as a wealth

of cultural information, original elicitation, and at least half of one traditional story and a connected text on cultural practices (the only longer texts recorded in Mutsun that we know of so far). Although his focus on re-elicitation maintains the topic choices of the previous sources, it is useful for clearing up their inaccurate transcriptions. The main problem with the Harrington material is accessibility. Most of it is in Harrington's very poor handwriting, and there is no way to locate particular information within the data.

2.4. Analyses of original sources

There are two major secondary analyses. Mason (1916) translated Arroyo's Spanish to English and analyzed the Arroyo sentences into morphemes. He created a Mutsun–English dictionary from this data. Mason was unable to work with a native speaker of Mutsun on this project, so all the glosses have been through double translation without consultation with a native Mutsun speaker or the author of the Spanish glosses (Arroyo). (This is also true of our translations of Harrington's Spanish glosses.)

Although Mason's analysis of Arroyo's material is highly accurate and insightful, the Mason dictionary is undermined by the poor quality of Arroyo's original data. The Arroyo data is so inaccurate, especially for phonological forms, that no analysis of it can form a good basis for a Mutsun dictionary.

The other analysis of a source is Okrand's (1977) grammar of Mutsun, which he compiled based on a small portion of the Harrington data (a few boxes of notes that were at UC Berkeley at the time, before the microfilmed notes were available). This grammar is an excellent descriptive analysis. Okrand is careful not to neutralize any distinctions, and he presents the evidence clearly in cases of inconclusive data. After Okrand

completed his Mutsun grammar, he had an opportunity to work with more of the Harrington notes on Mutsun at the Smithsonian. He compiled some thousands of slip notes beyond what he used for his dissertation, but never published this work.

There are a few minor original sources, as well, but these are mostly short word lists, and most information in them is likely to be in the other sources as well. For more details on all of the sources, see Warner et al. (in preparation-a). From this collection of sources, representing various stages of language attrition and various strengths among the field workers, the modern Mutsun community is attempting to bring their ancestral language back. This article discusses how we create a dictionary for community use from the information in these sources.

3. Structure of the lexical database

We originally chose The Linguist's Shoebox (SIL 1998) as the lexical database software for the Mutsun project, and we are now switching to SIL's LinguaLinks (SIL 2002). The need to keep track of information from multiple archival sources, which were all originally written in different transcription systems, requires some changes from a lexical database one might construct if one were collecting the data from a living speaker. For each lexical entry, we choose the phonological form which we consider to be the most reliable of all the forms given by the various sources (see 4.2 below). This form is the one that we recommend to the community, and it is the headword. In the lexical database, forms from all sources are recorded only as converted to the practical orthography (the orthography developed for use within the community (Warner et al. in preparation-a)), not in the transcription system of the original source. In a separate text

database (see section four), both the original transcription (an exact representation of what is in the original source document) as well as the conversion to practical orthography are recorded. (Because the two databases are directly linked, particularly in the newer LinguaLinks software, this does not lead to problems in updating them both.)

A sample lexical entry which illustrates most of the information the database contains (in Shoebox format) appears in Figure 1. How the information for each field is determined and encoded is discussed in the appropriate sections below.

INSERT FIGURE 1 HERE

The `\lx` field shows the main form of the lexeme (see 4.2 and 4.5). `\ps` shows the part of speech (section 4.4). `\ge` gives the translation(s). `\xv`, `\xe` pairs give example sentences and their 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 (see section 4.2), and show which source the variant came from (Mason=Ma/Ar, Okrand=O/H, etc.). `\pdl`, `\pdv`, `\pde` triples list any irregular grammatical forms of the word. By using the `\va`, `\ve` fields vs. the `\pdl`, `\pdv`, `\pde` fields, we can distinguish 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. The `\np` field shows in which sources the main form listed in the `\lx` field is attested (here, Okrand and Merriam both have /neppe/, see section 4.2 below). `\ns` shows who has edited the entry and when. Additional fields not used for this entry show regular allomorphs of the main

form (section 4.5.5), usage information (e.g. “rare” or “vulgar”), and what language the word is borrowed from if it is not native to Mutsun. We also initially kept track of what pages of the sources the word appeared on, but the development of a searchable text database (section six) has made this unnecessary.

As of the writing of this paper, the lexical database contains approximately 3000 entries. It contains all the forms in the Mason dictionary, the Merriam notes, and the Okrand dissertation, as well as all forms in the portion of the Harrington notes that we have been able to analyze thus far. It does not contain the information from the majority of the Harrington data, for which analysis is ongoing, or information in the original Arroyo (1861, 1862) publications that is not in the Mason dictionary.

4. Extracting data from the sources

4.1. Converting transcriptions to practical orthography consistently

Converting from each source to the practical orthography presents its own challenges.² Arroyo's transcriptions, while they have the advantage of being based on Spanish orthography, are not very consistent. Mason (1916: 404) finds that Arroyo spells the retroflex stop phoneme *T* as any of 'tr, th, thr, thrs, trs', and we have also found 'ths, tchr, tshs, strs'.³ The prevalence of the letter 's' in Arroyo's spellings of retroflex *T* reflects its affricated pronunciation (Okrand 1977: 21-23), which is evidenced by frequent confusions between *T* and *c* in all the sources and a rarity of confusions between *T* and *t*.⁴ Arroyo's frequent use of the letter 's' as part of his spelling of the retroflex stop thus provides evidence, but it also means that sequences such as 'sthr' in his material are ambiguous between *sT* and simply *T*.

Another example is Arroyo's spelled 'g', which is sometimes the phoneme *w* (when followed by spelled 'u' or 'ü'), but before a consonant is ambiguous between *k* and *h*. Arroyo writes the letter 'h' often, but this can be variously a vowel length marker (so that his spelled 'ihi' can indicate *ii*), nothing at all, or possibly even the phoneme *h*. Following Spanish usage, Arroyo writes 'i' for the phoneme *y* and (in some environments) 'u' for the phoneme *w*. The glide vs. vowel status of a sound can sometimes be recovered from the syllable structure, since the Harrington data shows that Mutsun is maximally CVC (Okrand 1977) and has very few VV sequences, but some cases remain ambiguous. This problem is compounded by Arroyo's spelling 'si' for the alveopalatal fricative phoneme *S*. He similarly spells other palatalized consonant phonemes as consonant-i sequences. Thus, a postvocalic sequence 'sia' in Arroyo's spelling could phonemically be any of *Sa*, *sya*, or *sia*, and *Sya* would also not be unreasonable. In entering the Mason and Arroyo data into the database, we have tried to develop consistent and objective ways to convert Arroyo's transcriptions to a phonemic string. Some information is simply not recoverable, however: Arroyo often neutralized vowel and consonant length distinctions.

The Merriam data is the most difficult to convert accurately to a phonemic representation because Merriam's transcriptions are based on English spelling, and his ability to transcribe vowels was poor. Although Mutsun has only five vowels, each either long or short, Merriam uses a wide variety of vowel symbols very inconsistently. He wrote at least two guides to his transcription system, but neither gives a thorough description. For example, Merriam's explanation does not specify the open/closed syllable pattern, in which a single vowel letter (e.g. 'e') in open syllables usually has its English 'long' value (e.g. IPA /i/ for 'e'), and in closed syllables it has its English 'short'

value (e.g. IPA /ɛ/ for 'e'). In a closed syllable, Merriam writes a double letter vowel (e.g. 'ee' for IPA /i/) to symbolize the 'long' value. The second syllables of the pair 'she-ne' (Merriam's spelling) for phonemic *sinni* 'child' and 'she-neek-ma' (Merriam's spelling) for phonemic *sinnikma* 'children' provide particularly clear evidence of this syllable-structure-conditioned variation. Since Merriam worked with the mother of Harrington's informant, differences between these two sources cannot easily be ascribed to dialect differences or language change, and are more likely to reflect Merriam's inconsistencies or errors. Table I shows part of the conversion system the first author derived for Merriam's Mutsun notes.

INSERT TABLE 1 ABOUT HERE

Even after these conversions, comparison with Harrington and Mason/Arroyo has revealed some forms in which Merriam's vowels appear incorrect. For example, Merriam spells 'ā'-ne-nah' for "blackberry", which should convert to phonemic *enina*, but Okrand finds *eenena* in Harrington. (Neutralization of vowel length in the first syllable is not surprising, but the second syllable vowel is an inexplicable transcription discrepancy.) Luckily, Arroyo usually represents vowel quality, Merriam's weakest point, accurately.

Compared to Mason/Arroyo and Merriam, the Harrington data is easy to convert to practical orthography. The main difficulty with Harrington is that he notoriously represents fine phonetic detail rather than a phonemic level (Hinton 1994: 201), for example using approximately six different symbols for sibilant fricatives, of which Mutsun only distinguishes two (*s* and *S*). This much phonetic detail is not useful to

community members trying to learn how to pronounce a word.⁵ Fortunately, Harrington's symbols map onto the phonemes fairly predictably. Harrington's usage of a variety of accent marks is less consistent or clear (Okrand 1977: 95), but the practical orthography does not mark stress (see Warner et al. in preparation-a), so this is not a problem.

Because Harrington re-elicited most or all of the data in the earlier sources, and the Harrington data is the most phonologically accurate source, it might seem unnecessary to expend effort on deciding how to convert the older source data accurately. However, because Mrs. Solorsano was the last speaker of the language, and attrition was taking place, she was not able to give corrected forms of many words from the earlier sources. She simply did not know all the words of the language. Furthermore, projects on most other languages probably do not have a phonologically accurate source that re-elicited all previous material. This particular feature of the Mutsun sources is unusual, and we therefore discuss the process of extracting data from a variety of sources even though the Harrington data will confirm or correct much of the earlier data.

4.2. Choosing the most reliable form

In the Mutsun data, it is very common to find more than one phonological form for a single word, whether in different sources or a single source. For example, for the word "salt (n.)", Mason/Arroyo give both *akkes* and *awes*, and Merriam gives *akkis*. Some of these alternate forms are the result of known processes of allomorphy, but most are simply transcriptional inaccuracies or unexplained disagreements on phonological form. In our current lexical database, approximately thirty-nine percent of the 3053 Mutsun

headword entries have at least one variant form. Many have more than one variant, so that there are approximately three fifths as many variants as there are (non-variant, main) headwords. As we continue to enter more information from the Harrington notes into the database, we expect that the number of variants will increase dramatically.

When there is more than one phonological form available, we must choose which form to use as the citation form and recommend to the community for language learning, and count any other forms as variants. The primary criteria for choosing the most reliable form are syllable structure (maximally CVC) and allowable verb templates, because the Harrington data and Okrand analysis show the phonotactics of the language to be very clear and robust. For example, Harrington's form *hanni* "where" is preferred to Mason/Arroyo's phonotactically illegal *ann*, and Mason/Arroyo's *lophe* "become moldy" is preferred to their other form *lopkti* (which contains a phonotactically illegal cluster, and *h* and *k* are frequently confused in the Arroyo data because of Spanish orthographic norms).

Okrand (1977) found that almost all verb stems in Mutsun fall into one of seven disyllabic verb templates, all of which have a vowel-final allomorph as the form that appears in isolation and before most suffixes. The Arroyo/Mason forms violate these constraints with abandon. For example, Mason gives *hairmurnikkwi* for "to lift, to pick up", with five syllables rather than two and a CVC syllable structure violation. (*hai* could be taken as *hay*, leaving four syllables and two syllable structure violations instead.) Some violations are not as drastic: Arroyo gives *akkara* for "to look up", which has one syllable too many (and *-ra* is not a known suffix). Harrington obtained the phonotactically legal *akra* for this form.⁶ Some discrepancies between Arroyo and the

later sources might reflect language change or dialect differences. However, the Harrington data so clearly follows the phonotactic constraints in the vast majority of lexical items, and the Arroyo data violates them with such frequency, that we are inclined to lay the blame on the earlier transcriber.

In choosing which form to use if more than one form is phonotactically legal, we give more weight to Okrand/Harrington than to the other sources, since Harrington was an excellent phonetician. We consider how many of the sources agree about (parts of) the form (majority rule), but we also consider what types of transcription errors are common in which sources. Hence, Okrand's *tappur* "wood, tree" is preferred to *tapur*, given by both Mason/Arroyo and Merriam, not only because Okrand is more reliable overall, but also because Arroyo and Merriam often replace long consonants with short.

Deciding whether two forms are one word with a transcription difference or two distinct words is often difficult. Most of the glosses have been through double translation by way of Spanish, so meanings are not precise, and both Arroyo's and Merriam's transcriptions allow for considerable phonological leeway. For example, is Mason/Arroyo's *amun* "concerning" the same word as Okrand's *amuu* "such that"? The typesetter for the Arroyo text frequently confused *u* and *n*, so these are probably the same. Mason/Arroyo's *ahenyakken* and Okrand's *haahen*, both translated as "flee", are less clear. The phonological disparity is large, but these are probably different recordings of the same word with some unrecognized suffixes in the longer form.⁷ There is, of course, no definitive answer to whether two forms are forms of the same word.

When deciding whether to combine two forms into one entry or keep them as separate entries, we attempt to strike a balance between neutralizing potentially real

distinctions that could be useful when trying to communicate in Mutsun, and asking community members to memorize arbitrary lexical distinctions. As a comparison, if one can imagine a hypothetical linguist in the distant future trying to reconstruct English from incomplete records, would 'lie' and 'lay' be considered separate words, mistranscriptions, free variants, or dialectal variants of the same word?

4.3. *Reconstructing a phonotactically legal form*

Many words from Mason/Arroyo and Merriam are of suspicious form in that they violate phonotactic constraints or the verb templates, and it is not desirable to ask the community to learn words that are probably incorrect (e.g. *hairmurnikkwi* above). The phonotactic patterns of the language are among the best documented and most consistent aspects of the language in the Harrington data, and Harrington is without question the most phonologically accurate source, so we prefer to offer forms that conform to known phonotactic constraints. Furthermore, if a verb appears only as a consonant-final form, it is impossible to add some suffixes (those with two initial consonants) without further violating CVC syllable structure. One can often determine from the evidence of related forms, morphological alternations, and transcription conventions what a likely correct form is, even if no alternate forms of the word itself are attested.

We decided to provide reconstructed forms that obey the phonotactic constraints pending finding the more accurate form in the Harrington data, since it could take years to analyze it all, and the community needs a more reliable dictionary now.⁸ Of course, we record in the database the fact that a form is an unattested reconstruction, we list the attested form as a variant, and we describe the evidence used to do the reconstruction in

the notes field. Sometimes, we later find a phonotactically legal form in the Harrington notes, and this either confirms or replaces our reconstruction.

Some reconstructions are simply a matter of further transcription conversion: Mason/Arroyo record 'henkotrstrsse' (their spelling), which we reconstruct phonemically as *henkoTTe*, "silence", based on Arroyo's frequent use of 's' with the retroflex *T* phoneme. In some cases one can recognize a morpheme that is well-attested elsewhere: Mason/Arroyo give *enekmin* for "letter", but Okrand finds *enne* for "to paint, write", and *-kmin* is a known nominalizer. Thus, unattested *ennekmin* is more likely. By considering morphological alternations, one can reconstruct the phonemic string *rorSo* from Arroyo's transcription 'rorois' "to play". We do have an attestation of *roroSpu* "amuse oneself", and the reflexive *-pu* takes the metathesized consonant-final allomorph of the verb (Okrand 1977). Thus, the stem must be *rorSo*, and 'is' for the phoneme *S* is plausible as an Arroyo transcription.

Some reconstructions are less certain: there are multiple possibilities for the form Arroyo spells as 'aoepaeis' "perhaps", but such long sequences of vowels are very unlikely in Mutsun. We reconstruct the phonemic string *epeS*, based on the facts that Arroyo also spells 'epaeis' as a form of the same word, that in Arroyo's spelling 'is' for the phoneme *S* is likely, and that Arroyo's spelled 'ae' seems to be more often *e* than *a* where we have comparative data from other sources. We reconstruct *surni* "to heat, warm" for Mason/Arroyo's *surn*, but this is very unsure. The final *-n* cannot be the inherent reflexive suffix, as the final *-n* in many verbs is, first because it would attach to a vowel-final form, and second because the single syllable *sur* is too short for a verb stem. More likely, a final vowel was simply not heard and omitted, but there is no good way to know

which vowel. We chose *i* as a vowel perhaps easily missed, but we recorded in the database that this reconstruction was very uncertain.

4.4. *Determining the part of speech*

Part of speech is listed in the Mutsun dictionary, and we feel this is necessary because English translations are so often ambiguous as to noun vs. verb status, and Mutsun nouns and verbs are not overtly marked as such in any obvious way. Marking part of speech in a language with no living speakers brings up two questions. First, it is not entirely clear what the relevant parts of speech are in Mutsun. Second, there may be insufficient data to judge part of speech for particular words.

Mutsun clearly has nouns and verbs, since these take different sets of suffixes. Mason (1916) lists a category of adjectives in his dictionary. However, most of Mason's adjectives are just verbs with stative meanings, such as *awli* "to be salty, to be saline". Okrand finds that there is no adjective part of speech in Mutsun (1977).

The part of speech categories we use are: noun, verb, perfective, imperative, pronoun, adverb, quantifier, numeral, question word, demonstrative, conjunction, exclamation, and suffix. It is not clear that all of these function in syntactically or morphologically distinct ways, but when the entire set of Harrington data has been analyzed, it will be possible to determine this. The perfective part of speech refers to verbs ending with the perfective suffix *-Ste* (e.g. *kommeSte* "tired"). Such forms are clearly just suffixed verbs, not a truly separate part of speech, but they often require a different English translation than the stem does (e.g. *komme* presumably means "to get tired out, to tire"). These forms' translations are useful to list as headwords in the

English-Mutsun dictionary so that users can look up "tired", "thirsty", etc. The category "imperative" is used only for a small number of inherently imperative verbs that lack non-imperative forms (e.g. *iSke!* "go!, run!, scat!, *someone goes").

Mason (1916) occasionally listed a word with an incorrect part of speech, and Merriam gave neither parts of speech nor sentential context. We determine part of speech from context in attested sentences, suffix usage, and, when no other information is available, Spanish and English glosses. Because of the uncertainty of part of speech, we allow "n?" and "v?" as part of speech labels for uncertain nouns and verbs.

4.5. *What should be a headword?*

4.5.1. *Type of morphology.* Others have discussed the problem of what to choose as headwords for community-oriented dictionaries, particularly if the verb stem is not easy to relate to the whole verb word (Young and Morgan 1987; Hill 1996; Munro 1996, 2002; Abbott 1998; Hinton and Weigel 2002; Pulte and Feeling 2002). Munro (1996) gives an example from Cherokee in which the string of morphemes for the simplest possible whole-word citation form of a particular word is *gà-hnéég-a*, the surface form is *kànééga*, and the root is the bound morpheme *-hnéég-*. Since most community members will not want to look up bound roots rather than whole words (Young and Morgan 1987), it is important to choose which inflected form to use as the citation form.

This problem is not as difficult for a language like Mutsun as it is for most of the examples discussed by the previous authors. First, in Mutsun, most verb and noun stems can occur in isolation, without overt inflections. Thus, the stem *amma* "eat" means that

someone eats, non-past tense. Person and number are marked with clitic or independent pronouns. The stem *sii* simply means "water", and it is a word by itself.

Second, Mutsun is (almost) entirely suffixing, so one can look up even highly suffixed forms under their stems. Finally, Mutsun does not have pervasive opaque phonology. In Navajo and many other Native American languages, the string of underlying morphemes sometimes bears little resemblance to the surface form of the entire word, so listing entire words as headwords obscures the relationships among morphemes, while listing morphemes makes it difficult to find entries. In Mutsun, all morphemes, including suffixes, are for the most part transparently recognizable. For example, the attested word *wattimpihte-k* "3rd-sg. is carried" consists of *watti* "go", *-mpi* "causative", *-hte* "passive", and *-k* "3rd sg. clitic". If *watti*, *wattimpi*, and *wattimpihte* are all listed, the relationship among them will be transparent.

Some have referred in previous literature to the difficulty of choosing headwords for polysynthetic languages (Abbott 1998). It may be that what makes it difficult to choose headwords is a specific combination of features of a language, such as being prefixing rather than suffixing, having mostly bound stems, and having many opaque phonological alternations. Mutsun is agglutinative, and it can use both nominalizers and verbalizers to make a verb into a noun and then back into a verb, just as Abbott (1998) demonstrates for Oneida. However, because it is suffixing, does not depend on bound stems, and does not have difficult phonology, headwords can be whole words without being difficult to look up.

4.5.2. *Forms with suffixes as headwords.* Corris et al. (2004) mention that some of their Australian indigenous language dictionary users wanted to have all inflectional forms of each word listed in the dictionary, and found it difficult to use any inflected form that was not listed in the dictionary in its entirety. As they discuss, it is not practical to have every inflected form of every word listed, especially in an agglutinative language. The third author of the current article, who is not a linguist but is a community language revitalization leader, has had no problem with generating regular inflected forms from stems listed in the dictionary. Ability to generate inflected forms when only whole-word uninflected forms are listed may vary greatly among communities and individuals.

We do in many cases list as headwords forms with derivational suffixes. To be listed as a headword, the form must be attested with that suffix. Even if we consider it very likely that *pio* "to brush teeth" could occur with *-pu* (reflexive) to make *piopu* "to brush one's teeth", we will not list this form if it is not attested, because our own understanding of suffix usage is imperfect.

Among attested derived forms, we list the suffixed form as a headword either if the meaning with the suffix is not transparent, or if we judge that dictionary users might not realize they could add this suffix to this stem. That is, we include more derived headwords than one would in a dictionary for fluent speakers. For example, we include all forms that appear with the inherent reflexive suffix *-n*, because its contribution to the meaning of a word is often not transparent (e.g. *pirkan* "to scratch (of birds)", cf. *pirka* "to dig up"). We also include as a headword *peTTempi* "to cause to become stuck together, to stick together" (cf. *peTTe* "to stick together, for example with glue", *-mpi* "causative", and *peTTen* "to get stuck together" with inherent reflexive). Someone who

understands how a causative suffix works and knows that the causative is fairly productive in Mutsun might not need *peTTempi*, with its transparent meaning, to be listed separately. A native Mutsun speaker might not need such an entry either, but many learners would.

4.5.3. *Potentially bound stems.* Although most verb and noun stems in Mutsun are not bound, many may only occur with derivational suffixes in the available data. Many of these are probably accidental gaps, but Mrs. Solorsano told Harrington that a few verb stems could not be used without derivational suffixes, including for example *istun* "to dream" (*istu-n*, with the inherent reflexive) (Okrand 1977: 214). *istu* appears in the Mason (1916) dictionary without suffixes, also translated as "to dream". However, Arroyo (Mason's source) never found *istu* without derivational suffixes: he found it several times with *-n* and once with the reflexive *-pu* instead. Thus, the dictionary lists a form which Mrs. Solorsano denied as possible, and which should probably be bound, because Mason assumed the suffix could be removed. Similarly, Okrand often cites stems in isolation regardless of whether they occurred unaffixed in his data. Even if a stem occurs only affixed in the data, because of the lack of living fluent speakers, one cannot usually determine whether these are accidental gaps.

For purposes of the lexical database, we have decided to enter stems even if the stem is only attested with a suffix, as long as the meaning of the stem is reliably derivable from the attested forms. For forms with the inherent reflexive suffix, it is often difficult to predict the meaning of the unaffixed stem (as with *istun* "to dream" above), and the reflexive suffix often participates in forms with idiomatic meanings (e.g. *annan* "to

pardon" with the inherent reflexive and *annapu* "to pity" with the reflexive *-pu*). If we cannot reasonably infer the meaning of the stem, we do not enter it as a headword.

In some cases, Mason failed to remove even productive inflectional suffixes from items he listed in his dictionary. For example, he listed the word for "language, speech" as 'ritca-se' (his spelling), which is clearly the objective form *riicase* of the stem *riica* ('tc' is his equivalent of the *c* of our practical orthography). The hyphen shows that he recognized the suffix, but he did not list the stem in his dictionary. Mason also gives 'ixiras' (no hyphen, his spelling), which Harrington data eventually allowed us to identify as *yiira* "kick" plus the remote past tense suffix *-s*. ('ixi' is probably an Arroyo spelling for *ii*, and deletion of *y* before *i* is typical for Arroyo.) When we can identify inflectional suffixes, we of course remove them and list the stem in dictionary.

4.5.4. *Irregular forms of words.* Mutsun has only a few known irregular forms of words (as opposed to regular allomorphy or unexplained variants of the same word form). The example in Figure 1 above, *neppe* "this", has two object forms, *neppese* (regular, with *-se* object marker) and *neppes* (an irregular reduction).⁹ This form is listed under the headword *neppe* as an irregular form (Figure 1). In this case, we have also chosen to list the regular instrumental and ablative forms *neppesum* and *neppewas* under the headword *neppe*, because Mutsun demonstratives rarely take case suffixes other than the objective.

One further irregular form of *neppe* does not appear as an irregular form: *nepkam* "these" (pl.). (The regular plural would be **neppekma*.) This is listed as a separate headword meaning "these" as well as being listed as a related word in the \cf field. We chose to make *nepkam* a separate headword because English also has an irregular form

for the plural of "this". A user of the dictionary wanting to find out how to say "these" in Mutsun would most likely look under "these" rather than "this".

4.5.5. *Citation forms vs. underlying forms.* Because the Mutsun dictionary is primarily for community use, we choose headword citation forms based on the surface form of a word as it is used in isolation, rather than its underlying form according to linguistic analysis. Mutsun has several cases of allomorphy, such as vowel length alternations conditioned by syllable structure and position in the word, metathesis conditioned by the following suffix, final *h*-deletion, etc. We encode regularly conditioned allomorphs in a field of the lexical database entry, and an introduction to the dictionary explains their usage, while chapters of the separate language textbook (in progress) introduce the allomorphy in use.

For example, many verbs (Okrand's verb classes two, four, and six, 1977: 114) alternate between a vowel-final and a consonant-final form of the verb (sometimes with additional changes in length of a segment), with the consonant-final form appearing only before the reflexive suffix *-pu* or the reciprocal *-mu*. An example is *mattalpu* "to lie on one's stomach", but *matlanu* "to put someone else in a prone position" with the positional causative *-nu* (Okrand 1977: 225). Despite its limited distribution, Okrand chooses the consonant-final form as the underlying form based on which forms can be predicted from which. However, since the consonant-final form appears only before two suffixes, we choose the vowel-final form as the citation form, and the morphology field contains the consonant-final form so that a reader can see that this is an alternating verb.

Mason/Arroyo and Merriam record many of the distinctions involved in allomorphy inconsistently (e.g. presence or absence of *h*, vowel length). From the verb templates (Okrand 1977: 114), it is often clear that a Mason/Arroyo or Merriam verb form must have a long vowel that was not recorded. For example, Mason lists a verb *homo* "to skin, to take off hide". The verb templates with just two non-identical consonants are C_1VC_2VV , C_1VVC_2V , and $C_1VC_2C_2V$. Thus, if Okrand's templates are correct, as they seem to be for Harrington's data, the underlying form could be *homoo*, *hoomo*, or *hommo*, but not *homo*. Since word-final long vowels are shortened in polysyllabic words (Okrand 1977), an underlying form *homoo* could indeed surface as the recorded form *homo*. However, unless one finds an attestation of this stem followed by an appropriate suffix (e.g. *homoona*, *hoomona*, or *hommona* "go to skin" with andative *-na*), one cannot tell whether a final underlying long vowel has been shortened (as it should be), or whether Arroyo has simply neglected the vowel or consonant length. In this case, Arroyo uses the stem in a word *homonin* "skinned", so the short vowel is probably a transcription error. There is no way to determine which segment should be long at least until more data is entered into the database, so we use the attested form (here *homo*) as the citation form, even though it does not conform to the possible verb templates. As more data is entered into the text database (discussed below), each form is checked against current entries and such corrections are made.

4.5.6. *Whether to list suffixes as headwords.* Should suffixes also be listed as headwords in the dictionary? The primary medium for learning of the suffixes will be a language textbook and language classes, but it might be helpful to also be able to look up suffixes

in the dictionary. For linguists' purposes, it seems that each morpheme of the language should have an entry in the dictionary.

Suffixes require more explanation of how to use them than an entire word, and it is often difficult to translate them without using grammatical terminology that most users will not understand. We currently list suffixes as headwords in the dictionary. We use the 'usage' field to include an explanation of how to use the suffix (e.g. "attach to a noun to make a verb that means you have the noun" for the possessive verbalizer *-te*), and an example sentence is also included. When possible, we give the English translation in non-technical prose rather than in grammatical terms (e.g. "at, to, toward" instead of "locative"). However, we are currently considering moving all of the suffix headwords to a separate suffix listing outside the main body of the dictionary. Particularly for the English-Mutsun part of the dictionary, having entries such as "benefactive" alphabetized between content words is not helpful.

4.5.7. Borrowings. Lichtenberk (2003) discusses whether to include words borrowed from a dominant language in dictionaries of indigenous languages. He argues that omitting borrowings misrepresents the language as it is spoken. For his dictionary of Toqabaqita, he chooses to include those borrowings that are pronounced with nativized phonology, and to consider non-nativized words to be code-switching rather than borrowing. Duval and Kuiper (2001) argue for construction of a dictionary of borrowings used in Maori, as a separate dictionary.

For Mutsun, even the Arroyo data (collected in 1815, within a few years of European contact) includes some borrowings from Spanish, such as the forms *kiripire*,

kiriwire, *kirivire* for "write" from Spanish *escribir*. Harrington found *kiriwiiri* and *eskriwiiri*. *kirivire* contains the non-native segment *v*, and *eskriwiiri* has a three-consonant cluster, but the other forms are phonologically nativized. We have decided to include Spanish borrowings which were in use in Mutsun when it was still spoken. It is sometimes better to have a borrowed but attested way to express a concept than to have no way to express it. (See Warner et al. (in preparation-b) for decisions on when to use borrowings.) The borrowings are also useful for research on nativization strategies.

We include Spanish borrowings which appear within an otherwise Mutsun sentence, but not Spanish forms from sentences where the speaker appears to code-switch to Spanish for a larger part of the sentence. There are currently 102 borrowed entries in the dictionary, 100 of them from Spanish. (One is from a related language or dialect, and the remaining one appears to be perhaps from English.) The vast majority of these appear only in the later sources (Harrington and Merriam). All but four of these occur at least once as a phonologically nativized form.

4.5.8. *Newly constructed words*. Because Mutsun stopped being spoken regularly before a great many modern objects were introduced, the revitalization project is in the process of constructing large numbers of new words. We have decided to include these words in the dictionary, at least once community members have used them enough times to be satisfied with them. Since the vast majority of the words in the dictionary are from original sources, it might seem strange to list *anSa-ennes* "e-mail" ("far write" plus the instrumental nominalizer) or *ammamsa* "restaurant" ("eat" with locative nominalizer) along with traditional words. However, if Mutsun is to come into daily use again, such

words will be common, and community members need to be able to look them up (Hinton and Weigel 2002). These entries are, of course, marked as "constructed".

5. The hard copy dictionary

5.1. Mutsun-English and English-Mutsun dictionaries

From the lexical database, we can readily generate a hard copy dictionary. We have had to make a few modifications to the Shoebox Multi-Dictionary-Formatter output (SIL, 1998) to meet the needs of both linguists and the community. The Shoebox program generates a professionally formatted dictionary very easily and automatically. However, it primarily generates the language-to-English dictionary. It will generate an English-Mutsun word list as well, but this is intended only as a word lookup list to remind the user where to look in the Mutsun-English dictionary, and it includes only a few fields of the database. Thus, a Mutsun person wishing to find out how to say something in Mutsun would have to look each word up twice.

Corris et al. (2004) discuss the inconvenience of limited finder lists for native communities and learners. For endangered (let alone dormant) languages, the section of the dictionary going from the daily use language to the heritage language is the part that community members most want (Corris et al. 2004). This is logical: if no one is currently speaking Mutsun to you, but you want to learn to talk in Mutsun, and your daily use language is English, you need an English–Mutsun dictionary. Therefore, we force Shoebox to generate a complete English-to-Mutsun dictionary by exporting the Mutsun-to-English database as a text file, using a specially written computer script to reverse and reorganize certain fields, and re-importing the new file into Shoebox, with English as the

field language and Mutsun as the language of translation. We then generate dictionaries separately from the two files. This process has to be repeated each time we wish to put out an updated version of the dictionary, but it provides a thorough English–Mutsun dictionary, and still allows us to organize the main database by Mutsun words rather than English words and to maintain just one database.

One complication involved in creating the reverse dictionary is that most Mutsun words have more than one English translation, and we wish to have all the English translations present in each entry. Therefore, our reversal process splits each Mutsun entry that has multiple translations into several entries, and puts the additional translations into a separate field, which then appears in the dictionary entry preceded by "Other mng.:". This is not standard in most bilingual dictionaries, but we find it very useful. Newmark (1999) also takes this approach for an English–Albanian dictionary. The dictionary entries (Mutsun–English and English–Mutsun) corresponding to the database entry in Figure 1 are shown in Figure 2.

INSERT FIGURE 2 ABOUT HERE

Phrasal translations in English also present a question: under what headword, or headwords, should one list a word meaning "small stone for splitting acorns" (*paakayukis*) or one meaning "act like a flea" (*porti*)?¹⁰ Newmark (1999) discusses the ease with which such translations can be accommodated as multiple entries in a reverse dictionary. We usually list phrasal English meanings under each of the content words of the phrase, unless this puts multiple entries for the same Mutsun word very close to each

other on a page. Thus, *paaka-yuukis* appears under each of "small, stone, split, acorns". (The format is "split – small stone for splitting acorns" etc.) However, *yuu*, translated as "and", "and so", and "and therefore" does not need to appear as three consecutive nearly identical entries, so it is listed only under "and".

Newmark (1999) created his English–Albanian dictionary by automatically reversing a lexical database, as we do. He discovered benefits of automatic reversal in finding sets of words with related semantic content that would not otherwise be listed near each other (e.g. several idioms involving jinxes). In the Mutsun case, not enough is known about lexical semantics for detailed sets of meanings to appear. However, the automatic creation of the reverse dictionary did immediately make it apparent that the Mutsun data includes sets of words with identical English translations and not enough information to distinguish them, such as large numbers of words translated simply as "bird" or another set translated as "break" (see 5.2 below). The lack of detailed lexical semantic information for most words is a major obstacle to reconstructing the language as it was formerly spoken. Lexical semantics is perhaps the most poorly documented aspect of the language. Warner et al. (in preparation-b) discuss the ethical issue of imperfect revitalization and the Mutsun community's choices regarding incomplete information.

5.2. *Which information to include*

Some information which must be maintained in the lexical database for research purposes is not useful for non-linguists learning the language. As researchers, we often need to know which pages of which sources a word occurs on. For research purposes, one also wants to be able to easily compare all variant forms of the word within one record, so all

variants, as well as which source they came from, must appear. We originally included variant forms below headwords in the dictionary entries, but we found that this made the entries cluttered and difficult to read. Corris et al. (2004) document how entries with too much information can prevent beginning dictionary users from extracting any information at all from the dictionary. The notes field kept by the researchers and information on which researchers have edited an entry are of course also not included in the dictionary. All of this information will be included in an electronically searchable database when the project is complete.

We do retain information in the hard copy dictionary about which source the main form of an entry came from, and this can be helpful in language learning.¹¹ There are often multiple choices for a particular English meaning, such as "clean (v.)", for which the current dictionary lists *itko*, *hapu*, *hatki* (all Mason/Arroyo) and *hitwi* (attested by Okrand, Harrington, and Mason/Arroyo). Although the forms are all phonotactically legal, *hitwi* would be the best choice, because it is attested by multiple sources, including Harrington, the most reliable. There is also more information about *hitwi* than the others, such as an example sentence and several other English translations, whereas the other forms have little information, signaling their unreliability. On the other hand, if one were looking up a word for a particular type of plant or animal, one should favor the word given by Merriam (a naturalist).

Currently, the introduction to the dictionary instructs readers in how to choose a form if multiple headwords are available for a single English meaning. However, we are considering eventually removing poorly attested headwords from the print dictionary where better attested forms for the same English translation are available and there is no

additional information to distinguish the forms from each other. It may be dangerous to have four unrelated forms listed for "to clean" with no information to distinguish them. If some community members memorize *itko* and others happen to choose *hitwi*, this is likely to lead to confusion and slow the spread of fluency in the community. However, until all the information from all sources is analyzed, we do not wish to take the drastic step of removing entries from the dictionary. (All information will always be retained in the electronic version, of course.) This problem, specific to dormant or dead languages, is very common in the Mutsun data because Harrington re-elicited the Arroyo material, and his speaker often gave different words from Arroyo's.

We do currently include part of speech information in the dictionary. Because English translations are so often ambiguous as to whether they are nouns or verbs (e.g. "stick", "wash", "hurt"), some way to convey at least noun vs. verb is critical. We considered removing explicit part of speech labels such as "n.", "v." and changing the English translations to "to stick" or "a stick", "to wash" or "the wash", etc. However, this proved impractical because of mass nouns, abstract nouns that do not take "a" or "the", and low frequency noun or verb meanings. The benefit of removing explicit part of speech labels would be that community members with little formal education would not have to interpret part of speech labels, but we judged that the confusion that would result from translations that do not take "a", "the", or "to" easily was not worth this benefit. Corris et al. (2004) mention that many of their beginning dictionary users did not understand that "to" shows that the following word is a verb, so "a/the" and "to" may not provide more help than "n." and "v." do, anyway.

Other than the information we have discussed excluding here, all other fields of the lexical database are included in the hard copy dictionary (as shown in Figure 2). This can lead to complicated dictionary entries, which may prove to contain too much information for most beginning language learners (see 5.4 below).

Munro (1996) discusses the importance of including lexically specific syntactic information in dictionaries of Native American languages. For example, for Zapotec, she finds that the dictionary must show for each adjective which of several constructions the adjective can appear in. For a dormant language, such detailed information is often not available. As was discussed for potentially bound stems (4.5.3 above), we must simply assume that everything in Mutsun is productive and regular unless we have evidence to the contrary.

5.3. Current status of Mutsun hard copy dictionaries

We have generated several intermediate, temporary versions of the dictionary thus far (in 1998, 2002, 2004, and 2005), as data has been entered. The 2005 version reflects a recent major revision of the practical orthography, and is now being distributed in the community. This version includes all the entries in the Mason dictionary, all of the Merriam notes, all forms cited in the Okrand grammar, and the contents of approximately 500 pages of Harrington data. In the long term, we plan to include all recorded information from all sources in the dictionary.

5.4. Usability in the community: Alternative versions and formats

Corris et al. (2004) discuss making a beginners' dictionary that consists of a simple word list, and an intermediate or advanced learners' dictionary with morphological and usage information, example sentences, etc. They, as well as Lichtenberk (2003), mention the difficulty of generating even one dictionary with the limited resources usually available for work on endangered languages. However, the Shoebox software makes it very easy to generate multiple versions of the hard copy dictionary with different fields of the lexical database included in each. This leaves only the cost of photocopying two versions, which is less than the cost of copying an equal number of copies of the larger intermediate version of the dictionary, unless the dictionary is being formally published. Based on Corris and colleagues' observation of the usefulness of a simple word list for beginning learners, we recently generated a simplified beginner's dictionary containing only the Mutsun and English words, part of speech, and example sentences. This is not quite the word list Corris et al. (2004) suggest, but the community leader (the third author) felt that example sentences would be helpful even for beginners. We will test this version in the community and revise it based on community members' feedback.

Corris et al. (2004) found that their beginning users of Australian language dictionaries were unable to cope with much information in the entries at all, and that even advanced learners were confused by any grammatical terms beyond parts of speech. They also found that some learners needed to have all forms of all words, even inflectional forms, listed to be useful. Nesi (1996) shows that even advanced non-native learners of English residing in England are not able to extract grammatical information from example sentences in an English dictionary. Their learners are far more advanced than most learners of dormant languages. Corris et al. conclude that community

members must be trained in dictionary use. We feel that one cannot expect community members to put large amounts of effort into learning dictionary usage when they are already taking on the daunting task of learning a dormant language. Even in community language teaching, overt grammar explanations must usually be minimized in favor of immersion methods. However, it is also clear that without some training in dictionary use, many aspects of the dictionary will not be useful. This issue probably requires a careful balance of how much grammatical information is presented to which learners through the dictionary, and how much through various styles of language teaching.

Electronic versions of the dictionary for community use might work better for beginning dictionary users. Corris et al. (2004) discuss an electronic interface to a Warlpiri dictionary that offers a variety of formats for looking up information and even integrates language puzzles. They found that community members were able to use the electronic tool effectively and enjoyed it. With recent advances in technology, one can imagine many innovative formats for audiovisual dictionaries. However, designing such tools and taking them beyond the stage of simple labeled pictures on a CD, to tools that truly promote fluency and are usable within the community, will require a great investment of resources. Of course, sound files to accompany the dictionary could be a useful feature. For Mutsun, we are currently focussing on providing a hard copy dictionary and providing both audio and text teaching materials, rather than on providing audio with the dictionary itself. Strategies for teaching materials and use of non-print media are discussed in Warner et al. (in preparation-a).

6. The text database and text collection

We began with the lexical database, with words or morphemes as the unit of analysis. However, one would like to be able to make a concordance for any given morpheme, so that one could examine all known examples of the morpheme's use instead of only a few example sentences. We therefore developed a text database, using Shoebox's (SIL 1998) text database structure and automatic parser. To accommodate work with multiple archival sources, we added fields to record the original spelling and original translation, exactly as written on the page, for each source. Entries (e.g. Figure 3) also include the practical orthographic rendering of the form, the corresponding headword for each morpheme, our translation, the parsed morpheme-by-morpheme gloss, and the parts of speech of each morpheme. Location of the sentence in the original source is also encoded, as is information about who has worked on the entry when.

INSERT FIGURE 3 ABOUT HERE

Separate fields for the original spelling, its conversion into practical orthography, and the headword for each morpheme are necessary because of the profusion of variant forms in the data. The conversion into practical orthography converts only predictable transcriptional differences, not differences in how the researcher appears to have heard the word. That is, variant forms are maintained in the practical orthography field. This allows for future research on how particular sources used particular variants, while also relating the original spelling to the main form of the morpheme (the headword) so that information about original spelling does not also have to be kept in the lexical database.

When adding new sentences to the text database, we run the automatic parser, then use knowledge of Mutsun, Spanish, and general linguistics to verify or correct the parse. New headwords and new information about existing headwords are added to the lexical database until it contains all information necessary for a correct parse. Occasional Harrington sentences and many Arroyo sentences simply do not contain enough information to determine an accurate parse. When consultation among the researchers fails to resolve the problem, the entry in the text database is marked "unsolvable" pending further data, and an explanation is added to the notes field. Thus far though, only twenty out of approximately 2500 analyzed entries have been left unresolved. When all of the original source data has been entered into the text database and analyzed, products from the text database will include a hard copy interlinearized text collection and a searchable electronic database. A morpheme concordance may also be useful, although the searchable database may make this unnecessary.

7. Conclusions

We conclude that it is both possible and very important to make dictionaries of endangered or dormant languages that are designed to be usable by non-linguist community members. One must often present information in a different format, with different headwords, or in a different transcription than one would for a research-oriented linguistic dictionary. However, linguists have enough practice with interpreting information about language that learning to use a community-oriented dictionary is only a minor inconvenience for them, while a research-oriented dictionary may be useless in the community. Our project demonstrates that one and the same database can generate

products for several levels of community language learners (depending on how many database fields are included in the print dictionary) and for research purposes (through a searchable electronic database including all information). In addition, a concordance or search of the text database will eventually supplement the dictionary for research purposes, providing more information than can be included in the dictionary.

The problems encountered in creating a dictionary for community purposes, especially when it is for a dormant language and when the information must come from several disparate archival sources, will vary greatly depending on the language and the original sources. There are similarities between this type of project and both community-oriented endangered language lexicography and lexicography for dead languages. We hope that by providing an example of how we have analyzed and organized the information on one such language, Mutsun, others will be encouraged to undertake similar community-oriented lexicography projects.

Acknowledgements

We would like to thank the following individuals and organizations, all of whom have contributed to the Mutsun dictionary. They are listed alphabetically: Keith Alcock, The Amah Mutsun Tribal Band, Paul Barney, Luis Barragan, Tania Granadillo, Birgit Hellwig, Leanne Hinton, the Mutsun Language Foundation, Alina Twist, the University of Arizona (Department of Linguistics, the Library), and the University of California Berkeley (Department of Linguistics, the Bancroft Library). We also thank two anonymous reviewers for their helpful comments. The Woodrow Wilson Foundation and the National Endowment for Humanities have provided partial funding for the project. Finally, we express our appreciation to all the members of the Mutsun community who are participating in relearning Mutsun.

Notes

¹ We prefer the words "dormant" or "sleeping" to "extinct" or "dead" for languages with a community but no living speakers. Hinton (2001) uses "sleeping" to describe these languages. We distinguish between "dormant" languages such as Mutsun and "dead" languages such as, as far as we know, Gothic: dormant languages have a heritage community that would like to bring the language back into use, and there is enough data about the language that bringing some form of it back might be possible. Furthermore, community members often do not like to hear their languages or cultures described as "extinct," and "extinction" implies a permanent status.

² The practical orthography currently in use for Mutsun is shown here. Symbols can be written double to indicate distinctive length. See Warner et al. (in preparation-a) for discussion of how this orthography was developed and why these symbols were chosen.

	Labial	Dental	Palatal-ized	Alveo-palatal	Retro-flex	Velar	Glottal
Stops	p	t	tY		T	k	'
Affricates		ts		c			
Fricatives		s		S			h
Nasals	m	n	N				
Liquids		l	L				
Flap/tap		r					
Glides	w		y				

	front	central	back
high	i		u
(lower-)mid	e		o
low		a	

- denotes separation between stems and clitics, and between parts of compounds.

³ Throughout the paper, forms in the practical orthography (which is also a phonemic transcription) are printed in italics. Where the original spelling used by a source is at issue, original spellings are in single quotation marks. The practical orthography does

not use IPA symbols, but it is a phonemic transcription, and maintains all distinctions of the language.

⁴ Even Harrington, an excellent phonetician, often re-elicited a word with *T* or *c* multiple times to be sure of which sound he was hearing.

⁵ Although Harrington's phonetic level transcription is not problematic for linguists, it causes problems for community members working with the original notes. Non-linguists tend to feel that each symbol should be pronounced differently. When the differing symbols are representing primarily free variation, literal interpretation of the transcriptions will lead community members to attempt to produce a wide range of artificial distinctions, occurring arbitrarily in any particular word. It is important to train community members who work with original sources not to take every detail of every transcription as a requirement, but also to train them in which non-English distinctions are important to maintain.

⁶ A reviewer points out that it may be surprising to find older forms apparently inserting sounds. However, Mutsun is known to have had some type of short, unclear epenthetic vowel in some consonant clusters (Okrand 1977 p. 66-67), and Harrington generally did not transcribe these, while Arroyo often did. This may explain the extra vowel in *akkara*, and many of the other longer Arroyo forms include suffixes or otherwise represent a string of morphemes.

⁷ *ahen* may correspond to *haahen*, with omission of *h* and of vowel length. The final *-n* in *haahen* is the final allomorph of the inherent reflexive suffix *-ni*, so the original spelling *aheniakken* might be correct, with a vowel rather than a glide. The *-ak* could be the third person singular clitic pronoun. However, this leaves the final *-ken* completely

unclear, and if *-ak* is the clitic pronoun, then the *-n* allomorph, rather than the *-ni* allomorph, would be expected. *aheniakken* is far too long for the disyllabic verb templates, so there must be additional morphemes in this form.

⁸ Providing forms that may be changed after further data analysis means that learners must often relearn corrected forms of words they have learned in the previous form. This is obviously inconvenient, but the Mutsun community prefers this to waiting until all archival analysis is complete before beginning to learn the language. Indeed, we doubt that motivation for the project could be sustained if the community were asked to wait on revitalization for several years while linguists complete archival analysis (Warner et al. in preparation-a).

⁹ This reduction is also found with some of the objective pronouns. Its apparent restriction to function words suggests that it may have been conditioned by word frequency or prosodic reduction.

¹⁰ *porti* consists of *poor* 'flea' and the verbalizer *-ti*, so the literal translation is "to be a flea".

¹¹ This does not provide all the information about sources. If the main phonological form comes from Harrington, but Mason/Arroyo and Merriam also have the word, in a variant form, the source information will mention only Harrington. Since variant forms are not included in the hard copy dictionary, a dictionary user would not be able to see that this word is attested by three sources. This is unlikely to be a problem: any dictionary user who is advanced enough to use information about how many variants are attested by which sources could be working with the electronic version containing all information.

References

A. Dictionaries

Young, R. W., and Morgan, W., Sr. 1987. *The Navajo Language: A Grammar and Colloquial Dictionary* (Revised Edition). Albuquerque: University of New Mexico Press.

B. Other literature

Abbott, C. 1998. 'Lessons from an Oneida Dictionary.' *Dictionaries* 19: 124-134.

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, Inc., New York, and subsequently by Coyote Press, Salinas, CA.)

Arroyo de la Cuesta, F. 1862. *A Vocabulary or Phrase Book of the Mutsun Language of Alta California*. *Shea's Library of American Linguistics*, 8. New York: Cramoisy Press. (Reprinted 1970 by AMS Press, Inc., New York, and subsequently by Coyote Press, Salinas, CA.)

Berman, H. 2002. 'Merriam's Palewyami Vocabulary.' *International Journal of American Linguistics* 68: 428-461.

Callaghan, C. 1975. 'J. P. Harrington: California's Great Linguist.' *Journal of California Anthropology* 2: 183-187.

Callaghan, C. 1997. 'Evidence for Yok-Utian.' *International Journal of American Linguistics* 63: 18-64.

- Callaghan, C.** 2001. 'More Evidence for Yok-Utian: A Reanalysis of the Dixon and Kroeber Sets.' *International Journal of American Linguistics* 67: 313-345.
- Corris, M., Manning, C., Poetsch, S., and Simpson, J.** 2004. 'How Useful And Usable Are Dictionaries For Speakers of Australian Indigenous Languages?' *International Journal of Lexicography* 17: 33-68.
- Costa, D. J.** 1991. 'Approaching the Sources on Miami-Illinois.' In *Papers of the Algonquian Conference/Actes du congres des algonquinistes, 1991*, 22: 30-47
- Costa, D. J.** 2003. *The Miami-Illinois Language*. Lincoln: University of Nebraska Press.
- Duval, T., and Kuiper, K.** 2001. 'Maori Dictionaries and Maori Loanwords.' *International Journal of Lexicography* 14: 243-260.
- Harrington, J. P.** 1922, 1929-1930. The Papers of John Peabody Harrington in the Smithsonian Institution. Microfilm volume II, reels 37-61, and volume IX, reel 17.
- Hill, K. C.** 1996. 'Technical Report On the Hopi Dictionary Project.' *Dictionaries* 17: 156-179.
- Hinton, L.** 1994. *Flutes of Fire: Essays on California Indian Languages*. Berkeley, Calif.: Heyday Press.
- Hinton, L.** 2001. 'Sleeping Languages: Can they be Awakened?' In L. Hinton, K. Hale (eds.) *The Green Book of Language Revitalization in Practice*. Academic Press. 413-417.
- Hinton, L., and Weigel, W. F.** 2002. 'A Dictionary for Whom? Tensions between Academic and Nonacademic Functions of Bilingual Dictionaries.' In W. Frawley,

- K. C. Hill, P. Munro (eds.) *Making Dictionaries: Preserving Indigenous Languages of the Americas*. Berkeley and Los Angeles, CA: University of California Press. 155-170.
- Lichtenberk, F.** 2003. 'To List or Not to List: Writing a Dictionary of a Language Undergoing Rapid and Extensive Lexical Changes.' *International Journal of Lexicography* 16: 387-401.
- Mason, J. A.** 1916. 'The Mutsun Dialect of Costanoan Based on the Vocabulary of De La Cuesta.' *University of California Publications in American Archaeology and Ethnology* 11(7): 399-472.
- Merriam, C. H.** 1902. C. Hart Merriam Papers. Microfilm, Bancroft Library, University of California, Berkeley. Collection number: BANC MSS 80/18 c. Negative number: BNEG Box 1556, reels 69, 74.
- Mills, E. L.** 1981. *The Papers of John Peabody Harrington in the Smithsonian Institution 1907-1957, A Guide to the Field Notes*, vol. 2 (Northern and Central California). White Plains, NY: Kraus International Publications.
- Munro, P.** 1996. 'Making a Zapotec Dictionary.' *Dictionaries* 17: 131-55.
- Munro, P.** 2002. 'Entries for Verbs in American Indian Language Dictionaries.' In W. Frawley, K. C. Hill, and P. Munro (eds.), *Making Dictionaries: Preserving Indigenous Languages of the Americas*. Berkeley and Los Angeles, CA: University of California Press, 86-107.
- Nesi, H.** 1996. 'The Role of Illustrative Examples in Productive Dictionary Use.' *Dictionaries* 17: 198-206.

- Newmark, L.** 1999. 'Reversing a One-Way Bilingual Dictionary.' *Dictionaries* 20: 37-48.
- Okrand, M.** 1977. *Mutsun Grammar*. Unpublished Ph.D. dissertation, University of California, Berkeley.
- Pulte, W., and Feeling, D.** 2002. 'Morphology in Cherokee Lexicography: The Cherokee-English Dictionary.' In W. Frawley, K. C. Hill, and P. Munro (eds.), *Making Dictionaries: Preserving Indigenous Languages of the Americas*. Berkeley and Los Angeles, CA: University of California Press, 60-68.
- SIL (Summer Institute of Linguistics).** 1998. *The Linguist's Shoebox Version 4.02 for Macintosh*. <http://www.sil.org/>.
- SIL (Summer Institute of Linguistics).** 2002. *LinguaLinks: Electronic Helps for Language Field Work*. <http://www.sil.org/>.
- Warner, N., Luna-Costillas, Q. , and Butler, L.** In preparation-a. Language revitalization: Problems and potential solutions from the perspective of the Mutsun language.
- Warner, N., Luna-Costillas, Q. , and Butler, L.** In preparation-b. Ethics and revitalization of dormant languages: The Mutsun language.

Table 1. A subset of our rules for conversion from Merriam's transcription to the Mutsun practical orthography, based on Merriam's description of his transcription and comparison of data with other sources. All rules involving the letter 'e' in Merriam's transcription are listed. Items with more specific environments supercede those with more general environments, hence the rule for 'ew' overrides the general rule for 'e'. The examples listed in the practical orthography are based on attestations from other sources. Differences involving only vowel length are ignored, as Merriam was very inconsistent in transcribing length.

Merriam's transcription	In syllable type	Practical orthography	Example (Merriam)	Example (pract. orth.)
e	open	i, y	ah'-men-ne oo-e-kah	ammani "rain (n.)" uyka "yesterday"
e	closed	e	men	men "your"
ě(h)		e	tre'-pě se'-rěh'	Tippe "knife" sire "liver"
ē		i	ē'r-ā'k	irek "rock"
ee, ēē		i	heen	hiin "eye"
ew		iw (even in closed syll.)	tew-yen	tiwyen "antelope"

```

\lx neppe
\ps dem
\ge 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 nepkam
\ce these (irreg. pl. of neppe)
\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

```

Figure 1. Lexical database entry for the word *neppe* "this".

this *dem. neppe. wattin-ka neppe rukkatkatum.* I go away from this house. **makkese neppe uTTasi.** This one cares for us. *Restrict:* close by; can modify a noun or stand alone as a pronoun. *See:* **nepkam** 'these (irreg. pl. of neppe)'; **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.]

neppe *dem. this. wattin-ka neppe rukkatkatum.* I go away from this house. **makkese neppe uTTasi.** This one cares for us. *Restrict:* close by; can modify a noun or stand alone as a pronoun. *See:* **nepkam** 'these (irreg. pl. of neppe)'; **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.]

Figure 2. Dictionary entries corresponding to the lexical database entry in Figure 1.

```

\id 77
\idA 33
\osA ?Ara inthrisnane rotes?
\t ara inTis nane rotes
\m aru hinTis 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

```

Figure 3. A sentence from Arroyo, as entered and analyzed in the text database. \id: sentence ID number. \idA: sentence location in the source. \osA: original spelling. \t: practical orthographic representation. \m: main form (headword) of each morpheme. \g: gloss. \p: part of speech. \f: free translation. \otA: original translation. \nt: general notes. \ns: record of who has worked on the entry when.