**1. OVERVIEW**

This proposal for the Nahuatl Language Documentation Project (NLDP): Sierra Norte de Puebla represents the continuation and expansion into the Sierra Norte de Puebla of a previous NSF three-year project (#0504164) that has focused on documentation (archival quality digital recordings and time-coded transcriptions along with enrichment of an already extensive lexicon of 10,000 entries) of Guerrero Nahuatl. It will provide important high-quality comparative data (digitally recorded corpus, time-coded transcription, and lexicon) of a second Nahuatl language.[1] The Sierra Norte initiative involves close collaboration among a team of highly qualified and dedicated personnel and institutions (see section 6):

- partnership with Tosepan Titataniske (http://www.tosepan.org; support docs., pp. 1–2), an extremely active indigenous cooperative that has built a cultural center—Kaltaixpetaniloyan, with a library, computer center, and electronically equipped meeting room—that will serve as a documentation production center and house the project results, making them available locally.
- Ojo de Agua Comunicaciones (http://www.laneta.apc.org/ojodeagua; support docs., pp. 3–5), an award-winning video collective (with both national and international awards) from Oaxaca, Mexico, specializing in partnering with indigenous villages both in making educational videos and in training personnel to take control of their own multimedia production processes.
- two highly qualified computational linguists (Mike Maxwell and Bill Poser; see biographical sketches) who will develop and make publicly available an electronic toolset that will provide needed computational support to this project and language documentation efforts in general.
- a renowned ethnobotanist (Robert Bye; support docs., p. 24), ex-director of the Jardín Botánico at the Universidad Nacional Autónoma de México, who will work with Tosepan to create an ethnobotanical documentation center with herbarium specimens and digital recordings.
- a PI (Jonathan Amith) fluent in Balsas Guerrero Nahuatl (Ameyaltepec and Oapan variants) who has extensive experience in Nahuatl language documentation and description, in ethnobiological research, and in working with indigenous communities and training speakers in documentation.

The results of this three-year project will be the following:

- an extensive set of archival quality audio and video recordings (about 125 hours in each format).
- time-coded transcriptions of 125 hours (50%) of the total audio (125 hours) and video (125 hours) material recorded. The archived material will utilize to full effect unique electronic tools—a transducer and text morphological annotation tool—that have been developed to facilitate archiving of time-coded transcriptions in a form accessible to researchers with limited knowledge of Nahuatl: three-line interlinear format in which each text is accompanied by a "mini-dictionary" comprising all the lemmas found in that particular text.
- five 27-minute educational videos in Nahuatl (with English and Spanish subtitles) on a range of topics. These documentaries will be derived from 125 hours of original footage (25 hours per final 27-minute edited distribution format) that will be securely archived.
- a detailed lexicon, similar to that produced by Amith for Oapan and Ameyaltepec Nahuatl,[2] with a minimum coverage of all lemmas in the transcribed texts (estimated to be approximately 7–8,000), along with detailed semantic, grammatical, and etymological information in each entry.
- Web pages (which may be either locally or remotely accessed) with line-by-line playback of documentary recordings, created from Transcriber files by WebScriber (see methodology section).

---

[1] *Nahuatl* is perhaps best considered a group of closely related, but separate languages. *Ethnologue* (2004) lists 26 different Nahuatl languages; the new INALI catalogue (2007) lists 28. Despite deficient methodologies, the gist of these efforts is that Nahuatl is highly diversified with low levels of intelligibility among the many languages or variants. Other than those made available by Amith, few archival recordings/transcriptions of Nahuatl are available to scholars, students, or speakers.

[2] See http://nahuatl.ldc.upenn.edu Username: oapan / Password: nahuatl

- a software toolset (WebScriber, Text Morphological Annotation Tool, Prompter/Segmenter, and Shoebox Utilities, useful to language documentation efforts worldwide.
- Nahuatl materials that will be utilized in less-commonly-taught language courses in the Title VI programs at Yale, U. Chicago, and Columbia, at the least (support docs., p. 14 and p. 15 ).
- three native speakers from the Sierra Norte de Puebla fully trained in documentation skills.
- a regional indigenous cultural center, museum, and herbarium, housed and administered by Tosepan, which has committed the necessary space and personnel (support docs., pp. 1–2)

Tosepan Titataniske has 5,800 members and a network of 30 *promotores* who will facilitate documentation activity by identifying those speakers with the greatest skills in endangered genres of discourse or the most complete knowledge of disappearing domains of cultural knowledge. Tosepan is committed not only to full participation in the documentation aspect of this project; they will contribute important human and physical resources to develop and maintain a resource center and library for all documentary materials; establish a space for housing ethnobiological specimens and assign a person to oversee acquisition, access, and maintenance; and provide a space for the storing and exhibition of cultural materials that will form part of a regional indigenous museum. Partnership with Tosepan offers a unique opportunity not only for successful documentation of endangered genres of discourse and vanishing domains of cultural and linguistic knowledge but for the proper archiving of linguistic and cultural material to benefit the indigenous and nonindigenous population of the region.

**2. RESULTS OF WORK FROM AN ACTIVE NSF GRANT (RESULTS OF FIRST TWO YEARS OF A THREE-YEAR GRANT)**

In 2005, Amith was awarded a Documenting Endangered Language grant (#0504164) for a three-year project (October 2005–September 2008) entitled "Guerrero Nahuatl Language Documentation and Lexicon Enrichment." Years 1 and 2, now concluding, were dedicated to language documentation in the primary region of study: the Balsas River valley of central Guerrero, comprising 15 Nahuatl-speaking communities. Year 3 will expand documentation efforts to neighboring regions to the east. This new proposed project will build upon the experiences acquired and the human and electronic resources developed in the Guerrero project and apply them to a Nahuatl spoken in the Sierra Norte de Puebla.

***Recording, Archiving, and Transcription:*** Archiving indigenous language materials at both the Archive of Indigenous Languages of Latin America (AILLA) and the University of Chicago Language Archives (UC-LA) has been a continual process throughout the grant period. At present the AILLA includes 275 items archived by Amith, all but 54 were transferred as a direct result of the NSF grant. Much additional material will be uploaded in the next several months. Of 280 Nahuatl recordings presently available on the AILLA server, 275 items (98%) were archived by Amith, attesting to the importance of this material and the fact that Amith quickly makes his material available to other researchers. A DEL grant to AILLA to digitize and archive tape recordings made by Jane and Ken Hill (many of which are of Nahuatl) will make other Nahuatl material available. But with continuing fieldwork by Amith it will remain the case that the vast amount of archived and accessible Nahuatl material will be the result of his efforts.

Transcriptions are being made available as they are proofed for errors. The resulting formatted (Word and .pdf documents) transcriptions, derived from a time-coded transcription, will be reimported into time-coded Transcriber format through the WebScriber program (see methodology section). The goal is to provide formatted, time-coded, three-line interlinear transcriptions of all relevant recordings. The final format of each transcription will be archived with a computer-generated mini-dictionary comprising entries for all the lemmas contained in the text.

***Publications:*** Amith will continue collaborating closely with Susan Guion (Univ. of Oregon, Linguistics; support docs., p. 23) to study the unusual tonogenetic phenomenon that characterizes two variants of Nahuatl spoken in the Balsas River valley. An article presenting their initial results of their joint efforts—"Historical origins and acoustic correlates of stress and tonal phenomena in Balsas Nahuatl"—was submitted to the *Journal of Phonetics* by Guion, Amith, Christopher Doty, and Irina Shport (the latter two

are graduate linguistic students at Oregon). It is now being divided into two separate articles for resubmission. These articles will hopefully be an important contribution to an unusual Nahuatl phenomenon. One reviewer commented to this effect:

> At its core, this is a great paper…. The experiments have been well constructed, the acoustic data appropriately (and meticulously) measured, and the relevant statistics very competently executed and reported…. The authors should be encouraged to take a very, very good manuscript and transform it into a potentially landmark paper for those interested in Native American linguistics and issues of tonogenesis.

Another article by Amith alone ("Tone and tonogenesis in Balsas Nahuatl: Accentual patterns from coda *h") is currently under review at the *International Journal of American Linguistics*.

Amith has reached a publication agreement with the Instituto Nacional de Lenguas Indígenas (INALI) to edit the initial volume of a monolingual series of Nahuatl recordings and transcriptions: *Ok nemi totlahtōl. Volumen 1: Estado de Guerrero: San Agustín Oapan, Ameyaltepec, San Francisco Ozomatlán.* The volume comprises six CDs (two from each village) of high quality digital field recordings of children's stories (sample audio at http://www.balsas-nahuatl.org/toolset/seerakokoneetl.mp3), a book of transcriptions, and a brief grammatical introduction to each variant (support docs., p. 10, p. 11, and pp. 12–13). Tentatively, the INALI director, Fernando Nava, has suggested publishing 15,000 copies of this first volume, to be distributed free to Nahuatl-speaking communities throughout Mexico. Funding permitting, the goal is to publish six volumes, approximately 25 percent of the material recorded and transcribed with NSF support. The texts published in this series will be the only ones with audio and transcriptions that fully and accurately represent vowel length in such a wide range of Nahuatl material.

***Native speaker training and human resource development:*** Amith has trained native speakers in language documentation. Three are actively working on the NSF project: Jeremías Cabrera and Emiliana Domínguez (Oapan) and Inocencio Díaz (Ameyaltepec). Inocencio Díaz is a religious official in his community and is not continuing his education outside of the NSF project. Jeremías Cabrera and Emiliana Domínguez, however, are both planning on pursuing degrees in linguistics.

Jeremías Cabrera is the fulltime coordinator of the Guerrero project. He recently graduated with a bachelors degree in information technology; as part of the collaboration of the NSF project with INALI he completed his required six-month social service with this institution. He has trained other native speakers (including Rey Castillo and Plutarco Mejía, Mixtec and Tlapanec speakers, respectively, supported under a Ford Foundation grant to Amith) in language documentation (including the use of Shoebox and Transcriber, field research and recording techniques, and basic sound editing). Cabrera will be a key instructor in the Sierra Norte de Puebla. Subsequently, he plans to enter the masters program in linguistics at the Centro de Investigación y Estudios Superiores en Antropología Social (CIESAS, Mexico City) while continuing as a part-time collaborator on the NSF grant. Cabrera has the potential to complete a doctorate and contribute immensely to Nahuatl language documentation and research.

Emiliana Domínguez had just graduated from middle (secondary) school, with no further educational goals, when she joined the NSF project. Now an accomplished fieldworker and transcriber, she will finish her high school education next year. Domínguez has expressed a strong interest in continuing her education by pursuing a bachelors degree in linguistics at the Escuela Nacional de Antropología e Historia. For her age and educational level, she is probably one of the most proficient native-speaking language documentation workers in Mexico.

The NSF project has supported training native Nahuatl speakers working with John Sullivan at the Universidad de Zacatecas. Three were invited to a two-week intensive workshop on Nahuatl grammar, lexicography, and documentation. The workshop continued with fieldwork in the Huasteca (Hidalgo and northern Veracruz) where 72 recordings were made (now archived at AILLA). Throughout the fall and spring 2006–7 Amith worked with the speakers on transcriptions and a lexicon; he provided them with five months of grant support (at 40% FTE for each) and conducted weekly half-day workshops on

Huastecan Nahuatl lexicography and grammar through Internet teleconferencing (support docs., p. 16, pp. 17–18, from two of these individuals). Amith supported two speakers, Victoriano de la Cruz and Manuel de la Cruz (both from the Chicontepec municipality of Veracruz) in entering the masters degree program in linguistics at CIESAS. He will continue to work with them as they pursue advanced degrees.

Finally, Christopher Doty, a graduate student in phonetics at University of Oregon and a coauthor on the aforementioned paper on Balsas Nahuatl tonogenesis, will pursue a doctorate degree on the phonetics of Nahuatl in Ahuelicán. He has completed the total immersion course in Oapan, Guerrero, that Amith offers every summer through Yale University and in which his documentation materials are used in a pedagogical context. Doty's thesis promises to be one of the first prosodic analyses of Nahuatl, of particular interest in Ahuelicán because of its tonogenetic phenomenon. He has agreed to contribute material to the documentation efforts in Balsas Nahuatl (support docs., pp. 30–31).

***Additional funding and expanded work in other endangered languages:*** Amith has leveraged his NSF grant to obtain complementary funding for language documentation and the dissemination of materials to native speakers. The Ford Foundation has supported him with two grants: "Community and School Outreach for Nahuatl Education and Literacy" (2006; for $49,000) and "Las lenguas indígenas de La Montaña, Guerrero" (2007; for $89,000). The first promotes indigenous literacy and cultural knowledge through Nahuatl textbooks. These funds can be used to disseminate the results of this Sierra Nahuatl project. The second grant allows Amith to extend his work to Tlapanec and Mixtec language communities in Guerrero. For example, Rey Castillo, a fluent Mixtec (Tu'un sávi) speaker from Yoloxochitl, Guerrero, who recently graduated from the masters degree program at CIESAS with a path breaking phonetic and phonological study of the variant of this language spoken in his native village, will work half-time over the next two years on Tu'un sávi documentation and lexicography. Eighteen texts of approximately four hours have already been digitally recorded by Amith and transcribed in time-coded format by Castillo. Both sound and transcription will soon be made publicly available at AILLA.

***Permanent archiving of related nonlanguage materials:*** The NLDP, both the expiring grant focused on Guerrero and this new application focusing on the Sierra Norte de Puebla, stresses documenting endangered knowledge about the natural environment and material culture by recording and transcribing hundreds of texts dealing with these increasingly endangered semantic domains. A concomitant to this methodology is the need to provide for secure and professional depositing of associated cultural and environmental materials (i.e., items about which the texts comment). To date the Nahuatl project has collected 1,151 botanical specimens (all professionally mounted), about 300 entomological specimens, 19 fish, 24 arachnids, and 22 reptiles, and some 50 objects of material culture (nets, baskets, arrows, rope, etc.). Most are linked to relevant recorded and transcribed texts (e.g., the folklore associated with different types of grasshoppers, or how to make and use arrows). Permanent storage of all such material objects will be at the Smithsonian Institution (support docs., p. 9 on zoological specimens) and at the Jardín Botánico, Mexico City. An agreement with the collections manager of the U.S. National Herbarium (support docs., p. 8) will ensure that plants in the Nahuatl ethnobotanical collection will remain accessible as a unit to future researchers. Material cultural artifacts will be deposited in the National Anthropological Archives and duly cross-referenced on the language metadata.

**3. EASTERN SIERRA NORTE DE PUEBLA: THE REGION AND PREVIOUS WORK**

In evaluating Amith's previous NSF proposal—Guerrero Nahuatl Language Documentation and Lexicon Enrichment Project—one reviewer commented:

> In his proposal the PI argues profusely—and convincingly—for the urgent need to document Nahuatl of Guerrero. However, for the exact same reasons *there is no doubt that Nahuatl is equally endangered in all the areas where it is spoken*. It is also likely that *detailed studies of other varieties of Nahuatl will uncover equally interesting typological features peculiar to each one of them, and every single area deserves thorough documentation* [emphasis added].

Likewise, the NSF review panel summary statement for the Guerrero project states that "perhaps some future development can be thought of where some other areas be brought in for comparison or extension, to broaden the impact of the research."

The present project takes up these calls to document Nahuatl "in all the areas where it is spoken" by applying the documentation methodology and the commitment to community and archiving characteristic of Amith's work in Guerrero to the Sierra Norte de Puebla, a region where the indigenous collective and partner in this documentation effort, Tosepan Titataniske, is centered and has concentrated its efforts in indigenous cultural and economic activities. This partnership ensures not only the logistical support necessary for the present project but that the results will be made available locally in a self-sufficient extant cultural center.

**4. PREVIOUS WORK ON SIERRA NORTE DE PUEBLA NAHUATL: DICTIONARIES, GRAMMARS AND CORPORA**

***Dictionaries:*** There are three dictionaries for the Sierra Norte de Puebla. Unfortunately, all of them are deficient in one way or another. Vowel length, if recorded, is unreliable, and often the aspirated phoneme [h] is inaccurately written, if at all. The three relevant dictionaries are Key and Key (1953b) and Toumi (1984) for Zacapoaxtla and San Miguel Tzinacapan, respectively, in the eastern Sierra Norte de Puebla, and Brockway, Hershey de Brockway and Santos Valdés (2000) for Tlaxpanaloya in the western Sierra Norte de Puebla.

The Zacapoaxtla dictionary is best described as a simple word list of some 4-5,000 entries (Nahuatl–Spanish side). No example sentences or grammatical information is given (e.g., parts of speech). A representative entry is "mahuiltia / juega (de jugar)." Vowel length is, as an author admitted, "botched."[3] Toumi's dictionary is similarly flawed; it too is basically a word list, with headwords often followed by a single Spanish gloss and grammatical information as well as example sentences absent. Vowel length is marked though, as a brief examination reveals, probably as haphazardly as in the Zacapoaxtla work. Word-final aspiration ([h], written /j/) is likewise inconsistently and unreliably represented.

The work of Brockway, Hershey de Brockway and Santos Valdés (2000) from the western Sierra Norte near Hidalgo is more complete. Basic grammatical categories are represented and example sentences are included in each entry. However, vowel length representation is not attempted (perhaps given the unsuccessful efforts of the two previously mentioned works) and final aspiration is not accurately represented (e.g., in denominal adjectivals ending in *-yoh*; cf. entry for *soquiyo* 'lodoso' p. 195).

In sum, there is no lexicon that adequately meets the requirements of language documentation: accurate phonological representation; electronic (XML) database; and full grammatical, semantic, and etymological information. The dictionary that will result from this project not only will include accurate representation of vowel length and final aspirations; a detailed semantic account of each headword; illustrative sentences; and sound files (developed with the Prompter and Segmenter described below) linked to each headword and most illustrative sentences. It will also be archived in electronic format following best practices now recognized (e.g., with XML tags). A model of the result is the Nahuatl Learning Environment lexicon being built by Amith.[4]

***Grammars:*** A short, 50-page grammatical sketch accompanies the Brockway, Hershey de Brockway, and Santos Valdés (2000) dictionary from Tlaxpanaloya. Brockway (1979) presents a grammatical sketch, apparently of the same language, structured to answer a specific set of questions designed by the editor of the book, Ronald Langacker. Robinson (1970) presents a tagmemic grammar of the Zacapoaxtla area. Relevant to the present project's focus on the Cuetzalan area, he mentions that "this grammar does not

---

[3] Key Ritche [2000?] comments: "When [our informant] and my husband went over the vocabulary file to write in length for publishing … he probably just wrote it where it felt comfortable…. In my dictionary of Zacapoaxtla Nahuat, which I take full responsibility for, the recording of length is botched so badly!"

[4] See n. 1; links to Lexicon > Dictionary Search Page; see Tutorial. The dictionary is still in development.

attempt to cover in detail the significant variations in speech in the Cuetzalan and Chignahuapan areas" (p. vii). An unpublished masters thesis (Troiani, 1979; obtained in photocopy) is deficient in many ways.

***Texts and recordings:*** It is in the realm of textual production that the Cuetzalan region stands out, due almost entirely to the efforts of the Taller de Tradición Oral, founded in 1973, a group of approximately ten to twelve individuals, most of whom are native speakers of Nahuat from San Miguel Tzinacapan. Two anthropologists—Alfonso Reynoso Rábago and Pierre Beaucage—have collaborated extensively with this group, first Reynoso, who was instrumental in the foundation of the Taller, and then Beaucage, who has worked closely with the Taller since 1984. Three works are important:[5]

- A series of 12 bilingual pamphlets, each with a different Nahuatl story (average about 1500 words) transcribed from recordings (see Taller de Tradición Oral del CEPEC, 1982 [?]).
- An bilingual anthology of these and other stories published by the Taller as *Tejuan tikintenkakiliayaj in toueyitatajuan = Les oíamos contar a nuestros abuelos: ethnohistoria de San Miguel Tzinacapan* (Taller de Tradición Oral del CEPEC, 1994).
- *Maseulaxiujpajmej, Kuesalan, Puebla: Plantas Medicinales Indígenas, Cuetzalan, Puebla* (Taller de Tradición Oral, 1988) a bilingual description of 150 useful plants.

The work undertaken by the Taller de Tradición Oral, with the exemplary collaboration of two nonindigenous anthropologists (Reynoso and Beaucage) is impressive both in its scope and in its methodology. For example, the first two works cited above were based on recordings transcribed with only minimal change to the oral discourse, a documentation methodology well ahead of its time. And the ethnobotanical work is striking in its direct presentation of indigenous texts specifically targeting endangered domains of knowledge. It is work that needs to be built upon.

Nevertheless, there are facets to the Taller project that limit its use for language documentation: (a) there is no audio component; the original cassette recordings have deteriorated over the years to a point beyond recovery;[6] (b) vowel length is not documented in the transcriptions, though it is clearly noticeable in speech; (c) there is no accompanying dictionary or grammar; (d) the material is hard to get and limited in quantity (the original pamphlets are out of print and in no library while the 1994 compilation is found in only 36 libraries worldwide); and (e) there is no durable archival or electronic record.[7]

**5. DOCUMENTATION IN THE EASTERN SIERRA NORTE DE PUEBLA: ARGUMENTS FOR URGENCY:**

This section will present four arguments for the urgency of a documentation project in the Eastern Sierra Norte de Puebla: (1) typological/linguistic; (2) cultural/ethnographical; (3) archival and methodological; and (4) practical or synergistic.

***Typological/linguistic:*** One purpose of documentation is to collect *new* data to help determine how a particular language fits into a universal typological scheme or how it relates to similar languages within a larger group (such as the Nahuatl languages). Much such information will result from the present project and it would be premature (and beyond space limitations) to offer a detailed description of the typological or dialectal position of Sierra Nahuatl (for Nahuatl dialectology see Canger 1980, 1997). One unusual feature is the effect of syllable weight and structure on word formation.

---

[5] More academic works are Beaucage (1973a, 1973b, 1974, 1985, 1995, 1997), Beaucage and the Taller de Tradición Oral (1997), Taller de Tradición Oral and Pierre Beaucage (1987, 1988, 1990), and Reynoso Rábago (2006), an extensive analysis of the recordings that, nevertheless, has little actual Nahuatl text.

[6] This became apparent when Amith gave a workshop to Taller members in Tzinacapan and tried to digitize some of the material. Amith has visited Cuetzalan twice, for 20 days total, to lead workshops on Nahuatl documentation with native speakers of the Taller and Tosepan (support docs., pp. 1–2 and p. 19)

[7] Only one short recording of any Nahuatl from Puebla is archived (an interview and transcription deposited by Jane and Ken Hill at AILLA from La Resurrección Tepetitlan, in the center of the state). For some additional published works on Sierra Norte de Puebla Nahuatl, see bibliography, section c.

For example, in eastern Sierra Norte de Puebla Nahuatl, stems that end in /l/ lose the absolutive *-li* after bimoraic monosyllabic roots (*ta:l, chi:l*, 'land' and 'chile', respectively) or after disyllabic roots (*taxkal* and *tekol*, 'tortilla' and 'charcoal', respectively). The absolutive is retained, however, after monomoraic roots (*pili* and *kali*; the geminate sequence **ll* is reduced to /l/) in both unpossessed and possessed forms (thus *pili* 'child' and *nopili* 'my child'). When the stem-final consonant is not /l/ the absolutive *-ti* is retained after unpossessed monosyllabic stems regardless of weight (e.g., *ohti*, *tanti*, *xi:kti*, 'road', 'tooth' and 'navel', respectively; note that a coda does not affect weight). Possessed forms lose the absolutive, even if the resultant stem is monomoraic (*nooh* and *notan*, 'my path' and 'my tooth', respectively).

Verbal inflection in the perfective likewise manifests an unusual patterning based on an apparent constraint on monosyllabic stems. Thus the verbs *chi:wa* 'to do' and *ki:sa* 'to emerge' both add the marker *-k* without loss of the stem-final vowel: *kichi:wak* 's/he did it' and *ki:sak* s/he/it emerged' both of which maintain the disyllabic stem. However, when the stem is reduplicated the stem-final vowel is now lost in the perfective, as in *kichihchi*:w 'he made it'. Likewise when *ki:sa* 'to emerge' is used as an aspectual or associated motion ending, the stem-final vowel is also lost in the perfective: *takwahtiki:s* 'he ate on the way there.' Preliminary evidence thus suggests the importance of syllabic and metrical considerations in Sierra Nahuatl inflectional paradigms. However, the patterns of absolutive loss/retention in nouns and stem-final vowel loss in perfective verbs is not altogether clear and may well vary in different variants or among speakers. The needed research could benefit immensely from a large-scale documentation project.

***Cultural:*** The NSF Program Solicitation notes that "each endangered language embodies unique local knowledge of the cultures and natural systems in the region in which it is spoken," thus clearly recognizing that language documentation is not simply a question of linguistic or typological relevance but of cultural and ethnographic relevance as well. The NLDP has focused and will focus heavily on textual material (recordings and transcriptions) of *endangered genres of discourse and vanishing cultural knowledge and traditions*. It does so by applying, in addition to recording stories, ritual discourse, life histories and testimonials, a methodology that may be called "cultural lexicography," in which native speaker exegesis is recorded and transcribed on hundreds of lexical items. To date (and mostly for Guerrero) these include short ethnobiological texts on each of the hundreds of terms for plants and animals, descriptions of diseases and cures, cooking instructions, and accounts of the production of material culture (e.g., snare and net hunting, lime burning, cotton spinning and weaving).[8] Also urgently being documented—if only for the suddenness of loss that occurs with the death of the last fluent practitioner—is the disappearance of ritual discourse in a wide range of genres. Memorized dance relations or poems that date to the colonial period, bride-asking speeches and admonitions to newlyweds that abound in metaphors for courtship and marriage, shamanistic prayers to the invisible *aires* who seize peoples souls until assuaged to let go, and music for syncretistic line dances that originated in the Middle Ages—all of these disappear with the death of the last practitioner. These ritual texts are not lost piece-by-piece, like the specialized vocabulary of fading cultural practices, but suddenly vanish from one day to the next, as has occurred over the last 50 years with the death of the last "traditional" generation. The loss of endangered genres of discourse and cultural domains of knowledge, be they ritual texts or specialized vocabulary of a dying material culture or lore about the habitat and use of certain plants or animals, is an urgent area for documentation, and one that Amith has already begun as part of the Nahuatl Learning Environment. Significantly, a focus on endangered genres of discourse and fading cultural knowledge in a grassroots educational project will provide an excellent means to articulate the needs of Nahuatl-speaking villagers with the linguistic and ethnographic goals of language and cultural documentation.

***Archival: Permanent and Local:*** In her letter of support Heidi Johnson, project manager at AILLA, refers to Amith as "a model researcher… diligent in archiving his language documentation materials as soon as

[8] Hill and Hill (1980:345) note the narrowing functional range of Nahuatl in the Malinche region and see this as "probably a transitory stage which will lead rapidly to language obsolescence."

possible…. His recordings are of exceptionally high quality [and] the breadth of content in his corpus is also exemplary." She suggests that it might be one of the two most diverse collections in the archive, that "Amith's recordings are always delivered with complete metadata" and that he is "in a class of only three super-depositors." Amith will continue this practice of great textual diversity and prompt archiving in a new NSF project, ensuring that the documentation results are quickly available to researchers.

Amith is, however, trying to go beyond an archiving practice that simply meets the best accepted criteria for storing audio and text material: 48KHz sampling rate, 16-bit recordings; metadata that conforms to OLAC standards; transcriptions in XML format; and careful attention to signing rights agreements with all narrators. He is also working toward archival goals that will ensure that the Nahuatl material is not simply "durable" but easily intelligible to future generations who might want to access the recordings and texts well after Nahuatl has become a dead language. Two facets of archiving strategies will be discussed in the methodology section: (1) to ensure the intelligibility of the archived material the NLDP will aim to archive all sound material with three-line (surface form—morphological parse—gloss) interlinear time-coded transcriptions accompanied by a mini-dictionary comprising all the lemmas found in the particular text; (2) a parallel text format (e.g., with proper punctuation and sentence/paragraph breaks) for use by native speakers and any others for whom the three-line interlinear format is unwieldy.[9]

The NLDP will also ensure that all materials are locally available and maintained. Tosepan, which has a long-standing and growing presence in the Sierra Norte de Puebla, is unique in this respect: it has the resources and institutional continuity to guarantee long-term, secure storage and effective distribution through its cultural center, Kaltaixpetaniloyan. This center will house the results of the present project—language and cultural documentation along with including educational videos and a herbarium.

***Practical:*** The Sierra Norte documentation project is eminently practical: it is assured success and both local and nonlocal impact given the synergistic relationship among all the well qualified participants: Tosepan (an indigenous cooperative), Ojo de Agua (a filmmakers collective), Maxwell and Poser (computational linguists), Roberto Bye (ethnobotanist), AILLA and UC-LA (archiving institutions); Jeremías Cabrera and Emiliana Domínguez (Nahuatl-speaking documentation specialists already trained in the Guerrero project) and Amith (PI). The qualifications of these individuals and institutions (see section 7 and p. 3) will ensure that the project will be highly successful in meeting its documentation, descriptive, educational, and archival goals.

**6. QUALITY OF PERSONNEL: PARTICIPANTS AND INSTITUTIONS**

**JONATHAN AMITH, PROJECT PI:** Amith is an independent scholar whose research and writing is multidisciplinary (anthropology, linguistics, history). He has experience in sound processing, translation, photography, and editing that will enable him to prepare high quality documentation materials. For over 25 years he has worked in Balsas Nahua communities. This includes five years of virtually uninterrupted residence in Ameyaltepec and Oapan, and periodic visits throughout Guerrero and other states. He is very familiar with those genres of discourse that are dying out and the specialized cultural knowledge that is disappearing. Besides Amith's descriptive and documentation work, he is fluent in the Nahuatl spoken in Ameyaltepec and Oapan and familiar with the particularities of variants from many other villages, including those of the Sierra Norte de Puebla. His long-standing commitment to Nahuatl documentation goes far beyond an academic interest. In the mid-1990s he worked with indigenous artist friends in producing an award-winning book of political art to protest the proposed construction of a hydroelectric dam near San Juan Tetelcingo. Foreseeing long-term involvement in the area, he has bought land in Oapan and has constructed a one-room house with an additional building that has been equipped for use in language documentation and literacy training. Nahuatl documentation is integral to his professional life and career goals. He has various qualifications to carry out documentation projects,

[9] On the necessity, yet challenges, of archiving for different users, see Trilsbeek and Wittenburg (2006).

particularly those enumerated by Himmelmann (1998).[10] His skills and dedication to documentation are attested by many linguists, educators, and archivists (support docs., p. 14, p. 20, pp. 21–22, and p. 23).

**MIKE MAXWELL AND BILL POSER, PROGRAMMERS:** Maxwell and Poser (see biosketches) are highly qualified senior scholars. They will collaborate in areas of their expertise: Maxwell in computational morphology and Poser in sound processing and advanced programming. Maxwell is a senior research scientist at the Center for Advanced Study of Language, University of Maryland, and an international linguistics advisor for the Summer Institute of Linguistics. He has worked closely with Amith on a morphological transducer for Oapan Nahuatl since 2003, most recently as a co-researcher on a Department of Education Grant to the Linguistic Data Consortium (where Maxwell formerly worked) entitled "Teaching and Learning Aids for Morphologically Complex Languages." Poser is phonetician and expert programmer (see his software at http://www.billposer.org/software.html). He has taught linguistics and computational methods for linguists at Stanford (where he had tenure) and the University of Pennsylvania. He has been appointed to the Linguistic Society of America Technology Advisory Committee.

**ROBERT BYE, ETHNOBOTANIST:** Bye has published extensively in ethnobotany and is the project director for a long-term ethnobotanical research project among the Tarahumara in the northern state of Chihuahua, Mexico. He has twice served as director of Mexico's national botanical garden and herbarium. Since 2003 he has collaborated with Amith on research in the Balsas River valley, Guerrero, and is presently developing an ethnobotanical research project, focused on biodiversity and sustainable development, with Amith in the Sierra Norte de Puebla.

**TOSEPAN TITATANISKE:** Tosepan is now in its thirtieth year of operation (see Bartra, Cobo, and Paz Paredes, 2004). It's 5,800 members and 30 *promotores* are both Nahuatl and Totonac speakers, concentrated in the eastern Sierra Norte de Puebla. It has established a credit union (now with 5.5 million dollars in loans to its members, an increase from 160,000 dollars at its inception), a coffee cooperative for organic and fair trade commercialization, and a training and cultural center where it organizes educational activities to benefit its members. Its firm and active commitment to this project is evidenced by the human and material resources it is offering (support docs., pp. 1–2).

**OJO DE AGUA COMUNICACIONES:** Ojo de Agua is an award-winning collective of independent filmmakers based in Oaxaca, Mexico. They have won over a dozen national and international awards, (e.g., two documentaries that tied for first place in the Latin American Environmental Film Festival, New Orleans [2005]). Since 2003 they have been producing a television series "Pueblos de México," now with over 35 episodes. They not only document indigenous culture but have organized over 50 training workshops in indigenous communities to empower them to realize their own multimedia presentations. They have already produced videos with Tosepan (see bibliography, section b). The equipment Ojo de Agua uses is broadcast quality; sound will be recorded in archival quality 48KHz/16-bit (support docs., pp. 3–5).

**AILLA AND UC-LA:** Two of the leading archives for permanent and secure storage of linguistic data (recordings and texts) are the Archive of the Indigenous Languages of Latin America and the University of Chicago Language Archives. Amith has contributed extensively to both, which have duplicate holdings of all his material. He will continue to promptly store his field recordings, metadata, and transcriptions in both locations (support docs., p. 6 and p. 7).

**NATIONAL MUSEUM OF NATURAL HISTORY:** Amith has not only made arrangements to donate all ethno-biological specimens to the NMNH, but to maintain a database record of these so that his ethnobotanical

---

[10] "Ideally, the person in charge of the compilation speaks the language fluently and knows the cultural and linguistic practices in the speech community very well. This, in general, implies that the compiler has lived in the community for a considerable amount of time. Furthermore, the compiler should be familiar with a broad variety of approaches to language and capable of analyzing linguistics practices from a variety of points of view. These demands will only rarely be met by a single individual" (p. 171).

and ethnozoological specimens can be located at any time in the future. The NMNH has ample experience in housing ethnological collections of biological specimens (support docs., p. 8 and p. 9).

**7. METHODOLOGY**

***Field recording, metadata, and intellectual property rights:*** All audio recordings are digital, at 48K and 16-bit, using either a Sonifex Courier or Marantz PMD670 or 671. A unidirectional headset microphone (ATM-75) is generally used, for stereo both speakers are miked. Video recording will be with a Sony DV Cam, model DRS 370 of broadcast quality and operated by professional filmmakers. Wireless Sennheiser microphones will be used, lavalier for speakers and a boom mike for ambience. All audio on videos will be recorded at 48K and 16-bit. Postproduction (estimated at 2 weeks for each 27-minute segment) will yield broadcast quality educational videos and archival quality sound. The edited video and original footage will be permanently archived (see letter from AILLA on standards). Of the estimated 125 hours of video recording about half will be selected for transcription over the three-year project duration.

Metadata will be logged at the time of recording following OLAC standards (on the quality of Amith's metadata to AILLA, support docs., p. 6) and then transferred to a database. A separate database file will be kept on all contributors/narrators including age, sex, origin (with documentation if speaker has migrated to a new community). Intellectual property rights will be carefully registered. The purpose and use of the recording will be carefully explained to all contributors, who will be asked to sign a contract that allows for educational, archival, and academic nonprofit use. Rights to any commercial or lucrative venture will remain with the narrator (see example contract, support docs., p. 25 and p. 26). He or she will specifically be asked whether the material can be archived and accessed locally for community and school education. Communities or schools that are given copies of the documentation materials are required to sign a receipt agreeing to use it for educational purposes only (support docs., pp. 27–28).

***Ethnobiology:*** Given the extremely endangered situation of ethnobiological knowledge[11] and the difficulty of obtaining high quality material and data, Amith has paid particular attention to this facet of documentation. Care is taken in obtaining high-quality collections, a fact appreciated by biologists asked to provide scientific determinations.[12] The denotata of ethnobiological knowledge must be identified through scientific determinations in binominal nomenclature. This establishes a type of etic grid that identifies the species to researchers and allows them to engage in comparative study across languages and language variants. To ensure proper scientific determinations, Amith has established a support network embracing over one hundred botanists and biologists around the world, all renowned specialists for particular families or genera. This has already led to the discovery of one new species, a *Ficus* described by Cornelius Berg, an expert on the Moraceae family in Mexico. Recently two other botanists, Tom Daniel (Acanthaceae) and Robert Cruden (Anthericaceae) have mentioned that specimens Amith collected were unlike anything they had seen and are probably new species.

***Time-coded and publishable transcriptions*:** Transcription will be done in either Transcriber or ELAN by native speakers trained in this project. Best practices will be followed in terms of accurate representation of all "so called filled pauses, false starts and self-repair and repetitions."[13] However, experience, based on transcribing over 100 hours of texts, has shown that understanding this initial time-coded line-by-line transcription is often challenging given that it is extremely difficult to segment a sound file according to

---

[11] On the loss of botanical knowledge, see various articles in Maffi (2001). See bibliography, section d.

[12] In joining the advisory board, Paul Berry, director of the University of Michigan herbarium, mentioned that the "specimens in the past have been of top-notchy quality, not only in the botanical sense, but also because of the extensive linguistic and ethnobotanical information associated with them" (support docs. p. 29). Linguists have also commented on the care placed in working with ethnobiological texts and knowledge (see letter of Thomas Smith-Stark, support docs., pp. 21–22).

[13] See Schultze-Berndt (2006).

meaningful and consistent semantic criteria. Rather, prosodic considerations orient segmentation and, therefore, the length and content of transcribed lines. Another problem is that playback of time-coded Transcriber and ELAN files requires either that the programs be installed or that the files be converted to some other form (such as an HTML Webpage).

WebScriber was developed to solve these two problems. To facilitate user-friendly access to the time-coded transcription, WebScriber contains a Transcriber-to-HTML transformation utility that creates Web pages viewable through a browser. The content can either be remotely or locally based; in the latter case it is burned onto a CD. At http://nahuatl.ldc.upenn.edu/webscriber[14] one can view how transformations are created or modified, or link to an actual page so created (e.g., by clicking on the link to Mundo Ramírez Reyes). To solve the problem that the breaks in an initial time-coded transcription do not usually reflect syntactic divisions (thus making the transcription at times extremely hard to read and understand, particularly if lacking correct punctuation), a WebScriber utility was developed by which text can be imported into a time code. This allows a documentation specialist to export a time-coded transcription into text format (removing original time codes); edit the text according to syntactic considerations, (the result of which can be distributed to native speakers as published texts for language learning and revitalization, as the INALI publication will do); and reimport the finished text, segmented according to syntactic divisions, it into a new time code in which the line divisions/sound segments can be clauses, sentences, or even short paragraphs. The result is a line-by-line time-coded transcription with full punctuation that can then be transformed into HTML by the Transcriber-to-HTML utility in WebScriber.
***Three-line interlinear transcription*:** As the project nears completion the reimported time-coded transcriptions will be processed through Maxwell's XSLT transducer by the Text Morphological Annotation Tool to create the annotation file (parse and gloss) and, subsequently, the three-line interlinear text in archival format. A separate file will explain all glosses (e.g., 1sgS for first-person singular subject). In cases in which usage deviates from a recognized ontology (e.g., GOLD) this deviation will be explained (e.g., that "absolutive" is used in Nahuatl to refer to a suffix marking unpossessed singular nouns, unlike its use in ergative-absolutive languages to mark subjects of intransitive verbs or objects of transitive verbs). Maxwell's transducer has been constructed to facilitate parsing of different Nahuatl languages (see description below) and can be easily adapted for Sierra Norte de Puebla once the lexicon is built and the morphological grammar written (as they will be during the initial stages of this project). For any token in the text, the annotation tool permits notation of spelling problems (which can be immediately corrected), parser failures, and missing dictionary entries. Parser failures and missing dictionary entries can be appropriately tagged and the problem corrected (by fixing the morphological grammar or enhancing the lexicon) before the text is resubmitted to the annotator.

Annotation across a wide range of documentary transcriptions will not only quickly produce a large corpus of time-coded three-line interlinear texts. It will yield a morphological grammar that has been tested against, and reflects, actual speech. And it will enrich a dictionary that at a minimum will contain lexical entries for all words in the corpus. The glosses will include a unique reference number that will be used to automate dictionary lookup of each stem. (For example, the parse and gloss of *nicho:kas* 'I will cry' will be: n-cho:ka-s / 1sgS-to.cry[03840]-fut.sg). The unique record identifier will facilitate automated dictionary look-up. Thus the entire text can be processed so that each stem is looked up, the relevant entry pulled and then stored as an addendum to the parsed text. Future researchers will be presented with a parsed and glossed time-coded transcription with dictionary entries for all words in the text.
***Orthography*:** Amith's orthography is an adaptation of one variant of the system used by several state-level public education systems (by acronym, SEP; Mexico's educational system is decentralized): [h] is represented by /h/ and [w] by /w/.[15] In typing, vowel length is represented by a colon (easier to type and

[14] Site logon: username *oapan* / pwd *nahuatl*. WebScriber logon: username *smedero* / pwd *password*.
[15] Some state systems use /j/ and /u/ (as an onset) and /uh/ (as a coda).

search); in published texts by a macron (easier to read). Amith distinguishes [ku] and [k$^{w}$] by /ku/ and /kw/, here deviating from the SEP orthography. The INALI publication will use Amith's orthography.

Character translation is easy; Amith's orthography can easily be converted to other systems at time of publication or archiving, should a publisher or archivist decide that it is best to this. Poser will use his XLIT program (http://www.billposer.org/Software/xlit.html) to batch process orthographic changes.

***Lexicon:*** Throughout the project an XML lexicon will be developed by native speakers using Shoebox. The goal will be to cover all lemmas found in the corpus. In annotating the texts toward the end of the project all words that do not parse because the stem is not in the lexicon will be tagged as such. The stem will then be added to the dictionary along with the relevant grammatical and semantic information. The dictionary will follow the format used by Amith in his Guerrero work.[16] Bill Poser will develop a Shoebox Utilities toolset for maintaining data integrity (see description below). He will also work with Amith on a schema to ensure consistency in the XML structure (e.g., that all illustrative sentences are followed by an English *and* Spanish translation). Based on the results of running the utilities, the Shoebox file will be corrected and then archived in XML format.

The morphological grammar and the lexicon are basic to the proper functioning of the morphological transducer. The morphological grammar will be written to accommodate variation. For example, in Balsas Nahuatl Oapan has *k* → *h* before all consonants whereas Ameyaltepec has *k* → *h* only before homorganic consonants. Each rule is written into the executable grammar and activated depending on the variant of Nahuatl being processed (see below, discussion of dialect-specific rules in the transducer).

The lexicon must contain all lemmas in the texts to be parsed along with the necessary grammatical information. Thus for the lexicon to be effective across minor variation, headwords must be identified (tagged) according to the Nahuatl variant where they are found. Thus in the Guerrero dictionary *te:postla:lia* is marked as an Ameyaltepec term and *tepostla:lilia* is marked as a Oapan term. They are synonyms for 'to brand' (as an animal). The Sierra Norte de Puebla electronic dictionary will continue to use this multivariant methodology to tag lemmas specific to a given Nahuatl dialect or subdialect.

The transducer also relies on grammatical information (e.g., verb class) to properly parse and generate forms. If in different dialects an identical word belongs to different classes, this will be encoded in the lexicon so it can be accessed by the transducer. For example, in Balsas Nahuatl *pati* 'to get better' (historically **pahti*) is found with the same meaning in Ameyaltepec and Oapan. But in Ameyaltepec the perfective loses the stem-final vowel (*o:pat* 's/he got better') whereas in Oapan it adds the *-k* marker (*o:patik*). The lexicon encodes these dialectal differences in verb class. The transducer uses the tag to place the verb in the correct class for each variant, facilitating correct parsing and generation.

Finally, as texts are run through the transducer at the end of the project, some words will continue to not parse. If they are deemed unparseable because they are absent from the lexicon then the lexicon will be augmented. If they are deemed unparseable because the grammar is deficient, it will be corrected. Thus at this final stage of documentation the text morphological annotation tool will facilitate lexical enrichment and the perfecting of the morphological grammar.

## 8. TIMELINE

**Training period: 8 months:** During this initial 8-month (32-week) training period—the shared responsibility of Amith and Cabrera—some documentary activity will take place. The basic lexicon frame (headwords and grammatical categories, some sense definitions) will be created. The target will be a 5,000-entry word-list at the conclusion of this period (average 150 words/week). The list will serve to orient spelling, particularly proper placement of vowel length. Trainees will learn recording techniques and metadata requirements. Approximately one-half of each day will be dedicated to lexicon

[16] For sample entry go to http://nahuatl.ldc.upenn.edu > Lexico > Dictionary Search Page and type in for Nahuatl word—begins with— "kis". The username and passwords are *oapan* and *nahuatl*.

development and grammar study, one-quarter to transcription, and one-quarter to recording and local archiving and database management. Amith and Maxwell will begin constructing an executable morphological grammar that Maxwell will operationalize in Xerox Finite State Tools.
**Intensive documentation: 18 months (1.5 years):** During this approximately 82-week period focus will be on intensive language documentation and transcription as well as lexicon enrichment (words that are new will be added to the lexicon, senses and illustrative phrases will be added and enhanced). Each documentation expert will produce one-hour of highly accurate transcription every two weeks. This liberal estimate includes time for field recordings, first transcriptions and proofing, and lexicon enrichment. The total transcription production will be about 125 hours. During this same period raw footage for the five documentaries will be filmed and postproduction will begin. Amith and Maxwell will intensify development of the transducer, taking full advantage of work already done on Balsas Nahuatl.
**Review: 10 months.** The first 8 months of this period will be dedicated to review and preparing materials for final format. The time-coded transcriptions will be formatted (punctuation, sentence and paragraph breaks, etc.). Each text will then be exported into a time-coded shell using WebScriber. It is estimated that each hour of transcription will require about 30 hours for processing (syntactic formatting and reposting the text into time-coded Transcriber format). The three native-speaker documentation experts should be able to accomplish this in eight months. The final two months will utilize the transducer and text morphological annotation tool to produce three-line interlinear transcriptions (estimated time is one day for each hour of transcription). Amith will fully participate in this phase. If the transducer is accurate there should be no difficulty in creating the annotation files that will generate the interlinear format.
**General:** At all times Amith will be able to conference over the Internet and will dedicate 4 hours a week, every Monday, to a review of the previous week's activity. After the training period Amith will spend 3 months each year in Cuetzalan. Cabrera will spend a similar amount of time, with one month overlap. Thus direct project coordination will occur for 5 months of the year. Ethnobiological work will take place throughout the 3-year duration; intensive collection will be frontloaded to the beginning of the project given the time it sometimes takes in obtaining scientific determinations.

### 9. ELECTRONIC TOOLSET

Of the tools necessary to achieve the archiving and annotation goals set forth in the previous section only the morphological transducer is language specific. The remainder are generic and will be open source.[17]
***WebScriber:*** WebScriber is a Web-based application that presents Transcriber files to an end-user in HTML for line-by-line playback. While Transcriber is a free tool available on many platforms, it still requires downloading and installing software that is not commonly found on machines in the classroom. Additionally, Transcriber is aimed at the creation of a time-stamped transcription but not ideally suited for viewing the output. The complex interface for transcribing audio is an unnecessary burden on end-users who wish merely to view the results. Transforming a Transcriber file to HTML provides a logical means of quickly presenting transcriptions for pedagogical and research purposes. The HTML page can either be created on a server, in which the desired segment is streamed from a single source file, or offline (on a CD), in which case the original source file will have been segmented into individual audio files linked to each line of text.

The uploading of the original audio file (.mp3 or .wav) used during transcription will include a playback option for each Transcriber segment in the HTML interface. Additional options are available to provide a light-weight transcription interface useful for pedagogical purposes such as evaluating a student's transcription or interpretation of individual segments of speech. WebScriber can also import text into a time-coded text created in Transcriber, creating a Transcriber file that can then itself be transformed to HTML.

---

[17] For fuller accounts of the toolset, see http://www.balsas-nahuatl.org/toolset (same username/password).

***Morphological Transducer:*** Maxwell and Amith are currently developing a computational grammar of several variants of Balsas Nahuatl. The morphological and phonological complexity is considerable. In addition to having multiple orders of prefixes and suffixes, there are several kinds of reduplication, and a considerable amount of stem allomorphy that must be accounted for.

The computational grammars are being developed and tested using the Xerox Finite State Tools (Beeseley and Karttunen, 2003). These tools provide a transducer: a tool that can be used for both parsing and generation. One advantage of the Xerox tools over most other morphological parsing or generation tools (including such SIL tools as AMPLE and STAMP, Shoebox/ Toolbox, or PC-KIMMO) is that the Xerox tools allow a linguist to build the grammar using such traditional linguistic concepts as allomorphs and ordered phonological rules. Lauri Karttunen has provided Maxwell with a revised version of the Xerox tools that works with Unicode.

The grammar development mechanism uses a single lexicon and grammar for all Balsas Nahuatl dialects presently under investigation. Dialect-particular lexical variation is stored in separate fields in the lexicon and can thus be extracted automatically when building a parser for a particular variant. While the vast majority of the grammatical and phonological rules are the same for all dialects in the study, there are some dialect-specific rules and affixes. To account for this, dialect-specific rules or affixes are stored next to each other in the grammar files, but flagged by a dialect code. At compile time, a code representing the desired dialect is assigned to a variable, and the appropriate grammar rules are then extracted automatically. This same methodology will be applied to the transducer for Sierra Norte de Puebla Nahuatl, which is morphologically very close to Balsas.

As with any grammar development project, an improvement to one part of the grammar can result in breaking some other part previously working. (This is particularly true of changes to phonological rules.) It is therefore necessary to track the effect of changes to the grammar rules and to lexical entries. For this purpose, Maxwell has a large test suite that is run whenever any major change is made (and frequently at other times). In addition to lists of wordforms that are simply parsed, Maxwell and Amith have found generating paradigms to be useful. These paradigms are output as XML files, from which useful displays (such as HTML) can be automatically built. Both the parsed wordforms and the generated paradigms are compared with previous runs of the test suite, using a version control system to track changes.

The parser is being tested on natural texts, which often contain a word that is a minor variant of a more common form. Fortunately, most such variation can be described by optional phonological rules. We can treat these rules (e.g., intervocalic *k*-deletion) as applying optionally when parsing, but as either not used, or are used obligatorily, when generating. Thus generation yields a canonical form (the most common) but the transducer can parse multiple (including less common) forms. The rules in question are flagged for parsing or generation using a mechanism similar to that used for dialectal variation.

***Text Morphological Annotation Tool:*** The Text Morphological Annotation Tool enables a researcher to interact with a tokenizer and a parser to produce a fully tokenized, fully-glossed and -parsed version of a text. When the user opens a plain text file, it is automatically tokenized. The user reviews the tokenization and, if it is correct, the parser is run over the entire document. A color-coded display indicates whether each token is non-parsing (it should not be submitted to the parser), unparsed (the parser did not return any potential parses), ambiguously parsed (the parser returned more than one potential parse), unambiguously parsed (the parser returned exactly one potential parse), or selected (the user has identified one potential parse as the correct one for this token). Finally, the user steps through each token with one or more potential parses and identifies the correct one.

At each stage, errors can be manually corrected. The user can directly edit the text of a token, or can split or merge adjacent tokens. Once corrections have been made, the parser can be re-run over individual tokens or over the entire document. Error sources such as incorrect tokenizations, parser failures, and missing dictionary entries can be noted in problem notes for each token to provide feedback to the developers of these resources who can use it to correct the morphological grammar or lexicon as needed.

The tool is both configurable and extensible. A configuration file controlling encoding, display, and plug-in settings can be exported to XML and then imported on another machine so that users can share settings. Interlinear display of parse and gloss information can be toggled on and off. At the same time, new interfaces to tokenizers and parsers can be rapidly deployed through a plug-in system. Currently, plug-ins exist for tools that run on a (local) command-line and tools that run on a remote server and are accessed using XML-RPC server. New plug-ins are simple to program and are automatically detected by the annotation tool, making it possible to quickly integrate new tokenizers and parsers into the tool.

The result of processing a text with this tool is an annotation file (in XML, using annotation graph toolkit, a framework for representing linguistic annotations of time series data) that can easily be displayed or printed in a three-line interlinear form: Surface representation—Parse—Gloss.

***Prompter/Segmenter:*** Prompter and Segmenter are a pair of programs that together address the problem of recording large numbers of words and phrases to accompany a lexicon, to which they may easily be linked. Prompter is used to elicit the desired items, resulting in a single audio recording containing all items as well as a log file containing a time stamp for each item. Segmenter is then used to isolate each item in the recording and write it out in a separate, informatively named, sound file.

Prompter reads the items to be elicited from a lexicon, which may be in CSV, Shoebox, or XML format and applies user-specified filters to restrict attention to the desired subset of lexicon entries. It then presents one item at a time, holding the prompt until the user presses the "Next" button. From one to three components may be displayed for each item, e.g. the headword, an example sentence, and the lexical category of the word. The choice of fields to display and which field is associated with which position in the display is freely configurable.

The order in which items are presented is determined by a sort on the user's choice of two keys, e.g. semantic field and citation form. Information about the timing of the recorded items is provided in two ways. First, a log entry is created for each item containing a timestamp with respect to the onset of the recording. Second, at the start of each batch a tone is emitted that will be picked up by the recorder.

Segmenter is a specialized audio editor designed especially for the efficient fragmentation of long recordings into small pieces. It provides multiple views of the audio waveform at different resolutions, the ability to move the window onto the audio waveform in a number of ways, automated setting of cut points at zero-crossings, and automatic generation of output file names or use of names read from a list. Segmenter identifies the synchronization tone emitted by Prompter and can read Prompter log files so as to obtain a list of the timestamps of the beginnings of each item. Two of the mechanisms for positioning the window provided by Segmenter are of particular importance. First, the position of the left window edge can be set to a beginning-of-item timestamp. Second, the window may be positioned with respect to transitions between regions of sound and silence. The regions selected may be written out to sound files directly or placed in a list to be written out later, after further examination.

***Shoebox Utilities Kit:*** The "Shoebox" format and minor variations on it are widely used for lexical databases, not only by users of the Shoebox/Toolbox programs but by others as well. Such databases are frequently edited by human beings, without strong controls by a database manager, so it is common for them to deviate from canonical format. Some such deviations interfere with the use of computational tools; others are likely to reflect errors or missing information. The Shoebox Normalizer will identify deviations from canonical format and help users restructure the database so that it is conformant. It will:

- reorder fields to a canonical order, respecting hierarchy;
- generate a histogram of tags used for use in merging fields not intended to be distinct;
- check the uniqueness of unique identifiers and report duplicates and gaps in the sequence;
- check for the presence of obligatory fields and report records lacking them;
- check for duplication of fields required to be unique and report records with duplicate fields;
- check the validity of cross-references;
- check implicational relationships among fields (e.g., that English glosses are also in Spanish).