Squib

The Formosan Language Archive: Development of a Multimedia Tool to Salvage the Languages and Oral Traditions of the Indigenous Tribes of Taiwan

Elizabeth Zeitoun, Ching-hua Yu, and Cui-xia Weng

INSTITUTE OF LINGUISTICS (PREPARATORY OFFICE), ACADEMIA SINICA

We introduce the newly developed Formosan Language Archive and show how it has been built up through examples drawn from Rukai, a Formosan language spoken across southern Taiwan and including six main dialects (Mantauran, Maga, Tona, Budai, Labuan, and Tanan) that exhibit great variation. After displaying the layout of the archive, we explain how texts and sound files are recorded and digitalized, how words are tagged, and what the purpose of the search system is. Last, we compare the Formosan language archive to other well-established language archives and show how and why we have adopted a layout that is somewhat different but enables us to capture, through thorough linguistic analysis, the variations displayed in the Rukai dialects (and in Formosan languages in general).

1. INTRODUCTION. Out of the 24 or so Formosan languages known to have been spoken up to the twentieth century in Taiwan (Keta[n]galan, Taokas, Papora, Babuza, Favorlang, Hoanya, Siraya, Makattao, Taivoan, Kavalan, Pazeh, Thao, Atayal, Saisiyat, Bunun, Tsou, Saaroa, Kanakanavu, Rukai, Paiwan, Puyuma, Amis, Seediq, Yami),[1] nearly half (the first nine mentioned) are already extinct, and the others are declining rapidly. The possible reasons for language death in Taiwan are diverse but to some extent interrelated: early sinicization of the plain tribes, loss of the languages as a legitimate means of daily communication under a fifty-year governmental policy imposing Mandarin Chinese as the only official language, the passing away of elderly speakers in linguistically still-extant communities, and emigration of younger villagers to neighboring towns.

The Formosan languages exhibit great variation that is still not well understood. Until the mid 1990s, their research was rather neglected. Preliminary studies were made during the Japanese occupation. These laid the foundations for more detailed descriptions. They were followed by a series of descriptions on the synchronic and diachronic phonologies of the Formosan languages as well as discussions of their genetic classification. In the past few years, a renewed surge of interest has caused an

1. Yami is not a Formosan language, but belongs to the Batanic group. It is included in this discussion of Formosan languages because Orchid Island where it is spoken relates to Taiwan politically.

Oceanic Linguistics, Volume 42, no. 1 (June 2003)

influx of studies that have been carried out within different theoretical orientations. However, in this community-shared attempt to salvage the cultures and languages of the Formosan tribes, we are faced with two major contradictions: first, data collection remains a lone enterprise, whose results are usually not shared among the linguistic community. What is published is the product of fieldwork, that is, linguistic descriptions and analyses. Second, due to practical reasons such as time constraints, difficulty in accessing the material at hand, pressure from academic institutions to publish theoretically relevant analyses rather than text collections and other such materials, linguists working on the Formosan languages do not usually transcribe texts but content themselves with recording unrelated sets of sentences. As a result, very few text collections have been published for the Formosan languages.[2]

The Formosan Language Archive has been developed within Academia Sinica under the auspices of the National Science Council. One of its purposes is to collect, conserve, edit, and disseminate via the worldwide web a virtual library of language and linguistic resources permitting access to recorded and transcribed Formosan text collections. A pilot study was conducted in 2001. From 2002 this project has been granted national status and the first project span is five years. It is hoped that by 2006, text archiving of at least nine out of the fifteen extant Formosan languages (including Rukai, Yami, Saisiyat, Tsou, Atayal, Bunun, Paiwan, Amis, and Puyuma) will have been carried out with the help of linguists, engineers, and speakers of the languages themselves.

This paper presents an overview of this newly developed archive and illustrates its build-up through examples drawn from Rukai. Rukai is spoken across the south of Taiwan and includes six main dialects: Mantauran, Maga, Tona, Budai, Labuan, and Tanan, each of which exhibits variation. A comparison of the Formosan language archive with other well-established language archives shows how and why we have adopted a unique layout that enables us to capture, through thorough linguistic analysis, the variation displayed by the Rukai dialects as well as by the Formosan languages in general.[3]

2. BROWSING THE FORMOSAN LANGUAGE ARCHIVE. The Formosan language archive—which includes both Chinese and English browsing display—contains three main types of information databases that we believe will be helpful to the Formosanist and the Austronesianist community as a whole: (1) texts (collected from different languages/dialects), (2) a database of geographic information, and (3) four bibliographical databases. These respective databases make possible all kinds of research and are briefly introduced below.

2.1 TEXTS WITH ANNOTATIONS

2.1.1 Search system. The search system,[4] which consists of relational databases with a web-based interface, enables users to select a single dialect or language and to download recorded texts. Each text is divided into paragraphs and sentences that are

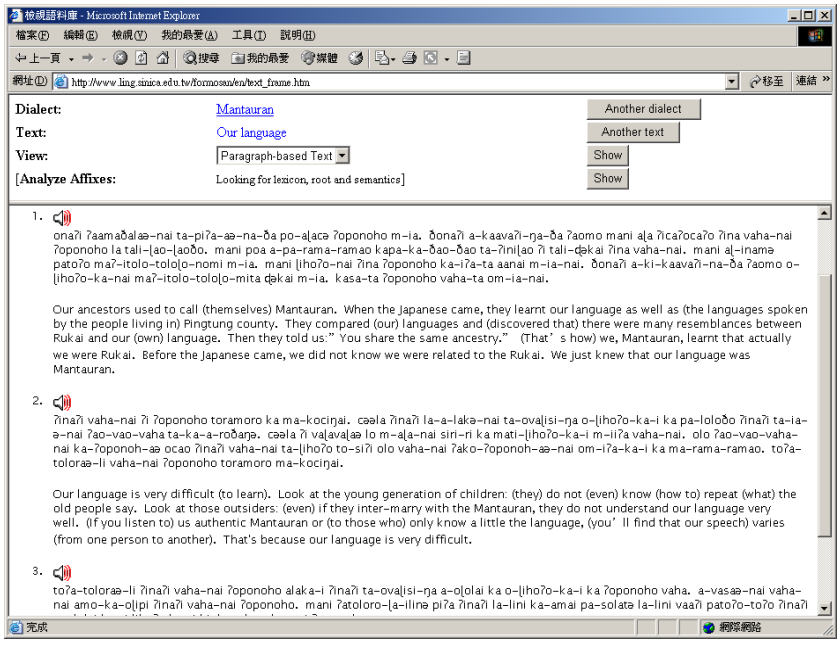
2. To fill that gap, it is hoped that the texts included in the Archive will also be published; see, for instance, Zeitoun and Lin 2003.
3. We would like to express our appreciation to Sander Adelaar, Feng-fu Tsao, and David Blundell for reading and commenting on an earlier draft of this squib.

translated in Chinese and English and can be heard on sound files. In each sentence, every word is glossed, analyzed, and segmented into morphemes. Thus, a user can choose to view a paragraph with its transcription with or without its translation, the transcription of a sentence alone, or the transcription with word glosses and/or translation, and so forth (see figures 1 and 2).

The search system also allows us to exemplify the occurrence and distribution of all the affixes and lexical items recorded in the texts, and to identify their lexical category. These various databases (including textual annotations and mark-ups, a list of affixes, and tagged lexical categories) can be cross-referenced. That is, if a user intends to look for the distribution of a particular affix, examples will be drawn from the main text archive.

As the first-year project (2001) was only a pilot study, only data on Mantauran (Rukai) were analyzed and displayed on the web. The addition of three more Rukai dialects (Tona, Maga, and Tanan) has led us to make changes to the original search system. First, the reedition of Li's (1975) texts has forced us to include in the archive a set of "raw" data, that is, data with partial linguistic annotations, including lexical but no morphemic glosses. Users will soon be able to cross-reference both "raw" and "linguistically annotated/reedited" data. Second, because some Formosan languages (e.g.,

FIGURE 1. PARAGRAPH DISPLAY



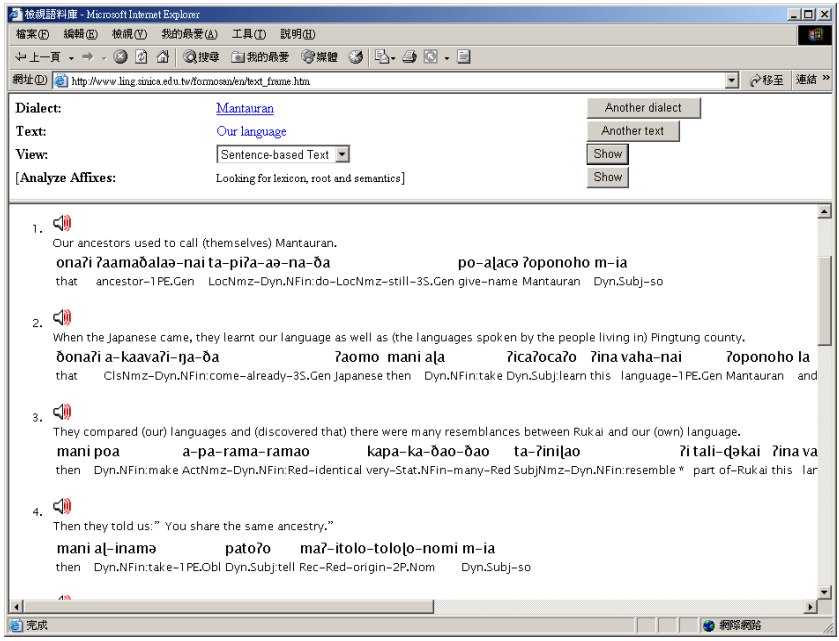
4. The text archive display is to some extent based on the LACITO Archive project directed by Boyd Michailovsky, Michel Jacobson, and John B. Lowe. We are grateful to them for the discussions we had during the preliminary stage of the build-up of this archive.

Maga Rukai) exhibit much morphophonemic alternation that renders opaque any systematic morphemic analysis, both a phonetic transcription (transcripts of sound files) and a phonemic transcription will be made available on the web for such languages.

2.1.2 Metadata. A set of metadata was also developed with the assistance of MAAT[5] in order to facilitate data retrieval and resource sharing during our project in the second year. Metadata is "structured data about data" and it can help archivists manage their data collection through structured archiving, labeling, and cataloging of each text. Users, on the other hand, can access and retrieve data more easily from various archives.

The metadata description is mainly concerned with information regarding the hypertext. That is, it includes information on (i) the profile of the text (title of the text, language and dialect recorded, genre), (ii) fieldwork activity (informants, fieldworkers/transcribers/editors/translators, dates of data collection/editing/translating, credentials, and awards), and (iii) management statements (copyright). In order to unify the term usages in some of the entries of the metadata and in order to reduce the cataloging burden of each text, a language database and a personal information database were built separately. The language database includes information on geographical distribution, dialects and dialectal variations, population, and linguistic practices, while the personal information database provides the name(s) of the infor-

FIGURE 2. SENTENCE DISPLAY



5. MAAT refers to the Metadata Architecture and Application Team affiliated with the National Digital Archives Program in Taiwan, under which Weng Cui-xia works as an assistant.

mants; the language group they belong to; their age, language ability, place of birth, and so on; and basic information on the fieldworkers/transcribers/editors/translators. Hyperlinks further enable users to retrieve more information from the elements "language," "informant," and "participant" in the metadata text.

In order to achieve interoperability across all language repositories, all language resource holders need to have their metadata conform to a standard so that the standard-conformant repositories can share their holdings. Meanwhile, end-users can get access to all the resources that the repositories hold because of the identical metadata sets. OLACMS (Open Language Archives Community Metadata Set) is an emerging metadata standard that has increasingly been applied by language archives. Because we are members of OLAC, our local metadata set is mapped according to the OLACMS principles and guidelines to conform to the community-shared standard, and also to serve as a basis for further data sharing.

2.2 GEOGRAPHIC INFORMATION DATABASE. The geographic search system enables users to determine the geographical distribution of each language and dialect. It is hoped that in the future we will be able to build this system in such a way that it will be possible to observe the expansion or decrease of a particular linguistic community over the last hundred years. Another goal is to provide the mapping of phonemes, lexical items (arranged in different semantic fields), and grammatical words that may allow users to spot the distribution of cognates within the Formosan languages and identify areal features.

2.3 FOUR BIBLIOGRAPHIC DATABASES. The search system is allowed to access the latest information on publications about Formosan languages pertaining to linguistics, language teaching, literature, and music. These four databases are regularly updated.

2.4 FUTURE DATABASES. In the future, we hope to be able to develop other secondary databases, enabling users to make queries regarding Formosan language grammars, conventions (abbreviations, linguistic terms and their definitions) used among Formosanists and on-line dictionaries. Below, we deal specifically with the work associated with text archiving.

3. FROM FIELDWORK TO LINGUISTIC ANNOTATIONS. Before archiving and digitalizing text documents, a prior stage consists, of course, in collecting, recording, translating, and annotating data in the field. This work results in field data collections that include texts, vocabulary lists, grammar notes with examples, and so on. What the Archive contains, at this point, is mainly recorded and translated texts. Thus lexical words and affixes are only retrievable if found in the texts (see 2.1.1).

3.1 THE CORPUS. The corpora consist of texts that have been recorded by trained linguists or linguistically trained native speakers. They include folktales and narratives (mostly the recalling of traditional customs), but as of yet no conversations, songs, or elicited sentences. A total of 117 stories (narratives and folktales) have been recorded

in two languages (Rukai and Yami) that amount to over 600 minutes of sound recordings (see table 1) and are currently being annotated by three linguists (Elizabeth Zeitoun, Tien-hsin Hsin, and Victoria Rau) with the help of two native speaker assistants (Hui-chuan Lin for Mantauran Rukai and Ma-nyu Tung for Yami).[6]

3.2 WHY RUKAI FIRST? Field materials and texts, unless published, are inaccessible to the linguistic community. As head of the Formosan Language Archive project, the first author decided to work on Rukai first because it is a language that has been studied extensively by scholars at Academia Sinica such as Paul Jen-kuei Li (cf. Li 1973, 1975, 1977, 1997), by a former postdoctoral student Tien-hsin Hsin (cf. Hsin 2000, 2002), and by Zeitoun and her assistant (cf. Zeitoun 1995, 1997a,b, 2000a,b, 2002; Lin 1999; Zeitoun and Lin to appear). Victoria Rau also agreed to participate in this project and share the Yami materials she has been collecting over the past several years. Since January 2002, the English translation and the morphemic annotation of the Yami texts has thus been carried out by herself and Ma-nyu Tung (see Tung and Rau 2001, 2002).

3.3 TRANSCRIPTIONS. All the reedited and newly collected texts are transcribed with IPA symbols. There are several reasons for this: (1) there is no standardized writing system for Formosan languages: for example, in Amis, *d* represents the fricative lateral [ɬ]; in other languages such as Paiwan, it is used for the alveolodental [d]; in Thao, *c* is used by Blust (to appear) to transcribe the interdental fricative [θ], while others record [θ] as *th;* (2) use of the IPA corresponds to international standards, as this is the practice in other Archives projects, including the Rosetta Project (see list of Archive Sites following References). For texts that are recorded with a romanized writing system (e.g., Yami), cross-references will be devised to facilitate the correspondence between graphemes and IPA symbols.

3.4 LEXICAL SEGMENTATION: WHY A MORPHEMIC ANALYSIS? The structure of an annotated text is familiar to linguists: it comprises the transcription of the original language, divided into utterances, clauses, or sentences; glosses; and free translations. An example is given in (1). Glosses (or tagsets) can be provided at the word level (stems) or at the morphemic level (roots and affixes). The major difference

TABLE 1. AVAILABLE TEXTS LINGUISTICALLY ANNOTATED IN CHINESE AND ENGLISH AND ON DISPLAY BY DECEMBER 2002

LANGUAGE	DIALECT		FIELDWORKER(S)	CORPUS
	Mantauran	(1)	E. Zeitoun & H.-c. Lin (1992–2001)	14 stories
		(2)	E. Zeitoun & H.-c. Lin (1992–2001)	21 stories
Rukai	Maga		Tien-hsin Hsin (2001)	24 stories
	Tona		E. Zeitoun (1993–2001)	12 stories
	Tanan		Paul Li (1975), reedited by E. Zeitoun (2002)	26 stories
Yami	—		V. Rau and Ma-nyu Tung (1998–2001)	20 stories

6. Text analysis for Saisiyat, Atayal, and Amis is also on-going, with the help of Tai-hua Chu (Saisiyat), Yu-ting Ye (Atayal), and Cui-wei Lin (Amis).

between these two types of annotations lies in the fact that glosses at the word level might provide only a vague interpretation of a word; in the texts that have been collected for Formosan languages, we have found that this interpretation is most often context-based (i.e., subject to the context of the whole sentence). At the morphemic level, on the other hand, roots and affixes, and morphological alternations as well, must be identified and further analyzed.

(1) ðonaʔi a-kaavaʔi-ŋa-ða ʔaomo
那 子句名物化 - 動態 . 非限定 : 來 - 已經 - 他 . 屬格 日本人
that ClsNmz-Dyn.NFin:come-already-3S.Gen Japanese

mani a⌊a ʔicaʔocaʔo ʔina vaha-nai
就 動態 . 非限定 : 拿 動態 . 虛擬式 : 學 這 話 - 我們 . 屬格
then Dyn.NFin:take Dyn.Subj:learn this language-1PE.Gen

ʔoponoho la tali-⌊ao-⌊aoðo.
萬山 和 屬於 - 重疊 - 下面
Mantauran and part+of-Red-below

日本人來了以後就 （開始）學我們萬山和屏東縣的話。
When the Japanese came, they learnt our language as well as (the languages spoken by the people living in) Pingtung county.

We have adopted a morphemic analysis for the annotations of all the Rukai texts for various reasons. First, it enables the annotator to be consistent—words are not "contextually" glossed but their "core" meaning is sought out. Second, it helps to determine the distribution and meaning of nearly every affix, permitting us to construct the affixal database. Third, it deepens one's understanding of the grammar of a specific language, making it easier to identify major lexical and syntactic categories.

For Rukai, the abbreviations used are given in table 2. This list has been adopted for other Formosan languages with minor deletions or additions. Portmanteau morphemes (i.e., single morphemes performing two or more grammatical functions) are indicated by a dot ".", e.g., Mantauran *ða* '3S.Obl' (s/he); distinguishable but not easily breakable affixes or morphemes are separated by ":"; bound forms are linked by a hyphen "-". There has been no attempt to distinguish between affixes and clitics, because such a distinction remains controversial among scholars working on these languages.

3.5 LEXICAL CATEGORIES. For the tagging of major lexical categories, we are following—but with some reservations—the standardization established by China Knowledge Information Processing (CKIP), the unit in charge of the Academia Sinica Chinese Corpora. Not all lexical categories devised by CKIP are found in the Formosan languages, and conversely some lexical categories not listed by CKIP are necessary to describe the Formosan languages, as shown in table 3.

3.6 INITIAL TAGGING AND VALIDATION OF THE TAGGED CORPUS. The corpus is annotated by linguists and/or trained native speakers. There has been no attempt to develop an automatic or semi-automatic linguistic tagger. The reasons for not developing such a tool are as follows. The linguistic data collected for a single dialect are limited—while the number of sentences in recorded texts ranges from

450 to 1,500 sentences, which is about 7,000 words at the most, at least 20,000 words would be necessary to develop an automatic tagger. There are numerous homonymous morphemes. There are various problems regarding the recognition of word and sentence boundaries. Finally, variation among these dialects/languages is tremendous.

Once the data are annotated and tagged and prefixes and lexical categories are identified, it is often necessary for the analyst or the fieldworker to make two or three more fieldtrips to check data inconsistencies and unknown roots or morphophonemic alternations, and to verify linguistic hypotheses on which the tagging of the corpus is based. It takes six months to a year to validate a tagged corpus. A computer program helps to validate the consistency rate (e.g., the number of tagged words or translated sentences in English and in Chinese must be identical) but it is up to the linguist to go through the data frequently and make corrections and revisions.

TABLE 2. ABBREVIATIONS AND OTHER CONVENTIONS

ActNmz	動態名物化	Action nominalization
AgtNmz	主事名物化	Agentive nominalization
Caus	使役	Causative
ClsNmz	子句名物化	Clausal nominalization
Cnc	讓步	Concessive
Cntrfct	違反事實	Counterfactual
Dyn	動態	Dynamic
E	排除式 （= 我們）	Exclusive
Fin	限定	Finite
Gen	屬格	Genitive
Imp	祈使	Imperative
Imprs	無人稱	Impersonal pronoun
I	包含式 （= 咱們）	Inclusive
InstNom	工具名物化	Instrument nominalization
LocNom	處所名物化	Locative nominalization
Nom	主格	Nominative
Neg	否定	Negation
NegImp	否定命令	Negative Imperative
NFin	非限定	Non-Finite
ObjNom	受事名物化	Objective nominalization
Obl	斜格	Oblique
Pass	被動	Passive
P, plur	複數	Plural
Ref	反身	Reflexive
Rec	相互	Reciprocal
Red	重疊	Reduplication
S	單數	Singular
Stat	狀態	Stative
StatNmz	狀態名物化	State nominalization
Subj	虛擬式	Subjunctive
SubjNmz	主語名物化	Subject nominalization
Sup	最高級	Superlative
TempNmz	時間名物化	Temporal nominalization
Top	主題	Topic
1	我 （們）	1st person
2	你 （們）	2nd person
3	他 （們）	3rd person
“.”	帶著兩種功能之詞素	Portmanteau morpheme
“:”	（可區分之）詞綴	(divisible) affix
“-”	接詞	Affix or clitic

3.7 LINGUISTIC AND ANALYTIC PROBLEMS

3.7.1 Linguistic problems. Every text consists of sentences translated from a Formosan language into English and Chinese. The translation from one language to the other must be consistent and equivalent. Such equivalence is sometimes difficult to achieve because of the lexical, grammatical, and typological differences that Chinese and English exhibit as opposed to the Formosan languages.

Because the initial tagging is followed by human validation, the glosses are not without errors, but errors can be reduced through browsing and careful comparison of each lexical item listed in the index derived from each text.

When texts are recorded for dialects of the same language and are further compared, annotator(s) must identify cognates and determine whether or not any semantic shift has taken place. As just one example, *ɬəpəŋə* is glossed as 'finish' in Mantauran and Tona, but as 'all' in Tanan.

3.7.2 Analytic problems. The second problem is analytic in nature and includes the following: the need for (i) proper transcription of texts, (ii) correct identification of sentences and clauses, (iii) appropriate classification of morphemes as bound or free, and (iv) proper recognition of roots and prefixes.

The annotator must, of course, be well trained in linguistics to understand grammatical structures of languages or dialects, considering that a lack of fluency in the language under investigation increases dramatically the difficulties s/he is faced with in annotating the corpus.

3.8 BUILDING UP THE DATABASES. So far, the build-up of one relational database (using Microsoft Access) has allowed us to store our relatively small archive. In such a database, each text, identified as 01 (first text), 02 (second text), and so on, is broken into paragraphs (001, 002, 003) and sentences (a, b, c), associated with Chinese/English translation and sound files. Words occurring in each sentence are allocated a word order and a Chinese-English gloss. Cross-references are made possible through the computational manipulation and recognition of (i) identical numbering of paragraphs or sentences, (ii) orthography, (iii) glosses. To display IPA symbols on the web, a program allows the automatic conversion of IPA symbols into unicode numbers (e.g., [ʔ] = ʔ). A parse program exports our machine-readable format files (i.e., the original files) into the database. (See tables 4 and 5.)

Once the data are converted from a flat-file format into this relational format (see 3.11), the query system is easily executed through the well-defined SQL (Structured Query Language). However, we are aware of the limitations of such a system and as the corpora grow, it will become necessary to migrate our archive into a more powerful client/server database.

3.9 SEARCHING AND BROWSING THE TEXTS. To facilitate the dissemination of the data, the search system is built around Active Server Page (ASP) scripts and standard web browsers. This represents one of the most powerful features of our database. In practice, texts are accessible to any user equipped with a standard browser, a sound card, and the installed Unicode font file. The entry display screen is

shown in figure 3. The user can type a word and view the search result, as shown in figure 4. The search output includes all the sentences containing that word. It is worth noticing that if the word includes IPA symbols—which cannot be found on the computer keyboard—a web-based interface allows the keying of such symbols through the use of a numeric code.

TABLE 3. COMPARISON OF LEXICAL CATEGORIES EXISTING IN CHINESE AND IN THE FORMOSAN LANGUAGES (those in the shaded area are not listed for Chinese but are needed for Formosan languages)*

ABBR.	CHINESE	RUKAI	OTHER FORMOSAN LANGUAGES
1. A	Adjective	—	(+)
2. C	Conjunction	✓	✓
3. ADV	Adverb	—	(+)
4. ASP	Aspect	✓	✓
5. N	Noun	✓	✓
6. DET	Determiner	—	—
7. M	Measure	✓	✓
8. T	Particle	✓	✓
9. P	Preposition	✓	✓
10. Vi	Intransitive Verb	✓	✓
11. Vt	Transitive Verb	✓	✓
12. Post	Postposition	—	✓
13. FW	Foreign Word	✓	✓
14. U	Undecided	✓	✓
15. AUX	Auxiliary	—	(+)
16. NEG	Negator	✓	✓
17. TOP	Topic	✓	✓
18. Tns	Tense	—	(+)
19. MOD	Mood	✓	✓

* ✓= lexical category found in Rukai and/or other Formosan languages; (+) = rare; — = nonexistent.

TABLE 4. DATABASE AT THE SENTENCE LEVEL

LOCATION	C_FREETRAN	E_FREETRAN	SOUNDPATH	TEXTID
001a	我們的祖先自稱是萬山人。	Our ancestors used to call (themselves) Mantauran.	Mantauran/001a.mp3	1

TABLE 5. DATABASE AT THE WORD LEVEL

LOCATION	WORD-ORDER	ORTHOG	CGLS	EGLS	TEXT-ID
001a	0	onaʔi	那	that	1
001a	1	ʔaamaðalaə-nai	祖先 - 我們 . 屬格	ancestor-1PE.Gen	1
001a	2	ta-piʔa-aə-na-ða	處所名物化 - 動態 . 非限定 : 做 - 處所名物化 - 還 - 他 . 屬格	LocNmz-Dyn.NFin:do-Loc-Nmz-still-3S.Gen	1
001a	3	po-aɭacə	取 - 名	give-name	1
001a	4	ʔoponoho	萬山	Mantauran	1
001a	5	m-ia	動態 . 虛擬式 - 這樣	Dyn.Subj-so	1

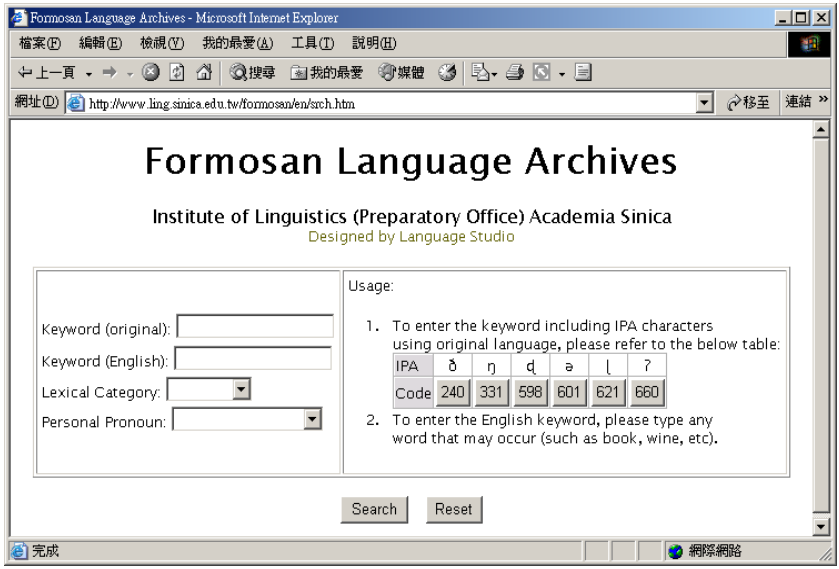
3.10 AUDIO OUTPUT. Audio output is another feature of our archive. To download recorded sentences, we have created an MP3 file for each sentence digitally recorded in the original file. Thus, linguistic workers are able to select, view, and analyze the sound spectrographs, and process (and eventually revise) the sound data using the sound editing software of their choice (such as SoundForge).

3.11 USING XML FOR DATA INTERCHANGE. We have developed a useful tool for converting the archived documents into XML files that can be imported selectively into the database, because XML documents are well-known cross-platform text files for data interchange. A section of one corpus file converted into XML-specific format is shown in figure 5.

4. COMPARISON WITH OTHER LANGUAGE ARCHIVES. Here we make a brief comparison between the Formosan Language Archive and other language archives; the latter are focused on the dissemination of languages with oral literature, on the one hand, and on archiving well-documented languages (such as French, English, or Mandarin Chinese), on the other.

4.1 ARCHIVES ON LANGUAGES WITH ORAL LITERATURE (FIELDWORK IS A "MUST"). In archives that we were able to browse, we found that sound files do not always correspond to the transcriptions given on the texts or are not aligned with the sentences/paragraphs of these texts. In our own archive, we try whenever possible to transcribe texts according to the recorded tapes so that the user can follow the transcription of the sentence/paragraph while listening to the audio

FIGURE 3. THE ENTRY SCREEN OF THE KEYWORD SEARCH



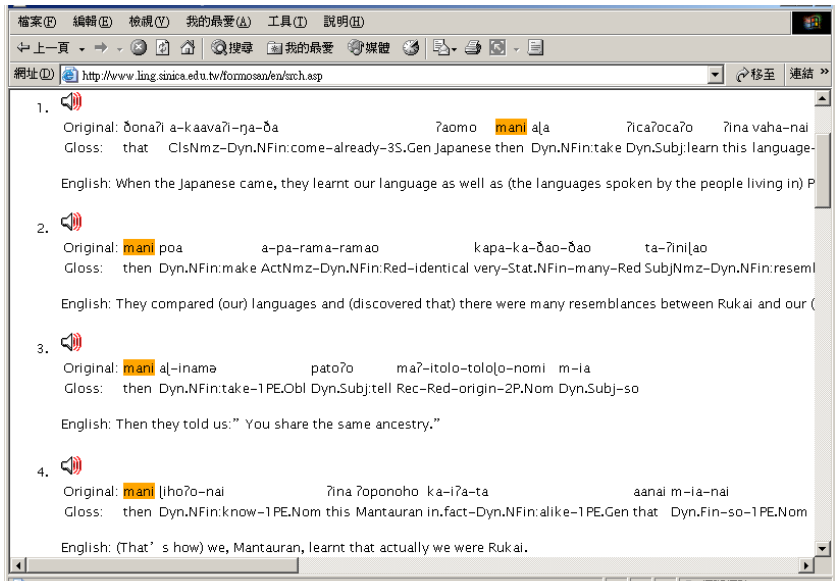
file. Furthermore, texts are sometimes in an Acrobat Reader format, which makes it difficult to browse the data. In this kind of display, the search function for a particular paragraph or sentence is usually not available and there is usually no attempt at cross-referencing the data in the texts to a list of affixes or lexical categories. A search system to find the distribution of lexical items is usually not available either.

These archives, on the other hand, provide more complete extralinguistic information (e.g., additional references on the culture of a particular linguistic community), and some take full advantage of the latest technological developments by providing not only audio files but also video clips. So far, no attempts have been made in either direction in the Formosan Language Archive, due to lack of time and financial resources.

4.2 ARCHIVES ON WRITTEN LANGUAGES (CHINESE, ENGLISH, AND FRENCH). The main differences between well-established archives on Chinese, English, and French, on the one hand, and ours, on the other, lie in the fact that they have access to resources we have been unable to use to date for the documentation of Formosan languages, due to either practical or theoretical reasons.

In their internal format, tagsets are introduced within the texts (as in [2] on page 230), rendering the reading of each text difficult, but still possible for some readers, for instance, native speakers. We believe this kind of display in the web page would decrease dramatically the user's ability to access texts, the structure or meaning of which he or she might not readily understand.

FIGURE 4. SEARCH RESULT



(2) <G>Au_cours_de</G><L>au_cours_de</L></EE><C>PCD></C>
Au cours de ...
In the course of (where G: form, L: root, EE: tag, C: component)
—Extracted from the Talana corpus (web page)

These types of corpora are usually tagged automatically, and then corrections are made by a human resource. We have explained above why we cannot achieve such tagging in our own project.

Finally, linguists working on well-documented languages have explored the possibility of developing a word net (with research on near synonyms, problems of homonymy, problems of translation, etc.), as well as developing treebanks, that is, syntactically annotated corpora. We are still unable to conduct such tasks at this preliminary stage.

5. CONCLUSION. The heritage of the Formosan languages is not only important for historical and typological reasons but also because they exhibit an incredibly rich linguistic variety. Yet we know very little about them, because few texts for each of these languages (not to mention their dialects) have been recorded. The best way for linguists to contribute to Formosan linguistics and to understand the grammatical diversity of these languages is to do fieldwork, and to collect, record, and analyze texts of all kinds.

FIGURE 5. CONVERSION OF A PORTION OF A CORPUS FILE INTO AN XML-SPECIFIC FORMAT

```
<?xml version="1.0" encoding="BIG5" ?>
<header>
…
</header>
<text id="rukai1" lang="rukai">
<s id="001a" textid="1">
 <transcr>
  <w><form>ona&#660;i</form><cgls> 那 </cgls><egls>that</egls></w>
  <w><form>&#660;aama&#240;ala&#601;-nai</form><cgls> 祖先 - 我們 . 屬格
          </cgls><egls>ancestor-1PE.Gen</egls></w>
  <w><form>ta-pi&#660;a-a&#601;-na-&#240;a</form><cgls>處所名物化 - 動態 . 非限定 : 做 - 處
          所名物化 - 還 - 他 . 屬格 </cgls><egls>LocNmz-Dyn.NFin:do-LocNmz-still-
          3S.Gen</egls></w>
  <w><form>po-a&#621;ac&#601;</form><cgls> 取 - 名 </cgls><egls>give-name</egls></w>
  <w><form>&#660;oponoho</form><cgls> 萬山 </cgls><egls>Mantauran</egls></w>
  <w><form>m-ia</form><cgls> 動態 . 虛擬式 - 這樣 </cgls><egls>Dyn.Subj-so</egls></w>
 </transcr>
 <freetran lang="Chinese"> 我們的祖先萬山 （自己是）萬山人。 </freetran>
 <freetran lang="English">Our ancestors used to call (themselves) Mantauran.</freetran>
</s>
…
</text>
</xml>
```

The Formosan Language Archive is a long-term project that can only be carried out if Formosanists agree to cooperate and work together in sharing annotated data they have previously recorded. This task is not simply urgent; it has become crucial.

REFERENCES

Blust, Robert. To appear. Thao dictionary. Language and Linguistics Monograph Series, No. A5. Taipei: Institute of Linguistics (Preparatory Office), Academia Sinica.

Hsin, Tien-hsin. 2000. Aspects of Maga phonology. Ph.D. dissertation, University of Connecticut.

———. 2002. Maga (Rukai) texts. MS.

Li, Paul Jen-kuei. 1973. *Rukai Structure.* Institute of History and Philology, Special Publication No. 64. Taipei: Academia Sinica.

———. 1975. *Rukai Texts*. Institute of History and Philology, Special Publication No. 64–2. Taipei: Academia Sinica.

———. 1977. The internal relationships of Rukai. *Bulletin of the Institute of History and Philology*, 48.1:1–42.

———. 1996. The pronominal systems in Rukai. In *Reconstruction, classification, description: Festschrift in honour of Professor Isidore Dyen*, ed. by Bernd Nothofer, 209–230. Hamburg: Abera Verlag.

Li, Jen-kuei, ed. 1997. *The Austronesian languages of Kaoshiung county* (in Chinese). Series on the materials regarding Kaoshiung county, 7. Kaoshiung: Kaoshiung county government.

Li, Paul Jen-kuei, and Shigeru Tsuchida. 2002. *Pazih texts and songs.* Language and Linguistics Monograph Series No. A2-2. Taipei: Institute of Linguistics (Preparatory Office), Academia Sinica.

Lin, Hui-chuan. 1999. *Let's talk Mantauran*, 1–6. Taipei: Crane Publishing.

Rau, Victoria. 2002. Nominalization in Yami. *Language and Linguistics* 3.2: 165–195.

Tung, Ma-nyu, and Victoria Rau. 2001. *Yami texts* (in Chinese). Taipei: Crane Publishing.

———. 2002. Yami texts. MS.

Yu, Ching-hua. 2002. Discussion on the digitization of the Formosan Language Archive: Building up of the architecture of the archive. Paper presented at the first workshop on the Digital Library Projects, Taipei, July 25–26.

Zeitoun, Elizabeth. 1995. Problèmes de linguistique dans les langues aborigènes de Taiwan. [English version: Issues in Formosan linguistics]. Ph.D. disssertation, Université René Diderot Paris 7.

———. 1997a. Coding of grammatical relations in Mantauran. *Bulletin of the Institute of History and Philology*, 68.1: 249–281.

———. 1997b.The pronominal system of Mantauran (Rukai). *Oceanic Linguistics*, 36: 114–148.

———. 2000a. *A reference grammar of Rukai* (in Chinese). Series on Formosan Languages, 8. Taipei: Yuanliou Pub. Co.

———. 2000b. Dynamic vs. stative verbs in Mantauran (Rukai). *Oceanic Linguistics*, 39:415–427.

———. 2002. Nominalization in Mantauran (Rukai). *Language and Linguistics*, 3.2: 241–282.

Zeitoun, Elizabeth, and Hui-Chuan Lin. 2001. We should not forget the stories of the Mantauran, vol.2: Traditional folktales. MS.

———. 2003. *We should not forget the stories of the Mantauran,* vol.1: *Memories of the past*. Language and Linguistics Monograph Series, No. A4. Taipei: Institute of Linguistics (Preparatory Office), Academia Sinica.

ARCHIVE SITES

http://ckip.iis.sinica.edu.tw/CKIP/
http://lacito.archivage.vjf.cnrs.fr/
http://sino-tibetan.cityu.edu.hk/rda/
http://www.ailla.org/pc/mainindex.html
http://www.ling.sinica.edu.tw/formosan
http://www.rosettaproject.org:8080/live/
http://www.sinica.edu.tw/SinicaCorpus
http://wwwtalana.linguist.jussieu.fr

Elizabeth Zeitoun
hsez@ccvax.sinica.edu.tw

Ching-hua Yu
harryyu@gate.sinica.edu.tw

Cui-xia Weng
cxw@gate.sinica.edu.tw