Computational Linguistics and Chinese Language Processing
Vol. 10, No. 2, June 2005, pp. 167-200

The Formosan Language Archive: Linguistic Analysis and Language Processing

Elizabeth Zeitoun* and Ching-Hua Yu*

Abstract

In this paper, we deal with the linguistic analysis approach adopted in the Formosan Language Corpora, one of the three main information databases included in the Formosan Language Archive, and the language processing programs that have been built upon it. We first discuss problems related to the transcription of different language corpora. We then deal with annotation rules and standards. We go on to explain the linguistic identification of clauses, sentences and paragraphs, and the computer programs used to obtain an alignment of words, glosses and sentences in Chinese and English. We finally show how we try to cope with analytic inconsistencies through programming. This paper is a complement to Zeitoun *et al.* [2003] in which we provided an overview of the whole architecture of the Formosan Language Archive.

Keywords: Formosan languages, Formosan Language Archive, corpora, linguistic analysis, language processing

1. Introduction[1]

The Formosan Language Archive at Academia Sinica[2], Taipei, is part of the Language

* Institute of Linguistics, Academia Sinica, Taipei, Taiwan
E-Mail: {hsez, harryyu}@gate.sinica.edu.tw

[1] Earlier drafts of this manuscript were presented in different occasions and among others, at The Fourth Workshop on Asian Language Resources, March 25, 2004, Sanya, Hainan Island, China and at the Tuesday Seminar at the Institute of Linguistics, University of Hawai'i at Mānoa on Feb.1, 2005. We are thankful to all the participants for their helpful suggestions and comments.We are also grateful to two reviewers for their insightful comments that helped us revise an earlier version of this manuscript.

[2] The Formosan Language Archive is located at: http://formosan.sinica.edu.tw/. The project work team (headed by E. Zeitoun) includes/included* the following assistants and members.

- Language analysis: E. Zeitoun, Hui-chuan Lin, *Tien-hsin Hsin (Rukai)
  Tai-hwa Chu, E. Zeitoun (Saisiyat)
  Yu-ting Yeh, E. Zeitoun
  *Cui-wei Lin (Amis)
  Jia-jing Hua, E. Zeitoun (Paiwan)

[Received November 25, 2004; Revised March 11, 2005; Accepted March 11, 2005]

Archives Project, developed within the five-year National Digital Archives Program (NDAP) and launched in 2002 under the auspices of the National Science Council of Taiwan. A pilot study was conducted in 2001.

The main purpose of our project is to collect, preserve, edit and disseminate via the World Wide Web a virtual library of language and linguistic resources permitting access to recorded and transcribed Formosan text collections, comparative data and related references. Its goal is two-fold: (i) to provide a platform upon which research on various linguistic phenomena can be done through search in the Language Archive and (ii) to develop a pedagogical tool. The first goal has been partially achieved, as will be demonstrated below, but we need more funding and help from the linguistic and non-linguistic community to reach the second goal.

The Formosan Language Archive includes both Chinese and English browsing display on the Internet, and contains three main types of information databases: (1) the corpora of nine Formosan languages with annotated texts[3] (see Table 1 for a list of the digitized texts, as of June 2005) and audio files if available, (2) a geographic information system and (3) four bibliographical databases. The Formosan Language Corpora consist of a trilingual platform, with Formosan texts glossed and translated into Chinese and English. Texts are assigned to four categories: (i) folktales, (ii) narratives, (iii) conversations and (iv) songs with audio files

---

E. Zeitoun, Qiu-yun Liu, Bukun Ismahasan (Bunun)
Lin Zhi-xian (typest)

- GIS: Jia-jing Hua & *Bai Bing-ling
- Engineering: Ching-hua Yu
- Metadata: *Weng Cui-xia, E. Zeitoun, Ching-hua Yu
- References: E. Zeitoun, Qiu-yun Liu, Jia-jing Hua

Most of the assistants working on language analysis (Hui-chuan Lin, Tai-hwa Chu, Yu-ting Yeh, Cui-wei Lin, Jia-jing Hua, Bukun Ismahasan) are aboriginal and have been trained for years (since 1997-1998) in recording and analyzing their own language, i.e., they know how to transcribe and annotate a corpus. All the analyses are supervised by the project director.

[3] The Formosan languages belong to the Austronesian language family, which includes a diversity of languages stretching west to east from Madagascar to Easter Island and north to south from Taiwan to New Zealand. There are still fourteen extant Formosan languages, five of which are moribund and are preceded with an asterisk in the list that follows. While population statistics are available, it is rather difficult to identify the number of speakers for each community. The languages include: Atayal, Amis, Bunun, *Kanakanavu (about a hundred speakers left), *Kavalan, Paiwan, *Pazih (one speaker left), Puyuma, Rukai, *Saaroa (about a hundred speakers left), Saisiyat, Seediq, *Thao (about twenty speakers left), and Tsou. Yami, spoken on Orchid Island (politically part of Taiwan) is genetically closer to the Philippine languages (the Batanic subgroup).

transcribed as faithfully as possible[4]. The Formosan Language Corpora provide different types of search systems -- sentence-based, paragraph-based, concordance-based, keyword-based, affix-based and lexical category-based -- and preserve the original work recorded by earlier scholars by providing two kinds of display, *cf.,* "original data" and "re-edited data," which can be viewed separately or conjointly. The geographic search system permits users to determine the geographical distribution of each language/dialect. It is hoped that in the future, we will be able to further develop this system so that it will be possible to observe the expansion/decrease of a particular linguistic community over the last hundred years. Another goal is to provide mappings of phonemes, lexical items (arranged in different semantic fields) and grammatical words to allow users to see the distribution of cognates within the Formosan languages and identify areal features. The search system for the four bibliographical databases allows access to the latest information in publications about Formosan languages pertaining to linguistics, language teaching, literature and music. The display of the Archive will not be further discussed in this paper, as it has been reported in more detail elsewhere (see Zeitoun *et al*. [2003]).

Table 1. Digitized texts in Chinese and English, as of June 2005

Language	Dialect	Fieldworker and/or analyst	Texts (Stories)	Words	Sentences	Voice file (mp3)	Web Display available
Rukai	Mantauran	1) E. Zeitoun & Hui-chuan Lin [2003]	14	6598	764	60MB	✓
		2) E. Zeitoun & Hui-chuan Lin [1999-2004]	21	7000	1200	65MB	
	Maga	Tien-hsin Hsin [2002]	24	3945	419	50MB	✓
	Tona	1) E. Zeitoun [1993-2001]	12	11281	899	60MB	✓
		2) E. Zeitoun [2003-2004]	8	3400	500	35MB	
	Labuan	E. Zeitoun [2003]	9	650	200	14MB	
	Tanan	Paul Li [1975]	26	10656	1237	--	
Saisiyat	Tungho	1) Chu Tai-hwa [2003], supervised by E. Zeitoun	14	4479	374	30MB	✓
		2) Chu Tai-hwa [2004~], supervised by E. Zeitoun	3	800	250	15MB	
Atayal	Squliq	Ye Yu-ting [2003] supervised by E. Zeitoun	20	10439	1476	80MB	✓
Tsou	Tfuya	Tung *et al.* [1964]	48	9088	1362	70MB	✓
	Tapangu		57	8334	1003	66MB	✓
	Duhtu		29	5589	661	43MB	✓

[4] Texts are recorded in the villages where the informants live (usually either inside or outside their houses). Texts recorded in the Paiwan language have also been video-taped. The informants are free to record narratives, folktales or songs. Conversations only include two speakers.

Amis	Central	Fey *et al.* [1993]	25	50000	1780	200MB	✓
Bunun	Isbukun	Tseng *et al.* [1998]	49	35089	1265	--	✓
Kanakanavu	--	Tsuchida [2003]	10	5961	781	--	✓
Pazih	--	Li & Tsuchida [2001]	31	7590	991	--	✓
Paiwan	Southern	Hua Jia-jing [2004~2005] supervised by E. Zeitoun	20	12000	800	55MB	

The goal of the present paper is to discuss the linguistic analysis approach adopted in the Formosan Language Corpora and the processing programs that have been developed for it. Indeed, the digitization of various Formosan languages and dialects has posed numerous challenges on both the linguistic and computational levels. We have had to develop not only a uniform annotation system to account for language variation and typology but also processing tools for annotating the growing corpus and retrieving and displaying the data from/on the Internet.

This paper is organized as follows. In section 2, we discuss problems related to the transcription of different corpora. In section 3, we deal with annotation rules and standards. In section 4, we turn to the notion of text structure. In section 5, we discuss problems related to analytic and programming consistency. Conclusions are drawn in section 6.

2. Transcriptions

In this section, we first deal with the orthographic system adopted in the Archive and then discuss IPA conversions from one operating environment (Word) to another (Web).

2.1 Orthographic system adopted in the Archive

We first outline the phonemic inventory of the Formosan languages. We then provide an overview of the diverse writing systems that have been used to transcribe the Formosan languages. Finally, we deal with the problems raised by these writing systems, and explain our preference for using IPA for standardized transcription.

2.1.1 Outline of the phonemic inventory of the Formosan languages

The Formosan languages exhibit fairly simple phonemic inventory systems consisting usually of no more than twenty consonants and four vowels, which typically include a series of voiceless and voiced stops: /p, t, k, q, ʔ, b, d, g/; an affricate: /ts/; fricatives: /s, z/; a series of nasals: /m, n, ŋ/; liquids: /l, r/; and four vowels: /a, i, u, ə/. Of course, there is great variation among these languages which has arisen through phonological changes. They will not be detailed in the present paper. Most noticeably, Paiwan has developed a series of palatals: /c, ɟ, ʎ/; Rukai, Paiwan and Puyuma exhibit a partial/full series of retroflexes: /ʈ, ɖ, ɭ/. Atayal, Seediq, Bunun, Paiwan and Thao distinguish between velar and pharyngeal sounds, while

Amis differentiates glottal and epi-glottal sounds [Li 1999]. A few languages such as Squliq Atayal, Tsou, Maga Rukai and Saisiyat have developed more complex vocalic systems. All the consonants and vowels found in the Formosan languages are given in Table 2 below.

Table 2. The phonemic inventory of the Formosan languages

【CONSONANTS】

		labial	Dental	palatal	retroflex	velar	pharyngeal	epi-glottal	glottal
stop	-vd	p	t	c	ʈ	k	q	ʡ	ʔ
	+vd	b ɓ	d ɗ	ɟ	ɖ	g			
affricate			ts						
fricative	-vd	f ɸ	θ s	ʃ	ʂ	x	χ	ħ	h
	+vd	β v	ð z	ʒ	ʐ	ɣ	ʁ		
nasal		m	n			ŋ			
liquid			l ɬ [lh]	ʎ	ɭ				
trill/flap			r ɾ [r]						
glide		w		y					

【VOWELS】

	front	central	back
high	i	ɨ ʉ	u
mid	e	ə, œ	o
low	æ	a	

The basic syllable structure in most languages is CVC, though both Rukai and Tsou now exhibit a CV syllable structure. Consonant clusters occur in only a few languages (e.g., Tsou, Maga Rukai, Thao and Atayal). Stress is usually non-phonemic.

2.1.2 Writing systems adopted to transcribe the Formosan languages during the past four hundred years

Different writing systems (alphabetic, syllabic and logographic) have been adopted to transcribe the Formosan languages during the past four hundred years. Four stages can be distinguished that reflect the history of Taiwan. The last of them is the most complex.

Dutch colonization (1629-1661):

The Roman alphabet was first used in Taiwan in the 17th century by Dutch missionaries to record Siraya and Favorlang. They devised a Romanization system based on the Dutch spelling, which at the time had not yet been standardized.

Chinese colonization (1661-1895):

With the colonization of Taiwan by the Chinese, many land contracts, songs, place or family names and reports were transcribed with Chinese characters. The phonetic value of these Chinese characters is somewhat complex, sometimes referring to Mandarin Chinese and at other times to the Minnan pronunciation.

Japanese colonization (1895-1945):

From 1895 to 1945, Taiwan was a Japanese colony. Aboriginal children were enrolled in schools (up to the age of 12) and learnt Japanese, so they were able, in later years, to transcribe their own language in katakana.

Post-1945:

With the arrival of the Nationalist Chinese under the leadership of Chiang Kai-shek, the Chinese government imposed Mandarin Chinese as the sole official language. The Zhuyin fuhao system more popularly known as Bopomofo, was introduced and used in textbooks, dictionaries etc. At one time, it was also used to transcribe the Formosan languages (e.g., the Bible, songs and textbooks). Bopomofo consists of 37 symbols derived from Chinese characters, and some of these symbols were slightly altered to convey sounds recorded in the Formosan languages that are not found in Chinese. Different writing systems (all Romanized) were devised by the Catholic and the Protestant Church and used during the same period. The lack of adherence to common principles had the unfortunate consequence of producing different writing systems for different tribes. Diacritics were introduced: in Amis, for instance, ^ is used to represent a glottal stop.

In 1991, Prof. Li Jen-kuei [Li 1992] was asked by the Ministry of Education of Taiwan to devise writing systems for the Formosan languages and proposed different solutions (e.g., replacing IPA symbols such as *ŋ* with a capital letter *N* or with two symbols, *ng*).

In 2002, linguists were asked by the Council of [Taiwan] Indigenous Peoples, Executive Yuan, to work in collaboration with each tribe according to their individual expertise and finalize the orthographic system(s) of all the aboriginal languages of Taiwan. This has also led to a variety of Romanized systems that try to improve on the Romanization systems of the Catholic and Protestant Church while taking into account Li's [Li 1992] recommendations.

2.1.3 Problems raised by more recent Romanized writing systems

We will not discuss problems with earlier writing systems (the Dutch-based transcription system and the use of Chinese characters and symbols) as these have been addressed elsewhere (see, for instance, Adelaar [1999] and Rau [1995]). We will, rather, focus on the inconsistencies in the Romanization systems, devised either by missionaries or linguists.

The various Romanization systems devised by missionaries were not usually based on a phonemic representation of the language being transcribed. This had, in many cases, an unfortunate consequence: a relevant phonemic contrast was not represented while other non-distinctive features were taken into account. Early *et al.* [2003:15] showed, for instance, that in Paiwan the orthography used by two Swiss Catholic missionaries did not distinguish between the two phonemes /tj/ and /dj/ but represented both phonemes with a single graph, cf., *tj*. Li [1992:21] also noted that an orthographic system was devised for Paiwan whereby a distinction between /θ/ and /s/ was made, but such a contrast does not exist in that language.

No common principles have been applied to the Formosan languages nor have they been consistently adopted among linguists. In Amis, for instance, *d* represents the lateral fricative /ɬ/, but in all the other Formosan languages, it refers to a dental /d/. Blust (see Blust [2003]) transcribes [θ] as *c* (to show phonological change, PAN *C > Thao θ), while most other linguists transcribe [θ] as *th*. Table 3 provides a comparison of the various symbols used to transcribe the Formosan languages along with their IPA equivalents.

Table 3. Comparison of the various graphs used to transcribe the Formosan languages along with ipa equivalents

【VOWELS】

IPA	GRAPH
i̇	i̇
ɨ̇	ɨ̇
ʉ	ʉ ɨ U
u	u
e	e, é[5]
œ	oe
ə	e
o	o
æ	ae
a	a

[5] This graph is used for Maga Rukai, which has /ə/ and /e/ as distinctive vowels.

【CONSONANTS】

IPA	GRAPH	IPA	GRAPH
p	p	s	s
t	t	ʂ	sh
ts	ts	ʃ	sh
	c		S
ʈ	tr	x	x
	T	χ	h
c	tj	h	h
	t́	ħ	
	t̲	v	v
k	k	ð	dh
q	q		z
	’	z	z
ʔ	’		rh
ʔ	’	ʐ	z
	^	ʁ	R
b	b	m	m
ɓ		n	n
β		ŋ	ng
d	d	ɬ	d
ɗ			l
ɖ	dr		lh
	D	l	l
	rh	ɭ	lr
	r̲		L
ɟ	dj	ʎ	ɬ
	d́		l̲
	d̲		l
g	g		lj
ɣ		r	r
f	f	ɾ	
ɸ		w	w
θ	th	j	y
	c		

To overcome the problem of non-standardization in the current writing systems, we decided to record or re-edit texts in IPA, a recommended standard used in many Archive projects (e.g., the Rosetta Project). However, to preserve the integrity of earlier recorded data, we keep intact original materials recorded with certain Romanization systems and produce

new versions of these based on our own standardized format.

It became necessary for us to make changes in our corpus, as we were including more and more languages. The first languages we started to digitize and to annotate were Rukai, Atayal and Tsou. The commonly accepted use of *c* in Formosan linguistics as a replacement for [ts] seemed at the time the best choice[6], as there are consonant clusters in three of these languages, cf., Maga Rukai, Squliq Atayal and Tsou. However, the introduction of Paiwan, in which there is a distinction between palatalized and non-palatalized sounds, forced us to change our writing policy though, as *c* is the standard IPA symbol used to represent a palatal stop. We thus changed the earlier *c* to *ts*, to distinguish the affricate [ts] from the palatal [c].

Other changes may be needed in the future as we include more languages, but we plan to keep them to a minimum.

2.2 Using Unicode IPA symbols

To convert IPA symbols from Word documents (in which texts are typed) to the Web, we make use of the Unicode encoding system, which offers the possibility of displaying symbols uniformly across browser platforms. In Unicode, each IPA character is assigned a standardized encoding number so as to avoid using the same code for two different symbols. In theory, Unicode represents the best way to display IPA characters on the Web. In practice, it requires an initial configuration. Displaying IPA symbols on certain platforms is sometimes difficult as will be shown below (Webster [2002]).

This section discusses how we use IPA in our two working environments (Word and the Web) and how we convert IPA symbols into a computer-readable form.

Three things are required to convert IPA symbols from word processing documents into HTML files:

1. An operating system that supports Unicode (e.g., Windows 2000/XP).
2. An installed Unicode font that includes IPA (e.g., *SILDoulosIPA* for Microsoft Word, and *Lucida Sans Unicode* for the Web).
3. A Unicode-compliant application (e.g., Microsoft Word or Internet Explorer).

2.2.1 Creating Word processing documents

All the texts included in the Formosan Language Corpora contain different kinds of information: metadata information, utterance identifications, orthographic transcriptions, interlinear word-glosses and free translations. Specific IPA symbols are introduced in the files whenever necessary. We make use of the Unicode-compliant font *SILDoulosIPA*, made

[6] At the same time, we started to analyze data on Saisiyat, a language that has no affricate.

available through SIL. The data is typed as follows:

(1) **ʔinaʔi vaha-nai ʔi ʔoponoho toramoro ka ma-kotsiŋai.**
這 話-我們.屬格 * 萬山 很 * 狀態.虛擬式-難
this language-1PE.Gen * Mantauran very * Stat.Subj-difficult
我們萬山話很難(學)。
Our language is very difficult (to learn). (Zeitoun and Lin [2003, ex. 01-002-a])

Strictly speaking, a Word document is not an ASCII text file, as it may contain formatting code (e.g., indenting, italics, etc.) and IPA symbols, which are challenging for computer processing. It is thus necessary to convert these phonetic symbols into computer-readable forms. Thus the interoperability can be achieved on another application or platform. A macro can be used to transliterate IPA symbols as decimal numeric entities. For instance, the **ð** character is rendered by the HTML code **&#240**. Each IPA symbol is automatically converted into its corresponding numeric reference entity throughout a document. When this operation is finished, we import these alphanumeric characters into the textual database. Once the database has been established, the query operation can be performed as desired.

2.2.2 Creating Unicode IPA Web pages

To display IPA symbols in Web pages, some preliminary work must be done by the user, i.e., his/her computer must be configured with a Unicode IPA font and a Unicode-compliant browser for viewing IPA symbols on the Web. Internet Explorer automatically views web pages encoded with UTF-8, an encoding standard, provided that an appropriate font is installed. As for the font, most Windows 2000/XP machines make use of the *Lucida Sans Unicode* font, which contains the Unicode IPA symbols.

In order to display Unicode IPA Web pages, we declare that the HTML page is using:

(2)
```
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
...
</head>
```

Then, we need to either specify the name of the font locally, e.g., *Lucida Sans Unicode* as:

(3) ð

or declare it globally in the <head> element of the HTML page, for example:

```
(4) <head>
        ...
      <style type="text/css">
        BODY {font-family: Lucida Sans Unicode; font-size: 10pt;}
      </style>
    ...
    </head>
```

Our database keeps track of all the graphs and symbols in each corpus. In other words, not only the Romanization systems but also the numeric reference entities for IPA symbols are stored. This means that when a query is issued from the user's machine, the request is then sent to the server application, which sends the query command to the back-end database, producing a query result that satisfies the initial criteria. The result is then sent back to the server program, which finally produces the HTML output for the user. Our web application is oriented to both browsing and searching the corpus. Either method displays the HTML output, including the IPA codes (if any), and finally displays it in the client browser. If the client computer has the appropriate font installed, e.g., *Lucida Sans Unicode*, then the IPA symbols are guaranteed to be displayed correctly; if not, the user's web browser will display "????" or empty boxes □□□.

2.2.3 Keyword search with IPA symbols

As briefly outlined above (see section 1), the Formosan Language Archive not only permits the browsing of texts, but also allows for searching based on (i) keywords, (ii) list of affixes and (iii) lexical categories. While the search through affixes and lexical categories is rather simple, as the user browses a separate database[7], keyword search is one of the most important features of the Formosan Language Archive. The search can be made by typing a word in any

[7] These two databases can be cross-referenced, i.e., if a user intends to look for the distribution of a particular affix, then examples will be drawn from the main text archive.

of the Formosan languages included in the corpora, its Chinese or English translation or glosses. Of interest for us is searching performed by typing a word in a specific Formosan language. Since each corpus includes IPA symbols, the type of search must also handle these.

The two applications we are using, Microsoft Word and Internet Explorer, do not allow the automatic insertion of Unicode IPA. However, it is easier to insert manually IPA symbols in Word than in Internet Explorer. The insertion of IPA symbols will first be discussed here with respect to these two environments. We will then explain how we devised a keyboard mapping mechanism that allows the insertion of IPA symbols on the Web.

In Microsoft Word (e.g., 2000/XP), there are several ways to insert a Unicode IPA symbol. The first is the well-known *Insert…Symbol* menu command. After Insert…Symbol is chosen, a Unicode font is then selected, the pull-down list on the right displays all of the Unicode code points (such as "IPA Extension") included in that font. The second method consists of using the AutoCorrect feature, which is designed to replace mnemonic abbreviations with their Unicode IPA equivalents. This method is handy, but a constraint is placed on codes. They must all begin and end with a non-alphabetic character (see Webster [2002]). A third method consists of inserting IPA symbols using the find/replace function.

It is extremely difficult, if not impossible to insert Unicode IPA symbols when using Web browsers like Internet Explorer. Such symbols, if inserted, usually become empty boxes in the field. To display such symbols, we decided to design a keyboard in which all occurring IPA symbols (so far, 15) along with their numeric equivalents could be displayed (Figure 1). When the user clicks on one of the IPA buttons, the reference code is inserted into the field automatically, and the code is enclosed by "less than" and "greater than" marks (e.g., <660>). The reason for not inserting the typical reference entity (e.g., ʔ) directly is that the ampersand character is significant for Web processing. When the field data is posted onto the server, the Web application can manipulate it due to its computer-readability. In the server, each of the posted IPA symbols is converted back into the standard entity (e.g., $#660;). During this process, we can guarantee that the search string is kept undistorted when sent to the server. It should be noted, however, that a few IPA symbols are able to appear AS IS in the field. Even so, these symbols would be urlencoded[8] into unexpected character strings which would be hard for the program to parse.

When we started digitizing data on the Formosan languages and were confronted with the insertion of IPA symbols on the Web, we found the above method most acceptable. The sole limitation is that users must have installed Unicode IPA symbols beforehand to take advantage of this type of input mechanism.

---

[8] This method is normally used when the browser sends form data to a Web server. It replaces spaces with "+" signs, and unsafe ASCII characters with "%" signs followed by their hexadecimal equivalents.

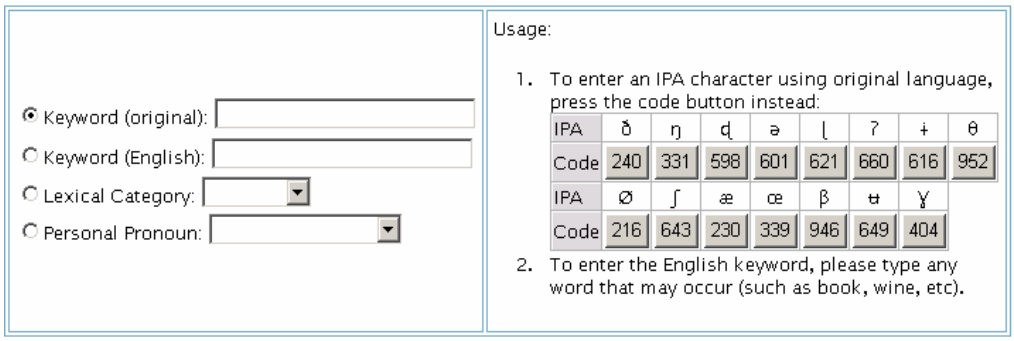


Figure 1. IPA Keyboard Mapping

3. Annotation rules and standards

3.1 Ontology of different Formosan languages

The use of language codes is necessary when constructing the ontology of different Formosan language families included in the corpora. Our coding system is actually based on the latest version of Ethnologue, which was developed by the Summer Institute of Linguistics and is available on the Internet (e.g., DRU for Rukai and BNN for Bunun). As the SIL website does not provide abbreviated names for dialects. We use a two-letter code based on the dialect name itself (e.g., Mn from *Mantauran* Rukai). Thus, the language and the dialect codes form distinct entries in our database.

The codes used for the Formosan languages (along with the dialects they include) that are being archived are shown in Table 4.

Table 4. The code system used in the Formosan language archive

Language	SIL Code	Dialect	Code
Rukai	DRU	Mantauran Maga Tona Budai Tanan Labuan	Mn Mg To Bu Ta La
Saisiyat	SAI	Taai Tungho	Ta Tu
Atayal	TAY	Squliq C'uli'	Sq Cu
Tsou	TSY	Duhtu Tfuya Tapangu	Du Tf Ta

Amis	ALV	Sakizaya Northern Tavalong-Vata'an Central Southern	Sa No Ta Ce So
Bunun	BNN	Takituduh Takibakha Takbanuaz Takivatan Isbukun	Td Th Tb Tn Is
Kanakanavu	QNB	Kanakanavu	Ka
Puyuma	PYU	Nanwang Kapitul	Na Ka
Paiwan	PWN	t-dialect tj-dialect	Td Tj
Pazih	PZH	Pazih Kaxabu	Pzh Kx

3.2 Rules for annotating the corpus in English and Chinese

The Formosan languages are morphosyntactically heterogenous, and though the literature on a number of Formosan languages is now much more abundant than it used to be, many grammatical phenomena have yet to be clarified or need to be further investigated. This poses a challenge for the analysis of each Formosan language corpus that we deal with, as will be explained below.

As pointed out by Zeitoun *et al.* [2003], each text is annotated based on linguistic annotation standards. The transcription of a text in the original language is divided into utterances, sentences and clauses. Words are glossed, and sentences are given free translations. Glosses (or tagset) can be provided at two different levels: the word level (stems) and at the morphemic level (roots and affixes). The major difference 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 and render its word formation opaque. In the texts that have been collected for the Formosan languages (e.g., Tung [1964], Li [1975]), we find 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 as well as morphological alternations must be identified and further analyzed.

Since we started our research in 2001, we have applied a morphemic analysis to annotate all the texts that have been recorded or re-analyzed by ourselves. This method has many advantages in spite of its shortcomings (see below). First, the linguist can annotate the corpus consistently, i.e., words are not "contextually" glossed but their "core" meaning is sought. Second, it helps to determine the distribution and meaning of nearly each affix, thus allowing

construction of an affix database. Third, it deepens one's understanding of the grammar of a specific language, making it easier to identify major lexical and syntactic categories (also included in a database).

The first corpus was annotated in 2001 and focused on only one dialect of Rukai, Mantauran. Over the past four years, as different languages have been annotated, we have been obliged to add more abbreviations to our original list, taking into account morphosyntactic distinctions that exist in these languages. This does not pose a problem, as far as linguistic analysis is concerned, because we know that the Formosan languages exhibit much typological variation. As our abbreviation list was discussed in Zeitoun *et al.* [2003], we will only deal in this section with problems that have arisen due to inclusion of more languages in our corpora.

The addition of new abbreviations has had two different consequences: (i) the use of particular glosses for a single language, and (ii) the insertion of new symbols to distinguish different types of affixes. We will discuss these two consequences in turn below.

Some of these abbreviations are (so far) only used for one language. In Atayal, for instance, there is a distinction between the immediate progressive and remote progressive (cf., *nyux* vs. *cyux*). As progressive auxiliary verbs have grammaticalized from earlier existential verbs that still co-occur productively in this language, the same immediate/remote distinction is also found in these existential verbs. This dichotomy has been reported in Seediq, a language from which collections of texts ready for digitization have not yet been retrieved. Atayal is, thus, the only language in our corpora that makes use of these four abbreviations. Other abbreviations, e.g., AF, PF, Red and LocNmz, are much more common and widely spread cross-linguistically.

One of the most important changes we have had to make has been the insertion of brackets <>, commonly used to delimit infixes and recommended by the Max Planck Institute, Leipzig[9]. Initially, that symbol was not used in our glosses because in the languages that we were annotating (Rukai, Tsou, Atayal and Saisiyat), two infixes barely co-occur simultaneously. In Saisiyat, for instance, though the combination ***ʃ<om><in>βət*** [beat<AF><Perf>beat] 'beat' is grammatically correct, it was not found in our corpus. Originally, if we had a word like *ʃomβət* 'beat' to annotate, we would use hyphens to show its word formation, cf., ***ʃ-om-****βət* [beat-AF-beat], following a common practice among Formosanists. The introduction of two new languages, Bunun and Paiwan, forced us to use brackets instead, as the occurrence of two infixes in these languages is quite productive.

[9] Abbreviations and glosses recommended by the Max Planck Institute, Leipzig (www.eva.mpg.de/lingua/files/morpheme.html) were made available to the public following the creation of our own archive.

Our newest abbreviation list is shown in Table 5. Abbreviations are given both in English and in Chinese, as one of the major goals of the Formosan Language Archive is to build a multilingual corpora in which the original orthography and Chinese-English translations co-exist.

Table 5. Abbreviations used in the Corpora

ABBREVIATION	CHINESE	ENGLISH
ActNmz	動態名物化	Action nominalization
AF	主事焦點	Agent Focus
Asp	時貌 (或 動貌)	Aspect
Caus	使役	Causative
ClsNmz	分句名物化	Clausal nominalization
Cnc	讓步	Concessive
Cntrfct	違反事實	Counterfactual
Dyn	動態	Dynamic
E	排除式 （= 我們）	Exclusive
EP	強調助詞	Emphatic Particle
Excl	驚嘆語	Exclamation
Ext.Imm	存在.近距	Existential Immediate
Ext.Rem	存在.遠距	Existential Remote
Fill	填充語	Filler
Fin	限定	Finite
FP	語尾助詞	Final Particle
Fut	未來	Futute
Gen	屬格	Genitive Case
Hab	習慣	Habitual
HP	勸建助詞	Hortative Particle
I	包含式 (=咱們)	Inclusive
IF	工具焦點	Instrumental Focus
Imp	命令	Imperative
Imprs	無人稱	Impersonal pronoun
InstNmz	工具名物化	Instrument nominalization
Irr	非實現	Irrealis
LF	處所焦點	Locative Focus
LF.Hort	處所焦點.勸建	Locative Focus Hortative
Lig	連繫詞	Ligature
Loc	處所格	Locative Case
LocNmz	處所名物化	Locative nominalization
NAgPass	非主事被動	Non agentive passive
Neg	否定	Negation
NegImp	否定命令	Negative Imperative
NFin	非限定	Non-Finite

NSpec	未指定	Non-specific
Nom	主格	Nominative Case
ObjNmz	受事名物化	Objective Nominalization
Obl	斜格	Oblique
Pass	被動	Passive
P, plur	複數	plural
Perf	完成貌	Perfective
PF	受事焦點	Patient Focus
PF.Hort	受事焦點.勸建	Patient Focus Hortative
Prfct	完成進行	Perfect
Prog.Imm	進行.近距	Progressive Immediate
Prog.Rem	進行.遠距	Progressive Remote
QP	引述助詞	Quotative Particle
Real	實現	Realis
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
Spec	指定	Specific
Subj	虛擬式	Subjunctive
SubjNmz	主語名物化	Subjective 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
<>	中綴	Infix
*	無法確定構詞語法功能	Morphosyntactic function undetermined

Our annotation system is not without shortcomings which we are well aware of. First, though our morphemic analysis allows for the development of different (re)search tools (e.g., keywords, list of affixes and lexical categories), the reading of a word without a whole translation of the sentence is nearly impossible for someone not familiar with Formosan languages. To cope with this problem, lists of lexical categories have been made for each

corpus that allow the user to search for a word, to determine its word formation, to check for related words and to understand its meaning. Second, morphemic analysis can be performed only if a language is well understood by the analyst. Though the project leader trained for many years aboriginal assistants in linguistic analysis, and is supervising the development of each corpus to help make the consistency rate higher through the use of the same terminology, it has become clear that to overcome analytical problems, the participation of more language specialists in the development of each different corpus is crucial. Third, while users can cross-reference rather easily both "original" and "linguistically re-annotated/re-edited" data files, our system can not display the phonetic/phonemic transcriptions of languages, as in the case of Maga Rukai for instance, where morphophonemic alternations render systematic morphemic analysis opaque. This inoperability of our system results from the fact that only a few languages exhibit such dense internal variation so that it is hard to generalize a program for the whole corpora. But this limitation has been solved by adding columns pertaining to morphophonemic alternations in the databases for lexical categories.

Other shortcomings (e.g., inconsistencies in glosses or "wrong" analyses) have been either remedied through the development of new programs or can be resolved through follow-up revisions and corrections of earlier corpora.

3.3 AnnoTool: An Annotation Tool for Formosan Languages

To help with annotation of the corpora, a program called **AnnoTool** (see Figure 2) has been developed. It has two main functions: it facilitates the tagging of texts and the translation of the linguistic terminology from English to Chinese or vice versa.

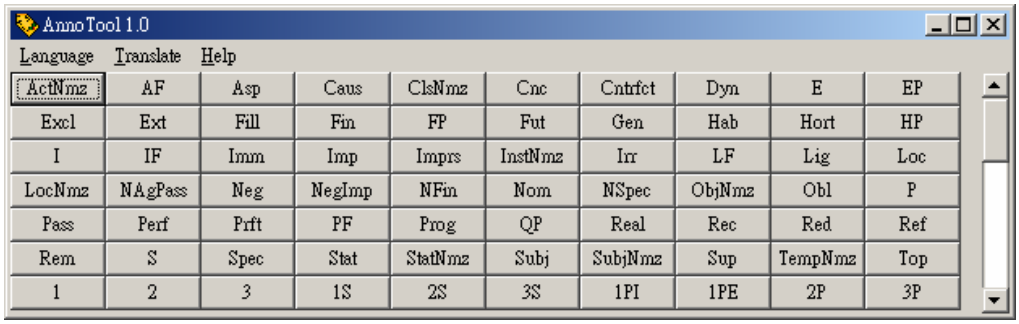


Figure 2. A screenshot of AnnoTool

When launched, the program pops up the list of morphosyntactic abbreviations used to annotate each corpus. To satisfy the requirements outlined in section 3.1, **AnnoTool** allows the expansion of abbreviations used by linguists. Its second major function is translating annotation tags from English into Chinese – or vice versa – in order to reduce the work involved in glossing each text. The above two facets of the program are explained below.

AnnoTool has been designed to work with Microsoft Word. The user can have both programs running concurrently. However, it is necessary to arrange the desktop so that the two windows do not overlap each other. As shown in Figure 3, **AnnoTool** usually occupies one-third of the screen, and Word two-thirds. When the user clicks on one of the buttons in **AnnoTool**, a tag is inserted into the Word document automatically. This method makes linguistic analysis more efficient and more accurate. It is more efficient because the linguist can view the on-screen list and stick to pre-defined terminology. It is more accurate because the likehood of introducing typos is kept to a minimum.

Labels can be translated from English into Chinese, or vice versa. To do so, the user must first select a single term or an entire line from a document and then switch to **AnnoTool** and click on English→Chinese (or Chinese→English) in the Translate menu. Accordingly, the selected sequence in Word can be translated into one of these two languages.

We are conscious that one limitation of **AnnoTool** is that it has been programmed to handle a specific terminological set. It does not deal with the literal translation of lexical words or phrases. Nevertheless, using this tool makes our linguistic analysis easier than it used to be.

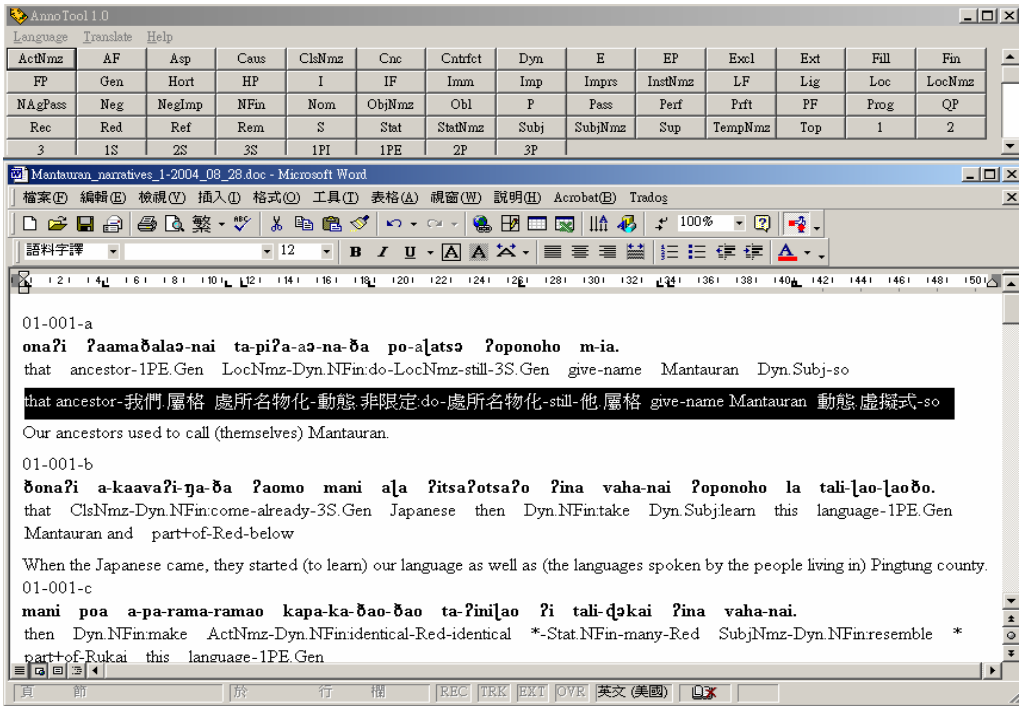


Figure 3. Using AnnoTool with Word

3.4 Affixes and lexical categories

For the tagging of major lexical categories, we follow – though with some reservations – the standardization established by CKIP in charge of the Academia Sinica Chinese Corpora. Not all of the 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 illustrated in Tables 6 and 7. The set of lexical categories has been improved since two more languages (Atayal and Saisiyat) other than Rukai[10] were tagged.

4. Text structure

In this section, we deal with linguistic "recognition" of clause/sentence and paragraph boundaries, and the programs that have been developed to obtain from the Internet a parallel alignment of words, glosses and sentences both in Chinese and in English.

Table 6. A comparison of existing lexical categories in Chinese and in Formosan languages

✓: lexical category found in Rukai or in other Formosan languages

(+) : rare

—: non-existent

Abbreviated basic lexical categories for Chinese	Non-abbreviated basic lexical categories for Chinese	Basic lexical categories in Rukai	Basic lexical categories in 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 Words	✓	✓
14. U	Undecided	✓	✓

[10] We are grateful to the two assistants, Yu-ting Yeh and Tai-hwa Chu, in charge of the Atayal and Saisiyat corpora for their help in improving the databases for lexical categories.

Table 7. Unlisted lexical categories for Chinese that must be included in our description of the Formosan languages

Other basic lexical categories not listed for Chinese	Non-abbreviated basic lexical categories	Basic lexical categories in Rukai	Basic lexical categories in other Formosan languages
1. AUX	Auxiliary	—	(✓)
2. NEG	Negator	✓	✓
3. TOP	Topic	✓	✓
4. TNS	Tense	—	(✓)
5. MOD	Mood	✓	✓
6. CM	Case marker	✓	✓
7. INT	Interrogative word	✓	✓
8. LIG	Ligature	✓	✓
9. EXC	Exclamation word	✓	✓
10. ONOM	Onomatopoeia	✓	✓

4.1 Clause/sentence and paragraph boundaries

As far as linguistic data is concerned, two major factors help in the recognition of clauses/sentences: intonation and syntactic structure. We transcribe every text based on voice files that are recorded and digitized. Though we have not taken into account nor have we tried to provide the duration of each word, intonation plays quite an important role in the detection of clause/sentence boundaries. The analyst's knowledge of the language also helps him/her determine the beginning and the end of a clause vs. that of a sentence. To give but one example, in Tona Rukai, *si* 'and' can appear at the end of a sentence or between two nouns or two clauses. Syntactically speaking, it thus functions as a phrasal or causal coordinator/conjunction. In terms of discursive practices, it is used to mark a pause. That pause can be perceived as "long" (as in (5)), in which case the analyst has to treat the clause as a full sentence, or as short (as in (6)), in which case, two clauses will be treated as being coordinated and forming a longer sentence.

(5) Tona Rukai

la	**ʔabəə**	**m-wa**	**nakay**	**baivi**	***si...***(where ...= pause)
then	Dyn.NFin:return	Dyn.Subj-go	this	village	and

'They returned to the village and…' (Zeitoun [2004, ex. 01-002-e])

(6) Tona Rukai

la	**wa**	**waməcə**	**na**	**bəkəʔə**	**na**	**caŋacaŋaə**
then	Dyn.NFin:go	Dyn.Subj:take	*	pig	*	white and black
la	**paowa**	**po-ʔadʝiŋi**		***si***	**la**	**so**
then	Caus:Dyn.NFin:go	Caus:to-inside		and	then	just
doo	**ki**	**paŋətədʝə**		**ʔaboalə**		***si...***
Dyn.NFin:can	*	person name		Dyn.Subj:come out		and

'Then they brought a black flecked with white pig, put it inside (the hole) and Pangetede could come out.' (Zeitoun [2004, ex. 01-004-b])

4.2 Design of programs to recognize words, sentences and paragraphs

In accordance with annotating conventions, the transcription of a text is divided into sentences, which are further segmented into space-delimited words. There are two types of translations: glosses at the word level and free translations at the sentence level. Sentences are numbered for reference purposes. The encoded format of the reference number is xx-xxx-x, where the first part refers to the text id, the second indicates the paragraph id, and the third corresponds to the sentence id[11].

Each utterance or sentence contributes to the concept of "one block." A block thus includes: (i) the reference information, (ii) the original utterance or sentence, (iii) word glosses and (iv) free sentential translations.

The annotated data has a three-level hierarchy. It includes the "text," the "word" and the "sentence." Transcriptions, glosses and translations are associated with one of these three levels. Metadata is associated at the text level. The structure is hierarchical in that a text contains sentences and words. Based on this hierarchical structure, it is easy for a computer to handle a text as an object (see Jacobson *et al.* [2000]).

A parse program was written to extract sentence and word objects from each corpus. The location of each sentence, their translations and other related information are stored in the sentence-level database (Figure 4). The locations of words, their transcriptions, their word order, Chinese and English word glosses, and punctuation are stored into the word-level database (Figure 5). The *location* field, as a primary key, is used to relate one database to another.

[11] In the corpus files, we simply use the three-part encoding format to represent a sentence location. In the implementation of the database, however, we prefix it with a language id. It then has the final format of xxxxx-xx-xxx-x, in which the first part stands for a language and its dialect.

翻譯 : 資料表

location	c_freetran	e_freetran
DRUMn_01_001_a	我們的祖先萬山自稱是萬山人。	Our ancestors used to call (themselves) Mantauran.
DRUMn_01_001_b	日本人來了以後就（開始）學我們	When the Japanese came, they started (to learn) our language as we
DRUMn_01_001_c	他們比較這兩種語言後，（就發現	They compared (our) languages and (discovered that) there were n
DRUMn_01_001_d	然後他們告訴我們說：「你們是同	Then they told us: "You share the same ancestry."
DRUMn_01_001_e	我們萬山人（才）知道「原來咱們	(That's how) we, Mantauran, learnt that actually we were Rukai.
DRUMn_01_001_f	日本人還沒來之前，我們不知道我	Before the Japanese came, we did not know we were related to the

Figure 4. Sentence-level database

原文與註解 : 資料表

location	wordorde	orthog	punct	pul	pur	cgls	egls
DRUMn_01_001_a	0	onaʔi				那	that
DRUMn_01_001_a	1	ʔaamaðala&				祖先-我們.屬格	ancestor-1PE.Gen
DRUMn_01_001_a	2	ta-piʔa-aə-n				處所名物化-動態-	LocNmz-Dyn.NFin:do-
DRUMn_01_001_a	3	po-aɭatsə				取-名	give-name
DRUMn_01_001_a	4	ʔpponoho				萬山	Mantauran
DRUMn_01_001_a	5	m-ia	.			動態.虛擬式-這樣	Dyn.Subj-so

Figure 5. Word-level database

There is no translation at a higher level than the sentence, so there is no need for a paragraph-level table. The free, sentence-level translations can be strung together and arranged in the original order, and they serve as intelligible, if not always smooth or elegant, translations of the whole text.

4.3 Bilingual translation and alignment

Our project consists of multilingual parallel corpora, which in turn consist of Formosan utterances and bilingual translations. At the sentence level, a source segment and two translations of this source are included. At the word level, each lexical unit and their bilingual glosses are included. From the typographic format, two linking correspondences can be inferred from the text: sentence alignment and word alignment. We developed a morphological program to convert the implicit structure of the text into the XML format, which is now the commonly used standard for corpus encoding (Figure 6), as well as a database format (Figures 4 & 5), which can be utilized by the relational method and is accessed by using the structured query language (SQL). When corpus information has been encoded in such formats, it is easier to handle the alignment problems.

```
<?XML version="1.0" encoding="BIG5" ?>
<TEXT id="01" code="DRUMn" lang="Rukai" dial="Mantauran">
<HEAD>
```

```
    <TITLE>Our language</TITLE>
    ...
  </HEAD>
  <BODY>
  <S id="01-001-a">
    <TRANSCR>
      <W><FORM WO="0")>ona&#660;i</FORM><CGLS>那</CGLS><EGLS>that</EGLS>
           </W>
      <W><FORM WO="1">&#660;aama&#240;ala&#601;-nai</FORM><CGLS>祖先-我們.屬格
          </CGLS><EGLS>ancestor-1PE.Gen</EGLS></w>
      <W><FORM WO="2">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 WO="3">po-a&#621;ac&#601;</FORM><CGLS>取-名</CGLS>
           <EGLS>give-name</EGLS> </W>
      <W><FORM WO="4">&#660;oponoho</FORM><CGLS>萬山</CGLS>
           <EGLS>Mantauran</EGLS></W>
      <W><FORM WO="5">m-ia</FORM><CGLS>動態.虛擬式-這樣</CGLS>
           <EGLS>Dyn.Subj-so</EGLS></W>
      <PUNCT>.</PUNCT>
    </TRANSCR>
    <FREETRAN lang="Chinese">我們的祖先自稱萬山人。</FREETRAN>
    <FREETRAN lang="English">Our ancestors used to call (themselves) Mantauran.
     </FREETRAN>
  </S>
  …
  </BODY>
  </TEXT>
</XML>
```

Figure 6. XML markup of a linguistic text

4.3.1 Sentence alignment

According to our conventional notations, sentences have been aligned since the first corpus (that for Mantauran Rukai) was initially built on a sentence-by-sentence basis. Then the Chinese and English translations were appended. They are clearly distinguishable for distinct line position in the file:

(7) Mantauran Rukai

o-kaoðo ka aɭa-nai toʔonai.

我們沒有任何(東西)

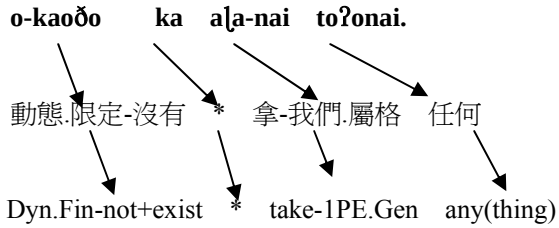
We did not have anything. (Zeitoun and Lin [2003, ex. 013-165-b])

To keep the sentences aligned, our approach maps the linking relationships of the sentence segments and stores or encodes them in a standard format. In other words, the sentential information is stored in the individual fields of a record or in certain XML node elements.

4.3.2 Word alignment

In the Formosan Language Corpora, each uttered word is space-delimited and owns its bilingual glosses appear below it. If no gloss is available, then an asterisk * replaces it:

(8) Mantauran Rukai



Interlinear morpheme-by-morpheme glosses provide most of the information necessary to build a word alignment database. In the database design, each record is based on a transcribed word. This lexical unit includes important pieces of information, such as a unique identifier (here called a *location*), a spelled form, a specific word order, and glosses. Word order plays a major role in word arrangement. It allows words (along with their glosses) to be pieced together and to reappear in the same order as in the original format.

Word alignment provides a basis for the extraction of bilingual lexicons. Using the alignment database, we can get the full index of a particular language. However, as each word is deliberately cut into pieces corresponding to morphemes rather than being given a literal meaning, it is impractical to put them together in the order of the source language since the result would be incomprehensible gibberish. That is why we provide a lexical category search, which allows the user to browse the meaning of each word, and a reference to its word formation (see section 3.1).

Our aligning strategy thus consists quite simply of arranging words with their

corresponding glosses and attaching bilingual translations in a correct order. It will be possible to enhance this approach in the future so that it can be used for other processing tasks.

5. Consistency

The consistency of transcribed words and aligned glosses is one of our major concerns in the construction of each corpus. Even though the various types of corpora collected follow standard notations (IPA transcriptions, interlinear glosses and sentence translations), a certain degree of inconsistency can be found in each corpus.

Inconsistencies can be found at different levels: lexical (i.e., inconsistency in transcriptions; word glossing problems), morphological (incorrect identification of morpheme boundaries, hesitation regarding the distribution of certain affixes or roots), and syntactic (differences in syntactic structures between different dialects or languages that may yield incorrect interpretations of the data at hand). We provide examples of these three types of inconsistencies below and show how we are able to deal with them.

5.1 Transcriptions and word glosses

Certain incorrect transcriptions are easily "repaired," e.g., in Tanan Rukai, Li [1975] recorded 'very' as *aramor* and *ʔaramor*, but later fieldwork showed that the second instance is the correct one. Other discrepancies are more difficult to account for. In Mantauran Rukai, the term *ivoko* 'male friend' contrasts with *la-ʔivoko* 'male friends'. We checked both words many times, and both forms are correct (the first without a glottal, and the second with a glottal).

Word glosses pose another challenge to the linguist, who, for one thing, must be familiar with the culture of the language in question. We are confronted with two interrelated problems: (i) the analyst must decide on the "core meaning" of a word, but at the same time be aware of instances of polysemy or homophony; (ii) there must be a concordance between Chinese and English, but that concordance will sometimes be difficult to reach. To give one example, we were confronted with a series of words in Mantauran Rukai that have to do with social organization, cf., *vaɭovaɭo* 'young (between 15 and 30 years old), maiden woman', *savarə* 'young (between 15 and 30 years old), unmarried man', *titina* 'young or middle-aged (between 15 and 45 years old) married woman with children' (also referring to one's aunt or a woman of the same age as one's mother), *tamatama* 'young or middle-aged (between 15 and 45 years old) married father with children' (also referring to one's uncle or a man of the same age as one's father). When glossing these terms, we had to make decisions about the most linguistically meaningful and culturally relevant aspects of these words and also be able to find the equivalence between English and Chinese. We finally decided to use such glosses as 'young woman', 'young man', 'middle-aged woman' and 'middle-aged man', which have

been/are being adopted for other Rukai dialects and other languages whenever necessary.

5.2 Morphology

Morphology plays a crucial role in understanding the Formosan languages, and the morphemic method we have adopted to annotate each corpus has forced us to deal even more carefully with word formation. The analyst is confronted with two major problems, (i) the incorrect identification of morpheme boundaries, and (ii) the restricted distribution of certain affixes or roots that might render their use and functions opaque.

5.2.1 Morpheme boundaries

Blust [Forthcoming] states that "most AN languages can be characterized as agglutinative-synthetic." Our assumption is based on the fact that morphemes can either be free or bound and can include roots, function words, clitics and affixes and on the fact that morpheme boundaries are usually clear. However, morpheme boundaries might also be difficult to identify, and linguists sometimes propose different approaches to analyzing for the same words. The first problem that has to be settled is whether a word is composed of one or two morphemes. It happens that in some languages/dialects, certain words are no longer divisible, though historically, an affix could be identified. That is the case with the word *ʔoponoho* 'name of a tribe (Mantauran) or the place they inhabit', which derives from the prefixation of *ʔo-* (<*swa-* from) to *ponoho* (<*ponogo* 'name place').

Different analyses from ours are found in the literature, and we must take them into consideration. In Saisiyat (Chu [2003]), for instance, we analyze *ʔi/ʔik* as a ligature, i.e., a grammatical word that carries no lexical meaning. These two morphemes occur in complementary distribution and must be glossed slightly differently, *ʔi* as 'Lig' and *ʔik* as *ʔi-k* 'Lig-Stat'. The first occurs before dynamic verbs and the second before stative verbs. Li [1999], on the other hand, has analyzed both morphemes as sometimes bound and sometimes free, and translated them as 'not'.

5.2.2 Distribution of affixes and roots

Some morphemes are invariable. Because their distribution is very much restricted and their morphophonemic/morphemic alternations are nonexistent, it might be difficult to determine their roots, their origins, their lexical categories. This is the case with Mantauran Rukai *tila!* which translates as 'Leave/Go away' but is actually formed with a first person plural pronoun *t(a)-* adjoined to what was originally the root *ila*. This type of analysis can only be drawn on external evidence, and as mentioned above, necessitates a good understanding of the language being investigated.

Likewise, some affixes are very non-productive, and it might be difficult to determine their meaning. This is the case with Mantauran Rukai *ta ʔa ða ʔanə* 'house warning' (< *ða ʔanə* 'house'); the meaning of *ta ʔa* is still poorly understood.

5.2.3 Syntactic structures

The major problem that the linguist must be aware of regarding syntactic structures has to do with typological diversity. For instance, in Mantauran and Labuan Rukai, though subordinate temporal clauses are superficially identical, in the former, the subject is marked by the genitive, and in the latter, it is marked by the nominative.

(9) Mantauran Rukai

onaʔi	**ʔiða**	**a-paka-kanə-ŋa-*li***
that	yesterday	ClsNmz-Dyn.NFin:finish-eat-already-1S.Gen
(ʔa)	**o-ðavacə-ŋa-ɭao.**	
Top	Dyn.Fin-leave-already1S.Nom	

'Yesterday, after I had eaten, I left.'

(10) Labuan Rukai

sa	**maka-kanə-ŋ-*ako***	**ko**	**aga**	**ka**
when	Dyn.Fin:finish-eat-already-1S.Nom	Acc	rice	Top
w-a-davac-ako.				
Dyn.Fin-Real-leave-1S.Nom				

'Yesterday, after I had eaten, I left.'

5.3 Programs developed to remedy analytic inconsistencies

From the processing perspective, a hyphen is used as a morpheme boundary and as such provides morphemic information that can be used to parse word tokens (e.g., *om-ia-nai* 'Dyn.Fin-so-1PE.Nom') without difficulty. To remedy inconsistencies in transcriptions and glosses, all the words can be extracted from the corpus data to create an index. This index list (or finderlist) enables the analyst to compare all the words in order to minimize incorrect spelling or glosses. This program can also output a frequency list of morphemes (Hockey [1998]).

Initially, the design of each database had to take punctuation into account. We treat a

space between two words as a punctuation mark, so every word can be said to have an associated punctuation mark. Although this mark indicates a boundary between a group of words, in practice it is connected to the preceding word. Following this approach, we can treat punctuation as a field of the preceding word, as shown in Figure 5 (Leech *et al.* [1995]).

5.3.1 Word-by-word alignment consistency checker

At a very early stage in the development of the Formosan Language Archive, a program called **Chkgloss** was designed to verify the rigid structure of the corpus by comparing the number of orthographic words with that of their glosses (see Figure 7). In each corpus, transcribed words are aligned vertically with their interlinear glosses. For example:

(11) **o-kaoðo ka aɭa-nai toʔonai.**
動態.限定-沒有 * 動態.非限定:拿-我們.屬格 任何
Dyn.Fin-not+exist * Dyn.NFin:take-1PE.Gen any(thing)
我們沒有任何（東西）。
We did not have anything.

As mentioned above, to guarantee that transcribed words are the same in number as their glosses, an asterisk is used to represent an empty word (whose meaning or morphosyntactic function remains opaque). It is only after the verification process is completed that a text can undergo whole transformation and be displayed on the Internet.

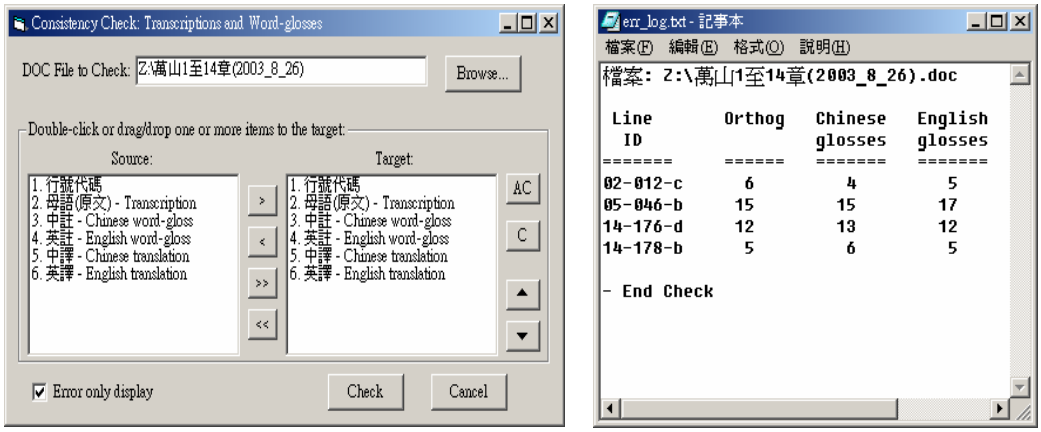


Figure 7. A Screenshot of Chkgloss

Chkgloss is helpful for identifying errors because it provides the consistency rate between (i) each tagged word and its gloss and (ii) each sentence and its bilingual translation. In most

cases, a corpus has to undergo back-and-forth processing several times before it can be deemed to be valid (Figure 8).

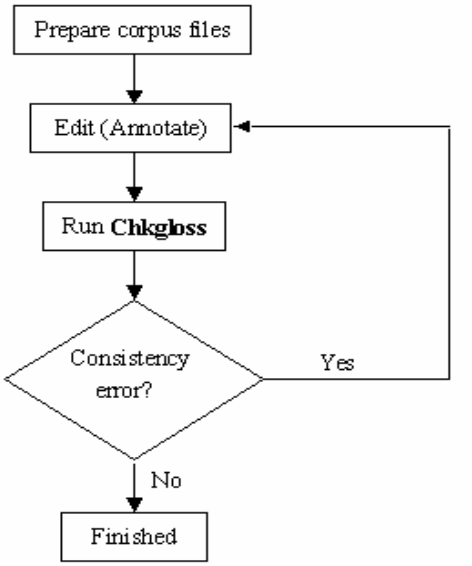


Figure 8. The workflow of using Chkgloss

6. Conclusion

The Formosan Language Archive is a useful tool for conducting research on the Formosan languages. The multilingual comparable corpora have begun to find their way in linguistic applications and natural language processing. As far as linguistic applications are concerned, each corpus features well-analyzed data that can serve as a basis for more in-depth studies. There are a number of advantages in providing word alignment, sentence alignment, linguistic annotations and bilingual translations. Computer-aided linguistic research is being carried out using tools and techniques that improve the work of the analyst. Applications that were developed for the Formosan Language Archive include Unicode IPA symbols, **AnnoTool**, **Chkgloss** and **Indexer**.

Drawbacks are inevitable, however. If suitable electronic text versions had been available, progress would have been more rapid. Admittedly, a lot of time has been spent on reformatting the legacy data to make it computer readable. In addition, electronic versions of earlier published materials have to be made from scratch, since there were previously no electronic files (e.g., Li [1975], Tung *et al.* [1964], Fey *et al.* [1993] etc.).

It is the purpose of our project to collect, analyze and digitize data on many, if not all, of the Formosan languages for which texts are available, but more corpora need to be included to refine the original architecture of the archive. On the other hand, we also need to think about

how to develop new tools, make use of existing tools described in the literature (cf., Szakos *et al.* [2004]) and process the voice files for further research (e.g., phonetic and discursive studies)[12]. We might also be able at some point to conduct an experiment on natural language processing (e.g., corpus-based machine translation).

References

Adelaar, K.A., "Retrieving Siraya phonology: a new spelling for a dead language," In *Selected Papers from the Eighth International Conference of Austronesian Linguistics*, ed. by E. Zeitoun and P. J.-K. Li, Symposium Series of the Institute of Linguistics (Preparatory Office), No. 1, Institute of Linguistics, (Preparatory Office), Academia Sinica, Taipei, 1999, pp. 313-354.

Bird, S. and G. Simons, "Seven dimensions of portability for language documentation and description," *Language*, 79(3), 2003, pp.557-582.

Blust, R., *Thao dictionary*, Language and Linguistics Monograph Series, No. A5 Institute of Linguistics, (Preparatory Office), Academia Sinica, Taipei, 2003.

Blust, R., *The Austronesian Languages*, Ms., Forthcoming.

Bow, C., B. Hughes and S. Bird, "Towards a general model of interlinear text," In *Proceeding of EMELD 2004: the Workshop on Linguistic databases and best practice*, Detroit, Michigan, July 15-18, 2004.

Chu, T.-H., *Saisiyat texts*, ms., 2003.

Early, R. and J. Whitehorn, *One hundred Paiwan texts.* Pacific Linguistics, 542, Research School of Pacific and Asian Studies, The Australian National University, Canberra, 2003.

Fey, V., *et al*., *O'Orip no' - Amis Ameizu wenhua - Amis Culture*. Taiwan Bible Society, Taipei, 1993.

Hockey, S., "Textual Database," *Using Computers in Linguistics: A Practical Guide*, ed. by J. Lawler and H. Aristar-Dry, Routledge, London, 1998, pp.101-133.

Hua J.-J., *Southern Paiwan texts*, ms., 2005.

Hsin, T.-H., *Maga (Rukai) texts*, ms., 2002.

Jacobson, M., B. Michailovsky and B. Lowe, "Linguistic documents synchronizing sound and text," *Speech Communication*, 33, 2000, pp.79-96.

Leech, G., G. Myers and J. Thomas, *Spoken English on Computer,* Longman Publishing, New York, 1995, pp. 208-219.

Li, P. J.-K., *Rukai Texts*, Institute of History and Philology, Special Publication No. 64-2, Academia Sinica, Taipei, 1975.

[12] In the case of phonetic studies, the manual re-segmentation of the files would be necessary.

Li, P. J.-K., *Orthographic systems for Formosan languages*, Ministry of Education, Taipei, 1992. [In Chinese]

Li, P. J.-K., *The history of Formosan aborigines: Linguistic, Nantou,* The Historical Research Commission of Taiwan Province, Taiwan, 1999. [In Chinese]

Li, P. J.-K. and S. Tsuchida, *Pazih texts and songs*, Language and Linguistics Monograph Series, No. A2-2, Institute of Linguistics, Preparatory Office, Academia Sinica, Taipei, 2002.

Rau, V. D., "The scientific and social principles of the orthographic symbols of the aboriginal languages of Taiwan: a case study of Atayal," *The Languages of the Austronesian tribes of Taiwan,* ed. by P. J.-K. Li and Y.-C. Lin, Ministry of Education, ROC, Taipei, 1995, pp.31-47. [In Chinese]

Szakos, J. and U. Glavitsch, "Portability, modularity and seamless speech-corpus indexing and retrieval: a new software for documenting (not only) the endangered Formosan aboriginal languages. Paper read at the Workshop on Linguistic databases and best practice," In *Proceeding of EMELD 2004: the Workshop on Linguistic databases and best practice,* Detroit, Michigan, July 15-18, 2004.

Tseng, S.-Q. *et al*., *zouguo shikong de yueliang*, Chen-hsing Publ. Co., Taipei, 1998. [In Chinese]

Tsuchida, S., *Kanakanavu Texts (Austronesian Formosan)*, Endangered Language of the Pacific Rim. ELPR Publication Series A3-104, Nakanishi Printing Co., Kyoto, 2003.

Tung, M.-N. and V. Rau, *Yami text,*. ms., 2002.

Tung, T.-H. *et al*., *A descriptive study of the Tsou Language, Formosa,* Institute of History and Philology, Special Publication 48, Academia Sinica, Taipei, 1964.

Venezky, R. L., "Principles for the design of practical writing systems," *Anthropological Linguistics*, 12(7), 1970, pp.256-270.

Webster, G., "Using Unicode IPA on the web and in word processing," University of Washington Language Learning Center, (paper available at : http//depts.washington.edu/llc/help/presentations/index.php)2002.

Ye, Y.-T., *Atayal texts*, ms., 2003.

Yu, C.-H., "Discussion on the digitization of the Formosan Language Archive – building up of the architecture of the archive." In *Proceeding of the First workshop on the Digital Library Projects*, July 25-26, 2002, Taipei.

Zeitoun, E., *Tona texts*, ms., 2004.

Zeitoun, E., C.-H. Yu and C.-X. Weng, "The Formosan Language Archive: development of a multimedia tool to salvage the languages and oral traditions of the indigenous tribes of Taiwan," *Oceanic Linguistics*, 42(1), 2003, pp.218-232.

Zeitoun, E. and C.-H. Yu, "The Formosan Language Archive: Language Processing and Linguistic Analysis," In *Proceeding of 1st International Joint Conference on Language Language Processing (IJCNLP-04) Fourth Workshop on Asian Language Resources*, March 25, 2004, Sanya, Hainan Island, China.

Zeitoun, E. and H.-C. Lin, *We should not forget the stories of the Mantauran : Memories of the past,* Language and Linguistics Monograph Series, No. A4, Institute of Linguistics, Preparatory Office, Taipei, 2003.

Zeitoun, E. and H.-C. Lin, *We should not forget the stories of the Mantauran: Traditional folktales,* ms., 2004.

Zeitoun, E., *The Formosan Language Archive: Linguistic analysis and language processing*, Invited talk at Tuesday Seminar of the Institute of Linguistics, University of Hawai'i at Mānoa, January 25, 2005.

Related web pages:

http://www.ailla.org/pc/mainindex.html

http://lacito.archivage.vjf.cnrs.fr/

http://www.emeld.org/workshop/2004/paper.html

http://www.language-archives.org

http://www.rosettaproject.org:8080/live/

http://www.sinica.edu.tw/SinicaCorpus

http://wwwtalana.linguist.jussieu.fr

http://sino-tibetan.cityu.edu.hk/rda/ -- no longer available, later moved to:

http://victoria.linguistlist.org/~lapolla/RDA/index.html

http://paradisec.org.au