Taiwan Journal of Linguistics
Vol. 6.2, 79-118, 2008
doi:10.6519/TJL.2008.6(2).4

DEVELOPING AN ONLINE CORPUS OF FORMOSAN LANGUAGES*

Li-May Sung, Lily I-wen Su, Fuhui Hsieh and Zhemin Lin

ABSTRACT

Information technologies have now matured to the point of enabling researchers to create a repository of language resources, especially for those languages facing the crisis of endangerment. The development of an online platform of corpora, made possible by recent advances in data storage, character-encoding and web technology, has profound consequences for the accessibility, quantity, quality and interoperability of linguistic field data. This is of particular significance for Formosan languages in Taiwan, many of which are on the verge of extinction. As a response to the recognition of this burgeoning problem, the key objectives of the establishment of the *NTU Corpus of Formosan Languages* aim to document and thus preserve valuable linguistic data, as well as relevant ethnological and cultural information.[1] This paper will introduce some of the theoretical bases behind this initiative, as well as the procedures, transcription conventions,

*Parts of this paper were presented at the Workshop on Corpora in Taiwan, February 2, 2007, National United University, and at the Joint NTU-Rice Linguistic Workshop 2008, March 3-4, 2008, National Taiwan University. We would like to thank the audiences there and two anonymous reviewers for their valuable comments and suggestions.

[1] The NTU Corpus of Formosan Languages started as one of the sub-projects of the Multimedia Laboratory established by the Center for Information and Electronics Technologies, which integrates the professional and academic work of various departments and colleges at the National Taiwan University, with a view to establishing a standard for the creation of linguistic corpora databases through the application of information technology on linguistics research. The NTU Corpus of Formosan Languages is currently receiving funding to expand its data from the Center for Humanities Research, National Science Council and the Center for Austronesian Studies (96R0502-07), National Taiwan University. Our website is as follows: http://corpus.linguistics.ntu.edu.tw. A brief introduction of our corpus is also available in Su, Sung, Huang, Hsieh and Lin (2008).

database normalization, in-house system and three special features in the creation of this corpus.

Key words: Formosan languages, Taiwan, corpus, database normalization, discourse, intonation unit (IU), *Pear* story, *Frog* story, cross-referencing retrievability, multilingual search, interoperability

1. INTRODUCTION

The recent advances in the field of natural language processing techniques and information technologies, which enable researchers to submit, browse and search linguistic materials, have generated considerable interest in the construction of an integrated platform. To place narratives and field notes in a corpus is, at present, the most efficient and effective way to document, and thereby preserve precious linguistic and cultural data. This is of particular significance, now that the number of languages in the world is diminishing at an unexpected and startling speed; as many as half of the estimated 6,000 languages spoken on earth are, as pointed out in Krauss (1992), 'moribund'. Ninety percent of the existing languages today are likely to become extinct or, at the very least, seriously endangered in the near future, and the phenomenon of dying languages is being seen to occur at a particularly ferocious rate in America, Africa, Australia and Southeast Asia (Brenzinger 1992; Crawford 1995; Robins and Uhlenbeck 1991; Schmidt 1990).

Formosan languages in Taiwan nowadays are also confronted with the same growing crisis of language endangerment. With a total population of only 2.31% of the population of Taiwan as a whole,[2] economic, social and other pressures have caused the Formosan communities to cease to speak their native languages and turn to more dominant languages, such as Mandarin Chinese, Taiwanese or Hakka. As a response to the recognition of the growing problem of endangerment, the key objectives of the establishment of the NTU

[2] This estimation of populations is based on data taken from the website of the National Institute of Educational Resources and Research, Taiwan, retrieved June 13, 2008. Seven out of the fourteen tribes consist of less than 10,000 people, including Tsou, Saisiyat, Yami, Thao, Kavalan, Sakizaya, and Seediq. Among them, the number of Thao is less than one thousand. (http://3d.nioerar.edu.tw/2d/native/course/course_0101.asp)

Corpus of Formosan Languages are dedicated to the documentation and preservation of valuable linguistic data, as well as to provide the best possible record of these endangered languages, for the benefit of related linguistic research and for the pursuit of knowledge within the field of social science or in language communities.[3]

The NTU Corpus of Formosan Languages (henceforth the NTU Corpus) introduced in this paper is a long-term attempt to systematically create image, sound and video recordings and integrate them with text materials. After the recordings are transcribed, tagged and analyzed, an online database system is then a necessary platform to store and to represent data, for further use of the collected data. The following guidelines are thus proposed: our system shall (a) be easy to customize for most Formosan languages, (b) specify standardized procedures of segmentation, transcription, translation and annotation, (c) develop automatic extraction of morpho-syntactic information to reduce the need for human labour, (d) provide a web-based presentation of data in a multi-media format, (e) create an accessible, searchable, and multi-lingual platform, and (f) transform existing data to other formats that are interoperable with other systems.

At present, a total of 32 texts, running up to 2 hours and 30 minutes, are available online from the NTU corpus including five Formosan languages: Kavalan, Saisiyat, Tsou, Amis and Sakizaya.[4] The data contain various spoken genres such as conversations, stories, songs, and folktales.[5] The organization of this paper is structured as follows.

[3] To our best knowledge, at present there are two other online corpora of Formosan languages. They are the *Academia Sinica Formosan Language Archive* and the *Digital Archiving Yami Language Documentation*. The *Academia Sinica Formosan Language Archive* is available at http://formosan.sinica.edu.tw/. See also Zeitoun et al. (2003) and Zeitoun and Yu (2005) for an introduction. The *Digital Archiving Yami Language Documentation* is available at http://yamiproject.cs.pu.edu.tw/yami/en_index_flash.htm.

[4] There are further 98 texts, which have either not been transcribed yet or have only been roughly transcribed, together running approximately up to 8 hours and 24 minutes. Signed consent forms have also been obtained from our speakers, allowing us to use the elicited texts, photos, sound and video recordings etc. for this project.

[5] Some of the texts are collected in the course *Linguistic Fieldwork* that the first author, Li-May Sung, has been teaching in the past few years and some are partial research results of several integrated NSC projects including those of Huang, Su and Sung (2001, 2002, 2003, 2004), Huang and Sung (2005), and Huang, Szakos and Sung (1999). These texts were either recorded during class sessions or in the field at various villages during fieldtrips.

Section 2 provides some of the theoretical bases underlying the construction of a natural spoken database. Section 3 outlines the procedures for text transcription and database normalizations. Section 4 is a description of our in-house system, including back-end programs, the POS-tagger and a unified output interface. Section 5 introduces three specially-designed features. Some concluding remarks are given in Section 6.

2. THEORETICAL ASSUMPTIONS

In this section, we will elaborate upon the theoretical bases behind the design of the NTU Corpus, arguing for the importance and necessity of studying spoken data. All the texts in this corpus are comprised of first-hand, naturalistic data that were collected during field research using a digital recorder and a digital camcorder.[6] It is now commonly acknowledged that the investigation of spoken speech beyond sentential level has provided new insights into probing the nature of language. Spoken data in actual discourse context, in contrast to written texts, is of particular significance in revealing communicative, psychological and cognitive aspects of language in use. By examining natural spoken data, we can gain a clearer picture of certain linguistic phenomena, such as issues related to the identification of lexical category (cf. Hopper and Thompson 1984), ergativity (cf. Du Bois 1987; DeLancey 1981) and transitivity (cf. Hopper and Thompson 1980), etc. In addition to syntactic knowledge, 'socially shared cognition', as referred to in Schegloff (1991), can also be better depicted by naturalistic data.

To faithfully present the above-mentioned discourse information, all utterances in the corpus are segmented and coded on the basis of the so-called Intonation Unit (IU) (cf. Chafe 1987), which may allow researchers to obtain rather different results from studies based on constructed sentences. The intonation unit (IU), a basic prosodic chunk, displays a detailed representation of interactional linguistic features such

[6] A hand-held mobile audio recorder (Korg MR-1) is used at present, which features a 20G byte internal hard drive, offering up to six hours of recording at the highest audio quality (1-bit 2.8224 MHz stereo). The Korg MR-1 also provides high-speed USB 2.0 connectivity for fast and easy transfer of sound files to the computer. The digital camcorder used is the Hitachi DZ-MV580A which can take photos and record video clips to DVD-RW discs.

as pauses, repetitions, repairs, false starts, fillers, hesitants and intonation etc, none of which are epiphenomenal. On the contrary, they reveal important linguistic aspects of the language itself, as well as the cognitive and psychological aspects of language use in production. For example, in-depth studies on conversations have shown that fillers, repairs and repetitions are important interactional strategies used by speech participants to hold the floor, plan speech in advance, perform word-searches, and so on. In addition, by investigating Q-A pairs and sequential structures, such as *pre*'s, greetings, and openings and closings, we can learn more about the pragmatic and cultural aspects of a language (Schegloff 1980, 1988; Sack, Schegloff and Jefferson 1974).

Aside from recorded spoken data, the NTU Corpus is also tailored to contain relevant ethnological, sociological and cultural field notes which have been gathered in the course of doing field work. These pieces of information are precious as well, for they aid in revealing not only the structure of a language but also show how each grammatical construction relates to the kind of social action that can be found in the language community. For example, data related to how spatial orientation is expressed in various societies (Bernd Heine 1997; Tuan 1974) help us to understand the way perception and conceptualization in different cultures influences languages.

All in all, the theoretical rationale underlying the NTU Corpus aims at providing the best possible naturalistic record for studying the interaction between grammar, discourse and cognition in Formosan languages. The well-known quote from Du Bois (1985) best explains our discourse-oriented approach here: "Grammar codes best what speakers do the most." We believe that grammar is an inventory of grammaticalized recurrent patterns which emerges from the daily use of language (Ono and Thompson 1995; Lerner 1996; Schegloff 1996; Couper-Kuhlen and Thompson 2000).

3. PRODUCING TEXT TRANSCRIPTIONS

Any establishment of corpora involves much more than the mere storage of digital texts together; rather, it concerns the nature of the transcription conventions adopted, how the texts are managed and the standard operations that are required before/upon submission. The procedures to handle data in our corpus are designed following the core

idea of low coupling, in order to reduce complexity, avoid inconsistency, and facilitate efficiency by the use of in-house programs. The following sections outline the standard procedures for creating a text transcription in our corpus.

3.1 IU Segmentation

The intonation unit (IU), mentioned above, serves as the basic transcribing unit during the process of the collection of materials. In the practice of segmenting work, many of our student transcribers[7], however, face challenges in identifying prosodic cues or when attempting to determine the boundaries of intonation units. Therefore, transcribers are trained to listen to sound files, using both the *Praat*[8] software and MS word processor, in order to segment the IUs, and thereby produce text transcriptions. The problems that transcribers often encounter are minimized with the visual aid of the wave form and spectrogram in *Praat*, and the segmentation is thus more accurate. In addition, five salient prosodic features are specified in the work manual for handling Formosan languages, to ensure consistency among transcribers and facilitate the subsequent processing of data. They include (a) pauses, (b) change of duration (lengthening), (c) change of intensity (loudness), (d) change of pitch (falling or rising pitch contours), and (e) breath reset. These diagnostics for intonation unit boundaries are primarily based on the discussions of Du Bois et al. (1993), Chafe (1987, 1994), Tao (1996) and Himmelmann (2006). Example (1) and Figure 1 illustrate a long pause[9] between two intonation units and the lengthening of the IU-final particle *u*. Example (2) and Figure 2 indicate that the IU boundary has a falling terminal contour and that there is a change of intensity occurring

[7] The student transcribers are part-time assistants recruited from our institute for this corpus project. They all have first-hand experience and training in working with one Formosan language from the course *Linguistic Fieldwork.*

[8] Praat is a programmable phonetic analyser written by Paul Boersma and David Weenink of the Institute of Phonetic Sciences, University of Amsterdam, which is commonly used to aid in transcriptions of digitized sound files. It is available for free under the GNU General Public License. See the website http://www.fon.hum.uva.nl/praat/ for details.

[9] The length of pauses varies. Pauses of less 0.3 second, between 0.3 second and 0.6 second, and above 0.7 second are transcribed respectively as two dots (..), three dots (…), and three dots followed in parenthesis by a measurement of the pause length (…(N)).

on the syllable *za* inside the word *mahiza*.[10] Finally, pitch resetting is usually manifested in the form of rising at the beginning of an IU in contrast with the end of the preceding IU. This is shown in example (3) and Figure 3, in which the pitch level at the beginning of IU33 *tu* is set higher than at the end of IU32 *paluma-han=tu*.

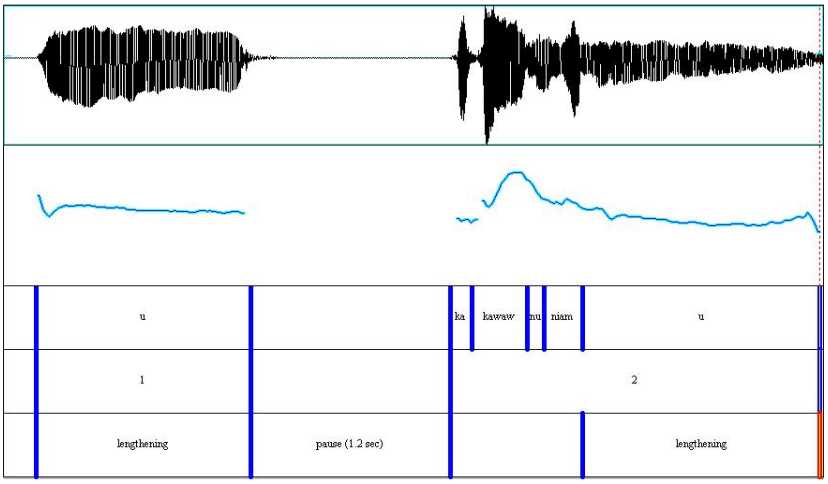
(1) An Excerpt from Sakizaya (Skzy_ta'on_story: IU 45-47)

45.	...(1.9) u==,_				
	CN				
	CN				
46.	...(1.2) ka-kawaw	nu		niam	u== ,_
	RED-thing	GEN		1EPL.POSS	CN
	重疊-事情	屬格		1PL.排除.所有	CN
47.	.. babalaki	haw	sa	.\	
	elder	HAW	DM		
	老人	HAW	DM		

We old people do everything well.
我們老人做事做的好。

[10] A caret '^' is used in our corpus to indicate the change of intensity (or loudness), which is perceived as a focused (or "emphasized") prominent syllable which sometimes does not coincide with the stressed pattern of a word.

Figure 1. Long pause between IUs and lengthening

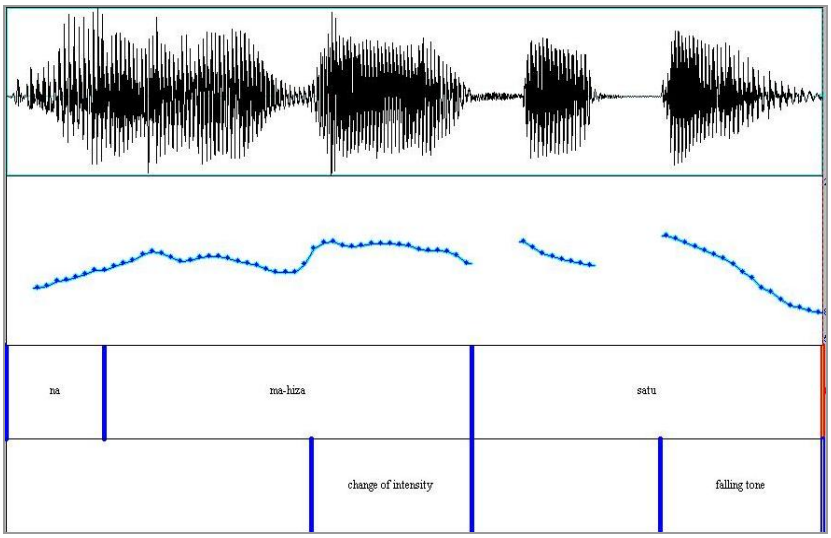


(2) An Excerpt from Sakizaya (Skzy_amui_putong: IU 48)

48. ...(1.9) na=mahi^za satu ,\
past=that.way DM
過去=那樣 DM

That's how it was.
過去是那樣

Figure 2. Change of intensity coupled with a falling contour



(3) An Excerpt from Sakizaya (Skzy_amui_putong: IU 32-34)

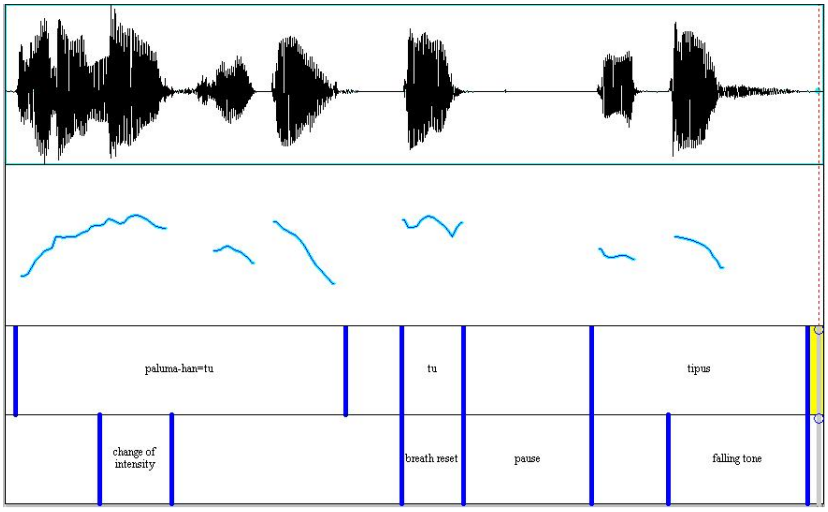
32. ...(1.4) paluma-han=tu ,\
plant-PF=PFV
種-受焦=完成

33. .. tu ,\
OBL
斜格

34. ... tipus ,\
rice
稻米

Rice was grown.
稻米種了

Figure 3. Resetting of the baseline pitch level



3.2 Text Standardization

With the aid of *Praat*, the collected data are segmented and presented in the form of a sequence of numbered IUs. When transcribing, we follow the orthographic system proclaimed together by the Council of Indigenous Peoples and the Ministry of Education in 2005, which has been generally accepted by the Formosan language community.[11] The topics contained in the data range from daily conversations, songs, narratives of the *Pear* story and the *Frog* story, and folktales or legends such as weaving, flooding or molting, etc.

Both the *Pear*, a six-minute colour-mute film made by Wallace Chafe in the mid-1970's, and the *Frog*, a wordless pictorial book sketched by Mercer Mayer (Mayer 1980), have been used as a basis to elicit texts for comparative and cross-linguistic studies in various languages by a wide range of scholars (cf. Chafe 1980; Berman and

[11] Details of the proclamation and lists of orthographies adopted by each Formosan language are obtainable from the website of the Council of Indigenous Peoples, Executive Yuan:
http://www.apc.gov.tw/chinese/docDetail/detail_TCA.jsp?docid=PA000000000154&linkRoot=1&linkParent=0&url=

Slobin 1994; Strömqvist and Verhoeven 2003).[12] The collection of the *Pear* and *Frog* narratives is part of a typology research project to study contextual variations of narrative constructions within and across Formosan languages.

For the genre of folktales or legends, our native speakers were asked individually to narrate any traditional or ecotypified historicized stories that he or she was familiar with. The data from these stories may reflect socio-cultural values from within the speaker's language community and are of significant value in aiding the preservation of the Formosan language heritage. An excerpt in (4) illustrates a legend of how Saisiyat people in the ancient times underwent the process of molting as they grew old:

(4) An Excerpt from Saisiyat (Molaw: IU27-31)

27. ... 'am==
FUT
未來

28. ...(1.0) 'am 'al'alak naehaen .\
FUT young again
未來 年輕 又

29. ... hoe==pay
AF.be.tired
主焦.累

30. ... o: biSbiS 'atomalan
INT AF.painful very
感嘆 主焦.痛苦 非常

31. ...(0.8) hinghae' ki Sibai' hara haysani Sibai'
same and snake like now snake
相同 與 蛇 像 現在 蛇

[12] Our speakers are asked to tell stories either after having watched the *Pear* film or while looking at the pictorial book *Frog*. *Pear* is about a farmer and a boy who stole some of the farmer's pears. The boy then has a series of adventures with other children. The wordless pictorial book *Frog* tells the adventure story of a little boy and his dog searching for his frog that went off into the woods.

> They would grow old and their skin would molt; (it was like that) for generations. It was very tiring and painful, just like a snake molting.
> 這樣一代又一代, 他們老了蛻皮又變年輕了.非常的累又非常的痛苦.就跟蛇一樣,像現在的蛇一樣.

When transcribing the data, a set of metadata is specified at the head of every text. Metadata has been considered important for effective archiving and for describing the characteristics of recorded materials.[13] The metadata we catalogue in our corpus contains 12 fields including file name, topic, type, language, dialect, speaker, duration, number of IUs, time of recording, time of revising, names of transcribers and final inspectors. An example of the head is given in (5), and the description of the field schema is shown in Table 1.

(5)

File name: pear_imuy
Topic: Pear story
Type: Narrative
Language: Kavalan
Dialect: Xin-she (新社)
Speaker: Imuy, 潘金妹, Female, 51 yrs
Duration: 00:01:15
Total IUs: 31
Collected: 2003-05-30
Revised: 2006-03-01
Transcribed by: 葉俞廷(Yu-ting Yeh), 王以勤(I-chin Wang)
Verified by: 鍾曉芳(Siaw-Fong Chung), 沈嘉琪(Chia-chi Shen), 葉俞廷(Yu-ting Yeh),謝富惠(Fu-Hui Hsieh)

[13] Several websites provided useful discussions and tools for creating the metadata structure during the establishment of the corpus. Some of them are: EMELD (http://emeld.org/index.cfm), OLAC (http://www.language-archives.org/), the Linguist List (http://www.linguistlist.org/tools/index.html), SIL International (http://www.sil.org/linguistics/computing.html) and IMDI (http://www.mpi.nl/IMDI/). Following the suggestion of Dr. Hsueh-hua Chen of the Dept of Library and Information Science, NTU, one of the collaborators in this project, the metadata sets specified by OLAC and by the Academia Sinica Formosan Language Archive were used as references during the creation of the sets for our corpus. Also see Chen (2007) for a discussion of the metadata infrastructure we have adopted.

Table 1. Metadata schema

Field name	Description	Format
File name	File name of the text	String (e.g., pear_imuy)
Topic	Topic of the text	String (e.g., Pear Story)
Type	Style of the text	String (e.g., Narrative)
Language	Language of the text	String, first letter capitalized
Dialect	Dialect or district	String
Speaker	Information about speaker's native name, Chinese name, gender and age	String (Chinese string), string, numeric
Duration	Length of recording	hh:mm:ss
Total IUs	Number of IUs in text	Numeric
Collected	Date of recording	yyyy-mm-dd
Revised	Date of latest revision	yyyy-mm-dd
Transcribed by	Transcriber(s) and annotator(s)	String (different strings are separated by commas)
Verified by	Final inspector(s)	String (different strings are separated by commas)

The text transcription following the metadata is described in Table 2 below.[14]

[14] According to our rough estimation, one-minute of raw data requires ten to twelve hours of working time.

Table 2. Part of a text excerpted from Kavalan (frog_buya: IU 1-4)

IU#	text, glossing, translation & notes	Description
1.	..ma==sang nani,\	[IU #, with a period at the end; native words separated by spaces]
	before DM	[English gloss separated by spaces]
	以前 DM	[Chinese gloss separated by spaces]
2.	yau usiq sunis 'nay,_	[IU #, with a period at the end; native words separated by spaces]
	EXIST one child that	[English gloss separated by spaces]
	存在 一 小孩 那	[Chinese gloss separated by spaces]
3.	...(1.7) atu wasu-na,/	[IU #, with a period at the end; native words separated by spaces]
	CONJ dog-3SG.GEN	[English gloss separated by spaces]
	連接詞 狗-3SG.屬格	[Chinese gloss separated by spaces]
4.	..kin-awsa nani.\	[IU #, with a period at the end; native words separated by spaces]
	CLF.HUM-two DM	[English gloss separated by spaces]
	人量詞-兩 DM	[Chinese gloss separated by spaces]
	#e Long time ago, there were a child and a dog.	[English translation of IU 1 to IU 4]
	#c 從前，有一個小孩和一隻狗。	[Chinese translation of IU 1 to IU 4]
	#n	[Field notes; etymological, cultural or ethnological notes]

As shown in Table 2, the text is displayed in the form of intonation units, preserving details of discourse linguistic features such as pauses, repetitions and fillers etc. IUs are numbered according to their sequence as they occur in the context. The text, glosses, translation and notes are transcribed into the native language, English and Chinese, so as to fit the needs of both local and international communities. The alignment of native words and English/Chinese glosses is handled automatically by the system. Morpheme boundaries, morphological information and word meanings are also extracted automatically by the back-end programs, to be discussed in Section 4. Lines beginning with a sharp (#) are processor instructions (PI) for the programs. The symbol "#e" indicates a line of English translation of a meaningful chunk, usually a sentence, composed of several IUs. The symbol "#c" marks a Chinese translation and "#n"

marks an elicitation, etymological or ethnological note. It is possible for there to be more than one note. The following excerpt from the Pear story in Kavalan is an example. It contains a field note that provides a further explanation of the word *baqi* which was obtained from our language consultant.

(6) An Excerpt from Kavalan (pear_buya: IU1-2)

1. yau	baqi-AN	‘nay	usiq._
EXIST	elder.male-AN	that	one
存在	老人-AN	那	一

2. ...(2.4) matiw ta		ni-paluma-an-na		tu	sinsuli.
AF.go	LOC	NI-plant-LF-3SG.GEN		OBL	plum
去	處格	NI-種-處焦-3SG.屬格		斜格	李子

There was an old man who went to the orchard where he grew plums.
有一個老人去他種李子的果園。

Note: “yau usiq baqi-AN ‘nay” is also acceptable. The word “baqian” refers to any male elder. The word “baqi” is a kinship term, usually referring to a grandfather or grandchild/grandson, and is sometimes used to mean any senior relative; when used in “baqi=ay siqay”, it means a hundred-pacer, a kind of snake that is the one most honoured by the Kavalan people.

註解: “yau usiq baqi-AN ‘nay”也可以.baqi-an泛指老人,baqi是親屬詞,指祖父或孫子(通常跟sunis 小孩一起用),有時也指長輩,老人家,也可以指百步蛇(baqi=ay siqay). 噶瑪蘭人最尊敬的一種蛇.

The discourse features are transcribed according to Du Bois et al. (1993), a *de facto* standard in the linguistic community (see Appendix, Table 1). The only three differences are the symbols for truncated words (-), lengthening (=) and primary accent (^). Since there are many affixes and clitics in Formosan languages, which are coded with the symbols ‘-’ and ‘=’, we thus reserve the symbols, ‘-’ and ‘=’ for the grammatical coding of affixes and clitics, and use the symbols ‘--’ and ‘==’ to code truncated words and lengthening, respectively. The symbol ‘^’, originally used for primary accent in Du Bois et al. (1993), is now reserved to indicate a

change of intensity in the present corpus, where such change is perceived as a focused (or "emphasized") prominent syllable which sometimes does not coincide with the stressed pattern of a word.

We mostly follow a standard set of conventions, known as the *Leipzig Glossing Rules* in providing interlinear morpheme-to-morpheme glosses and grammatical information. These conventions were jointly developed by the Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology in conjunction with the Department of Linguistics of the University of Leipzig.[15] Some slight modifications, however, have been made due to the spontaneous speech nature of our data. For example, discourse function utterances indicating false start, pause filler, or uncertain hearing are glossed as FS, FIL, and X respectively. Another example is that the use of AF, PF, RF (IF), and LF to represent Agent Focus, Patient Focus, Referential Focus (Instrumental Focus), and Locative Focus[16] are added in addition to those specified in the *Leipzig Glossing Rules* so as to meet the specific conventional needs of the community of Formosan linguists.[17]

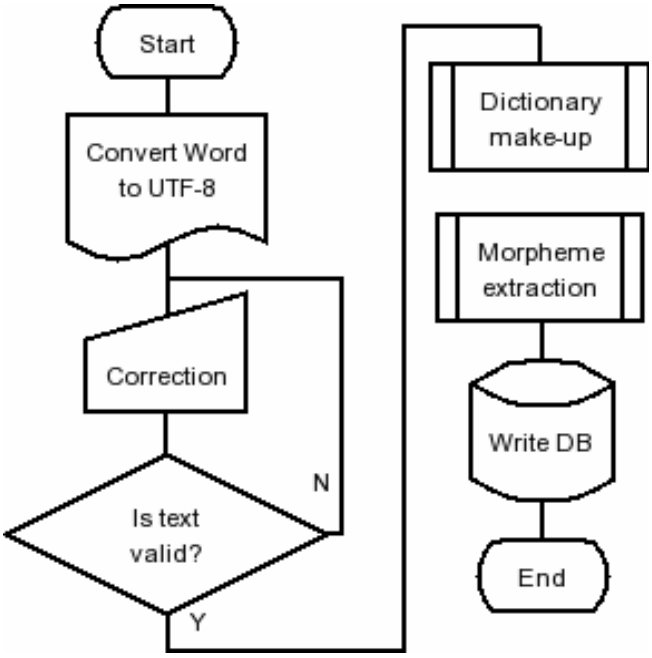
Once the transcription process is complete, the data is given to the database maintenance engineer for processing and storage. At this point, the online corpus can easily be accessed via the internet, for people to browse and search the corpus. A standard operation is also set for the database maintainer to handle fieldwork collections, as shown in Figure 4 below.

[15] See their website http://www.eva.mpg.de/lingua/files/morpheme.html for details.

[16] The term "focus" refers to a set of verbal morphology that signals the semantic role of the grammatical subject in a clause.

[17] See our website for a list of the grammatical markings we employ in our corpus.

Figure 4. Standard operation of text submission (Lin 2005:71)



3.3 Database Normalization

The choice of the right database engine greatly affects efficiency in data anagement and information retrieval. We use SQLite[18] rather than other enterprise database engines, for the following reasons in order to simplify the programming logic and allow for high-speed queries:[19]

[18] See http://www.sqlite.org for an introduction to this database engine. All relational database engines that follow the SQL92 standard can be used in the implementation of the schema.

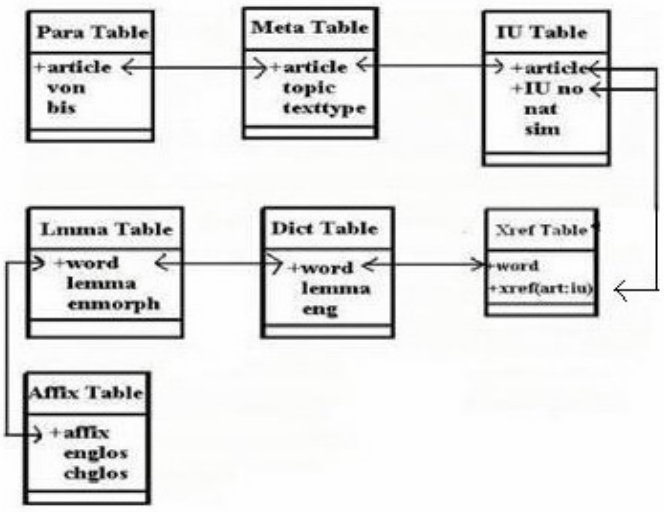
[19] To our best knowledge, different linguistic communities have used different database engines to manage text data. For example, Toolbox, developed by the SIL International, is used by some field linguists. The Academia Sinica Formosan Language Archive is run on Microsoft Access and the Jakarta Field Station of the Max Planck Institute for Evolutionary Anthropology uses customized FileMaker database software.

(7)

a. It is light-weight, fast and platform independent.
b. A database of one native language is stored in a single file, and so files are easy to maintain and administer.
c. It supports UTF-8 encoding.
d. It is free software.

In addition to a good database engine, database normalization is an important process to improve storage efficiency and data integrity. Each Formosan language is placed in an individual database which can be cross-related to other databases, and is stored in one single file in our corpus. The schema for every Formosan language is the same, including normalized tables such as Metadata Table, IU Table, Dictionary Table and Affix Table etc. Each table has a primary key value, which is important to cross-link the tables together when there is a query about different pieces of information. The relationships between the tables in the database are shown in Figure 5.

Figure 5. Relationships between the normalized tables in the database (Lin 2005:75)



The texts are mainly stored in IU Table format, in contrast to the word-based or sentence-based design in other archives. Each intonation unit is stored in one row, as illustrated in Table 3.

Table 3. IU table: Storage of a single intonation unit

Field name	Description
article	Pear 3 (file name)
no	2 (IU #)
nat	...(1.2) ima h-oem-angaw kaSna'itol ray kaehoey babaw (native words in Saisiyat)
sim	ima homangaw kasnaitol ray kahoy babaw (simplified native words)
eng	PROG <AF>set.a.ladder AF.move.up LOC tree above (English gloss)
chn	進行 <主焦>放樓梯 主焦.爬上去 處格 樹 上面 (Chinese gloss)

A simple query of "**%keyword%**", directed to every field listed above, produces certain search results. Variations in the spelling of the same word which result from possible variations in the pronunciation are also stored in one row for searching requests. A more detailed account of this will be provided in the next section. Words in the database are separated by a single space, so that they can be easily processed in programs by a single function (explode () in PHP and split () in Python). When no proper gloss is available when transcribing, capitalized forms of corresponding native morphemes or words are inserted temporarily. Thus, words and glosses may be automatically aligned, one-to-one, across the fields in our system.

Another special table structure is the Lemma Table, which is provided below:

Table 4. Lemma table: Coding schema for lemma

Field name	Format	Description
word	varchar(80)	Word (foreign key to an element of dict.lemma)
lemma	varchar(80)	Prefix-#stem#-suffix
enmorph	varchar(255)	Morphological marks in English, separated by commas
chmorph	varchar(255)	Morphological marks in Chinese, separated by commas
enstem	varchar(255)	Meaning of the stem in English, separated by commas
chstem	varchar(255)	Meaning of the stem in Chinese, separated by commas

This table illustrates how the various morphological parts of a single word are encoded in our corpus and in the dictionary. Each entry is marked and tagged with morphological information such as the stem and affixation, so that each affix and stem can be properly retrieved when a search is made for the word. For example, *kapapama'an* 'bicycle' in Saisiyat is stored as *ka-#papama'#-an* 'KA-ride-NMZ' in the table. If one looks for the prefix *ka-*, the suffix *–an* or the stem *papama'*, one can obtain the correct result by searching for the morphological elements before or after the first sharp sign (#) or after the second one.

4. IN-HOUSE SYSTEM

4.1 Back-end Programs and the POS-tagger

So far we have discussed how corpus design involves much more than just storing digital texts together. More specifically, the texts stored in our database need to undergo a great deal of pre-processing. This is done via a batch of back-end programs which are designed in-house and written in Python[20]. The submission of texts is handled through the use of these programs by the database maintainer, following the standard operation procedure described above in Figure 4, Section 3.2. Mismatches in alignment or failure in automated morphological analysis can be corrected immediately. Below is a list of a brief description of our back-end programs.

[20] See the following website for this programming language: http://www.python.org.

Table 5. Some back-end programs in the system

Name of back-end program	What the program does
features.py	It defines language-specific feature vectors and provides proper connection to the DSN (database source name).
simplify.py	It contains lists of parallel spellings occurring in different Formosan languages to reduce the possible duplication of entries.
tidy.py	It converts Chinese punctuations into ASCII and removes unnecessary Microsoft Word mark-ups.
canon.py	It checks the input validity, including metadata and text formats etc. It automatically stores the input in the database after it has passed the check.
extractmorph.py	It defines morphological and discoursal codes and extracts them from the texts.
makedict.py	It extracts information from submitted texts and updates the online glossary and dictionary.
mp3splt.py/mpgsplt.py	It segments sound and video files according to the IU time-label files created by the Praat TextGrid.

The coupling of the modules is fairly low. The programs *features.py*, *simplify.py* and *tidy.py* provide necessary first steps before all other programs.

Since each speaker may carry a slightly different dialectal accent, different transcribers may be inconsistent in spelling from word to word. Therefore, a set of feature vectors has been configured for each Formosan language so that the system of retrieval will not be confused by these possible pronunciation variations. The language-specific vectors transform the variants into (unified) simplified spelling forms, which help to speed up the query search. These are handled by the *features.py* and *simplify.py* programs. For example, the pronunciations of *a* and *ae* [æ] of the same word vary among Saisiyat speakers. Another common example is that glottal stops across many Formosan languages are either constantly dropped by some native speakers in fast paced speech, or else

are not so easily detectable by our transcribers; thus the word *'aehae'* "one" in Saisiyat could sometimes be spelled as *'ahae* or as *aehae*, and the word *'ima'*, a progressive marker, either as *'ima'* or as *ima*. Below are some feature vectors of Saisiyat.[21]

(8)
Saisiyat: ae [æ] → a, oe [œ] → o, S [ʃ] → s, ' [ʔ] → ø

A string substitution is executed before any other operation in the database in order to reduce the possible duplication of entries; otherwise full-text searches may result in incomplete information retrieval.

Most of our collected texts are transcribed and annotated in MS Word format, which is the most familiar and most frequently used system. Nevertheless, all the files done in this system automatically generate a Unicode BOM (byte-order-mark, U+FEFF) at the very beginning, which is absent in Unix-based systems. Such a byte-order mark (BOM) may cause a potential problem when reading files edited in different operating systems. This problem is solved by the program *tidy.py* in our design as soon as the text is submitted in order to fulfil the criteria of platform-independence.

Once the texts are submitted and processed by our system, the texts are stored in Unicode (UTF-8 encoding); the advantage of using such an encoding form is that it is easy to incorporate other languages which adopt other writing or phonetic systems, such as the IPA in the annotations. If some of the tribes decide to adopt non-ASCII letters, such as "ɖ ʈ ɽ ɬ ʔ", into their writing systems, our programming design will be able to process them correctly, without the need for modification or transformation.

In addition to the back-end programs, pre-processing of raw texts by part-of-speech (POS) tagging is also important in helping reducing the amount of human labour. A POS tagger has been designed especially for our Formosan corpus, and it is mainly based on Eric Brill's Transformation Based Learning (TBL) approach (Brill 1995).[22] The use of the POS tagger is still on trial at the current stage. It is hoped that, in

[21] Note that this is not to claim that *æ* and the glottal stop etc. are not phonemically important.

[22] See Lin and Sung (2004) and Lin (2005), who use our database of Saisiyat as a testbed for POS tagging training and parsing, for a detailed discussion of the scope of tag sets, TBL algorithm, accuracy rate and other related issues.

the future, as raw texts are uploaded into the database, they can be automatically tagged by the TBL tagger, and the online glossary and dictionary are also updated simultaneously. Any time that the database maintainer finds an error in the tagged corpus, it can be corrected online with immediate feedback to the tagger and the tagger can be retrained later.

4.2 Unified Output Interface

The development of internet browsers has enabled the creation of bodies of text, image, sound, and video that can easily be searched for, allowing a wide range of users to quickly and easily access the information provided in the corpus. For this to happen, we have chosen the most economical channel to offer the widest possible access to our data: namely, the Internet. Thus an open web-based resource dissemination system has been created, and a unified user-friendly web interface has been developed in PHP for this goal. Our corpus system follows the HTML 4.01 specifications proposed by the World-Wide Web Consortium (W3C)[23] and is designed to be browsed with all major browsers for the purpose of accessibility. For a dynamic and interactive representation, either the Document Object Model (DOM) [24] or JavaScript 1.2 is preferred. Other popular browsers, such as Internet Explorer 5.0, Mozilla 1.7, Firefox 0.9 and Opera 4, are also compliant to these standards. Figure 6 is a screen screenshot of our corpus output.

[23] See http://www.w3.org/TR/REC-html40/ for the latest version of HTML 4.01.
[24] See http://www.w3.org/DOM/ for an explanation of the DOM.

Figure 6. A screenshot from our website

As shown in Figure 6, the displayed output is composed of the following parts: (i) a set of metadata on the top frame of the displayed output, (ii) the speaker's photo at the bottom-left frame where, audio or video clips can be played, with several switches placed underneath to adjust the browsing effects, and (iii) a frame at the bottom-right of the screen for the selected texts following our transcription standards. At the end of each IU and each sentence which may contain more than one IU, corresponding segmented audio or video clips are provided for the users to listen to or watch the actual recording of the setting. Due to the bandwidth limits at different web hosts, when uploaded online our audio and video data are transferred and stored in 16kbps (or 24 kbps) 11kHz MPEG-1 Audio Layer 3 (MP3) and MPEG-1 video.

In addition, our corpus features a cross-referencing capability: an online glossary, which can be readily retrieved every time a word token in the text is clicked, as seen in Figure 7, and which provides users links to more sample sentences and to the source texts.

Figure 7. Pop-up glossary with cross-references

5. THREE SPECIAL FEATURES OF THE NTU CORPUS

As we have described so far, the corpus has been set up to enable any user, not only a linguist or the language community, but anybody who is interested in Formosan languages and cultures, to access the valuable linguistic data available through a diverse array of formats via the most convenient means possible. Three special features, among many, are included within the NTU Corpus: (i) an online dictionary, (ii) a multilingual search function and (iii) interoperability, which we will illustrate in details below.

5.1 Online Dictionary

One special feature of our database is that it can automatically generate an online dictionary, accompanied by a word token count.[25] The counting of tokens is updated as new texts are uploaded. As shown in Figure 8, when choosing a language and entering the corpus, the user

[25] This function is currently restricted to internal use by transcribers and primary investigators of this project. We plan to make it available for the general public in the future.

can find the "Generate a dictionary" function above the list of the texts. The contents of the dictionary can be printed out and also cut and pasted for any sort of linguistic analysis. Figure 9 below is an excerpt from the Saisiyat dictionary in the Saisiyat corpus after pressing the [Generate a dictionary] icon.

Figure 8. A screen snapshot of the Saisiyat corpus

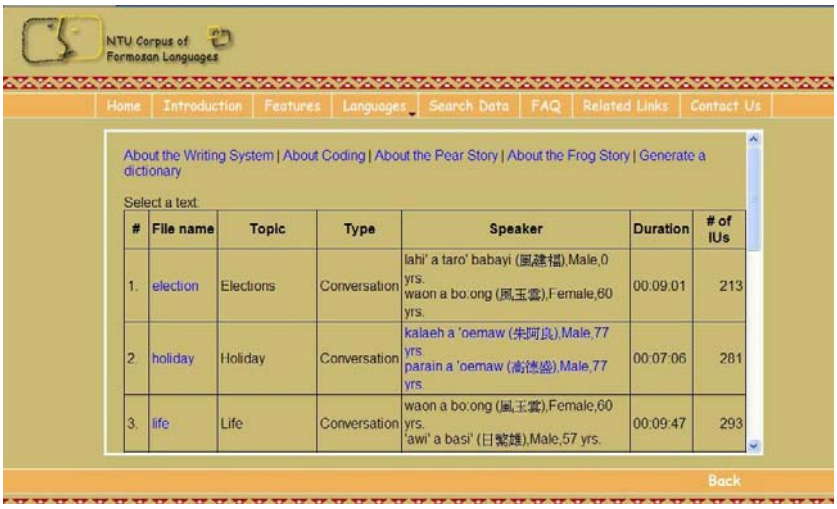
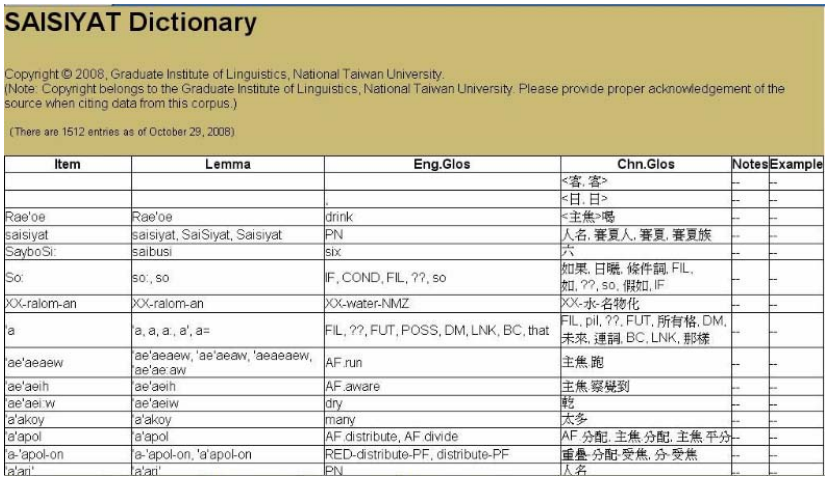


Figure 9. Online dictionary of the Saisiyat corpus



SAISIYAT Dictionary

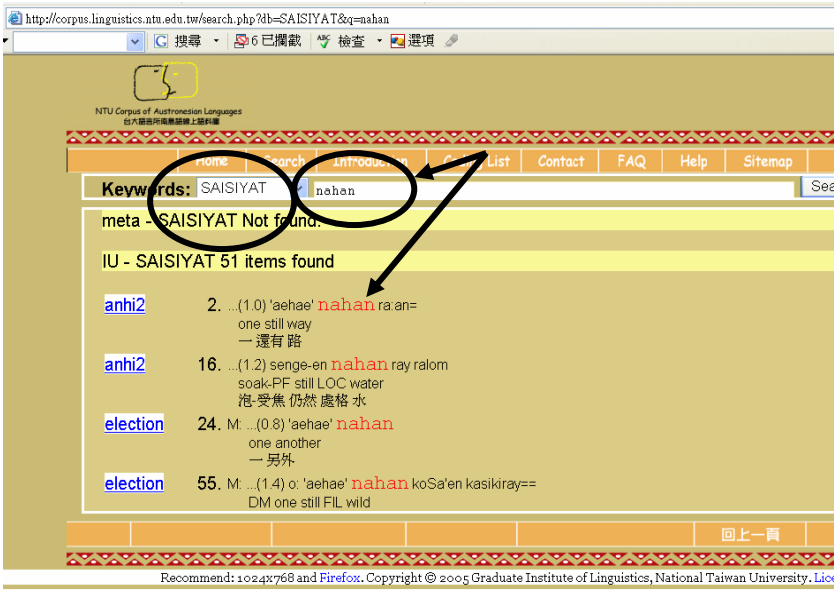
(There are 1512 entries as of October 29, 2008)

Item	Lemma	Eng.Glos	Chn.Glos	Notes	Example
			<客, 客>	--	--
		.	<日, 日>	--	--
Rae'oe	Rae'oe	drink	<主焦>喝	--	--
saisiyat	saisiyat, SaiSiyat, Saisiyat	PN	人名, 賽夏人, 賽夏, 賽夏族	--	--
SayboSi:	saibusi	six	六	--	--
So:	so:, so	IF, COND, FIL, ??, so	如果, 日曬, 條件詞, FIL, 如, ??, so, 假如, IF	--	--
XX-ralom-an	XX-ralom-an	XX-water-NMZ	XX-水-名物化	--	--
'a	'a, a, a:, a', a=	FIL, ??, FUT, POSS, DM, LNK, BC, that	FIL, pil, ??, FUT, 所有格, DM, 未來, 連詞, BC, LNK, 那樣	--	--
'ae'aeaew	'ae'aeaew, 'ae'aeaw, 'aeaeaew, 'ae'ae:aw	AF.run	主焦.跑	--	--
'ae'aeih	'ae'aeih	AF.aware	主焦.察覺到	--	--
'ae'aei:w	'ae'aeiw	dry	乾	--	--
a'akoy	a'akoy	many	太多	--	--
a'apol	a'apol	AF.distribute, AF.divide	AF.分配, 主焦.分配, 主焦.平分	--	--
a-'apol-on	'a-'apol-on, 'a'apol-on	RED-distribute-PF, distribute-PF	重疊.分配.受焦, 分.受焦	--	--
a'ari'	a'ari'	PN	人名	--	--

5.2 Multilingual Search Function

Another special feature is that our corpus allows web users to search for any specific target word or morpheme in English, Chinese or any of the Formosan languages contained in the corpus. When a search is made for a lexeme, e.g., *nahan* in Saisiyat, the search function can retrieve all occurrences of the lexeme specified. The search result for *nahan* is shown in Figure 10.

Figure 10. Search result for the lexeme *nahan* in the Saisiyat corpus



Users can also search for one particular lexeme across the Formosan language corpora. For example, by typing the key morpheme *ma-* and selecting the "all languages" option, all related instances in different texts of all Formosan languages will be generated, as shown below in Figures 11, 12, and 13.

Figure 11. Search result for the prefix *ma-* (Page 1 of 31 pages)

Figure 12. Search result for the prefix *ma-* (Page 2 of 31 pages)

Figure 13. Search result for the prefix *ma-* (Page 3 of 31 pages)

Each text and each Formosan language in our corpus is stored in a cross-related file with the same normalized tables, and cross-text and cross-language searches can be executed with a single command. The number of tokens of the queried lexeme is also revealed at the same time The search function is geared to provide cross-referencing retrieval, which may lead web users directly to the text context where each instance occurs for further research.

5.3 Interoperability

The final special feature of our corpus regards the issue of interoperability. Given the amount of time, money, and labour spent in constructing the corpus, it is of no use if the encoded data cannot be shared within the linguistic community. Thus, it is important for the system to provide interoperable functions so that the data can be exchanged between various systems. To achieve this, the Extensible Mark-up Language (XML)[26], now a *de facto* standard on the web and the

[26] Extensible Markup Language (XML) is a simple, very flexible text format designed to handle the exchange of a wide variety of data. See http://www.w3.org/XML/ for a detailed explanation.

most preferred data format for archiving, is employed to facilitate data interchange or transformation between different systems. When requested by other institutions, our program can export the text materials as XML, carrying every morphological detail, aligned glosses and part-of-speech of each word in a uniform format. In so doing, all researchers of natural language processing are able to profit from our linguistic fieldwork. An example of an exported format is given below.

(10)

```
<?xml version="1.0" encoding="utf-8" ?>
<article id="pear_imuy">
  <topic>Pear Story</topic>
  <language>Kavalan</language>
  <dialect>Xinshe</dialect>
  <speaker>
    <natname>imuy</natname>
    <chnname>潘金妹</chnname>
    <gender>F</gender>
    <age-of-record>51</age-of-record>
  </speaker>
  <duration>00:01:15</duration>
  <total-iu>31</total-iu>
  <collected>2003-05-30</collected>
  <revised>2006-03-01</revised>
  <transcriber>葉俞廷(Yu-Ting Yeh)</transcriber>
  <transcriber>王以勤(Yi-Qin Wang)</transcriber>
  <transcriber>鍾曉芳(Siaw-Fong Chung) </transcriber>
  <transcriber>沈嘉琪(Chia-chi Shen) </transcriber>
  <transcriber>謝富惠(Fuhui Hsieh)</transcriber>
  <doublecheck>林東毅(Dong-yi Lin)</doublecheck>
  <text>
    <iu id="iu_1">
      <word>
        <nat>tangi</nat>
        <sim>tangi</sim>
        <eng>today</eng>
        <chn>今天</chn>
        <pos>RB</pos>
      </word>
```

```
            <word>
                ...
            </word>
        </iu>
        <iu id="iu_2"> ...   </iu>
        ...
        <para von="1" bis="4">
            <eng>I just saw a person there ...</eng>
            <chn>我剛剛看到 ...</chn>
            <notes>Some field notes</notes>
        </para>
        ...
    </text>
</article>
```

6. CONCLUDING REMARKS

With the advances in storage of data, character-encoding and web technology, the development of an online platform of corpora has profound consequences for the accessibility, quantity, and quality of linguistic field data. This is of particular significance for Formosan languages in Taiwan, which are confronted with an immediate and growing crisis of endangerment. As a response to the recognition of this burgeoning problem, the key objectives of the construction of the NTU Corpus of Formosan Languages aim to document and thus preserve valuable spoken data in actual discourse context, as well as relevant ethnological and cultural information. It is designed to be extendible in order to include the processing of other Formosan languages, once the proper feature vectors are set for every language. It has been developed to house virtually unlimited quantities of easily accessible linguistic data. A total of 4186 IUs are available at present from our corpus online. It is the goal of our team to continue to expand our corpus. More texts of Kavalan, Tsou, Saisiyat, Amis, Sakizaya, Seediq and Bunun are now being annotated, and are scheduled to be uploaded online in the future.

As our system is designed with an emphasis on multilinguality, cross-referencing retrievability and interoperability, it will bring great benefit to linguists, language instructors and native speakers of Formosan languages who care about the conservation and transmission

of their own languages and cultures. As pointed out in Diamond (2000), Formosan languages are Taiwan's gift to the world. The development of the NTU Corpus of Formosan Languages is tailored to preserve this precious treasure. Furthermore, many Formosan languages are on the verge of becoming endangered, or even extinct in the near future. Given this impetus, more researchers and local Formosan language communities are urged to participate in this very worthwhile project, to use and to promote the enlargement of the corpora.

APPENDIX

Table 1. Symbols for discourse transcription (adopted from Du Bois et al.(1993))

Meaning	Marker
Units	
Intonation Unit	((newline))
Truncated IU	--
Word	((space))
Truncated word	--
Speaker identity / turn start	:
Speech Overlap	[ ]
Transitional Continuity	
Final	.
Continuing	,
Appeal	?
Terminal Pitch Direction	
Fall	\
Rise	/
Level	_
Accent and Lengthening	
Primary accent	^
Secondary accent	`
High booster	!
Low booster	;
Lengthening	= =
Tone	
Fall	\
Rise	/
Fall-Rise	\/
Rise-fall	/\
Level	_

Pause	
Long	...(N)
Medium	...
Short	..
Latching	0
Vocal Noises	
Vocal noises	(CAPITAL LETTERS)
Inhalation	(H)
Exhalation	(Hx)
Glottal stop	%
Laughter	@
Quality	
Quality	<Y Y>
Laugh quality	<@ @>
Quotation quality	<Q Q>
Phonetics	
Phonetic / phonemic transcription	(/ /)
Transcriber's Perspective	
Researcher's comment	(( ))
Uncertain hearing	<X X>
Indecipherable syllable	X
Specialized Notations	
Duration	(N)
Intonation unit continued	&
Accent unit boundary	|
Embedded IU	<| |>
Restart	{Capital Initial}
False start	< >
Code switching	<L2 L2>
Nontranscription line	$
Reserved Symbols	
Phonetic / orthographic symbols	'
Morphosyntactic coding	+ * # { }
User-definable symbols	" ~

REFERENCES

Berman, Ruth and Dan Slobin. 1994. *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Brenzinger, M. (ed.) 1992. *Language Death: Factual and Theoretical Explorations with Special Reference to East Africa*. Berlin: Mouton de Gruyter.

Brill, Eric. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 21.4:543-565.

Chafe, Wallace L. (ed.) 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, N.J.: Ablex Publishing Corp.

Chafe, Wallace L. 1987. Cognitive constraints on information flow. *Coherence and Grounding in Discourse*, ed. by Russell S. Tomlin, 21-51. Amsterdam: John Benjamins.

Chafe, Wallace L. 1994. *Discourse, Consciousness and Time*. Chicago: University of Chicago Press.

Chen, Hsueh-hua. 2007. Taiwan nandaoyu yuliaoku quanshi ziliao fazhan zhi yanjiu [The study of metadata creation for Formosan language archives]. *Yuanzhuminzu Yuyan Fazhan Luncong [Studies on the Development of Indigenous Languages]*, 43-62. Taipei: Council of Indigenous Peoples.

Couper-Kuhlen, E., and Thompson, S. 2000. Concessive patterns in conversation. *Cause-condition-concession-contrast: Cognitive and Discourse Perspectives*, ed. by E Couper-Kuhlen and B. Koremann, 381-410. Berlin: Mouton de Gruyter.

Crawford, J. 1995. Endangered native American languages: What is to be done and why? *The Bilingual Research Journal* 19.1:17-38.

DeLancey, Scott. 1981. An interpretation of split ergativity and related patterns. *Language* 57.3:626-657.

Diamond, Jared M. 2000. Linguistics: Taiwan's gift to the world. *Nature* 403:709-710.

Du Bois, J. W. 1985. Competing motivations. *Iconicity in Syntax*, ed. by John Haiman, 343-365. Amsterdam: Benjamins.

Du Bois, J. W. 1987. The discourse basis of ergativity. *Language* 63:805-855.

Du Bois, J. W, Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. *Talking Data: Transcription and Coding for Language Research,* ed. by J. A. Edwards and M. D. Lampert, 45-90. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Heine, Bernd. 1997. *Cognitive Foundations of Grammar*. Oxford/New York: Oxford University Press.

Himmelmann, Nikolaus. 2006. The challenges of segmenting spoken language. *Essentials of Language Documentation*, ed. by Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel, 253-274, Berlin: Mouton de Gruyter

Hopper, Paul J., and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56:251-299.

Hopper, Paul J., and Sandra A. Thompson. 1984. The discourse basis for lexical categories in Universal Grammar. *Language* 60.4:702-752.

Huang, Shuanfan, Lily I-wen Su, and Li-May Sung. 2001. *A Reference Grammar of Tsou*. NSC Technical Report (NSC 89-2411-H-002-007-M7). National Taiwan University.

Huang, Shuanfan, Lily I-wen Su, and Li-May Sung. 2002. *A Functional Reference Grammar of Saisiyat*. NSC Technical Report. (NSC 90-2411-H-002-049). National Taiwan University.

Huang, Shuanfan, Lily I-wen Su, and Li-May Sung. 2003. *A Functional Reference Grammar of Saisiyat*. NSC Technical Report (NSC 91-2411-H-002-089). National Taiwan University.

Huang, Shuanfan, Lily I-wen Su, and Li-May Sung. 2004. *A Functional Reference Grammar of Saisiyat*. NSC Technical Report (NSC 92-2411-H-002-077). National Taiwan University.

Huang, Shuanfan, and Li-May Sung. 2005. *A Functional Reference Grammar of Kavalan*. NSC Technical Report (NSC 93-2411-H-002-095-MG). National Taiwan University.

Huang, Shuanfan, Jozsef Szakos, and Li-May Sung. 1999. *A Functional Reference Grammar of Tsou*. NSC Technical Report (NSC 88-2411-H-002-050). National Taiwan University.

Krauss, M. 1992. The world's languages in crisis. *Language* 68:6-10.

Lerner, Gene H., and Takagi, T. 1999. On the place of linguistic resources in the organization of talk-in-interaction. *Journal of Pragmatics* 31:49-75.

Lin, Zhemin and Li-May Sung. 2004. Tiny corpus applications with transformation-based error-driven learning: Evaluations of automatic grammar induction and partial parsing of SaiSiyat. *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation*, 197-204. Tokyo: Waseda University.

Lin, Zhemin. 2005. *Automatic processing of languages with small-scaled corpus: Part-of-speech tagging and partial parsing SaiSiyat and applications*. M.A. Thesis. National Taiwan University.

Mayer, Mercer. 1980. *Frog, Where Are You?* New York: Dial Books.

Ono, Tsuyoshi, and S. A. Thompson. 1995. What can conversation tell us about syntax? *Alternative Linguistics: Descriptive and Theoretical Modes*, ed. by P. W. Davis, 213-271. Amsterdam: John Benjamins.

Robins, R. H., and Uhlenbeck, E. (eds.) 1991. *Endangered Languages*. Oxford: Berg.

Sack, Harvey, Emanuel A. Schegloff and Gail Jefferson. 1974. A simplest system for the organization of turn-taking for conversation. *Language* 50:696-735.

Schegloff, Emanuel A. 1980. Preliminaries to preliminaries: "Can I ask you a question?" *Sociological Inquiry* 50:104-151.

Schegloff, Emanuel A. 1988. Pre-sequences and Indirection. *Journal of Pragmatics* 12:55-62.

Schegloff, Emanuel A. 1991. Conversation analysis and socially shared cognition. *Perspectives on Socially Shared Cognition*, ed. by Lauren B. Resnick, John M. Levine and Stephanie D. Teasley, 150-171. Washington D.C.: American Psychological Association.

Schegloff, Emanuel A. 1996. Turn organization. *Interaction and Grammar*, ed. by E. Ochs, E. Schegloff and S. Thompson, 52-133. New York: Cambridge University Press.

Schmidt, A. 1990. *The Loss of Australia's Aboriginal Language Heritage*. Canberra: Aboriginal Studies Press.

Strömqvist, Sven and Ludo Verhoeven. (ed.) 2003. *Relating Events in Narrative, Volume 2: Typological and Contextual Perspectives*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Su, Lily I-wen, Li-May Sung, Shuping Huang, Fuhui Hsieh and Zhemin Lin. 2008. NTU Corpus of Formosan Languages: A state-of-the-art report. *Corpus Linguistics and Linguistic Theory* 4-2, 291-294.

Tao, Hongyin. 1996. *Units in Mandarin Conversation: Prosody, Discourse and Grammar*. Amsterdam: John Benjamins.

Tuan, Yi-Fu. 1974. *Topophilia: A Study of Environmental Perception, Attitudes, and Values.* Englewood Cliffs, N.J.: Prentice Hall.

Zeitoun, Elizabeth, Ching-hua Yu, and Cui-xia Weng. 2003. The Formosan language archive: Development of a multimedia tool to salvage the languages and oral traditions of the indigenous tribes of Taiwan. *Oceanic Linguistics* 42.1:218-232.

Zeitoun, Elizabeth, and Ching-Hua Yu. 2005. The Formosan language archive: Linguistic analysis and language processing. *Computational Linguistics and Chinese Language Processing* 10.2:167-200

Li-May Sung
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan 106, ROC
limay@ntu.edu.tw

Developing An Online Corpus of Formosan Languages

Lily I-wen Su
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan 106, ROC
iwensu@ntu.edu.tw

Fuhui Hsieh
Dept. of Applied Foreign Languages
Tatung University
Taipei, Taiwan 104, ROC
hsiehfh@ttu.edu.tw

Zhemin Lin
Winstron Corporation
philippe_lin@wistron.com