Documentation of Formosan Languages

Lily I-wen Su
National Taiwan University

0. Introduction

Formosan languages spoken by the aborigines of Taiwan are rapidly losing their speakers. With almost no exception, speakers of the Formosan languages are limited to the older generations. This loss of language entails, of course, loss of its respective cultural heritage and even that of ethnic identity.

The study of the Formosan languages started some 65 years ago by Japanese scholars, during the time when Taiwan was still under the occupation by the Japanese. Their works on the Formosan languages pave quite a solid foundation for later investigation of these languages. Such studies use however the traditional field method of data solicitation from individual informants. Most of the data are not available to later researchers. Even if the data were well preserved, they are texts or isolated sentences transcribed in the conventions used by researchers. Such practice was in use until we began our project series, first on Seediq, then Tsou, and now Saisiat.

In addition to sentences elicited, our data comprise mainly of spoken discourse produced by our informants. The isolated sentences, recorded via the traditional question-answer method between the informant and the field worker, are not totally out of its usage contexts. The extended spoken texts include discourse of two major types: narratives and face-to-face conversations between two native speakers recorded in their most natural settings. The narratives come mainly from the informants' recounting of the story right after watching the famous Pear Stories, a film made in early 80s used widely by the discourse analysts. A small fraction of our narratives are obtained by the informants' retelling based on some of the written texts prepared by the previous researchers.

What singles us out from other research groups working on the Formosan languages lies in the way in which the data are transcribed. As truly believers of the so-called "emergent grammar," we think grammar is a result of "frozen pragmatics." For this reason, any possible "deviations" found in the data solicited, be it a prolonged pause, a false start, a repetition, or a "corrected error," is considered of highly significance in the understanding of the grammar of the language. A narrow transcription method designed by DuBois is therefore adopted for the purpose of our study.

Since there is no standardized writing system for any of the Formosan languages, we transcribe the data elicited according to the system customized for the purpose of our study. The system is made up primarily of the alphabets, in conjunction with

diacritics necessary for the sake of pronunciation.

1. Method of Transcription

The data elicited for our projects are transcribed according to the transcription convention proposed by DuBois et al (1993). Some details are omitted, as they might not be of direct importance and relevance to our research. The data, all spoken in nature, are not recorded based on the traditional concept of a "clause," in the sense normally associated with the written language. Instead, each unit in our corpus is delineated as an intonation unit that resembles the production of natural language, bordered with pauses or intonational differences. Some minor phonetic details pertaining to each intonation unit are ignored, but pitch usually serves as a reliable cue to delineate one unit from the other if no pause is observed. Linguistic phenomena such as lengthening, truncation, self-correction are also jogged down in order to best capture the production of spontaneous speech.

The table below summarizes the symbols and their functions as they are used in our databases:

Units:

Intonation unit	{carriage return}
Truncated intonation unit	--
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	^
Lengthening	=

Pause:

Long	...(N)
Medium	...
Short	..
Latching	(0)

Vocal Noises:

Vocal noises	( )

Inhalation	(H)
Exhalation	(Hx)
Glottal Stop	%
Laughter	@

Quality:

Quality	<Y Y>
Laugh quality	<@ @>
Quotation quality	<Q Q>

Phonetics/phonemic transcription:

Phonetic/phonemic transcription	(/ /)

Transcriber's Perspective:

Researcher's comment	(( ))
Uncertain hearing	<X X>
Indecipherable syllable	X

Specialized Notations:

Intonation unit continued	&
Code switching	<L2 L2>

2. Types of Data

The Taita Spoken Formosan databases are composed of data from various sources, most of them from field sessions via oral elicitation. The majority of them fall under the genres of narratives and face-to-face conversations. In addition, we also elicit data of personal interest via individual field sessions.

2.1. Narratives

The narratives consist of three major categories: legends and folklores, pear stories, and elicited stories. Each narrative is told by one single speaker, first recorded then transcribed with the help of an experienced informant who helps to identify the sound and meaning of some difficult lexical items.

2.1.1. Legends and Folklores

Legends and folklores related to the tribes are a good source of the narratives to be elicited. It is not an easy task for only some of the older speakers of the language are proficient and knowledgeable enough to recount such stories. Incidentally, most of the Formosan speakers are also pious Christians, and are also good narrators of the Bible stories. All of the narratives in our databases are obtained via the primary source. There are, however, some narratives recounted by competent speakers based on legends or folklores available from the field notes or

published works of our predecessors. Whatever the source, all are transcribed by the transcription method proposed by DuBois, as mentioned above.

2.1.2. Pear Story

This category comprises of data based on the pear film (Chafe 1980). Subjects are first asked to watch the pear film (the colored version), then describe what they have just seen after 10 minutes of recess in their native languages, either in Seediq, Tsou or Saisiat since they are the only three we have studied so far. The data is then narrowly transcribed in the same method mentioned above. All the linguistic and paralinguistic details are recorded, including speech errors, false starts, pauses, repetitions, repairs, significant voice qualities etc.

2.1.3. Elicited Stories

This category comprises of data elicited with instruction given in Mandarin Chinese. Subjects are first offered a story recounted in Mandarin Chinese (self-fabricated by the researchers). They are then, after 10 minutes of recess, asked to recall the story in their native languages. The transcription method is the same as mentioned above. The idea is not to retell the story correctly in terms of its plot. We are simply interested in the elicitation of spontaneous, unplanned discourse.

2.2. Conversation

For conversational data, we record then transcribe, following the same method, the face-to-face interaction between two native Formosan language speakers on practically any topic that is of interest to the two participants. In the presence of a recorder, the speech participants are normally too nervous to produce natural language in use. This being the situation, we often end up by using only part of the recorded conversation, leaving behind the very beginning portion of the talk exchange. As with any other types of data, what interests us is not the content of the talk exchange – it is the linguistic interaction that is of relevance to our exploration. The end product of our research – the functional grammar of the language under study – should be of descriptive as well as explanatory adequacies.

2.3 Field Notes

Much of our research data are acquired through the numerous field sessions with our informants. These informants are usually experienced language specialists referred to us by other researchers or by others familiar with the tribes. Each session is conducted by different researchers and tailored in accordance with his/her particular needs. That is, the session is designed in such a way so that the data elicited can be of help to meet his/her research interest. Field notes from these sessions are not made public, which is one of the greatest loss and waste of our research effort.

3. The Problems Faced

There are however technical problems regarding the way our data are currently stored and retrieved: For one thing, keyword search is prohibited from our somewhat

primitive databases. We now store our transcribed data in MS-Word format. Via this format, we can easily access, by the FIND function, the context where a linguistic target occurs in a particular file. We are nevertheless unable to get hold of all the occurrences, against the entire database, of this very item with one single command. One needs to calculate, file after file, the occurrences of this item manually in order to know its frequency. It becomes troublesome indeed if one were to get a hard copy of all the occurrences! This poses, of course, no big problem with the Key-Word-In-Context (KWIC) corpus.

Even if Corpus Wizard KWIC alternative were available, our current technology would not allow us to get a good sense of the context in which the target item occurs. The fact that our data is transcribed in the so-called intonation unit (IU) makes it sometimes difficult to view the target item beyond the IU where it is contained. The software stipulates that the maximum of immediate context that can be called upon is limited to an approximation of only twenty-five words surrounding the target item for written corpus. This will definitely entail far less for spoken corpus. Given such restriction, we will have to sacrifice the advantage of contextual information, which is of great importance to our discourse analysis methodology, in exchange for the convenience of keyword search.

It will thus be ideal if the beauty of keyword search and the provision of a wider context can be combined. Within this new program, we hope that every time a keyword is searched, we may simply click on the keyword and the link to its linguistic context can become available automatically.

As our data are not tagged, it is inconvenient and time-consuming as far as the searching process goes. It would be nice if wordlist can be generated automatically once the input of a lexical item, with grammatical information and meaning, is completed. It would be even nicer if this wordlist can work bi-directionally: once a linguistic item from the source language is entered, its grammatical information and meaning can be generated from the wordlist. One software named Shoebox seems to hint at such a possibility, but it is not quite user-friendly.

Tagging is a must in the future, but it has to be done under a unified system. If not, problems still exist regarding the application of the data. At present, discrepancies in our glass are observed due to a disintegrated coding (or tagging) system. Basically, our data are given in phonetic alphabets accompanied in the second line by their English equivalents, coupled with grammatical information such as focus, case, person, number, and tense. Yet, problems are encountered because glossing is not done consistently. The following Tsou example may illustrate this point.

In Tsou, the verb *ako* (AF)*/aka* (PF) is glossed either as "keep on; continue" (1, line 108), or as "always" (2, line 21):

(1)

106 …(0.7)ine mio

at that time

107 …isi cu=

NAF-3rd

108 (0) **aka** ta'totohUngva eainca ma sia na mo meo'eo'i,
keep_on think-NAF say-NAF who Nm AF steal-AF
"(And) at that time he kept thinking of who might have stolen the pears."
(pear3:106-108)

(2)
21 A: 'ahUeU **ako** na'no eahioa 'o c'o.....
stubborn **always** very exist-work

'a mo na'no **ako** atutumzo
AF very **always** painful
"She was so old and very stubborn. She always went to work. In fact, she always felt pain."(07062000:20-21)

Similarly, *asngUcU* (AF)*/asngUcva* (PF) is glossed either as "always" (3, line 82) or "continuously" (4, line 118):

(3)
82 …(1.1)isi **asngUcva** tiatatvia no cmoi ho easasa ho
NAF-3rd. **always** carry_with_fingers Obl bear and drag_along and

83 …tesi akoeva no
FUT-3rd.intend to
84 …p'aeni to oko-si
feed Obl child-3rd.
"He is always carried and dragged by the bear. The bear intends to feed its child." (bear:82-84)

(4)
117 mo maica 'o mo angu UmnU ci piepiya
AF like_this Nm AF too good Rl soul
'The soul has been too good. (Sarcastically. How unfortunate she is!)'
118 **asngUcU** easiungu mevavoezuhu mevcongU
continuously 安 marry_many_times marry
'Many times she got married, all with easiyungu安.' (dailylife:117-118)

And the verb *smoeoa* is glossed both as "fear" (5, line 28) and "afraid" (6, line364).

(5)
27 …(3.1) mo mainci=
AF why
28 …aac'o eainca ateueuna **smoeoa** ci eoi ta ceoa.
above_all say-PF all_together **fear** Rl insect Gen earth
"But why above all did they say that this is the most fearful insect on earth?"

(6)
363 …ho isi cu eaica
when NAF-3rd Perf say-NAF
364 ..aana te s'a ahta **smoeoa** ho isi cu afu'a
again(Neg) Fut Adv longer **afraid** conj NAF-3.s.g Perf once
-NAF -NAF
365 ….sia to fucu no meemeno.
put-NAF Obl bag Gen iron
"Once put the snake into an iron net, people would not be afraid of it any more."

This multiple notations of a single word by different transcribers may blur the authentic use we intend to explore. And the different parts of speech noted down by different transcribers may create more difficulties in the retrieval of the data.

We also encounter problems in glossing when a word may appear in more than one spoken forms. For instance, many words in Tsou may be pronounced in either their full or abbreviated forms. The verb *'a'usni* (7, line 140) may be abbreviated as *'a'uni* (7, line 141). The situation can be chaotic if it is glossed with different English translation.

(7)
138.…(1.4)t'aunana
consider-PF
139 …(0.9)osi cu c'o nana asngUca eainca no koeu-si:
NAF-3rd. only uninterruptedly say Nom his ear
140 …(0.7)a'UmtU.. 'ana nte'o s'a.. eatatiskova.
really-AF no_longer n-Fut-1st. man
tesi te'o cu anana'va eainca **'a'uni** no eainca
Fut-1st. indeed say **suppress-PF** (tesi, repaired by te'o)
"The man was thinking that (if I had not kicked the bear), I would have not been a man/living."
141 ..te-si cu anana'va eainca **'a'usni** no cmoi.. maitan'e.
Fut-3rd. indeed say **defeat-PF** Obl bear now
"I would have been defeated/killed by the bear (if I hadn't done that)."

Similarly, different dialects may have different morphological form of the same verb. For example, the Tsou verb *cohivi* means "to know" in the Tfuya dialect. Some native speakers in Tapang may refer to the same word with another variant *cohivite*. Such confusion may be avoided should field notes be included or cross-references be provided.

All the problems stated here may result in difficulty if statistics of the distributional patterns is of crucial importance. A good dictionary compiled out of the data collected via our fieldwork may be the final answer to the problem. This dictionary should cross-reference all the possible variations, dialectal or not, of a given word. Ideally, syntactic and semantic aspects, or even notes on how the word may be used should also be included. If possible, anthropological and etymological should also be furnished.

References

Chafe, Wallace (ed) 1980. The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production. Norwood, NJ: Ablex.

DuBois, John W., Stephan Schuetze-Cobum, Susanna Cumming, and Danae Paolino. 1993. Outline of Discourse Transcription. In J. A. Edwards and M. S. Lampert (eds). *Talking Data: Transcription and Coding for Language Research*. Hillsdale, NJ: Lawrence Erlbaum.

Huang, Shuanfan, Lily I-wen Su & Limay Sung. 2002. Tsou Reference Grammar. NSC Project Report.