6

Documenting and Revitalizing Kavalan

Fuhui Hsieh and Shuanfan Huang

Tatung University and National Taiwan University

The purpose of this paper is to provide a two-dimensional approach to language documentation (Himmelmann 1998). In addition to building a database, we also conducted a sociolinguistic survey designed to document the state of health of a language in a particular spatio-temporal frame. Our goal is to share our fieldwork experience of documenting Kavalan, a seriously endangered language in southeastern Taiwan now spoken by fewer than just a few dozen speakers. We first discuss our field experiences in working with speakers of Kavalan in Sinshe village, the only significant Kavalan settlement left in Taiwan, and the state of the Kavalan language, based in part on Huang and Chang's (1995) earlier sociolinguistic survey, and in part on a recent more in-depth village-wide survey of language use in the community.Next, we introduce the NTU Corpus of Formosan Languages, part of which incorporates our corpus data in Kavalan. The NTU Corpus of Formosan Languages aims to establish a standard for the creation of linguistic corpus databases through the application of information technology to linguistic research. The creation of this linguistic database enables us both to preserve valuable linguistic data and to provide a systematic recording of these languages, for the benefit of future linguistic research.

1. INTRODUCTION. As the world's languages are dying at an unprecedented speed, language documentation has now been widely recognized as an important aspect of linguistic research.[1] As many as half of the estimated 6,000 languages spoken in the world are 'moribund' (Krauss 1992); in other words, these languages are spoken by adults only and are not being passed on to the next generation (cf. Crawford 1995). It has been estimated that ninety percent of the existing languages today are likely to die or become seriously endangered in the near future; and this phenomenon is particularly acute in Americas, Africa, Australia and Southeast Asia (Brenzinger 1992; Robins and Uhlenbeck 1991; Schmidt 1990). For most of the Austronesian languages spoken in Taiwan, the language crisis is imminent, and that makes documentation of Formosan languages all the more urgent. According to a most recent census report of the Council of the Indigenous Peoples in Taiwan, in August 2007, of the twelve indigenous tribes in Taiwan, there are five tribes whose populations number less than ten thousand each: Tsou (6,432), Saisiyat (5,514), Yami (3,125),

[1] We would like to thank the head of the Development Association of the Sinshe Community, Mr. Yinhua Pan (潘銀華), Mr. Jinlong Pan (潘金龍) and all the people we met in Sinshe for their help to make our field trip an enjoyable experience. We are also grateful to Professors Mingyi Wu (吳明義) and Shihhui Lin (林蒔慧) for their enthusiastic help to find Kavalan seniors in Sincheng Township for our language survey. Finally, we thank Ms. Fengyuan Yeh (葉鳳園) and Yuyuan Fang (方渝苑) in the Administration of Residents and Residence of Fengbin Township, Hualien for their assistance in providing us with valuable information we needed. The research reported here was supported by a research grant from National Science Council to the second author.

Kavalan (1,078) and Thao (619).[2] Their languages are long known to be at varying degrees of decline and thus potential extinction driven by the twin forces of rapid urbanization and Sinicization. Competent speakers of Kavalan, for example, are now estimated at less than just a few dozens.

The purpose of this paper is to provide a more in-depth approach to language documentation. In the words of Himmelmann (1998:165), "language documentation may be characterized as radically expanded text collection", and its purpose is to represent the language for both linguists and the uninitiated, who do not have access to the language itself. Although text collection may help preserve (some aspects of) a language itself, it cannot reveal the true picture of actual language use by a community of speakers in a particular spatio-temporal setting. As a language is in decline, one may want to know what the various forces are which may be contributing to its endangered status, how the language is losing ground in the home and the work domains. A sociolinguistic study, appropriately designed, may certainly help provide, at least in part, answers to these questions.

In this paper, we report our field experience of documenting Kavalan, a seriously endangered Formosan language spoken in southeastern Taiwan. In section 2 we give a brief introduction to the history of the Kavalan people. In the second part of the section we report on our recent field trip to Sinshe Village and present some of the data regarding the state of health of the Kavalan language in this village. In Section 3, we introduce the NTU (National Taiwan University) Corpus of Formosan Languages, which is built with an attempt to establish a standard for the creation of linguistic corpus databases through the application of information technology to linguistic research. As pointed out by Lehmann (2001:87), "an important specific purpose of language documentation is to serve as a record of the past and as an element of ethnic identity for future members of the community that has lost its identity as a speech community but which still recalls that their ancestors had a language of their own." The creation of this linguistic database thus enables both experts and common users to preserve valuable linguistic data and to provide a systematic recording of these languages.

2. CURRENT STATE OF THE KAVALAN LANGUAGE. In this section, we will first give a brief introduction to the history of the Kavalan people. Then, we will report on our recent sociolinguistic survey conducted in Sinshe Village and present some of the data regarding the state of health of the Kavalan language in this village.

2.1 A BRIEF HISTORY OF KAVALAN. For many centuries, the Kavalan people inhabited the present-day Ilan area in northeast Taiwan, which was known as *kap-a-lan* (蛤仔難or甲仔難), a transliteration of the word 'Kavalan'. These people called themselves Kavalan, meaning "people living in the plains", to distinguish themselves from the other aboriginal people in the mountain areas, e.g., the Atayals. Into this fertile land of tranquility toward the end of the eighteenth century came hordes of Han Chinese, which soon set in motion a series of arduous and probably sometimes heart-rending southward migration by the Kavalan people.

[2] According to the census of the Council of the Indigenous Peoples in Taiwan, the total population of the indigenous peoples in Taiwan in August 2007 is 481,119, which is less than 2% of the population of Taiwan. http://www.apc.gov.tw/chinese/docDetail/

The first migration took place during the period between 1830 and 1840. In 1796, the first group of the Han Chinese, led by Wu Sha (吳沙), moved in and opened up their first settlement in Toucheng (頭城).[3] As more and more Han Chinese followed into the area and took over, by force and by craft, the Kavalans' land, a number of sporadic migrations occurred, principally to Sanshing (三星) and Suao (蘇澳). Between 1830 and 1840, as they were losing their land and thus their socio-economic dependence, the Kavalan people, led by Kaliwan tribe living in Tongshan Township (冬山鄉), underwent a massive southward migration, and took up residence in Sincheng, Hualien (新城,花蓮), where their settlement was known as Kaliwan Village (加禮宛社).[4]

The Kavalan people did not stay there for long because of the Kaliwan Incident (加禮宛事) in 1878. There are two different versions of what triggered the Kaliwan Incident. One version holds that the Han Chinese, led by a businessman Wenli Chen (陳文禮), invaded and took over the land belonging to the Kaliwan Village. Another version holds that the Qing official Huihuang Chen (陳輝煌) swindled the Kaliwan people out of lots of money, and people in the Kaliwan Village, aided by Sakizaya, rose up to fight against the Qing soldiers. Many Kavalan people died in the battle. Fearing possible retaliation by the allied power of the aboriginal peoples, the Qing government forced the Kavalan people to move out of the area. This forced migration resulted in most of the remaining Kavalan people finally settling further south in Sinshe (Hsinshe, Xinshe, 新社) Village, a little village facing the Pacific Ocean, with Ocean Mountain Range at their back (See Picture 1). Others chose to settle in villages even further south along the Pacific coast, principally in Jangyuan, Taitung (台東樟原). Map 1 shows the two migration routes of the Kavalan people.

Sinshe Village--a sparsely populated village facing the Pacific Ocean on the southeast coast

[3] The Romanizations of the place names mentioned in this paper are conventional English translations adopted by each local government.

[4] Taiwan was at that time governed by the Qing Dynasty, but that was in name only. Many indigenous tribes were independent entities and did not come under the jurisdiction of the Qing Empire. For example, it was not until 1810 that a government office was set up to try to rule over the 36 Kavalan tribes.

The population of the Kavalan people has changed drastically over the last three hundred years. As shown in Table 1, the Kavalan population has over the centuries steadily declined in their homeland, Ilan, with just four Kavalans living there now, according to the official census of the Council of Indigenous Peoples, Executive Yuan.

TABLE 1. Population Change of the Kavalan in Ilan County

Year	1650 *	1852 *	1896*	1935 *	1969*	2007**
Population	9770	5507	2780	1544	app. 800	4

* Data taken from Huang & Chang (1995:2)

** Data taken from the Council of Indigenous Peoples, Executive Yuan (02/2007

MAP 1. Migration Routes of the Kavalan people

Sinicization and the eventual loss of identity have combined to produce a sharp population decline in the two other counties where the Kavalan now live, i.e., Hualien and Taitung. As shown in Table 2, there are now just 650 Kavalan people residing in these two counties, 568 in Hualien and 82 in Taitung. In other words, 87% of the Kavalan people now live in Hualien County, with most of them concentrated in Fengbin Township (豐濱鄉).

Table 2. Population Change of the Kavalan in Hualien and Taitung

Year	1897*	1966*	2007**
Population	About 1000	1289	650

* Data taken from Huang & Chang (1995:2)

** Data taken from the Council of Indigenous Peoples, Executive Yuan (02/2007)

There are five villages in Fengbin Township: Gangkou (港口村), Sinshe (新社村), Jingpu (靜浦村), Jici (磯崎村), and Fengbin (豐濱村). As shown in Table 3, Sinshe (新社村) and Fengbin (豐濱村) villages are the only two significant settlements for the Kavalan. However, even in these two villages, the Kavalan are vastly outnumbered by the Amis (14 % vs. 86%), as shown in Table 4, which accounts in part for the decline of the state of health of the Kavalan language.

Table 3. Kavalan Population in Fengbin Township (豐濱鄉) (04/2007)

Village	**Male**	**Female**	**Total**
Gangkou港口	7	12	**19**
Sinshe新社	105	86	**191**
Jingpu靜浦	0	0	**0**
Jici磯崎	6	5	**11**
Fengbin 豐濱	65	63	**128**
Total	**183**	**166**	**349**

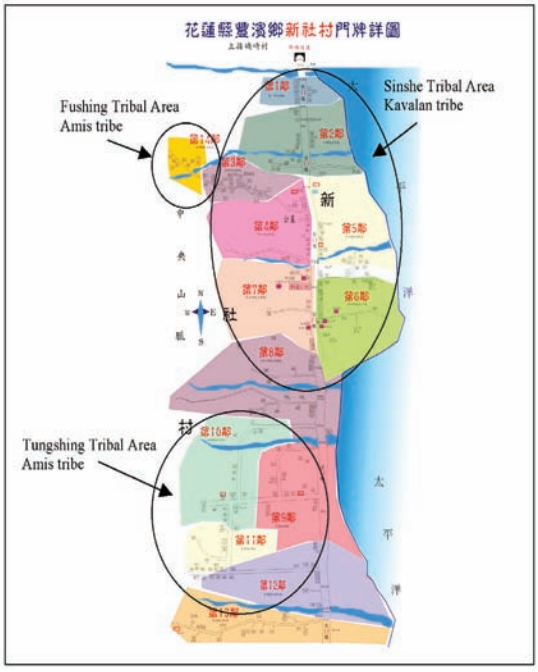
Table 4. Kavalan and Amis populations in Fengbin Township (04/2007)

Village \ Population	Male		Female		Total	
	Kavalan	Amis	Kavalan	Amis	Kavalan	Amis
Gangkou港口	7	437	12	334	19	771
Sinshe新社	105	249	86	211	**191**	**460**
Jingpu靜浦	0	442	0	296		738
Jici磯崎	6	98	5	110	11	208
Fengbin 豐濱	65	853	63	681	**128**	**1534**
Total	183	2079	166	1632	349	3711

2.2 THE STATE OF HEALTH OF THE KAVALAN LANGUAGE

2.2.1 BACKGROUND. An early, and in fact the only sociolinguistic study on the Kavalan language was conducted more than a decade ago (Huang and Chang 1995). To update on the current state of health of the language, we did a village-wide sociolinguistic survey in Sinshe Village in April, 2007.

As shown in Map 2, the Kavalan people spread over eight neighborhood areas (from the First Neighborhood to the Eighth Neighborhood), which are collectively called Sinshe Tribal Area (新社部落). The Fourteenth Neighborhood is the Fushing Tribal Area (復興部落), where all the residents are Amis speakers. Another Amis tribal area, called Tungshing Tribal Area (東興部落), covers the Ninth to Thirteenth Neighborhoods.



MAP 2. A Detailed Map of Sinshe Village
(taken from the website of the Administration of Residents and Residence of Fengbin Township, Hualien, at http://www.hl.gov.tw/977/upload/upload//index02.htm)

2.2.2 METHODOLOGY. As mentioned earlier, the Kavalan population has over the centuries steadily declined. In Sinshe Village there were just 191 people in the official household registration record, which means the actual number of residents should be far less, since many people work and live in the cities, but still retain their names in the household registration. The Village is characterized by a high proportion of inter-marriages and a low proportion of the younger generation. Consequently, the respondents in this survey were mostly senior citizens. The criterion we used to select our respondents was that they have at least one Kavalan parent. There were in this survey a total of 12 female and 11 male

respondents, with a mean age of 63 and an average age of 62.87; the oldest participant was 87 and the youngest 42. Nearly all of our respondents were multi-lingual in Kavalan, Amis, Mandarin, and Taiwanese, and a few even spoke a fifth language, Sakizaya.[5] All of them were married, and had children.

In this sociolinguistic survey, we focus our attention on the language ability of the Kavalan people in the Sinshe Village. Language ability in this study is defined as "being able to use the language to communicate with the family members or with the community members." In other words, we evaluate a speaker's language ability in terms of whether s/he can use the language to communicate rather than whether s/he can spell a word correctly. Therefore, the questions we asked our respondents were three. (1) "What language or languages do you use as a means of communication when you talk to your parents (last generation) and when you talk to your children (next generation)?" (2) "In what language or languages do your parents talk to you?" (3) "In what language or languages do your children talk to you?"

We interviewed our respondents in their homes or at a place where they usually gather for social occasions, such as churches, grocery stores or weaving classroom. At each interview session, we first explained our intention to our respondents that we would like to know how and when they used Kavalan to communicate with other villagers. Then, we recorded each respondent's bio-data, i.e., age, marital status, number of children, the tribes that his/her parents and grandparents belong to, and the tribes that his/her spouse belongs to. Next, we asked the three questions listed above to determine their language ability. For each language the respondent mentioned, we marked one point in that language column. Since all the respondents in this study were multilingual, their answers may not be limited to a single language.

2.2.3 RESULTS AND DISCUSSIONS. We list our findings in Table 5 below. In Table 5, the abbreviations K, A, T, M, and O stand for the Kavalan, Amis, Taiwanese, Mandarin and other languages, such as Sakizaya, Japanese or Hakka, respectively.

As shown in Table 5, all the respondents, i.e., 23/23, used Kavalan at home. 78.26% of their parents (i.e., 36 of the 46) used Kavalan, but 86.96% of their parents (i.e., 40/46) spoke Amis at home; since some of their parents were Amis. 73.91% of their children now in their 30s-40s also used Kavalan without difficulty. All of the respondents' next generation also spoke Mandarin. Note the comparatively lower percentages among the respondents and their last generation on this measure.

In the home domain, all our senior respondents reported that they spoke Kavalan to the family members. However, if they talked to the younger family members, those below 20, they talked in Kavalan and the younger family members replied in Mandarin or Taiwanese.

[5] Although Kavalan is still actively used in the village, outside the home domains the Kavalans speak different languages to different co-participants, as expected. When they go grocery shopping in a Taiwanese's store, they speak Taiwanese. When they meet their Amis neighbors, they chat in Amis. Although Kavalan and Amis live in different parts of the Village, as shown in Map 2 above, they go to the same churches, Catholic or Christian, there being only one each in Sinshe. In the churches, especially the Catholic, the Kavalans tend to switch to Amis, since the priest is an Amis, and the Bible is in Amis. At present there is no Kavalan Bible. All these suggest that Kavalan is in some sense a minority language even in Sinshe Village.

TABLE 5. Language Ability of the Kavalan people in Sinshe Village

	Last Generation					**Respondents**					**Next Generation**				
Language	K	A	T	M	O	K	A	T	M	O	K	A	T	M	O
	36/ 46	40/ 46	23/ 46	16/ 46	23/ 46	23/ 23	21/ 23	14/ 23	16/ 23	9/ 23	17/ 23	9/ 23	14/ 23	23/ 23	5/ 23
Percentage	78.26%	86.96%	50%	34.78%	50%	100%	91.3%	60.87%	69.57%	39.13%	73.91%	39.13%	60.87%	100%	21.74%

K: Kavalan; A: Amis; T: Taiwanese; M: Mandarin; O: Others (esp. Sakizaya, Japanese, and Hakka)

TABLE 6. Language Abilities of the Kavalan (comparative)

	Last Generation					**Respondents**					**Next Generation**				
Languageg.	K	A	T	M	O	K	A	T	M	O	K	A	T	M	O
51 ↑ *	24/ 32	27/ 32	12/ 32	8/ 32	11/ 32	16/16	15/ 16	8/ 16	9/ 16	8/ 16	**15/ 16**	8/ 16	13/ 16	16/16	5/ 16
	75%	84.38%	37.5%	25%	34.38%	100%	93.75%	50%	56.25%	50%	**93.75%**	50%	81.25%	100%	31.25%
50 ↓ **	12/14	13/14	11/14	8/14	12/14	7/7	6/7	6/7	7/7	1/7	**2/7**	1/7	1/7	7/7	0/7
	85.71%	92.86%	78.57%	57.12%	85.71%	100%	85.71%	85.71%	100%	14.29%	**28.57%**	14.29%	14.29%	100%	0%

*Respondents who are older than 51 years old.
** Respondents who are younger than 50 years old.

All of the respondents reported that almost all the children younger than 20 could hardly understand Kavalan, let alone speak fluent Kavalan. Consequently, the seniors often had to switch to Mandarin or Taiwanese to help the process along. The situation is predictably considerably worse for families residing in cities where environmental support for the use of Kavalan is non-existent (Huang and Chang 1995).[6]

We tried to find some young respondents in their 20s in Sinshe, but to no avail. People in their 20s-40s tend to live and work in the cities, since there are simply no jobs available in the village. Those still staying in the village are those older than 50, or those younger than 15, youngsters who are still in school.

There were a total of seven respondents younger than 50 in our survey, which means their children might be in their 20s. We thus separated these respondents from the others, and we arrived at Table 6.

As shown in Table 6, almost all of the children, about 94%, of the respondents older than 51 speak Kavalan, while only less than 30% of the children of the interviewees younger than 50 do. Given the limitedness of our sample size, the true picture of the rate of language transmission among these children is probably somewhere between these two extremes.

At this point, it may be of some interest to compare our findings with those in Huang and Chang's (1995), as shown in Table 7.

TABLE 7. Huang and Chang's Survey (1995:8)

		Last Generation			**Respondents**			**Next Generation**		
	Language	K	A	T	K	A	T	K	A	T
Sinshe		23/28	16/28	23/28	13/14	13/14	12/14	**12/14**	6/14	14/14
	Percentage	82.1%	57.1%	82.1%	93%	93%	85.7%	**85.7%**	46.1%	100%
Taipei		64/64	63/64	54/64	31/32	28/32	32/32	**3/27**	2/27	21/27
	Percentage	100%	98.4%	84.4%	96.8%	87.5%	100%	**11.1%**	7.4%	77.7%

In their study, Huang and Chang conducted two surveys, one in Sinshe and one in Taipei.[7] The average age of Huang and Chang's (ibid: 5) Taipei respondents was 36.5, and that of their Sinshe respondents was 58.5; therefore, the average age of the next generation of the Taipei respondents was estimated to be 17 or younger, and that of the Sinshe respondents was about 30 or older. Huang and Chang's (ibid) survey shows that the language transmission rate in Sinshe stood at 86%, while that in Taipei was a low of 12%. Their finding is quite similar to ours, since our survey shows that the older respondents were able to transmit their mother language to their next generation (about 94%), while the younger respondents can barely do so (only about 29%).

Nonetheless, there are two exceptional cases in our survey. One case involves two

[6] The environmental support here means the family members, friends, the community or even the school and the work places.

[7] Most of the Kavalan migrating into the Taipei area reside in Banqiao (板橋) and Shulin (樹林) areas.

little children, a four-year-old boy and his 2-year-old younger brother, who live with their Kavalan-speaking grandmother in Sinshe, and are acquiring the language as their first language, as their parents, the Kavalan father married to a Vietnamese bride, have to work in Taipei and do not have time to take care of them. In the other case, two teens are living with their parents and can speak Kavalan because their father, one of our younger respondents, 42, deliberately "creates the home environment" for them; that is, their 'language crisis-conscious' father asks them to speak Kavalan in the home.

As suggested above, another settlement area for the Kavalan people is Jiali Village of Sincheng Township in Hualien County (花蓮縣新城鄉), the midway homeland to many of the ancestors of those now living in Sinshe. As shown in Table 8, there are now only 105 Kavalans living in the town of Sincheng, 65 of whom live in Jiali Village (嘉里村), where the Kaliwan Incident took place and was once known as Kaliwan Village.

TABLE 8. The Kavalan Population in Sincheng Township, Hualien (03/2007)

Tribe / Village	**Amis**	**Kavalan**
Dahan (大漢村)	983	5
Beipu (北埔村)	772	15
Jialin (佳林村)	146	2
Kangle (康樂村)	398	1
Shunan (順安村)	67	0
Sincheng (新城村)	81	3
Jiali (嘉里村)	492	**65**
Jiasin (嘉新村)	336	14
Total	**3275**	**105**

We were able to locate and interview two Kavalan seniors in the Village. One gentleman was Tuyaw (李抵搖), 88, who is married to a Sakizaya from Yuli (玉里) and speaks Kavalan, Sakizaya, Japanese, Amis and Taiwanese. Tuyaw remembered that his grandfather came from Sanshing, Ilan (宜蘭三星), and that his grandparents and parents could speak fluent Kavalan, Amis and Sakizaya. He could not recall the approximate time period when everyone in the neighborhood started to discontinue the use of Kavalan altogether. The most common languages he uses in his daily routines are Taiwanese, Amis, and, occasionally, Japanese. He hardly ever has any opportunity to speak Kavalan now, since no one speaks the language in the neighborhood. None of his seven children speaks either Sakizaya or Amis, much less Kavalan, the minority among the minority languages in the town.

Another respondent was the Kavalan wife of a former *Tapang no yaro* (chief of the Local Amis, 阿美族頭目) Jiang Jia-zou (江加走). Mrs. Jiang, 71, speaks fluent Taiwanese like a native Taiwanese speaker, and does not keep a Kavalan name. She spoke Kavalan while she was little, but, like Tuyaw, has hardly ever used it since then, since her neigh-

bors, relatives, or even her four children do not speak the language. While Kavalan is thriving forty miles down south in Sinshe, it is reasonable to suggest that Kavalan was already languishing in Sincheng when Mrs. Jiang was young and soon became extinct in the neighborhood as it lost its status even as a home language or a language of local social communication.

Huang and Chang (1995:7) explained that the rapid decline and eventual loss of the Kavalan language in areas outside of Sinshe Village is triggered by a combination of factors, namely, a high percentage of intermarriages, the 'destabilizing effect' of the more prestigious languages, i.e., Mandarin, Taiwanese and Amis, used in the wider communities, and, concomitantly, the relatively low degree of self-identity. These elements are still exerting their powerful forces. To be sure, indigenous issues have had a high profile over the past decade and the government has encouraged their rights to self-identity and the resumption of indigenous names and a number of policy changes in the direction of greater respect for the indigenous languages and heritage. 'The indigenous language nest program' has also been in effect for a number of years now--- these changes mean removal of social and political forces that have helped shape the language inequality in the first place and may yet provide structural conditions for the preservation, or at least slow down the process of decline, of indigenous languages.

2.3 PROPOSED STRATEGIES FOR REVITALIZING THE KAVALAN LANGUAGE. As part of a new educational policy, indigenous languages have been taught in the elementary schools since September 2001. Although these language lessons typically run only 1 or 2 hours a week, they are a step forward toward preserving the indigenous languages (cf. Lillian Huang 2007). Nonetheless, the current state of the Kavalan language teaching in the Village does not bode well for Kavalan. In Sinshe Village, there is only one elementary school, and it teaches both Amis and Kavalan. The students taking Amis far outnumber those taking Kavalan, since the Amis population is the majority in the village. The teacher teaching Amis is an Amis, who is also a regular member of the school faculty, while the Kavalan teachers are not. When the students go on to junior high school, in the neighboring village, only Amis but not Kavalan is taught. Furthermore, there is no other way for these Kavalan children to continue their education in Kavalan, except from their families.

Although the Council of Indigenous Peoples (CIP) has launched a six-year program since 2003 to help revitalize indigenous languages in both rural and urban areas (cf. Lillian Huang 2007), school or the community leaders in Sinshe Village have failed to propose any language revitalizing programs with which to apply for grants from the CIP, thus forestalling Kavalan language revitalizing efforts in the village, since without government grants, they could not set up community-based language classrooms to train language teachers.

Concerned linguists and community leaders must join hands to help come up with language revitalizing programs and apply for funding from the CIP (i) to set up public Kavalan classrooms, and (ii) to train more Kavalan teachers committed to Kavalan language teaching. To be sure, if a language is not spoken in the home, classroom teaching might seem to be a superficial and cosmetic measure, but it can have other positive functions. Many Kavalan parents expect their children to get good grades in school in order to go to a better high school and college, and, eventually, get a better job. Classroom teaching can

thus make learning Kavalan matter in this regard. Moreover, classroom teaching is also a direct way for the Kavalan children to know about their own language, culture and history. Armed with this knowledge they can then go on to document their legends, folklores and songs and even work on their own bible. Although there are now a number of studies related to the Kavalan languages (Li 2007), including a recently published Kavalan Dictionary (Li and Tsuchida 2006), no Kavalan version of the Bible is yet available, and so the Kavalan churchgoers in Sinshe have to use the Amis version, to the detriment of the Kavalan language in the community. Compiling a Kavalan version of the Bible is sorely needed.

In the next section, we will introduce the NTU Corpus of Formosan Languages, which is built with an attempt to document some of the most seriously endangered Formosan Languages.

3. NTU CORPUS OF FORMOSAN LANGUAGES. Digital Archiving of the Yami Language at Providence University (Rau and Yang 2007; Rau, Yang and Dong 2007) stands as the first attempt to provide public access to the language. In addition, Academia Sinica's Formosan Language Archive (Zeitoun et al. 2003; Zeitoun and Yu 2005) appears to be a large-scale digital archive with an attempt to document all the Formosan languages, including their dialects.[8]

The NTU (National Taiwan University) Corpus of Formosan Languages demonstrates our attempt not only to document some seriously endangered Formosan languages, but to provide further public and user-friendly access to both specialists and non-specialists.[9] There are two special features of the NTU Corpus, i.e., dictionary and search, which we discuss in detail below. Moreover, the NTU Corpus of Formosan Languages is the first corpus to document spoken data in terms of prosodic units, i.e., the Intonation Unit (IU), which is defined as a stretch of discourse falling under a single coherent intonation contour (Chafe 1987, 1993, 1994; Du Bois et al. 1993; Schuetze-Coburn 1993; Schuetze-Coburn et al. 1991; Tao 1993). Although natural spoken language is often found to have a high proportion of pauses, hesitations, fillers, repetitions, and false starts, they are important for us to learn more about the pragmatic and cultural aspects of a spoken language. A detailed study of fragments of conversation reveals that fillers, repairs and repetitions are important interactional strategies used by the speech participants to hold the floor, to plan for language production, to do lexical searching, and so on. Fillers, repairs and repetitions are thus essential in that they enable the conversation to go on without much difficulty by sending out signals of the speaker's next move and intention (Schegloff 1980, 1988, 1991, 1996; Sack, Schegloff and Jefferson 1974; Huang 1993, 1999).

3.1 AN OVERVIEW OF NTU CORPUS OF FORMOSAN LANGUAGES. The main purpose of the NTU Corpus of Formosan Languages is to document some of the endangered languages spoken in Taiwan, such as Kavalan, Saisiyat and Tsou. It was part of the projects of the Multimedia Laboratory operated by the Center for Information and

[8] At present, Academia Sinica's digital archiving is still under construction; many language corpora are listed but contain no text.

[9] See Sung et al. (submitted) for a detailed description of the NTU Corpus of Formosan Languages.

Electronics Technologies at National Taiwan University,[10] with an attempt to establish a standard for the creation of linguistic databases through the application of information technology to linguistics research.

The corpus contains face-to-face conversations and narratives, including both natural and elicited narratives for cross-linguistic research. The materials for elicited narratives are based on the Pear Story (Chafe 1980) and Frog Story (Mayer 1980).[11] The NTU corpus is composed of spoken texts in Saisiyat, Kavalan, Amis and Tsou.[12] At present, a small corpus of Kavalan with just four narrative texts has been placed online, which runs to about ten minutes, for a total of 228 IUs, though we have collected and transcribed a total of 21 Kavalan texts, including 15 narratives and 6 conversations, which together run to about 136 minutes. The narrative texts are 11 elicited narratives, including four pear stories and seven frog stories, and 4 other narratives. The conversations cover six face-to-face conversations between acquaintances or relatives, which together run to about 61 minutes. All these texts will be put on line in the near future.

3.2 TRANSCRIBING SPOKEN DATA. The process of transcribing spoken data is both tedious and time-consuming.[13] We first sound-record or/and video-tape spoken texts. After a spoken text is collected, our graduate assistants then help transcribe the raw data, and tag and annotate the transcribed texts according to a prepared coding list. Then other assistants double check the annotated text.

Since our corpus data are natural spoken narratives and conversations, in order to reflect and record the discourse information, such as pauses, false starts, repairs, and so on, the transcription needs to meet not only grammatical but also discourse coding standards. Our transcription of the discourse information mostly follows Du Bois et al. (1993), a *de facto* standard in the linguistic community. After the tagging, annotating and double-checking, the transcribed texts are submitted to the corpus programmer before they are finally put on line. The whole transcription process can be diagrammed as in Figure 1.

The texts in our system are stored in Unicode (UTF-8 encoding); the advantage of using such an encoding form is that it is easy to incorporate other languages into our an-

---

[10] They are Graduate Institute of Linguistics, Department of Information Management, Department of Library and Information Science, Department of Computer Science and Information Engineering, Department of Electrical Engineering, Department of Journalism, and Department of Drama and Theater) and colleges (College of Electrical Engineering and Computer Science, College of Liberal Arts, College of Social Sciences, and College of Management)

[11] Frog stories are elicited narratives: our informants are asked to tell a story while looking at the pictorial book *Frog, Where Are You?* (Mayer 1969, reprinted in 1980). This pictorial wordless book tells the adventure story of a little boy and his dog searching for his frog that got out into the woods.

[12] The corpus of Saisiyat has 22 texts, including three conversations, five Pear stories, eight Frog stories, four Saisiyat legends and two daily life narratives, which together run to 118 minutes, for a total of 3437 IUs. The Corpora of Amis and Tsou each contain two narratives, one Frog story and one Pear story, which run to five minutes with a total of 138 IUs and eight minutes with a total of 237 IUs, respectively.

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

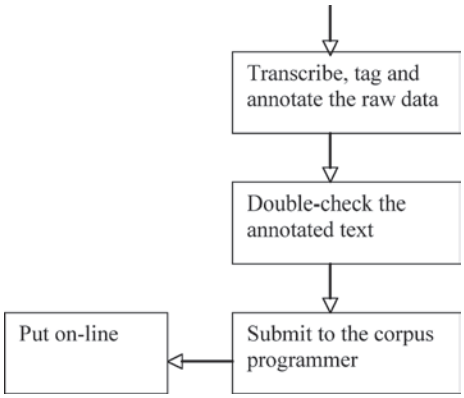


FIGURE 1. Procedure of data transcription and submission

notation system which adopt other writing or phonetic systems, such as IPA. If some of the tribes decide to adopt non-ASCII letters, such as "ê Â ú Ò Ä", into their writing systems, our programs can process them correctly with no need of modification or transformation.

3.3 ACCESSIBILITY AND SPECIAL FEATURES. There are two special features of the NTU Corpus, i.e., dictionary and search, which we will illustrate with examples below. One special feature is that our system can automatically generate an online dictionary, with the information of the total number of word tokens. The count of tokens is updated as new texts are uploaded. As shown in Figure 2, when choosing a language and entering the corpus, the user can find a "Dump dictionary" function on the top of the list of the texts. The dictionary can be printed out at a marginal cost, and it can also be cut and pasted for any linguistic analysis.

Another special feature of our corpus is that it allows users to search for any specific target word or morpheme in English, Chinese or any Formosan languages. For example, if a user wants to know how to say the English word *know* in these Formosan languages, he may type in the English word 'know' and then he may find that in Kavalan, the equivalent word is *supaR*, while in Saisiyat, there are two words equivalent to the English word *know*, *sekela* and *ra:am*. In Amis, the equivalent is *ma-fana'*; and in Tsou, it is *cohivi*.

In the Search function page, one can also search for a particular lexeme in a corpus; for example, one may want to search for the distribution of the lexeme *Rayngu* 'not know; not able to' in Kavalan. He may select the language Kavalan, and type in the key word, *Rayngu*; then, he can find all the related data, as shown in Figure 3.

FIGURE 2. The function [Dump dictionary] in the Kavalan Corpus.

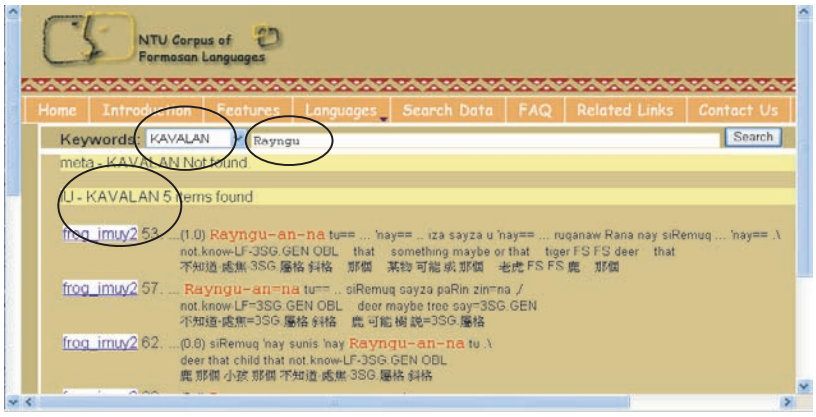


FIGURE 3. Pop-up search result of the lexeme *Rayngu* in the Kavalan corpus

Moreover, users can search for a particular lexeme across the languages in the corpora, e.g., *ma-* in Kavalan, Amis, and Saisiyat. Since each text and each Formosa language in our corpus is stored in a cross-related file with the same normalized tables, cross-text and cross-language search can be executed in a single command (cf. Sung et al., submitted). The number of tokens of the searched lexeme is shown at the same time. For example, if a user wants to investigate the syntactic behaviour of the marker *ma-* in Formosan languages (say, Kavalan, Saisiyat, Amis and so on), he/she can type in the key morpheme *ma-* in the search page, wait for a second and then all the related data in different texts and (Formosan) languages will come out, as shown below in Figure 4.

FIGURE 4. Pop-up search results of the prefix *ma-* (Page 1 of 31 pages)

With the rapid progress in internet technology and the processing techniques of natural linguistic database, the creation of a language database has become a most effective means of recording and preserving precious linguistic data. The NTU Corpus is structured in a way that enables any user, linguist or not, who is interested in Austronesian languages and culture, to gain access to the rich and valuable linguistic data available through a diverse array of format in the most convenient means.

5. CONCLUSION. In the preceding sections we have provided a fairly detailed survey of the state of health of the Kavalan language. We have also looked at the workings of the NTU Corpus of Formosan Languages, which is built with an attempt to document some of the endangered languages spoken in Taiwan, languages that are Taiwan's gifts to the world (Diamond 2000). Documenting Kavalan as well as other Formosan languages is an ongoing project for us. Budgetary constraints and shortage of staff have meant that we have not been able to proceed at a pace we would have liked it to be. When we are done, however, we hope the NTU Corpus of Formosan Languages will ultimately prove to be a valuable research tool to the academic community.

References

Brenzinger, Matthias, ed. 1992. Language death: Factual and theoretical explorations with special reference to East Africa. Berlin: Mouton de Gruyter.

Chafe, Wallace L. 1980. The deployment of consciousness in the production of a narrative. In The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, ed. by Wallace Chafe, 9-50. Norwood, N.J.: Ablex Publishing Corp.

Chafe, Wallace L. 1987. Cognitive constrains on information flow. In Coherence and grounding in discourse, ed. by Russell S. Tomlin, 21-51. Norwood, N.J.: Ablex.

Chafe, Wallace L. 1993. Prosodic and functional units of language. In Talking data: Transcription and coding in discourse research, ed. by Jane A. Edwards and Martin D. Lampert, 33-43. Hillsdale, N.J.: L. Erlbaum.

Chafe, Wallace L. 1994. Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing. Chicago: University of Chicago Press.

Crawford, James. 1995. Endangered native American languages: What is to be done, and why? The Bilingual Research Journal 19(1):17-38.

Diamond, Jared M. 2000. Linguistics: Taiwan's gift to the world. Nature 403:709-10.

Du Bois, J. W, Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. In Talking data: Transcription and coding for language research, ed. by Jane A. Edwards and Martin D. Lampert, 45-90. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. Linguistics 36:161-95.

Huang, Lillian M. 2007. Strategies in revitalizing indigenous languages in Taiwan. Pre-Conference Proceedings of the International Conference on Austronesian Endangered Language Documentation, 25-45. Taichung: Providence University.

Huang, Shuanfan. 1993. Pause as a window on the mind and the grammar—evidence from spoken Chinese discourse. Paper presented at the Workshop on Interfaces and the Chinese Language, June 30-August 6, 1993, in Ohio State University.

Huang, Shuanfan. 1999. The emergence of a grammatical category definite article in spoken Chinese. Journal of Pragmatics 31(1):77-94.

Huang, Shuanfan, and H. 1995. Kavalan: A sociolinguistic study. Ilan Wenxian (Ilan Journal of History) 14:1-13. (In Chinese)

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

Lehmann, Christian. 2001. Language documentation: A program. In Aspects of typology and universals, ed. by Walter Bisang, 83-97. Berlin: Akademie Verlag.

Li, Paul Jen-kuei. 2007. Documentation of the most endangered Formosan languages. Pre-Conference Proceedings of the International Conference on Austronesian Endangered Language Documentation, 1-12. Taichung: Providence University.

Li, Paul Jen-kuei, and Shigeru Tsuchida. 2006. Kavalan Dictionary. Taipei: Academia Sinica.

Mayer, Mercer. 1980. Frog, where are you? N.Y.: Dial Books.

Rau, D. Victoria, and Meng-Chien Yang. 2007. e-Learning in endangered language documentation and revitalization. Pre-Conference Proceedings of the International Conference on Austronesian Endangered Language Documentation, 101-22. Taichung: Providence University.

Rau, D. Victoria, Meng-Chien Yang, and Maa-Neu Dong. 2007. Endangered language documentation and transmission. Journal of National Council of Less Commonly Taught Languages (NCOLCTL). University of Wisconsin at Madison. 53-76.

Robins, Robert. H., and Eugenius M. Uhlenbeck, eds. 1991. Endangered languages. Oxford: Berg.
Sack, Harvey, Emanuel A. Schegloff and Gail Jefferson. 1974. A simplest systematics 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 50:104-51.
Schegloff, Emanuel A. 1988. Pre-sequences and indirection. Journal of Pragmatics 12:55-62.
Schegloff, Emanuel A. 1991. Conversation analysis and socially share cognition. In Perspectives on socially shared cognition, ed. by Lauren B. Resnick, John M. Levine and Stephanie D. Teasley, 150-71. Washington D.C.: American Psychological Association.
Schegloff, Emanuel A. 1996. Turn organization. In Interaction and grammar, ed. by Elinor Ochs, Emanuel A. Schegloff and Sandra A. Thompson, 52-133. Cambridge University Press.
Schmidt, Annette. 1990. The loss of Australia's aboriginal language heritage. Canberra: Aboriginal Studies Press.
Schuetze-Coburn, Stephan. 1993. Prosody, grammar, and discourse pragmatics: Organizational principles of information flow in German conversational narratives. PhD diss., University of California at Los Angeles.
Schuetze-Coburn, Stephan, Marian Shapley and Elizabeth Weber. 1991. Units of intonation in discourse: A comparison of acoustic and auditory analyses. Language and Speech 34(2):207-34.
Sung, Li-May, Lily I-wen Su, Fuhui Hsieh and Zhemin Lin. Submitted. Design of a multimedia corpus of Formosan languages.
Tao, Hongyin. 1996. Units in Mandarin conversation: Prosody, discourse, and grammar. Amsterdam/Philadelphia: John Benjamins Publishing Company.
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):218-32.
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

Fuhui Hsieh
hsiehfh@ttu.edu.tw

Shuanfan Huang
sfhuang@ntu.edu.tw