CHAPTER 4

SOCIOLINGUISTIC SURVEY DESIGN

This chapter describes the design of the sociolinguistic survey undertaken in February and March 2006. The survey was designed by the author. The sociolinguistic questionnaires used in the survey were pilot tested in January 2006. The fieldwork was carried out by a team of four comprising two survey specialists, a Shan interpreter and a general assistant.

4.1 Goals of Survey and Research Questions

The survey was designed by first identifying four broad goals for the survey and then, within each goal, specific research questions were formulated. The survey instruments were then designed to collect sufficient information to provide at least a partial answer to each research question. In considering the details of this particular survey it is good to bear in mind the broader context, namely the ongoing assessment of the Khuen language development programme.

The following sections list the main goals and the research questions related to each goal.

4.1.1 Goal 1: Assess the need for literacy in Khuen.

The research questions listed below examine the potential of Khuen speakers to use one or more of the Languages of Wider Communication in use the region, namely Shan, Burmese or Lue.

Research Question 1.1: Do Khuen speakers master Shan adequately?

Research Question 1.2: Do Khuen speakers have any negative attitudes that would inhibit their use of Shan for written or oral communication?

Research Question 1.3: What is the current literacy rate in Shan among Khuen speakers?

Research Question 1.4: Do Khuen speakers master Burmese adequately?

Research Question 1.5: Do Khuen speakers have any negative attitudes that would inhibit their use of Burmese for written or oral communication?

Research Question 1.6: What is the current literacy rate in Burmese among Khuen speakers?

Research Question 1.7: Do Khuen speakers master Lue adequately?

Research Question 1.8: Do Khuen speakers have any negative attitudes that would inhibit their use of Lue for written or oral communication?

Research Question 1.9: What is the current literacy rate in Lue among Khuen speakers?

4.1.2 Goal 2: Assess the potential demand for literacy in Khuen.

Language vitality is central to any consideration of the likely interest in Khuen literacy as are attitudes to Khuen culture and literacy in Khuen. These issues are probed in the following research questions.

Research Question 2.1: Does it appear likely that Khuen will continue to be spoken by future generations?

Research Question 2.2: Do Khuen speakers have positive attitudes towards Khuen culture?

Research Question 2.3: Do Khuen speakers have positive attitudes towards Khuen literacy?

Research Question 2.4: What is the current literacy rate in Khuen among Khuen speakers?

4.1.3 Goal 3: Determine which Khuen variety is the most suitable for use as a written standard.

Both linguistic issues such as phonology and the lexicon and sociolinguistic issues such as the prestige of particular varieties are significant in evaluating the use of a particular variety as a written standard. Such considerations are used to formulate the research questions listed below.

Research Question 3.1: What are the groupings of Khuen varieties based on lexical similarity?

Research Question 3.2: What are the groupings of Khuen varieties based on phonological criteria?

Research Question 3.3: What groupings of Khuen varieties are perceived by Khuen speakers?

Research Question 3.4: Which Khuen variety has most prestige in the eyes of Khuen speakers?

4.1.4 Goal 4: Determine which Tai languages are most closely related to Khuen.

There are at least two reasons why it is of interest to study the relationships Khuen fits in the Southwestern branch of the Tai language family. Firstly, in seeking to increase the body of literature available in Khuen, the potential to use as source texts for translation into Khuen materials from a closely related language is a great benefit. The closer the relationship between Khuen and that other language, the easier it should be to use such materials. Secondly, to the author's knowledge no previous study has compared Khuen with Lue Murng Yorhng.

Research Question 4.1: Which Tai languages are most closely related to Khuen based on lexical similarity?

Research Question 4.2: Which Tai languages are most closely related to Khuen based on phonological features?

The following sections describe the methodology used to conduct the survey. First the three survey instruments – which were designed specifically for this survey – are described in turn. Section 4.2 describes the Khuen 406 wordlist. The Individual Sociolinguistic Questionnaire (ISQ) and the Knowledgeable Insider Questionnaire (KIQ) are described respectively in Sections 4.3 and 4.4. Sections 4.5 and 4.6 describe respectively the method of selecting the survey sites and the subjects at each site. A summary of the total body of data collected from all survey sites is presented in Section 4.7 and the methods of analysing the information gathered are described in Section 4.8.

4.2 Khuen 406 Wordlist

The design of the Khuen 406 wordlist is described in outline in this section. Further details of the a priori evaluation of the wordlist are given in Appendix 2. The full wordlist itself is given in Appendix 3.

4.2.1 Wordlist design

The Khuen 406 wordlist has two main purposes, namely to facilitate cross-linguistic comparison as well as to give clear evidence of all the significant speech sounds in Khuen. The list was compiled in two parts; the first part consisted of the 200 highest-ranked words on Mann's (2004) wordlist which provide a standard corpus of culturally relevant words that facilitate cross-linguistic comparison. The second part consisted of 206 words chosen[26] to provide a strong basis for the analysis of the phonology of Khuen varieties. (The original plan was for a 400-item list: the extra six items were included to allow for the strong possibility that some items would need to be discarded because of some shortcoming in the

[26] The wordlists described in Section 3.1.1 as well as previous publications about Khuen described in Section 2.3 were used as references in selecting words for the list.

translation into the language of elicitation and/or the elicitation process itself.) Since time is normally very limited in survey fieldwork, ease of elicitation was a significant factor in the choice of words included in the list. The items in the second part were chosen incrementally in the light of the phonetic features of the words in the first part. The aim was to achieve a balanced coverage of all the significant phonological features of Khuen. These features were largely those identified by Woranut (1978) and listed in Table 11. In addition, since Tai Lue is generally held to be closely related to Khuen and in some areas there is contact between the two languages, what Li refers to as the 'nasal umlaut' feature is also included (Li 1964:12). The final wordlist is evaluated from the point of view of each of the features in Appendix 2.

4.2.2 Grouping and ordering the items on the wordlist

Having selected the words to be included in the list, the ordering of the words must be determined. The principle used in this case was to group items according to the semantic universe of the LRPs. This general principle follows Milliken and Milliken (n.d.) and the SIL Mainland Southeast Asia Group (n.d.). After initially grouping the words into broad categories, the words were carefully ordered within the groups in an attempt to maximise the associations between words close to each other in the list. This grouping and ordering was refined with the help of a survey specialist with much experience in eliciting wordlists in similar cultural situations. The broad categories are labelled as follows:

1. The world around
2. Plants and food
3. Animals
4. The human body
5. People
6. Around the home
7. Individual human activities
8. Human interaction

9. Numbers
10. Dimensions
11. Appearance

One important effect of ordering along semantic lines is that grammatical categories are mixed. For example in category 3 ‘Animals’ the noun ‘dog’ (item 108) is immediately followed by the verbs ‘lick’, ‘bite’ and ‘hurt’. The categories themselves are ordered so that more concrete concepts occur at the beginning and more abstract concepts towards the end but given this constraint, there is much scope for rearrangement should the present ordering cause problems in practice.

4.2.3 Elicitation

Having chosen the individual lexical items to be included in the list, it is important to develop a robust method of elicitation. For simplicity the words are initially elicited on their own so as not to distract from obtaining the word whose meaning most closely matches the item in the list. This word is transcribed in the field situation. When all of the words have been elicited, they are recorded not on their own but in an elicitation frame. Such frames are especially important for tonal languages but also languages where vowel length is also a research issue (Pike 1948). Care is needed when developing the frame because of the possibility of tone sandhi, that is, the possibility that the tones of the words of the elicitation frame might change the tone of the target word. Moreover, even if there is no sandhi caused by a particular frame in the language of elicitation, it is still possible that there might be problems in the target language. Bearing this in mind, two frames were developed based on previous knowledge of Khuen – one for nouns and one for verbs and adjectival predicates. These are given in (1) and (2). By having both the field transcription and the recording in the frames it is possible to check for consistency – it obviously inspires more confidence if the two agree.

(1) Noun frame

ʔan^{33} *niː*^{32} _____ *ʔɤː*^{32}

CLF this PRT

'This is a _____'

(2) Verb frame

man^{33} _____ *ʔɤː*^{32}

3s PRT

'He/she/it is _____-ing'

Because the wordlist has verbs, nouns and adjectives mixed together the particular elicitation frame used often needs to alternate between (1) and (2). For convenience the wordlist has a column specifically for the tone frame and indicates each place where a change of frame is needed. The full wordlist is given in Appendix 3.

The following two sections describe two types of questionnaire used on the survey. Both were designed using the Research And Instrument Design Tool (RAID) (Nahhas et al. n. d.). The main advantages of this were the systematic linking of the individual questions on the questionnaires to the research questions and the fact that most of the questions in the RAID data bank had already been tested in field situations elsewhere.

4.3 Individual Sociolinguistic Questionnaire (ISQ)

The purpose of this questionnaire is to gather information relating to the research questions under investigation in the survey. Related questions are grouped together to make for a logical progression of questions. The sections are described in Table 18 below and the full questionnaire is given in Appendix 4. The Shan translation of the ISQ is given in Appendix 5.

Section Title	Summary of Information Elicited
Subject Demographics	information about the subject and his/her family
Domains of Language Use	information about which language(s) the subject uses in various situations
Ethnolinguistic Identity	information about distinctive Khuen customs and the subject's attitudes to them
Bilingual Proficiency Evaluation	information about the subject's self-assessment of his/her ability to speak and/or comprehend languages of wider communication in use in the region
Language Attitudes	information about the subject's attitudes to other languages and speakers of those languages
Dialect Perceptions	information about how the subject perceives any groupings of Khuen varieties and whether there is any variety that carries more prestige than other varieties

Table 18 Description of Sections on Individual Sociolinguistic Questionnaire

The sections do not match the research questions exactly and so answers to some questions will be relevant to more than one research question. For example, questions in both the 'Bilingual Proficiency'; 'Language Attitudes'; and 'Domains of Use' sections are relevant to Research Question 2.1 concerning Khuen language vitality.

4.4 Knowledgeable Insider Questionnaire (KIQ)

The purpose of this questionnaire is to collect information about the community as a whole. It is expected that this questionnaire will only need to be administered once in each location. The village leader is usually the person chosen to answer the questions on this questionnaire. The questions are grouped into sections which are described in Table 19. The full questionnaire (in English) is given in Appendix 6 and the Shan version in Appendix 7.

Section Title	Summary of Information Elicited
Subject Demographics	information about the subject and his/her family
Village Name, Population	information about how the village is referred to by the inhabitants themselves and outsiders as well as the ethnic mix of households in the village
Languages and Ethnic Groups	information about how outsiders refer to the language used/ inhabitants of the village and the ethnic profile of the village
Village History	information about migration (if any) of the village community as a whole
People Group History	information about the history of the Khuen people
School	information about the language mix in the village school (if any) and whether children go elsewhere for some or all of their schooling
Languages of Wider Communication	information about the use of each of the LWC's used in the village
Intermarriage	information about extent and conventions of intermarriage between Khuen and other ethnic groups

Table 19 Description of sections on Knowledgeable Insider Questionnaire

4.5 Site Selection

The selection of locations to collect data for the survey was guided by the main goals of the survey described in Section 4.1 as well as the factors described in Mann (2006). Because of time constraints, it was decided to focus on the 10 village tracts where Khuen is strongest. From the point of view of phonology, there are clear benefits from collecting wordlists from communities where the language is used every day. By limiting the survey sites to these village tracts however, the scope of the inferences from the study is also limited to the Khuen population in these village tracts. In other words the results of this survey will not apply to mixed communities where Khuen speakers are in a minority.

The sites were selected in consultation with the Khuen Cultural and Literature Committee (KCLC). The criteria used for the selection of sites were geographical distribution, prestige and perceived dialectal differences. The names of the sites (villages) and which of the 10 Khuen-majority village tracts they lie in are listed in Table 20. As can be seen from the table, a total of 7 survey sites were included representing 6 different village tracts.

Village Tract Reference	Name of Village Tract	Name of Village included in Sample	Reason for Selecting Village
1	Yang Lorh	Yang Lorh	Prestige
2	Dory Lohng		
3	Wat Sahw		
4	Kaht Taw	Pa Jahm	Reported dialectal differences
5	Kart Thay		
6	Yang Kehng	Wan Jorhn	Reported dialectal differences
7	Kaht Fah		
8	Murng Lang	i. Yang Kway ii. Wan Jay	Reported dialectal differences
9	Murng Lahp	Wan Kahng	Geographic location
10	Murng Jem	Murng Jem	Geographic location

Table 20 Names of survey sites

A full listing of the village tract names written in Khuen script as well as IPA and Roman script is given in Appendix 1. Also listed in Table 20 are the reasons for selecting each village as a survey site. Yang Lorh village was chosen because it was recognised as having prestige. It is a large village near to Keng Tung and home to a number of families with high social status. It is also home to a highly respected monk. The villages of Pa Jahm and Wan Jorhn were selected because of reported dialectal differences. Murng Lang village tract has the strongest reported evidence of dialectal differences, so two sites were selected there, namely Yang Kway and Wan Jay. Yang Kway is reached relatively easily from Keng Tung, whereas Wan Jay is considerably more remote. The villages of Wan Kahng and Murng Jem were chosen because they lie to the south of Keng Tung in contrast to the other sites.

The map in Figure 39 shows the village tracts in Keng Tung township immediately around Keng Tung. This map is taken from the map in Figure 5. The approximate locations of the survey sites are signified by ▲ on the map.

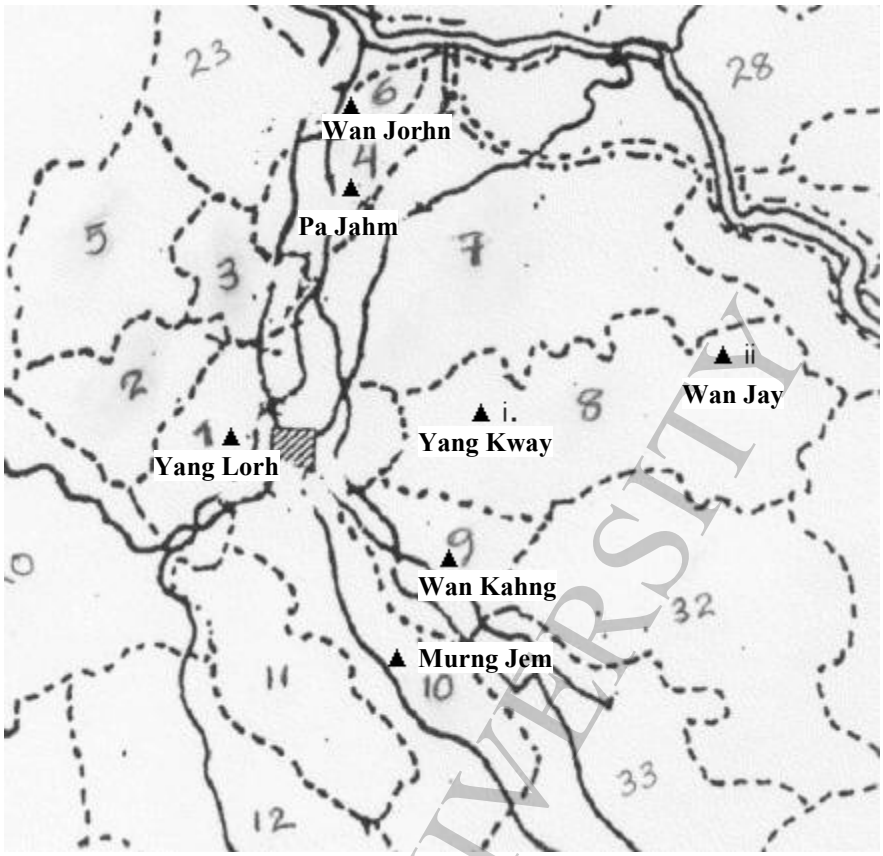

Figure 39 Approximate locations of survey sites

4.6 Subject Selection

The method of selection of subjects is of great significance for the validity of the study, both in terms of obtaining quality data and in minimising bias in the sample. Criteria for subject selection are described in the following two sections – first for subjects answering the Individual Sociolinguistic Questionnaires and secondly for wordlist language resource persons (LRPs).

4.6.1 Questionnaire subject selection

The methodology used to select subjects for the Individual Sociolinguistic Questionnaires was a quota sample of twelve adult Khuen speakers from each location. Two subjects in each of six age-gender categories were included in the sample. The age categories were chosen to ensure that the data collected came

from a broad range of the adult population. Gender was also chosen as a factor because there are significant differences in social roles between men and women which are expected to affect the linguistic features under investigation in the survey. Quakenbush (1992) compared reported language proficiency with proficiency assessed by means of oral proficiency interview. One notable feature of those cases where there was a significant discrepancy between reported and tested scores was that women underestimated their proficiency (Quakenbush 1992:70). This is further motivation to include gender as a factor in the design. Table 21 summarises the sampling design for a single survey site.

		Age			Total
		15-30	**31-45**	**46+**	
Gender	**Male**	2	2	2	6
	Female	2	2	2	6
Total		4	4	4	12

Table 21 Sampling design for a single survey site

Due to time constraints no subjects answered Individual Sociolinguistic Questionnaires in Wan Jorhn village. The full sampling design was carried out at each of the other 6 sites meaning that a total of 6 x 12 = 72 Individual Sociolinguistic Questionnaires were collected.

The survey team made a preliminary visit to each village where they contacted the village chief and explained the purpose of the survey. The team were generally well received and found the village chiefs most helpful. The team explained that they would like to interview Khuen speakers in the various age-gender categories and the village chief arranged for individuals to come and take part in the survey.

The quota sampling method was chosen for efficiency and simplicity. However, since it is not a random sampling technique, care must be exercised when making inferences from the sample. Nahhas (2007a) describes the pros and cons of various random and non-random techniques and the information/assumptions necessary to be able to generalise from the sample to the population as a whole. For example, in order to generalise from this sample to the whole adult male

Khuen speaking population, the proportions of the whole male Khuen speaking population in the three different age groups would need to be known. Since this information is not available, separate estimates will be reported for the three different age categories. Since men and women have distinctly different social roles, it is felt that any quantities of interest should be estimated separately for men and women.

Even in generalising from the sample in a particular age-gender category to the whole of that category in the population requires certain assumptions to be made. The key issue is that the subjects are chosen in an unbiased manner. Since human judgement was involved in selection of the subjects, certain steps were taken by the survey team to mitigate any bias. The preliminary visit gave time for the village chief to select and forewarn particular subjects, thus making it more likely that those individuals would be available when the survey team came to conduct the interviews. This reduced the chance of bias from the unavailability of the subjects of choice. The team also asked the village chief to not choose people from the same household so as to include subjects from as wide a range of households as possible.

4.6.2 Wordlist Language Resource Person (LRP) selection

The village leader in each location was asked for assistance in identifying at least two LRPs who satisfy the criteria specified below.

- Each LRP should be a mother tongue speaker of Khuen whose parents spoke Khuen with them.
- Each LRP should have lived in the village all or most of their life.
- At least one LRP should be 45 years old or over.
- Each LRP should have all their front teeth and no impairment to their hearing.

Having more than one LRP increases the reliability of the responses given since the LRPs can discuss any issues that may arise or confer over the exact meaning or usage of a particular word. Choosing LRPs over the age of 45 has the advantage that the LRP should be able to comment on whether there have been any changes in the language during their lifetime, that is, they have a broader perspective than a younger person. One potential disadvantage is that more recent developments in the language might be underrepresented.

At each of the 7 survey sites the wordlist was elicited once. Each time between 2 and 4 LRPs participated in the elicitation process so each of the 7 wordlists represents a consensus from among LRPs selected by the village chief as being good representatives of the language spoken in that village.

4.7 Summary of Data Collected on Sociolinguistic Survey

Having described the selection of survey sites and subjects in the previous two sections, in this section the results of the data collection process are described. Table 22 presents a summary of the total body of data collected from all of the survey sites.

Village Tract Reference	Name of Village Tract	Name of Village in Sample	Data Collected in Village			Ages of LRPs contributing to WL
1	Yang Lorh	Yang Lorh	KIQ; WL; ISQ (x 12)			40, 50, 52
2	Dory Lohng					
3	Wat Sahw					
4	Kaht Taw	Pa Jahm	KIQ; WL; ISQ (x 12)			62, 66, 69
5	Kart Thay					
6	Yang Kehng	Wan Jorhn	KIQ; WL			48, 52, 54
7	Kaht Fah					
8	Murng Lang	iii. Yang Kway iv. Wan Jay	KIQ; WL; ISQ (x 12) KIQ; WL; ISQ (x 12)			30, 46, 49, 72 61, 71
9	Murng Lahp	Wan Kahng	KIQ; WL; ISQ (x 12)			45, 60
10	Murng Jem	Murng Jem	KIQ; WL; ISQ (x 12)			46, 52, 60
Total Data Collected			**KIQ**	**WL**	**ISQ**	
			7	7	72	

Table 22 Summary of data collected on sociolinguistic survey

In Table 22 KIQ denotes the Knowledgeable Insider Questionnaire; WL denotes the Khuen 406 wordlist and ISQ denotes the Individual Sociolinguistic Questionnaire. As seen in the table, the survey team collected a wordlist and KIQ at each survey site making a total of 7 wordlists and 7 KIQs. Due to time constraints, no ISQs were collected in Wan Jorhn. Thus a total of 6 x 12 = 72 ISQs were collected from the other six survey sites. The final column of Table 22 shows the ages of the LRPs who participated in the elicitation process in each location.

4.8 Methods of Analysis

In this section methods of analysis are described and a priori criteria are specified for evaluating the information elicited by each of the survey instruments.

4.8.1 Analysis of wordlists

Recordings of the wordlists were transcribed using the International Phonetic Alphabet (International Phonetic Association 1999) and the Chao (1930) system of transcribing tone. Speech Analyzer[27] was used in the transcription process, particularly in the transcription of tone. Wordlists from the various survey sites were compared with each other and with wordlists from other Tai languages of the region. The wordlists were compared in two ways. Firstly the wordlists were used to deduce a tentative word-level phonology for each speech variety. The consonants, vowels and tone systems for each speech variety were then compared using Robinson's (1994:141) hierarchy of phonological features. The hierarchy of phonological features readily defines a classification of speech varieties.

Secondly, a common core of items from the wordlists was compared using lexicostatistics. The higher the percentage of lexically similar items for any two varieties, the closer the relationship between those two varieties was taken to be.

[27] Speech Analyzer was developed by JAARS (an affiliate with SIL International). This instrument has pitch and spectro analysis capability.

Interrelationships between all speech varieties were represented in a phenogram created by using the Unweighted Pair-Group Method using Arithmetic Average algorithm (UPGMA). The proximity of two speech varieties in the resulting phenogram is proportional to the degree of similarity of those speech varieties. Groupings of speech varieties corresponding to significant differences in lexicostatistical percentages were considered more credible than groupings that have greater likelihood of occurring by chance. The perceptions of individuals about different Khuen varieties were compared to the evidence from the wordlist data to ascertain whether or not the perceptions relate to identifiable phonological features or lexical similarity percentages.

4.8.2 Analysis of questionnaire responses

Categorical data was presented in contingency tables and tested for statistical significance as follows. For 2×2 tables (i.e. tables with 2 categorical variables each of which can take one of two levels) a Fisher Exact test was used. For 2×3 tables (i.e. tables with 2 variables one of which can take one of two levels and the other one of three levels) a Freeman-Halton test was used. For tables with more than two variables log-linear analysis was used. Richard Lowry's online facility[28] for statistical computation was used for each of these tests as well as computing confidence intervals for estimates of literacy rates. Numerical data was plotted and analysed using Minitab Statistical Software (Minitab Inc. n.d.).

The results were interpreted with respect to each of the specific research questions as well as the broad goals of interest in the study. Results that would support a need for literacy in Khuen would be questionnaire responses indicating high Khuen language vitality and low proficiency in the LWCs used in the region. Low Khuen language vitality would likely contraindicate a need for literacy in Khuen. High bilingual proficiency may contraindicate a need for literacy in Khuen if

[28] VassarStats: Web Site for Statistical Computation at http://faculty.vassar.edu/lowry/VassarStats.html

attitudes toward the other language are such that literacy in that language would be accepted.

Results that would support the promotion of one variety for use as the written standard would be widespread acknowledgement that speakers from the area where that variety is spoken speak Khuen the best.

The data collected on the survey is summarised and analysed in the following two chapters. Chapter 5 is focused more on sociolinguistic issues, as expressed in Goals 1 and 2 and Chapter 6 covers the areas of phonology and lexical similarity as expressed in Goals 3 and 4.