

CHAPTER 3

LEXICOSTATISTICS

3.0 Introduction

As stated in the first chapter, the goal of this thesis is to identify and compare the varieties of Kuy spoken in Cambodia using four different approaches. The first approach looked at sociolinguistic factors, or self-reported information about dialects and language use. The second approach, covered in this chapter, is *lexicostatistics*. This is a statistical method employed to measure the percentage of similar lexical items between two varieties. Details of the methodology used to determine lexical similarity and the results in terms of dialect groupings for the Kuy varieties of Cambodia will be presented in this chapter.

Lexicostatistics can be used to indicate whether speakers of two varieties would *not* be able to understand each other (if a small percentage of lexical items are similar) or whether speakers of those varieties could *potentially* understand each other (when a higher lexical similarity percentage is found). Methods other than lexicostatistics must be used to assess more clearly the level of intelligibility between varieties. By itself, lexicostatistics is accurate only for relative, not absolute comparisons, since several other factors besides lexical similarity affect intelligibility. This is discussed by J. Grimes (1988:19):

At the low end of the scale there is a constant relationship: comprehension is always poor when vocabulary similarity is low. But that relationship does not hold up at the high end of the scale... The reason why high similarity is a poor predictor of high intelligibility is that there are other factors besides similarity in vocabulary that influence intelligibility...e.g., differences in function words and affixes, syntactic and morphological rearrangements, certain kinds of regular sound shifts, and semantic shifts in both genetically derived vocabulary and loans.

Huffman (1976a:552) experimented with lexicostatistics using different lengths of wordlists in several Mon-Khmer languages (including Kuy), and says:

My conclusions...are that 1. basic vocabulary is highly specific to individual groups of languages, or perhaps to individual cultures, and that 2. lexicostatistics...is useful in showing relative distance between languages within a given group of languages and using a given corpus of vocabulary, but that absolute percentages are meaningless.

Before delving into methodology, it should be pointed out that there is often some confusion between *lexicostatistics* and *glottochronology*, which are sometimes mistakenly thought to be the same. Trask (1996:362) notes that Swadesh introduced a “time element” into lexicostatistics in the 1940s, and in the 1950s Robert Lees derived an equation for this process of glottochronology. Lexicostatistics, therefore, is one part of glottochronology, but glottochronology goes beyond this, as described in the following definition:

Glottochronology: Subdiscipline of lexicostatistics founded by M. Swadesh that investigates historically comparable vocabularies using statistical methods. The aim of glottochronology is to determine the degree of relatedness between languages as well as an approximate dating of their common origin and divergent development. This process was developed in analogy to the carbon-14 method, in which the age of organic substances can be determined based on the decay of the radio-active isotopes contained within them. Similarly, glottochronology is used to determine the ‘life span’ of words in their respective vocabularies.... The methods and conflicting results of glottochronology have come under criticism. (Bussmann 1990:192)

On the other hand, lexicostatistics can successfully be used by itself for a preliminary indication of language relationships, as the following linguists state:

[The term lexicostatistics] is most particularly applied to a simple procedure for estimating the degree of linguistic distance between genetically related languages. The central idea is that individual words in any language are steadily replaced over time. Thus, if we have several languages which we know are related, then we can choose a representative sample of the vocabularies of all of them and calculate the percentage of shared vocabulary items. Languages which share a larger proportion of their vocabularies are presumably more closely related than those sharing a smaller proportion.... This is admittedly a rather crude approach, but it may sometimes yield results of interest. (Trask 1996:361-2)

Lexicostatistics is not a precision tool. Careful phonological reconstruction is necessary if one desires detailed information about language relationships. Lexicostatistics is useful, however, for giving a quick general picture of language groupings. (Thomas & Headley 1970:411)

It is this latter use of the term *lexicostatistics*, that of “giving a quick general picture of language groupings,” which is employed in this chapter. In the present study, instead of looking for language groupings, lexicostatistics is applied to the Kuy varieties studied in Cambodia in order to identify dialect groupings. There is no glottochronological attempt at determining the dates when varieties separated from a common ancestor.

3.1 Methodology

A key to lexicostatistics is the choice of the items to be compared. It is essential that *basic* or *core vocabulary* (or the “vocabulary of ordinary life,” J. Grimes 1995:9) be selected, words which are not normally expected to change easily or be borrowed. This includes body-part names, nature terms, low numerals, simple verbs and adjectives, and pronouns. Morris Swadesh (1955:124) identified 100 words thought to be universal and culture-free.²² This list is very useful, though not perfect, for in any given language, one or more of the words on this 100-item list may be borrowed, may have more than one neutral equivalent or some other problem. There have been several adaptations to these Swadesh wordlists. Other

²² This 100-item list is a revision of a 200-item wordlist. See Section 1.4.2, especially Footnote 15.

lists are often based on the Swadesh 100- or 200-item wordlists and adapted to be appropriate to the language family and culture of the linguistic groups under study (see quote by Huffman in Section 3.0).

There has been some debate on the number of words to compare when calculating lexical similarity. Several studies have used varying lengths of wordlists (for example, Smith 1981, Huffman 1976a). In these studies, the absolute results change depending on the length of the list, but the overall relative comparisons remain. The 100 words chosen for the present study came primarily from the Swadesh 100 list, but also were cross-referenced with common items on several other lists. These words were already present in the full 566-item list used for the current research.²³

In order to find a non-biased, somewhat standardized list from which to choose words for lexicostatistical comparison, Mann (2003) created a hybrid wordlist by comparing four different lists often used for Southeast Asian languages. He prioritized all of the words by the number of lists on which they appeared; for example, those words which appear on all four wordlists appear first on the hybrid list, in alphabetical order by the English gloss, followed by those appearing on three lists, and so on. The first 100 items on this hybrid list are then chosen for comparison, except that if there is insufficient data for an item, the next word on the hybrid list is chosen. This process draws on the experience of other linguists in Southeast Asia and helps to prevent biased selection of words which could skew results. The 100 words chosen for comparison of the Kuy varieties in this study follows the list in Mann's paper.

The 100 words thus chosen are then compared across speech varieties according to set criteria. All twelve wordlists collected in this study are used for lexicostatistical comparison, so that groupings based on lexicostatistics can be compared with those groupings made by native speakers of how varieties relate to

²³ See Section 1.4.2 for more information on development of the full wordlist for this thesis.

each other. (See Section 3.3, Analysis of results, and Section 2.2.1, Reported information on varieties of Kuy.)

The methodology used for the lexicostatistical comparison is adapted from that described in Blair (1990:30-33). An important step in lexical comparison is to have criteria to follow consistently to determine whether a pair of lexical items is lexically similar or not. The criteria employed should be relevant to the language family. Using the method described in Blair, items are compared on a phone by phone basis, rather than by an overall guess. These criteria, which are applied to each pair of phones compared, are grouped into three categories: category one indicates phones are the same or closely related, category two indicates phones are somewhat related, while category three indicates phones are very different. The specific criteria used in this study are presented in Table 4.

Category	Criteria
Category 1	a. exact matches b. vowels differing by one feature (includes diphthongs) c. vowels differing by length d. regular sound correspondences
Category 2	a. phonetically similar consonants b. vowels differing by two or more features (includes length and diphthongs)
Category 3	a. non-phonetically similar consonants b. non-regularly occurring deletions
Ignore	a. supplemental semantic morphemes b. presyllables c. regularly-occurring deletions d. reduplicated syllables e. register distinctions

Table 4. Criteria for lexicostatistics

An example of the process of applying these criteria is given in Appendix C. Vowels are favored in this process, meaning that vowels can differ by more features than consonants can before affecting cognate status. In comparative phonological studies it is generally true that vowels tend to change more easily than consonants. In category 1 of Table 4, “vowels differing by one feature” indicates that the two vowel phones compared are only one step apart on a Kuy vowel chart. The terms “regular” and “regularly occurring” indicate that a pattern

is found three or more times in the data, while “non-regularly occurring” suggests less than three instances. “Phonetically similar consonants” are those consonants which differ by one feature (such as adjacent place or contrasting manner of articulation features) or those phones which are historically connected.

The “Ignore” criteria are used to filter out features which are not relevant to lexical similarity. Supplemental semantic morphemes are extra syllables attached to an item which have an additional, but somewhat separate meaning, such as the word for ‘fruit’ being given along with each particular kind of fruit, or the basic word for ‘tree’ attached to all parts of the tree, such as ‘root’, ‘branch’, ‘leaf’, etc. These morphemes should be ignored when counting lexical similarity, so that only the main root form for each item is compared. In many Mon-Khmer languages, including Kuy, there is a tendency for presyllables to change or decay over time. Sidwell (2000:40) states in regard to South Bahnaric word forms (using the term *minorsyllables* to refer to *presyllables*), “minorsyllables are inherently less stable than mainsyllables. This is because they are always unstressed, and can be dropped in speech if it does not cause ambiguity.” This applies to the Kuy varieties investigated in this study and leads to differences in presyllable forms where some varieties have retained the form largely as it was in the proto-language, some have a modified form and other varieties have lost the presyllable entirely. The presyllables are thus ignored in the lexicostatistical comparison, in order not to bias the results against similarity. Regularly-occurring deletions (those with three or more examples in the data) are ignored, for example, final [r] deletion. Reduplicated syllables, which are often added for emphasis, provide no additional phonological information and may be ignored, especially as they are often optional.

Register distinctions, as discussed in Section 4.4, are very relevant to the historical development of current Kuy varieties. However, an extensive study of Kuy register is beyond the scope of this thesis. In the initial collection of the wordlists, it is not clear if register was transcribed consistently or not. At this point, register is ignored in the lexicostatistical process, but should be included in future studies

if possible, as register provides a fuller understanding of the vowels in Kuy. Kuy is generally thought to have two registers, one clear and one breathy (see further discussion in Section 4.4). It is not yet evident whether some phonetic features present on some vowels in the data for this study, such as nasalization, tenseness, or diphthongization (in one variety), are a result of register or not. For the purposes of this study, all of these features, including breathiness, are ignored in the lexicostatistical process.

After each pair of phones has been compared according to the preceding criteria, the number of matches in each category are tallied for that lexical item pair. A matrix is then used to determine systematically whether the lexical items are considered similar or not, such that items are considered cognate if at least half of the phones are in category one, while of the remaining phones, at least half are in category two, as seen in Table 5.

Phones		Category 1	Category 2	Category 3
2	=	2	0	0
3	=	2	1	0
4	=	2	1	1
5	=	3	1	1
6	=	3	2	1
7	=	4	2	1

Table 5. Phone table for minimum lexical similarity

(Blair 1990:32)

The next step is to count the number of cognates between each pair of varieties. Missing items are not counted, so that a final percentage of lexical similarity between two varieties results from the number of cognates between those two varieties divided by the total number of items compared for the two varieties. In the present analysis, the total number of items compared between any two wordlists varies from 94 to 100 total items.

3.2 Lexicostatistic similarity between Kuy varieties

Following the methodology described in the preceding section, the percentages of lexical similarity between Kuy varieties were calculated using 100 basic, or core,

words. (The 100 words in all twelve wordlists can be found in Appendix D.) These percentages of lexical similarity are summarized in Table 6.

Tralaek											
98	Srae										
98	100	Rumchek									
98	100	100	Chi Aok								
96	98	98	98	Svay Damnak							
90	92	92	92	92	Pal Hal						
90	92	91	92	92	100	Prame					
90	92	92	92	92	100	100	Anlong Svay				
89	90	90	90	90	98	98	98	Samraong			
88	90	90	90	90	93	93	94	91	Chranaol		
87	89	89	89	88	90	90	91	91	96	Thmei	
85	84	84	85	84	90	88	87	86	85	82	Krala Peas

Table 6. Lexical similarity between Kuy varieties

The percentages in Table 6 are organized with the lower numbers to the bottom and left, and the higher numbers farthest towards the center diagonal. This allows for groupings of related varieties to be seen easily, as will be discussed in Section 3.3.

To better visualize the relationships between the varieties, an average link cluster analysis was calculated, following the Unweighted Pair-Group Method using Arithmetic Average, or UPGMA (as discussed in J. Grimes 1995:69). This analysis was generated using the “Neighbor” program which is part of the PHYLIP 3.6 suite of programs (Felsenstein 2002). The results of this cluster analysis for the Kuy varieties are shown in Table 7.

Percentage of similarity	Names of languages
100.0	Srae, Rumchek
100.0	Srae, Rumchek, Chi Aok
100.0	Pal Hal, Prame
100.0	Pal Hal, Prame, Anlong Svay
98.0	Tralaek, Srae, Rumchek, Chi Aok
98.0	Pal Hal, Prame, Anlong Svay, Samraong
97.5	Tralaek, Srae, Rumchek, Chi Aok, Svay Damnak
96	Chranaol, Thmei
91.6	Pal Hal, Prame, Anlong Svay, Samraong, Chranaol, Thmei
90.4	Tralaek, Srae, Rumchek, Chi Aok, Svay Damnak, Pal Hal, Prame, Anlong Svay, Samraong, Chranaol, Thmei
85.5	Tralaek, Srae, Rumchek, Chi Aok, Svay Damnak, Pal Hal, Prame, Anlong Svay, Samraong, Chranaol, Thmei, Krala Peas

Table 7. Cluster analysis by average link method

Table 7 presents the levels of average similarity at which each variety is related to the other varieties. There is a correlation figure of 0.953 on these calculations, indicating that the diagram conforms well with the actual data. “A correlation value close to 1.0 (such as .828) means that the tree representation is a reasonable distortion-free way of representing the similarity information” (J. Grimes 1995:72).

The data in Table 7 is graphically represented in the tree diagram shown in Figure 12, generated using the “Drawgram” program which is part of the PHYLIP 3.6 suite of programs (Feldsenstein 2002).

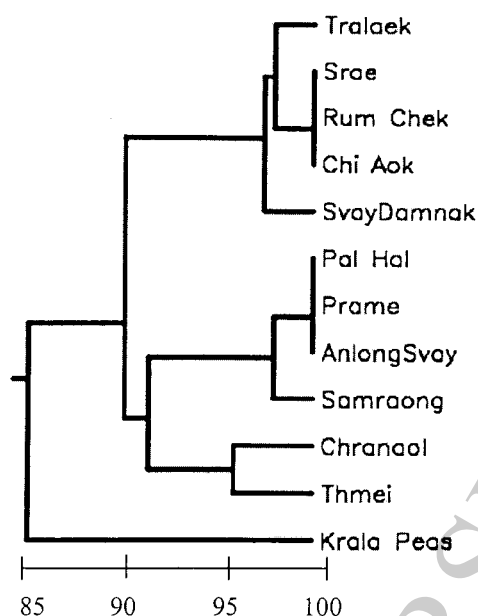


Figure 12. Kuy language tree

This tree diagram shows that the Kuy varieties spoken in Tralaek, Srae, Rumchek, Chi Aok and Svay Damnak (all of which are reported to be Kuy Ntra) are closely related to each other, since they cluster closely together on the tree. Similarly, the varieties spoken in Pal Hal, Prame, Anlong Svay and Samraong (all reported to be Kuy Ntua) are also closely related to each other. The varieties spoken in Chranaol and Thmei are grouped closely together, though slightly less closely compared to those within Kuy Ntra or within Kuy Ntua. Krala Peas is shown in the tree to be relatively less closely related to other varieties spoken in Cambodia. However, Krala Peas still has an average level of 85.5% similarity with the other Kuy varieties.

3.3 Analysis of results

All of the percentages of lexical similarity among the Kuy varieties collected in Cambodia are over 80%. The grouping of percentages of lexical similarity as reported in Tables 6 and 7 and Figure 12 is consistent with the self-reported information given by speakers of Kuy in Cambodia (see Section 2.2). The consistency of the lexical findings coupled with the self-reported information on

different types of Kuy is the basis to consider these groupings to be different dialects of Kuy (subject to confirmation by comprehension testing and comparative phonological reconstruction). The term *dialect* will subsequently be applied throughout the rest of this thesis. The table of lexical similarity percentages is reproduced here as Table 8 with the names of the dialects, as given by native speakers, and with shading, the darkest shade denoting similarity percentages from 100-96%, medium shade indicating 95-91%, light shade indicating 90-86% and no shading for 85-81%.

98	98																	NTRA
98	100																	
98	100	100																
96	98	98	98															
90	92	92	92	92														
90	92	91	92	92	100													
90	92	92	92	92	100	100												
89	90	90	90	90	98	98	98											
88	90	90	90	90	93	93	94	91										
87	89	89	89	88	90	90	91	91	96									
85	84	84	85	84	90	88	87	86	85	82								

Table 8. Lexical similarity between Kuy varieties,
with shading

The blocks of darkest shading in Table 8 indicate Kuy speech varieties which can be considered together as dialect groupings. These groupings correspond with self-reported information (Section 2.2.1), such that wordlists from those locations indicated by native speakers as being of the same variety or dialect have 96% or higher similarity with each other. For example, the five speech varieties reported to be Kuy Ntra are shown in Table 8 to group together with percentages ranging from 96% to 100% similarity. The four reportedly Kuy Ntua speech varieties group at 98% to 100% similarity. The speech varieties in Chranaol and Thmei show 96% similarity with each other. These two were reported simply as Kuy, with no dialect name given (though ‘Mai’ is used in this thesis based on the word for ‘what’ in Chranaol and Thmei). Speakers in both Chranaol and Thmei did report that these two villages speak the same variety of Kuy.

Krala Peas is the most different from the other varieties, ranging from 82% to 90% similar. Speakers in Krala Peas indicated that they knew of no other villages which spoke the same as them, but noted Prame and Pal Hal as being somewhat similar to their variety (and also close geographically). In Table 8, Krala Peas shows a slightly higher similarity with these two villages than with the other speech varieties, though not enough to be considered the same dialect.

These results show that the Kuy people in Cambodia are able to distinguish Kuy varieties, which is supported by lexical differences. However, in regard to the subvarieties (or subdialects) of Kuy Ntra, namely Kuy Wa' and Kuy O', reported information about different varieties is not necessarily reinforced by lexical similarity percentages. Kuy Wa' was reported to be spoken in Tralaek (though not reported by residents interviewed in Tralaek itself), and Kuy O' in Svay Damnak (spoken by a few older residents). As seen in Table 8, these two speech varieties have 96% lexical similarity, and each has 98% lexical similarity with locations reported to be Kuy Ntra. The high percentages of lexical similarity with other Ntra wordlists seems to confirm preliminary observations that the varieties are basically the same except for more frequent use of certain final particles. Perhaps also a change has been in progress, where the reported information reflects the past, and lexical similarity percentages reflect the present situation. Residents of Svay Damnak and Tralaek each suggested that even though their ancestors used the respective particles frequently, the current residents "no longer" speak O' or Wa', respectively, but now speak Ntra.

The preceding Table 8 can be used to find an average lexical percentage shared for each Kuy dialect. As an intermediate step, Table 9 provides the numbers from the varieties given in Table 8, grouped into blocks by dialect.

98	Ntra										
98	100										
98	100	100									
96	98	98	98								
90	92	92	92	92	Ntua						
90	92	91	92	92	100						
90	92	92	92	92	100	100					
89	90	90	90	90	98	98	98				
88	90	90	90	90	93	93	94	91	Mai		
87	89	89	89	88	90	90	91	91	96		
85	84	84	85	84	90	88	87	86	85	82	Mla

Table 9. Lexical similarity between Kuy varieties,
grouped by dialect

The average within each block is then calculated. First the average among wordlists of the same dialect is calculated; that is, the average is found within the Ntra block and within the Ntua block of Table 9. The Mai block has only one percentage. Only one Mla wordlist exists in the current data so comparisons cannot be made within Kuy Mla. The averages within dialects are presented in Table 10.

	Ntra	Ntua	Mai
Average within dialect	98	99	96

Table 10. Average lexical similarity within dialects

These averages range from 96% to 99% similarity. The average lexical similarity percentages between dialects can also be calculated based on Table 9. For example, there are five Ntra wordlists, so the five percentages in the bottom left block can be averaged to give a single percentage of lexical similarity between Kuy Ntra and Kuy Mla. Likewise, the percentages in each other block can be averaged. The results are presented in Table 11.

Ntra					
91	Ntua				
89	92	Mai			
84	88	84	Mla		

Table 11. Average lexical similarity between dialects

In Table 11, the average lexical similarity between dialects ranges from a low of 84% between Ntra-Mla and Mai-Mla to a high of 92% between Ntua-Mai dialects. Mla is somewhat less closely related to the other dialects, with average lexical similarity percentages below 90% with each of the other three dialects. The Ntra, Ntua and Mai dialects seem slightly more closely related to each other, with lexical similarity percentages ranging from 89% to 92%.

The average similarity percentages within dialects in Table 10, which are 96% and above, are all higher than the average similarity percentages between dialects, which range from 84% to 92%, in Table 11. This provides extra support for viewing these varieties as distinct dialects.

Further evidence comes from lexicostatistical comparisons within the Katuic branch. Exact percentages cannot really be compared, since these comparisons were done by different researchers²⁴ on different varieties; however, the general range is still useful. Kuy as spoken in Thailand is 70-85% lexically similar with other varieties in the West Katuic group (Nheu, Suai and Kuay), and 40-70% lexically similar with other Katuic languages (such as Bru, So, and Katu). Kuy is 32-40% lexically similar with Khmer.

3.4 Summary

As discussed in Section 3.0, lexicostatistics is useful for relative comparisons of varieties and for identification of dialect groupings. Preliminary indications from lexicostatistics as reported in this chapter are consistent with the self-reported information (Section 2.2) that there are four main dialects of Kuy spoken in Cambodia. The lexical similarity percentages between Kuy speech varieties in this study are 82% similar and above. Following Huffman's conclusion as discussed in Section 3.0, absolute percentages are not the focus, but relative similarity shows

²⁴ See the following studies: Thomas & Headley (1970), Huffman (1976a), Smith (1981), Migliazza (1992), Miller and Miller (1996), and Peiros (1996).

different groupings. Within this study, percentages above 95% seem to indicate the same dialect grouping. Four dialect groupings are evident among the twelve speech varieties studied. Three of these groupings are given the following dialect names by Kuy speakers (both those speaking the particular varieties and those speaking other varieties of Kuy): Kuy Ntra, Kuy Ntua and Kuy Mla. The fourth grouping, that of the two villages found in Kracheh province, was not given a specific name by native speakers, but is termed Kuy Mai in this thesis. Based on the lexical similarity results combined with sociolinguistic information, these four types of Kuy are considered dialects of Kuy in Cambodia.

An overview and phonological description of each of the Kuy dialects is provided in the next chapter. This is followed by comparative phonological reconstruction and results of comprehension testing, to further clarify the relationship between Kuy dialects.