Phonological Principles for Automatic Phonetic Transcription of Khmer Orthographic Words

Makara Sok
Payap University / Chiang Mai, Thailand
makara_sok@hotmail.com

Larin Adams
Payap University / Chiang Mai, Thailand
larin.adams@gmail.com

Abstract

This paper explores phonological regularities in the Khmer language which can be used to convert Khmer words written in Khmer script into both phonemic and close phonetic transcriptions. They involve *series assimilation, vowel modification*, and *sound change rules* which govern how a word should be pronounced. Based on these rules, a Thrax grammar was written to produce word transcriptions which closely approximate actual speech.

1. Introduction

The Khmer writing system is known as one of the most complex scripts because of its large alphabet inventory and because symbols (orthographic characters) are strung together in complex ways to form a word. The script often does not represent the inherent vowels whose pronunciation is dependent on the series of the surrounding consonants. Huffman (1970) claimed that the Khmer writing system is far more regular than that of English--the sounds (i.e. the phonological structure) and the symbols (i.e. the writing system) closely fit together leaving little to less ambiguity. However, no attempt to model this alleged regularity currently exists. Pali/Sanskrit loanwords are of exception, for they have their own pronunciation rules. It is also worth noting that the majority of Khmer native roots are either monosyllabic or disyllabic which is the main focus of this project.

2. Khmer Orthography

According to the Ministry of Education of Youth and Sport, there are 33 consonants, 23 dependent vowels[1] and 13 independent vowels--excluding some deprecated characters. Khmer words are written from left to right, though vowels could go before, after, above and around the base consonant. Subscripts can be placed under a base consonant to form a consonant cluster. Two types of special diacritics are used where needed to change consonant series and/or modify its inherent or dependent vowel; they are usually placed above a consonant. The order of character writing could be arbitrary in handwriting, but electronically they are fixed by the Unicode consortium.

In this research, characters are ordered as the following:

C (S) (D1) (V) (F) (D2)

where:
C -- any consonant
S -- any subscript (Coeng[2] + C)
D1 -- series shifter (◌៊/◌៉)
V -- dependent/inherent vowel
F -- any consonant in the final position
D2 -- Bantoc (◌់)

The orthographic syllable structure here shows that a word could be composed of just a single consonant. Since the inherent vowel is invisible, the orthographic vowel is optional.

[1] In modern Khmer system, there are three additional vowels.

[2] ◌្ (U+17D2) renders any consonant after it as a subscript in monosyllabic words, and in disyllabic words, it functions as a syllable break.

3. Khmer Phonology

	Labial	Alveolar	Palatal	Velar	Glottal
Plosives	p	t	c	k	ʔ
Asp. Plosives[3]	p^h	t^h	c^h	k^h	
Implosives	ɓ	ɗ			
Fricatives	(f)[4]	s			h
Nasals	m	n	ɲ	ŋ	
Semi-vowels	w		j		
Lateral		l			
Trill		r			

Table 1: Consonant Phonemes

	Front		Central		Back	
High	i	ii	ɨ	ɨɨ	u	uu
Mid	e	ee	ə	əə	o	oo
Mid-Low		εε	a	aa	ɔ	ɔɔ
Low			ɑ	ɑɑ		

Table 2: Monophthong

Diphthongs: iə, ɨə, uə, ae, aə, ao, oa, ie, ea

4. Dataset and Methodology

The dataset used in this experimentation is obtained from the official fifth edition of the Khmer monolingual dictionary published by the Buddhist Institute of Cambodia in 1967. The dictionary contains around 17,000 entries, which could not be used right away. It had to be cleaned of duplicates and stray characters such as: "!, ?, and ៗ". Pali/Sanskrit loanwords were identified and then removed for this stage of the project, leaving only native words for the conversion process (see Figure 1 below). Rounded rectangular balloons with fix border lines contain description of source/content of data which was put though each process labeled in heptagon balloons. On the right side, the rounded rectangular balloons with dotted border lines describes actions taken. For instance, the data used in the first process "cleanup.rb" was obtained from "Chuon Nath 2.0. original" lexicon. The "cleanup.rb" is a Ruby code file which was written using regular expression to carry out two cleanup actions: "removing duplicates" and "removing prefix" from the lexicon.

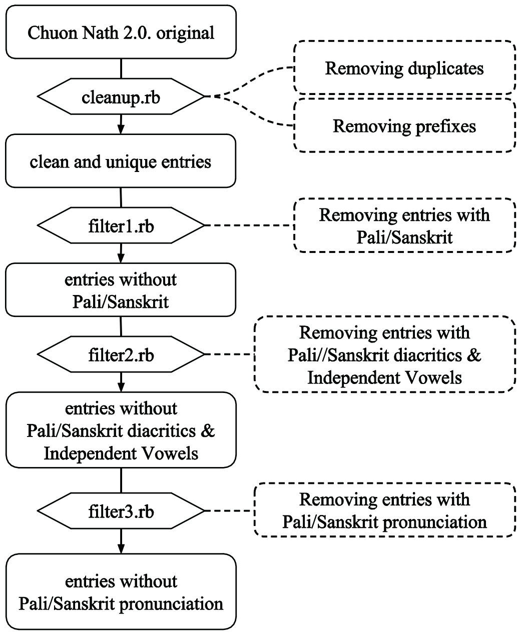
Chuon Nath 2.0. original
cleanup.rb
Removing duplicates
Removing prefixes
clean and unique entries
filter1.rb
Removing entries with Pali/Sanskrit
entries without Pali/Sanskrit
filter2.rb
Removing entries with Pali//Sanskrit diacritics & Independent Vowels
entries without Pali/Sanskrit diacritics & Independent Vowels
filter3.rb
Removing entries with Pali/Sanskrit pronunciation
entries without Pali/Sanskrit pronunciation

Figure 1: Data Cleanup

After the cleanup process, 11,109 words were identified as native words. The next step was to group words into their orthographic syllable structures: monosyllable disyllable, or double disyllable (see Section 6.3 below). There are four different types of disyllabic words, while there is only one type of monosyllabic words.

Grouping	Quantity
monosyllabic	3,969
disyllabic-type 1	1,249
disyllabic-type 2	815
disyllabic-type 3	769
disyllabic-type 4	864
double monosyllabic words	1,963
double disyllabic words	481
subtotal	10,110
remaining words[5]	999
Total	11,109

Table 3: Syllable Grouping

After preparing the dataset, character grouping and mapping were done. Then phonological regularities were explored and proposed.

In order to ensure the generated transcriptions are correct as expected, a set of validation data was built. 140 words were handpicked in order to assure that they cover all cases. Their expected phonemic transcription was taken from Headley's

[3] Ehrman (1972:4) included these aspirated consonants in the phonemic inventory.

[4] Only occurs in loan words.

[5] Any words which do not fit into the three groups were ignored, for they are compound words which is not yet covered at this stage.

Cambodian-English Dictionary (1997) and phonetic transcriptions were manually and carefully produced. The two outputs, phonemic and phonetic, generated by the grammar was then compared with the validation data. If the generated transcription match the validation data of the same orthographic word, it is considered as correct. Finally, phonemic transcriptions were automated for all 10,110 words in their respective syllable groups and compared with Headley's phonemic transcriptions.

5. Character Grouping and Mapping

Each orthographic consonant and vowel was mapped to an IPA character, while each diacritic was mapped to nothing. Consonants were grouped by their manner of articulations and series which would later be used to describe how the *series assimilation* works in initial consonant clusters.

- **Orthographic Consonants to Phonemes:**

Groupings		**Mapping**
Manner	**Series**	
Unaspirated Plosives	1st	ក > k ច > c ដ > ɗ ត > t ប > ɓ ប៉ > p * អ > ʔ
	2nd	គ > k ជ > c ឌ > ɗ ទ > t ប៊ > ɓ * ព > p អ៊ > ʔ *
Aspirated Plosives	1st	ខ > k^h ឆ > c^h ឋ > t^h ថ > t^h ផ > p^h
	2nd	ឃ > k^h ឈ > c^h ឍ > t^h ធ > t^h ភ > p^h
Fricatives	1st	ស > s ហ > h
	2nd	ស៊ > s * ហ៊ > h *
Nasals	1st	ង៉ > ŋ * ញ៉ > ɲ * ណ > n ម៉ > m *
	2nd	ង > ŋ ញ > ɲ ន > n ម > m
Approximants	1st	យ៉ > j * រ៉ > r * ឡ > l វ៉ > w *
	2nd	យ > j រ > r ល > l វ > w

Table 4: Character Grouping and Mapping

* marked consonants which have been modified by D1. These modification is done to fill in the

gap where certain consonants do not have the 1st/2nd series counterparts. For example,

- **Orthographic Vowels to Phonemes:**

Vowels	Series	
	1st	2nd
Inherent vowel	ɑɑ	ɔɔ
◌ា	aa	ie
◌ិ	eʔ	iʔ
◌ី	əj	ii
◌ឹ	əʔ	ɨʔ
◌ឺ	əə	ɨɨ
◌ុ	oʔ	uʔ
◌ូ	oo	uu
◌ួ	uə	uə
◌ើ	aə	əə
◌ឿ	ɨə	ɨə
◌ៀ	iə	iə
◌េ	ee	ee
◌ែ	ae	ɛɛ
◌ៃ	aj	ej
◌ោ	ao	oo
◌ៅ	aw	ɨw
◌ុំ	om	um
◌ំ	ɑm	um
◌ាំ	am	oam
◌ះ	ah	eah
◌ុះ	oh	uh
◌េះ	eh	eh
◌ោះ	ɑh	uəh

Table 5: Vowel Mapping

- **Diacritics (D1 and D2):**

D1 refers to one of the two series shifters: MUUSIKATOAN (U+17C9) and TRIISAP (U+17CA). The first changes the second series to the first series, and the later does the opposite.

D2, BANTOC (U+17CB), is placed on the final consonant to modified the vowel before it. Not any vowel could be modified by BANTOC. It is only applicable to the inherent vowel and the first dependent vowel (◌ា). See Table 6: Vowel Modification in section 6.3 below.

6. Orthography to Phonemic

One result of this process is a phonemic transcription which represent *careful speech*. It is important to note that it is always the case that the series of the consonant determines the series of the vowel attached to it. For instance, if the initial consonant is in the 1st series, the realization of the inherent vowel or the dependent vowel should also be in the 1st series. The same applies to the 2nd series initial consonant.

6.1. Monosyllable

Orthographic monosyllable structure:

C(S)(D1)(V)(D2)

The pronunciation rule is straight forward when the monosyllabic word does not contain initial consonant clusters or D1/D2 diacritics:

C	CF
CV	CVF

If an initial consonant cluster and/or D1/D2 diacritics are involved, special attention is needed. All syllable types of monosyllabic words containing initial consonant cluster and diacritic are listed below:

CS	CSF	CSFD2
CSV	CSVF	CSVFD2
CD1	CD1F	CD1FD2
CD1V	CD1VF	CD1VFD2
CSD1	CSD1F	CSD1FD2
CSD1V	CSD1VF	CSD1VFD2

In initial consonant clusters, there is a conflict of whether which series of the two consonants should be taken as the series of the cluster. Only when the series of the cluster is known, then the vowel attached to it would just follow. This is when *series assimilation* comes into play. Series Assimilation is a significant phenomenon in Khmer pronunciation which occurs in both monosyllabic and disyllabic words. The sonority hierarchy plays an important role in determining which series of the consonant cluster or the following syllable should be. The series of the least sonorous consonant (or in another word, strong consonantal) determines the series of the cluster regardless of its position in the cluster. Then the series of the vowel attached to it has to be in that same series. Note that the series of the syllable final consonant has no influence on the series of the vowel preceding it.

According to Hooper (1976:206) and Hogg and McCully (1983:33), the sonority hierarchy adapted to Khmer is as following: unaspirated plosives > aspirated plosives > fricatives > nasals > approximants (from the least sonorous to the most sonorous). Unaspirated plosives are the strongest consonant for they are the least sonorous, and the approximants are the weakest consonants for they are the most sonorous. Here are some examples:

(1)	ក្រាល 'to unroll'				
character sequence:	ក	្	រ	ា	ល
character mapping:	k		r	aa/ie	l
series of each character:	1		2	1/2	2
series of the cluster:	1				
phonemic transcription:	/kraal/				

(2)	ស្ពាន 'bridge'				
character sequence:	ស	្	ព	ា	ន
character mapping:	s		p	aa/ie	n
series of each character:	1		2	1/2	2
series of the cluster:	2				
phonemic transcription:	/spien/				

(3)	ព្រាល 'dim'				
character sequence:	ព	្	រ	ា	ល
character mapping:	p		r	aa/ie	l
series of each character:	2		2	1/2	2
series of the cluster:	2				
phonemic transcription:	/priel/				

(4)	ក្បាល 'head'				
character sequence:	ក	្	ប	ា	ល
character mapping:	k		ɓ	aa/ie	l
series of each character:	1		1	1/2	2
series of the cluster:	1				
phonemic transcription:	/kɓaal/				

Example (1) shows that the series of the cluster is 1st because ក /k/, an unaspirated plosive, is a stronger consonant (less sonorous) than រ /r/, an approximant; and the 1st series vowel is used. Example (2) illustrates the case a cluster of a fricative and a following less sonorous unaspirated plosive, the series of the cluster is 2nd; and the 2nd series vowel is used. Example (3) and (4) show the fact that if both consonant in the cluster are of the same series, the series of the cluster stay the same.

6.2. Disyllables

Orthographic disyllable structure:
Type 1: CN្+Monosyllabic Structure
Type 2: Cr+ Monosyllabic Structure
Type 3: C(D1)+ Monosyllabic Structure
Type 4: Cំ+Monosyllabic Structure

Where:
N -- any nasal consonant
r -- a subscript ្រ
ំ -- NIKAHIT, a vowel /ɑm/

Orthographically, disyllabic words are composed of two parts: a minor syllable and a major syllable (or a monosyllable which has previously been described). The major syllable no longer has *series assimilation* within its own syllable, but the series of the minor syllable does influence that of the major. The initial consonant of the minor syllable determines the series of the major syllable. It begins with a weak consonant.

For example,

(5)	សម្រាក 'to relax'					
character sequence:	ស	ម	្	រ	ា	ក
character mapping:	s	m		r	aa/ie	k
series of each character:	1	2		2	1/2	1
series of the minor-major:	1			1		
phonemic transcription:	/sɑm.raak/					

(6)	សម្គម 'skinny'				
character sequence:	ស	ម	្	គ	ម
character mapping:	s	m		k	m
series of each character:	1	2		2	2
series of the minor-major:	1			2	
phonemic transcription:	/sɑm.kɔɔm/				

Example (5) shows *series assimilation* cross syllable, while (6) illustrates just the opposite--dissimilation. The initial consonant (រ /r/) of the major syllable of example (5) is a weaker consonant, and (6) a stronger consonant--unaspirated stop គ /k/. The same rule applies to nasals (ង /ŋ/, ញ /ɲ/, ន /n/, ម /m/) as the initial consonant of the major syllable.

For example,

(7) សម្ងំ 'to remain quiet'

character sequence:	ស	ម ្	ង	ំ
character mapping:	s	m	ŋ	ɑm/um
series of each character:	1	2	2	1/2
series of the cluster:	1		1	
phonemic transcription:	/sɑm.ŋɑm/			

A question may arise as to where the open central unrounded vowel /ɑ/ in the minor syllable of each example above came from. This is exactly a case of the inherent vowel, but it has undergone a special treatment whereby the inherent vowel /ɑɑ/ is shorten to only /ɑ/.

6.3. Vowel Modification (D2)

Two vowels, the inherent vowels /ɑɑ/ɔɔ/ and the first vowel ា /aa/ie/ can be modified by adding a diacritic, BANTOC "់", to it. The BANTOC is used to shorten as well as change the vowel quality completely. It is attached to certain final consonants, such as: -ក /-k/, -ច /-c/, -ត /-t/, -ង /-ŋ/, -ញ /-ɲ/, -ន /-n/, -ល /-l/, -ស /-h/, -ប /-p/.

	Syllable Structures	Vowel Modification
1st series	CFD2 CD1FD2 CSFD2 CSD1FD2	/ɑɑ/ is shortened to /ɑ/ where F = -ក/-ង/-ច/-ញ/-ត/ -ន/-ល/-ស/-ប
2nd series		/ɔɔ/ is changed to /uə/ where F = -ក/-ស /ɔɔ/ is changed to /u/ (elsewhere)
1st series	CFD2 CD1FD2 CSFD2 CSD1FD2	/aa/ is shortened to /a/ where F = -ក/-ង/-ច/-ញ/-ត/-ន/- ល/-ស/-ប
2nd series		/ie/ is changed to /ea/ where F = -ក/-ង/-ច/-ញ /ie/ is changed to /oa/ (elsewhere)

Table 6: Vowel Modification

6.4. Sound Change Rules

The following sound change rules were adapted from a manuscript and implemented in the phonemic transcriptions:

- Final devoicing: voice implosives become voiceless in a syllable final position.
- Final unreleased: aspirated plosives become unaspirated in a syllable final position
- Implosive devoicing: voiced implosives become voiceless plosives in the initial position before another consonant.
- Fricative backing: alveolar fricatives become glottal fricatives in the syllable final position.
- Plosive backing: voiceless velar plosives become voiceless glottal plosives in the syllable final position.
- Trill deletion: alveolar trills get deleted in the syllable final position.
- Nasal deletion: nasals are deleted before another nasal in the syllable final position.
- Vowel backing: open and open-mid front unrounded vowel becomes central open and open-mid central unrounded vowel in closed syllable.

7. Phonemic to Phonetic

The result of this process is the phonetic transcription which is close to rapid speech or how people normally talk. A set of rules was observed and identified as followed:

- Aspiration is the transition between the first and second member of the cluster when the first member is one of the unaspirated plosives /p/t/c/k/ and the second member is one of these /p/t/c/k/m/n/ɲ/l/w/j/s.
  (Huffman 1972:55, Filippi 2009:164)
- Schwa is the transition if /m/l/ occurs before another consonant or ɓ/ɗ/ʔ occurs after any consonant. (Filippi 2009:165)
- Any sonorant consonants become voiceless when occurs after an aspirated plosive. (Filippi 2009:144-145)

- Plosives and nasals in the final position do not have audible releases. (Filippi 2009:144-145)
- The minor syllables of disyllabic words are subject to extreme reduction in rapid colloquial speech. They are usually reduced to just Cə-. (Huffman 1972:59)
    - Vowel of the minor syllable is reduced to a schwa [ə].
    - /rɔ-/ is reduced to [rə-] or [lə-].
    - /ɗɑ-/ and /sɑ/ are changed to [tə-].
    - /ɓɑ-/ is changed to [pə-].
    - /CrV-/ is reduced to [Cə-].
    - /ʔVN-/ is reduced to [N̩-].
    - /CVN-/ is reduced to [Cə].

8. Result

Based on the validation dataset of 140 selected words, the accuracy of phonemic transcription increases over time as more rules were added:

- character mapping: 2.14%
- series assimilation: 15%
- vowel modification: 50%
- sound change rules: 97.86%

The accuracy went up from 2.14% to 97.86%. The ~3% of errors is caused by the exception where certain words do not conform to any rules implemented.

Given that the phonemic transcription generated by the grammar satisfies the validation dataset, an attempt has been made to implement the grammar on the 10,110-word dataset in their respective syllable type. The results are:

Syllable Type	Accuracy Rate (%)
monosyllabic	99.06
disyllabic-type 1	98.32
disyllabic-type 2	97.06
disyllabic-type 3	96.62
disyllabic-type 4	98.61
Overall Accuracy Rate: 98.43%	

Table 7: Accuracy rate of the two syllable types

The accuracy rate of when the grammar was implemented on the monosyllabic and disyllabic words is comparable to the accuracy rate of when the grammar was implemented on the validation dataset. The phonetic rules are able to generate all 140 phonemic tokens.

The system was built using Thrax, a grammar compiler which compiles rules that consist of regular expression and context dependent rewrite rules into (FST) archive of weight finite state transducers. More rules can be added upon new encounters in order to improve the phonemic and phonetic transcription in an elegant way (Roark et. al. 2012:61).

9. Conclusion

The conversion does work as expected regardless of minor exceptional cases. There is always room of improvement. Pali/Sanskrit words should be studied and carefully incorporated into this current work.

Acknowledgements

I would like to express my very great appreciation to Dr. Larin Adams for his valuable contribution and patient guidance from the beginning to the end of this project. I would also like to thank Dr. Robert Batzinger for giving helpful advice and training me on how to do the data preparation using Ruby. My grateful thanks are also extended to Dr. Richard Sproat for introducing me to Thrax, to Dr. Thammanit Pipatsrisawat and Dr. Sopheakmungkol Sarin for their encouragement and help when I was struggling with Thrax grammar, and to Mr. Theeraphol Wattanavekin who helped me getting Thrax compiler up and running as well as providing instant support on Thrax when I needed it the most.

I would also like to extend my thanks to all professors and students at Payap University for providing helpful and constructive and helpful criticism on this project.

Finally, I wish to thank my mother and grandmother for their support emotionally and financially though out my study, without them I would not have come this far.

References

Chuon, Nath. 1967. Khmer-Khmer Dictionary. Phnom Penh: Buddhist Institute.

Ehrman, Madeline E., and Sos, Kern. 1972. Contemporary Cambodian: A Grammatical sketch. Washington. DC: Foreign Service Institute.

Filippi, Jean-Michael, and Hiep, Chanvichetr. 2009. ឯកសារណែនាំ អំពីសូរវិទ្យា Introduction to Phonetic. Phnom Penh: FUNAN.

Headley, Robert K., Rath Chim, and Ok Soeum. 1997. Modern Cambodian – English Dictionary. Kensington, Maryland: Dunwoody Press

Hogg, Richard M., and McCully, C. B.. 1987. Metrical Phonology: A coursebook. Cambridge: Cambridge University Press.

Hooper, Joan B. 1976. An introduction to Natural Generative Phonology. New York: Academic Press.

Huffman, Franklin E.. 1970. Cambodian System of Writing and Beginning Reader with Drills and Glossary. Adam Wood.

Huffman, Franklin E.. 1972. The Boundary between the Monosyllable and Disyllable in Cambodian. Lingua, 29.54-66.

Roark, Brian, Richard Sproat, Cyril Allauzen, Michael Riley, Jeffrey Sorensen and Terry Tai. 2012. The OpenGrm open-source finite-state grammar software libraries. In Proceedings of the ACL 2012 System Demonstrations, pp. 61-66.

The Unicode Consortium. 2015. The Unicode Standard, Version 8.0.0. Mountain View: CA http://www.unicode.org/versions/Unicode8.0.0/