A Multi-Study Analysis of an Unknown Southeast Asian Language for Establishing its Phonology

Geneviève Caelen-Haumont
Jean-Cyrille Ly Van Tu,
Alice Vittrant
Katarina Bartkova,

Mo Piu Project

- « Mo Piu » is a Southeast Asian language spoken in the North Vietnam mountains, near the Chinese border.
- Language in danger : in 2011, 237 speakers only
- Our previous studies have led us to postulate Mo Piu as a language belonging to the Hmong-Mien family
- Close collaboration between 4 linguists with different background and skills.
  - Focus on phonetic units
  - Instructions for labelling : each expert labels the same data by himself
  - Comparison of labelling
    - Measure of variability among experts
    - Unification of transcriptions
    - Quicker access to the Mo Piu phonological system

Mo Piu as a Hmong language

- *Spoken in an area known for its important linguistic diversity*
- *Belonging to Hmongic branch of Hmong-Mien languages has to be confirmed*

Clues for postulating Mo Piu as a Hmong language

1. Cultural features shared by other Hmongs :
   (rituals, bamboo pipes, traditional clothes, etc.)

Mo Piu bambou pipe player

Black Hmong of North-Vietnam
(photo Thanh xxx)

Mo Piu as a Hmong language

Clues for postulating Mo Piu as a Hmong language

Name of the language in Vietnamese (exonyme) : <mong xanh> [moŋ sɛŋ], i.e. *Green Hmong*

1. Name of the language in Mo Piu (endonyme) : [m̥õ$^{(ŋ)}$ **mbjo**] meaning *Green Hmong*
   cf . Ratliff (2010:256) Hmongic green/blue ***mpru**^{A}*
   Mienic green/blue **ʔmeŋ^{A}*

4. Words and reconstruction (1) : comparison between Mo Piu and Proto-Hmong-Mien/ Hmongic (from Ratliff 2010)

	Proto-Hg-Mn	Mo Piu
Sang/*blood*	*ntshjam^{X}	ntʃhaa
riz (cuit)/*rice*	*hnrɘaŋ^{H}	n̥õ$^{(ŋ)}$
manger/*eat*	*nuŋ^{A}	n̥õ$^{(ŋ)}$
langue/*tongue*	*mblet	mblɛ/mplɛ
queue/*tail*	*tɥeil^{X}	tœ(ʔœ) /tœ̰
nez/*nose*	*mbrui^{H}	mbjaa/mpjaa

Mo Piu as a Hmong language

- *Clues for postulating Mo Piu as a Hmong language*

4. Words and reconstruction (2) : comparison between Mo Piu and Proto-Hmong-Mien/ Hmongic (from Ratliff 2010 and *lingweb.ewa.mpg.de*)

Hmong Nzhuab, China		Dananshan Miao (Hmong Njua), China		**Mo Piu, Vietnam** *(transcription provisoire)*		*Reconstruction Proto-Hg-Mn*
1. ʔi^{54}	6. ɭou^{44}	1. ʔi^{43}	6. ʈəu^{44}	1. ɛ [æ]^{11}	6. ʈɔ^{11}	1. *ʔi
2. ʔau^{43}	7. ɕaŋ^{44}	2. ʔau^{43}	7. ɕaŋ^{14}	2. wa^{11}	7.dʑãŋ^{11}	2. *ʔu̯i
3. pei^{54}	8. zi^{22}	3. pe^{43}	8. zi^{24}	3. pa^{11}	8. ɟi^{11}	3. *pjɔu
4. plou^{54}	9. tɕuɑ^{42}	4. pləu^{43}	9. tɕua^{31}	4. plɔ^{11}	9. tɕo^{11}	4. *plei
5. tʃɻ^{31}	10. kou^{22}	5. tʂi^{43}	10. kəu^{24}	5. p(h)i^{11}	10. khɔ [ə]	5. *pra
						…
						10. *gjuɛp

http://lingweb.eva.mpg.de/numeral/

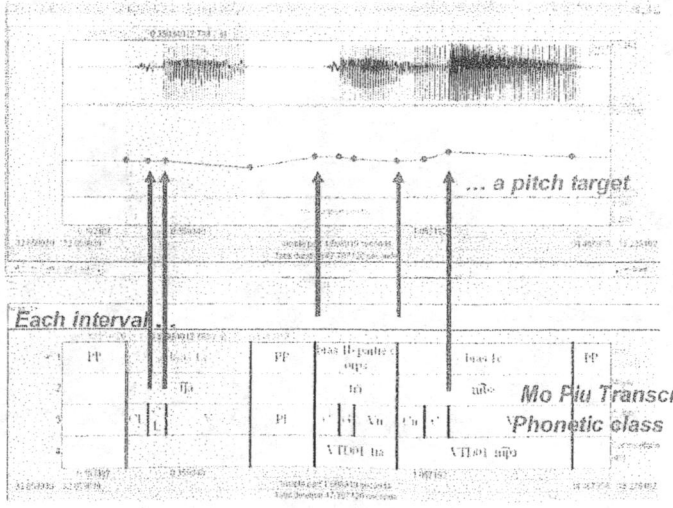
Data

- 3 fieldworks in 2009, 2010, 2011
- 36 hours of speech (and video)
  - 35 hours of speech (+ 1 hour of traditionnal songs)
    - * Continuous speech
      - « cultural » questionnaire
      - Life stories
      - Tales
      - Video comments (Projet Trajectoire – Fédération TUL-CNRS)
      - Stories told from pictures (frog story)
    - * Lists of words
      - Calmsea (from Matisoff 1978)
      - 500 mots *(500 words)*
      - Nouns in context (presentative phrases, classifiers etc.)

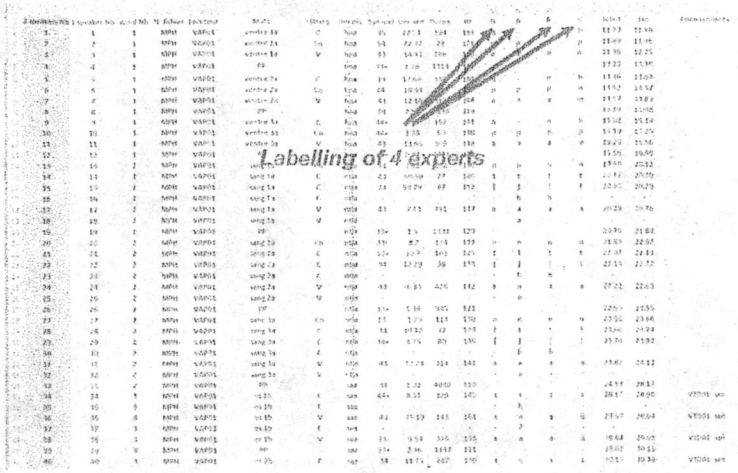
Data for this study

- One man (VAP01) and one woman (VTD01) were recorded
  - 44 words of *Calmsea* list of vocabulary (Matisoff 1978), body parts (repeated x 3)
  - All together, 418 lexical units in Mo piu, 4770 phonetic items transcribed
    - * 2541 vowels
    - * 2229 consonants
- Recording made by an engineer
  - 2 cameras : face and side
  - 2 microphones, 2 tracks *(#1- French + Vietnamese, #2-Mo piu)*
  - All monitored by computer
  - 25 speakers : 8 women + 17 men

Example of TextGrid with Praat-Momel



Example of xls file



Labelling of 4 experts

Aims of this study : 1- evaluating the variability between experts

- **Interesting study as the 4 experts do not have the same linguistic and cultural background**
  - Linguistic background
    - * 3 experts work in speech domain with speech technologies
    - * 2 experts have good skills in Oriental languages domain, and have been trained in Asian languages
    - * 1 expert has been raised in a Franco-Vietnamese culture
  - Different analysis methods : with or without the help of speech technology
- **Question : is this « cultural » difference is perceptible when labelling ? What about the variability between experts?**

Aims of this study : 1- evaluate the variability between experts

- **To evaluate this variability, we did :**
  - 2 confusion matrices for vowels and consonants
- **A confusion matrix is a table with 2 entries, where appear :**
  - on the horizontal axis: all the items found (after labelling by the 4 experts)
    - * The items are organised according to the proximity of their articulatory and / or acoustic features
  - on the vertical axis: all the items labelled by each expert

Ideologically, if the experts' transcriptions were corresponding totally

- there would be no dispersion
- i.e. only the diagonal would be filled

In practical terms, if the experts' transcriptions are highly matching

- The highest numbers of items are found in the diagonal line
- The dispersion occurs in the same articulatory class or acoustic class

Example of confusion matrix for the consonants

NB	p	t	k	b	d	g	ʈ	q	f	s	ʃ	z	ʒ	ç	j	h	m	n	ŋ	ɲ	l	ʎ
p	139	2		3	1																	
t	3	230			18		3		3	2	3	3	3			6						
k		2	25					3	2							2						
b	13			62	6													1				
d		1		3	33																	
g		1				13																
ʈ	5	49					54									3						
q	3	7	8					29	1									1				
f			1						58							9						
s										174	8			5		29						
ʃ											24											
z												12										

Example of confusion matrix for the consonants - *in percentage* -

%	p	t	k	b	d	g	t	q	f	s	ʃ	z	ʒ	ʧ	j	h	m	n	ŋ	ɲ	l	ʎ
p	95	1		2	0.7																	
t	1	83			6		1		1	0.7	1	1	1			2						
k		5	73					8	5							5						
b	15			75	7													1				
d		2		8	89																	
g		7				92																
t	4	44					48									2						
q	6	14	16					59	2									2				
f			1						85							13						
s										80	3			2		13						
ʃ											100											
z												100										

Aim 2 : Unifying our transcriptions

- **Method of phonetic encoding**
  - each expert transcribed the words in IPA
    - ★ visual and auditive analysis by using speech analysis and transcription software *(PRAAT-MOMEL, MISTRAL+)*
    - ★ *or auditive analysis only*
  - An xls file created with MISTRAL+ (MICA) enabled to gather the 4 transcriptions
    - ★ presentation by informant / mot 1… 44 (x 3 repetitions) / successive phonetic units
    - ★ with a whole set of prosodic parameters, manual or automatic
    - ★ tone analysis has been realized for another paper (cf. actes de TAL 2012)
  - 1815 phonetic units have been written down in an xls file (and 4470 annotated)

MISTRAL+ how diff from Praat

is this something you can do in Praat?

Melody Intonation Speaker Tonal Range Analysis using variable Levels

- et since INTSMEL script (2004) and MESLIM script (2008-2011)
- Dedicated to analysis of
- Enables to study:
  - ★ F0 modulation, derivatives, durations on phonetic, melodic an lexical levels
  - ★ Expressive, emotional, attitudinal speech of any language : accentuated, stressed, tonal or not
  - ★ The (even more important when the language is unknown) by a after tonal unit delimitation (syllable, vowel…)
-
  - ★ from the language structure
  - ★ from any linguistic theory
- during fastidious tasks

Method for phonetic notation converging

- **As mentioned above, there is a high conformity of the results on the confusion matrix.**
  - *Few dispersions*
  - Most of the dispersions occur in
- **Instructions for the notation**
  -
  -

A few examples of the inter-experts variability

- **Different traditions in notation**
  - The 3 speech experts have a notation which is more phonetic and acoustic.
  - The 4th expert annotates in a more phonological and more systemic manner
- **Divergences of notation among experts**
  - Voicelessness of nasales et liquids
    - n̥ / hn
    - l̥ / hl / ɬ
  - Transcription of clusters
    - tʃ / ç / ɕ
  - Transcription of glottalization
    - Under the segment transcription, indicating a creaky voice: [a̰]
    - Additional consonant in an intervocalic position: [aʔa]
    - Tone ?

A few particularities shared with the Hmong languages

- **Nasalization and glottalization are very frequent**
  - The 2 informants
- **Creaky and glottalized voice**

Technological analysis show a difference between

- ★A *creaky voice when the periodic structure is preserved despite the slowdown of the pulses*
- ★A *glottalized voice when the structure is not preserved*

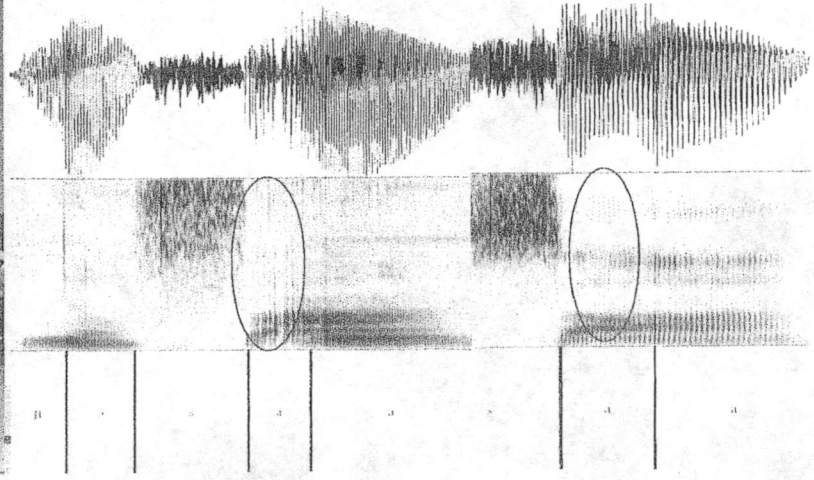
Is this distinction acoustic (informant variability / contextual) or phonological ?

A few particularities shared with the Hmong languages

Concerning the glottalization phenomenon, it still remains a variability too important between the 2 informants which prevents us from solving the problem.

Example: classifier [pɔ̃] bone [saa]

The female informant produces a … The male informant produces a …



Towards a phonological anaysis of Mo Piu

- By suggesting
  - A ***phonetically unified*** analysis of transcriptions and annotations
  - To structure the data in ***system***
  - To draw hypotheses much more established about the phonological system of the language
  - To ***make progress faster*** about this unknown language

Conclusion

- Word-by-word validation, phonetic unit by phonetic unit of the articulatory and acoustic matter
- Thanks to a selection of identical encodings common to all experts
- This analysis fulfills two goals
  - In the domain of [illegible]
    - First stage of the description of an endangered language
  - In the domain of [illegible]
    - [illegible], by being an unknown language…
    - …is used as a [illegible]
    - Automatic labelling and transcription from the whole set of phonetic units of various languages («language models »)

Surprisingly, when several associations of units belonging to Mandarin, Vietnamese, Cambodian, English and French were confronted, the combination Mandarin-French was the best one (presence of nasal vowels in French)

Conclusion

- Merci de votre attention
- Thanks for your attention

Geneviève Caelen-Haumont

Alice Vittrant

Aix Marseille université

LACITO

Une étude pluri-experte en vue d'établir la phonologie d'une langue inconnue de l'Asie du sud-est

ilpga

Jean-Cyrille Ly Van Tu

UNIVERSITÉ DE LORRAINE

Katarina Bartkova

institut de Linguistique et Phonétique

SORBONNE

atilf