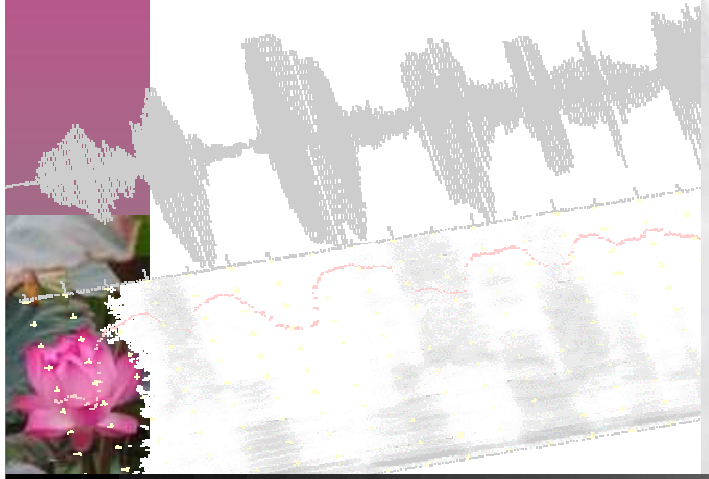
dépasser les frontières

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



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

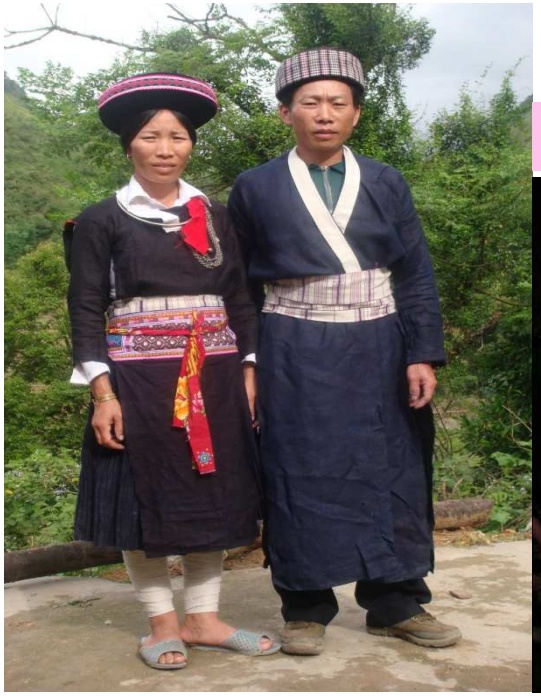
Le Projet Mo Piu

- **Le « *Mo Piu* » est une langue de *l'Asie du sud-est* : dans les montagnes du Vietnam du nord, près de la Chine**
- **Langue en danger de disparition : *en 2011, 237 personnes* la parlent**
- **Nos études antérieures ont permis de la rattacher aux *langues Hmong-Mien***
- **Collaboration étroite menée *par 4 linguistes, aux compétences diverses***
  - Focalisation sur les unités phonétiques
  - Consignes d'étiquetage : chacun étiquette pour soi le même corpus
  - Comparaison des étiquetages
    - ***mesure de la variabilité entre experts***
    - ***unification des transcriptions, soit un …***
    - …accès plus rapide à une vision plus cohérente du système et à sa ***représentation phonologique***

Le Mo Piu une langue Hmong

- Parlée dans une zone à forte diversité linguistique
- Rattachement très probable à la branche Hmongic (langues Hmong) des langues Hmong-Mien

- ***Les indices pour un rattachement du Mo Piu aux langues Hmongs***
  - Traits culturels communs aux Hmongs : *rituels, flûte à pipes de bambou, costume traditionnel, etc.*



Joueur de flûte Mo Piu

Hmong Noir du Nord-Vietnam

Le Mo Piu une langue Hmong

- ***Les indices pour un rattachement du Mo Piu aux langues Hmongs***
  - Nom de la langue (exonyme) en Vietnamien <mong xanh> [moŋ sɛŋ], càd. *Hmong vert*
  - Nom de la langue en Mo Piu (endonyme) : [m̥õ$^{(ŋ)}$ mbjo] signifiant *Hmong vert*

    cf . Ratliff (2010:256) Hmongic green/blue *mpru^{A}

    Mienic green/blue *ʔmeŋ^{A}
- **Mots et reconstruction : comparaison du Mo Piu avec le Proto-Hmong-Mien / Hmongic (d'après Ratliff 2010)**

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

Le Mo Piu une langue Hmong

- ***Les indices pour un rattachement du Mo Piu aux langues Hmongs***
  - Mots et reconstruction (suite) : comparaison du Mo Piu avec le Proto-Hmong-Mien/ Hmongic (d'après Ratliff 2010 *et 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. ʈou^{44}	1. ʔi^{43}	6. ʈəu^{44}	1. ɛ [æ]¤	6. ʈɔ¤	1. *ʔɨ
2. ʔau^{43}	7. ɕaŋ^{44}	2. ʔau^{43}	7. ɕaŋ^{44}	2. wa¤	7.dʑãŋ¤	2. *ʔu̯i
3. pei^{54}	8. ʑi^{22}	3. pe^{43}	8. ʑi^{24}	3. pa¤	8. ɟi¤	3. *pjɔu
4. plou^{54}	9. tɕuɑ^{42}	4. pləu^{43}	9. tɕua^{31}	4. plɔ¤	9. tɕo¤	4. *plei
5. tʃɻ^{54}	10. kou^{22}	5. tʂi^{43}	10. kəu^{24}	5. p(h)i¤	10. khɔ [ə]¤	5. *pra
						…
						10. *gju̯ɛp

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

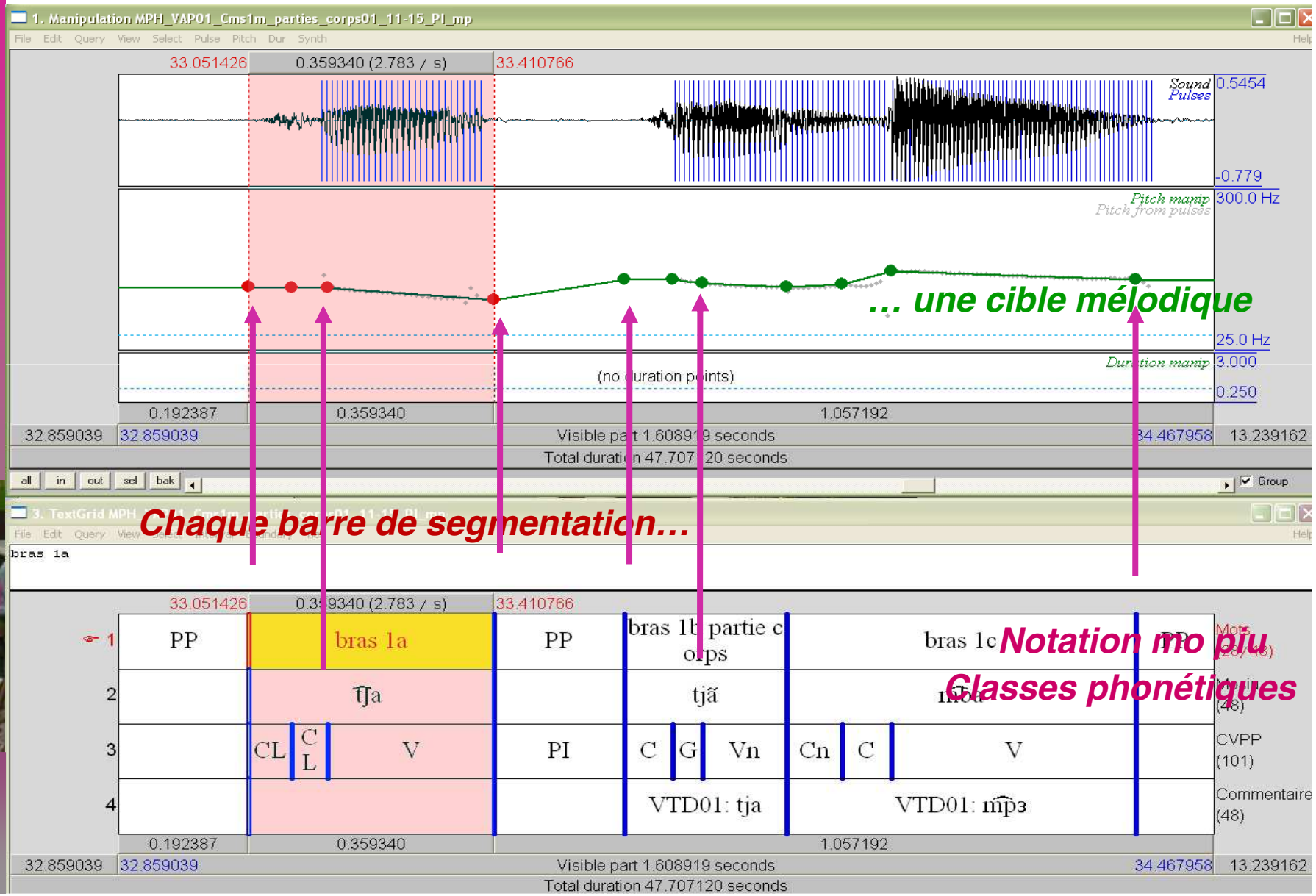
Le corpus

- **3 terrains en *2009, 2010, 2011***
- ***36h de parole (et video)***
  - 35h langue (+ 1h de chants traditionnels)
    - ***Parole continue***
      - questionnaire « culturel »
      - récits de vie
      - contes
      - commentaires de video (Projet Trajectoire – Fédération TUL-CNRS)
      - reconstitution d'histoire à partir d'images (conte de la grenouille)
    - ***Listes de mots***
      - Calmsea
      - 500 mots
      - Contextualisation des Noms (présentatif, classificateurs, etc.)

Corpus de l'étude présente

- **1 homme (VAP01) et une femme (VTD01) ont enregistré**
  - 44 mots français de la liste *Calmsea* (Matisoff 1978), parties du corps x 3 répétitions
  - En tout, ***418 unités lexicales mo piu, 4770 unités phonétiques transcrites***
    - 2541 voyelles
    - 2229 consonnes
- **Enregistrement réalisé par un ingénieur MICA**
  - ***2 caméras de face et profil***
  - 2 micros, 2 pistes (1- français + vietnamien, 2-mo piu)
  - le tout piloté par ordinateur
  - ***25 locuteurs : 8 femmes + 17 hommes***

Exemple de TextGrid sous Praat-Momel



Exemple de fichier xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	2 speakers Nb	1 speaker Nb	Word Nb	N° fichier	Locuteur	Mots	classes	mo piu	Syll mel	Derivee	Duree	Hz	G	A	B	C	Début	Fin	Commentaires
2	1	1	1	MPH	VAP01	ventre 1a	C	n͡ɲa	35	22.51	194	133	h			h	11.70	11.89	
3	2	2	1	MPH	VAP01	ventre 1a	Cn	n͡ɲa	54	-22.37	72	171	ɲ	ɲ	ɲ	ɲ	11.89	11.96	
4	3	3	1	MPH	VAP01	ventre 1a	V	n͡ɲa	43	-14.33	286	156	a	œ		œ	11.96	12.25	
5	4	4	1	MPH	VAP01	PP		n͡ɲa	33+	1.26	1114	133					12.25	13.36	
6	5	5	1	MPH	VAP01	ventre 2a	C	n͡ɲa	34	17.69	158	134	h		ɲ	h	13.36	13.52	
7	6	6	1	MPH	VAP01	ventre 2a	Cn	n͡ɲa	44-	-18.64	54	157	ɲ	ɲ	ɲ	ɲ	13.52	13.57	
8	7	7	1	MPH	VAP01	ventre 2a	V	n͡ɲa	43	-12.16	313	148	a	ɜ	a	œ	13.57	13.89	
9	8	8	1	MPH	VAP01	PP		n͡ɲa	34	2.56	1135	133					13.89	15.02	
10	9	9	1	MPH	VAP01	ventre 3a	C	n͡ɲa	44+	4.19	167	141	h	-	ɲ	h	15.02	15.19	
11	10	10	1	MPH	VAP01	ventre 3a	Cn	n͡ɲa	44+	3.33	66	146	ɲ	ɲ	ɲ	ɲ	15.19	15.25	
12	11	11	1	MPH	VAP01	ventre 3a	V	n͡ɲa	43	-11.65	309	148	a	ɜ	a	œ	15.25	15.56	
13	12	12	1	MPH	VAP01	PP		n͡ɲa	33+	0.02	1443	132					15.56	19.99	
14	13	13	2	MPH	VAP01	sang 1a	Cn	n͡ʃa	34	9.81	132	130	n	n	n	n	19.99	20.12	
15	14	14	2	MPH	VAP01	sang 1a	C	n͡ʃa	43	-50.59	77	140	t	t	t	t	20.12	20.20	
16	15	15	2	MPH	VAP01	sang 1a	C	n͡ʃa	34	53.79	87	112	ʃ	ʃ	ʃ	ʃ	20.20	20.29	
17	16	16	2	MPH	VAP01	sang 1a	C	n͡ʃa	-	-	-	-	-	h	h	-	-	-	
18	17	17	2	MPH	VAP01	sang 1a	V	n͡ʃa	43	-7.43	471	147	a	a	a	a	20.29	20.76	
19	18	18	2	MPH	VAP01	sang 1a	V	n͡ʃa	-	-	-	-	-	a	-	-	-	-	
20	19	19	2	MPH	VAP01	PP		n͡ʃa	33+	1.5	1131	120					20.76	21.89	
21	20	20	2	MPH	VAP01	sang 2a	Cn	n͡ʃa	33-	-8.2	134	133	n	n	n	n	21.89	22.02	
22	21	21	2	MPH	VAP01	sang 2a	C	n͡ʃa	33+	10.7	103	125	t	t	t	t	22.02	22.13	
23	22	22	2	MPH	VAP01	sang 2a	C	n͡ʃa	34	12.29	98	133	ʃ	ʃ	ʃ	ʃ	22.13	22.22	
24	23	23	2	MPH	VAP01	sang 2a	C	n͡ʃa	-	-	-	-	-	h	h	-	-	-	
25	24	24	2	MPH	VAP01	sang 2a	V	n͡ʃa	43	-6.33	426	142	a	a	a	a	22.22	22.65	
26	25	25	2	MPH	VAP01	sang 2a	V	n͡ʃa	-	-	-	-	-	a	-	-	-	-	
27	26	26	2	MPH	VAP01	PP		n͡ʃa	33+	1.34	895	121					22.65	23.55	
28	27	27	2	MPH	VAP01	sang 3a	Cn	n͡ʃa	33-	-1.75	114	130	n	n	n	n	23.55	23.66	
29	28	28	2	MPH	VAP01	sang 3a	C	n͡ʃa	34	10.42	77	129	t	t	t	t	23.66	23.74	
30	29	29	2	MPH	VAP01	sang 3a	C	n͡ʃa	44+	8.75	80	135	ʃ	ʃ	ʃ	ʃ	23.74	23.82	
31	30	30	2	MPH	VAP01	sang 3a	C	n͡ʃa	-	-	-	-	-	h	h	-	-	-	
32	31	31	2	MPH	VAP01	sang 3a	V	n͡ʃa	43	-12.29	314	141	a	a	a	a	23.82	24.13	
33	32	32	2	MPH	VAP01	sang 3a	V	n͡ʃa	-	-	-	-	-	a	-	-	-	-	
34	33	33	2	MPH	VAP01	PP		saa	34	1.02	4039	110					24.13	28.17	
35	34	34	3	MPH	VAP01	os 1b	C	saa	44+	8.51	329	140	s	s	s	s	28.17	28.50	VTD01: saã
36	35	35	3	MPH	VAP01	os 1b	C	saa	-	-	-	-	-	h	-	-	-	-	
37	36	36	3	MPH	VAP01	os 1b	V	saa	43	-25.19	143	164	a	a	ą̃	ã	28.50	28.64	VTD01: saã
38	37	37	3	MPH	VAP01	os 1b	C	saa	-	-	-	-	-	ʔ	-	-	-	-	
39	38	38	3	MPH	VAP01	os 1b	V	saa	33-	-8.54	375	133	a	a	a	ã	28.64	29.02	VTD01: saã
40	39	39	3	MPH	VAP01	PP		saa	33+	2.46	1137	111					29.02	30.15	
41	40	40	3	MPH	VAP01	os 2b	C	saa	34	11.75	247	130	s	s	s	s	30.15	30.40	VTD01: saã

Étiquetage des 4 experts

Les buts de l'étude : 1- évaluer la variabilité inter-experts

- **Etude intéressante car les 4 experts n'ont pas la même expérience / culture linguistique**
  - Culture linguistique
    - ⋆ ***3 appartiennent au domaine de la parole et manipulent les technologies de la parole***
    - ⋆ ***2 ont une pratique des langues asiatiques*** et/ou une solide formation dans le domaine des langues orientales
    - ⋆ ***1 est de culture franco-vietnamienne***
  - Méthodes d'analyse différentes : technologies de la parole ou pas
- **Question : est-ce que cette « culture » différente se ressent au niveau des étiquetages et de la variabilité inter-experts?**

Les buts de l'étude : 1- évaluer la variabilité inter-experts (suite)

- **Pour trancher la question, nous avons réalisé**
  - ***2 matrices de confusion pour les voyelles et consonnes***
- **Une matrice de confusion est un tableau à 2 entrées où sont consignées**
  - sur l'***axe horizontal toutes les unités*** rencontrées dans l'ensemble de nos étiquetages
    - classées par trait articulatoire et/ou acoustique proche
  - sur l'***axe vertical, toutes les unités étiquetées*** par chaque expert
- **dans l'idéal, si les experts avaient été totalement concordants**
  - il n'existerait aucune dispersion
  - c'est-à-dire que ***seule la diagonale serait remplie et elle seule***
- **Dans la pratique, si les experts montrent une bonne concordance**
  - Les ***unités les plus nombreuses*** se trouvent ***dans la diagonale***
  - La ***dispersion*** est localisée dans la ***même classe articulatoire ou acoustique***

Exemple de la matrice de confusion des consonnes

NB	p	t	k	b	d	g	ʈ	q	f	s	ʃ	z	ʒ	ɕ	j	h	m	n	ŋ	ɲ	l	l̪
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										
ʒ													11		1							
ɕ										8	7			20	13	3		1				
j														1	67							
h																87		2				
m																	163		1			
n																	4	159	3	1		
ŋ																		3	9			
ɲ															3	3	1	8	3	48		
l		2														1			1		164	
l̪		1																			4	4

Exemple de la matrice de confusion des consonnes *- en pourcentage -*

%	p	t	k	b	d	g	ʈ	q	f	s	ʃ	z	ʒ	ɕ	j	h	m	n	ŋ	ɲ	l	l̃
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																
ʈ	4	44					48									2						
q	6	14	16					59	2									2				
f			1						85							13						
s										80	3			2		13						
ʃ											100											
z												100										
ʒ													91		8							
ɕ										15	13			38	25	5		1				
j														1	98							
h																97		2				
m																	99		0.6			
n																	2	95	1	0.6		
ŋ																		25	75			
ɲ															4	4	1	12	4	72		
l		1														0.6			0.6		97	
l̃		11																			44	44

But 2 : unifier nos transcriptions

- **Méthode de codage phonétique**
  - chaque expert a transcrit les mots en ***codage API***
    - soit ***analyse auditive et visuelle*** en utilisant un logiciel de segmentation et étiquetage de la parole ***(PRAAT-MOMEL, MISTRAL+)***
    - soit ***simple analyse auditive***
  - Un fichier xls issu automatiquement de MISTRAL+ (MICA) a permis de rassembler les 4 étiquetages
    - présentation par locuteur / mot 1… 44 (x 3 répétitions) / unités phonétiques successives
    - avec tout un ensemble de paramètres prosodiques manuels et automatiques
    - une analyse tonale a été menée par ailleurs (cf actes de TAL 2012)
  - ***1915 unités phonétiques*** ont été ainsi consignées dans le fichier xls (et 4470 annotées)

MISTRAL+

- ***M**elody **I**ntonation **S**peaker **T**onal **R**ange **A**nalysis using variable **L**evels*
  - Réécriture complète et ajout de *nouvelles fonctionnalités* depuis les scripts INTSMEL (2004), MESLIM (2008-2011)
  - Dédié à l'analyse de la ***prosodie expressive / émotionnelle***
  - ***Annotation automatique des cibles mélodiques***
  - Permet d'étudier
    - Les modulations de F0, les dérivées, les durées aux niveaux phonétique, mélodique et lexical
    - Parole expressive, émotionnelle, attitudinale de n'importe quel langage : accentué, stressé, tonal ou pas
    - Le ***système tonal*** (d'autant plus précieux quand la langue est inconnue) par un ***étiquetage automatique*** après délimitation de l'unité tonale (syllabe, voyelle…)
  - ***Indépendant***
    - de la structure de la langue
    - de toute théorie linguistique
  - Permet d'***épargner beaucoup de temps*** dans les tâches fastidieuses

Méthode pour la convergence des notations phonétiques

- **Comme déjà mentionné, on lit sur la matrice de confusion / dispersion une *grande cohérence dans les résultats***
  - ***Peu de dispersion***
  - La grande majorité des dispersions appartiennent à ***la même classe articulatoire ou acoustique***

- **Consignes pour la notation**
  - ***pour une unité phonétique donnée dans un « mot »…***
  - ***…une notation identique par les 4 étiqueteurs est requise…***
  - ***…pour accorder le statut de phonème***

Quelques exemples de variabilité inter-experts

- **Tradition de notation différente**
  - ***Les 3 experts en parole ont une notation plus phonétique et acoustique***
  - ***Le 4e expert a une notation plus phonologique et systémique***
- **Divergences de notation entre experts**
  - **L'*assourdissement des nasales et liquides***
    - n̥ / hn
    - l̥ / hl / ɬ
  - Le ***marquage des sons complexes***
    - : tʃ / tç / tɕ
  - Le ***marquage de la glottalisation***
    - Segment
      - marquage sous la voyelle [a̰]
      - ajout d'une syllabe /consonne glottale [aʔa]
    - Relation au ton ?

Quelques particularités partagées avec les langues Hmong

- **Nasalisation et glottalisation très présentes**
  - Les 2 locuteurs

- **Voix craquée et glottalisée**
  - Les analyses technologiques montrent une différence entre
    - ⋆ Une ***voix craquée où la structure périodique est préservée malgré le ralentissement de la période***
    - ⋆ Une ***voix glottalisée où la structure n'est pas préservée***
  - ⇨ ***Cette distinction est-elle acoustique (variabilité locuteur / contextuelle) ou phonologique ?***

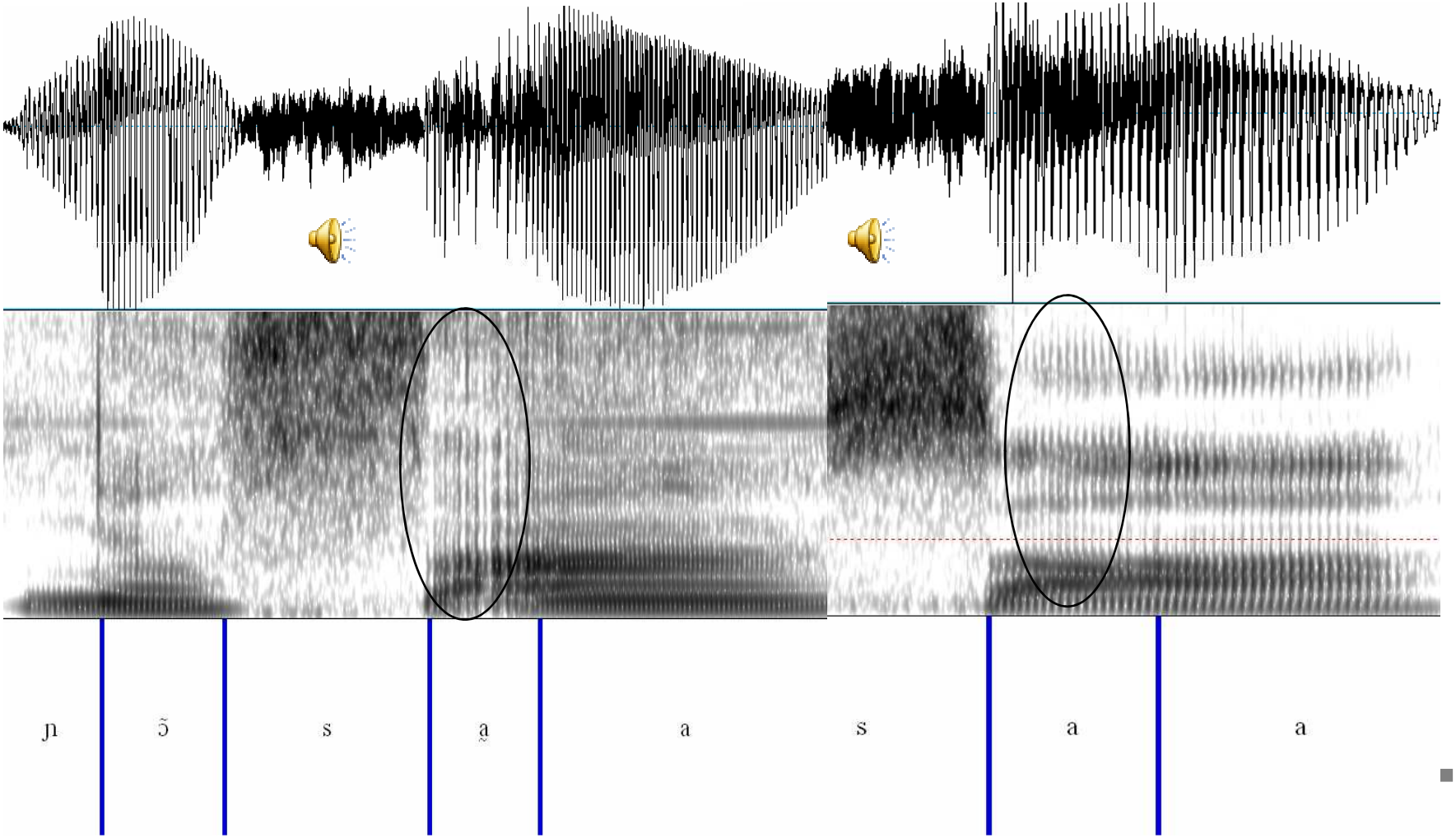
Quelques particularités partagées avec les langues Hmong

- ***Concernant la glottalisation, il existe à ce jour une trop grande variabilité entre les deux locuteurs étudiés pour trancher le problème***

Exemple: **classificateur [ɲɔ̃] os [saa]**

La locutrice produit une *voyelle glottalisée*

Le locuteur produit une *voyelle modale*



Vers une analyse phonologique du mo piu

- En proposant
  - une analyse ***phonétique unifiée*** des transcriptions et annotations
  - de structurer les données en ***système***
  - de poser des hypothèses mieux établies sur le système phonologique de la langue
  - de ***progresser plus rapidement*** dans notre étude de cette langue non connue

Conclusion

- **Validation mot après mot, unité phonétique après unité phonétique du contenu articulatoire et acoustique**
- **grâce à la sélection des codages identiques inter-étiqueteurs**
- **Cette analyse remplit 2 objectifs**
  - Dans le domaine de la ***linguistique de terrain***
    - premier stade de la description d'une langue en danger
  - Dans le domaine des ***technologies automatiques de la parole***
    - le ***mo piu*** étant une langue inconnue…
    - …sert de ***modèle pour valider nos méthodes de segmentation et annotation automatiques***
    - segmentation et annotation automatiques à partir de l'ensemble des unités phonétiques de plusieurs langues (« modèles de langue »)

Ainsi de manière assez surprenante en confrontant les diverses associations d'unités appartenant aux mandarin, vietnamien, khmer, anglais et français, c'est la combinaison mandarin + français qui est la meilleure (présence de voyelles nasales en français)

Conclusion

- Merci de votre attention
- Thanks for your attention

Geneviève Caelen-Haumont

Alice Vittrant

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

Jean-Cyrille Ly Van Tu

Katarina Bartkova