ភាសាខ្មែរ

«ការសិក្សាភាសាខ្មែរនៅសម័យបច្ចេកវិទ្យាបញ្ញាសិប្បនិម្មិត៖ ផ្ដើមចេញពីសូរវិទ្យាទៅដល់ប្រព័ន្ធសម្គាល់សម្ដី»

Khmer Linguistic Research in the Age of Artificial Intelligence:

From Phonetics to Automatic Speech Recognition

ថែម វត្ថុ

ថ្ងៃទី៤ ខែមិថុនា ឆ្នាំ២០២៦ • សាលទន្លេសាប អគារខេមរវិទូ • រាជបណ្ឌិត្យសភាកម្ពុជា

DRAFT

មាតិកា

Contents

ផ្នែកទី ១
ប្រវត្តិភាសាខ្មែរ
Mon-Khmer • បុរាណ–ទំនើប

ផ្នែកទី ២
ភាសាខ្មែរជាអ្វី?
សូរវិទ្យា វាក្យស័ព្ទ សម្ព័ន្ធ

ផ្នែកទី ៣
រចនាសម្ព័ន្ធសូរសព្ទ
ព្យញ្ជនៈ • ស្រៈ

ផ្នែកទី ៤
ភាសាខ្មែរនៅសម័យ AI
ទិន្នន័យសំឡេង • សំណេរ

ផ្នែកទី ៥
សេចក្ដីសន្និដ្ឋាន
យុទ្ធសាស្ត្រ និងការស្រាវជ្រាវ

DRAFT

ប្រធានបទ៖ ការសិក្សាភាសាខ្មែរនៅសម័យបញ្ញាសប្បនិម្មិត — ផ្ដើមចេញពីសូរវិទ្យាទៅដល់ប្រព័ន្ធសម្គាល់សម្ដី

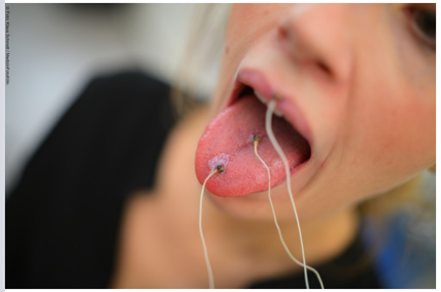
អ្វីដែលខ្ញុំនឹងមិននិយាយលម្អិត

- ពណ៌នា អំពីសំណេរ (សរសេរខុស សរសេរត្រូវ)
- វេយ្យាករណ៍ភាសាខ្មែរ
- ការរៀន និងបង្រៀន ភាសាខ្មែរ
- ប្រវត្តិសាស្ត្រ ចិត្តសាស្ត្រ និងនយោបាយ

DRAFT

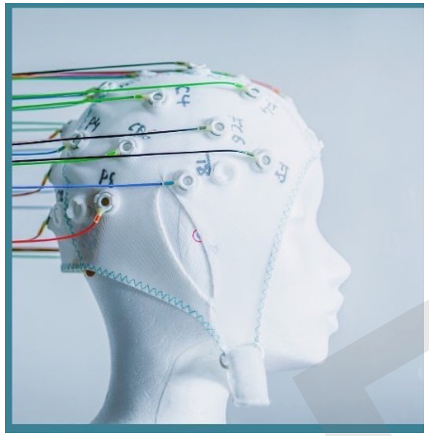
មន្ទីរពិសោធន៍ សូរវិទ្យានៅសាកលវិទ្យា កូឡូញ

Phonetics Lab, IFL, University of Cologne, Germany, Universität zu Köln, Deutschland



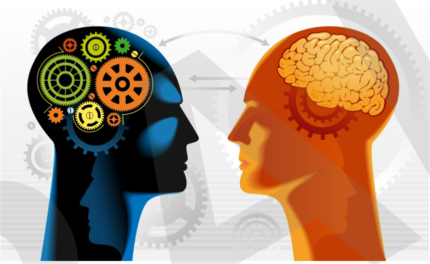
Electromagnetic articulography (EMA)

មន្ទីរពិសោធន៍សូរវិទ្យា



EEG (electroencephalogram)

ប្រសាទវិទ្យា



Statistics, probabilities, ML, DL

វិទ្យាសាស្ត្រកុំព្យូទ័រ និង AI

Speech Therapy, Autism speech

UniKlinik

ចិត្តវិទ្យា

-មុខវិជ្ជាមួយ មិនឯករាជ្យទៀតទេ
-ការសិក្សា ជាអន្តរវិស័យ
Interdesciplinary
-មុខវិជ្ជាមួយ មានមន្ទីរពិសោធន៍. មួយដែរ

សេចក្តីសន្និដ្ឋាន

ប្រធានបទ៖ ការសិក្សាភាសាខ្មែរនៅសម័យបញ្ញាសប្បនិម្មិត — ផ្តើមចេញពីសូរវិទ្យាទៅដល់ប្រព័ន្ធសម្គាល់សម្តី

DRAFT

All of these photos are licensed under CC BY-SA-NC

សន្និសីទ ជាតិ និងអន្តរជាតិ

SEALS៖៤ លើក ICAAL៖៤ លើក ICPHS៖ ១ លើក Geo-Ling៖ ១ លើក

Flows and Currents៖ ១ លើក និង ថ្នាក់ជាតិ៖ ៣ លើក

2025

-Flows and Currents in Southeast Asia Conference — London, England (3–4 July 2025)
-SEALS 34 (34th Southeast Asian Linguistics Society) — Bali, Indonesia (11–13 June 2025)
-Data Competence College by DKZ.2R — IT Center Aachen, Germany (27–28 March 2025)

2024

-SEALS 33 — National Tsing Hua University, Hsinchu, Taiwan (15–17 June 2024)

2023

-ICAAL 11 (11th International Conference on Austroasiatic Linguistics) — Chiang Mai, Thailand (26–27 October 2023)
-SEALS 32 — Chiang Mai University, Thailand (16–18 May 2023) *Funded by KAAD, Germany*

2022

-ICAAL 10 — Online, hosted by Living Tongues Institute for Endangered Languages, USA (30 November–2 December 2022)
-SEALS 31 — University of Hawaii at Manoa, USA (18–20 May 2022) *Funded by NSF, USA*

2019

-ICAAL 8 — Chiang Mai, Thailand (August 2019)
-ICPhS 2019 (International Congress of Phonetic Sciences) — Melbourne, Australia (August 2019)
-National Conference on Khmer Transcription and Transliteration — Phnom Penh, Cambodia

2017

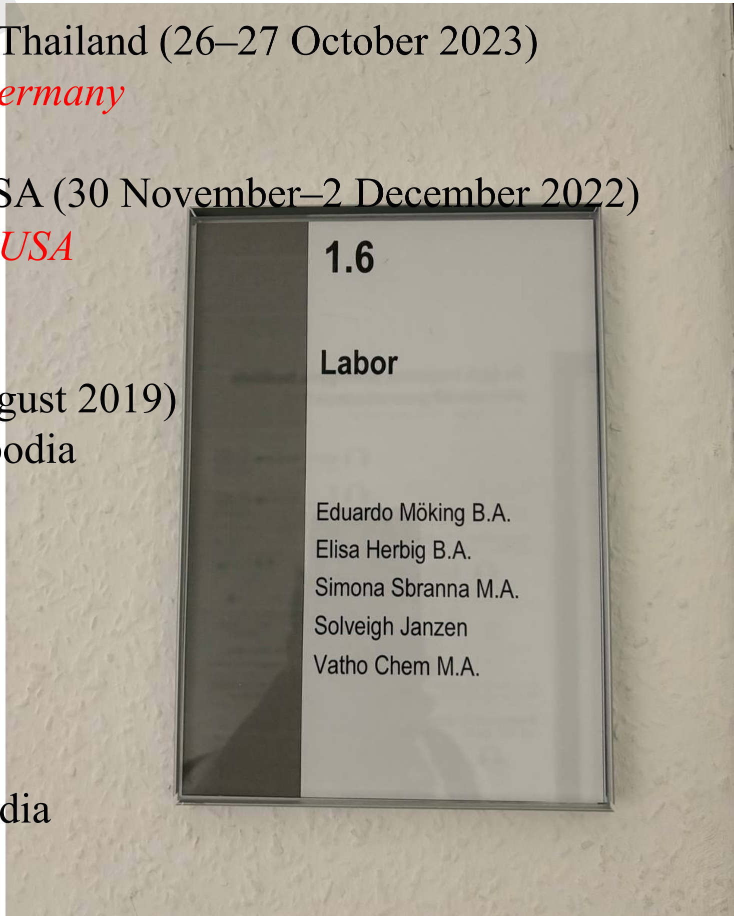
-National Conference on Khmer Language Diversity — Phnom Penh, Cambodia

2016

-3rd International Conference on Geo-Linguistics of Asia — Phnom Penh, Cambodia
-National Conference on Khmer Language — Phnom Penh, Cambodia

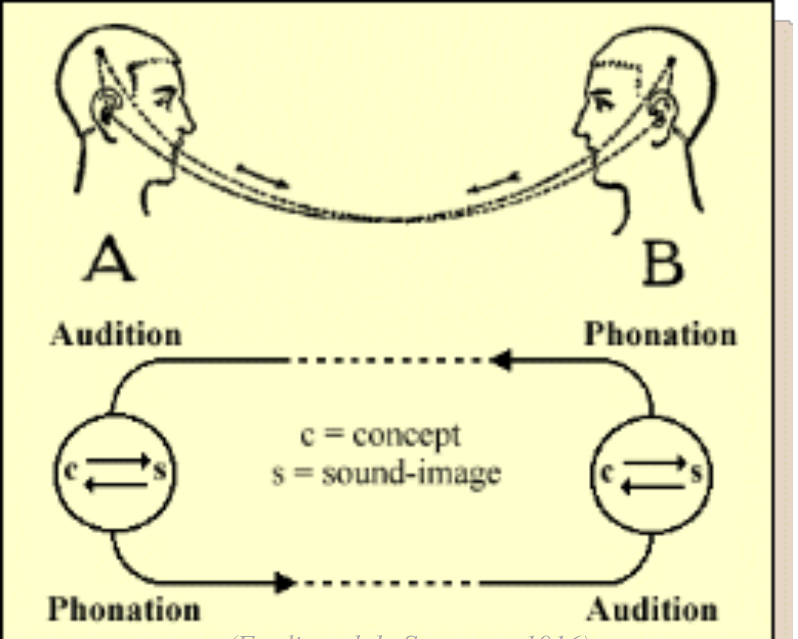
2015

-ICAAL 6 (International Conference on Austroasiatic Linguistics) — Siem Reap, Cambodia



DRAFT

ទ្រឹស្តីទំនាក់ទំនងគ្នា



(Ferdinand de Saussure, 1916)

ការទាក់ទងគ្នា ដោយប្រើប្រាស់ ភាសា យល់គ្នាបាន ៣០%

(Albert Mehrabian 1981, David Crystal1987)

»កែវភ្នែក

»កាយវិការ ដៃ ត្របកភ្នែក ជញ្ជើម ក្បាល បបូរមាត់

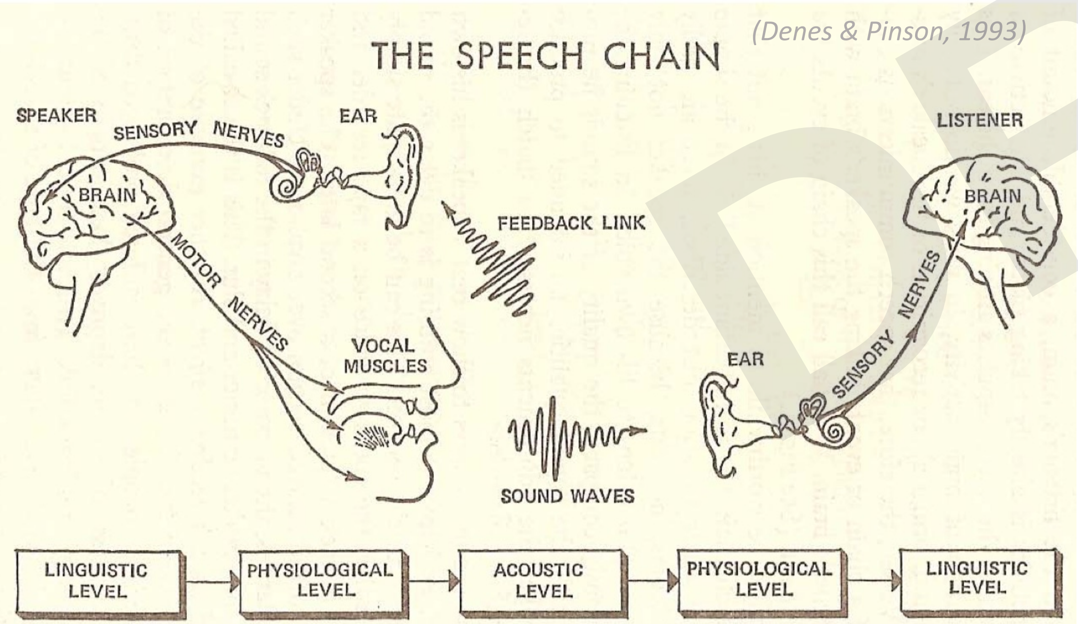
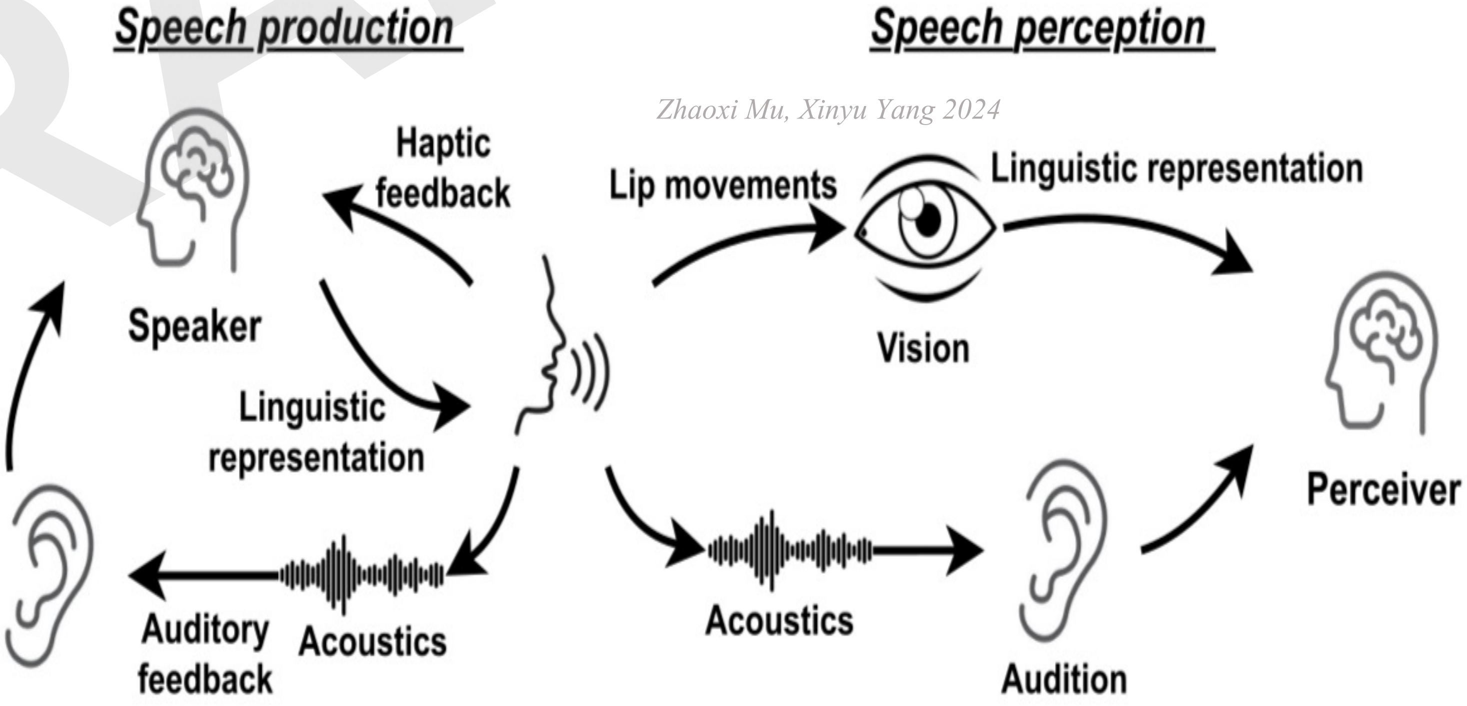


Fig. 1. The Speech Chain: the different forms in which a spoken message exists in its progress from the mind of the speaker to the mind of the listener.



DRAFT

Martine Grice

PhD · Professor at University of Cologne

The perception of post-focal prominence is (also) a matter of linguistic structure

August 2025 · Journal of the International Phonetic Association

DOI:10.1017/S0025100325100479

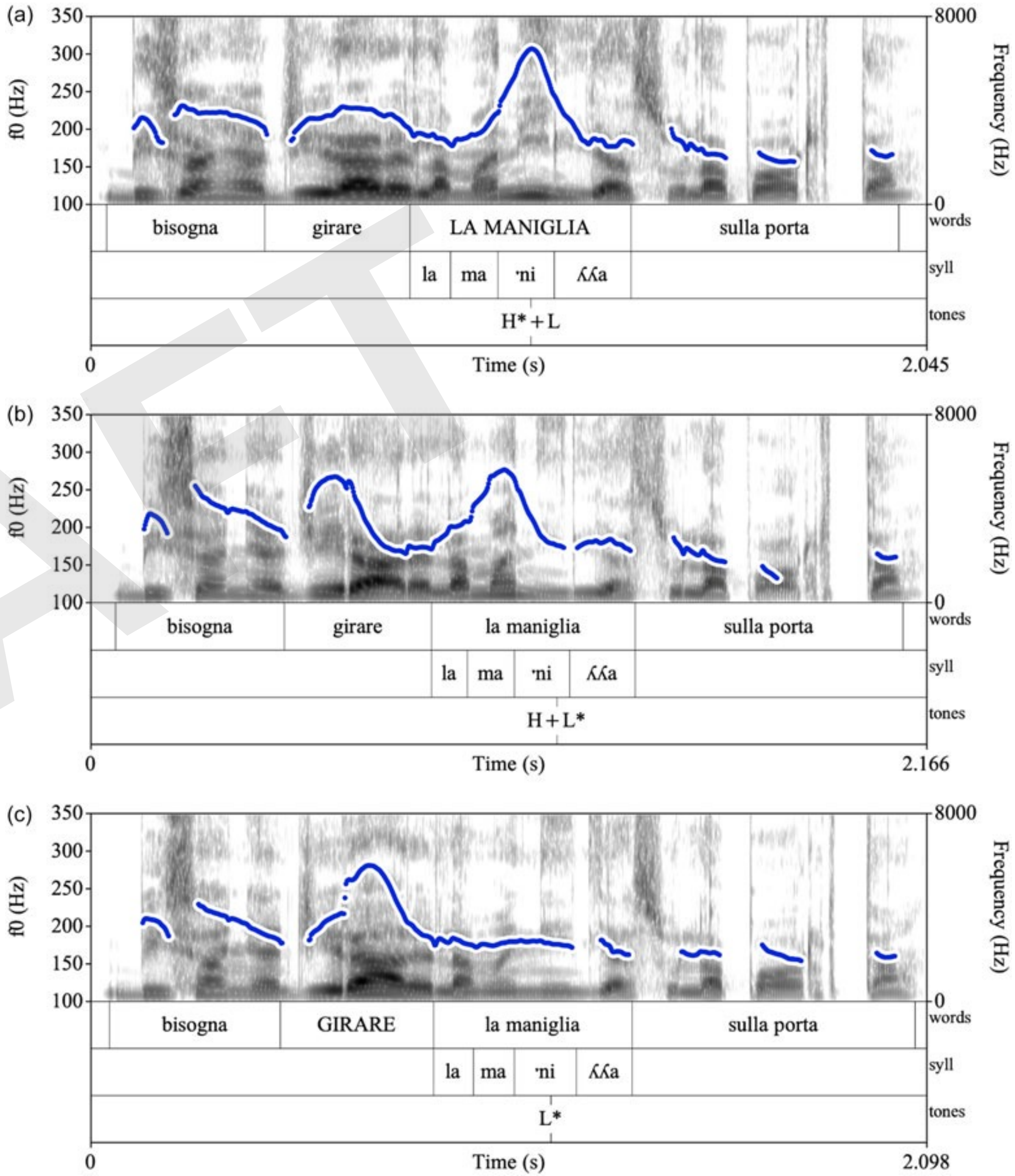
License · CC BY 4.0

Authors:

Martine Grice **Michelina Savino** **Maria Lialiou**

Aviad Albert **Petra B Schumacher**

Narrow Focus, Broad focus, post-focal prominence



DRAFT

Quantity-sensitivity affects recall performance of word stress

September 2024
DOI:10.21437/Interspeech.2024-182
Conference: Interspeech 2024

Authors:

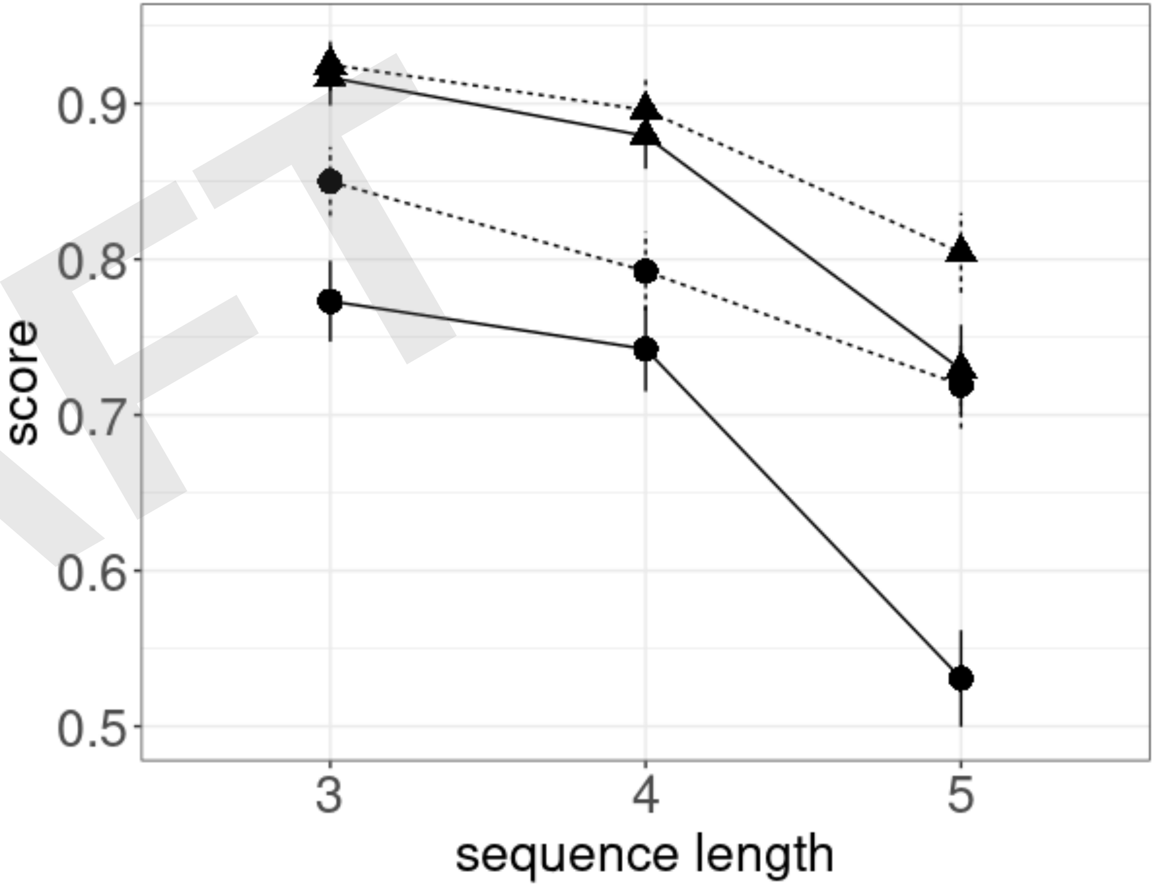
Constantijn Kaland

Maria Lialiou
University of Cologne

DRAFT

Table 3: *Overview of the factors and effects in the final GLMM.*

	β	SE	z	$p <$
(Intercept)	0.96	0.25	3.79	0.001
lg:Greek	1.16	0.37	3.18	0.01
contrast:ph	0.75	0.14	5.48	0.001
seq.length: 4-3	-0.39	0.19	-2.00	0.05
seq.length: 5-4	-0.96	0.18	-5.23	0.001
lg:Gr*contr:ph	-0.42	0.21	-2.00	0.05



Effect = change
No effect = no change

Simon Wehrle

PhD · PostDoc at University of Cologne

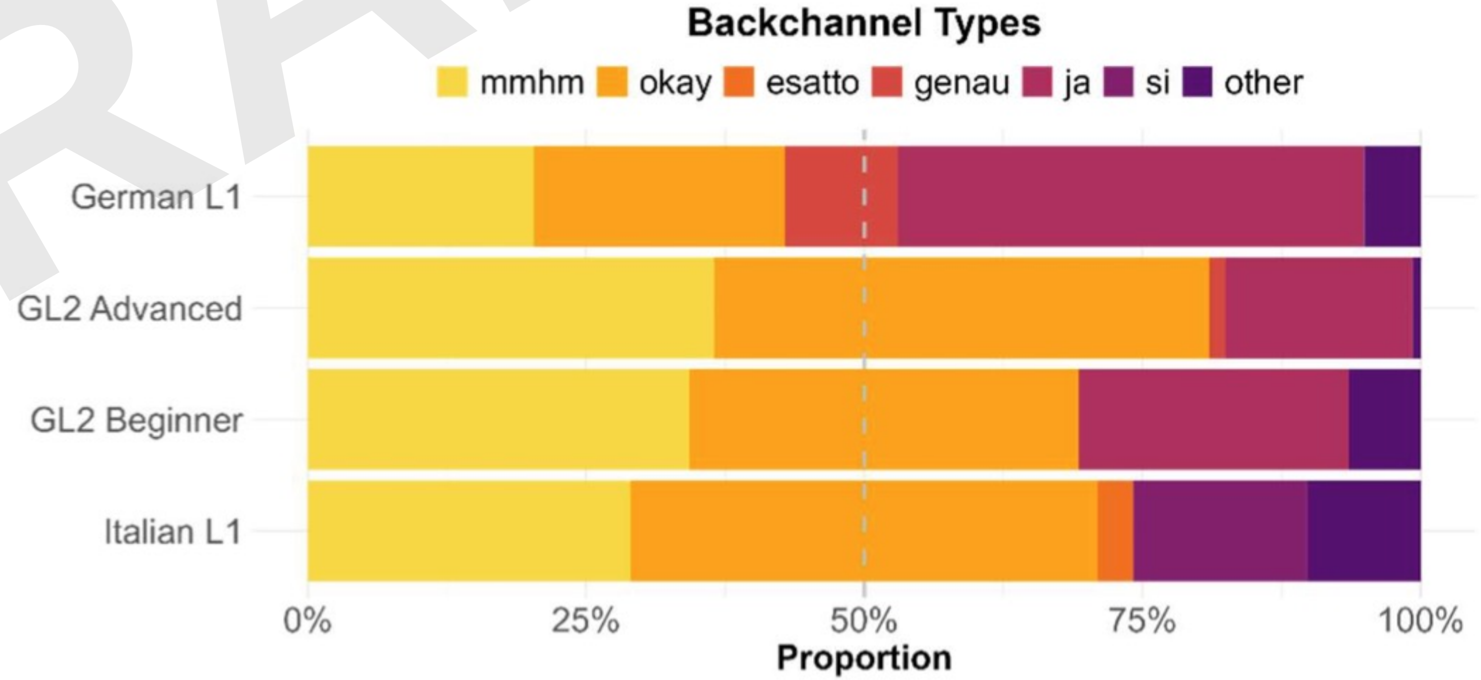
Currently Alexander von Humboldt Postdoctoral Fellow at Constantine the Philosopher University in Nitra, Slovakia

Backchannels in conversations between autistic adults are less frequent and less diverse prosodically and lexically

Published online by Cambridge University Press: **21 June 2023**

Simon Wehrle, Kai Vogeley and Martine Grice

Figure 5 - *Backchannel Type. Proportions of BC types are shown in percentages on the x-axis. Language groups are shown on the y-axis and are each assigned a bar. The mostly used BC types are listed in the legend and are colour-coded. The category "other" refers to types that rarely occurred.*



DRAFT

Measuring intonation style

Assessing the Intonation Style of Speakers with Autism Spectrum Disorder

Simon Wehrle[1], Francesco Cangemi[1], Harriet Hanekamp[1], Kai Vogeley[2,3] & Martine Grice[1]

New evidence for melodic speech in Autism Spectrum Disorder

Simon Wehrle[1], Francesco Cangemi[1], Kai Vogeley[2] & Martine Gr

OSF HOME

A brief tutorial for using Wiggliness and ... | Metadata | Files | Wiki

A brief tutorial for using Wiggliness and Spaciousness to measure intonation styles

Contributors: Simon Wehrle

| Last Updated: 2024-11-19 10:53 AM

Examining melodiousness in sarcasm: Wiggliness, spaciousness, and contour clustering

Csilla Tatár[1], Jonathan R. Brennan[1], Jelena Krivokapić[1], Ezra Keshet[1]

The interlocutor factor and intonational variation in collaborative dialogues

Tatiana Kachkovskaia

From wiggles to words: Acoustic and linguistic patterns in caregiver–child museum interactions

Haley McCreight; Jill C. Thorson

DRAFT

Individual variability in the attention-orienting function of edge tones in German: A pupillometry study

May 2026 · Laboratory Phonology 17(1)

DOI:10.16995/labphon.24022

Authors:

Maria Lialiou **Jesse A. Harris** **T. Mark Ellison**

Petra B Schumacher **Martine Grice**

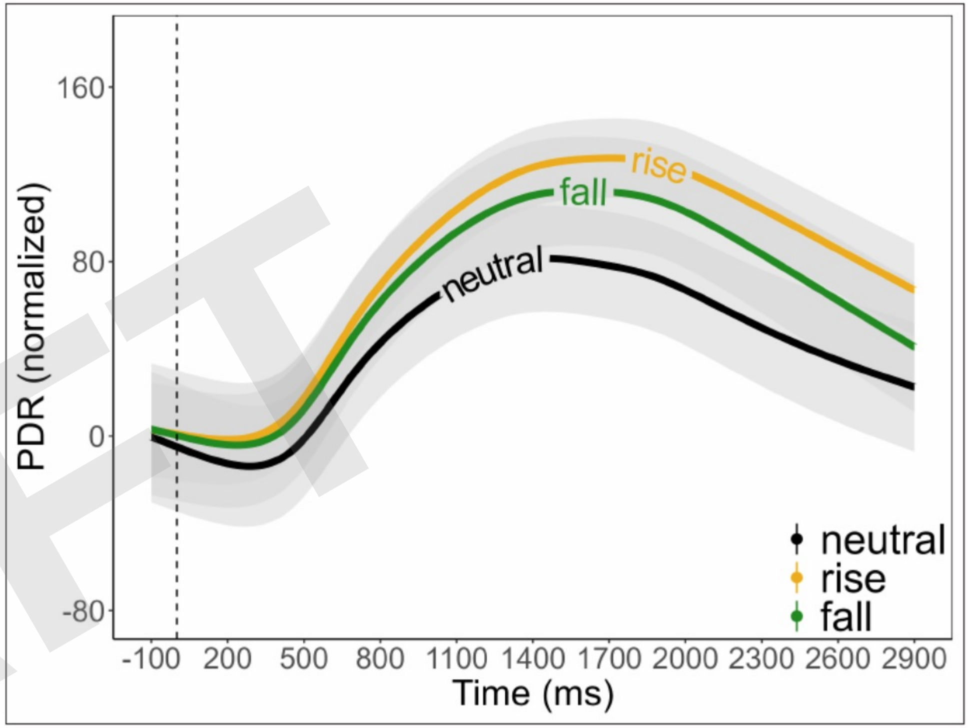


Figure 7: GAMM smooths of pupil dilation response (PDR; normalised) with 95% CIs, across prosodic conditions. PDRs are time-locked to the onset of the deviant (zero ms), as indicated by the vertical dashed line.

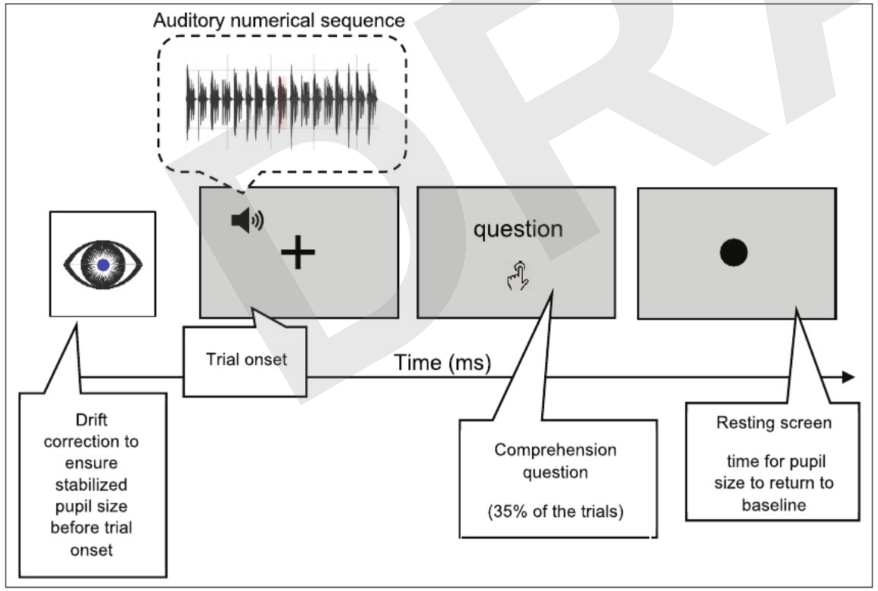


Figure 5: Schematic illustration of an experimental trial in the pupillometry.

DRAFT

Phonetic reduction in Central Khmer

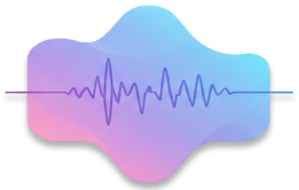
1. Deep transcription for the Khmer language: From phonetics to Automatic Speech Recognition
2. Phonetic reduction in Central Khmer
3. Reviewing Fundamental Frequency (f0) Estimation Algorithms

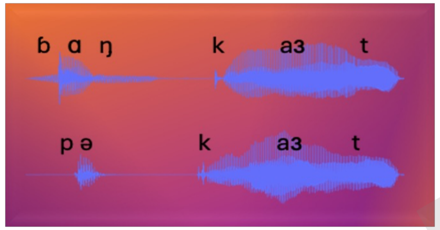
Universität zu Köln

DRAFT

Als Ergebnis der Prüfung wurde als Gesamtnote das Prädikat *magna cum laude* ("sehr gut" - 0,6) vergeben.

Phonetic Reduction in Central Khmer





Segmental Reduction

Vowel shortening

Consonant reduction

- Trill consonant reduction in C1 and C2 positions
- The reduction of voiced consonants (breathy voice)

Syllable reduction

Minor syllable

Pre-syllable

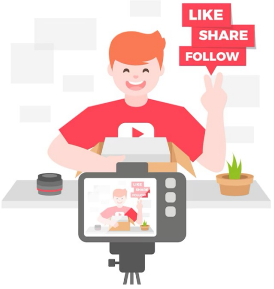
Word-level reduction

Reduction of content and function words

Lab vs. Naturalistic speech

Slow vs. Fast speech

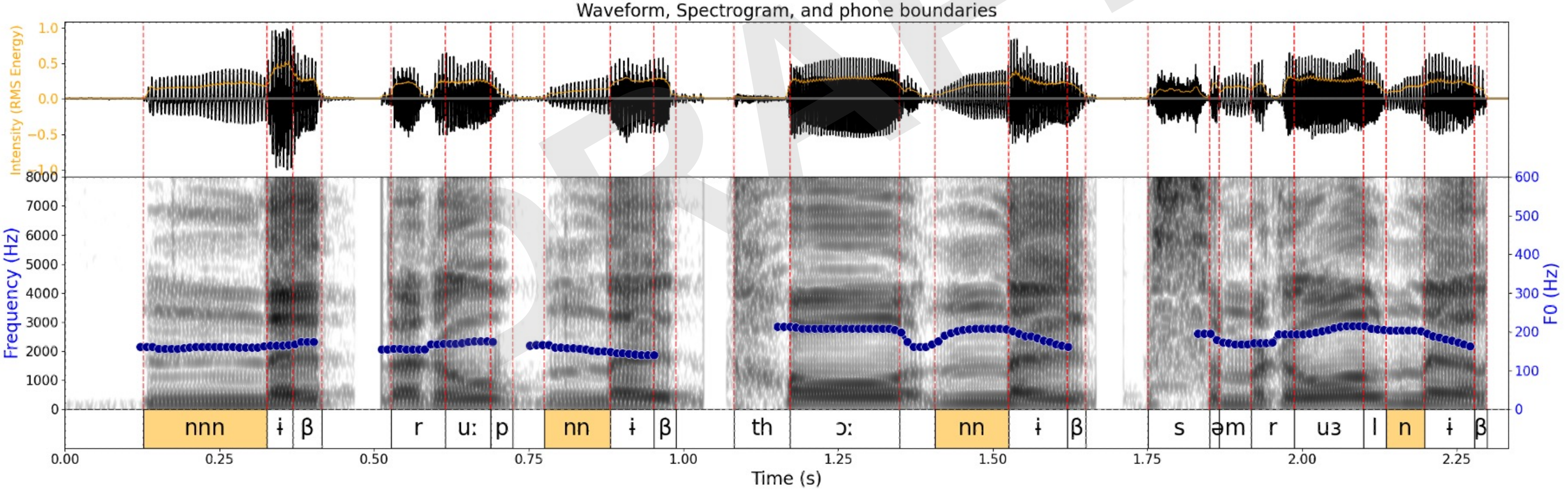
Careful vs. Casual speech



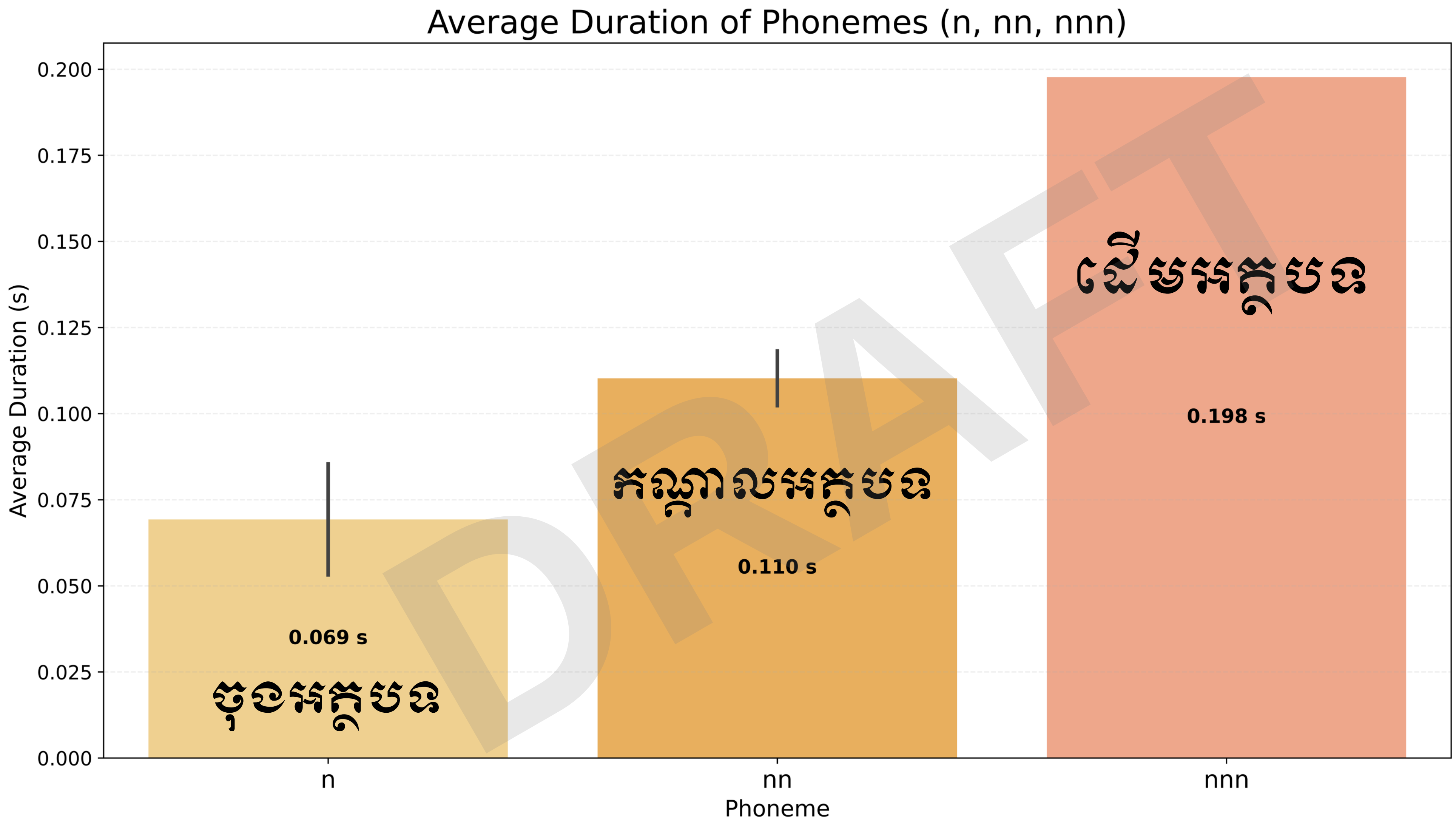
DRAFT

Temporal reduction: An overview

The word នៅ [nɨβ] "at; stay" in different positions, as extracted from an audio file, was taken from a Khmer news broadcast corpus featuring a male news reporter from Phnom Penh working at Radio France International: Khmer. There are at least three variations of the word នៅ [nɨβ] "at; stay" at the introduction of the reading passage (it was found at the beginning of the sentence), in the body (it was found at the beginning and in the middle of the sentence), and in the conclusion.



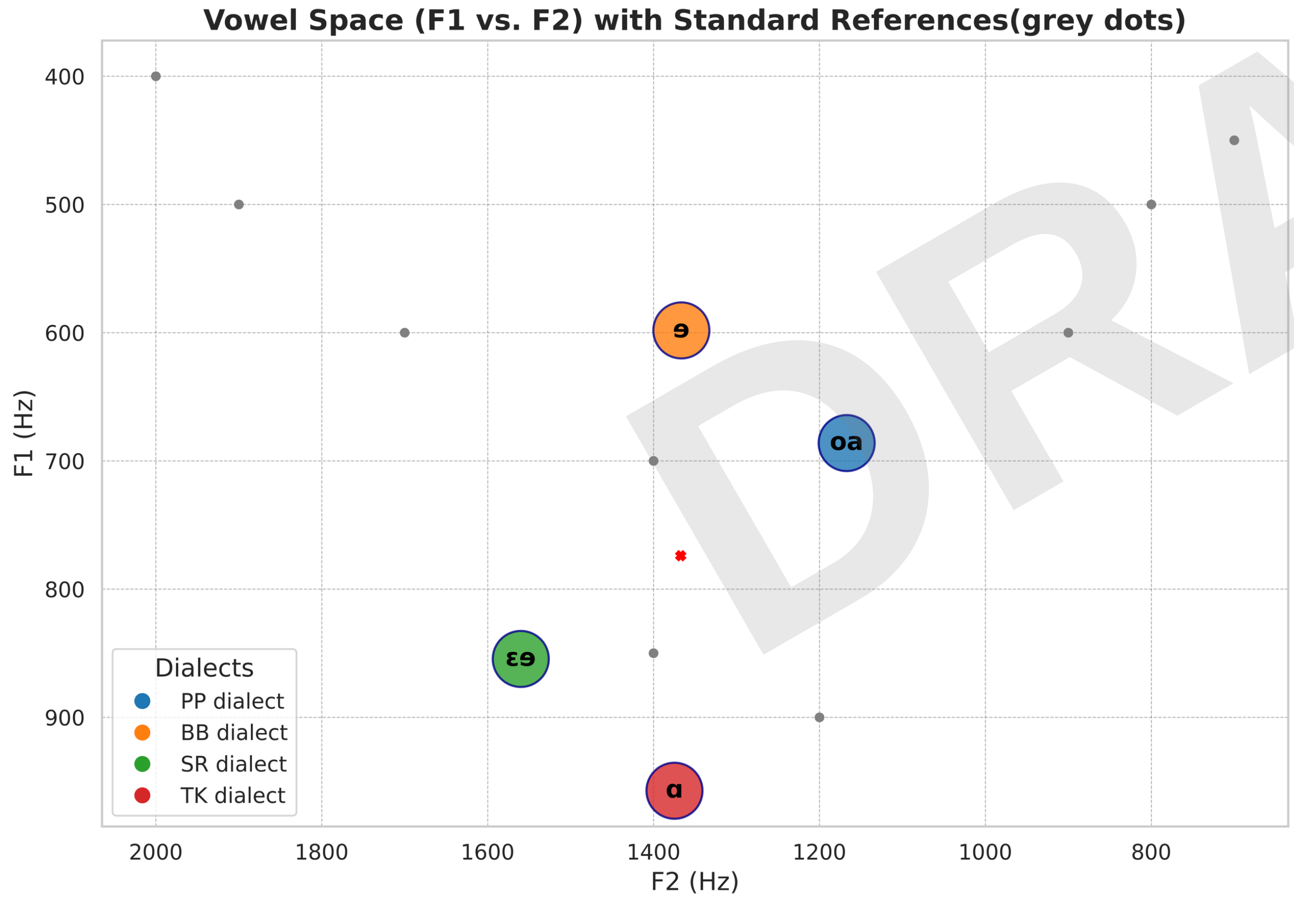
DRAFT

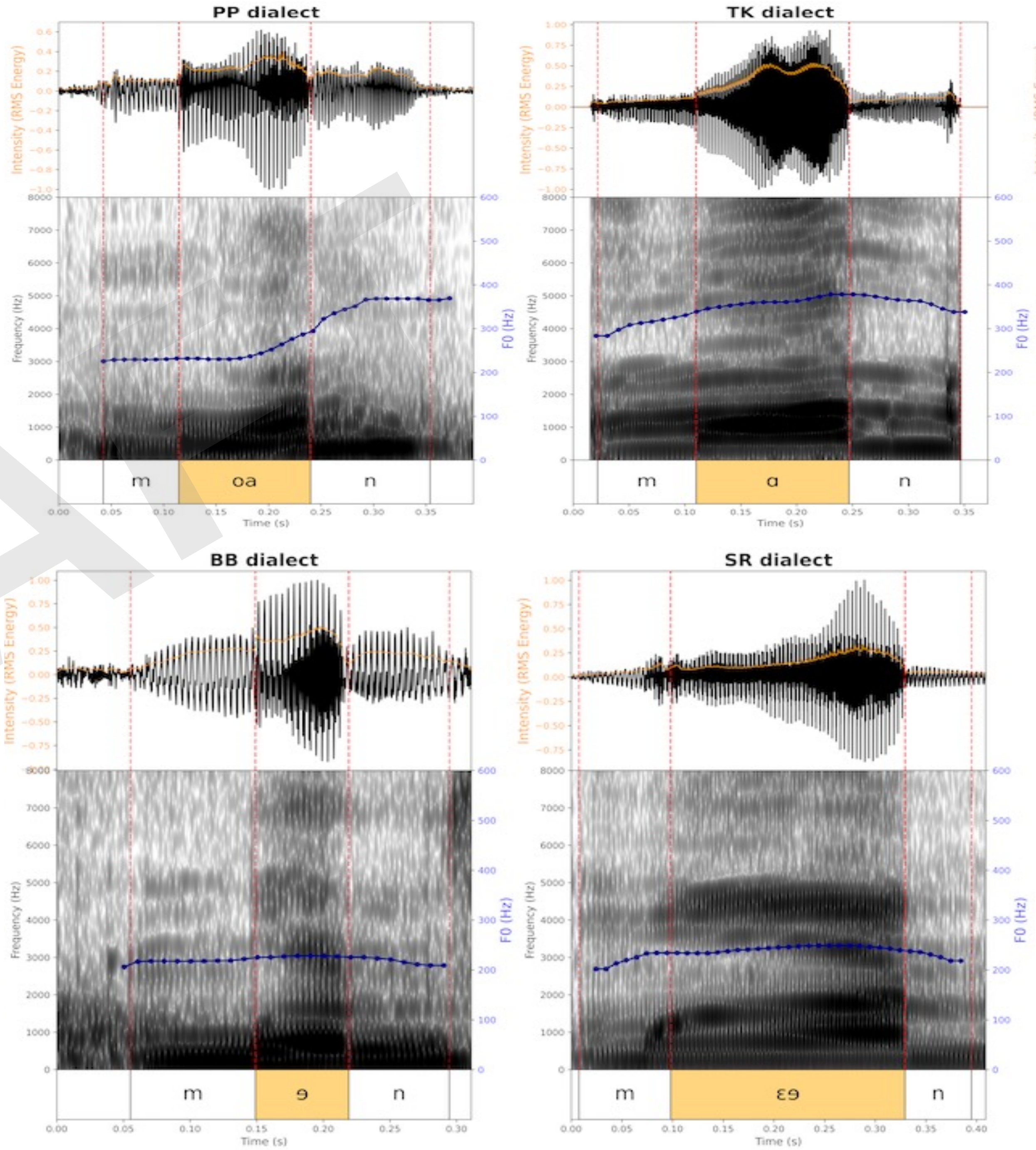


Average Duration of Phonemes (n, nn, nnn)
Average Duration (s)
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
DRAFT
ដើមអក្ខបទ
0.198 s
កណ្ដាលអក្ខបទ
0.110 s
0.069 s
ចុងអក្ខបទ
n
nn
nnn
Phoneme

Variation

Word	ST	PP	TK	BB	SR
មាន់	mɔːn	moan	mɑn	mən	mɛən





DRAFT

ការសក្សា ភាសា នៅអឺរ៉ុប

១. Cognitive and Generative Linguistics

២. **Computational and Corpus Linguistics**

៣. Sociolinguistics and Discourse Analysis

៤. Neurolinguistics and Psycholinguistics

៥. Probabilities and Statistics

DRAFT

ប្រវត្តិភាសាខ្មែរ

០១

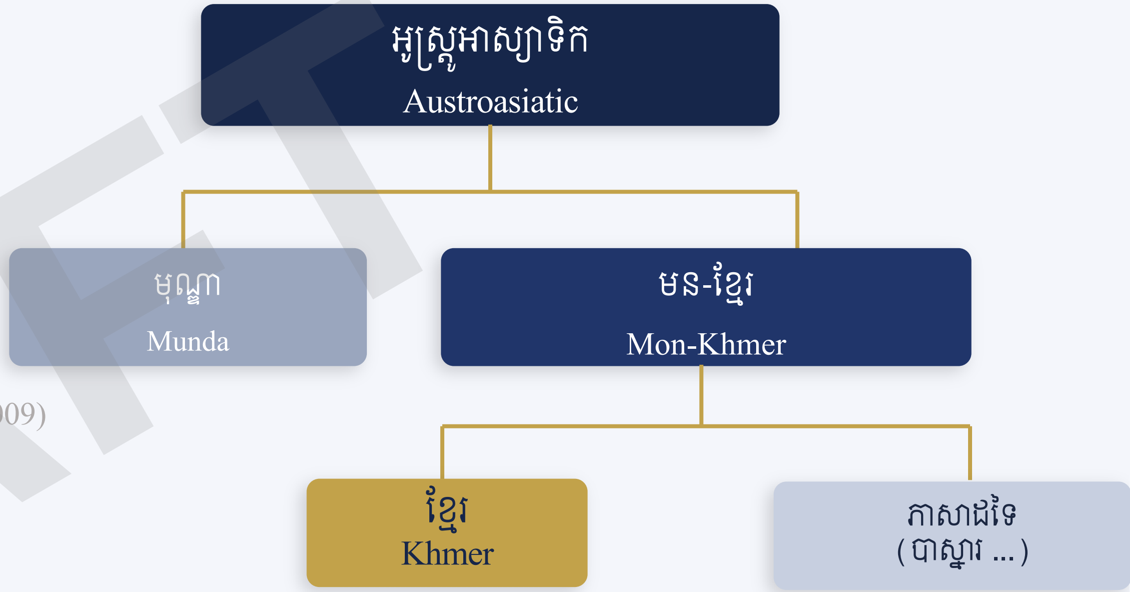
PART 01 · HISTORY OF THE KHMER LANGUAGE

DRAFT

ភាសាខ្មែរក្នុងគ្រួសារអូស្ត្រូអាស្យាទិក

ភាសាខ្មែរជាសមាជិកនៃអនុគ្រួសារ មន-ខ្មែរ (Mon-Khmer) ក្នុងគ្រួសារភាសាអូស្ត្រូអាស្យាទិក (Austroasiatic, AA) ដែលជាគ្រួសារភាសាចំណាស់ជាងគេមួយនៃតំបន់អាស៊ីអាគ្នេយ៍ដីគោក និង ឥណ្ឌា។ (Schmidt, 1904-6)

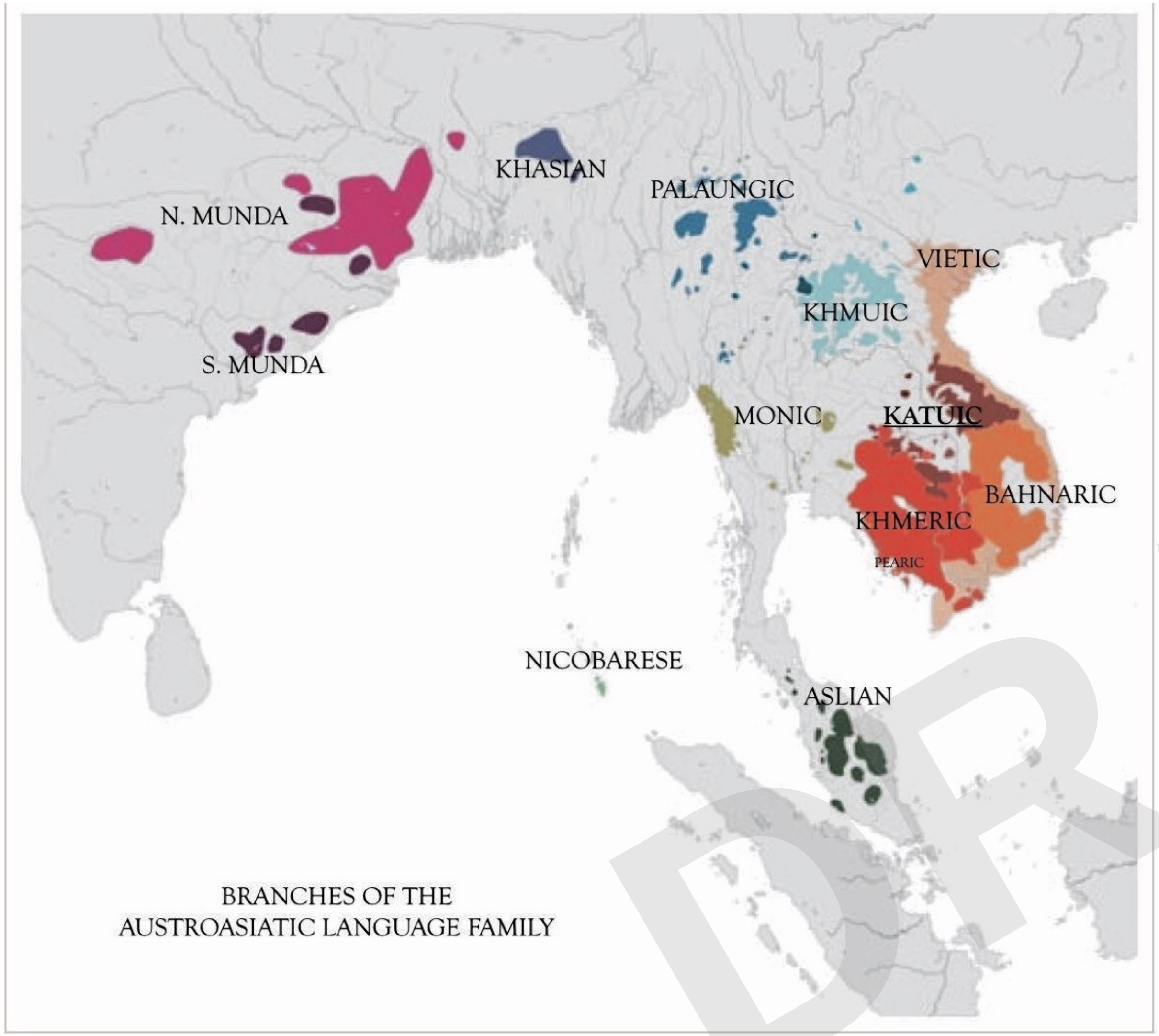
- ប្រមាណ ១៦៨ ភាសា ឃ្លាតគ្នា ខ្ចាត់ខ្ចាយពីឥណ្ឌាភាគឦសាន្ត ដល់ឧបទ្វីប ម៉ាឡេ (Diffloth 2005, Sidewell 2009)
- ការបែកខ្ញែកប្រមាណ ៤ ឬ ៥ ពាន់ ឆ្នាំមុន តាមសម្មតិកម្មនៃទន្លេមេគង្គ (Diffloth 2005)
- ភាសាខ្មែរមានអ្នកនិយាយច្រើនជាងគេទីពីរ បន្ទាប់ពីភាសាវៀតណាម (Schmidt 2007, Haudricourt 1953, Ferlus 2014, Diffloth 2005, Sidewell 2009)



(Schmidt 1907, Diffloth 2005, Sidewell 2009)

DRAFT

Sidwell (2009); Sidwell & Blench (2011), The Austroasiatic Urheimat: The Southeastern Riverine Hypothesis. Canberra: Pacific Linguistics.



Diffloth, Gérard, (2011), **Kuay in Cambodia: vocabulary with historical comments, UNESCO**

van Driem, George. (2017). The domestications and the domesticators of Asian rice. 10.1075/z.215.08van.

DRAFT

អំពីពាក្យថា«ភាសា» ក្នុងភាសាខ្មែរ

Latin

lingua (consonant shift : d > l) *មានន័យថា «អណ្ដាត»*

French

langue => French, English, Khmer, Thai…..

langage => language as a general phenomenon, or to the human ability to have language.

Parole => means 'speech, sometime it refers to written or spoken language.

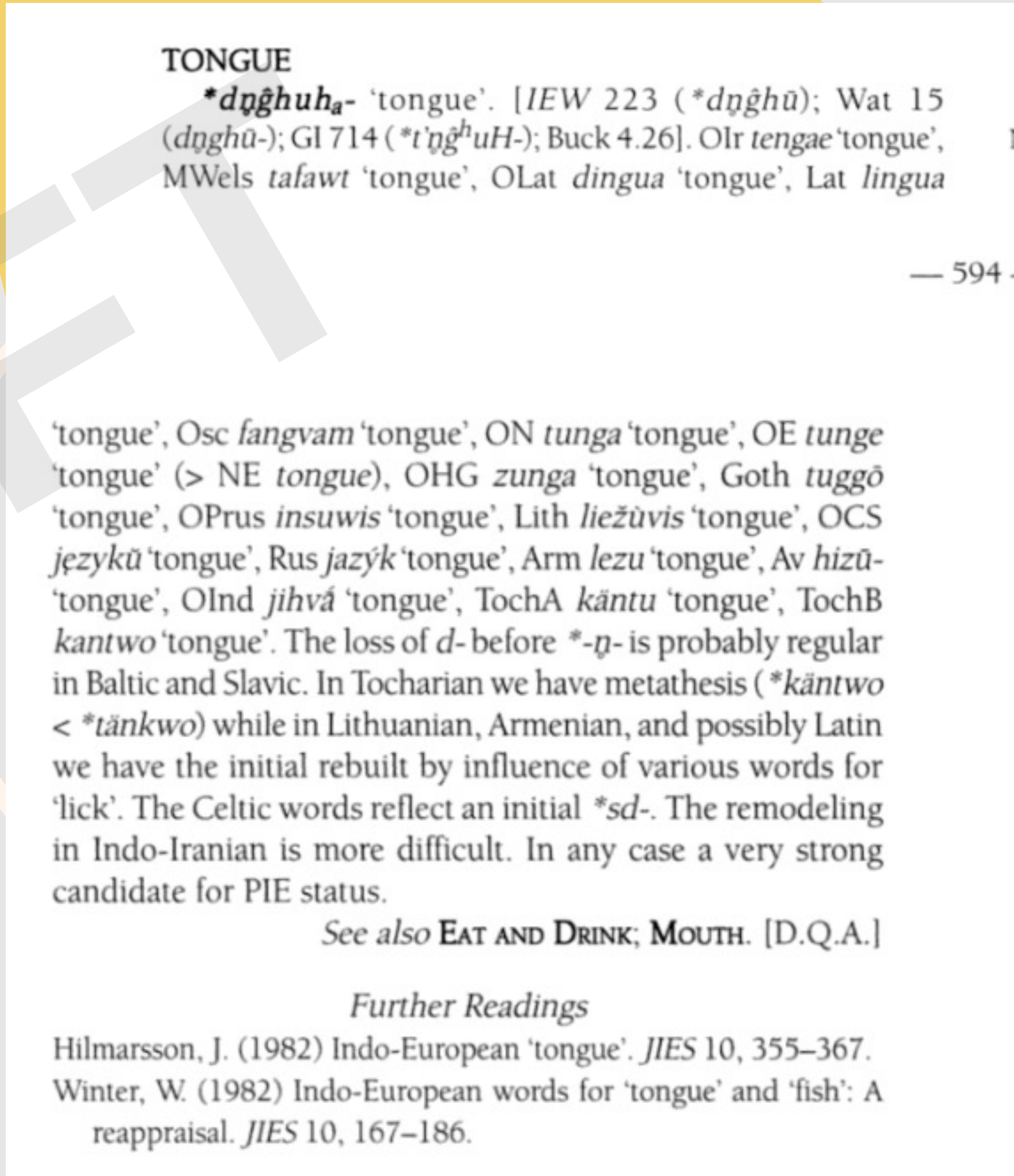
English

Language / tongue

ភាសា (សំ) Sanskrit भाषा (*bhāṣā,* "language") ; Bahasa ?

ចុះពាក្យខ្មែរ ដែលមានន័យថា ភាសា មានឬដែរទេ ?

ឧ. កុំមាត់ មនុស្សមាត់ច្រើន



TONGUE

***dn̥ĝhuh$_a$-** 'tongue'. [*IEW* 223 (**dn̥ĝhū*); Wat 15 (*dn̥ghū-*); Gl 714 (**t'n̥g^{h}uH-*); Buck 4.26]. OIr *tengae* 'tongue', MWels *tafawt* 'tongue', OLat *dingua* 'tongue', Lat *lingua*

— 594 —

'tongue', Osc *fangvam* 'tongue', ON *tunga* 'tongue', OE *tunge* 'tongue' (> NE *tongue*), OHG *zunga* 'tongue', Goth *tuggō* 'tongue', OPrus *insuwis* 'tongue', Lith *liežùvis* 'tongue', OCS *językŭ* 'tongue', Rus *jazýk* 'tongue', Arm *lezu* 'tongue', Av *hizū-* 'tongue', OInd *jihvā́* 'tongue', TochA *käntu* 'tongue', TochB *kantwo* 'tongue'. The loss of *d-* before **-n̥-* is probably regular in Baltic and Slavic. In Tocharian we have metathesis (**käntwo* < **tänkwo*) while in Lithuanian, Armenian, and possibly Latin we have the initial rebuilt by influence of various words for 'lick'. The Celtic words reflect an initial **sd-*. The remodeling in Indo-Iranian is more difficult. In any case a very strong candidate for PIE status.

See also Eat and Drink; Mouth. [D.Q.A.]

Further Readings

Hilmarsson, J. (1982) Indo-European 'tongue'. *JIES* 10, 355–367.

Winter, W. (1982) Indo-European words for 'tongue' and 'fish': A reappraisal. *JIES* 10, 167–186.

DRAFT

សម័យកាលនៃភាសាខ្មែរ

ភាសាខ្មែរបានឆ្លងកាត់ការវិវត្តជាបន្តបន្ទាប់ ចែកជាបួនសម័យកាលធំៗ តាមភស្តុតាងសិលាចារឹក និងឯកសារប្រវត្តិសាស្ត្រ៖

(Schmidt 1904-7, Huffman 1970, Pinnow 1977, Jenner 1967-2009, Jacob 1977, Diffloth 2005)

DRAFT

មុនគ.ស. – ស.វ.ទី៥

ប្រូតូ-ខ្មែរ

Proto-Khmer

ភាសាដែលកើតមានឡើង
តាំងពីមុនមានអក្សរ

ស.វ.ទី៦–១៤

ខ្មែរបុរាណ

Old Khmer

សម័យបុរេអង្គរ និងអង្គរ
• សិលាចារឹក
កម្ពុជាក្សរ

ស.វ.ទី១៥–១៩

ខ្មែរកណ្តាល

Middle Khmer

ការផ្លាស់ប្តូរសូរស្រៈ និងប្រព័ន្ធស៊េរី
ការបង្រួមសូរព្យញ្ជនៈឃោសៈ មកអឃោសៈ
[*b *d *j *g] → [*p *t *c *k]

ស.វ.ទី២០–បច្ចុប្បន្ន

ខ្មែរទំនើប

Modern Khmer

ភាសាខ្មែរស្តង់ដារ •
គ្រាមភ្នំពេញ
គ្រាមភាសាបាត់ដំបង
គ្រាមភាសាតាកែវកំពត

Ferlus (1992), Mon-Khmer Studies 21; Jenner (2009), A Dictionary of Pre-Angkorian / Angkorian Khmer. Diffloth (2005), The Contribution of Linguistic Paleontology to the Homeland of Austro-Asiatic

សម័យកាលនៃភាសាខ្មែរ

ស.វ.ទី១៥–១៩

ខ្មែរកណ្ដាល

Middle Khmer

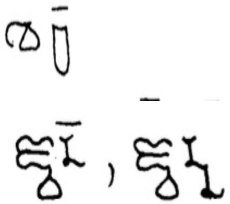
ការផ្លាស់ប្ដូរសូរស្រៈ និងប្រព័ន្ធព្យញ្ជនៈ ការបង្រួមសូរព្យញ្ជនៈឃោសៈ មក អឃោសៈ [*b *d *j *g] → [*p *t *c *k]

(Maspero 1915, Henderson 1952, Huffman 1970, Pinnow 1977, Headley 1985, Vickery 1991)

វ ◌្វ → ព ◌្ព ◌្ប ◌្វ ភ

វេរ → ពីរ

ជ្វន → ជួន



ver /វេរ/ "ពីរ" ម.ព. វ្យរ ផង។

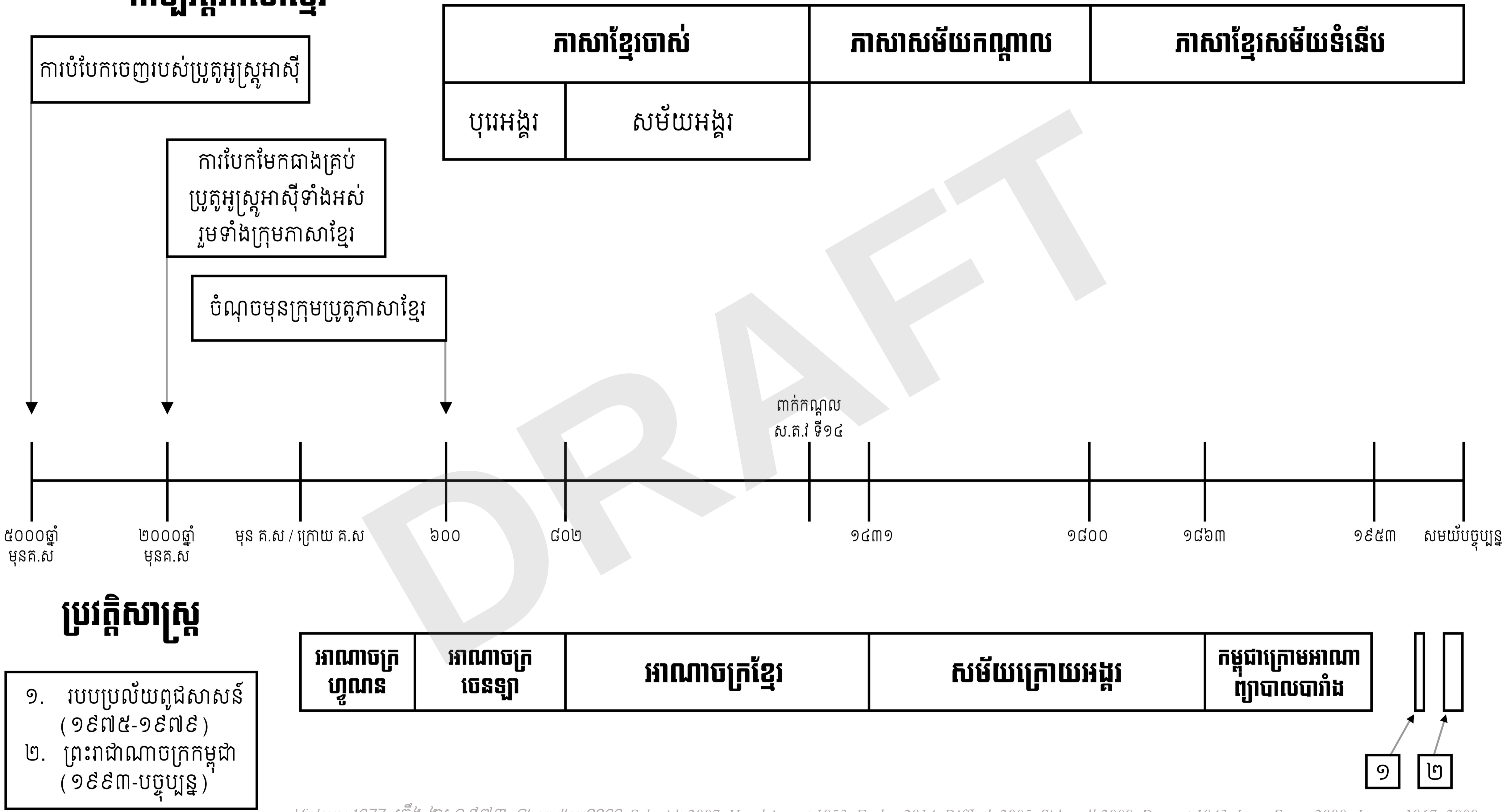
jvan ; jvann

(ឡុង សៀម ២០០០)

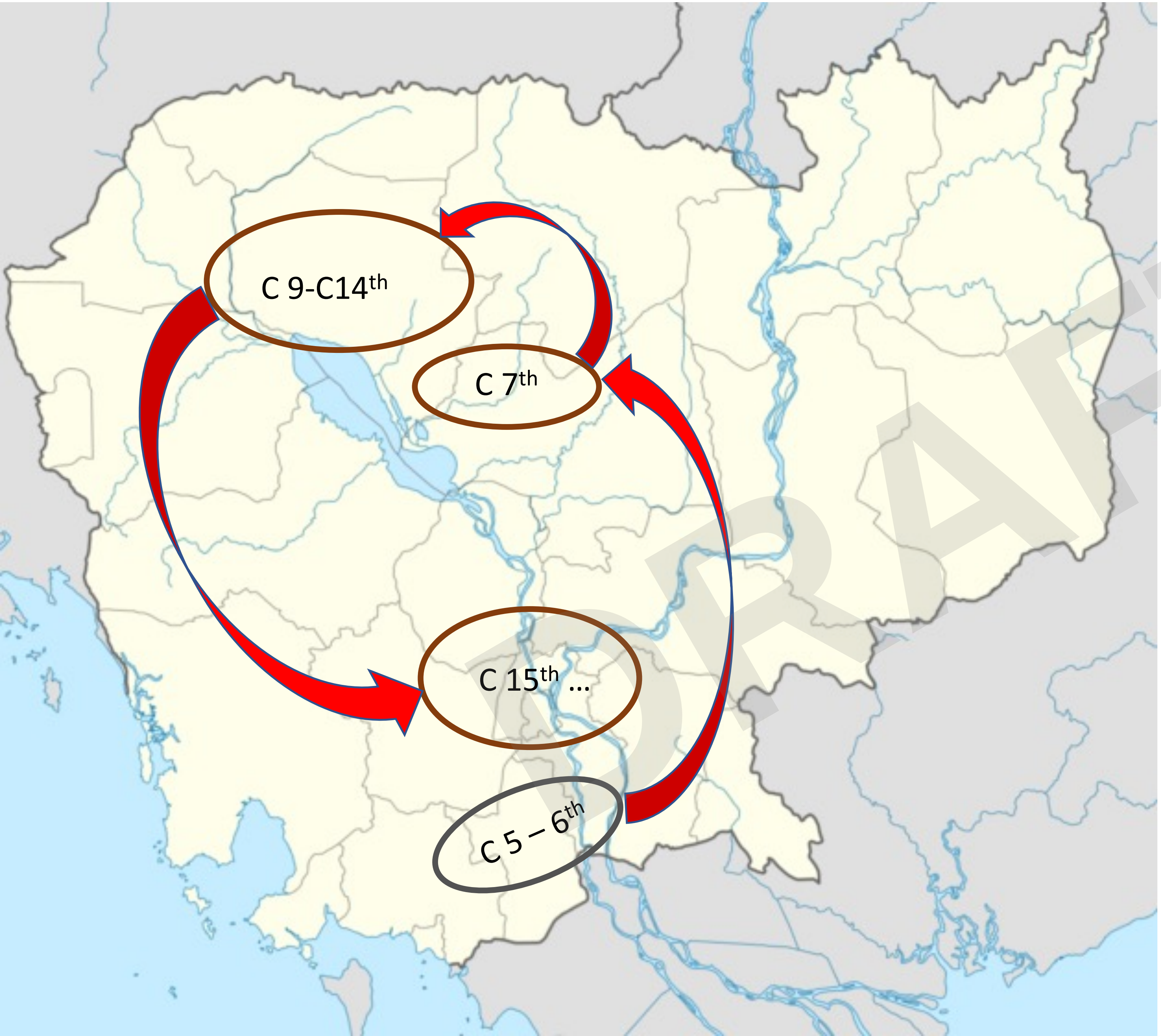
ការប្រែប្រួលសព្យញ្ជនៈ	ភាសាខ្មែរចាស់	ភាសាខ្មែរកណ្ដាល	ភាសាខ្មែរទំនើប
/b/→/p/ or /ph/	ប្លែង /ble:ŋ/	ភ្លើង /pᵊlə:ŋ/(?)	ភ្លើង /phlə:ŋ/
/d/→/t/ or /th/	ដារ /da:(r)/	ដារ /da̤:(r)/	ទារ /tiɜ/
/g/→/k/ or /kh/	អង្វយ /ʔəŋgu:j/	អង្វយ /ʔaŋku:j/	អង្គុយ /ʔaŋkuj/
/ɟ/→/c/ /ch/ /tɕ/	ជេង /ɟe:ŋ/	ជើង /cə:ŋ/	ជើង /tɕə:ŋ/

DRAFT

កាប្បវត្តិភាសាខ្មែរ



Vickery 1977, ត្រឹង ងារ ១៩៧៣, Chandler 2009, Schmidt 2007, Haudricourt 1953, Ferlus 2014, Diffloth 2005, Sidewell 2009, Dupont 1943, Long Seam 2000, Jenner 1967, 2009



ហ្វូណាន ឬ នគរភ្នំ

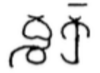
ចេនឡា ឬ សុវណ្ណភូមិ (?)

អាណាចក្រអង្គរ

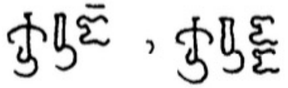
ស្រីសន្ធរ លង្វែក ឧត្តុង្គ ភ្នំពេញ

Vickery 1977, ត្រឹង ងារ ១៩៧៣, Chandler 2009,

vnaṃ

 dhmak

dalmāk

 kuruṅ, kuruṅṅ

វ្នំ => ភ្នំ

ទលិមាកិ => ទល្មាក់ ទ្រម័ក ទ្រមាក់ (ដំរី)

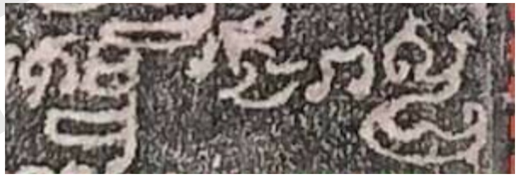
កុរុង => ក្រុង

Dupont 1943, Long Seam 2000, Jenner 1967, 2009

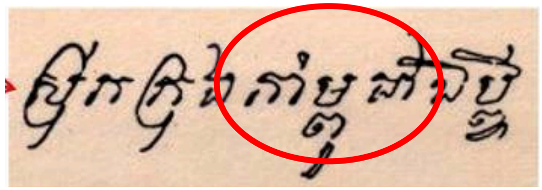
From early Inscriptions to the pagoda manuscripts and to the modern Khmer

C9th

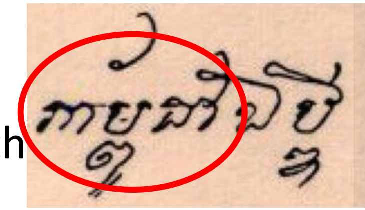
C11th

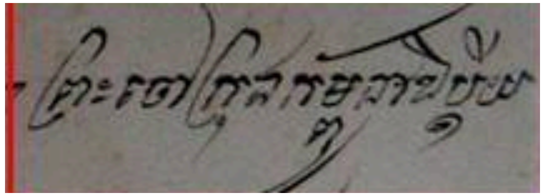


C15th

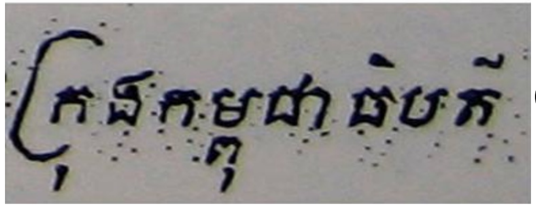


C16th





C19th



C20th

ភាសាខ្មែរពីទសវត្សរ៍ ១៩០០ ដល់បច្ចុប្បន្ន

ដំណាក់កាលនៃការកំណត់បទដ្ឋាន និងបង្កើតភាសាខ្មែរស្តង់ដារ — ពីការសិក្សាបែបវិទ្យាសាស្ត្រ ដល់ការប្រកាសវចនានុក្រមជាផ្លូវការ៖

១៩១៥

លោក G. Maspéro

«វេយ្យាករណ៍ភាសាខ្មែរ» ការសិក្សាបែបទំនើបដំបូង

១៩៣៨–១៩៦៧

សម្តេចសង្ឃរាជ ជួន ណាត

«វចនានុក្រមខ្មែរ» បទដ្ឋានអក្ខរាវិរុទ្ធផ្លូវការ

១៩៥២

លោកស្រី Henderson

ការពិពណ៌នាសូរសព្ទខ្មែរបែបវិទ្យាសាស្ត្រ

១៩៧០

លោក Huffman

ប្រព័ន្ធសរសេរ និងសូរសព្ទខ្មែរស្តង់ដារ

សម័យឌីជីថល

បច្ចុប្បន្ន

ភាសាខ្មែរលើបច្ចេកវិទ្យា • Unicode • AI

DRAFT

Maspéro (1915); Henderson (1952), BSOAS 14(1); Huffman (1970), Yale Univ. Press; ពុទ្ធសាសនបណ្ឌិត្យ, វចនានុក្រមខ្មែរ (១៩៦៧).

ការពណ៌នាអំពីលក្ខណៈនៃភាសាខ្មែរ

1889

Kuhn, Earst. 1889. *Beiträge Zur Sprachenkunde Hinterindiens. Sitzungsberichte Der Königlichen Bayerischen Akademie Der Wissenschaften. München: Philosophisch-philologische Classe II: 189–236.*

1902

Finot, L. 1902. « Notre transcription du cambodgien ». BEFEO, T. 2, PP. 1-15

1915

Maspero, G. 1915. Grammaire de la langue khmère (Cambodgien). Paris.

1946

Martini, François. 1946. Aperçu phonologique du cambodgien. Bulletin de la Société de Linguistique de Paris 42: 112–131.

1952

Henderson, Eugénie J. A. 1952. The main features of Cambodian pronunciation. Bulletin of the School of Oriental and African Studies 14: 149–174.

1966

Noss, Richard B. 1966. The treatment of */r/ in two modern Khmer dialects. In *Studies in comparative Austroasiatic linguistics*, ed. Norman H. Zide, 89–95. London: Mouton & Co.

1966

Gorgoniyev, Y. A. 1966. *The Khmer language*. Moscow : Nauka Publishing House

1967

Huffman, Franklin E. 1967. An outline of Cambodian grammar. PhD dissertation, Cornell University.

1968

Sakamoto, Yasuyuki. 1968. The standard Khmer language and its Phnom Penh dialect. *Tonan Aija Kenkyu* 6: 48–78.

2006

Filippi, Jean-Michel. 2006. Norme et phonétisme dans le Cambodge moderne. *Udaya* 7: 1–31.

2014

Kirby, James. 2014a. Incipient tonogenesis in Phnom Penh Khmer: Acoustic and perceptual studies. *Journal of Phonetics* 43: 69–85. doi:10.1016/j.wocn.2014.02.001.

DRAFT

ភាសាខ្មែរជាអ្វី?

០២

PART 02 · WHAT IS THE KHMER LANGUAGE?

DRAFT

ភាសាខ្មែរ ជាភាសាញែក/វិភាគ (Isolating / Analytic)

តាមរូបសព្ទវិទ្យា ភាសាខ្មែរជាភាសាប្រភេទញែក/វិភាគ (isolating / analytic) — ពាក្យភាគច្រើនជាឯកតាសូរតែមួយ (ឯកតាមួយជាឫសពាក្យ) ហើយទំនាក់ទំនងវេយ្យាករណ៍បង្ហាញតាមលំដាប់ពាក្យ និងពាក្យជាឧបករណ៍ ជាជាងការបំប្លែងទម្រង់ពាក្យ។ (Maspero 1915, Martini 1946, Gorgoniyev 1966, Jacob 1968, Huffman 1970, Bissag 2014)

ញែក / វិភាគ

Isolating & analytic language

-ពាក្យ ≈ ដាក់បន្តបន្ទាប់ ផ្តុំ ផ្សំ និងបូកបន្ថែម

-គ្មាន វិភត្តិ វចនៈ កាល លិង្គ បុរស ... (Maspero 1915, Martini 1946, Huffman 1970, Jacob 1968)

ប្រព័ន្ធសំណេរ

abugida, alphasyllabic script (Peter T. Daniels and William Bright, 1996)

-យកពីប្រព័ន្ធសំណេរឥណ្ឌា ព្រាហ្មី កំណត់ត្រាជាភាសាខ្មែរដំបូង ៦១១ (ហ៊ុន ឈុនតេង ២០២១)

-យកតួព្យញ្ជនៈ ជាគោល ចំណែកស្រៈ បន្ទាប់បន្សំ ដើម្បីបង្កើត ព្យាង្គ ឬ ពាក្យ (Bissag, 2014)

-សញ្ញា សរុប ដើម្បីបង្កើតសូរ មាន ៩៥ ឬ ៩៦ សញ្ញា។ (Nou, Sun , 2013)

លក្ខណៈ

-សម្បូរ ពាក្យ ១ ព្យាង្គ និង ២ ព្យាង្គ (ចំណែក ពាក្យច្រើនព្យាង្គ ខ្ចីពីខាងក្រៅ)

ឧ. កើត » ខ្នើត » កំណើត » បង្កើត » កើតឡើង (Henderson , 1952)

ឧ. ដំឡូងឈាមមាន់ ; ម៉ាស៊ីនកិនស្រូវ

ពហុសំយោគនៃសំណេរ

Polysynthetic

ពាក្យមួយ ≈ សរសេរបានច្រើនយ៉ាង

កងពល

«សាស្ត្រ» «សាសន៍»«សាសន» និង «សះ»

(វចនានុក្រមខ្មែរ, ១៩៦៧, ២០២២)

Comrie (1989), *Language Universals and Linguistic Typology*, 2nd ed.; Haiman (2011), *Cambodian: Khmer*. Amsterdam: John Benjamins.

DRAFT

វាក្យសម្ព័ន្ធ និងលំដាប់ពាក្យ

ភាសាខ្មែរមានលំដាប់ពាក្យមូលដ្ឋានជា ប្រធាន–កិរិយា–កម្ម (SVO)។ គុណនាមតាមក្រោយនាម គុណកិរិយាតាមក្រោយកិរិយា ហើយមិនមានការប្រែទម្រង់ពាក្យតាមពេលវេលា ឬ បុរស ឡើយ — អត្ថន័យវេយ្យាករណ៍កើតឡើងតាមលំដាប់ និងពាក្យជំនួយ។

Jacob (1968), Introduction to Cambodian. Oxford Univ. Press; Haiman (2011), Cambodian: Khmer, John Benjamins.

ប្រធាន	→	កិរិយាស័ព្ទ	→	កម្មបទ
កុមារ		**អាន**		**សៀវភៅ**
/koʔmaː/		*/ʔaːn/*		*/siəβ-phɨβ/*
‘child’		‘read’		‘book’
[272]		[267]		[286]

«កុមារអានសៀវភៅ។» — ‘The child reads a book។’

កុមារ

កុមារកំព្រា → ក្រមា

ស់ាភៅ

ស-ភៅ

tokens: ក ◌ុ ម ◌ា រ | អ ◌ា ន | ស ◌ៀ វ ភ ◌ៅ ។
ids : 5 14 8 13 9 4 12 13 6 4 11 15 10 7 16 17

DRAFT

តើអ្វីជាភាសាខ្មែរស្តង់ដា?

នៅក្នុងវិស័យសង្គមភាសាវិទ្យា ភាសាស្តង់ដា ត្រូវបានកំណត់និយមន័យថាជា វ៉ារ្យង់(variety) មួយនៃភាសា ដែលបានឆ្លងកាត់ការកូដកម្ម (codification) យ៉ាងស៊ីជម្រៅ ក្នុងវេយ្យាករណ៍ វាក្យសព្ទ ប្រព័ន្ធអក្សរ ឬលក្ខណៈផ្សេងទៀត ហើយដែលឈរលេចធ្លោជាងវ៉ារ្យង់ដទៃក្នុងសហគមន៍ភាសានោះ ដោយមានឋានៈឬកិត្យានុភាព (prestige) ខ្ពស់ជាងគេ។

(Richards & Schmidt 2010: 554; Finegan 2007: 14)

(

១. **ការជ្រើសរើសបទដ្ឋាន** (Selection of the norm)

— ការជ្រើសរើសគ្រាមភាសាមួយដែលជាធម្មតាមានកិត្យានុភាព សង្គម នយោបាយ ឬ សេដ្ឋកិច្ច។

២. **កូដកម្មនៃទម្រង់** (Codification of form)

— ការកត់ត្រាក្បួនវេយ្យាករណ៍ និងអក្ខរាវិរុទ្ធក្នុងវចនានុក្រម និងសៀវភៅវេយ្យាករណ៍។

៣. **ការពង្រីកមុខងារ** (Elaboration of function)

— ការពង្រីកធនធានភាសា (វាក្យសព្ទ ស្ទីលនៃការសរសេរ) ដើម្បីបម្រើដល់ដែនប្រើប្រាស់ផ្លូវការ ដូចជាច្បាប់ វិទ្យាសាស្ត្រ និងរដ្ឋបាល។

៤. **ការទទួលយក** (Acceptance)

— ការទទួលស្គាល់នៅក្នុងស្ថាប័នអប់រំ ប្រព័ន្ធផ្សព្វផ្សាយ និង ផ្នែករដ្ឋបាល។

(Haugen 1966, សង្ខេប៖ American Academy of Arts and Sciences, Daedalus 2023)

DRAFT

តើអ្វីជាភាសាខ្មែរស្តង់ដា?

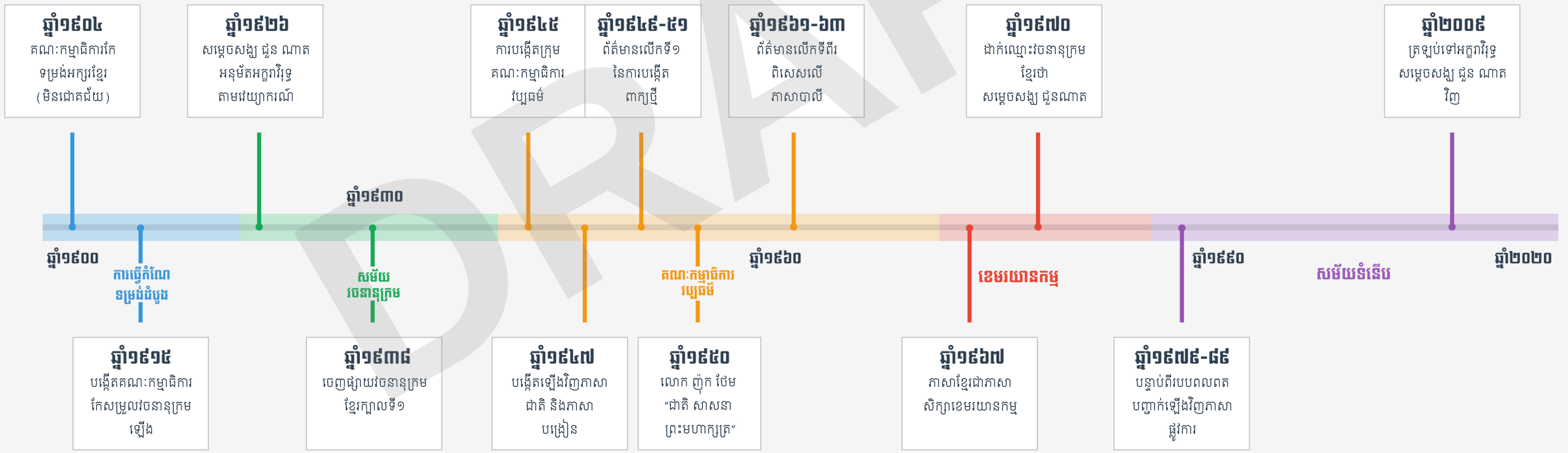
ការបង្កើតគណៈកម្មការភាសាខ្មែរសម័យទំនើប (១៩០៤-២០០៩)

បុគ្គលសំខាន់ៗ :

- សម្ដេចសង្ឃ ជួន ណាត : ព្រះសង្ឃ, អ្នកកែសម្រួលវចនានុក្រម
- សម្ដេចសង្ឃ ហួត តាត : សហការី, គណៈកម្មាធិការវប្បធម៌
- លោក ញ៉ុក ថែម : អ្នកនិពន្ធ “ជាតិ សាសនា ព្រះមហាក្សត្រ”
- លោក កេង វ៉ាន់សាក់ : អ្នកទំនើបកម្ម និង ដឹកនាំខេមរយានកម្ម

លទ្ធផលភាសាខ្មែរសម័យទំនើប :

- អក្ខរាវិរុទ្ធតាមវេយ្យាករណ៍ (សម្ដេចសង្ឃ ជួន ណាត)
- វាក្យស័ព្ទមកពីភាសាបាលី (គណៈកម្មាធិការវប្បធម៌)
- សាមញ្ញនីយកម្មនៃធាតុភាសា (ខេមរយានកម្ម)
- ភាពបតបែននៃការប្រើប្រាស់ភាសាខ្មែរសម័យទំនើប



ដំណាក់កាលនៃកាលប្បវត្តិភាសាខ្មែរ :

- ការកែទម្រង់ដំបូង(១៩០៤-១៩១៥)
- ការអភិវឌ្ឍវចនានុក្រម(១៩២៦-១៩៣៨)
- សម័យគណៈកម្មាធិការវប្បធម៌(១៩៤៥-១៩៦៣)
- ចលនាខេមរយានកម្ម(១៩៦៧-១៩៧០)
- ការឯកភាពគ្នានាសម័យទំនើប(១៩៧៩-២០០៩)

DRAFT

ភាសាខ្មែរ នៅលើប្រព័ន្ធផ្សព្វផ្សាយ ចន្លោះ ឆ្នាំ ១៩៦០ ដល់ ២០០០

1962

Mixed with BS dialects

/oβ/; /e/

Non-rhotic /ʁ/
Prefixe /pan/ បញ្ញា
No aspiration /p/,/t/
/c/,/k/ + nasals

977

/ɨβ/ ; /iɜ/

Non-rhotic /ʁ/

1985

Mixed with PP dialect

/ɨβ/ ; /iɜ/; /iə/

Non-rhotic /ʁ/

1999

Close to PP dialects
The strong vowel classes(ɑː, ɔː)
/ɨ/ ; /iɜ/
Notable allophone /k/ >[c,ʔ]
Rhotic /ɾ/, /ʁ/
Prefixes /paɲ/
Strong aspiration /p/,/t/ /c/,/k/ + nasals

DRAFT

តើអ្វីជាភាសាខ្មែរស្តង់ដារ?

DRAFT

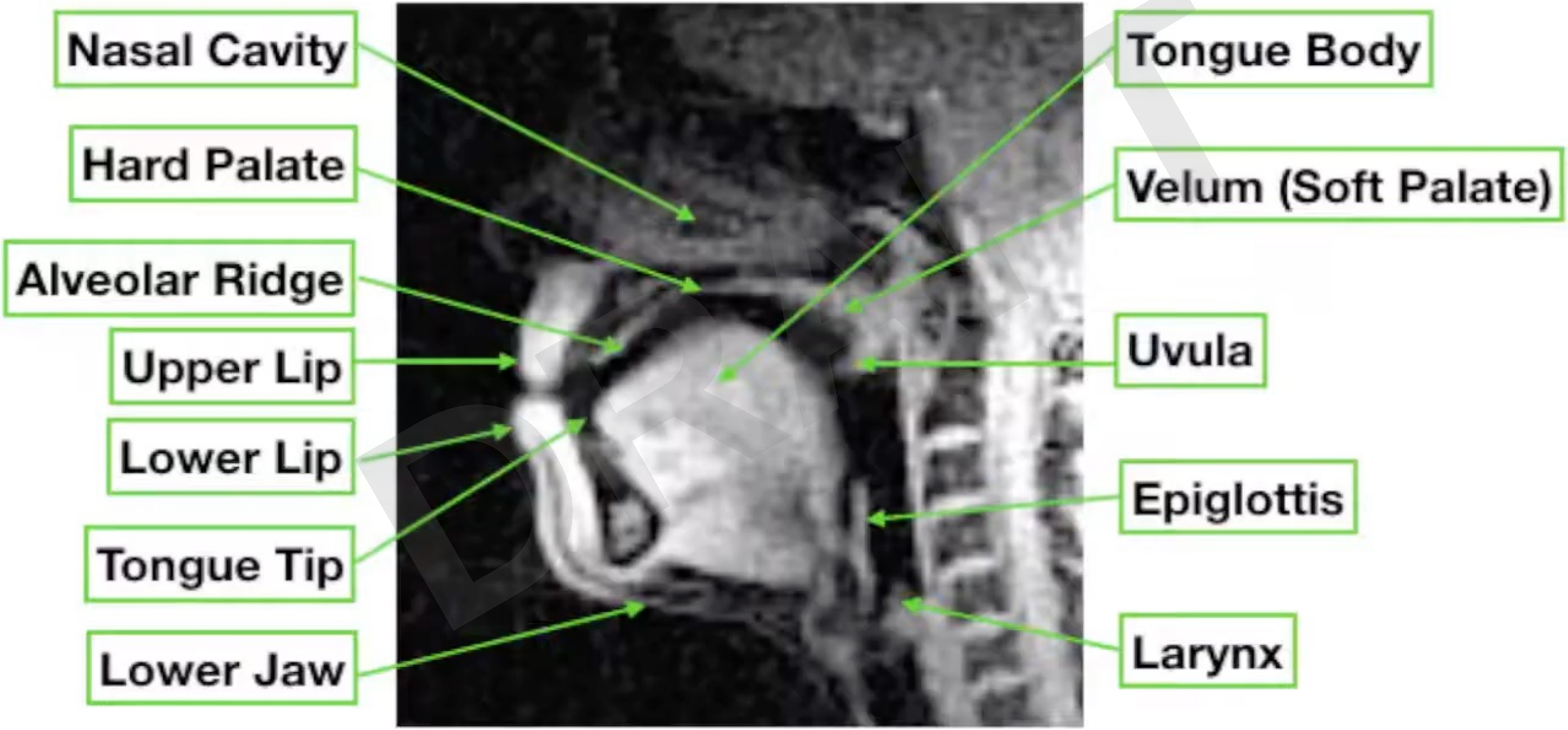
រចនាសម្ព័ន្ធសូរសព្ទ

០៣

PART 03 · KHMER PHONEMIC STRUCTURE

DRAFT

Real-time speech on Magnetic resonance imaging (MRI)



ព្យញ្ជនៈខ្មែរ៖ និយមន័យ និងរចនាសម្ព័ន្ធ

ព្យញ្ជនៈ គឺជាសូរសព្ទដែលបង្កើតឡើងដោយការរាំងខ្យល់ក្នុងផ្លូវសំឡេង។ ក្នុងប្រព័ន្ធសរសេរ ខ្មែរមានអក្សរព្យញ្ជនៈ ៣២ ឬ ៣៣ តួ ដែលចែកជាពីរសេរី តាមលក្ខណៈប្រវត្តិសាស្ត្រនៃការបន្លឺសំឡេង។ *(វចនានុក្រមខ្មែរ ១៩៦៧, ព្រំ ម៉ុល ២០០៣, នៅ ស៊ុន ២០១៣)*

សេរីទី១ · អឃោសៈ

a-series

ស្រៈពឹង 'អ'

ក ខ ច ឆ ដ ឋ ណ ត ថ ប ផ ស ហ ឡ អ

សេរីទី២ · ឃោសៈ

o-series

ស្រៈពឹង 'អ៊'

គ ឃ ង ជ ឈ ញ ឌ ឍ ទ ធ ន ព ភ ម យ រ ល វ

៣៣ អក្សរព្យញ្ជនៈ

២ សេរី (ឃោសៈ/អឃោសៈ)

~១៧–២១ សូរសព្ទព្យញ្ជនៈ

DRAFT

Henderson (1952), BSOAS 14(1):149–174; Huffman (1970), Cambodian System of Writing and Beginning Reader.

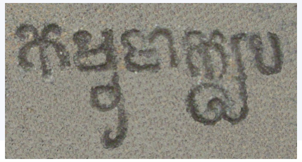
ព្យញ្ជនៈខ្មែរ និយមន័យ និងរចនាសម្ព័ន្ធ

អក្សរខ្មែរ បានវិវត្តន៍ចេញពី អក្សរ "ព្រាហ្មី"នៅសម័យ "បល្លវៈ" របស់ឥណ្ឌា

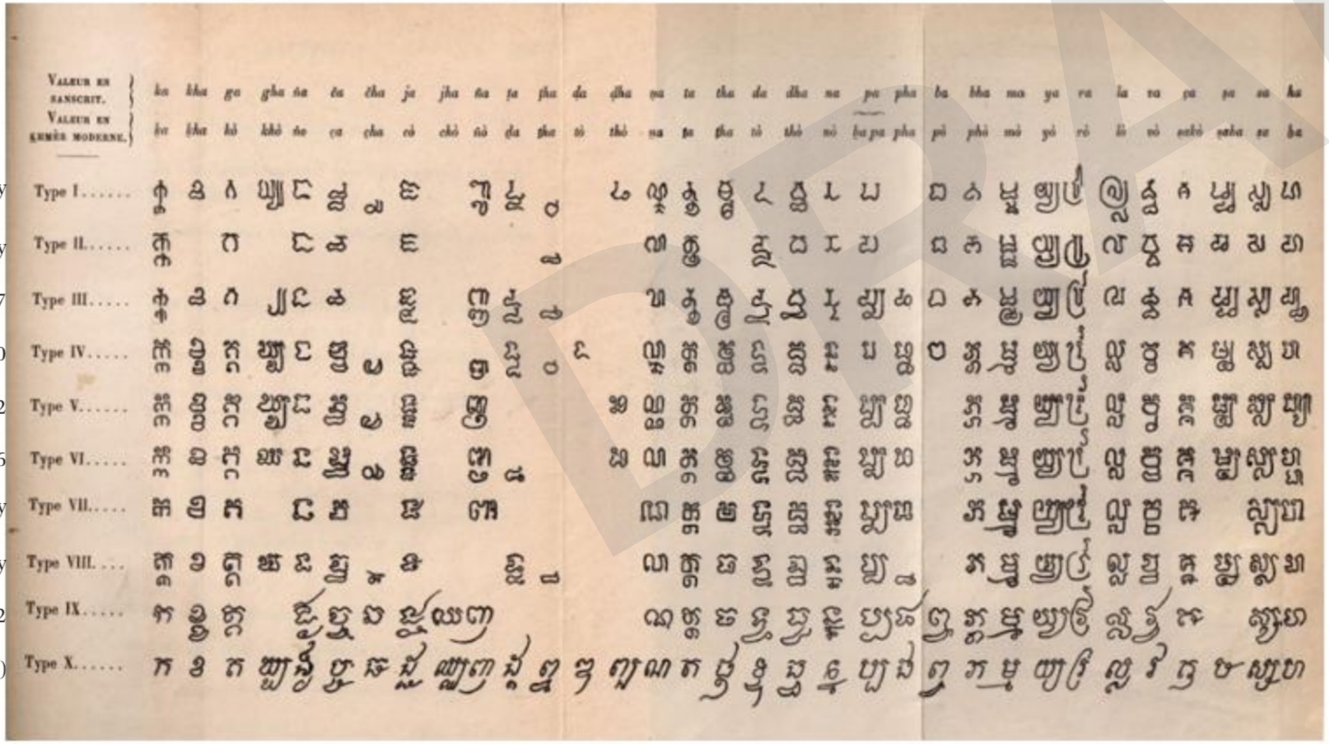
(Hendrik Kern 1879, Auguste Bard 1885, & Abel Bergaigne 1893)

នៅរាជ្យព្រះបាទយសោវម៌្មទី១ នាចុង ស.វ.ទី៩ ដើមស.វ.ទី១០ **កម្ពុជាក្សរ**

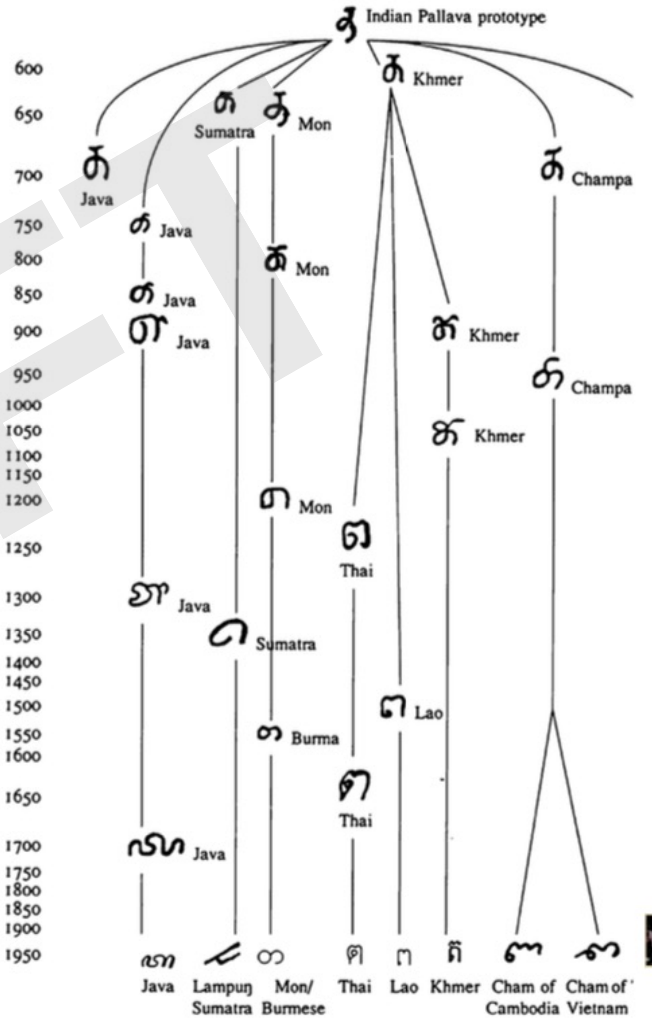
(អាំង ជូលាន ២០១៩, ហ៊ុន ឈុនតេង ២០២១)



K. 362



(G. Maspero 1915)



មហាសករាជ បើគិតមកជា គ្រឹសសករាជ ត្រូវបូកបន្ថែម ៧៨ឆ្នាំបន្ថែម (Dominique Soutif (EFEO), Chea Socheat (APSARA), Julia Estève (CRS, Mahidol) ។

DRAFT

ព្យញ្ជនៈខ្មែរ៖ និយមន័យ និងរចនាសម្ព័ន្ធ

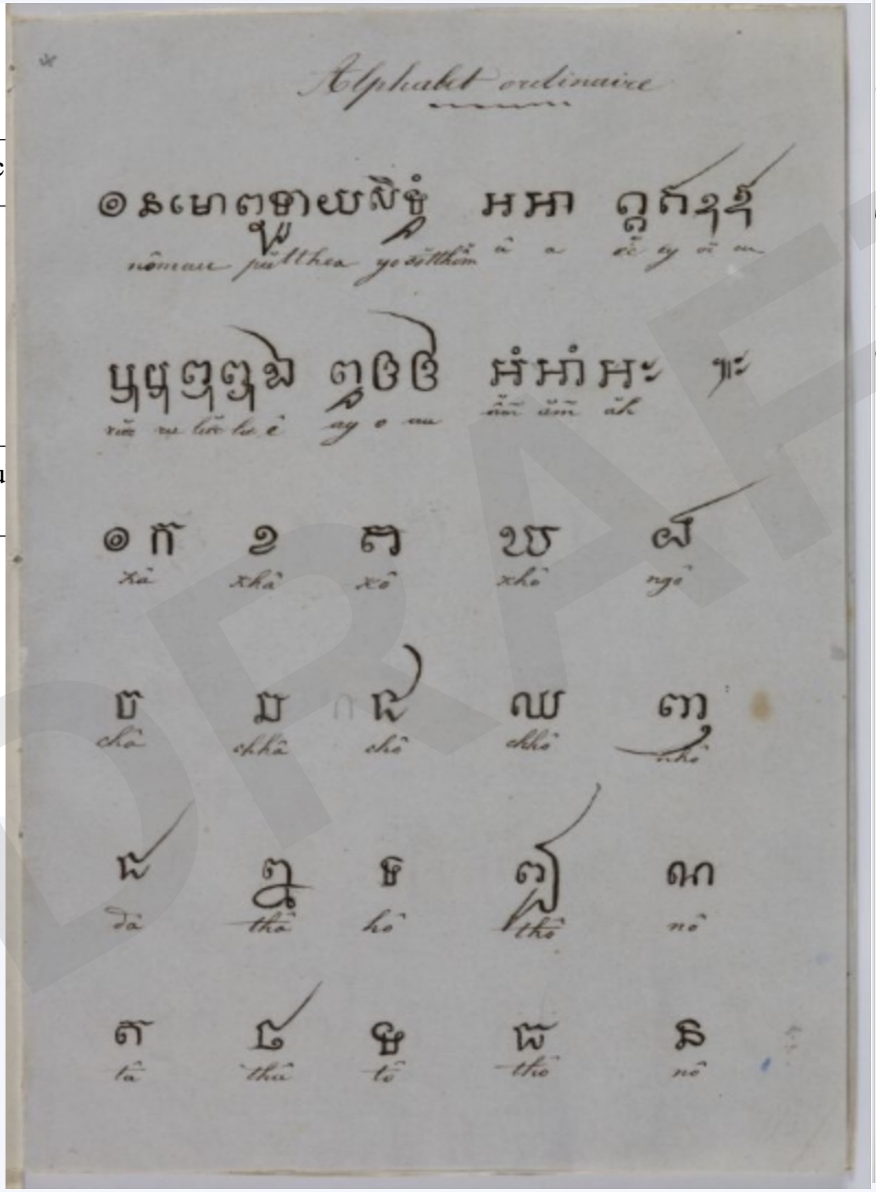
ក្ក	ខ្ខ	គ្គ	ឃ្ឃ	ង្ង	**Velars**
kaː	khaː	kɔː	khɔː	ŋɔː	
ច្ច	ឆ្ឆ	ជ្ជ	ឈ្ឈ	ញ្ញ	**Palatals**
caː	chaː	cɔː	chɔː	ɲɔː	
ដ្ដ	ឋ្ឋ	ឌ្ឌ	ឍ្ឍ	ណ្ណ	**Dentals**
ɗaː	thaː	ɗɔː	thɔː	naː	
ត្ត	ថ្ថ	ទ្ទ	ធ្ធ	ន្ន	**Dentals**
taː	thaː	tɔː	thɔː	nɔː	
ប្ប	ផ្ផ	ព្ព	ភ្ភ	ម្ម	**Bilabials**
ɓaː	phaː	pɔː	phɔː	mɔː	

យ្យ	រ្រ	ល្ល	វ្វ	ស្ស	ហ្ហ	ឡ្ឡ	អ្អ
jɔː	rɔː	lɔː	βɔː	saː	haː	laː	ʔaː

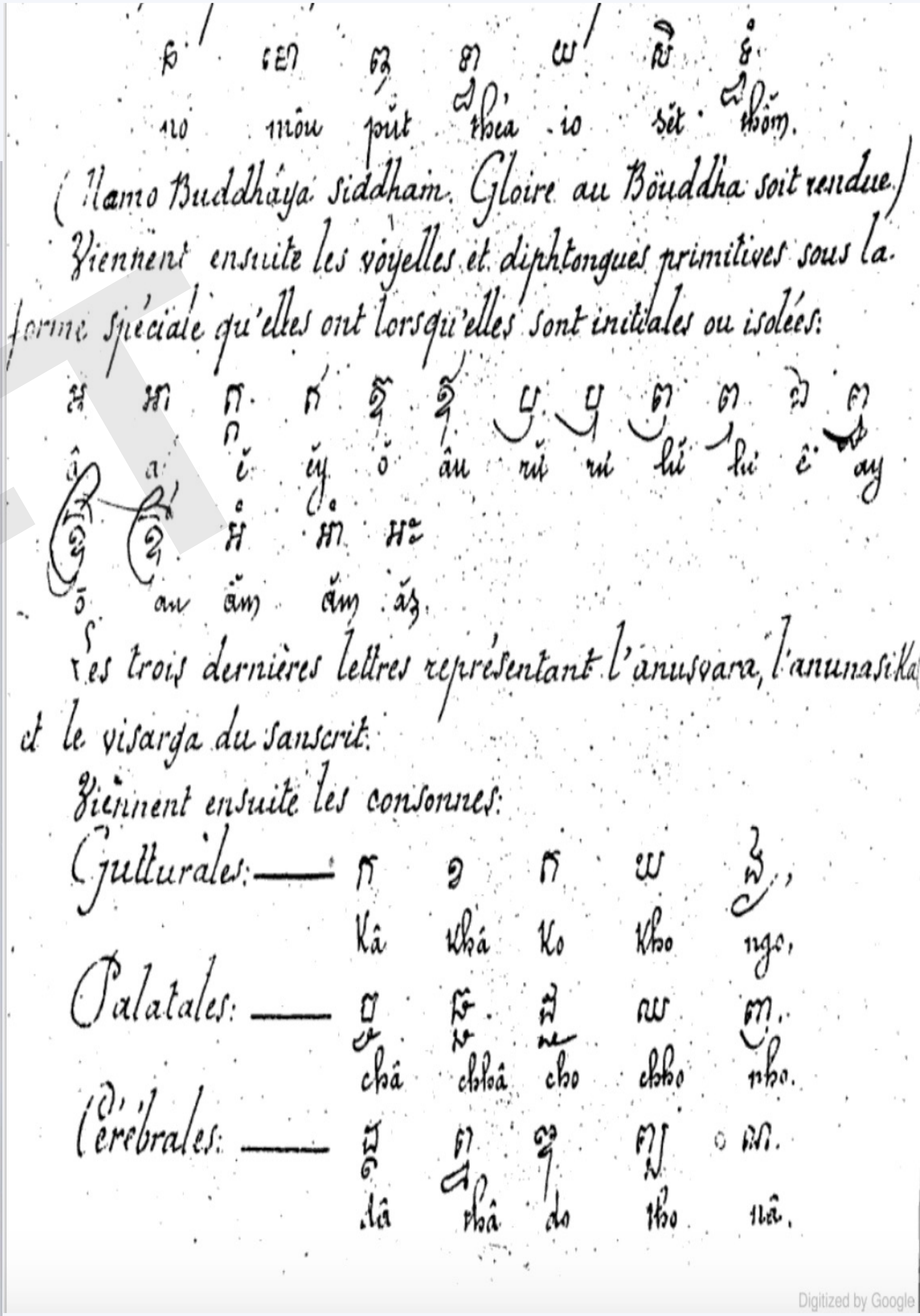
The c

The su

(Huffman 1970)



One of Mouhot's transcriptions of Khmer sacred text (Pali-Khmer) is kept at the British Library.


nô môu pout thea io sit thôm

(Namo Buddhâya siddham. Gloire au Bouddha soit rendue.)

Viennent ensuite les voyelles et diphtongues primitives sous la forme spéciale qu'elles ont lorsqu'elles sont initiales ou isolées:

Les trois dernières lettres représentant l'anusvara, l'anunasika et le visarga du sanscrit.

Viennent ensuite les consonnes:

Gutturales: — Kâ khâ Ko Kho ngo,

Palatales: — châ chhâ cho chho nho.

Cérébrales: — dâ thâ do tho nâ.

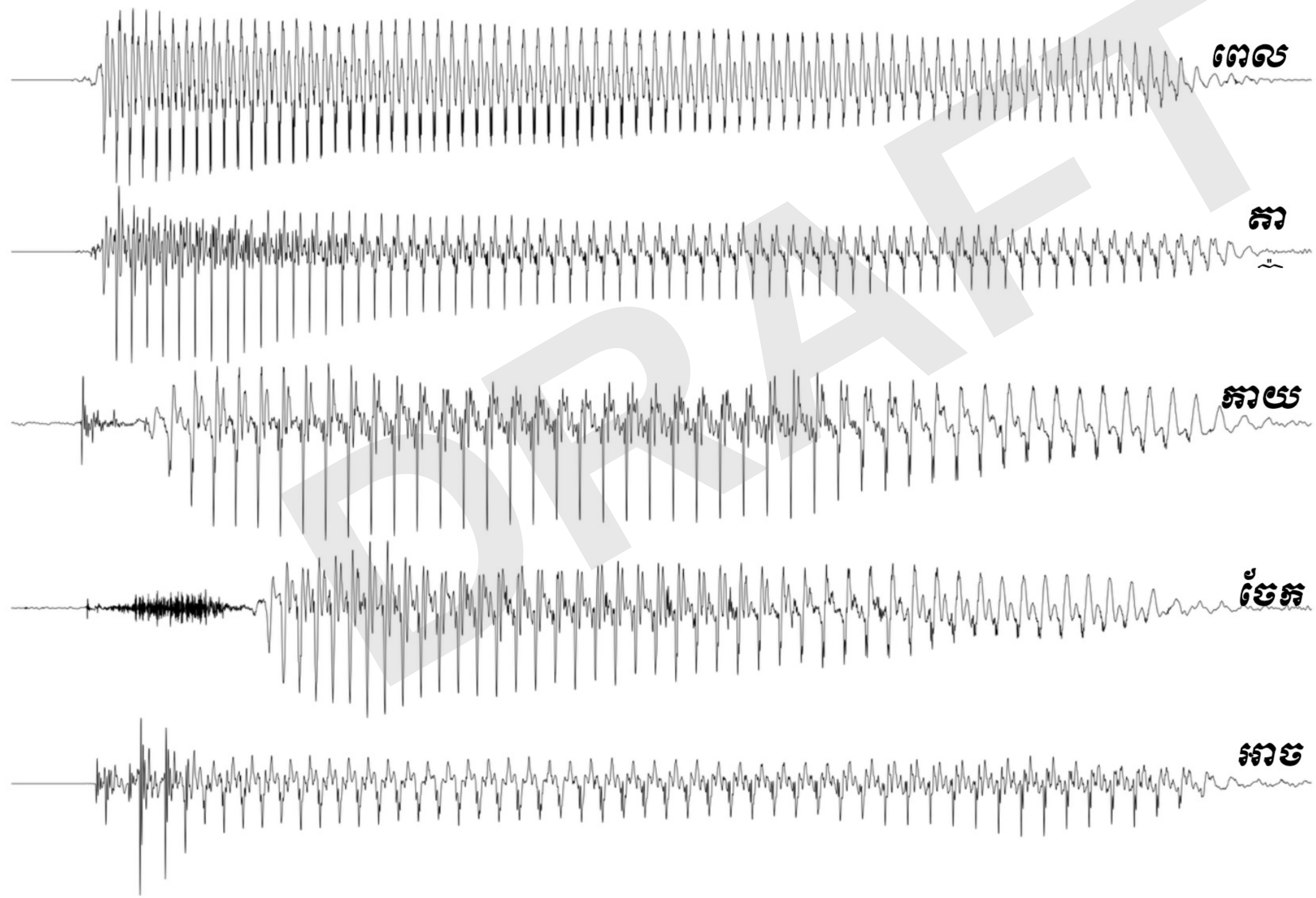
Digitized by Google

(Aymonier 1878)

DRAFT

ព្យញ្ជនៈខ្មែរ៖ និយមន័យ និងរចនាសម្ព័ន្ធ

ព្យញ្ជនៈខ្យលស្ទះ និងខ្យល់ស្ទះត្រដុស អាហ្វ្រីខិត [p t k tɕ ʔ]



ប៉ព /p/

តទ /t/

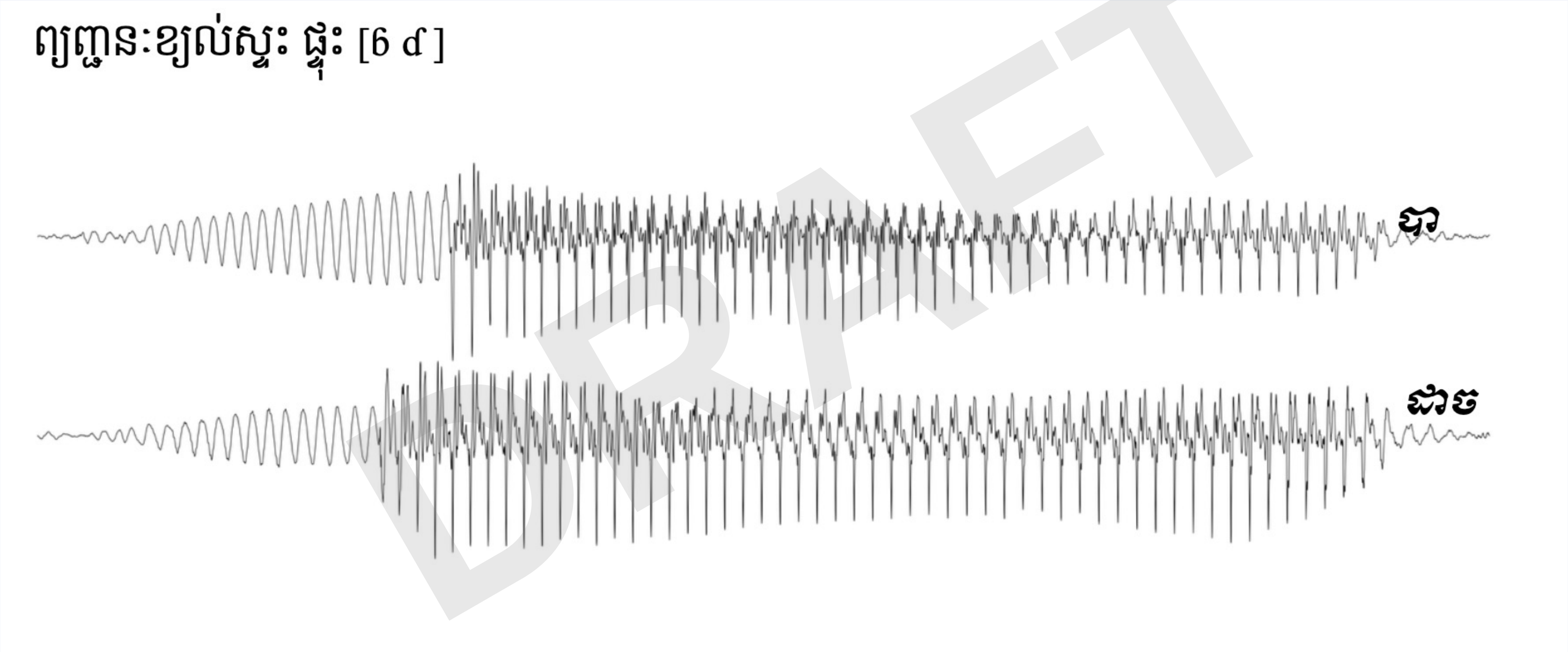
កគ /k/

ចជ /tɕ/

អអ៊ /ʔ/

DRAFT

ព្យញ្ជនៈខ្មែរ៖ និយមន័យ និងរចនាសម្ព័ន្ធ



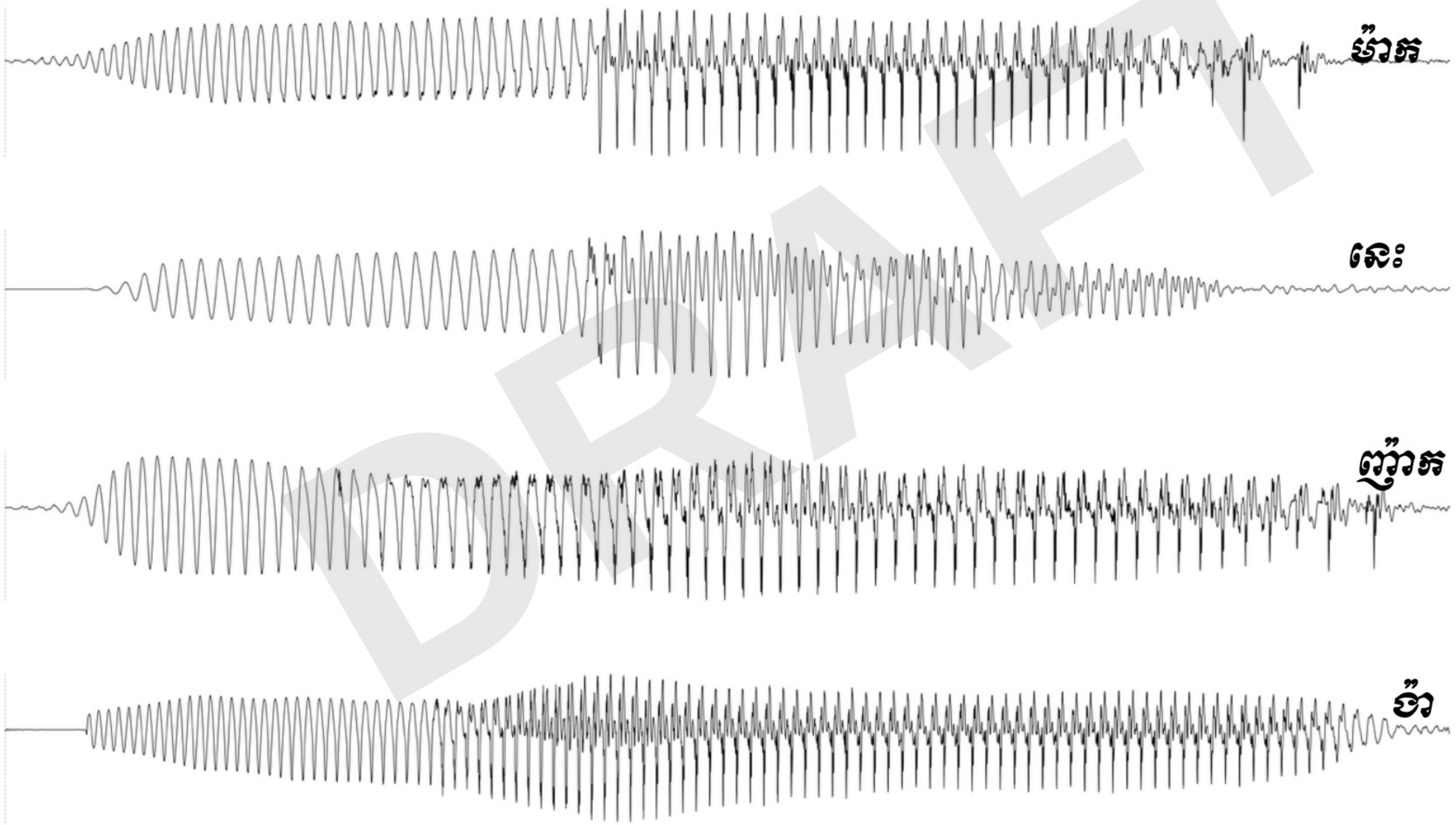
បប៊ /ɓ/

ដឌ /ɗ/

DRAFT

ព្យញ្ជនៈខ្មែរ៖ និយមន័យ និងរចនាសម្ព័ន្ធ

ព្យញ្ជនៈខ្យល់ច្រមុះ [m n ɲ ŋ]



DRAFT

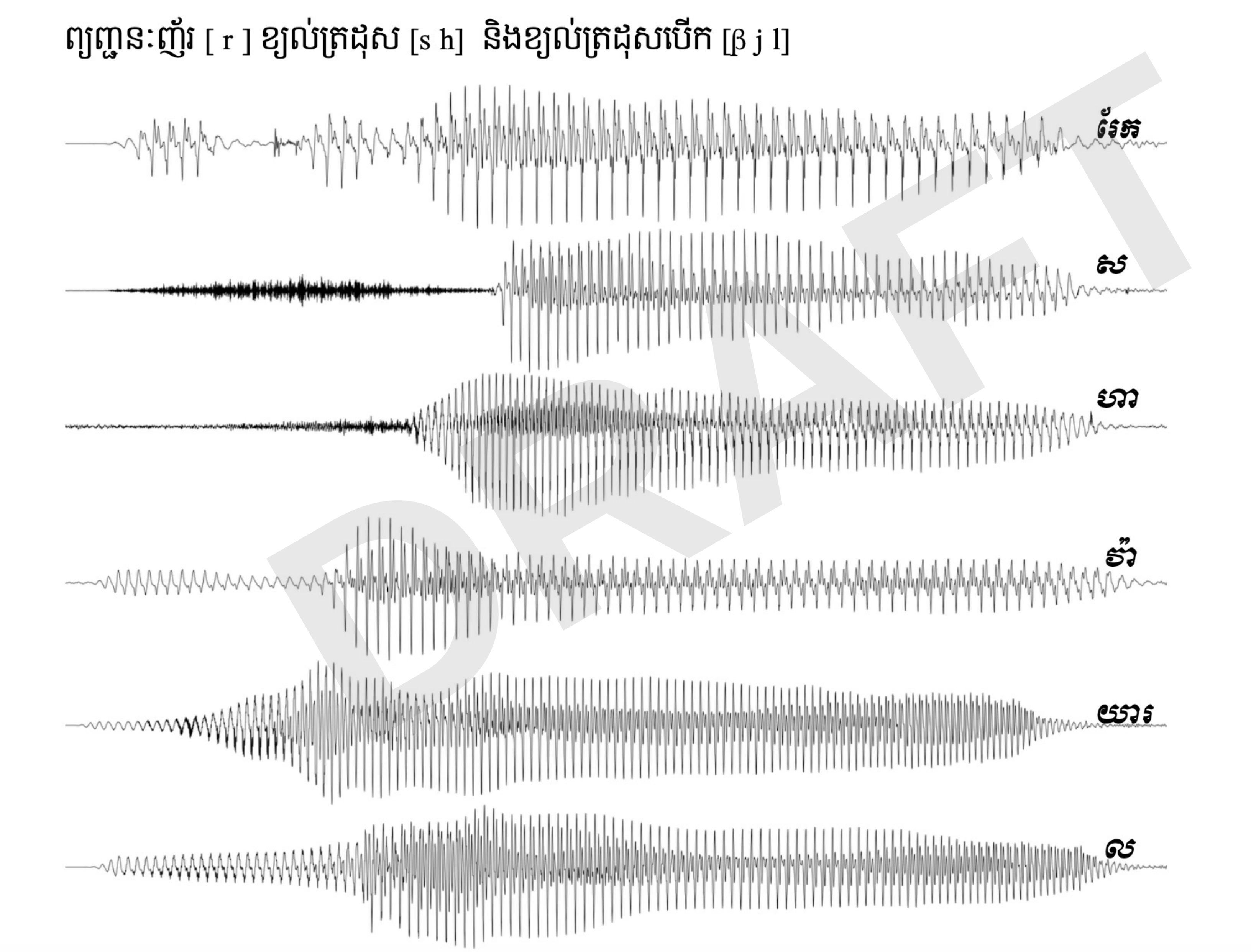
មម៉ /m/

នណ /n/

ញញ៉ /ɲ/

ងង៉ /ŋ/

ព្យញ្ជនៈខ្មែរ៖ និយមន័យ និងរចនាសម្ព័ន្ធ



DRAFT

រ រ៉ /r/ →/ɾ/→/ʁ/→/ɦ/

សស៊ /s/ → /-h/

ហហ៊ /h/→/-h/ (◌ះ)

វវ៉ /β̞/

យយ៉ /j/

លឡ /l/

Consonnes initiales

	sourdes (non aspirées)	sonores (non aspirées)	nasales
Glottale	ʼ		
Vélaires	k		ṅ
Palatales	c	y	ñ
Dentales	t	d	n
Labiales	p	b	m

Liquides

r, l

Spirantes

h, s, v, f

(Martini 1946)

Chart 1. Consonants

	Labial	Dental	Palatal	Velar	Glottal
Stops:					
Voiceless	-p	-t	-c	-k	-q
Voiced	b	d			
Spirants:					
Voiceless	(f)	s			-h
Continuants:					
Nasal	-m	-n	-ñ	-ŋ	
Semivocalic	-w		-y		
Lateral		-l			
Trilled		r			

(Huffman 1967)

Consonants:	Bilabial	Alveolar	Palatal	Velar	Glottal
Stop	p, b	t, d		k	
Fricative	f	s		h	
Nasal	m	n	nj	ng	
Lateral		l			
Rhotic		r			
Glide	v		j		

(Haiman 2011)

Occlusive	p, b	t, d	c	k	ʔ
Nasal	m	n	ɲ	ŋ	
Semivowel	v	r, l	y		
Spirant		s			h

(Henderson 1952)

	bilabial	dental	palatal	velar	glottal
Voiceless stops	p	t	c	k	q
Preglottalized stops	b	d	-	-	-
Nasal resonants	m	n	ɲ	ŋ	-
Spirants	-	s	-	-	h
Liquids	-	r l	-	-	-
Semivowels	w	-	j	-	-

(Jenner 1967)

	Labial	Dental	Palatal	Velar	Glottal
Plosive	p b	t d	c	k	ʔ
Nasal	m	n	ɲ	ŋ	
Trill		r			
Fricative		s			h
Approximant	β̞		j		
lateral approximant		L			

(Filippi & Hiep 2016)

DRAFT

ព្យញ្ជនៈខ្មែរ៖ និយមន័យ និងរចនាសម្ព័ន្ធ

ទីបន្លឺ / លក្ខណៈបន្លឺ	បបូរមាត់	ធ្មេញ-គន្លាក់ក្រអូម	ក្រអូម	រនាំងក្រអូម	ចន្លោះខ្សែសំឡេង
ព្យញ្ជនៈខ្យល់ស្ទះ	-p b- ប៉ ព ប ប៊	-t d- ត ទ ដ ឌ	-c ច ជ	-k ក គ	-ʔ អ អ៊
ព្យញ្ជនៈខ្យល់ស្ទះផ្ទុះ	ɓ- ប ប៊	ɗ- ដ ឌ			
ព្យញ្ជនៈខ្យល់ស្ទះត្រដុស			tɕ- ច ជ		
ព្យញ្ជនៈច្រមុះ	-m ម ម៉	-n ន ណ	-ɲ ញ ញ៉	-ŋ ង ង៉	
ព្យញ្ជនៈញ័រ		r- រ រ៉			
ព្យញ្ជនៈខ្យល់ត្រដុស		s- ស ស៊			-h ហ ហ៊
ព្យញ្ជនៈខ្យល់ត្រដុសបើក	-β̞ វ វ៉		-j យ យ៉		
ព្យញ្ជនៈខ្យល់ត្រដុសបើកខាង		-l ល ឡ			

b → ɓ (**បប៊**)

d → ɗ (**ដឌ**)

Plosive → implosive

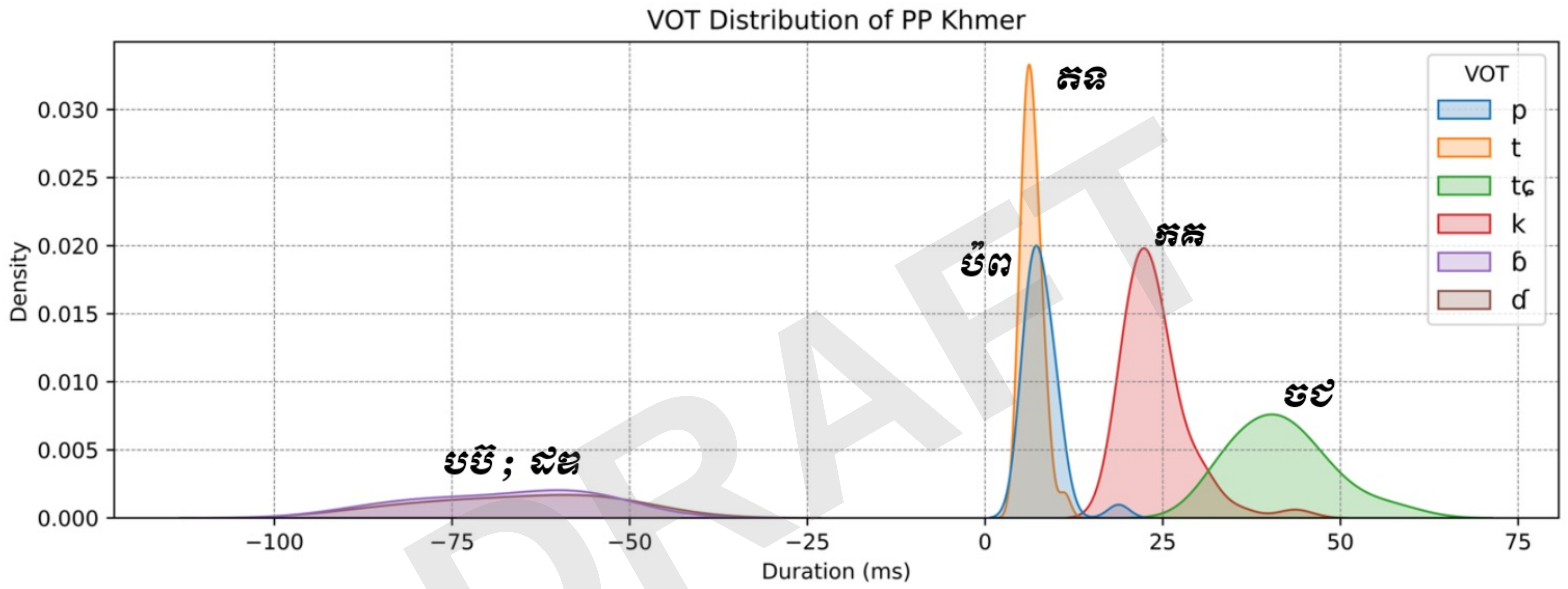
c → tɕ (**ចជ**)

Plosive → affricate

w, v → β̞ (**វវ៉**)

ធ្មេញលើ លើបបូរក្រោម
មកជាបបូររាងមូលវិញ។

DRAFT



រូបភាពទី ៖ រង្វាស់តម្រង់ជា ប្រអប់ ជាចំណុចទិន្នន័យ និង ជាវីយូឡុង នៃរង្វាស់ទីអូធី សម្រាប់អម្ពូរ ព្យញ្ជនៈខ្យល់ស្ទះផ្ទុះ(Implosive) មានសូរឃោសៈ ***បប៊ ដឌ*** *[ɓ ɗ] និងព្យញ្ជនៈអម្ពូរខ្យល់ស្ទះ ឥតសូរឃោសៈ (Voiceless)* ***ប៉ព តទ ចជ កគ*** *[p t tɕ k] ដោយថែម រង្វាស់ រយៈពេលបន្លឺសូរសំឡេង* ***ផភ ថធឆឈ ឆឈ ខឃ*** *[ph, th, tɕh, kh] ។*

DRAFT

ព្យញ្ជនៈខ្មែរ៖ និយមន័យ និងរចនាសម្ព័ន្ធ

1. **សញ្ញា ព្យញ្ជនៈ មានច្រើន(៣២ ឬ ៣៣) តែសូរសំឡេងតិច(១៧-១៩)**
   - ក និង គ មានសំឡេងតែមួយ /k/
2. **ព្យញ្ជនៈផ្សំគ្នា ដោយគ្មានវត្តមានរបស់ស្រៈ តែអាចបញ្ចេញសំឡេងបាន បង្កើតជាពាក្យបាន**
   - កងពល
3. **ព្យញ្ជនៈដើមព្យាង្គ មាន ១៧ យ៉ាងតិច តែព្យញ្ជនៈចុងព្យាង្គ មាន ១៣ ប៉ុណ្ណោះ**
   - ទេស s→h   កុដិ ɗ→t   កូរ r→ ɸ   កាប ɓ→p
4. **ព្យញ្ជនៈផ្សំគ្នា [ព+ព] មានចំនួន ៨៥ បន្សំ (ព្យញ្ជនៈផ្សំពីរ មាន ៨២ និង ផ្សំបី មាន ៣)**
   - ស្ត្រី សាស្ត្រា ល្ខោន ……

DRAFT

ស្រៈខ្មែរ៖ និយមន័យ និងរចនាសម្ព័ន្ធ

ស្រៈ គឺជាសូរសព្ទដែលខ្យល់ហូរចេញដោយសេរី។ ភាសាខ្មែរមានប្រព័ន្ធស្រៈធំល្បីល្បាញ — ប្រមាណ ៣០ ឯកតាស្រៈ រួមទាំងស្រៈឯក និងស្រៈផ្សំពីរ(diphthong) ដែលខុសគ្នាតាមប្រវែង (ខ្លី/វែង) និងសេរី (បន្លឺទី១/ទី២)។ *(Martini 1946, Henderson 1952, Huffman 1970, ឃួន សុខុម្ម 1970)*

២១~២៤	២៨~៣៦	២សេរី
ឯកតាសញ្ញាស្រៈ	សូរសំឡេងស្រៈ	មូលភាព អ អ៊

ស្រៈខ្មែរត្រូវបានចែកជាពីរ(a & b) ឆ្លុះបញ្ចាំងប្រព័ន្ធបន្លឺសំឡេងប្រវត្តិសាស្ត្រ (breathy/clear)។ *(Henderson 1952)*

អ៊ី →
- កី /kɛj/ → ɛ + j , ប្រជាប្រិយ (?)
- គី /kiː/ → iː , អាស៊ី (?)

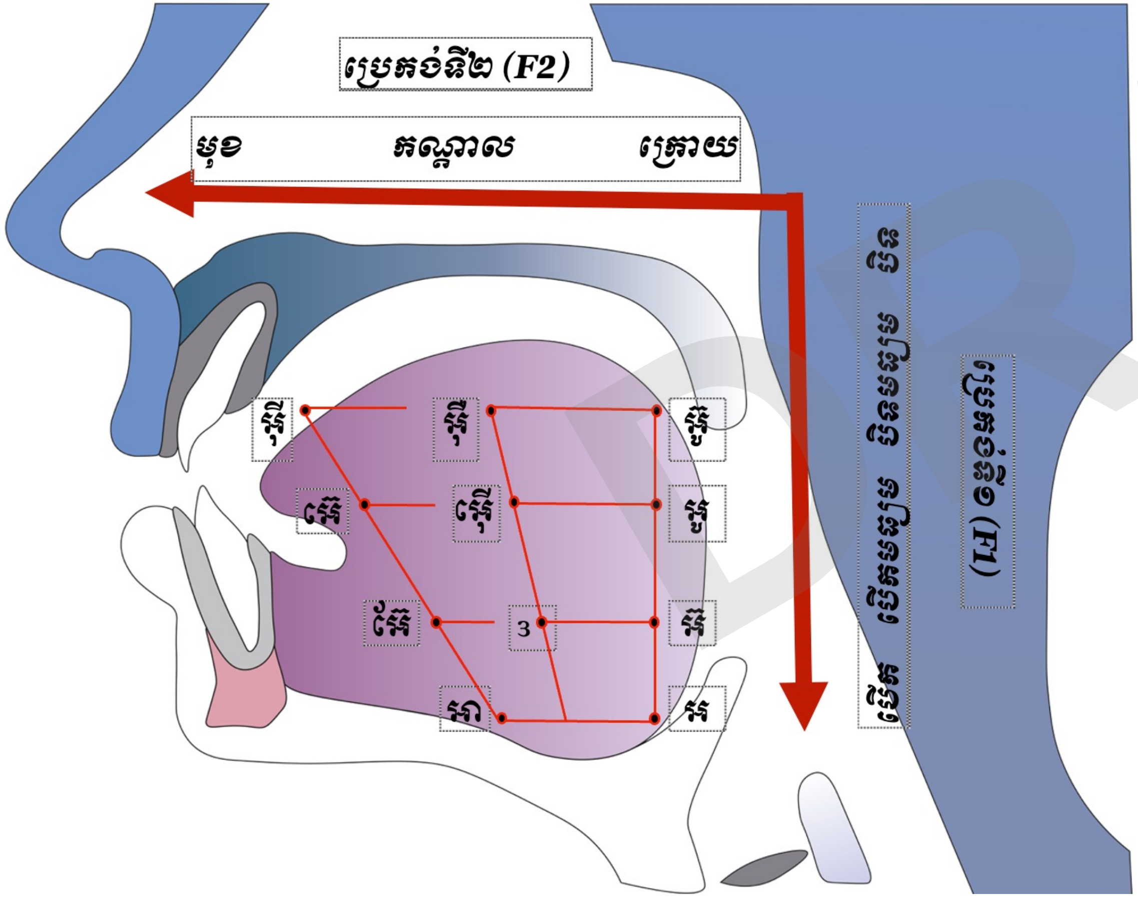
Wayland & Jongman (2003), J. Phonetics 31(2):181–201; Kirby (2014), J. Phonetics 43:69–85; Chem (2020), CJBAR 2(2).

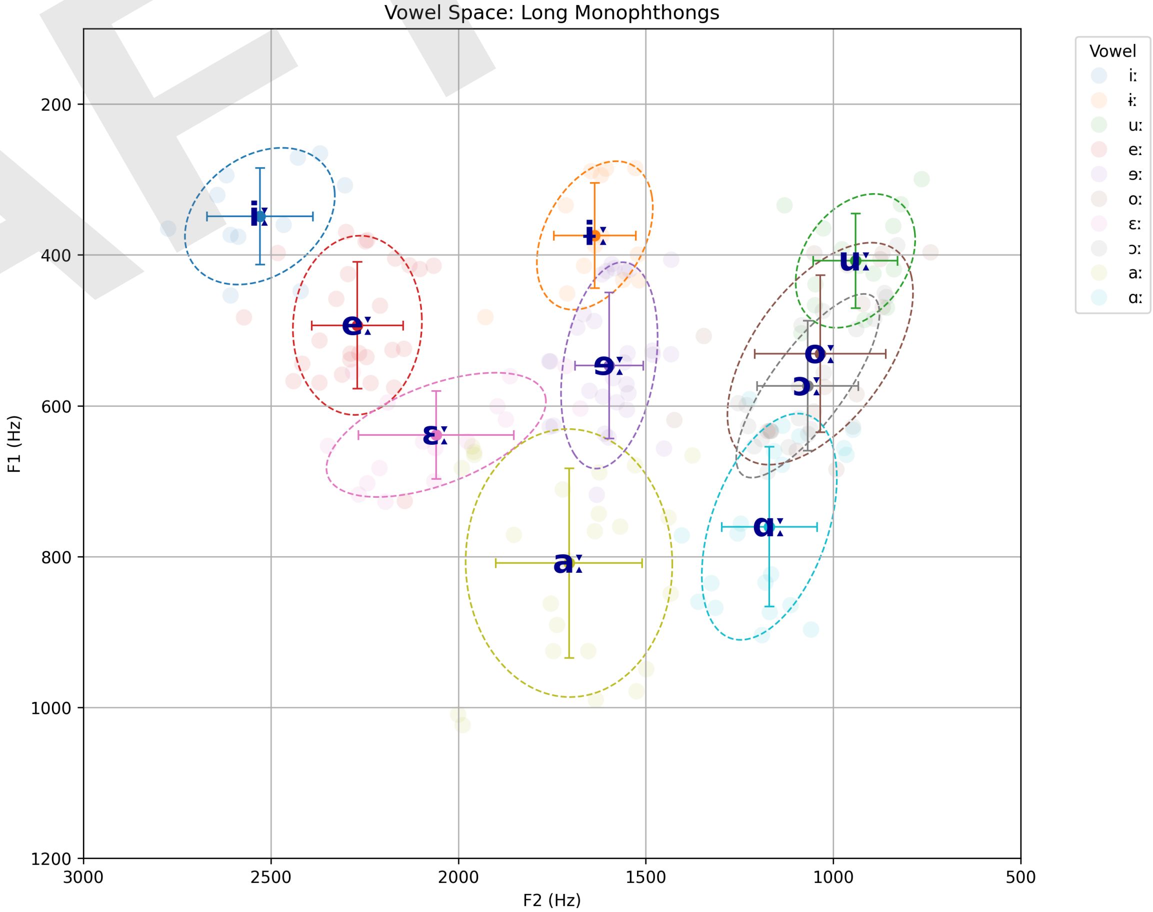
Vowel	1st Series Value		2nd Series Value[2]
អ	ɑː		ɔː
អា	aː		iɜ
អិ	e		i, ɪ
អី	ɛj		iː
អឹ	ɜ		ɨ
អឺ	əː		ɨː
អុ	o		u
អូ	oː		uː
អួ		uɜ	
អូវ	ɜβ		ɨβ
អើ	aɜ		əː
អឿ		ɨɜ	
អៀ		iə	
អេ		eː	
អែ	ae		ɛː
អៃ	aj		ɪj, ɨj
អោ	aɔ		oː
អៅ	aβ		əβ, ɨβ
អុំ	om		um
អំ	ɑm		um
អាំ	am		ɑm, oam
អះ	ah		eah
អិះ	eh		ih
អឹះ	ɜh		ɨh
អែះ	aɛh		ɛh
អុះ	oh		uh
អេះ	eh		ih, eh
អោះ	ɑh		uəh, ɔh

DRAFT

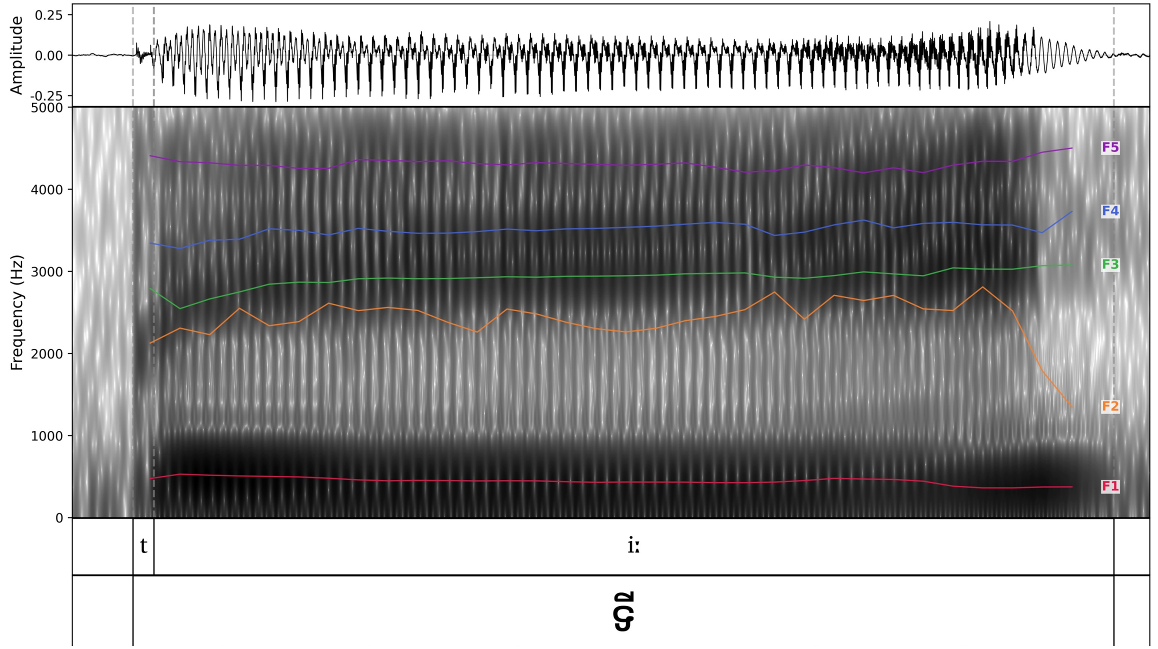
ស្រៈខ្មែរ៖ និយមន័យ និងរចនាសម្ព័ន្ធ

លំហស្រៈខ្មែរ (Khmer vowel space • F1 × F2)

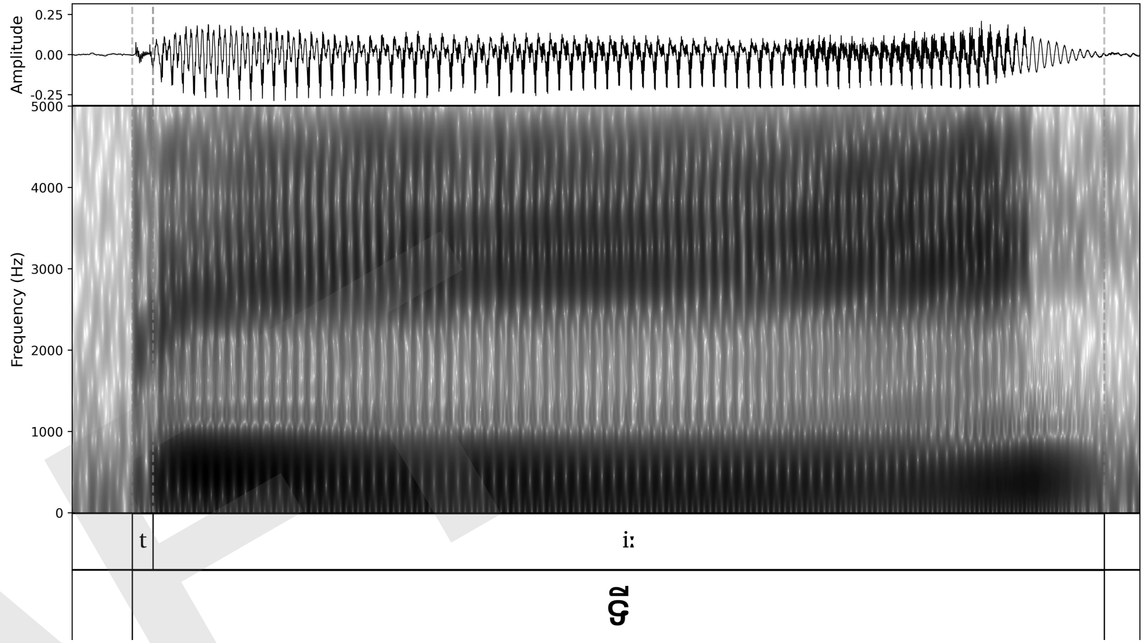




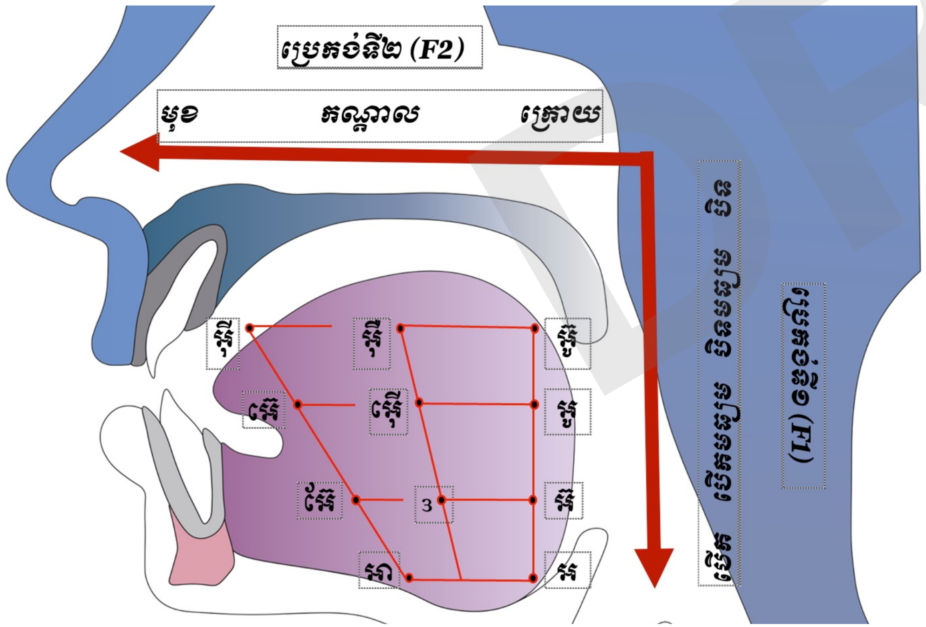
DRAFT



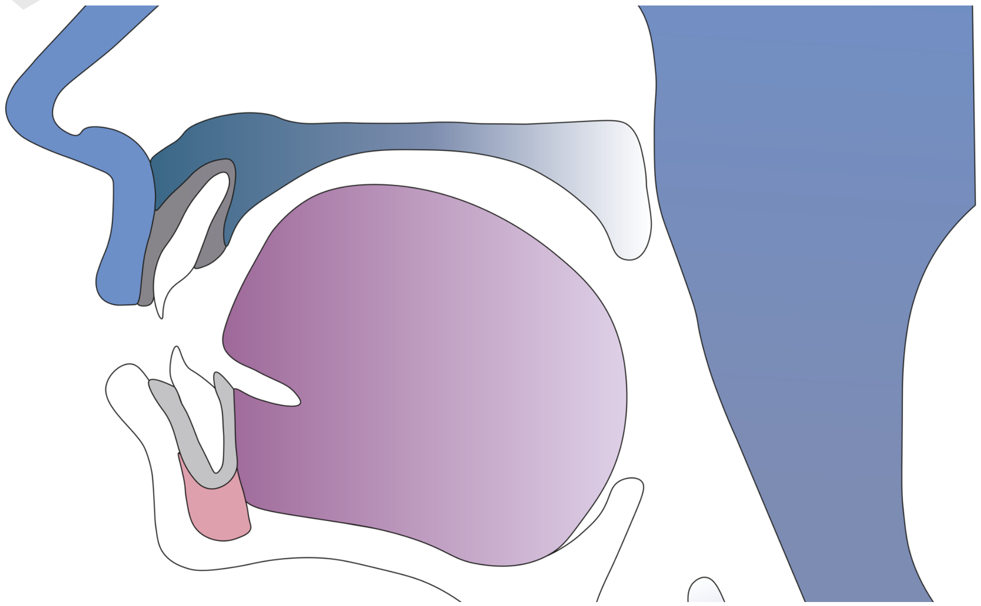
Amplitude
Frequency (Hz)
F5
F4
F3
F2
F1
t
iː



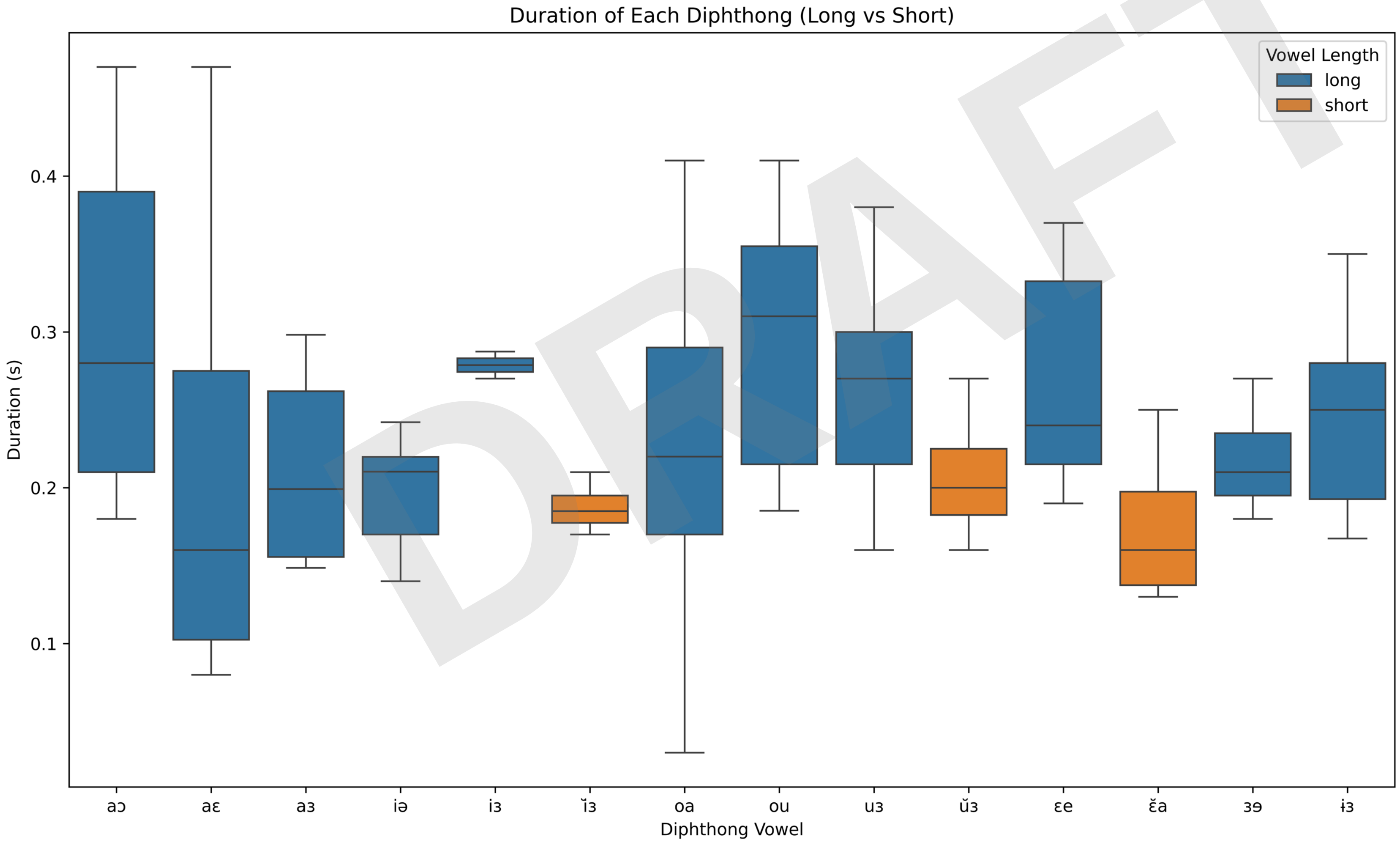
Amplitude
Frequency (Hz)
t
iː



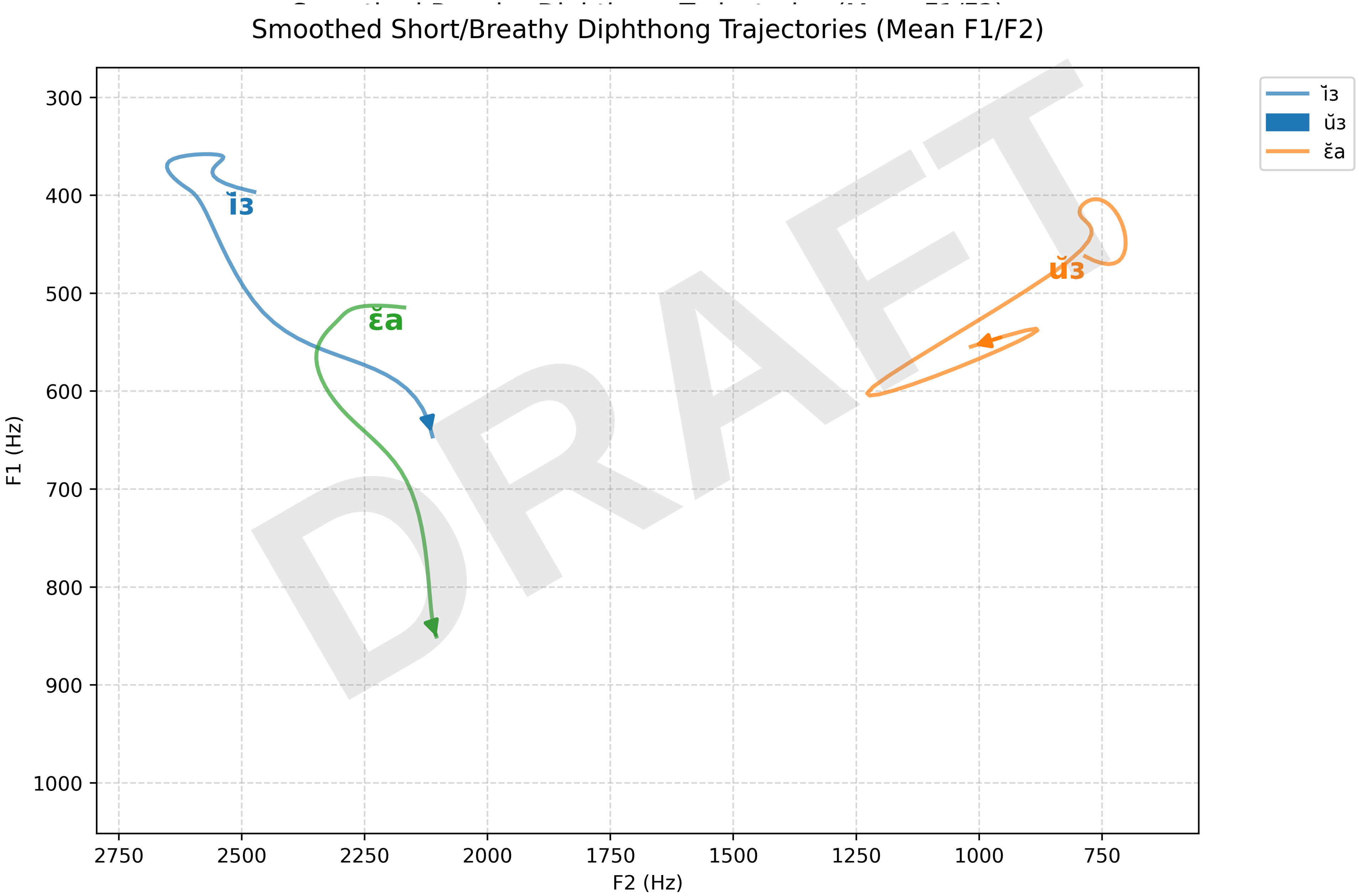
ប្រេកង់ទី២ (F2)
មុខ
កណ្ដាល
ក្រោយ
បិទ
បើក
ប្រេកង់ទី១ (F1)



ស្រះនោស និង ស្រះផ្សំពីរ

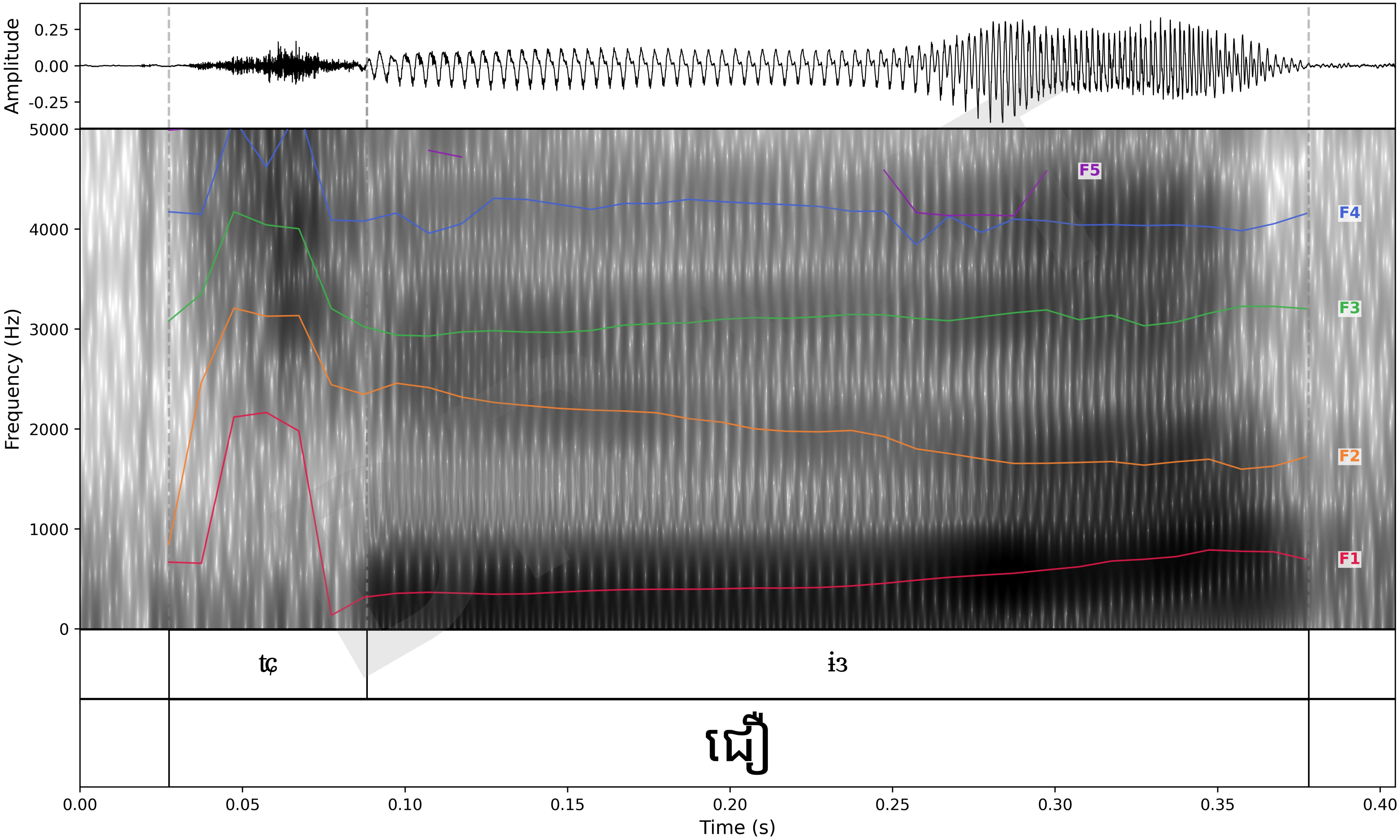


ស្រះផ្សំពីរ



Smoothed Short/Breathy Diphthong Trajectories (Mean F1/F2)
ĭɜ
ŭɜ
ɛ̆a
DRAFT
F1 (Hz)
F2 (Hz)
300
400
500
600
700
800
900
1000
2750
2500
2250
2000
1750
1500
1250
1000
750

ស្រះផ្សំពីរ

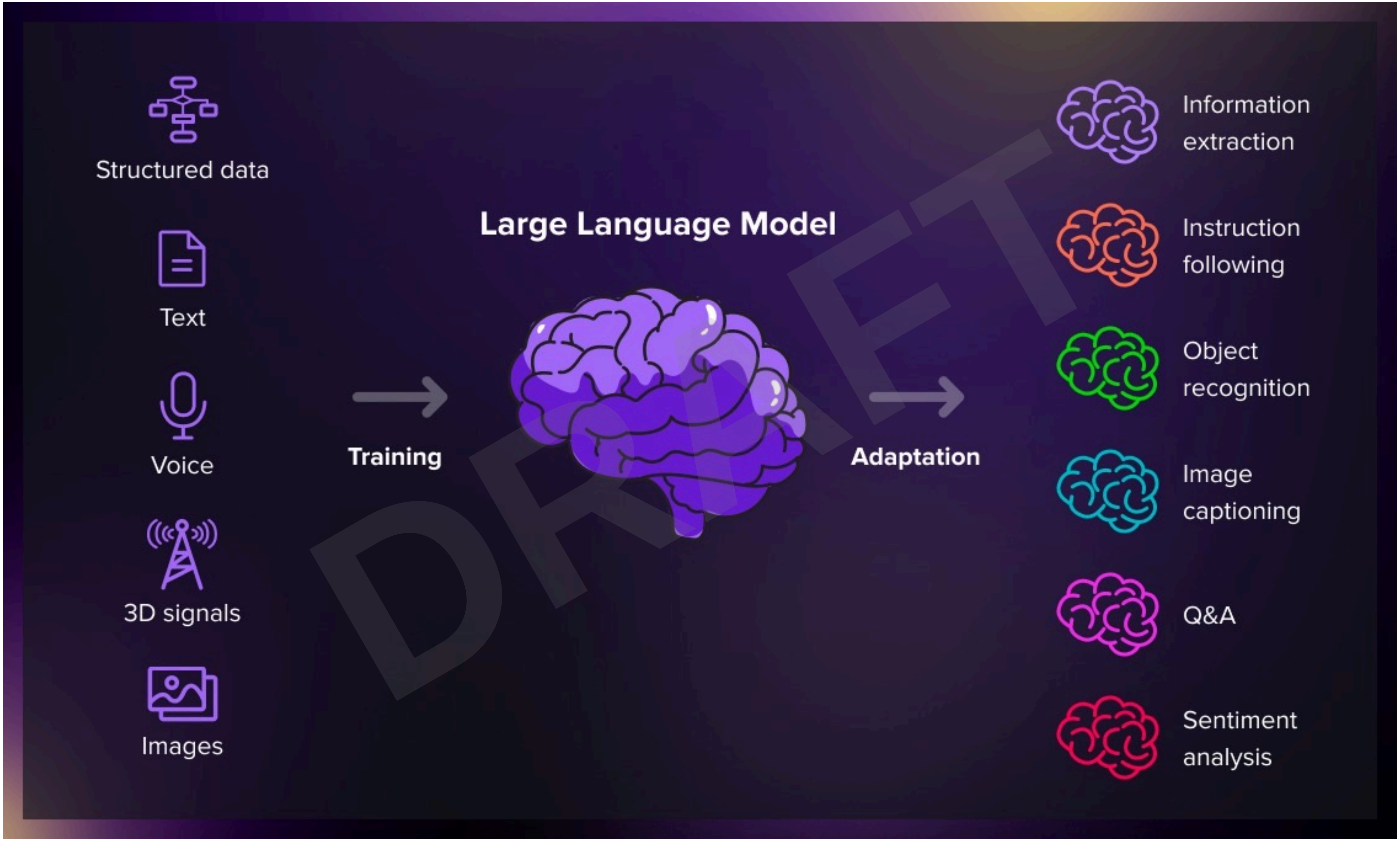

Amplitude
0.25
0.00
-0.25
Frequency (Hz)
5000
4000
3000
2000
1000
0
F5
F4
F3
F2
F1
tɕ
ɨə
ជឿ
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Time (s)

ភាសាខ្មែរនៅសម័យ AI

០៤

PART 04 · KHMER LANGUAGE STUDY IN THE AGE OF AI

DRAFT


Structured data
Text
Voice
3D signals
Images
Large Language Model
Training
Adaptation
Information extraction
Instruction following
Object recognition
Image captioning
Q&A
Sentiment analysis
DRAFT

អក្ខបទ

ខ្ញុំទៅផឹក ......

កាហ្វេ	70%
តែប៉ៃតង	30%
បាយ	0%

Token: ខ ្ ញ ុ ំ ទ ៅ ផ ឹ ក ក ា ហ ្ វ េ ។
Ids: 6 18 7 14 17 8 16 9 13 5 5 12 11 18 10 15 19

សំឡេង

អ [ɑː]

\+

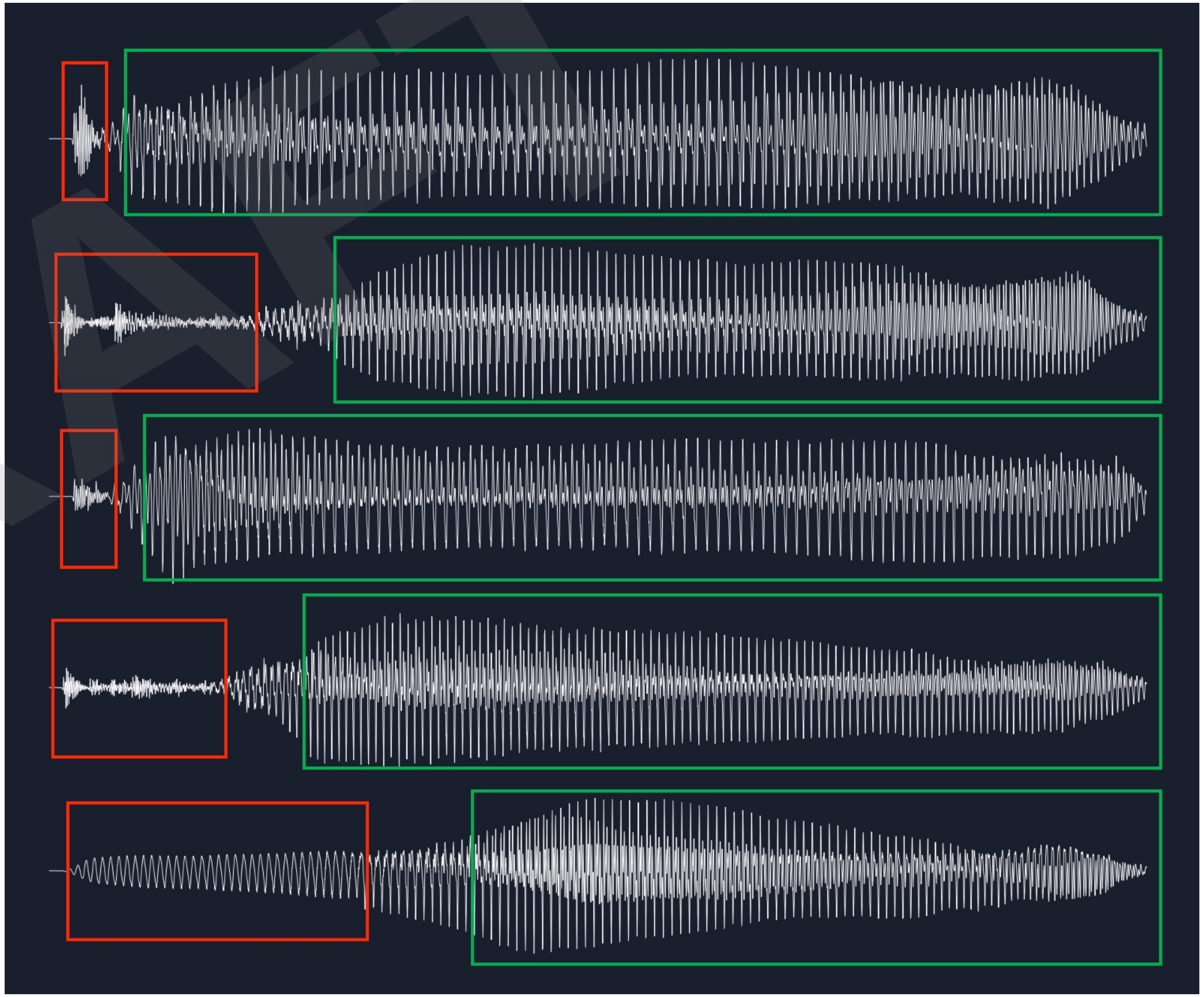
ក [k]

ខ

គ

ឃ

ង



DRAFT

៥០% នៃការផ្ញើសារជាសំឡេង ជាភាសាខ្មែរ

ហេតុអ្វីបានជាប្រទេសកម្ពុជាបង្កើតចរាចរណ៍សំឡេងពាក់កណ្ដាលនៃ Facebook Messenger ទូទាំងពិភពលោក?

នៅឆ្នាំ ២០១៨: Facebook បានរកឃើញថាអ្នកប្រើប្រាស់នៅកម្ពុជាបានបង្កើតចរាចរណ៍សំឡេងប្រហែល ៥០% នៃ Messenger ទូទាំងពិភពលោក។ គេបានសន្មតថាអក្ខរកម្មទាបជាមូលហេតុ ប៉ុន្តែអត្រាអក្ខរកម្មរបស់កម្ពុជាគឺមានរហូតដល់ ~៨០%។

ចម្លើយពិត: អក្សរខ្មែរមាន ៧៤ តួអក្សរ ដែលជាអក្ខរក្រមវែងជាងគេបំផុតនៅលើពិភពលោក។ ក្ដារចុចស្ដង់ដារមិនអាចផ្ទុកតួអក្សរទាំងអស់បានទេ ដែលតម្រូវឱ្យមានការប្ដូរស្រទាប់ (layer-switching) ជាបន្តបន្ទាប់។ ការវាយអក្សរខ្មែរលើទូរស័ព្ទដៃគឺមានការពិបាកខ្លាំង។

DRAFT

Digital Cambodia **2025**

Internet & Social Media Landscape

POPULATION
17.8M

MOBILE CONNECTIONS
25.3M
143% penetration

INTERNET USERS
10.8M
60.7% penetration

SOCIAL MEDIA USERS
12.9M
72.4% penetration

YoY Social Media Growth
+10.3%

PLATFORM USERS

Platform	Users	Share
Facebook	**12.9M**	72.4%
TikTok	**10.7M**	18+ users
Instagram	**1.65M**	9.3%
LinkedIn	**780K**	4.4%
X (Twitter)	**615K**	3.5%

Source: DataReportal Digital 2025, TRC (2025)

DRAFT

ទិន្នន័យភាសាខ្មែរគឺជាអ្វី?

«ទិន្នន័យភាសា» គឺជាសំឡេង និងអក្សរខ្មែរដែលត្រូវបានប្រមូល សម្អាត និងរៀបចំជាសារពើភណ្ឌ Corpus ដើម្បីឲ្យម៉ាស៊ីនអាចរៀនបាន។ ភាសាខ្មែរត្រូវបានចាត់ថ្នាក់ជា ភាសាមានធនធានតិច (low-resource) ដែលត្រូវការការវិនិយោគខ្លាំងផ្នែកស្រាវជ្រាវ។

DRAFT

Joshi et al. (2020), The State and Fate of Linguistic Diversity in the NLP World, Proc. ACL 2020:6282–6293.

ភាសាខ្មែរនៅសម័យ AI

ពីភាសាធនធានទាប ទៅភាសាដែលត្រៀមខ្លួនសម្រាប់បញ្ញាសិប្បនិម្មិត

ធាតុចូល

- សំឡេងនិយាយ
- រូបភាព / ឯកសារ
- អត្ថបទឌីជីថល

→

ប្រព័ន្ធ AI ភាសាខ្មែរ

ASR សម្គាល់សំឡេង	**TTS** សំយោគសំឡេង
OCR សម្គាល់អក្សរ	**NLP / LLM** ដំណើរការភាសា

→

ផលប៉ះពាល់

- ការអភិរក្សភាសា
- ការអប់រំ
- រដ្ឋាភិបាលឌីជីថល
- ភាពងាយប្រើ

បញ្ហាប្រឈមស្នូល៖ ស្គ្រីបស្មុគស្មាញ • គ្មានចន្លោះបំបែកពាក្យ • កង្វះ corpus ភាសា

DRAFT

ទិន្នន័យសម្ដី និងទិន្នន័យអក្សរ

ទិន្នន័យសម្ដី (Speech)

- ការថតសំឡេងនិយាយភ្ជាប់នឹងប្រតិចារិក (transcript)
- ការតម្រឹមសូរស័ព្ទដោយស្វ័យប្រវត្តិ (forced alignment)
- បម្រើការបង្កើតប្រព័ន្ធសម្គាល់សម្ដី (Khmer ASR) និងសំយោគសំឡេង (TTS)
- គំរូពហុភាសា (Wav2Vec2 · MMS) គ្របដណ្ដប់ភាសាខ្មែរហើយ

ទិន្នន័យអក្សរ (Text)

- អត្ថបទខ្មែរឌីជីថល · សារព័ត៌មានសរសេរដែលដាក់ស្លាក
- ការបម្លែងអក្សរស្វ័យប្រវត្តិ (Khmer OCR) ពីឯកសារស្កេន
- ការឌីជីថលភាវូបនីយកម្មឯកសាររដ្ឋបាល និងសលាចារឹកបុរាណ
- បម្រើការស្វែងរក · បកប្រែ · ប្រមូលផ្ដុំភាសា (NLP)

DRAFT

Pratap et al. (2024), Scaling Speech Technology to 1,000+ Languages, JMLR 25; Valy et al. (2017), SleukRith Set, Proc. HIP.

សេចក្ដីសន្និដ្ឋាន

០៥

PART 05 · CONCLUSION

DRAFT

ទិសដៅស្រាវជ្រាវ · អនាគត

អនាគតនៃការស្រាវជ្រាវភាសាខ្មែរដោយ AI

ឆ្ពោះទៅរកប្រព័ន្ធអេកូឡូស៊ីភាសាខ្មែរ ដែលគាំទ្រដោយបញ្ញាសិប្បនិម្មិត

DRAFT

គំរូភាសាធំខ្មែរ

បណ្តុះបណ្តាល foundation model សម្រាប់ការយល់ និងបង្កើតភាសាខ្មែរ

ធនធានបើកចំហ

ចងក្រង corpus រួម និងស្តង់ដារវាស់ស្ទង់សម្រាប់ភាសាធនធានទាប

អភិរក្សឯកសារបុរាណ

OCR និង AI សម្រាប់សិលាចារឹក និងសាស្ត្រាស្លឹករឹត

គ្រាមភាសា និងសូរស័ព្ទ

កត់ត្រាបម្រែបម្រួលតំបន់ និងបាតុភូតសូរស័ព្ទដោយស្វ័យប្រវត្តិ

ស្រាវជ្រាវពហុបែប

ផ្សំទិន្នន័យសំឡេង EEG និងចលនាភ្នែកក្នុង embedding តែមួយ

បច្ចេកវិទ្យាជំនួយ

TTS ការអានជំនួយ និងឧបករណ៍រៀនភាសាសម្រាប់ការអប់រំ

សេចក្ដីសន្និដ្ឋាន

ការសិក្សាភាសាខ្មែរនៅសម័យ AI មិនត្រឹមតែជាបញ្ហាបច្ចេកទេសប៉ុណ្ណោះទេ ប៉ុន្តែជាបញ្ហាយុទ្ធសាស្ត្រជាតិ។

ពីសូរវិទ្យាដល់សម្ដី

សូរវិទ្យាជាគ្រឹះនៃវិទ្យាសាស្ត្រភាសា — ការសិក្សាប្រកបដោយភាពម៉ត់ចត់ផ្ដល់មូលដ្ឋានសម្រាប់ប្រព័ន្ធសម្គាល់សម្ដីខ្មែរគុណភាពខ្ពស់

កាត់បន្ថយគម្លាតបច្ចេកវិទ្យា

ASR និង OCR ខ្មែរផ្ដល់សមភាពឱកាសដល់ប្រជាជនកម្ពុជាក្នុងការប្រើបច្ចេកវិទ្យាជាភាសាកំណើត

ការពារអត្តសញ្ញាណវប្បធម៌

ភាសាខ្មែរជាឃ្លាំងចំណេះដឹង និងប្រវត្តិសាស្ត្រ — AI ជួយផ្សព្វផ្សាយមរតកនេះទៅសកលលោក

ធនធានមនុស្ស និងសហការ

ភ្ជាប់ភាសាវិទ្យាប្រពៃណីនឹងវិទ្យាសាស្ត្រកុំព្យូទ័រ — សហការរវាងរាជបណ្ឌិត្យសភា សកលវិទ្យាល័យ និងរដ្ឋាភិបាល

DRAFT

ឯកសារយោង

Selected References

- Comrie, B. (1989). Language Universals and Linguistic Typology (2nd ed.). Oxford: Blackwell.
- Ferlus, M. (1992). Essai de phonétique historique du khmer. Mon-Khmer Studies, 21, 57–89.
- Haiman, J. (2011). Cambodian: Khmer. London Oriental and African Language Library 16. Amsterdam: John Benjamins.
- Henderson, E. J. A. (1952). The main features of Cambodian pronunciation. BSOAS, 14(1), 149–174.
- Huffman, F. E. (1970). Cambodian System of Writing and Beginning Reader. New Haven: Yale University Press.
- Jacob, J. M. (1968). Introduction to Cambodian. London: Oxford University Press.
- Jenner, P. N. (2009). A Dictionary of Pre-Angkorian / Angkorian Khmer. Canberra: Pacific Linguistics.
- Joshi, P., et al. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. Proc. ACL, 6282–6293.
- Kirby, J. P. (2014). Incipient tonogenesis in Phnom Penh Khmer. Journal of Phonetics, 43, 69–85.
- Maspéro, G. (1915). Grammaire de la langue khmère (cambodgien). Paris: Imprimerie nationale.
- Pratap, V., et al. (2024). Scaling Speech Technology to 1,000+ Languages. JMLR, 25(97), 1–52.
- Sidwell, P., & Blench, R. (2011). The Austroasiatic Urheimat. In N. J. Enfield (Ed.), Dynamics of Human Diversity, 315–343. Canberra: Pacific Linguistics.
- Valy, D., Verleysen, M., Chhun, S., & Burie, J.-C. (2017). A New Khmer Palm Leaf Manuscript Dataset – SleukRith Set. Proc. HIP, 1–6.
- Wayland, R., & Jongman, A. (2003). Acoustic correlates of breathy and clear vowels: Khmer. Journal of Phonetics, 31(2), 181–201.
- Chem, V. (2020). The Standard Khmer Vowel System: An Acoustic Study. Insight: Cambodia Journal of Basic and Applied Research, 2(2).
- ពុទ្ធសាសនបណ្ឌិត្យ. (១៩៦៧). វចនានុក្រមខ្មែរ (ស្នាដៃ សម្ដេចព្រះសង្ឃរាជ ជួន ណាត). ភ្នំពេញ.

DRAFT