Proceedings of the sixth International Conference on Asian Geolinguistics

EDITED BY
NOR HASHIMAH JALALUDDIN,
HIROYUKI SUZUKI, AND MITSUAKI ENDO

Geolinguistic Society of Japan

ISSN 2436-6471

Proceedings of the sixth International Conference on Asian Geolinguistics

EDITED BY
NOR HASHIMAH JALALUDDIN,
HIROYUKI SUZUKI, AND MITSUAKI ENDO

Geolinguistic Society of Japan
2025

Studies in Geolinguistics, Monograph series, No. 10
ISSN 2436-6471

Proceedings of the sixth International Conference on Asian Geolinguistics, edited by Nor Hashimah Jalaluddin, Hiroyuki Suzuki, and Mitsuaki Endo, 2025

Cover photo: Institute of the Malay world and civilisation © 2025 Mika Fukazawa

First published 2025
doi: https://doi.org/10.5281/zenodo.17204665

Published by: Geolinguistic Society of Japan
Website: https://geolinguistics.sakura.ne.jp/
Office address: Room G1305, Aoyama Gakuin Univeristy, 4-4-25 Shibuya, Shibuya-ku, Tokyo

Preface

This volume contains papers presented at the sixth International Conference on Asian Geolinguistics (ICAG) held at Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia, on the 5th of May, 2025. The previous ICAGs were held as follows: 1st, at Aoyama Gakuin University, Tokyo, 2012; 2nd, at Chulalongkorn University, Bangkok, 2014; 3rd, at Royal University of Phnom Penh, Phnom Penh, 2016; 4th, at Fakultas Ilmu Pengetahuan Budaya Universitas Indonesia, Jakarta, 2018; and 5th, at the University of Social Sciences and Humanities, VNU, Ha Noi, Vietnam, 2023.

The proceedings of each ICAG (PICAG) have been edited and published as open access documents; see below for bibliographical information. Before PICAG-5, the titles are denoted as *Papers from...*; as of PICAG-5, *Proceedings of...* is used.

PICAG-1: Endo, Mitsuaki (ed.) (2012) *Papers from the First International Conference on Asian Geolinguistics*. https://doi.org/10.5281/zenodo.6423581

PICAG-2: Endo, Mitsuaki (ed.) (2014) *Papers from the Second International Conference on Asian Geolinguistics*. https://doi.org/10.5281/zenodo.6423601

PICAG-3: Endo, Mitsuaki (ed.) (2016) *Papers from the Third International Conference on Asian Geolinguistics*. Fuchu: Research Institute for Languages and Cultures of Asian and Africa. https://publication.aa-ken.jp/papers_3IC_Asian_geolinguistics_2016.pdf

PICAG-4: Suzuki, Hiroyuki and Mitsuaki Endo (eds.) (2018) *Papers from the Fourth International Conference of Asian Geolinguistics*. Fuchu: Research Institute for Languages and Cultures of Asia and Africa. https://publication.aa-ken.jp/papers_4IC_Asian_geolinguistics_2018.pdf

PICAG-5: Trịnh, Cẩm Lan, Trần Thị Hồng Hạnh, Hiroyuki Suzuki, and Mitsuaki Endo (eds.) (2023) *Proceedings of the fifth International Conference on Asian Geolinguistics*. Tokyo: Geolinguistic Society of Japan. https://doi.org/10.5281/zenodo.8382130

The editors

TABLE OF CONTENTS

CONFERENCE PROGRAM

The Sixth International Conference on Asian Geolinguistics

Venue: Sudut Wacana, Level 3, Institut Alam dan Tamadun Melayu,
Universiti Kebangsaan Malaysia Kuala Lumpur, Malaysia
Online: Zoom Meeting

5th May 2025

Opening ceremony 8:55 – 9:00 Malaysia time
Nor Hashimah Jalaluddin (Universiti Kebangsaan Malaysia)
Welcoming speech

First session 9:00 – 10:30
Chair: *Nor Hashimah Jalaluddin (Universiti Kebangsaan Malaysia)*

1. *Chitsuko Fukushima (University of Niigata Prefecture)*,
   Revisiting 'Sun' in the Linguistic Atlas of Asia: Focusing on semantic extensions and differentiation
2. *Ian Joo (Otaru University of Commerce) and Yu-Yin Hsu (The Hong Kong Polytechnic University)*,
   Correlation between longitude and maximal lengths of onset and coda in Eurasia
3. *Mitsuaki Endo (Aoyama Gakuin University)*,
   Homelands, migration, and dispersion of the Kra-Dai

Second session 10:45 - 11:45
Chair: *Trịnh Cẩm Lan (University of Social Sciences and Humanities, VNU)*

4. [online] *Li Jinhua and Wang Lei (Nanjing University),*
   Geolinguistic Analysis of the Word Form for 'Cat' in Chinese Korean
5. *Trần Thị Hồng Hạnh and TRUONG Nhat Vinh (University of Social Sciences and Humanities, VNU),*
   A geolinguistic analysis of the word form for 'frog and toad' among Mon Khmer languages in Vietnam
6. [online] *Hiroyuki SUZUKI (Kyoto University),*
   Geolinguistic analysis of the word form for 'butterfly' in Tibetic languages of the eastern Tibetosphere

Third session 13:00 – 14:30
Chair: *Mitsuaki Endo (Aoyama Gakuin University)*

7. *Mika Fukazawa (National Ainu Museum),*
   Extraction of regularities and geographical patterns from the basic vocabulary of the Ainu language
8. [online] *Xiao, Nengping and Zeng, Xiaoyu (Nankai University),*
   On the nature of non-final T4 in Yangzhou (扬州) dialect
9. *Trịnh Cẩm Lan (University of Social Sciences and Humanities, VNU),*
   The Characteristics and distribution of l and n variants in the Red River delta, Vietnam

Fourth session 14:45 – 16:15
Chair: *Khairul Ashraaf Saari (Universiti Poly-Tech Malaysia)*

10. *Coretta Herliana Kura (Dewan Bahasa dan Pustaka), Nor Hashimah Jalaluddin and Siti Noraini Hamzah (Universiti Kebangsaan Malaysia),*
    The Distribution of Various Dialects in Lundu, Sarawak
11. *Atsuko Utsumi (Meisei University),*
    Diversity and distribution of Case-marking systems in the Philippines and Malay-speaking world

Fifth session 16:30 – 17:30
Chair: *Chitsuko Fukushima (University of Niigata Prefecture)*

12. *Sri Munawarah and Nazarudin (Universitas Indonesia),*
    Language Distribution and Variation in Urban Depok, Indonesia: A Geolinguistic Perspective
13. *Khairul Ashraaf Saari (Universiti Poly-Tech Malaysia), Nor Hashimah Jalaluddin, and Harishon Radzi (Universiti Kebangsaan Malaysia)*
    Dialect transition along the Perak river

❦

Revisiting 'sun' in the *Linguistic Atlas of Asia*: Focusing on semantic extensions and differentiation

Chitsuko Fukushima (University of Niigata Prefecture)

Abstract: The comments on 'sun' in Asia (Fukushima 2021) did not include maps for semantic extensions and differentiation, so mapping was attempted. The motivational map of 'sun' in the *Linguistic Atlas of Asia* shows the distribution of basic meanings, 'etymological motivations' and different categorical systems. The maps focusing on specific semantic variation tell us what changes happened in the area. Many languages in Southeast Asia have words meaning 'sun' and 'day'; thus polysemy. The compound forms in Southeast Asia and also in Turkic are the evidence for the change to monosemy.*

Key words: motivational map, etymological motivation, polysemy, monosemy, compound

1. Introduction

The author was involved in the cooperative research to make the *Linguistic Atlas of Asia* published in 2021. A group of researchers who were in charge of a language/ language group in Asia drew linguistic maps. The map of 'sun' in Asia is a synthesized map such as Figure 1 (Endo 2021). Comments on 'sun' in Asia that the author wrote (Fukushima 2021) included four sections: 1) the overall description, 2) semantic extensions and differentiation, 3) diffusion of word-formation patterns, and 4) reverential forms and sun worship. In the first section, Table 1 showed languages with the oldest monosyllabic forms, followed by disyllabic and compound forms: for example, *hi, taiyo, otentosama, nichirin*, etc. in Japanese (see Table 1). The second section included the following description but no maps: "The meaning of the word form denoting 'sun' is often extended to mean 'day.' Thus, the word forms for 'sun' and 'day' are not always distinguished clearly" (Fukushima 2021: 19). The third section

FUKUSHUMA, Chitsuko. 2025. Revisiting 'sun' in the *Linguistic Atlas of Asia*: Focusing on semantic extensions and differentiation. In Nor Hashimah Jalaluddin, Hiroyuki Suzuki and Mitsuaki Endo (eds.) *Proceedings of the sixth International Conference of Asian Geolinguistics*, 1–17. doi: https://doi.org/10.5281/zenodo.17204519

* This work has been supported by Tokyo University of Foreign Studies, April 2015–March 2018 ILCAA Joint Research Project: Studies in Asian Geolinguistics (jrp000210), with Mitsuaki Endo as coordinator.

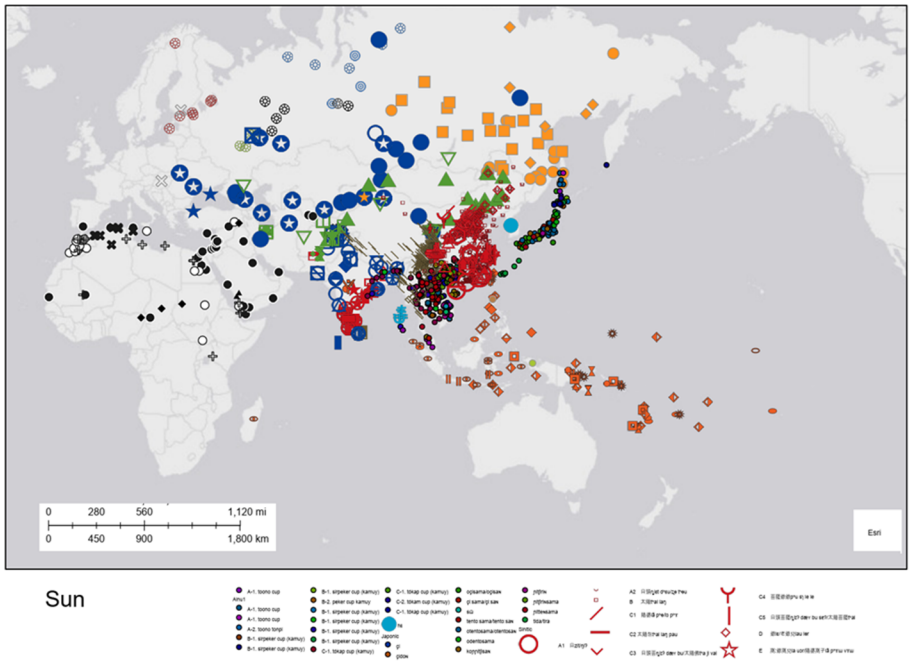


Figure 1: 'Sun' in Asia Source: Endo 2021

Table 1: Languages with oldest monosyllabic forms # loans from other languages

	monosyllabic	disyllabic	compound
Japanese	*hi*	*taiyō*# 太陽	*otentosama*#, *nichirin*# *etc.*
Korean	*hɛ*	*thɛjaŋ*# 太陽	*henim* reverential
Sinitics	*ri* 日	*Taiyang* 太陽	*+tou* 頭, *Bw-* 婆,菩,佛, *ye* 爺, *wo* 窩, *yan* 眼, *Kw-* 公, *di* 帝
Hmong-Mien	A *ɳV*1# B root is *naŋ* or *ntoŋ*	*ɳi tau*# *ni tau*# 日头(頭) or 热头(頭)	polysyllabic words + "hole," "father," "wife," "sky," "moon" etc.
Tibeto- Burman	Ax **nəy* A0 **g-nam*	various compound forms and plain forms	
Kra-Dai	*van*	*ta van* 'eye of day' *tang ugon* 'lamp of day'	
Austroasiatic	I A *ŋaj* oldest & others	"eye of day," "eye of sky," "eye of god" for all types	
Turkic	*kün*	*küneš*	*+karak* 'eye'

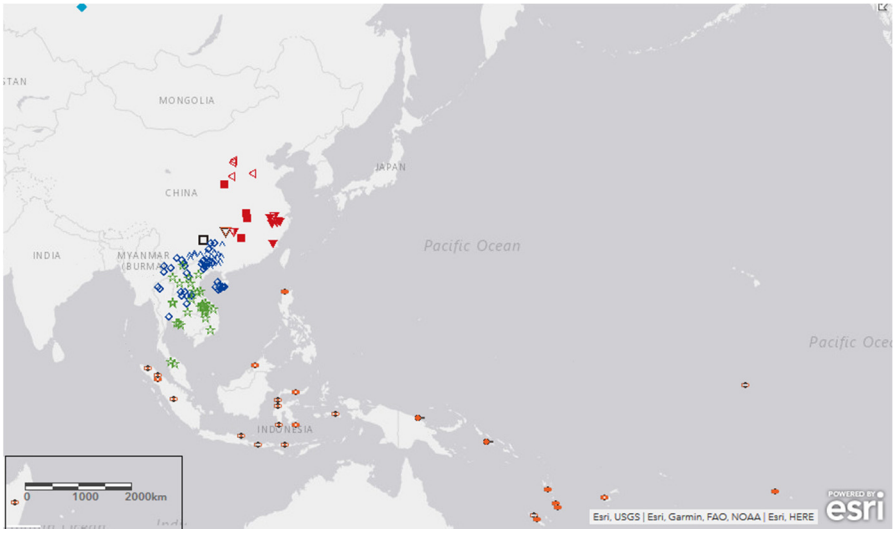


Figure 2: 'Eye' + 'day' compounds and their variants Source: Fukushima 2021

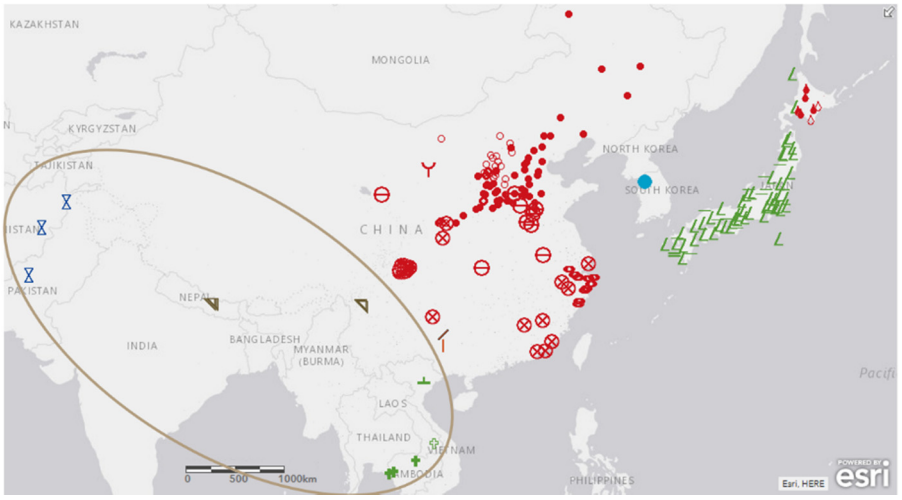


Figure 3: Reverential forms and use of names of god Source: Fukushima 2021

included a map of the 'eye' + 'day' compounds which are distributed in Southeast Asia (see Figure 2). The fourth section included a map of reverential forms or uses of personification distributed in East Asia and words denoting 'god, deity' or a name of the god used to mean 'sun' in South/Southeast Asia (see Figure 3; the latter feature circled in brown). These practices are seen as the evidence of sun worship.

The mapping of semantic extensions and differentiation in Asia was attempted and the results are reported here. ArcGIS online was used for mapping.

2. Semantic extensions and differentiation of 'sun' in Asia

2.1. Basic meanings of 'sun'

Meanings considered as basic meanings of 'sun' are shown below. Meanings b) and c) may be extended meanings. The word *hi* in Japanese has all these meanings.

a. sun/the star of the day — *hi ga noboru* 'the sun rises'
b. sunlight(sunshine)/the light of the day — *hi ga sashikomu* 'the sunlight shines into ...'
c. daytime/the bright time of the day — *hi ga nagakunaru* 'the days gets longer'
d. day — *yakusoku no hi* 'the day of appointment'

2.2. Distribution of semantic extensions and differentiation of 'sun' in Asia

2.2.1 A synthesized distribution

Based on the descriptions of 'sun' of languages/language groups in Asia, a motivational map was drawn to synthesize the variation of semantic extensions and differentiation of 'sun' in Asia (see Figures 4 & 5).

Figure 4: Languages/language groups in Asia

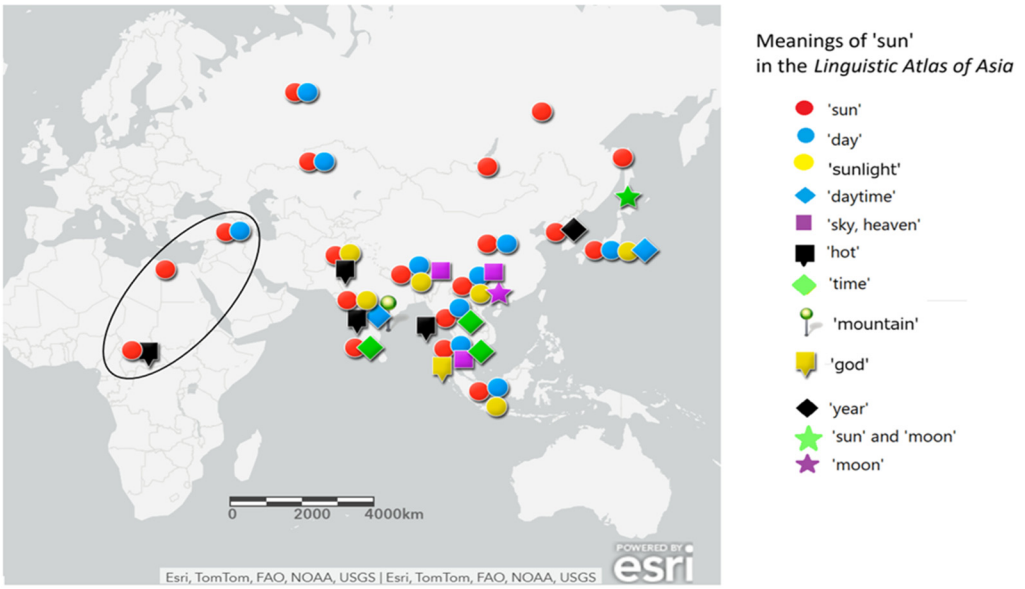


Figure 5: A synthesized map of semantic extensions and differentiation of 'sun' in Asia

Languages only with the meaning of 'sun' are those in Siberia, and those mostly with it are Arabic. The languages with the meaning of 'day' occupy the broad area in Asia, specifically Southeast Asia, East Asia, South Asia, and Western part of Eurasia. Languages with 'sunlight' and 'daytime' are found nearby 'day'.

Some of the variation such as 'sky, heaven', 'hot', 'time', 'mountain' and 'god' should rather be regarded as 'etymological motivations' (Tuaillon 1985: 7, motivations étymologiques). The words originally had these meanings and/or were part of a compound meaning 'sun'.

Others may be related with a different categorical system. For example, in Ainu, 'sun' is expressed by a compound which includes a noun meaning 'sun' and 'moon'. A Hmong-Mien language has a polysyllabic word with an element meaning 'moon'. In Korean, a word meaning 'sun' also means 'year' instead of 'day'.

2.2.2 Distributions of semantic extensions and differentiation of 'sun' in Asia

The maps showing the semantic distribution of 'sun' in Asia are introduced here. See Figure 6, a map of words meaning 'day'. Many languages from Southeast Asia to the central part of Eurasia have words which mean 'day'. These languages have words which mean 'sun' and 'day'; thus, polysemy.

See Figures 7 and 8, which show the words meaning 'sunshine, light' or 'daytime'. These distributions are found nearby that of 'day' shown by Figure 6. Figures 9-12

show the distributions of etymological motivations. Again, most distributions are found nearby that of 'day', specifically in Southeast and South Asia, except that in Arabic.

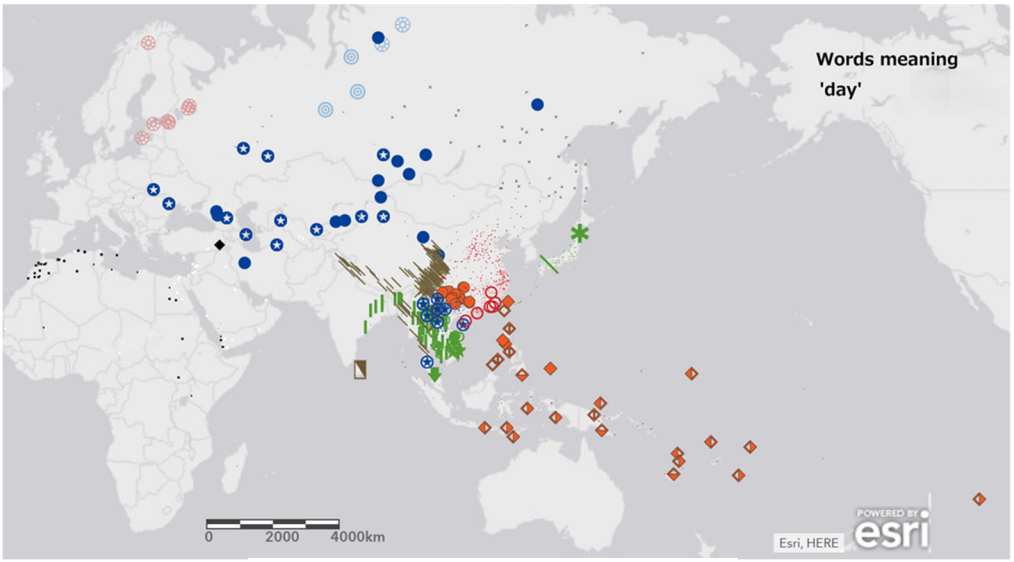


Figure 6: Words meaning 'day' in Asia

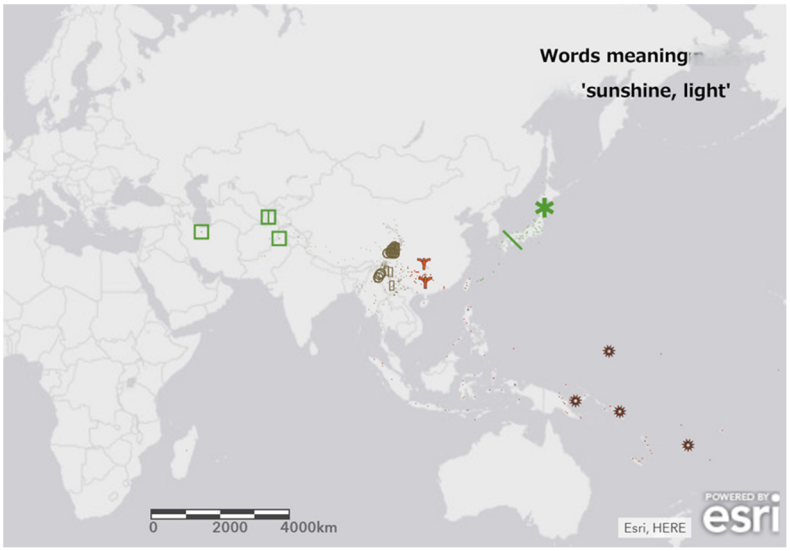


Figure 7: Words meaning 'sunshine, light' in Asia

Figure 8: Words meaning 'daytime' in Asia

Figure 9: Words meaning 'sky, heaven' in Asia

Figure 10: Words meaning 'time' in Asia

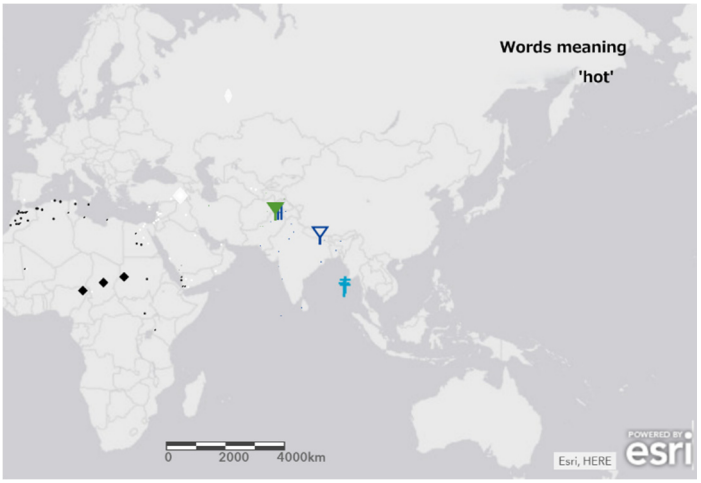


Figure 11: Words meaning 'hot' in Asia

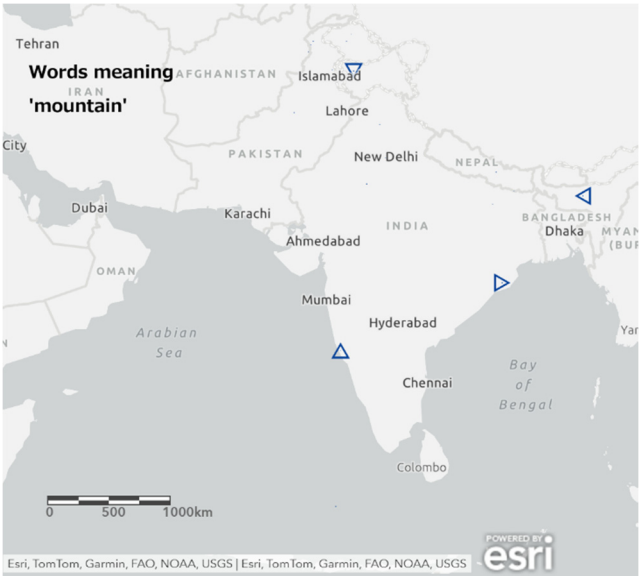


Figure 12: Words meaning 'mountain' in Asia

3. Semantic extensions and differentiation of 'sun' in individual languages/language groups in Asia

In this section, maps are introduced to show the variation in each language/language group. The comparison of maps shows the changes that happened.

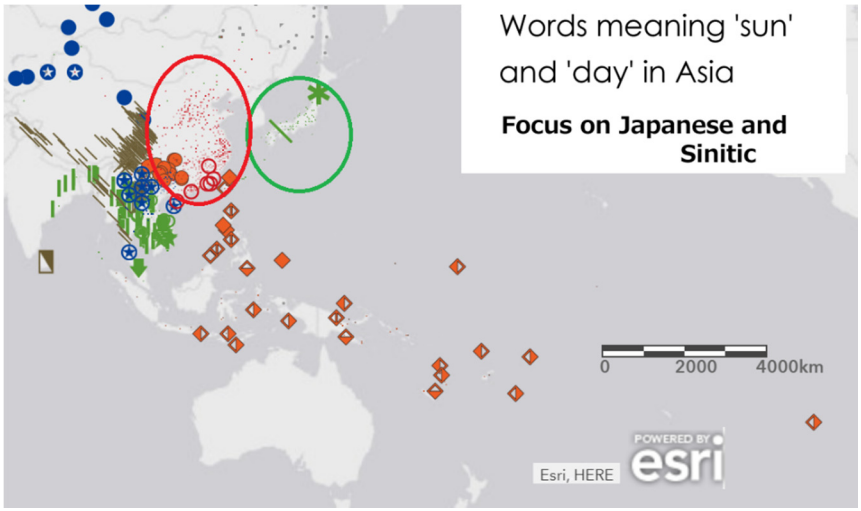


Figure 13: Words meaning 'day' in Asia: Focus on Japanese and Sinitic

3.1. Japanese and Sinitic

See Figure 13. Japanese shows a small distribution of *hi* which means 'day' since Chinese words (e.g. *taiyo*) and honorific words (e.g. *ohisama*) which mean only 'sun' are in everyday use. Sinitic shows the distribution of *ri* only in the South while disyllabic words and words of personification are used in many locations. This map shows the change from polysemy to monosemy in Japanese and Sinitic.

3.2. Austronesian

Figure 14 shows the distribution of words which mean 'day' and 'light'. Figure 15 shows the distribution of *mata* 'eye' and its compounds which mean only 'sun'. The distribution of Figure 15 seems newer compared with that of Figure 14. These maps also show the change from polysemy to monosemy.

3.3. Austroasiatic

Figure 16 shows the semantic variation and Figure 17 shows the distribution of the compound type. The maps also show the change from polysemy to monosemy.

3.4. Kra-Dai

Figure 18 shows the semantic variation of 'day' and 'time'. Figure 19 shows the distribution of the word meaning 'eye' and its compounds. The maps again show the change from polysemy to monosemy.

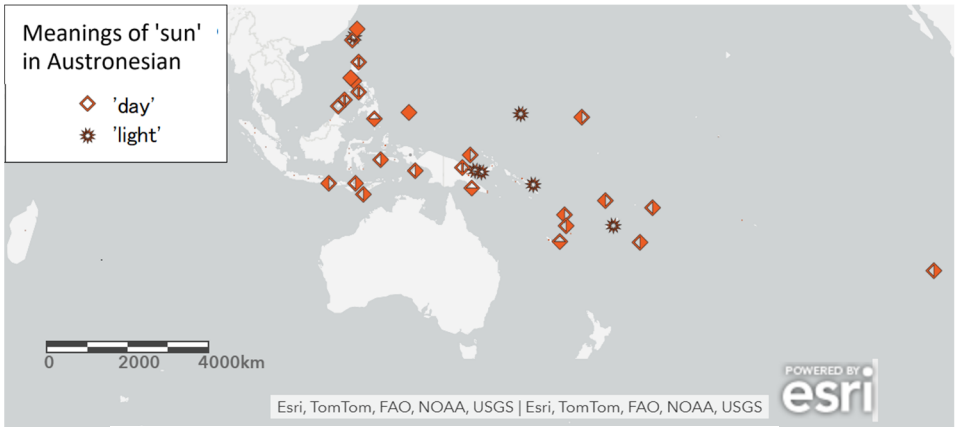


Figure 14: Semantic variation of 'sun' in Austronesian

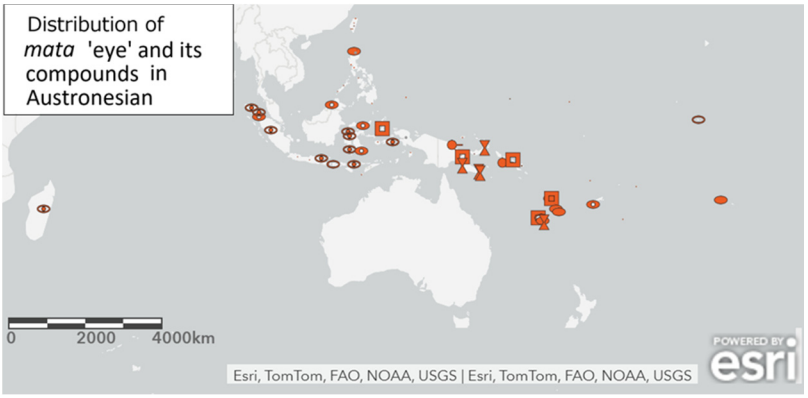


Figure 15: Distribution of *mata* 'eye' and its compounds in Austronesian

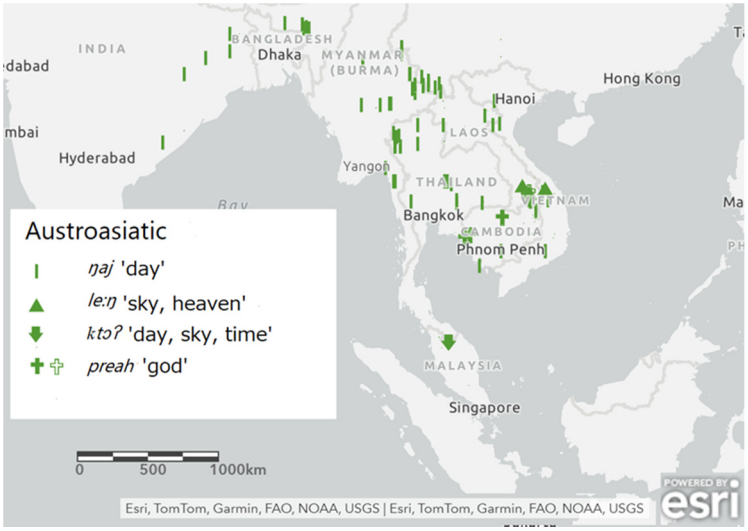


Figure 16: Semantic variation of 'sun' in Austroasiatic

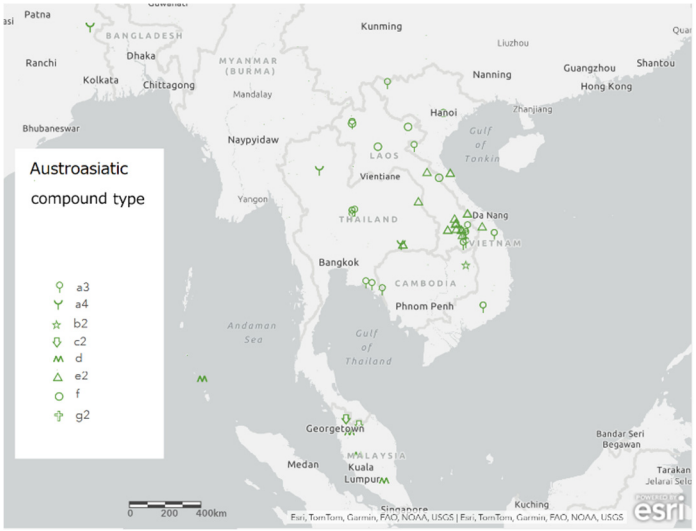


Figure 17: Distribution of the compound type in Austroasiatic

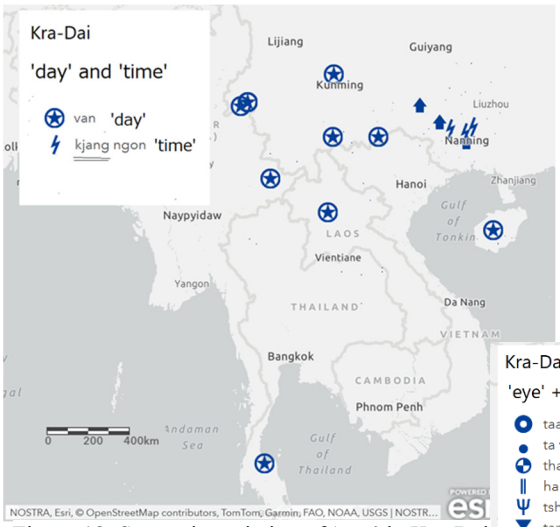


Figure 18: Semantic variation of 'sun' in Kra-Dai

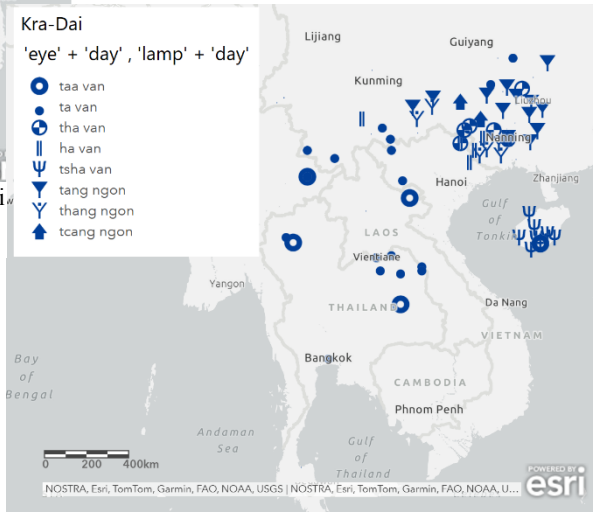


Figure 19: Distribution of 'eye' and its compounds in Kra-Dai

3.5. Tibeto-Burman

Figure 20 shows the following semantic variation: 'day, dwell', 'sunshine, bright', 'sky, heaven, clouds', and 'set (of the sun)'. Figure 21 shows the distribution of various types of compounds in the south-eastern part of the Tibeto-Burman area. The distribution of the compound type in Tibeto-Burman (Figure 21) is in succession with those of the compound type in Kra-Dai (figure 17) and in Austroasiatic (Figure 19).

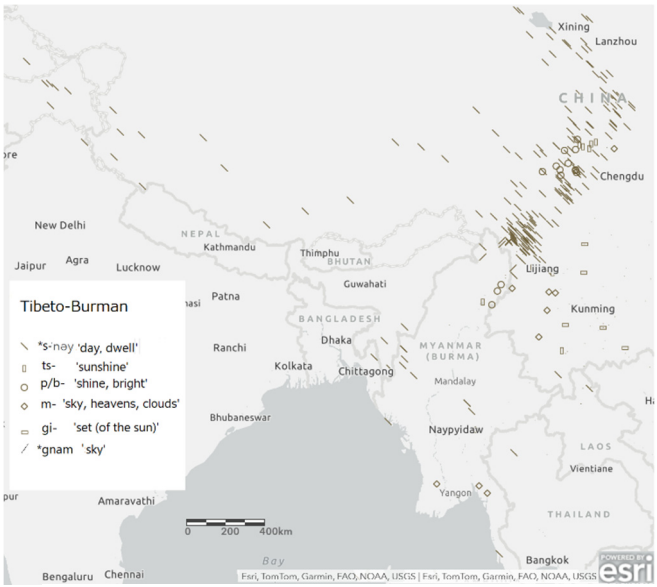


Figure 20: Semantic variation of 'sun' in Tibeto-Burman

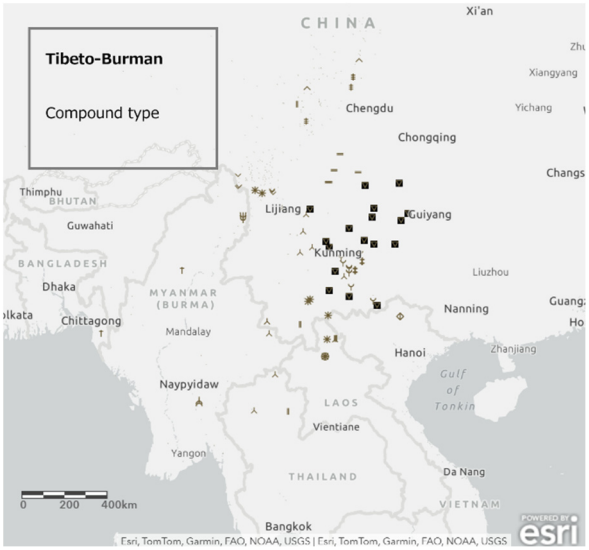


Figure 21: Distribution of the compound type in Tibeto-Burman <enlarged>

3.6. Turkic

Figures 22 shows the semantic variation in Turkic. The change from 'sun' and 'day' to 'sun' is attested. According to Saito (2021: 48):

> The form *kün*, which originally meant 'sun,' has extended to refer also to 'day.' … Chulym developed a compound word *kün kara ɣ ï* with the word *karak* 'eye.' … The type B word (*küneš* type) is used (almost?) exclusively for 'sun' in Turkish, Crimean Tatar, and Chuvash, while the type A word (*kün* type) is used for both 'sun' and 'day' in the other languages.

The 'eye of the day' compound is used in two distant locations (Chulym and Nogai[1]). This information is valuable since Urban (2010) identified the linguistic pattern only in the language families of Southeast Asia and Oceania.

Figure 22: Semantic variation of 'sun' in Turkic

3.7. Uralic

Figure 23 shows the semantic variation in Uralic. The languages with the meaning of 'day' are located in the north. This is a peripheral distribution, so the polysemy is older.

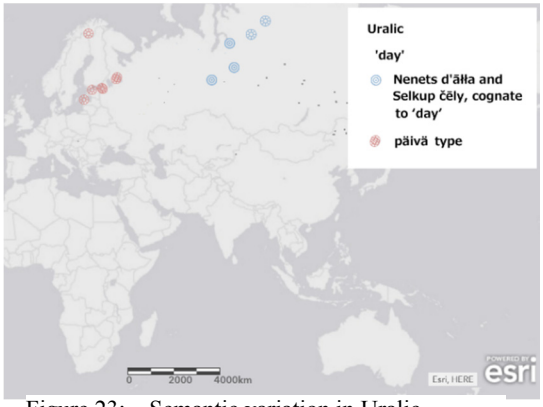


Figure 23: Semantic variation in Uralic

[1] *kün közi* (*köz* 'eye'). This information of Nogai has been recently provided by Prof. Yoshio Saitô.

3.8. South Asia

Figures 24 to 27 show the semantic variation in each language group in South Asia. Figure 24 shows the distribution of 'sunshine' and 'hot' in Iranian, Figure 25 that of 'sunshine', 'daytime', 'mountain', and 'hot' in Aryan, Figure 26 that of 'time' in Dravidian, and Figure 27 that of 'hot' in Andaman.

Figure 24: Semantic variation of 'sun' in Iranian

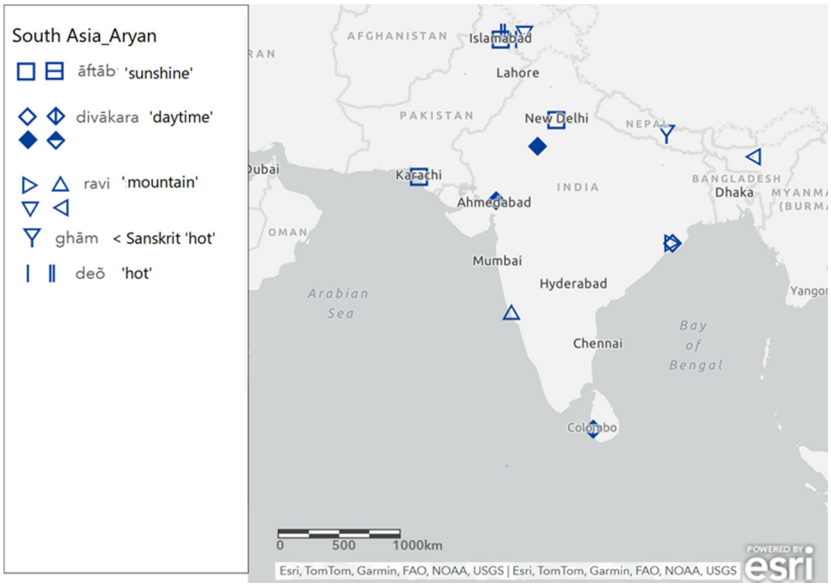


Figure 25: Semantic variation in Aryan

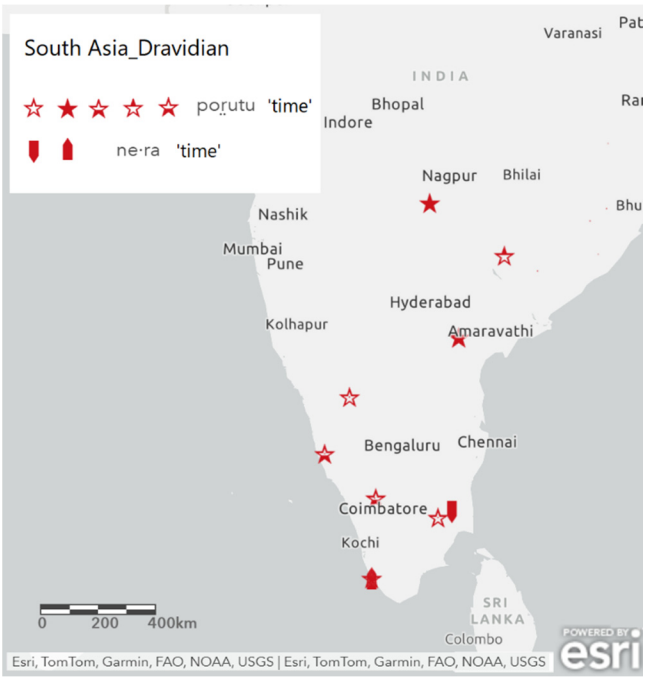


Figure 26: Semantic variation in Dravidian

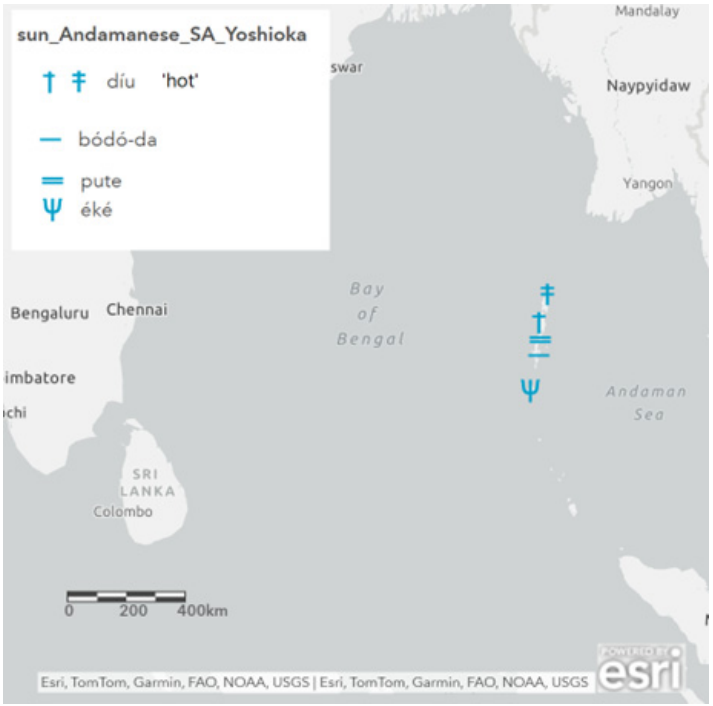


Figure 27: Semantic variation in Andaman

3.9. Arabic

Figures 28 shows the semantic variation in Arabic: 'hot' in three localities in Africa, but 'day' at one locality in Turkey. Most Arabic languages have only the meaning of 'sun'.

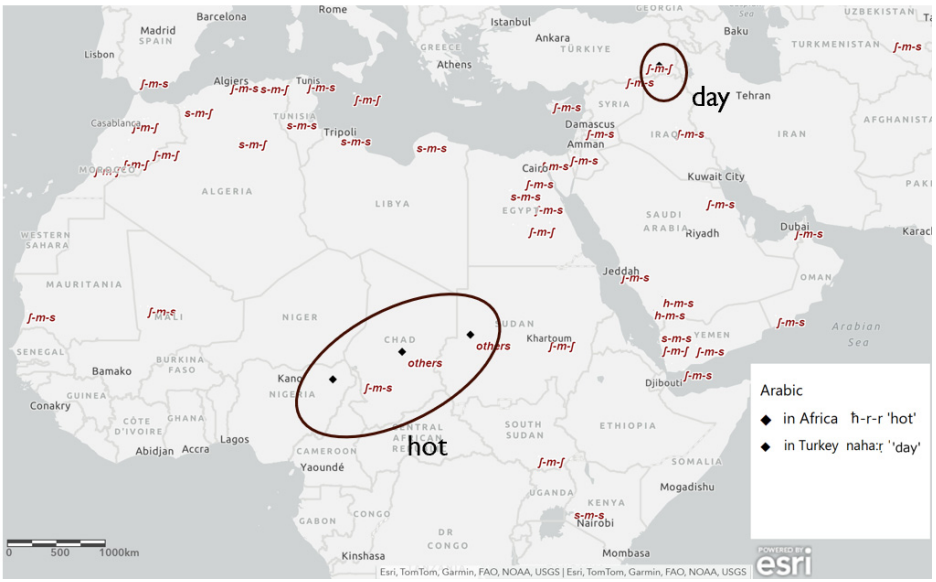


Figure 28: Semantic variation in Arabic

4. Conclusion

Based on the maps shown above, Table 1 and Figures 2 and 3 in Fukushima (2021) can be reinterpreted. Table 1: Monosyllabic forms are often polysemous, and disyllabic forms or compound forms are monosemous. Thus, the change from polysemy to monosemy is suggested by the table. Figures 2 and 3: The compounds and reverential forms are monosemous. Thus, the change from polysemy to monosemy is again suggested by these figures[2].

In the Introduction to the *Atlas linguarum Europae*, Alinei introduced two developments for analysis of lexicon (Alinei 1983: XX): onomasilogy (the study of different names for the same referent in a given area) and semasiology (the study of different meanings of the same etymon in a given area). The maps introduced today are drawn from the semasiological point of view and are motivational maps which show

[2] Most maps except Figures 4 & 5 were made by deleting irrelevant languages and words from the original linguistic map. This is the method used in Fukushima (2024), which mapped types of the sibling systems.

the distribution of linguistic motivations. When we examine lexical maps of linguistic atlas spanning across continents and covering many language groups, the semasiological point of view is necessary. This paper shows a good example. It is worth reexamining the linguistic variation in Asia from the viewpoint of linguistic motivations.

Acknowledgement

The members of the 2015-2018 research project involved in the investigations of the 'sun' data are the linguists listed below. I express my sincere thanks to them.

SHIRAISHI Hidetoshi (Nivkh), FUKAZAWA Mika (Ainu), KISHIE Shinsuke (Japanese), FUKUI Rei (Korean), UEYA Takashi, YAGI Kenji (Sinitic), TAGUCHI Yoshihisa (Hmong-Mien), ENDO Mitsuaki (Kra-Dai), SHIRAI Satoko, KURABE Keita, IWASA Kazue, SUZUKI Hiroyuki, EBIHARA Shiho (Tibeto-Burman), KONDO Mika (Austroasiatic), UTSUMI Atsuko (Austronesian), MATSUMOTO Ryo (Tungusic and Uralic), SAITÔ Yoshio (Mongolic and Turkic), YOSHIOKA Noboru (South Asia), NAGATO Youichi (Arabic)

References

Alinei, Mario (1983) Introduction. In: M. Alinei et al. (eds.), *Atlas Linguarum Europae*, Volume I: Premier fascicule, Commentaires, XV-XXXIX. Assen: Van Gorcum.

Endo, Mitsuaki, Makoto Minegishi, Satoko Shirai, Hiroyuki Suzuki, and Keita Kurabe (eds) (2021) *Linguistic Atlas of Asia*. Tokyo: Hituzi Syobo.

Fukushima, Chitsuko (2021) 'Sun' in Asia. In: M. Endo et al. (eds.), *Linguistic Atlas of Asia*, 18-21. Tokyo: Hituzi Syobo.

Fukushima, Chitsuko. (2024) Sibling terms in Asia and Africa: A geolinguistic approach to linguistic patterns. *Studies in Geolinguistics* 4: 145-155. https://doi.org/10.5281/zenodo.13948605

Saitô, Yoshio (2021) 'Sun' in Mongolic and Turkic. In: M. Endo et al. (eds.), *Linguistic Atlas of Asia*, 48-49. Tokyo: Hituzi Syobo.

Tuaillon, G. (1983) Soleil: Commentaire. In: M. Alinei et al. (eds.), *Atlas Linguarum Europae*, Volume I: Premier fascicule, Commentaires, 3-8. Assen: Van Gorcum.

Urban, Matthias (2010) 'Sun' = 'Eye of the Day': A Linguistic Pattern of Southeast Asia and Oceania. *Oceanic Linguistics* 49(2): 568-579. https://www.jstor.org/stable/40983980

Homelands, migration, and dispersion of the Kra-Dai

Mitsuaki Endo (Aoyama Gakuin University)

Abstract: Guangxi province was the homeland of the South-Western Tai, who migrated to Southeast Asia around the Mongol period. The Tai origin toponyms reveal the existence of the earlier Kra-Dai population in the central and eastern areas of Guangdong province. Moreover, the earlier Kra-Dai population generally lived in southeastern China.*
Key words: Kra-Dai, South-Western Tai, toponyms, migration, substratum

1. Introduction

This study discusses the geographical distribution of the Kra-Dai, which was wider than today and spread further in the whole Guangdong and Hunan provinces and even in southeastern China as a whole. Specifically, this study explores (1) the homeland of the Southwestern Tai, (2) the existence of the earlier Kra-Dai population in the central and eastern areas of Guangdong province as revealed by toponyms, and (3) the earlier Kra-Dai population in Changsha, Hunan province, also in southeastern China, the realm of the old Yue country.

Figure 1 shows Endo's (2022: 9) subgrouping of the Kra-Dai, while Figure 2 illustrates its hierarchy based on Liang and Zhang (1996: 13). With regard to the subbranches of the Tai branch, this study adopted Li's (1977) established classification.

Endo, Mitsuaki. 2025. Homelands, migration and dispersion of the Kra-Dai. In Nor Hashimah Jalaluddin, Hiroyuki Suzuki and Mitsuaki Endo (eds.) *Proceedings of the sixth International Conference of Asian Geolinguistics*, 18–32. doi: https://doi.org/10.5281/zenodo.17204531

* This work was supported by JSPS KAKENHI Grant Number JP23K25322.

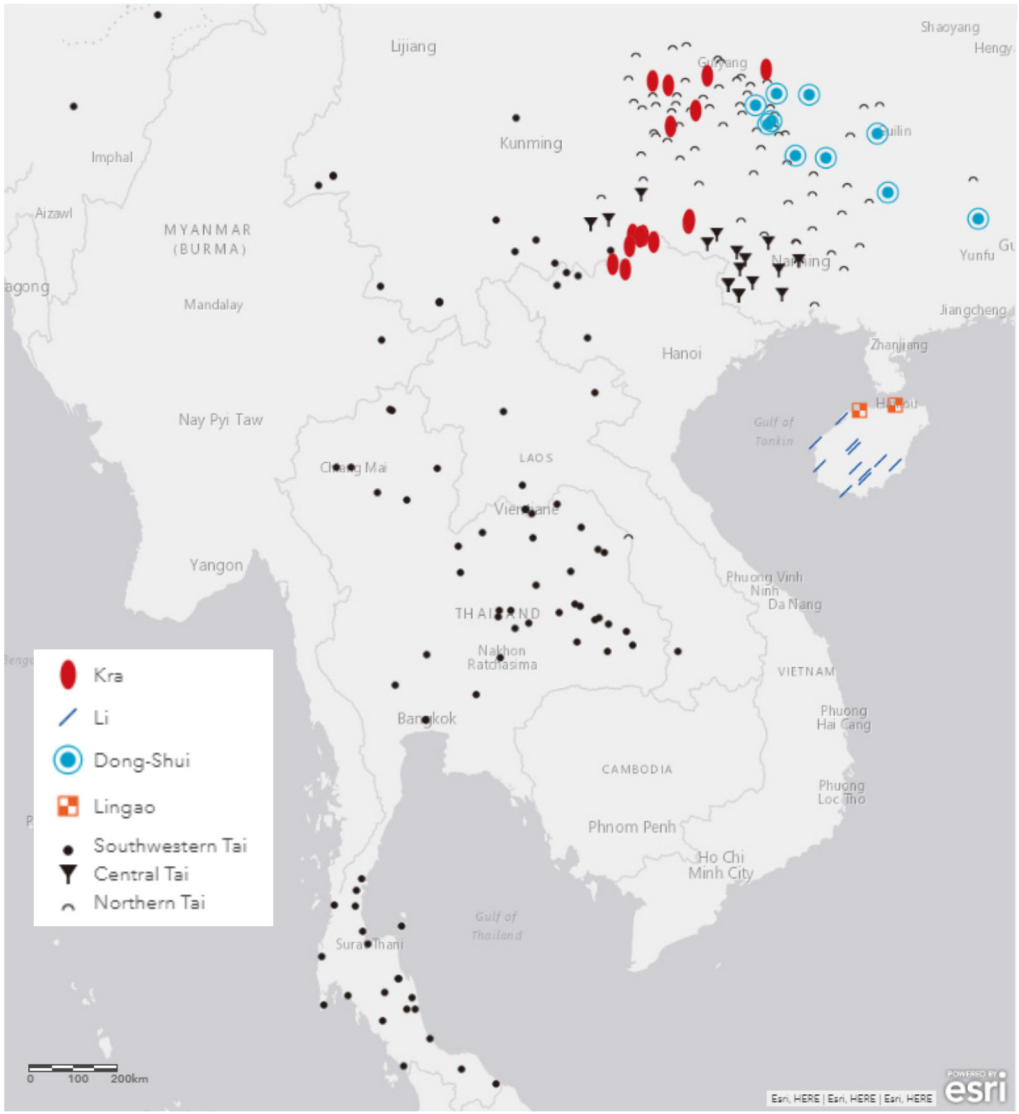


Figure 1: Subgrouping of the Kra-Dai

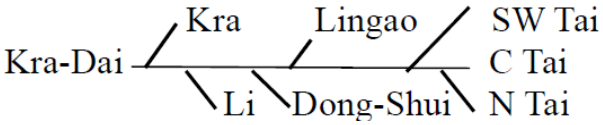


Figure 2: Hierarchy of branches

2. Homeland of the Southwestern Tai

2.1. Migration process proposed by Liang and Zhang (1996)

Southwestern Tai is a subbranch of the Tai branch in mainland Southeast Asia. Figure 3 shows a map of their migration routes according to Liang and Zhang (1996: 35), citing historical descriptions from the Qin Dynasty (246–206 BCE), during which they dwelled around the Ou River near Wenzhou, the present-day Zhejiang. After 756 CE, Zhuang people struggled against the Tang Dynasty and escaped to northwest Vietnam, avoiding the indigenous Vietnamese. From 1052 CE, Zhuang people migrated from Daxin County in western Guangxi toward Laos and northern Thailand, avoiding Dali 大理 County. In 1180 CE, Dai 傣 people migrated to southern Yunnan from the border area of China, Laos, and Thailand. Liang and Zhang (1996: 33–36) created a list of about 100 words that are common between the Southwestern Tai in inland Southeast Asia and southernmost Guangxi, and this vocabulary is different from the Zhuang dialects in central Guangxi.

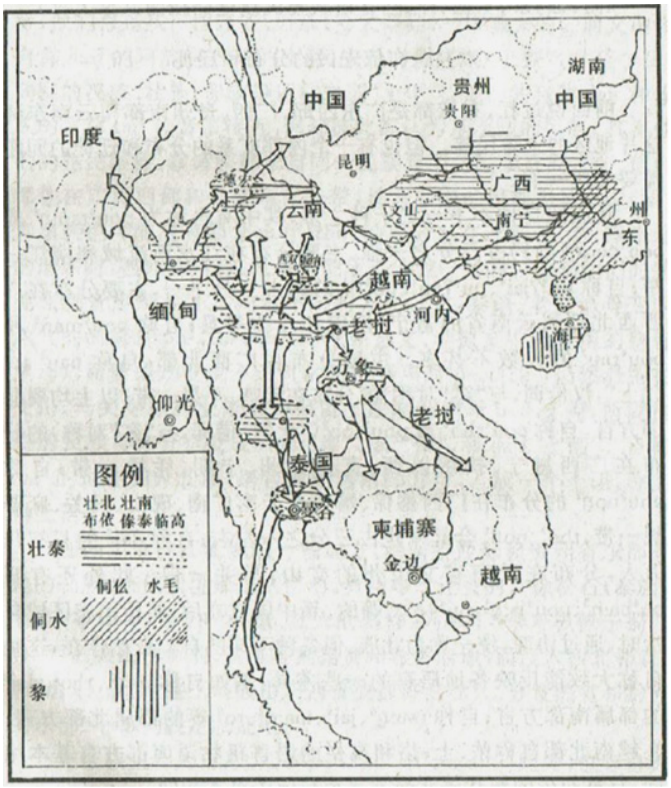


Figure 3: Migration routes of the Dai, Thai, Lao, Shan, and Ahom

2.2. Migration process proposed by Baker and Phongpaichit (2017)

Baker (2002) and Baker and Phongpaichit (2017: 25–42) described the Tai people's migration process from southern China to Southeast Asia as essentially consistent with Liang and Zhang's (1996) map in Figure 3. Figure 4 below was extracted from https://en.wikipedia.org/wiki/Tai_languages#/media/File:Tai_Migration.svg, last accessed 25 June 2025.

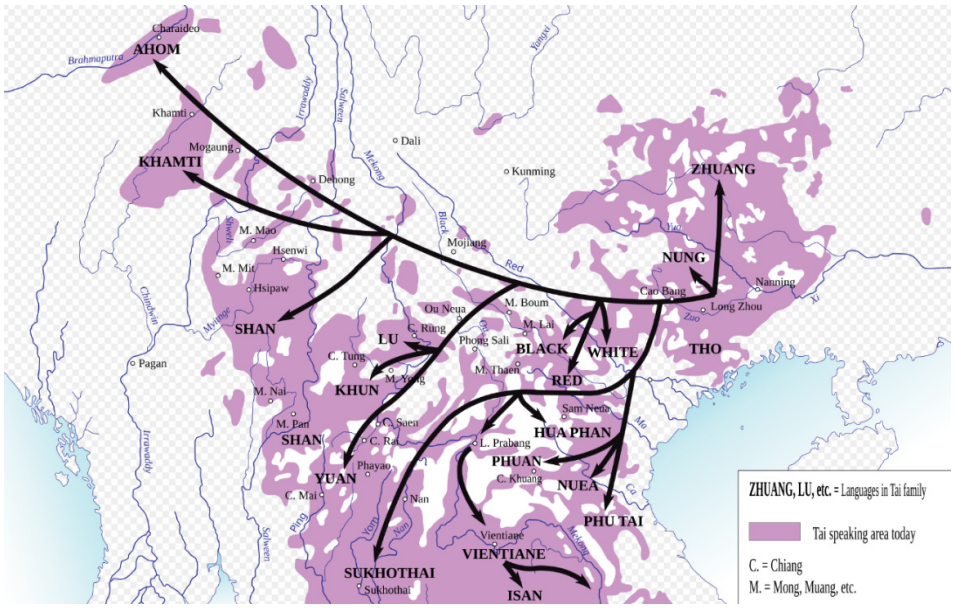


Figure 4: Dispersion route of Tai-speaking people in Southeast Asia

2.3. Some linguistic maps of the Southwestern Tai's migration process

From dozens of Kra-Dai linguistic maps, three words can help trace the migration route of the Southwestern Tai.

The first involves the geographical distribution of the initial consonants of the word for "six" (Figure 5), that is, *khr (*kr) as reconstructed by Yuan Jiahua (1963), and its reflections vary: l, j, r, ɣ, hj, h, ð, hl, khj, and z. Among these, h is distributed in southwestmost Guangxi and the entire Southwestern Tai area except for one location with r (Saek language). This exceptional reflection may be attributed to another migration wave.

The second is the word for "two" (Figure 6), that is, j or 0 (zero), which spread in the Longzhou and Southwestern Tai areas. This shows a scattered reflection of ɲ, which can also be treated as another wave of dispersion from the Guangxi area.

The third is the word for "iron" (Figure 7). In the map, types A1 and A2, including *lek*, go back to the Proto-Tai form *hlek. According to Endo (2017), "Sagart (1999: 200–201) pointed out that this proto Tai form is close to the Old Chinese form *^{a}hlik; hence it was borrowed into Tai as well as other Southeast Asian languages after the manufacturing of iron in this area took place. This occurred not earlier than 700–600 BCE, before the regular change *_hl-_ > _th-_ and *_-ik_ > _-it_ took place prior to the age of Middle Chinese." Types A1 and A2 are distributed throughout the Southwestern Tai area.

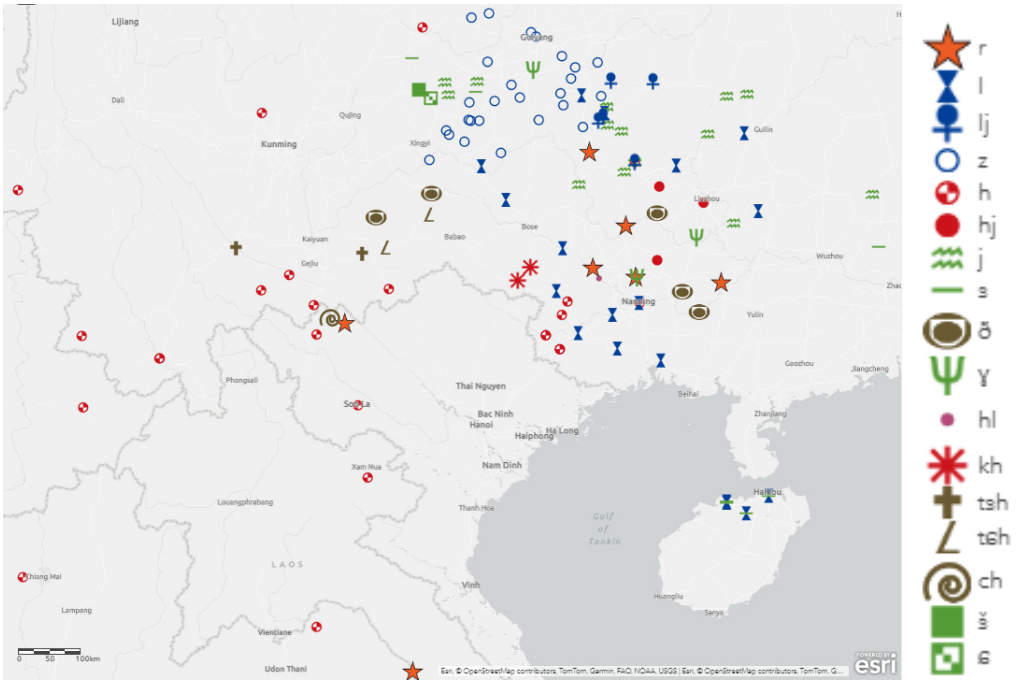


Figure 5: Initial consonants of the word for “six” in Kra-Dai

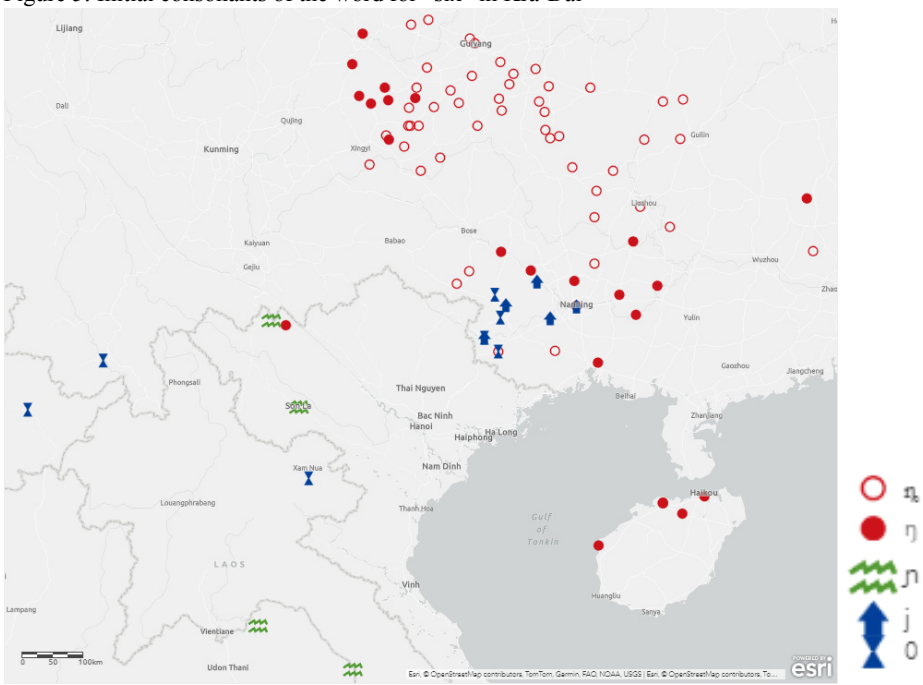


Figure 6: Initial consonants of the word for “two” in Kra-Dai

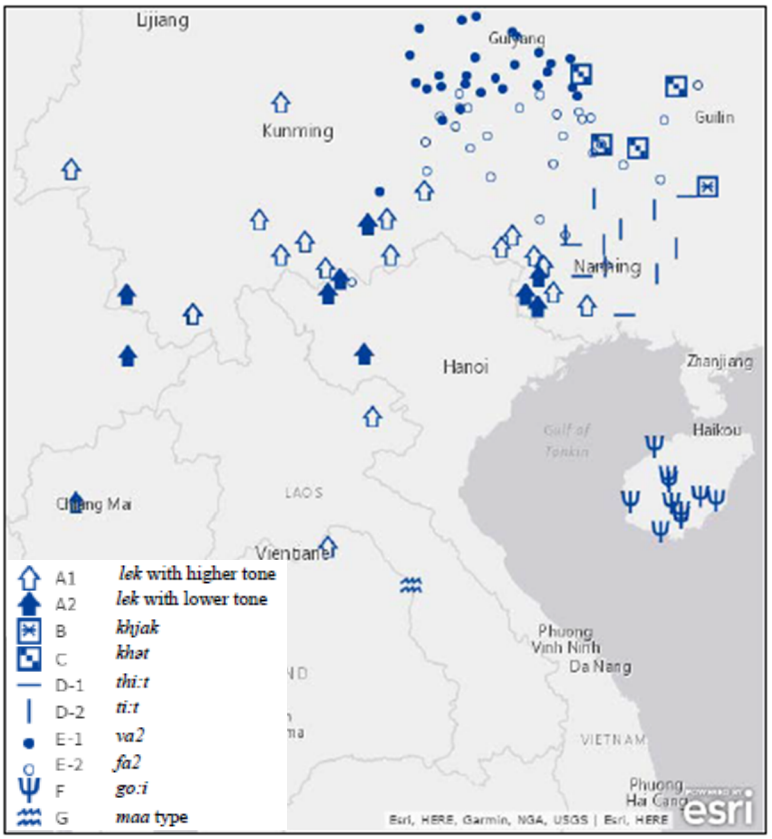


Figure 7: The word for "iron" in Kra-Dai

2.4. Evidence from genetics

Changmai et al. (2023), who analyzed the genome data of the Kra-Dai-speaking population in Thailand, found geographical differences in DNA proportion (Figure 8). The Lao element was generally predominant, implying that the mainstream Thai population came through Laos. In the northeastern part, the Cambodian element is observed in some places, indicating that mixture occurred between Kra-Dai-speaking people and indigenous Austroasiatic-speaking people. In fact, half of the population around the Surin area near the Cambodian border are Austronesian-speaking people. Thailand's northern district shows a considerably high ratio of Zhuang from China and Thai Vietnam, suggesting the homeland and stopover point of these human groups. The Dai element is also recognized in this area. Dai people dwelled in Yunnan, China. The Zhuang element occupies the main part of Hmong Daw in the Golden Triangle, which

means they experienced a language shift from Kra-Dai to Hmong. Thus, genetics provides not only evidence of population movements but also migration dating.

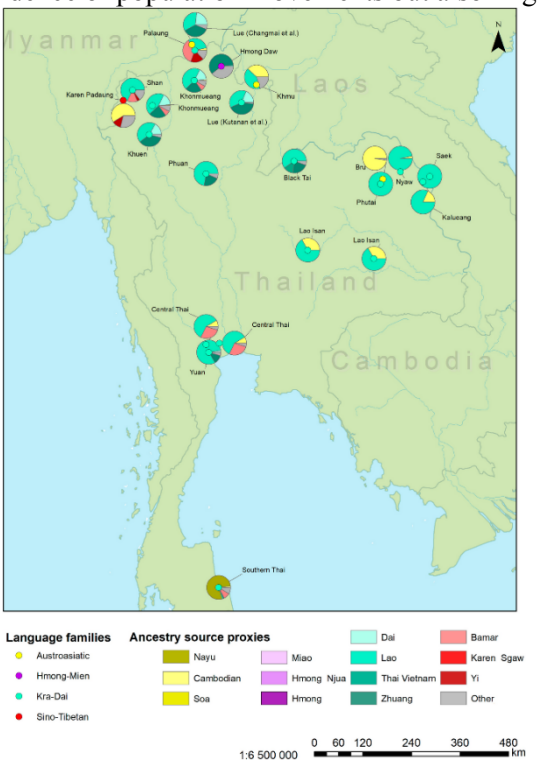


Figure 8: Sources of recent ancestry in Thai groups

3. Existence of the earlier Kra-Dai population in Guangdong province

Figure 1 shows how the Kra-Dai is exceptionally distributed in the Guangdong province today. However, because Kra-Dai origin toponyms are generally found in this area, Kra-Dai people can be said to have lived there in ancient times.

Xu (1939) enumerated toponyms with Kra-Dai elements in Guangdong province as shown in Figures 9–16, where the maps present different distribution patterns. Some toponyms meant "village" (Figures 11, 12, and 15) because of dialectal differences. Some of them are generally distributed in Hainan Island, covering the area where the Hlai people lived (Figures 9 and 16) or partially encompassing the area of the Lingao people (Figures 12 and 14).

Ostapirat (1998) reported a "Chinese" dialect in the Leizhou Peninsula in Guangdong, in which a small amount of the basic vocabulary is of Kra-Dai origin; otherwise, the whole linguistic system resembles a Chinese dialect and is a vestige of a mainland Lingao language, which is essentially the same as the nature of Cantonese dialects.

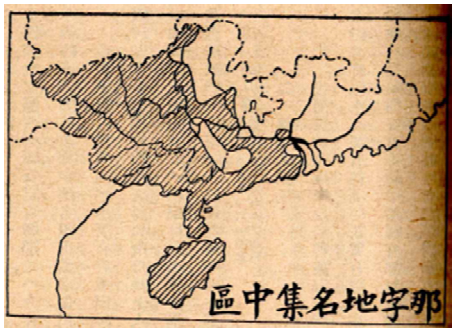


Figure 9: Toponyms with *na* (rice field)

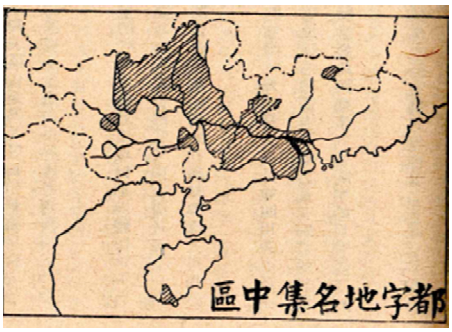


Figure 10: Toponyms with *du* (10 villages)

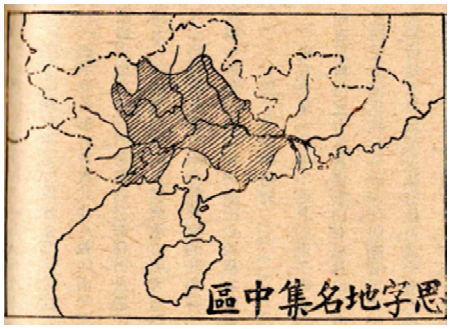


Figure 11: Toponyms with *si* (village)

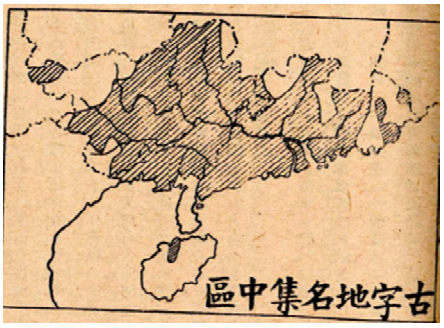


Figure 12: Toponyms with *gu* (village)

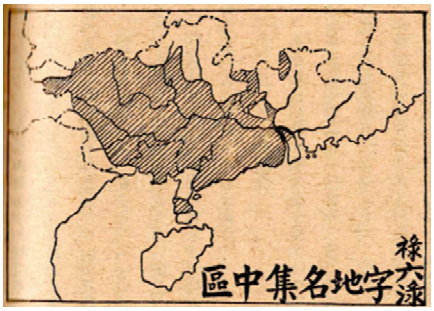


Figure 13: Toponyms wit *lu* (mountain)

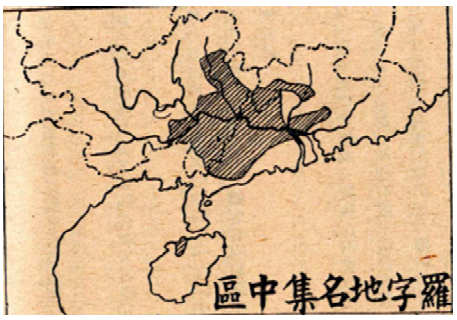


Figure 14: Toponyms with *luo* (various meanings)

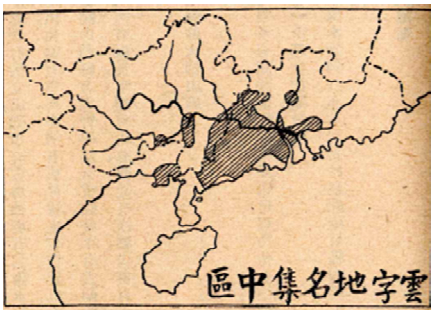


Figure 15: Toponyms with *yun* (*baan*, village)

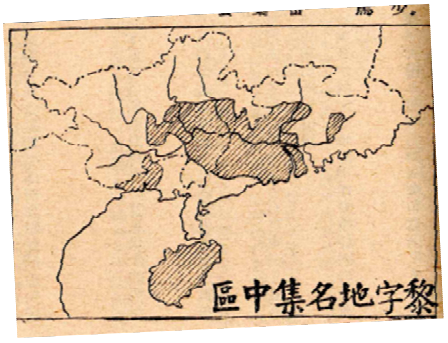


Figure 16: Toponyms with *li*

4. The Kra-Dai population in southeastern China as a whole

4.1. The earlier Kra-Dai population in Chu state during the Warring States period

Lin et al. (2004) examined □, meaning "one," in bamboo slip documents in the Chu state around the Hunan and Hubei areas dating back to the Warring States period (BC 5–3) and further compared it with [nɯŋ] "one" in modern Tai. They considered that some Tai-speaking people adopted the Tai sound 能, meaning "one," as a phonetic element.

4.2. Kra-Dai toponyms in southeastern China

Zhou and You (2019) expanded the scope of the Kra-Dai origin toponyms to the entire southeastern China. Figure 17 illustrates the geographical distribution of the Kra-Dai toponyms *lai* "stream" (◯), *luo* "mountain" (▲), and *tan* (no meaning provided). They are vestiges of the Kra-Dai in this area. Li (2001) reported that a dialect in the Daic substratum was used near Shanghai.

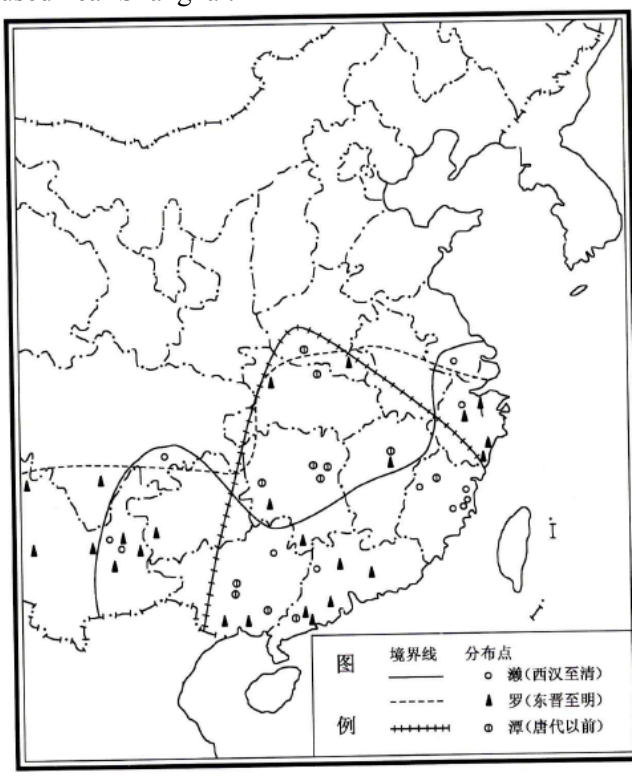


图 5-5 "濑、罗、潭"字地名分布示意图

Figure 17: Kra-Dai toponyms *lai*, *luo*, and *tan* in southern China (Zhou and You 2019)

5. Theories on the Kra-Dai migration process

5.1. Chamberlain (2016)

Chamberlain (2016: 67-70) provided a general picture of the Kra-Dai migration process as follows:

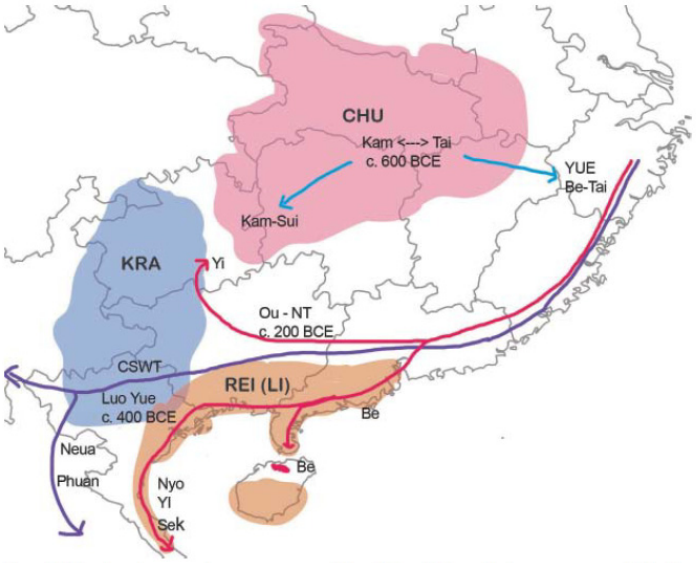


Figure 7. Map showing approximate movements of Kam-Tai and Tai peoples into spaces occupied by Kra and Rei. The dark purple line represents Luo-Yue (CSWT) and the red line Ou-Yue (NT).

Figure 18: The Kra-Dai migration process (Chamberlain 2016)

Phase 1 – 1100-800 BCE Shang ends, Zhou begins, Chŭ is born.

Kra separates from the Kra-Dai mainstream.

Phase 2 – Early Spring and Autumn Period, 771-685 BCE Political upheavals in Zhou causing capital to relocate Li splits off and moves south/southeast.

Phase 3 – Chŭ Hegemony – Late Spring and Autumn, 613-453 BCE King Zhuang expands Chŭ which becomes the most powerful state Yue (Be-Tai) separates from Kam-Sui moves east to the coast, conquers Wu.

Phase 4 – Warring States Period, 475-221 BCE

Chŭ annexes Yue, 333 BCE; Qin conquers Chŭ 223 BCE.

Yue royal families begin to move south, forming the Bai Yue.

Luo Yue (Central Southwestern Tai) overruns the lands of the Rei and the Kra in southern Lingnan and Annam.

Xi Ou (Northern Tai) follows and comes to dominate northern Lingnan, including some areas formerly held by Luo Yue in Jiuzhen.

Phase 5 – Qin Dynasty 223-206 BCE; Han Dynasty, 206 BCE–220 CE

Qin-Han begins colonization of the south, establishing commanderies at Canton, Jiaozhi, and Jiuzhen. Recorded history of the south begins.

Mobile Yue Central Southwestern Tai polities continue to establish chiefdoms, dominating the original Kra and populations in western Annam, and Rei in Jiuzhen.

Ou Yue and Yi Northern Tais push west from Nan Yue.

Be and Sek separate and move west and south from Nan Yue, respectively.

5.2. Blench (2018)

Blench (2018) discussed the genetic relation between Daic and Austronesian based on basic vocabulary and provided a scheme (Figure 19).

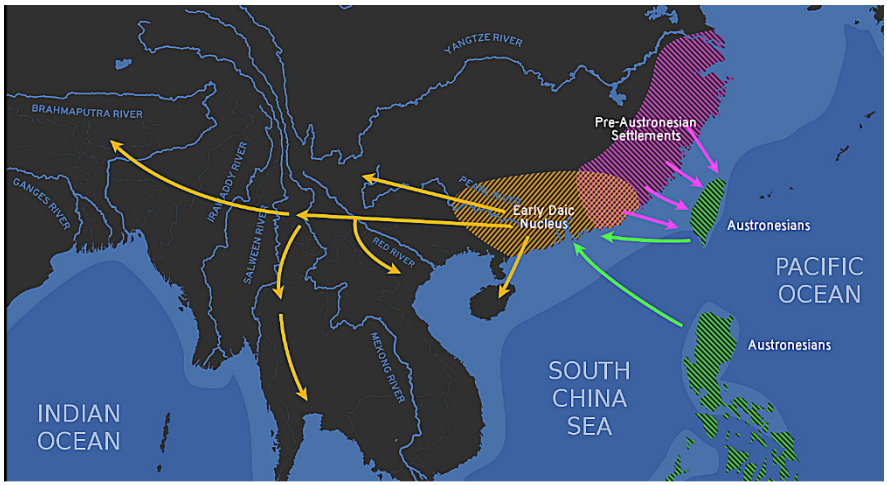


Figure 19: Genesis of Daic languages and their association with Austronesian (Blench 2018; https://en.wikipedia.org/wiki/Kra%E2%80%93Dai_languages#cite_note-Blench2018-42)

5.3. Tao et al. (2023)

Tao et al. (2023) inferred migration routes and chronology by applying Bayesian phylogenetic method to 600 vocabulary items of 100 Kra-Dai languages. However, the language dispersion process and migration routes do not necessarily always coincide. In addition, the chronology in Figure 20 shows mean dating, while the minimal datings in Figure 21 seem more plausible.

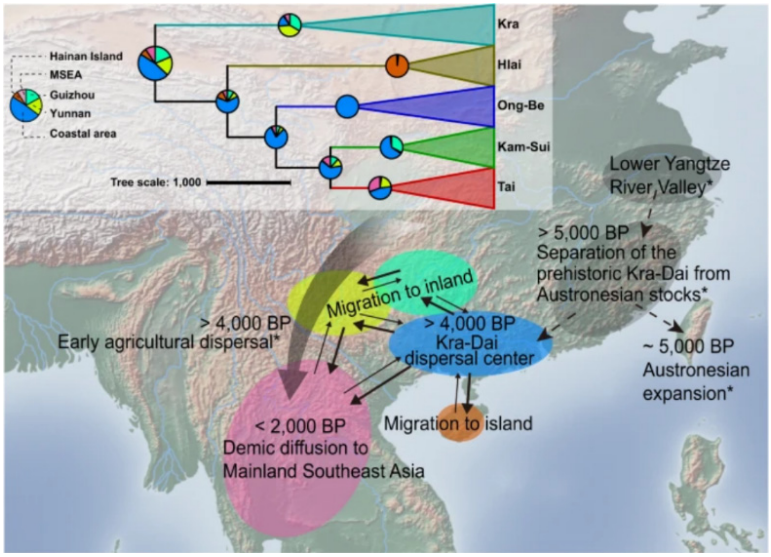


Figure 20: Inferred dispersal routes of Kra-Dai speakers and their languages in prehistory (Tao et al. 2023)

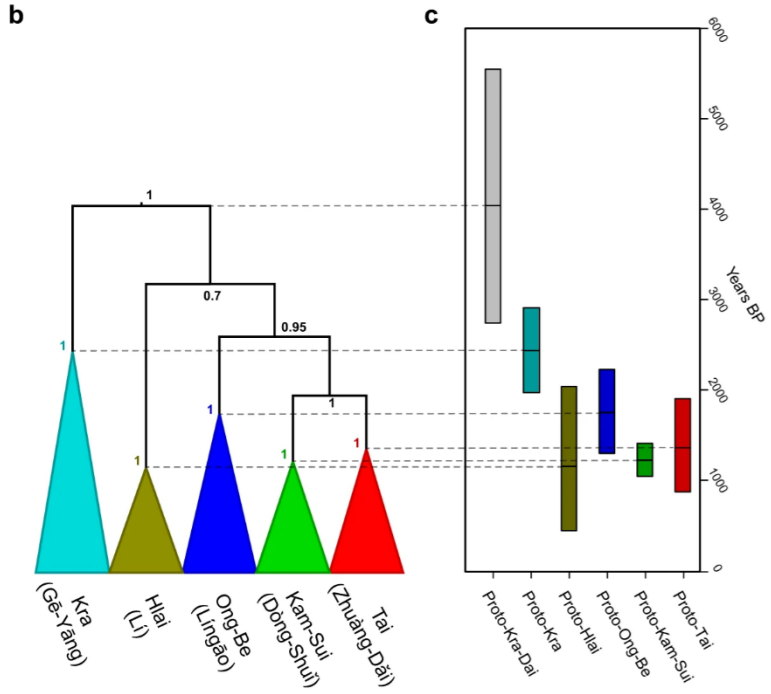


Figure 21: Maximum clade credibility tree with divergence time of the Kra-Dai languages (Tao et al. 2023)

6. Conclusion

In sum, Kra-Dai, as a branch of Austronesian, once spread in almost all areas in southeastern China. Following Sagart's Sino–Tibetan–Austronesian theory, it has further northwestern origins from the Gansu–Qinghai–Shaanxi area, migrating through Henan, Shandong, and Jiangsu and then to Fujian province. Endo (forthcoming) provides an in-depth discussion of the maritime migration routes of the Hlai and Kra branches of Kra-Dai. As observed in section 2.4 above, language shifts also occurred rather frequently. Li et al. (2010) showed that Mien (Hmong-Mien) is clustered with Sui, Hlai, Bouyei, and Thai (all are Kra-Dai) in terms of Y chromosomes, reflecting paternal origins, which suggests that Mien people are of mainly Kra-Dai origin and accepted Hmong through language contact. Such a language shift is rather attributed to language dispersion and not migration.

Even today, southern China is home to Austronesian, Austroasiatic, Hmong-Mien, and Sino-Tibetan people besides the Kra-Dai. After the Qin-Han period from 221 BC to 220 AD, the Chinese-speaking area significantly expanded, with human groups migrating, mixing, and changing their languages before and after. Their macro and micro histories can be successfully explained because of the collaboration between linguistics, archaeology, genetics, and other related disciplines.

References

Baker, Chris (2002) From Yue to Tai. *Journal of Siam Society* 90(1/2): 1–26. https://thesiamsociety.org/wp-content/uploads/2002/03/JSS_090_0b_Baker_YueToThai.pdf

Baker, Chris and Pasuk Phongpaichit (2017) *A History of Ayutthaya*. Cambridge: Cambridge University Press.

Blench, Roger (2018) Tai-Kadai and Austronesian are Related at Multiple Levels and their Archaeological Interpretation (draft). https://en.wikipedia.org/wiki/Kra%E2%80%93Dai_languages

Chamberlain, James R. (2016) Kra-Dai and proto history of South China and Vietnam, *Journal of Siam Society* 104: 27-77. https://thesiamsociety.org/wp-content/uploads/2016/04/JSS_104_0c_Chamberlain_KraDaiAndProtoHistoryofSouthChinaAndVietnam.pdf

Changmai, P., Phongbunchoo, Y., Kočí, J. et al. (2023) Reanalyzing the genetic history of Kra-Dai speakers from Thailand and new insights into their genetic interactions beyond Mainland Southeast Asia. *Scientific Reports* 13, 8371. https://doi.org/10.1038/s41598-023-35507-8

Endo, Mitsuaki (2016) Geographical distribution of the /r/ type sounds in Zhuang. *Papers from the Third International Conference on Asian Geolinguistics*. 46–71. URI: https://publication.aa-ken.jp/papers_3IC_Asian_geolinguistics_2016.pdf

Endo, Mitsuaki (2017) Iron: Tai-Kadai, *Studies in Asian Geolinguistics* 5: 15–16. https://publication.aa-ken.jp/sag5_iron_2017.pdf

Endo, Mitsuaki (2022) Subgrouping of Kra-Dai. In Hiroyuki Suzuki, Mika Fukazawa, Akiko Yokoyama, and Mitsuaki Endo (eds.) *Linguistic Atlas of Asia and Africa I*, 9. Tokyo: Geolinguistic Society of Japan. https://doi.org/10.5281/zenodo.7118188

Endo, Mitsuaki (forthcoming) Words for the numbers one to ten reflect the migration patterns of the Hlai and Kra people from Taiwan.

Li, Dongna et al. (2010) Genetic origin of Kadai-speaking Gelong people on Hainan island viewed from Y chromosomes. *Journal of Human Genetics* 55: 462–468. https://doi.org/10.1038/jhg.2010.50

Li, Fang Kuei (1977) *A handbook of comparative Tai*. Honolulu: University of Hawai'i Press.

Li, Hui [李辉] (2001) Shanghai Maqiao hua de Taiyu diceng cihui 上海马桥话的台语底层词汇 [Daic Background Vocabulary in Shanghai Maqiao Dialect]. *Qiong-Tai shaoshu minzu xueshu wenhua jiaoliu yantaohui lunwenji*《琼台少数民族学术文化交流研讨会论文集》[Proceedings for Conference of Minority Cultures in Hainan and Taiwan]. https://web.archive.org/web/20180327144856/http://loca.fudan.edu.cn/lh/Doc/D02.pdf

Liang, Min [梁敏] and Junru Zhang [张均如] (1996) *Dongtai Yuzu Gailun*《侗台语族概论》[*Synopsis of Kra-Dai*]. Beijing : Zhongguo Shehui Kexue Chubanshe.

Lin, Hongying [林虹瑛], Murase Nozomi [村瀬望], Furuya Akihiro [古屋昭弘] (2004) Sengoku moji 罷 ni tsuite [On the character 罷 during Warring States period]. *Kaipian*『開篇』23: 71–75.

Pittayawat, Pittayaporn (2009) *The phonology of proto-Tai*. Ph. D. Dissertation, Cornell University. https://ecommons.cornell.edu/server/api/core/bitstreams/6af02aa7-c444-481c-8d1b-ac0c25346f20/content

Ostapirat, Weera (1998) A mainland Be language?. *Journal of Chinese Linguistics* 26(2): 338–344.

Sagart, Laurent (1999) The roots of Old Chinese. Amsterdam : John Benjamins. https://starlingdb.org/Texts/Students/Sagart%2C%20Laurent/The%20Roots%20of%20Old%20Chinese%20%281999%29.pdf

Tao, Y., Wei, Y., Ge, J. et al. (2023) Phylogenetic evidence reveals early Kra-Dai divergence and dispersal in the late Holocene. *Nature Communications* 14, 6924. https://www.nature.com/articles/s41467-023-42761-x

Yuan, Jiahua [袁家骅] (1963) Zhuangyu / r / de Fangyin Duiying 壮语/ r /的方音对应 [Sound correspondences of / r / among the Zhuang dialects]. *Yuyanxue Luncong*《语言学论丛》5: 187–218.

Xu, Songshi [徐松石] (1939) *Yuejiang liuyu renmin shi*《粤江流域人民史》[History of people in Yuejiang river basin].

Zhou, Zhenhe [周振鹤] and You Rujie [游汝杰] (2019) *Fangyan yu Zhongguo wenhua*《方言与中国文化》[Dialect and Chinese culture], new edition. Shanghai: Shanghai Renmin Chubanshe.

A geographic distribution of the word form for 'frog/toad' among Mon-Khmer languages in Vietnam and its implication

TRẦN Thị Hồng Hạnh (USSH, VNU Hanoi)
TRƯƠNG Nhật Vinh (USSH, VNU Hanoi)

Abstract: This paper presents an etymological investigation into the history of the terms for 'frog' and 'toad' in Mon-Khmer languages spoken in Vietnam. The primary goals are to identify native etyma and discuss their sociocultural implications. Through cartographic representation of their geographical distribution and supplementary evidence from historical linguistics, archaeology, and genetics, the study confirms their Proto-Mon-Khmer origin and classifies them as part of the inherited lexical core. Notably, phonetic correspondences involving these terms are observed between Mon-Khmer and Tai-Kadai languages, suggesting a complex linguistic history. The data raises the question of whether these similarities point to historical connections and language contact phenomena that warrant further investigation, or whether they might instead result from independent lexical innovation, such as imitative onomatopoeia, as documented for certain other species in Asia.

Key words: Mon-Khmer, etymology, frog, toad, geolinguistics, archaeology, genetics

1. Introduction

When examining the history of languages in Southeast Asia, it's widely agreed that this is a region marked by "tremendous complexity and uncertainty" (Alves 2015, pg. 50). Therefore, matters of language phyla classification and timing, as well as interphyla borrowing within the region, remain a significant subject requiring extensive research to fill existing gaps. It is within this complex linguistic landscape that this study focuses on the etymology of 'frog' and 'toad' in Mon-Khmer languages in Vietnam. Animal etymology isn't just about tracing the origins of words. It's a powerful tool in historical linguistics, helping us reconstruct a comprehensive picture of the environment, culture, migration, and interactions among prehistoric communities, all through the lens of language. Therefore, for researchers interested in the prehistory of Southeast Asian

TRẦN, Thị Hồng Hạnh and TRƯƠNG, Nhật Vinh. 2025. A geographic distribution of the word form for 'frog/toad' among Mon-Khmer languages in Vietnam and its implication. In Nor Hashimah Jalaluddin, Hiroyuki Suzuki and Mitsuaki Endo (eds.) *Proceedings of the sixth International Conference of Asian Geolinguistics*, 33–49. doi: https://doi.org/10.5281/zenodo.17204540

languages in general, and Austroasiatic languages in particular, this area of inquiry is no exception to their interests.

These two species 'frog' and 'toad' can be classified as wild animals. In the online World Loanword Database (WOLD), borrowing rates for wild animals tend to be lower than those for domesticated animals, and are only higher than the borrowing rates for insects. The aim of this study is to approach these terms from a geolinguistic perspective, utilizing cartographic representation to illustrate their geographical distribution. Furthermore, evidence from historical linguistics, archaeology, and genetics will be employed to provide etymological notes on these terms. This research seeks to determine whether the etyma for 'frog' and 'toad' in Mon-Khmer languages truly represent inherited core vocabulary, and if they bear any relation to sociocultural historical events.

2. Literature review and theoretical framework

There have been studies on animal and plant terms in various languages from diverse perspectives. Cecil H. Brown's (1984) "*Language and Living Things: Uniformities in Folk Classification and Naming*" stands as an important work in ethnolinguistics and cognitive anthropology. Brown presents a strong argument for universal tendencies in how different cultures classify and name living things. He specifically focused on how languages worldwide name plants and animals, revealing general patterns in classification levels and the order of lexicalization (e.g., basic "life-form" terms like "tree" and "bird" are often named before more specific "generic" levels like "oak" and "sparrow"). His primary goal was to uncover the cognitive and structural universals governing human categorization of the natural world, emphasizing folk taxonomic hierarchy, psychological salience, and perceptual similarity. While not directly addressing 'frog' and 'toad' in Mon-Khmer languages, Brown's principles offer a robust analytical framework for this study. Brent Berlin's "*Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*" (1992) further solidifies this understanding. Building on and refining earlier works (including his own and that of Cecil H. Brown), the book systematically explores the universal principles underlying how traditional societies classify and name plants and animals. The central thesis of Berlin's work is that folk biological classification systems are not arbitrary cultural constructs, but rather reflect universal cognitive principles that are largely consistent across diverse cultures and languages. He posits that humans categorize the natural world in predictable ways due to shared cognitive capacities and interaction with the environment.

Challenging some aspects of these universal tendencies, Robert A. Blust's paper, "*The History of Faunal Terms in Austronesian Languages*" (2022) offers highly relevant methodological insights for reconstructing faunal terms and exploring their cultural implications. In this comprehensive study, Blust examined 277 etyma to provide an overview of reconstructed faunal terms primarily at the Proto-Austronesian, Proto-Malayo-Polynesian, and Proto-Western Malayo-Polynesian levels. An especially noteworthy insight for this study is the observation that the Proto-Malayo-Polynesian word for 'bird' (*manu-manuk) was almost certainly derived from the word for 'chicken' (*manuk) through reduplication. This finding challenges Brown's (1984) proposed life-form encoding sequence and opens new avenues for understanding semantic development in faunal terms.

Regarding animal and plant terms within Austroasiatic languages, G. van Driem's (2011), "*The ethnolinguistic identity of the domesticators of Asian rice*" serves as a key example of linguistic palaeontology. This work is highly pertinent to our study, as it shares a common concern for the prehistory of Southeast Asian languages, particularly the Austroasiatic family. G. van Driem employs a robust multidisciplinary approach, integrating evidence from historical linguistics, archaeology, and genetics. Through the analysis of vocabulary related to rice cultivation across various language families, he seeks to identify the ethno-linguistic groups most likely responsible for the initial domestication and dispersal of Asian rice. His findings strongly implicate Proto-Austroasiatic-speaking communities as key players in this agricultural revolution, offering a valuable model for tracing core vocabulary and discerning deep historical connections.

Complementing these linguistic perspectives, "*Phylogeographic history of plants and animals coexisting with humans in Asia*" edited by Osada et al. (2024) delves into phylogeography, the study of the geographical distribution of genetic lineages, and their historical evolution. While including the natural history of the Japanese Archipelago, its broader scope covers Asian biodiversity and human-environment interactions. Employing a similar multidisciplinary approach, this work integrates genomics, archaeological, and environmental evidence. Though not focused on Southeast Asian languages, its comprehensive data and methods directly complement our etymological analysis, especially for connecting vocabulary origins to broader cultural and environmental historical events in the region.

Some aforementioned review highlights that the study of animal terms in languages has garnered significant interdisciplinary interest and yielded important advancements. Despite these valuable contributions, a notable gap remains concerning

in-depth etymological data on specific animal terms specifically within the diverse Austroasiatic languages of Southeast Asia. In this context, Mark Alves stands out for his important and sustained contributions. His main interests lie in the languages of mainland Southeast Asia and southern China, especially Vietnamese and Vietic, Austroasiatic, Chinese languages, and neighboring languages, along with their historical linguistic issues. Consequently, he has conducted a series of studies on loanwords and etymology in these languages using an interdisciplinary approach. While Alves's contributions have significantly enhanced our understanding of lexical diffusion and ethnolinguistic history in mainland Southeast Asia, particularly concerning domesticated animals and material culture (see Alves 2009, 2015, 2022, 2023a, 2023b, 2025; Higham, C.F.W. & Alves, M.J 2025) a notable gap remains regarding in-depth etymological data for specific wild animal terms, such as 'frog' and 'toad', within the diverse Austroasiatic languages of Southeast Asia. This study aims to address this gap, specifically focusing on the etymological investigation of 'frog' and 'toad' within this language family.

This study adopts a geolinguistic framework supported by interdisciplinary data drawn from historical linguistics, archaeology, genetics, and ethnohistory. This approach is for three reasons. First, the spatial mapping of lexical variants facilitates the identification of lexical origins, diffusion pathways, and zones of linguistic convergence across mainland Southeast Asia. Second, lexical items denoting wild animals—such as 'frog' and 'toad'—are closely embedded within local ecologies and cultural taxonomies. A geolinguistic perspective thus provides critical insights into how environmental and sociocultural factors have influenced both the formation and spread of these terms. Third, geolinguistics offers an integrative platform that synthesizes linguistic data with findings from allied disciplines, thereby enabling a more comprehensive understanding of historical change of language and population in the region.

For the purpose of this study, the classification of language families isn't a primary research concern. Consequently, this paper operates on the understanding that the Austroasiatic language family is one of five major families in Southeast Asia, with Austronesian, Tai-Kadai, Hmong-Mien, and Sino-Tibetan being the others—a classification broadly accepted by scholars. According to Trần Trí Dõi (2022), some linguists designate the entire Austroasiatic family as Mon-Khmer. However, based on the aforementioned classification, Mon-Khmer is considered merely a branch within the Austroasiatic language family, further subdivided into nine groups comprising approximately 103 languages (Trần Trí Dõi 2022, p. 74). In Vietnam, this branch

includes five language groups: Khmuic, Vietic, Katuic, Bahnaric, and Khmeric. This is a crucial branch not only due to its substantial number of member languages and the even distribution of its speakers throughout mainland Southeast Asia, but also because it exhibits the most representative historical linguistic changes of the entire Austroasiatic family. Despite the understanding of Mon-Khmer as a branch, the term 'Mon-Khmer' used in the title and throughout the body of this paper stems from the fact that much of our corpus is drawn from other sources, notably Shorto's *Mon-Khmer Etymological Dictionary*. Therefore, this choice is intended to ensure consistency with those foundational resources.

3. Data and maps

The lexical data for this study is drawn from several databases. Our primary dataset consists of fieldwork-collected lexical data from the following locations: Hà Nội, Bắc Giang, Sơn La, Phú Thọ, Hòa Bình, Thanh Hóa, and Nghệ An. Additionally, this study incorporates lexical data compiled from various dictionaries and proto-language reconstructions published over the past several years. These sources include, but are not limited to, the *Katuic Comparative Dictionary* (Peiros 1996), *SEAlang Mon-Khmer Etymological Dictionary*, *Draft Manuscript on Chinese Dialects* (Endo 2001), *Từ điển Mường - Việt* (Nguyễn Văn Khang et al. 2002), *Austroasiatic Dataset for Phylogenetic Analysis* (Paul Sidwell 2015), various works by Alves on Mon-Khmer (see References), Pittayaporn (2009) for Proto-Tai, and the PanGloss collection, among other resources.

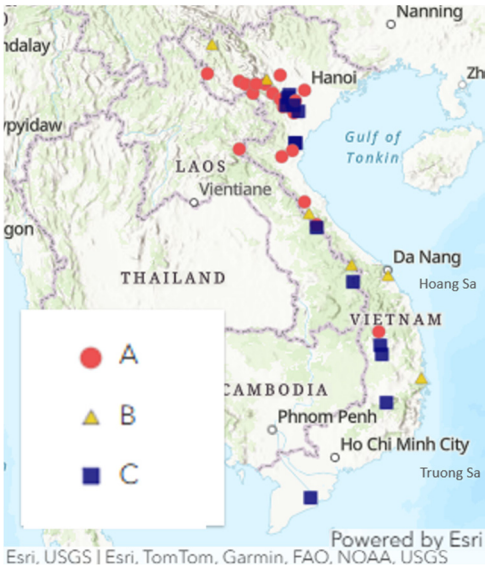
In Vietnamese, frog is *ếch* and toad is *cóc*. Both are very common amphibians in nature, but they have differences. Frogs typically have smooth, moist skin and long legs for jumping, and they live mostly near water. Toads, on the other hand, usually have dry, bumpy skin and shorter legs, making them more adapted to drier environments. Observations during the processing of materials reveal a notable point: while some certain sources differentiate between these two amphibian species, others lack a clear distinction. For instance, in Chứt, a language belonging to the Vietic group, the phonetic form kuàk refers to both 'frog' and 'toad'. Consequently, a rigorous differentiation between the forms for *ếch* (frog) and *cóc* (toad) is not being strictly maintained during the initial phase of data collection.

This data collection rule has allowed for the identification and documentation of diverse forms for "frog" and "toad" across Mon-Khmer languages in Vietnam, as presented in the following table. Specifically, these forms can be categorized into three main types, each with further subtypes.

Table 1: Types of Mon Khmer terms for 'frog' and 'toad'

Type	**Subtype**	**Pre-syllable**	**Main Syllable**			**Tone**
			Consonant	**Vowel**	**Final consonant**	
A Monosyllable CVC	A1		/ʔ/	/ɛ/ /eː/ /u//uː/ /o/ /ɔ/ /ɑ/, /ɒ/	/k/ /t/	*sắc* (rising tone), *nặng* (low dropping tone)
	A2		/k/, /kʰ/			
	A3		/ɣ/			
	A4		/r/, /ʐ/			
	A5		/tʰ/			
B Sesquisyllable CvCVC	B	/a/ /u/ /k/	/t/ /r/, /ʐ/			
C Disyllabic/ Reduplicative form	C	Reduplicative form	/k/ /r/	/o/ /ɔ/ /ɑ/, /ɒ/ /e/	/k/ /t/ **/∅/**	

Table 1 indeed reveals a significant phonological structural consistency across the 'frog' and 'toad' terms in Mon-Khmer languages, despite their classification into monosyllabic, sesquisyllabic, and disyllabic forms. Regarding the geographical distribution of these variants, Map 1 provides a visual representation.



Map 1: Distribution map of "toad" in Mon-Khmer languages in Vietnam.

Type A's distribution largely coincides with the primary habitation areas of Vietic language speakers. Type B is predominantly found in regions inhabited by conservative subgroups of Vietic, Khmuic, and Katuic languages. Type C, conversely, is more widely distributed across regions inhabited by Vietic, Bahnaric, Khmeric, and Katuic language communities. Despite these observed distributional differences among Type A, B, and C variants, their overall spatial arrangement exhibits considerable interspersion. This pattern may be indicative of historical internal evolution of subgroups within the Mon-Khmer branch under different contexts.

However, our fieldwork also revealed phonological forms similar to Type A variants in several Tai languages: Tày language in Vietnam and Zhuang language in Guangxi, China.

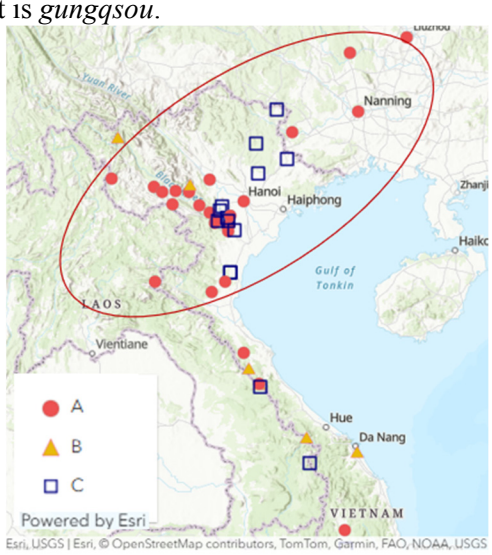
Table 2: 'Toad' in some Tai languages in Vietnam and Southern China

Word for 'toad'	Language/Dialect
/kɔk/	Zhuang (Liujiang, Liuzhou)
/kok/	Zhuang (Du'an, Hezhi)
/kok/	Zhuang (Longzhou, Chongzuo)
/kok/	Zhuang (Nanning)
cáy gộc /kaj ɣɔk/	Tày Bắc Giang
cáy rộc /kaj rɔk/	Tày Trùng Khánh
cấy cộc /kə̆i kok/	Tày Lạng Sơn
ca cộc /ka kok/	Nùng Bắc Cạn

This is quite compatible with the word 'frog' from the *Zhuang-Chinese-English Dictionary*, a resource primarily based on Nong Zhuang dialect from Zhetu District in Guangnan County, Yunnan Province. In this dictionary, the entry for 'frog' is annotated as:

frog: [n] *goap* (gaep gaep); [n] *gvej2*; [n] *gaeuh* 2
frog, large edible [n]: *goep*
frog, large male; bullfrog [n]: *goepcwz*, *goepndai*
frog, male of a small variety of edible [n]: *gvejndai*
frog, mountain [n]: *goep-ndoeng*

However, the dictionary's entry for 'toad' displays a phonological form that bears little clear connection, that is *gungqsou*.



Map 2: Distribution map of "toad" in Mon-Khmer and Taic languages in Vietnam and Southern China

It is worth noting that, in a previous study, Endo (2001) Endo mentioned the word for 'toad' in several dialects in southern China and stated that they are related to Tai languages, including Zhuang. In discussing the word '蛤' (frog/toad) in several Chinese dialects, he noted the word for toads *kop in southern dialects of Chinese, Yue (Cantonese), Min (Fukkienese), and Xiang (Hunanese). It also appeared in the Fangyan dating back to the Han dynasty with locality of the Guangxi Zhuang Autonomous Region now. According to Hashimoto (1976, as cited in Endo 2001, p. 146), it came from Tai languages. In Guangzhou dialect, '蛤' is pronounced [kɐp 下陰入/xià yīn rù/a low entering tone] and refers to a type of frog or toad. According to Hashimoto (1976, as cited in Endo 2001, p. 146), the word "frog" in Cantonese is found in Guangzhou [kap], Taishan [ka:p], Yangchun [kap], and Huaxian [kuop] (all with yin), which correspond to the Tai languages Boai [kop], Zhuang [kop], and Siamese [kop] (all with 7 tones), and is a very clear layer of Thai language. Although the Chinese Dialect Vocabulary lists words like '蜂蜂' and '田鶏' as equivalents for 'frog' in Cantonese, other dictionaries do indeed include terms containing '蛤'. Based on the characters, it would appear that Northern dialects also have '蛤蟆' (háma) for frog/toad, but phonologically, the Beijing Mandarin háma is better understood as deriving from '蝦' (pronounced with the "hu + jia" reading in Guangyun), rather than from '蛤' (with the 'gu + xiang' reading in Guangyun). In the Chinese Dialect Vocabulary entry for

'frog' (青蛙), dialects such as Yangjiang, Xiamen, Chaozhou, Fuzhou, and Changsha − primarily Yue and Hakka dialects − include terms that contain '蛤'.

Thus, a key question arises from this close geographic distribution presented on the Map 2: does it represent mere spatial coincidence, or does it suggest the occurrence of a specific historical linguistic event? To address this question and to investigate the etyma for 'frog' and 'toad' in the Mon-Khmer languages, the next section of this paper will continue the discussion, supported by evidence from historical linguistics, archaeology, and genetics.

4. Discussion

4.1. Historical linguistic data

Regarding the Vietnamese word *ếch* 'frog', the online World Loanword Database (WOLD) indicates no evidence of borrowing. Data for 'toad' is absent from the database. As Vietnamese is the only Mon-Khmer language represented in the WOLD corpus, its value for broader comparative analysis within the Mon-Khmer language family is consequently limited in the present discussion. This section therefore examines the reconstructed forms for 'frog' and 'toad' in Mon-Khmer languages, drawing primarily on data from the Mon-Khmer Etymological Dictionary (MKED). Although certain reconstruction results in the MKED remain open to scholarly debate (cf. Sidwell & Alves 2023), the database is nonetheless considered a highly reliable source for the historical linguistic study of Mon-Khmer languages.

Table 3: Reconstructed forms of 'frog' and 'toad' in Mon-Khmer languages

Word	Branch/group	Reconstructed forms
frog	Proto-Mon Khmer	*[ ]r[ɔ]k, *kit, *kiit, *ku[ə]t, *kət, *kəət
	Proto-Bahnaric	*ki(ː)t
	Proto-Katuic	*ʔaguut
	Proto-Khmuic	*ɟʔɛʔ
	Proto-Vietic	*gɔːt, *ʔeːk
toad	Proto-Mon Khmer	*[ ]r[ɔ]k
	Proto-Bahnaric	*-rɔk
	Proto-Katuic	*ʔaguut traak, *ʔagɔk, *ʔarɔk, *ʔaʔɔk

	Proto-Khmuic	*hro(ː)k
	Proto-Vietic	*-rɔk, *raːk, *roːk, *-duːt

It appears that Proto-Mon-Khmer likely possessed a single term ***[ ]r[ɔ]k** encompassing both 'frog' and 'toad'. This form represents a robust and well-supported reconstruction for 'toad' across several branches—including Bahnaric, Katuic, Khmuic, and Vietic—and for 'frog' at the deepest Proto-Mon Khmer level. As such, it stands out as a strong candidate for the common etymon of these amphibians. According to Alves (2009, p. 4), a form recognized in Proto-Mon-Khmer is not considered a loanword even when it appears in other language families, such as Tai-Kadai or Sino-Tibetan. Thus, based on this evidence, it can be asserted that the etymon for both 'frog' and 'toad' in Mon-Khmer languages is ***[ ]r[ɔ]k**.

This analysis leads us to revisit the question, introduced in Section 3, regarding any potential connection or implication between the phonetic forms of 'frog/toad' in Mon-Khmer and some Tai languages. The interaction and relationship between Mon-Khmer and Tai languages have been a significant and extensively researched topic in the historical linguistics of Southeast Asia. Although they belong to two distinct language families, their long history of geographical and cultural contact has led to considerable influences.

Alves (2009, p. 4) also stated that, based on somewhat tentative data—essentially a few dozen loanwords at most (Nguyễn Tài Cẩn 1995, p. 322, as cited in Alves 2009, p. 4)—Proto-Tai peoples may have come into contact with the ancestors of the Vietnamese before the Han Dynasty (200 BCE to 200 CE). At that time, they presumably shared with the Vietnamese some technology related to agriculture and animal husbandry and the associated vocabulary (e.g., *vịt* "duck" and *mương* "canal"). Pittayaporn (2019) determines Proto-Tai is the ancestor of the Tai languages of Mainland Southeast Asia. In his Proto-Tai reconstructions, the form for 'frog' is: ***kɤp^{D}**, and for 'small frog' is: ***krwe:^{C}**. This reconstructed form, however, does not provide significant insight into the question under consideration. Therefore, additional evidence is needed to resolve the issue.

4.2. Additional archaeological and genetic data

Drawing upon historical linguistic analysis and archaeological-genetic research findings, scholars such as Alves (2019, 2022b) and Trần Trí Dõi (2022) agree that speakers of Austroasiatic languages have been present in the Red River Delta since approximately 2000 BCE. During the Phùng Nguyên culture, the region's population

was predominantly Austroasiatic-speaking. By the Đông Sơn culture, this Austroasiatic-speaking populace gradually diversified into groups speaking languages of the Vietic branch, who likely engaged in varying degrees of contact with neighboring Tai-speaking populations. This period of initial linguistic contact between Vietic and Tai peoples likely commenced in the Iron Age, around 500 BCE, coinciding with the onset of the Đông Sơn culture and significant socio-cultural developments leading to the Cổ Loa site in northern Vietnam (Alves 2022a, p. 18). Trần Trí Dõi (2022) affirms that among various hypotheses regarding the language of Đông Sơn culture residents, the proposition that this community primarily spoke Vietic languages and engaged in contact with neighboring Tai-speaking populations seems to be strongly supported by linguistic, archaeological evidence. Consequently, this view is largely accepted by many researchers of Vietnamese linguistic, cultural, and societal history during the Đông Sơn period. This historical and linguistic understanding of the region's population dynamics is strongly corroborated by archaeological and recent genetic evidence.

Indeed, specific genetic analyses provide concrete support for this narrative. For instance, research by Vietnamese and international scholars (Dang Liu et al. 2019) indicates that the genetic make-up of the Vietnamese people is directly inherited from the Đông Sơn culture, showing only approximately 10-15% genetic divergence compared to ancient samples from the Đông Sơn culture site of Núi Nấp. Furthermore, evidence from craniometric measurements suggests that the Phùng Nguyên and Đông Sơn cultures were characterized by Austroasiatic-speaking populations (Matsumura et al., 2019). Such findings reinforce the long-standing presence and predominant influence of Austroasiatic-speaking groups in the Red River Delta during these formative periods.

When examining Đông Sơn culture, the bronze drum emerges as a profoundly significant archaeological artifact. Scholarly discourse concerning its origins primarily revolves around two ancient formative centers: Vietnam and Yunnan (China). Both are recognized as the foremost production hubs during the early development of bronze drum culture. As Chiou-Peng (2009) notes, "The data accumulated for more than half a century have now attested that comparable artifacts from the Dong Son region and Yunnan belonged to two cognate cultures in the Yue-based cultural sphere of southwest China and the Indochinese peninsula" (Chiou-Peng 2009, p. 34).

Regarding the classification of bronze drums, a slight divergence exists between Chinese and Vietnamese scholarly systems. Chinese scholars have categorized drums discovered within China and those held in Chinese collections into eight distinct types. This differs from the system proposed by Fr. Heger, currently adopted by Vietnamese

researchers, which includes one fewer type. Despite these classification differences, there is a broad consensus among researchers regarding the paramount importance of Heger Type I bronze drums, widely considered the earliest form. Notably, Vietnam possesses the largest quantity of these crucial artifacts, with their number nearly double that found in China and equivalent to the combined total from all other nations (Li & Huang 2016).

In Vietnam, Heger I bronze drums, also designated as Đông Sơn bronze drums, are further categorized by archaeologists into five chronologically ordered groups: A, B, C, D, and Đ. A detailed presentation of this system is omitted here, as bronze drum research is not the primary focus of this paper. Nevertheless, this classification is pertinent, as toad/frog statues, a key subject of interest, began to appear on the drum faces of those belonging to Group C (Phạm 2024, pg.48). A notable point of contention among scholars is the ongoing lack of consensus regarding the precise identification and nomenclature of these symbols—specifically, whether they represent frogs or toads (Phạm 2024, pg.44). This inconsistency in identification is entirely explainable. Firstly, from a visual standpoint, the shape of these symbols is not sufficiently distinctive, thereby hindering clear differentiation between toads and frogs. Secondly, within Vietnamese culture, both frogs and toads hold profound cultural symbolic significance, being intrinsically linked to the characteristics of wet-rice agriculture and the associated rain-praying beliefs. This also provides a foundational basis for understanding why only a single Proto Mon-Khmer reconstructed form exists for both 'frog' and 'toad'.

Numerous archaeological studies indicate that the discovery of Đông Sơn bronze drums (particularly Heger Type I) in areas of Southern China provides significant evidence for extensive trade and cultural exchange during and after the Đông Sơn culture. The widespread discovery of Heger Type I bronze drums and other artifacts across both Vietnam and the South China region has significant implications regarding cultural contact and the potential migrations of populations and language families in history.

Trịnh Sinh (2024) posits the extensive spread of Đông Sơn culture throughout Southeast Asia and Southern China and suggests insightful assumptions regarding the ancient routes of cultural exchange based on the discovery of Đông Sơn significant artifacts. Among the routes of Đông Sơn culture diffusion he proposes, there is a northeast coastal route to Guangdong and Zhejiang that facilitated trade and cultural interaction and a route along the Red River extending southwest, fostering exchanges with the Dien culture in Yunnan.

Figure 1: The coastal exchange route of Dong Son culture extended up to Zhejiang, evidenced by: 1. Nine bronze situlae in the Nanyue King's Tomb; 2. Mingqi drums (funerary drums) in the Shangmashan Tomb; 3. Bronze situlae and axes in Zhaoqing; 4. Daggers with human-figure hilts from Shumuling (Trịnh Sinh 2024, pg. 41)

Figure 2: The exchange routes of Dong Son culture with cultures in South China, evidenced by excavated and discovered Dong Son drums and situlae, include locations such as: 1. Shizhaishan; 2. Kaihua; 3. Azhangzhai; 4. Lijiashan; 5. Putuo; 6. Diandong; 7. Guixian; 8. Luobowan; 9. Xuzhuang; 10. Huili (Trịnh Sinh 2024, pg. 47)

The diffusion of Dong Son culture, as demonstrated by archaeologists, suggests a close contact between speakers of Taic and Vietic languages. This interaction likely stemmed not only from Taic southward migration but also from Dong Son inhabitants' trade and exchange with people in the North.

A hypothesis can be suggested: the exchange of artifacts, especially bronze drums featuring toad/frog symbols, facilitated the diffusion of Vietic phonological forms /**kok**/ and /**kɔk**/ into certain Taic languages. This diffusion subsequently led to the retention of these forms in some Taic dialects in Southern China today, as evidenced by earlier data.

Thus, building upon these archaeological and linguistic observations, a plausible scenario for language contact during the Dong Son period, particularly concerning the terms for 'toad' and 'frog', can be proposed: The Proto Mon-Khmer etymon for both 'toad' and 'frog' was a shared form ***[ ]r[ɔ]k**. As Mon-Khmer languages evolved, a distinction between 'toad' and 'frog' emerged in certain groups, though this differentiation wasn't always clear. By the Dong Son period, Mon-Khmer languages, primarily represented by the Vietic group, began linguistic contact with southward-migrating Taic-speaking populations. Simultaneously, with the diffusion of Dong Son culture, some Vietic speakers also interacted with Taic speakers to the north, allowing the proto-Vietic form for 'toad' ***ro**:**k** to spread to some Taic languages in that region. The shared initial consonants and syllable-final structures between Proto-Tai and Proto-Vietic phonological forms likely led to their interchangeable use across various Taic dialects.

However, this hypothesis remains tentative due to the limited Taic language data presented in this paper. Further in-depth research with a more comprehensive dataset is required to construct more detailed linguistic maps, which would enable a more definitive affirmation or refutation of this hypothesis.

Moreover, Alves (2015) highlights imitative onomatopoeia as a notable phenomenon in the coinage of terms for various bird species in Asia, both wild and domesticated. For instance, 'crow', 'pigeon/dove', and 'owl' have names clearly related to their calls. This mechanism is presented as a reasonably intuitive process, also seen in the term for 'cat', represented as MAO or a similar sounding word across Chinese, Tai, Vietnamese, and other regional languages (Alves 2014, pg. 40). Given that imitative onomatopoeia represents a significant universal lexical innovation, it is thus proposed that the phonological similarity between the Vietic word form for 'toad' and 'frog' (e.g., /**kok**/, /**kɔk**/) and its counterparts in some Taic dialects could also stem from this phenomenon. This suggests that these shared forms might have originated from the imitation of the amphibian's call, rather than exclusively through direct language contact and diffusion. "Such sound-symbolism adds uncertainty to claims of borrowing in a certain direction, especially at time depths of thousands of years, unless there is sufficient clarifying linguistic and extralinguistic evidence" (Alves 2015, pg. 41).

5. Conclusion

This paper investigates the possibility of linking geographical distribution and word form similarity to specific language interaction events, focusing on 'frog' and 'toad' terms. A preliminarily supported hypothesis proposed here suggests that the

geographical distribution of 'frog' and 'toad' variants in Taic languages, which are closely aligned with or even interspersed among Mon-Khmer languages, especially Vietic variants, resulted from language and ethnic contact during the Đông Sơn period. However, due to the limitation of current data, this remains a tentative hypothesis requiring further evidence. Additionally, the potential for the phonological similarity between Taic and Vietic variants to stem from imitative onomatopoeia must also be considered. Future research, therefore, must aim to thoroughly map these linguistic variations to definitively clarify the complex interplay between population movements, cultural exchange, and the evolution of shared lexical forms in ancient Southeast Asia.

References

Alves, Mark (2009) Loanwords in Vietnamese. In Martin Haspelmath and Uri Tadmor (eds). *Loanwords in the World's Languages: A Comparative Handbook*: 617-637. De Gruyter Mouton.

Alves, Mark (2015) Etyma for 'Chicken', 'Duck', & 'Goose' among Language Phyla in China & Southeast Asia. *Journal of the Southeast Asian Linguistics Society* 8:39-55. http://hdl.handle.net/1885/16086

Alves, Mark (2019) *Dữ liệu liên ngành: chi Vietic kết nối với văn hóa Đông Sơn* [*Data from Multiple Disciplines Connecting Vietic with the Dong Son Culture*]. Presentation at University of Social Science and Humanities, Vietnam National University of Hanoi. doi: https://doi.org/10.13140/RG.2.2.26690.22720

Alves, Mark (2022a) Lexical evidence of the Vietic household before and after language contact with Sinitic. Trang Phan, John Phan, and Mark Alves (eds) *Vietnamese Linguistics: State of the field, JSEALS Special Publication No9* Vol.15 (4): 15-58. University of Hawai'i Press. http://hdl.handle.net/10524/52500

Alves, Mark (2022b). The Đông Sơn Speech Community: Evidence for Vietic. *Crossroads*. doi: https://doi.org/10.1163/26662523-bja10002

Alves, Mark (2023a) Proto-Austroasiatic Etymologies of Words Related to Household Structures. In Paul Sidwell & Mark Alves (eds) *Proceedings of the 9th and 10th International Conferences of Austroasiatic Linguistics, JSEALS Special Publication No. 12*: 1-18. University of Hawai'i Press. http://hdl.handle.net/10524/52517

Alves, Mark (2023b) Preliminary Etymological Notes on Vietnamese Words for Pottery. In Alves, M. J., Lâm, Q. Đ., Trịnh, C. L., Trần, T. H. H., & Dương, X. Q. (eds) *Researching and Applying Linguistics and Vietnamese Language Studies* 1-22. Tokyo: Geolinguistic Society of Japan.

Alves, Mark (2024) An Etymological Study of Vietnamese Words for Weaving and Woven Objects. In Phan, T., Nguyen, TC., Shimizu, M. (eds) *Studies in Vietnamese Historical Linguistics. Global Vietnam: Across Time, Space and Community*. Singapore: Springer. doi: https://doi.org/10.1007/978-981-97-4314-8_2

Alves, Mark (2025a) First Millennium CE Mainland Southeast Asian Regional Loanwords Related to Material Culture. In Alves, M. (ed) *Papers from the 33rd Conference of the Southeast Asian Linguistics Society (2024)*: 170-189. University of Hawai'i Press. https://hdl.handle.net/10524/52539

Berlin, B. (1992) *Ethnobiological Classification: Principles of Categorization of Plants and Animals in Traditional Societies*. Princeton, NJ: Princeton University Press.

Blust, R. A. (2022) The History of Faunal Terms in Austronesian Languages. *Oceanic Linguistics* 41(1): 89-139. doi: https://doi.org/10.2307/3623329

Bui, Xuan Dinh (2015) Ethnic groups of Viet –Muong languages and Dong Son culture. *Vietnam Social Sciences*, No.4 (168), 82-92.

Chamberlain, James R. (2016) Kra-Dai and the Proto-History of South China and Vietnam. *Journal of the Siam Society* Vol. 104: 27-77.

Chiou-Peng, Tzehuey (2008) Dian Bronze Art: Its source and formation. *Bulletin of the Indo-Pacific Prehistory Association* Vol. 28: 34-43. https://doi.org/10.7152/bippa.v28i0.12013

Dang, Liu, Nguyen Thuy Duong, Nguyen Dang Ton, Nguyen Van Phong, Brigitte Pakendorf, Nong Van Hai, Mark Stoneking (202) Extensive Ethnolinguistic Diversity in Vietnam Reflects Multiple Sources of Genetic Diversity. Molecular Biology and Evolution Volume 37, Issue 9: 2503–2519. doi: https://doi.org/10.1093/molbev/msaa099

Driem, G. van (2011) The ethnolinguistic identity of the domesticators of Asian rice. *Comptes Rendus Palevol* 11 (Issues 2–3):117-132. doi: https://doi.org/10.1016/j.crpv.2011.07.004

Endo, Mitsuaki [遠藤光暁] (2001) *Kango Hōgen Ronkō*『漢語方言論稿』[Articles on Chinese Dialects]. Tokyo: Kohbun.

Haspelmath, Martin & Tadmor, Uri (eds.) (2009) *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wold.clld.org, Accessed on 28 June 2025).

Higham, C.F.W., Alves, M.J. (2025) The Southeast Asian prehistoric house: a correlation between archaeology and linguistics. *Asian Archaeology*. doi: https://doi.org/10.1007/s41826-025-00107-0

Li Kunsheng, Huang Derong [李昆声, 黄德荣] (2016) *Lun Hegeer Yi Xing Tonggu*《论黑格尔□型铜鼓》[On Heger Type I Bronze Drums].《考古学报》[Acta Archaeologica Sinica] 02:173-208.

Liu, Dang, Nguyen Thuy Duong, Nguyen Dang Ton, Nguyen Van Phong, Brigitte Pakendorf, Nong Van Hai, Mark Stoneking (2020). Extensive ethnolinguistic diversity in Vietnam reflects multiple sources of genetic diversity. *Molecular Biology and Evolution* Volume 37 (9): 2503–2519. https://doi.org/10.1093/molbev/msaa099

Luo, Liming (editor-in-chief), Lu Zhenyu (Editor), Chen Fulong (Editor) (2004) *Zhuang-Chinese-English Dictionary / Cuengh Gun Yingh Swzdenj*. Beijing: Nationality Press.

Matsumura, H., Hung, Hc., Higham, C. (et al.) 2(019) Craniometrics Reveal "Two Layers" of Prehistoric Human Dispersal in Eastern Eurasia. *Sci Rep* 9: 1451. https://doi.org/10.1038/s41598-018-35426-z

Nguyễn, Văn Khang (et al.), Bùi Chỉ, Hoàng Văn Hành (2002) *Từ điển Mường-Việt* [Dictionary of Mường-Vietnamese]. Hà Nội: Nhà xuất bản Văn hóa Dân tộc.

Osada, Naoki, Masahiko Kumagai, Hitoshi Suzuki, and Mitsuaki Endo (2024) *Phylogeographic History of Plants and Animals Coexisting with Humans in Asia*. Singapore: Springer.

SEALang Mon-Khmer Etymological Dictionary.http://www.sealang.net/monkhmer/dictionary/. (Accessed on 20 April, 2025).

Phạm, Huy Thông (eds) (2024) *Trống đồng Đông Sơn ở Việt Nam* [*Dong Son Drums in Vietnam*]. Hà Nội: Nhà xuất bản Đại học Sư phạm.

Peiros, Ilia (1996) *Katuic comparative dictionary (Southeast Asia)*. Canberra: Pacific Linguistics.

The Pangloss Collection. https://pangloss.cnrs.fr/?lang=en (Accessed on 10 April, 2025)

Pittayaporn, Pittayawat (2009) *The phonology of Proto-Tai*. Dissertation presented to the Faculty of the Graduate School of Cornell University.

Sidwell & Alves (2023) Re-evaluating Shorto's MKCD reconstructions. In Paul Sidwell & Mark Alves (eds) *Proceedings of the 9th and 10th International Conferences of Austroasiatic Linguistics, JSEALS Special Publication No. 12*: 98-126. University of Hawai'i Press.

Sidwell, Paul (2015) Austroasiatic dataset for phylogenetic analysis: 2015 version. *Mon-Khmer Studies Journal (Notes, Reviews, Data-Papers)* 44: lxviii-ccclvii. https://doi.org/10.15144/MKSJ-44.LXVIII

Trần, Trí Dõi (2022) *Lịch sử ngôn ngữ người Việt góp phần tìm hiểu văn hóa Việt Nam* [History of Vietnamese language in contributing to understanding Vietnamese culture]. Hà Nội: Nhà xuất bản Đại học Quốc gia Hà Nội.

Trịnh, Sinh (2024) Sự lan tỏa của văn hóa Đông Sơn [The diffusion of Dong Son culture]. *Thông báo khoa học Bảo tàng lịch sử quốc gia [Vietnam Museum of History Bulletin*] 2(2024): 38-59.

Geolinguistic patterns of the word form for 'butterfly' in Tibetic languages of the eastern Tibetosphere

Hiroyuki Suzuki (Kyoto University)

Abstract: This study examines the lexical forms of the word 'butterfly' in Tibetic languages spoken in the eastern Tibetosphere, specifically the Khams and Amdo regions. The word exhibits significant variation in both lexical form and syllable structure, presents cartographic representations of their geographic distribution, and offers interpretive analyses of their phonological and morphological features.*

Key words: Tibetic, eastern Tibetosphere, number of syllables, coalescence, motivation

1. Introduction

This study investigates the lexical forms of the word 'butterfly' in Tibetic languages of the eastern Tibetosphere, specifically in the Khams and Amdo regions. This word exhibits significant lexical forms across Tibetic languages; in Literary Tibetan (LT), the standard equivalent is *phye ma leb*, which can be interpreted folk-etymologically as 'open-or-flat,' referring to the flapping motion of a butterfly's wing. The analysis focuses on the morphological structure and syllabic composition of the lexical forms used in these languages.

Notably, the word for 'butterfly' displays striking morphological variation across local varieties. Tournadre and Suzuki (2023:872–873) provide word forms such as *phye ma kha leb*, *phye ma lab rtse*, *nyi ma leb leb*, and *cem ce lha mo*. These are written in LT spelling and are not consistently attested in LT dictionaries. This study focuses on the geographical distribution of such variants. Based on the data collected through fieldwork in the eastern Tibetosphere, this examines the geographical distribution of lexical forms by mapping their geographical variation and offering interpretive

SUZUKI, Hiroyuki. 2025. Geolinguistic patterns of the word form for 'butterfly' in Tibetic languages of the eastern Tibetosphere. In Nor Hashimah Jalaluddin, Hiroyuki Suzuki and Mitsuaki Endo (eds.) *Proceedings of the sixth International Conference of Asian Geolinguistics*, 50–61. doi: https://doi.org/10.5281/zenodo.17204557

* The work is part of the research outcomes of JSPS KAKENHI Grant Nos JP24K23937, JP25K00454, and JP25K04040. I should thank Kenzo Okawa for providing the beautiful photographs of butterflies from the eastern Tibetosphere. My thanks also goes to the participants who provided me with insightful comments during the initial oral presentation.

analyses of their phonological and morphological features. This also contribute to our understanding of the lexical diversity and distribution of multilexical root words in Tibetic languages.

Butterflies are familiar to the Tibetans in the eastern Tibetosphere. Figures 1 and 2 show two butterfly species from the rGyalrong region (bTsan-lha County, rNga-ba Prefecture, Sichuan). Therefore, Tibetic-speaking people are more familiar with live butterflies in their natural environment than with animals such as lions (LT *seng ge*) and tigers (LT *stag*). These figures are related to the keys to the present analysis. Figure 1 illustrates an open-wing posture of the butterfly, whereas Figure 2 shows a closed-wing posture.

Figure 1: *Parnassius cephalus*. From the rGyalrong region. © Kenzo Okawa

Figure 2: *Polyommatus eros*. From the rGyalrong region. © Kenzo Okawa

2. Morphological analysis

This section explores the etymology or folk etymology of lexical forms for 'butterfly', illustrating the LT equivalents for each morpheme that constitutes the word.

2.1. Forms with LT equivalents

The LT equivalent for 'butterfly' is a three-morpheme form: *phye ma leb*. This form can be analysed morpheme by morpheme as follows:

phye: a verb meaning 'open'
ma: conjunction of distinctive concepts[1]
leb: an adjective morpheme meaning 'flat'

[1] The morpheme *ma* also appears in *rgya-ma-bod* 'a person with the hybrid origins of Han Chinese and Tibetan' (literally, Han Chinese-conjunction-Tibetan) and *bya-ma-byi* 'bat' (literally, bird-conjunction-mouse). In many Tibetic languages, *ma* is necessary to connect two different categories as a compound.

In total, the trisyllabic form denotes 'an object which is not always open, not always closed'; more simply, 'open-or-flat', describing the butterfly's wing movements based on folk etymology.

The pronunciation of each morpheme is dependent on the phonetic correspondence of each variety. See Suzuki (2024) for cases in the eastern Tibetosphere.

A related form of *phye ma leb* is *phye ma ka leb* or *phye ma kha leb*, in which *ka leb* or *kha leb* is linked to the word for 'lid'. This quadrisyllabic form can be interpreted as 'opening-[and-closing]-or-lid', metaphorically describing the butterfly's wings to a lid that opens and closes.

A similar form that also employs the first two morphemes of *phye ma leb* is *phye ma lab rtse* or *phye ma la btsas*.[2] This quadrisyllabic form includes a disyllabic noun *lab rtse* or *la btsas*, both of which are written variants, indicating 'heaps of stones with flags in honour of local deities'. This object (Figure 3) is specific to the Tibetan culture, and no equivalents in other languages are available.

Figure 3: Labtse in Lhagang, Minyag Rabgang. © 2014 Author

[2] Two dictionaries of Amdo Tibetan, however, give spellings *phye ma leb rtse* (Hua and Klu-'bum-rgyal 1993:352) and *phye ma leb tse* (Geng et al. 2007:518).

This quadrisyllabic form can be analysed as 'open-or-*labtse*' describing the butterfly's wings with *labtse*-like antennae, as depicted in Figure 1. Although LT does not include a word form for 'butterfly containing *lab rtse* or *la btsas*', each morpheme of the quadrisyllabic form is independently attested in LT.

2.2. Other forms

Two additional types are identified: one with reduplication syllable, sometimes linked to LT equivalents; the other is miscellaneous.

The reduplicated type includes /ta pʰe: pʰe:/ (Lhagang), /ᵐbẽ bo/ (Sangdam), and /pi bi tʰa ro/ (dGonpa).[3] These examples contain the bilabial sounds /pʰ/ and /b/. This sound may reflect the iconicity of the action of flying, or to an onomatopoetic expression. Additionally, the form of sNyingthong, /ɕʰə mə pʰɔʔ/, appears to be a hybrid of LT *phye ma* and the onomatopoetic bilabial sound.

The miscellaneous type includes /ŋi mʌ le le/ (Zhollam), /ʔa ku pɛ leʔ/ (sKobsteng), /ʔɐ kʰo pe: mɐ/ (nKhorlo), /ⁿdza mo la tsa/ (sTaglo), /xʰɔ ᶮba ʰta pa/ (dGonpa), /ʔa ɕʰə ⁿ̥tʰe ma ka/ (Thopa), /ᶮgo xʰɯ tɕiʔ leʔ/ (Gadnagshod), /ᵐbɯ xu rɐ̀ / (Ongsum), *inter alia*. The forms of Zhollam, sKobsteng, and nKhorlo are loosely related. They belong to the Melung subgroup of Sems-kyi-nyila Tibetan: the syllable /le(ʔ)/, which may be related to the LT morpheme *leb* 'flat', occurs in the region between Zhollam and sKobsteng. The disyllabic morphemes /ʔa ku/ and /ʔɐ kʰo/, likely related to the word for 'uncle' (LT *a khu*), are found between sKobsteng and nKhorlo.

3. Mapping and discussions

Based on the morphological analysis presented in Section 2, cartographic representations are examined according to the classification outlined below.

A: *phye ma leb* - *phye ma ka leb* type (including LT equivalents of *phye* and *leb*)
- A1a: trisyllabic form without syllable coalescence
  - e.g., ɕʰe ma leʔ, ɕʰə ma le:
- A1b: trisyllabic form with some modifications
  - e.g., ᵖtsʰa: bɯ leʔ, ɕʰõ ɦa leʔ, ɕʰə wã la
- A1c: coalescent form from A2a
  - e.g., ɕʰɑ̃: ka ljɔʔ, ɕʰɑ̃: ka ljəʔ, ɕʰa: kə lɔʔ
- A2a: quadrisyllabic form

[3] The suprasegmental description is uniformly omitted when citing word forms.

e.g., ɕʰə mã kə ljɔʔ, ɕʰa ma ka lɑʔ, ɕʰe mɑ: ɦa lɤʔ, sʰo mɔ ka leʔ

A2b: quadrisyllabic form with some modifications

e.g., ɕʰe mɑ: ɦa lɤʔ, ɕʰɤ wã kə ljɤʔ, ʔa ma kə ljɯ

B: *phye ma lab rtse* - *phye ma la btsas* type (connected to *labtse*)

e.g., ɕʰi ma low ʰtse, xʰa mo la tse, ho mo la dza

C: reduplicative forms including labial initials (see Section 2)

D: miscellaneous (see Section 2)

3.1. Mapping the morpholofical structure

Figure 4 shows a map reflecting the aforementioned classifications. Type A is represented by the black symbols. Type B is sky blue, whereas Types C and D have different shapes in red. Notably, Type D includes forms not classified in the other types; hence, the distribution of Type D does not represent any geolinguistic significance.

Figure 4 indicates the red line dividing the target region into two parts. The north-eastern region (upper part of the red line) is traditionally referred to as Amdo, and the south-eastern region (lower part of the red line) is named Khams.[4] An overall observation of the distribution of the three types (A, B, and C) is as follows: Type A is mainly found in Khams, and Type B is dominant in Amdo. Type C is concentrated in the central area of the map, that is, the easternmost area (Suzuki and Sonam Wangmo 2015) of the Khams region.

It is noteworthy that Type B also forms a minority in the distribution area of Type A. Based on current reports (Suzuki and Sonam Wangmo 2016, 2017, 2019; Suzuki 2018, 2022), these are exclaves of Amdo Tibetan with oral histories which can relate the speakers' ancestors to their homeland in Amdo. The distribution of Type B in Khams denotes the retention of the lexical item for 'butterfly', showing neither influence from, nor exerting influence on, neighbouring Khams varieties.

Next, we focus on the distribution of Type A subclasses in the Khams region. Types A1a and A1b (trisyllabic forms) are predominant in northern Khams, whereas the others are mainly observed in southern Khams. The phenomenon of Type A1c belongs to a phonological change (coalescence of a LT suffix *po*, *pa*, *mo*, or *ma* and its preceding syllable), and hence does not appear only in this lexical item. Varieties with this phonological change are distributed in a given area (sPomborgang, Chaphreng, and Sems-kyi-nyila). The map indicates that Types A1a, A1c, A2a, and A2b coexisted in the central Khams area.

[4] For a more detailed information of the traditional description and language classification, refer to Ryavec (2015) and Tournadre and Suzuki (2023).

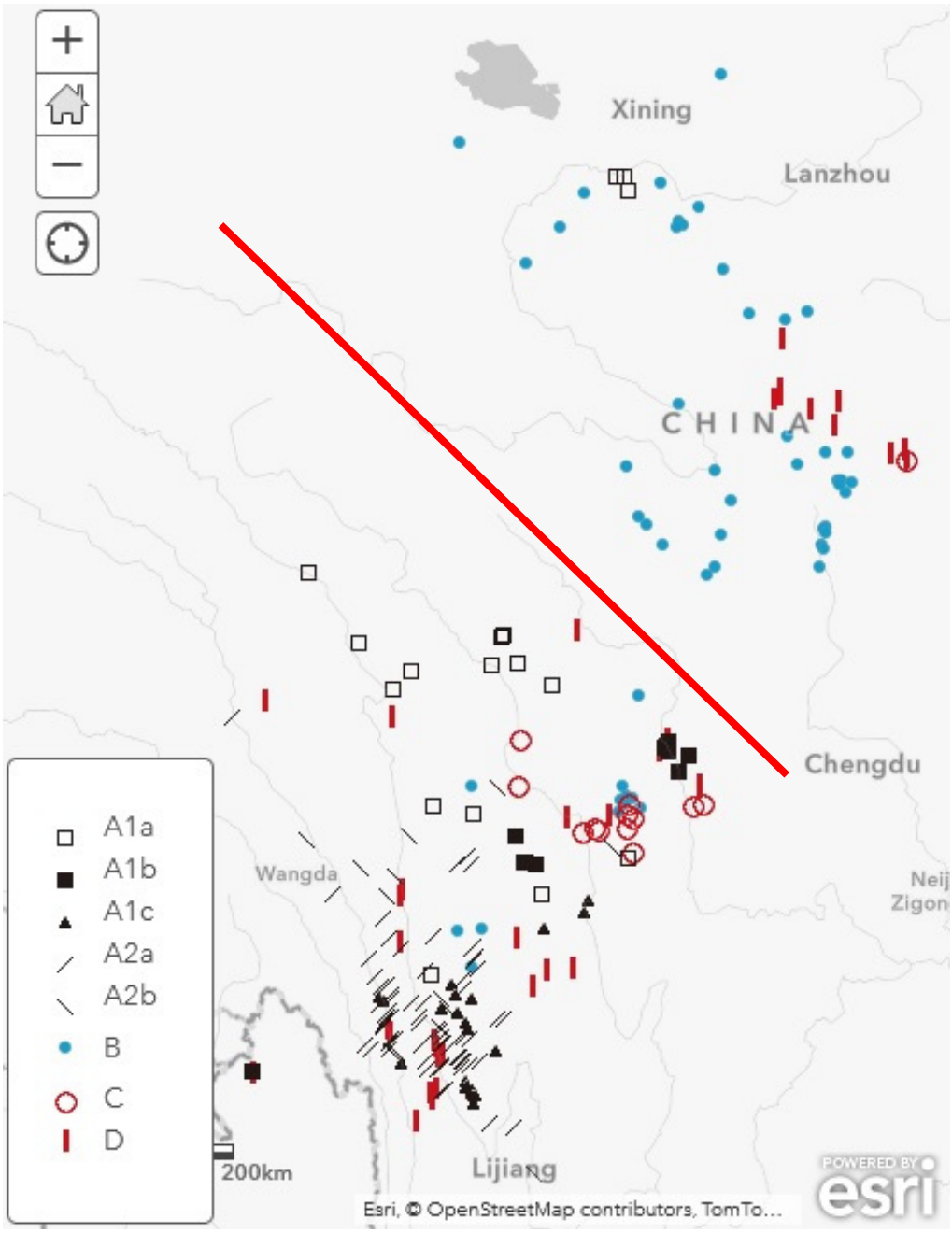


Figure 4: Map based on the morphological structure

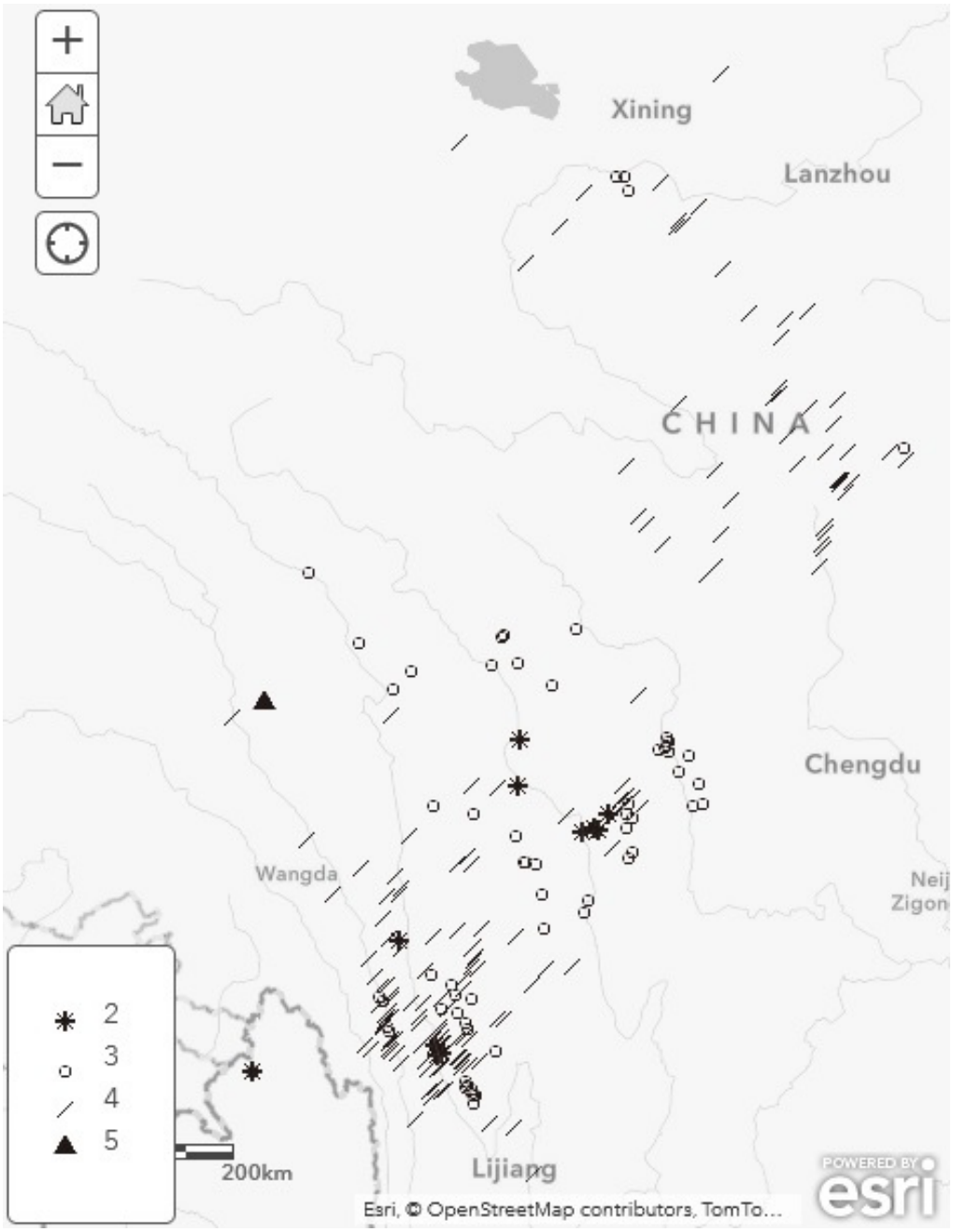


Figure 5: Map based on the syllable number

3.2. Mapping the number of syllables

Figure 5 indicates the number of syllables in word forms for 'butterfly', regardless of lexical type, ranging from two to five. Disyllabic forms appear centrally within a limited area, bordered by trisyllabic and quadrisyllabic forms, with some exceptions in Southern Khams. Trisyllabic forms span a broad central area, overlapping with quadrisyllabic forms, which dominate the eastern Tibetosphere. Quadrisyllabic forms are also found in both northern and southern regions, separated by disyllabic and trisyllabic zones. A cinquisyllabic form is found in isolation, in Thopa (/ʔa ɕʰə ⁿtʰe ma ka/).

Considering, language classification, quadrisyllabic forms, along with disyllabic and trisyllabic forms in the central areas, are found in Amdo Tibetan varieties (cf. the distribution of Type B in Figure 4). However, trisyllabic and quadrisyllabic forms also co-occur in several Khams Tibetan subgroups. Hence, it is essential to interpret the distribution in light of language affiliations and internal developments.

3.3. Discussions

Based on Figures 4 and 5, the following observations are made: language-level tendency, the ABA (concentric) distribution of tri- and quadrisyllabic forms is observed in Khams, reflecting a relationship between the trisyllabic and quadrisyllabic forms of Type A.

The language-level tendencies are conceptualised as follows: Amdo Tibetan varieties predominantly exhibit Type B forms. Even in in Khams speaking areas (the central region of the maps), Type B is retained. In contrast, Khams Tibetan varieties display disyllabic, trisyllabic, or quadrisyllabic forms of Type A, with no attestation of Type B.

In relation to syllable number, the ABA distribution of tri- and quadrisyllabic forms is attested in Khams. As indicated in Figure 5, the quadrisyllabic form appeared later in Khams. This suggests that Amdo-speaking people later came to the Khams area, which is historically correct (Suzuki and Sonam Wangmo 2019).

Based on the observations in Section 3.1, the relationship between the trisyllabic and quadrisyllabic forms of Type A, that is, A1c and A2a, is recognised. By definition, the trisyllabic form A1c is derived from the coalescence of two syllables within the quadrisyllabic form A2a. The condition for the coalescence of two syllables is that the second syllable is equivalent to the LT suffixes *pa*, *po*, *ma*, *mo*, etc. For the lexical form for 'butterfly', the first two syllables are *phye ma*, which fulfils a condition of the coalescence. This phenomenon has been attested in many Tibetic varieties, from Southern Khams to mNga'-ris (Jiang 2002:70–76). Particularly, the Southern Khams

varied across varieties. Coalescence does not always appear in each variety in a regular manner; some have many words with coalescence, whereas others do not. This observation suggests that coalescent forms for 'butterfly' display notable geographical continuity, as shown in Figure 5. This highlights the need to geolinguistically analyse each word from exhibiting syllabic coalescence.

3.4. Notes from the viewpoint of *motivation*

Motivation is a method of cross-linguistically effective geolinguistic analysis (Del Giudice & Brun-Trigaud 2024). From this standpoint, a brief analysis of Tibetic, Sinitic, and Japanese forms for 'butterfly' is provided.

Motivation is analysed by comparing morpheme-level meanings of the word forms. As provided in Section 2, Tibetic word forms includes the following semantic units: 'open' (*phye*), 'flat' (*leb*), 'lid' (*kha leb*), and 'labtse' (*la btsas*). Of them, LT *leb* 'flat' is cognate with Sinitic 蝶 *die* 'butterfly'. Its Old Chinese form is *l$^{\text{ʕ}}$ep (Baxter and Sagart 2014), and its Middle Chinese form is *depD*, which was actually borrowed into Old Japanese てふ *tehu*. Its reduplication form (てふてふ *tehutehu*) is the modern spoken form ちょうちょう *tyootyoo*. Therefore, from a viewpoint of motivation, the forms for 'butterfly' in Tibetic, Sinitic, and Japanese include a common semantic feature; however, its original meaning has been lost.

Another noteworthy feature is the use of labial sounds, such as /p/ and /($^{\text{m}}$)b/, which may reflect onomatopoeia iconicity of wing movement. Indeed, searching for word forms for 'butterfly' in Tibeto-Burman languages with the STEDT database (Matisoff 2015), we find many forms including /p/ with a Proto-Tibeto-Burman form pur ⪤ pwar (#355; BUTTERFLY). A detailed discussion is beyond the scope of this study, but future research could explore sound iconicity as part of the motivation in constructing lexical forms for 'butterfly'.

5. Concluding remarks

This article examined the word form for 'butterfly' in Tibetic languages of the eastern Tibetosphere by analysing variation in lexical forms and syllable count. It then presented linguistic maps for both features and offers an interpretive analysis of their distributions. Three principal lexical forms are recognised: *phye ma leb* - *phye ma ka leb* type, *phye ma lab rtse* - *phye ma la btsas* type, and reduplicative forms, including labial initials. The first two types are further characterised by the number of syllables. However, the number of syllables is not interpreted as reflecting ABA distribution for

the temporal differentiation, but rather as resulting from inherited sound change tendencies within a each variety.

This study broadens the perspective on lexical motivation by showing that Tibetic forms for 'butterfly' often include the LT morpheme *leb* 'flat', which may be cognate with a Chinese word for 'butterfly' and possibly related to a Japanese equivalent. Additionally, the frequent use of the labial initials suggests onomatopoetic iconicity linked to wing movement.

References

Baxter, William H. & Laurent Sagart (2014) Baxter-Sagart Old Chinese reconstruction, version 1.1. Unpublished manuscript, retrieved from https://sites.lsa.umich.edu/ocbaxtersagart/

Del Giudice, Philippe & Guylaine Brun-Trigaud (2024) Linguistic motives common to Japanese and Romance dialects: Two examples with maps. *Studies in Geolinguistics* 4: 20–39. https://doi.org/10.5281/zenodo.13948553

Geng, Xianzong, Junying Li, and Lhun-grub rDo-rje (2007) *Anduo Zangyu kouyu cidian*《安多藏语口语词典》[*Dictionary of spoken Amdo Tibetan*]. Lanzhou: Gansu Minzu Chubanshe.

Hua, Kan and Klu-'bum-rgyal (1993) *Anduo Zangyu kouyu cidian*《安多藏语口语词典》[*Dictionary of spoken Amdo Tibetan*]. Lanzhou: Gansu Minzu Chubanshe.

Jiang, Di [江荻] (1990) *Zangyu yuyinshi yanjiu*《藏语语音史研究》[*Study on Tibetan sound history*]. Beijing: Minzu Chubanshe.

Matisoff, James A. (2015) *The Sino-Tibetan etymological dictionary and thesaurus*. Berkeley: The Regents of the University of California. Data are also available as online STEDT Database. http://stedt.berkeley.edu/~stedt-cgi/rootcanal.pl

Ryavec, Karl E. (2015) *A historical atlas of Tibet*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226243948.001.0001

Suzuki, Hiroyuki [铃木博之] (2018) Litangxian ji qi zhoubian Zangzu yuyan xianzhuang diaocha yu fenxi 理塘县及其周边藏族语言现状调查与分析 [Research and analysis of the current statud of Tibetans' languages in Lithang County and its surroundings]. *Minzu Xuekan* 2: 35–44+106–109. https://doi.org/10.3969/j.issn.1674-9391.2018.02.005

Suzuki, Hiroyuki (2022) Amdo Tibetan-speaking Khampas in Lithang: Their language, identity, and migration history. Paper presented at 16th Seminar of the International Association for Tibetan Studies (Praha). https://doi.org/10.13140/RG.2.2.24354.50887

Suzuki, Hiroyuki (2024) Evolution of dorsal fricatives in rGyalthangic varieties of Khams Tibetan. *Journal of the Phonetic Society of Japan* 28(2): 107–118. https://doi.org/10.24467/onseikenkyu.28.2_107

Suzuki, Hiroyuki & Sonam Wangmo (2015) Challenge to discover endangered Tibetic varieties in the easternmost Tibetosphere: a case study on Dartsendo Tibetan. *Linguistics of the Tibeto-Burman Area* 38(2): 256–270. https://doi.org/10.1075/ltba.38.2.07suz

Suzuki, Hiroyuki & Sonam Wangmo (2016) Vocabulary of Shingnyag Tibetan: A dialect of Amdo Tibetan spoken in Lhagang, Khams Minyag. *Asian and African Languages and Linguistics (AALL)* 11: 101–127. https://doi.org/10.15026/89211

Suzuki, Hiroyuki & Sonam Wangmo (2017) Language evolution and vitality of Lhagang Tibetan: a Tibetic language as a minority in Minyag Rabgang. *International Journal of the Sociology of Language* 245: 63–90. https://doi.org/10.1515/ijsl-2017-0003

Suzuki, Hiroyuki & Sonam Wangmo (2019) Migration history of Amdo-speaking pastoralists in Lhagang, Khams Minyag, based on narratives and linguistic evidence. *Archiv Orientální* Supplementa XI: 203–222.

Tournadre, Nicolas & Hiroyuki Suzuki (2023) *The Tibetic languages: An introduction to the family of languages derived from Old Tibetan*. LACITO Publications. https://doi.org/10.5281/zenodo.10026628

Extraction of regularities and geographical patterns from the basic vocabulary of the Ainu language

Mika Fukazawa (National Ainu Museum)

Abstract: This paper demonstrates how to extract "regularities," such as phonological correspondences and morphological and grammatical rules, from basic vocabulary items and create relational maps in the Ainu language. Studies on basic Ainu vocabulary (Asai 1974), are based on Hattori & Chiri's (1960) list of 200 items taken from Swadesh's (1954, 1955) personal letters to Hattori (cf. Hattori 1954). This basic vocabulary contains information about lexical forms as well as various lexical correspondences and rules. For instance, the regularity of the semivowel (glide) type, -*iw*-:-*uy*, can be extracted from the items *ciw*:*cuy* for 'pierce (stab)' and *nociw*, *noociw*:*nocuy* for 'star' (Fukazawa & Ono 2025, Ono & Fukazawa 2025). When drawing a regularity map such as a lexical form map, it may be suggested that several lexical forms are integrated into one symbol with the corresponding regularity. However, as integrating symbols may exclude dialectal details, this study proposes drawing relational maps between different basic lexical maps.*

Key words: basic vocabulary, Ainu, phonological correspondences, lexical forms

1. Ainu language and the dialects

The Ainu language was (or still is, as a second language) spoken in Hokkaido, Sakhalin, the Kuril islands, and the northern part of mainland Japan. It is a language isolate that is typologically different from Japanese and other Asian languages (cf. Tamura 2000; Bugaeva 2022). Recently, it has been designated as an endangered language in Japan. It is now impossible to conduct a large-scale survey of the Ainu vocabulary. Even in 1960, a linguist of the Ainu language, Hattori Shirō, reported that some of his informants were among the last native speaker(s) of the Ainu dialect, and that they were very old. He also mentioned that some could speak the Ainu language fluently, while others knew only several words (Hattori & Chiri 1960).

FUKAZAWA, Mika. 2025. Extraction of regularities and geographical patterns from the basic vocabulary of the Ainu language. In Nor Hashimah Jalaluddin, Hiroyuki Suzuki and Mitsuaki Endo (eds.) *Proceedings of the sixth International Conference of Asian Geolinguistics*, 62–80. doi: https://doi.org/10.5281/zenodo.17204595

* This work was supported by JSPS KAKENHI Grant Numbers JP25K04114, JP23K25322. We would like to thank Editage (www.editage.jp) for English language editing.

When studying the Ainu dialects in the present, we have no choice but to use materials from the Ainu language of the past. The most reliable source of data for considering the classification of the Ainu dialects is the 200 (Asai's 202)[1] basic vocabulary items of the Ainu language, left by Hattori & Chiri (1960) and Asai (1974). Hattori & Chiri's (1960) data was mostly corrected by Hattori Shirō, Chiri Mashiho, and some collaborators from 1955 to 1956, and the data of Asai (1974) was based on the Hattori & Chiri (1960), his original survey of Asahikawa, Obihiro, and Kushiro dialects, and other written materials of Torii (1903), Murayama (1971), and Pinart (1872). Asai (1974) modified the data of Hattori & Chiri (1960) and added two dialects, Chitose and northern Kuril. See Table 1 and Figure 1 for the Ainu dialects used by Hattori & Chiri (1960) and Asai (1974).

Table 1: Ainu dialects

Hokkaido dialects	1. Yakumo, 2. Oshamambe, 3. Horobetsu, 4. Biratori, 5. Nukibetsu, 6. Niikappu, 7. Samani, 8. Obihiro, 9. Kushiro, 10. Bihoro, 11. Asahikawa, 12. Nayoro, 13. Soya, 21. Chitose
Sakhalin dialects	14. Ochiho, 15. Tarantomari, 16. Maoka, 17. Shiraura, 18. Raichishika, 19. Nairo
Northern Kuril dialect	20. Shumushu

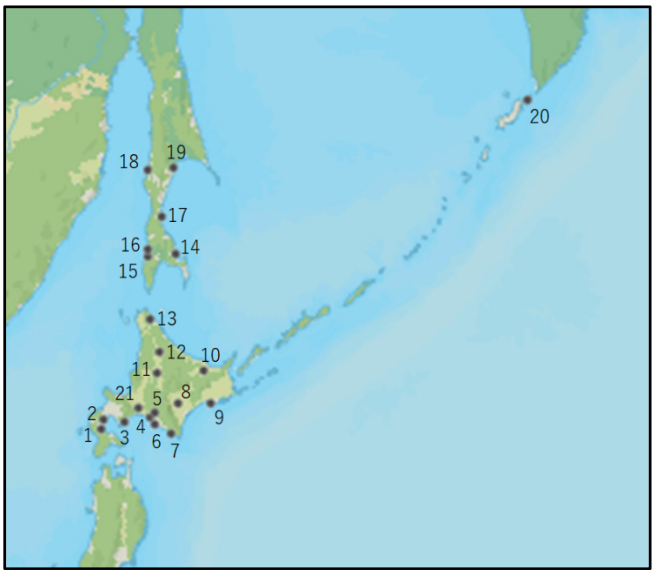


Figure 1: Map of the Ainu dialects[2]

[1] Asai (1974) divided the two basic vocabulary items of Hattori & Chiri (1960) into two respectively, increasing the total number of items from 200 to 202.

[2] The numbers correspond to those in Table 1.

It is important to note that differences existed not only between the datasets of Hattori & Chiri (1960) and Asai (1974), but also in their judgments of cognates on lexical items, which will be discussed in Section 2. The different judgments of cognates led to different available items with dialectal features and areal boundaries for their statistical approach. Thus, these two studies used different items from the list of 200 basic vocabulary items (or Asai's list of 202) in their analyses. Consequently, their conclusions on dialect classification differ. Hattori & Chiri (1960) demonstrate a Saru-Chitose (and Sakhalin) classification with an ABA distribution. In contrast, Asai (1974) demonstrates an Eastern-Western classification (see Figures 2 and 3, as well as the detailed discussion in Ono & Fukazawa 2024).

Nakagawa & Fukazawa (2022: 292-293) attempted to offer "a comprehensive categorization of Ainu dialects using indicators from vocabulary, phonetics, and morphosyntax, together with indicators of the geographical distribution of forms from a linguistic geography perspective." As examples of this, the article suggests that the Eastern-Western classification, also as known as the Northeast-Southwest division representatively corresponds with the distribution of the elision of word-initial /h/ and the pseudo sound correspondence *ca-*:*pa-*. The Saru-Chitose (and Sakhalin) classification shows a variety of differences in their person-marking systems, the morphology of interrogatives and existential verb, the lack of a juxtapositional particle, and the presence of a singular-plural distinction in the fourth person (cf. Nakagawa & Fukazawa 2022: 292-293).

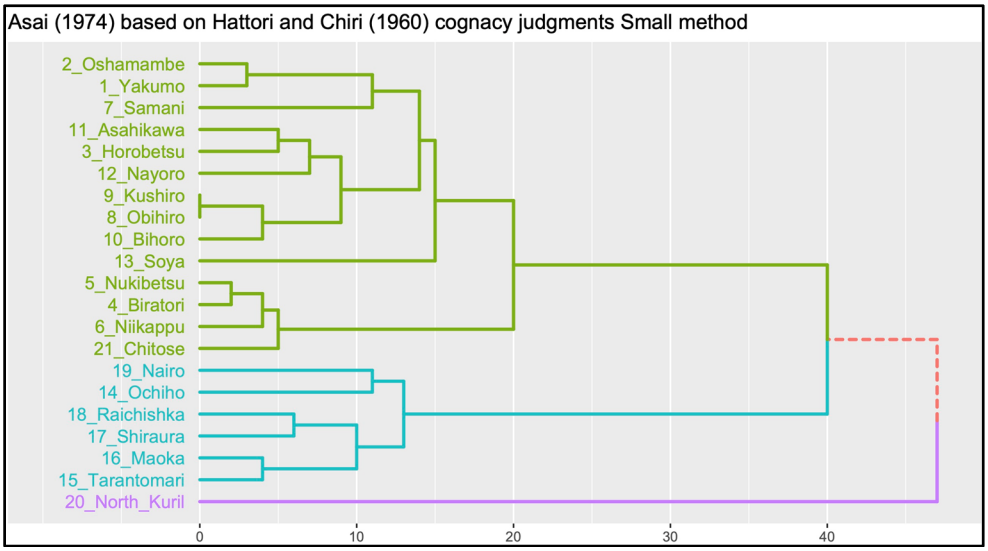


Figure 2: Hattori & Chiri's (1960) judgement on the Ainu dialects (Ono & Fukazawa 2024)

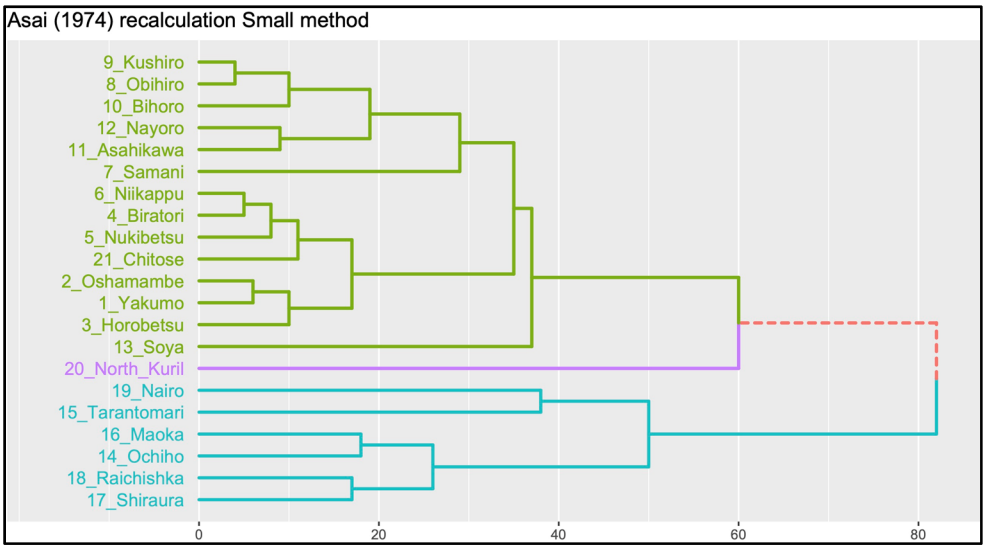


Figure 3: Asai's (1974) judgement on the Ainu dialects (Ono & Fukazawa 2024)

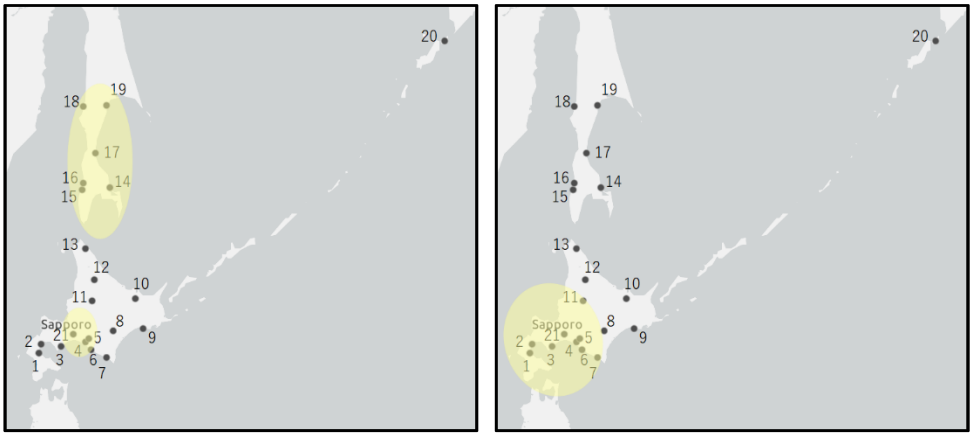


Figure 4: The maps of the Saru-Chitose (and Sakhalin) classification (Left) and the Eastern-Western classification (Right)

The next section demonstrates the classification of Ainu dialects using indicators from the vocabulary, phonetics, and morphosyntax of basic vocabulary items. These items are considered the most abundant and consistent data in the Ainu language.

2. Basic Vocabulary items of the Ainu language

In this section, we outline the basic vocabulary items of the Ainu language and discuss the challenges in creating geographical maps using them. Two hundred basic vocabulary items from Hattori & Chiri (1960) were taken from Swadesh's letters (1954, 1955) to Hattori (see Hattori 1954). The first 100 items in Hattori & Chiri (1960) are the same as those in Swadesh's (1955) "new list of 100 items" and have a relatively high retention rate. Therefore, it is difficult to borrow these items between dialects and languages (cf. Hattori 1954, Fukazawa 2025).

Both Hattori & Chiri (1960) and Asai (1974) are based on nearly identical items and word-form data. However, these studies differ in how they judge cognacy, or "similarity" among word forms of different dialects. Hattori & Chiri (1960) judged cognates based on shared roots. Asai (1974), on the other hand, judged cognates based on phonological "similarity." For example, the items of 'mouth' (No. 42), *paro, caro, cara,* and *caru*, are all cognates in Hattori & Chiri (1960), but non-cognates in Asai (1974). Similar issues arise when drawing geographical maps. This means that certain "similarity" rules must be established and adhered to when applying and adjusting the same symbols to the items.

2.1. Geographical maps of the first 100 items

Fukazawa (2018) attempted to create geographical maps of the first 100 items from Hattori & Chiri (1960). Fukazawa (2018) classified the item as the "monotonous type" if the lexical forms of a vocabulary item among all dialects had phonological correspondences and the same root. It was then suggested that 59 (including 5 possible) out of 100 basic items could be categorized as the monotonous type (see (1) below). This type is further classified into subtypes based on accent patterns and phonetic/phonological features.

(1) Monotonous type (Hattori & Chiri 1960; First 100 Items)
2. you [thou], (10. many), 11. one, 12. two, 13. big, 17. man, 18. person, 19. fish, (20. bird), 21. dog, 23. tree, 24. seed, 25. leaf, 28. skin, 29. meat, 30. blood, 31. bone, 32. grease, 34. horn, 36. feather, 39. ear, 40. eye, 41. nose, 43. tooth, 45. claw, 49. belly, 50. neck, 52. heart, 54. drink, 55. eat, 56. bite, 57. see, 58. hear, 60. sleep, 61. die, 62. kill, (63. swim), 65. walk, 66. come, 68. sit, 71. say, (72. sun), (73. moon), 75. water, 77. stone, 78. sand, 83. ashes, 84. burn, 85. path, 86. mountain, 87. red, 88. green, 89. yellow, 90. white, 92. night, 93. hot, 96. new, 97. good, 100. name (Fukazawa 2018: 78)

Thus, revising Fukazawa's (2018) monotonous-type list may be necessary. Ono & Fukazawa (2023) identify the items that Hattori & Chiri (1960) and Asai (1974) judged as cognates. Therefore, the revised monotonous-type list should also include a table comparing these items. The 55 cognate items are listed in (2). Of these, 36 items in parentheses were identified as cognates only by Hattori & Chiri (1960). This indicates that the Ainu language has small dialectal differences in lexical forms.

(2) Cognates judged by Hattori & Chiri (1960) and Asai (1974); First 100 items)
2. you [thou], (4. this), (11. one), (12. two), 13. big, 14. long, (17. man), 18. person, (19. fish), (21. dog), (23. tree), (24. seed), 29. meat, 30. blood, (31. bone), (32. grease), 34. horn, (36. feather), (37. hair), (38. head), (39. ear), (40. eye), (41. nose), (42. mouth), (43. tooth), (45. claw), 52. heart, 54. drink, 55. eat, 56. bite, 57. see, 58. hear, 60. sleep, (61. die), (65. walk), (66. come), (68. sit), (70. give), (71. say), (75. water), (77. stone), (78. sand), (79. earth), 83. ashes, (85. path), (86. mountain), 87. red, (90. white), (91. black), (92. night), (93. hot), (95. full), 96. new, 97. good, 100. name

A monotonous type is typically depicted as a monotonous distribution on a geographical map and is represented by a single symbol. Monotonous vocabulary items can be classified into subtypes with non-monotonous linguistic features. On a map of the subtypes, different symbols can be used to represent several vocabulary items with phonological, morphological, or grammatical features. Therefore, drawing maps of phonology, morphology, or syntax instead of maps of lexical forms is more meaningful for languages with small dialectal differences such as Ainu.

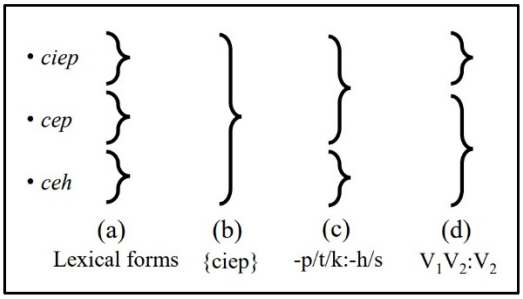
2.2. Issues of making regularity maps

This study will compile the vocabulary items that share patterns and characteristics into one map, as Fukazawa (2018) attempted previously, and refer to it as a "regularity map." While creating a regularity map, multiple lexical forms of a vocabulary item can be integrated into symbols with their corresponding regularities. However, these forms are not necessarily integrated into a single regularity map.

Here, consider the basic vocabulary item of 'fish' (No. 19), which Ono & Fukazawa (2025) and Fukazawa & Ono (2025) used to illustrate the regularity and division of its lexical forms. The vocabulary item for 'fish' (No. 19) has three lexical forms: *ciep*, *cep*, and *ceh*. The older form is *ciep*, which can be analyzed as *ci-*, the personal prefix of 1PL.EXCL.A, *e*, the transitive verb of 'to eat,' and *-p* (*-h* in Sakhalin),

the nominalizer of 'thing.' The forms *cep* and *ceh* occur through vowel reduction from *ciep* (cf. Chiri 1976 [1962]).

(3) The vocabulary item for 'fish' (No. 19)
(a) lexical forms: *ciep*, *cep*, and *ceh*
(b) {ciep}
(c) -p/t/k:-h/s − Regularity map
(d) V_1V_2:V_2 − Regularity map



If each of the four subtypes in (3a-d) were mapped, the maps would appear as shown in Figures 4 (a, b) and 5 (c, d). Integrating symbols, as in Figures 4 (b), 5 (c, d), may eliminate detailed dialect information compared to Figure 4 (a). In contrast, Figure 6 is a map with a single overlay of (c) and (d). While it shows most of the information, it is too complicated to discern the dialect features.

(a) (b)

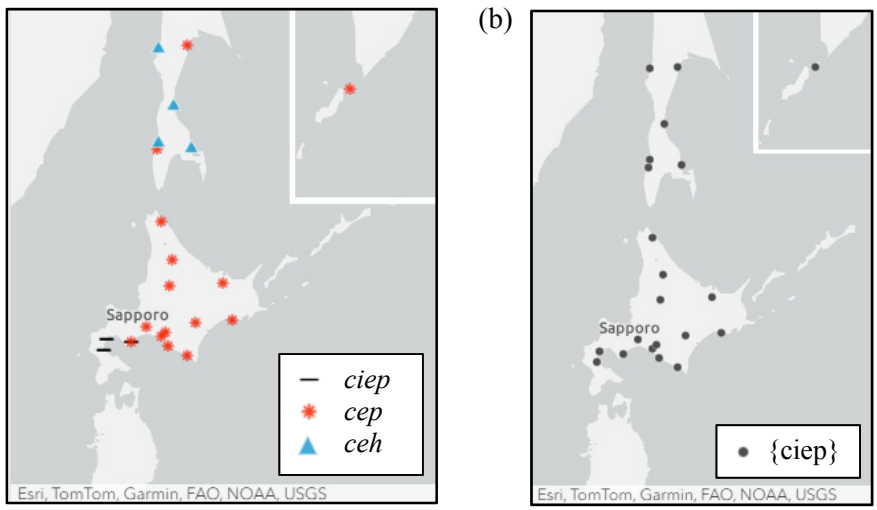


Figure 4: Maps of the (a) lexical forms of 'fish' and (b) {ciep}

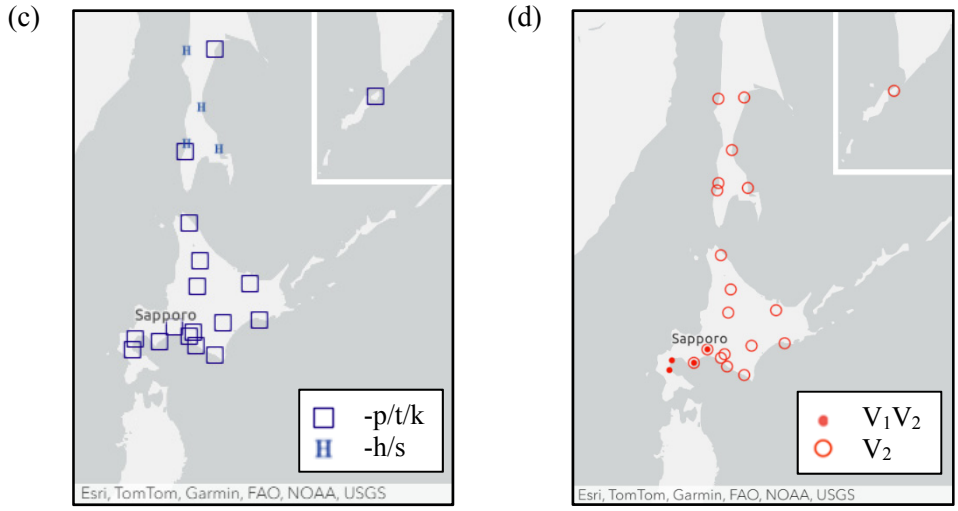


Figure 5: Maps of (c) -p/t/k:-h/s and (d) V_1V_2:V_2[3]

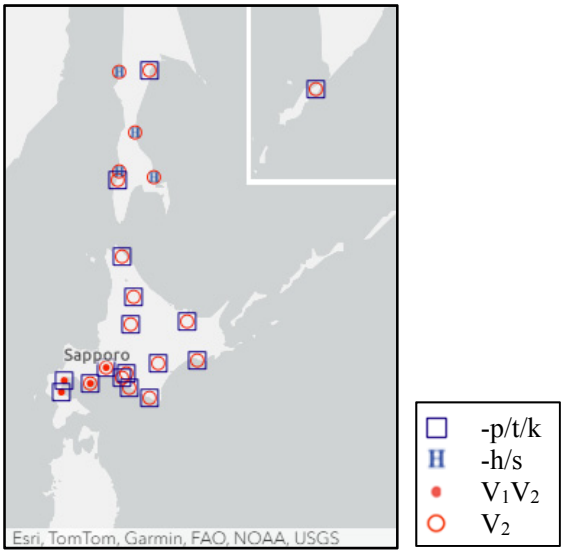


Figure 6: Layered Map of (c) -p/t/k:-h/s and (d) V_1V_2:V_2

The following section discusses how to draw regularity maps from basic vocabulary items that appropriately reflect dialect features and information while ensuring readability.

[3] Map of Figure 5 (d) is referred from Fukazawa & Ono (2025: Figure7).

3. Ways to draw regularity maps from basic vocabulary items

In this section, we will attempt to consider a method for drawing a regularity map of basic vocabulary items with the same patterns and characteristics. Basic vocabulary items contain information about lexical forms as well as various lexical correspondences and rules. The following examples use the basic vocabulary data from Hattori & Chiri (1960) and Asai (1974).

3.1. Regularity of the semivowel (glide) type

The regularity of the semi-vowel (glide) type *-iw*:*-uy* can be extracted from the items of (4).

(4) Regularity of semi-vowel (glide) types: *-iw* > *-uy* (metathesis form of *-iw*)
(a) Lexical forms of 'star' (No. 74): *nociw~noociw*:*nocuy*
(b) Lexical forms of 'pierce (stab)' (No. 179): *ciw*:*cuy*

Figure 7 shows the lexical forms of 'star' (No. 74). The lexical form *rikop* can be analyzed as *rik*, the noun 'sky,' *o*, the transitive verb 'be in,' and *-p*, the nominalizer of 'thing,' mening 'thing in the sky.' This form is probably newer than *keta* or *nociw*, although there is some controversy regarding which is older, *keta* or *nociw* (Fukazawa 2017: 100). The semivowel (glide) type *-iw*:*-uy* can be found in the lexical forms *nociw*, *noociw*, and *nocuy*. The *-uy* form of *nocuy* exists in the Kushiro dialect.

Figure 8 shows the lexical forms of 'pierce (stab)' (No. 179). Hattori (1964:146) recorded the lexical form *eciekara* in Asai (1974) as the *-iw* form *eciwkara*. The form of *ciri* in Asai (1974) appears to be a miswriting of *civi*, which can be assumed to be the *-iw* form of /ciwi/ (cf. Murayama 1971: 152). Thus, two distinct lexical formtypes are observed: {*otke*} and {*ciw*}. However, they may suggest a historical convergence from distinct semantic origins since {*ciw*} can also express the meaning of 'to ripple.' The semivowel (glide) type *-iw*:*-uy* can be found in the lexical forms *ciw*, *ciwi*, and *cuy*. The *-uy* form of *cuy* exists in the Bihoro dialect.

The regularity map in Figure 9 is derived from Figures 7 and 8. The *-uy* form is found only in eastern Hokkaido dialects, such as Kushiro and Bihoro.

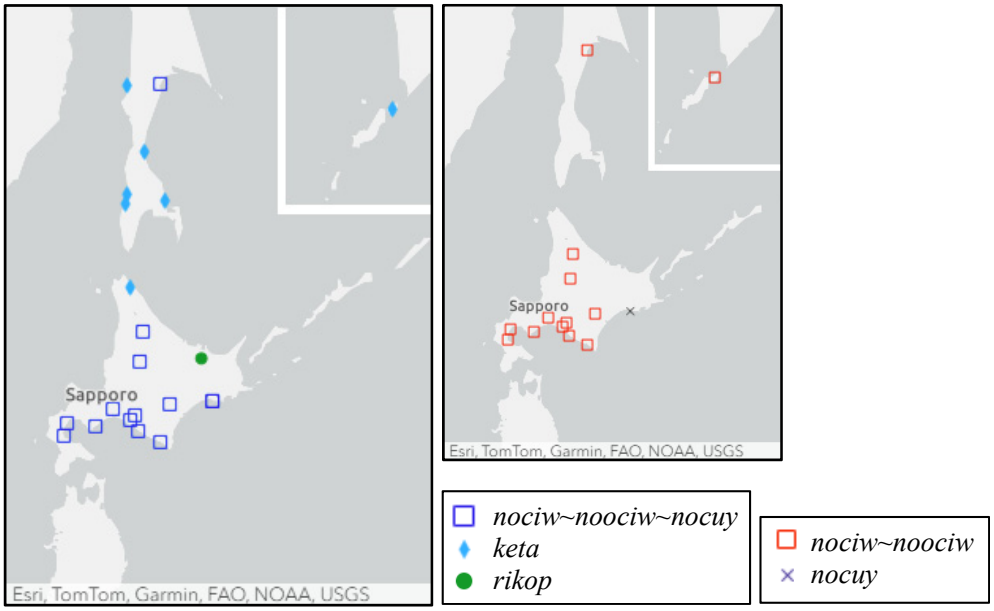


Figure 7: Maps of the lexical forms of 'star' (No. 74)

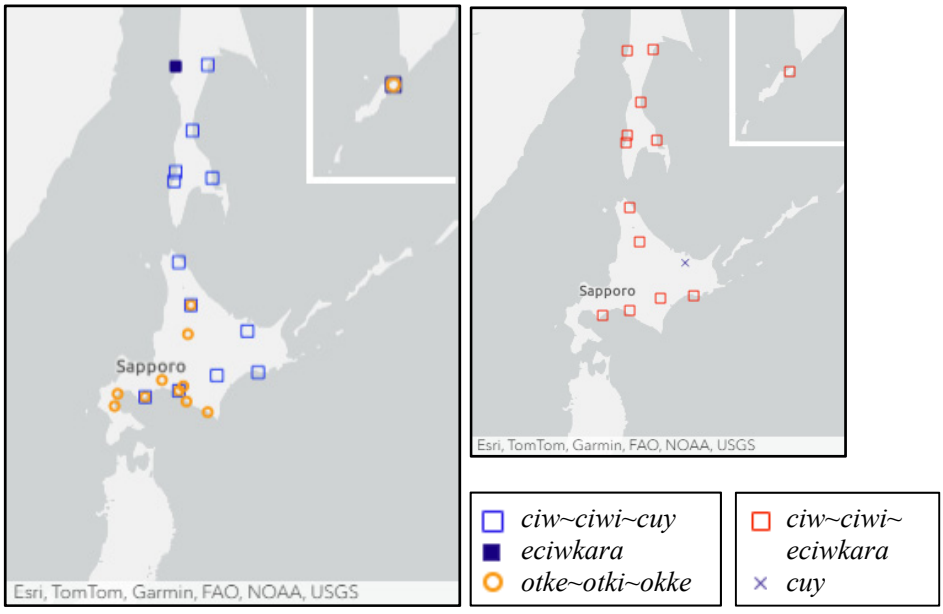


Figure 8: Maps of the lexical forms of 'pierce (stab)' (No. 179)

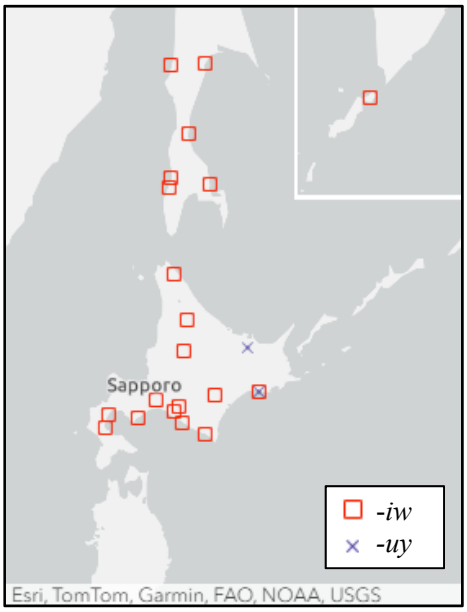


Figure 9: Regularity map of the semivowel (glide) type, *-iw*:*-uy*

3.2. Regularity of the consonant (C_1C_2:C_2C_2:hC_2) type

The lexical forms of 'pierce (stab)' (No. 179) also contain the consonant (C_1C_2:C_2C_2:hC_2) regularity, which is related to the other basic items, such as those for 'rain' (No. 76), as in (5).

(5) The regularity of the consonant (C_1C_2:C_2C_2:hC_2) type
(a) Lexical forms of 'rain' (No. 76): *apto*:*atto*:*ahto*
(b) Lexical forms of 'pierce (stab)' (No. 179): *otke~otki*:*okke*
(c) Lexical forms of 'lie' (No. 67): *hotke*:*hokke*

Figure 10 shows the maps of the lexical forms of 'rain' (No. 76). The lexical form *weni* in the westernmost Hokkaido dialects, Yakumo and Oshamambe, seems to be derived from the word *wen* for 'bad' (cf. Fukazawa 2021: 289). The lexical form *sirun* in Shumushu corresponds to *sirwen* 'bad weather' in the Hokkaido dialects. In the eastern Hokkaido dialects, the lexical forms *ruyanpe* and *ruwanpe* are used for 'rain,' while in the western Hokkaido dialects, they mean 'storm,' and the lexical form {apto} is used for 'rain.' The lexical form of C_1C_2 or C_2C_2, *apto* or *atto*, is also expected to exist in the eastern Hokkaido dialects but cannot be found.

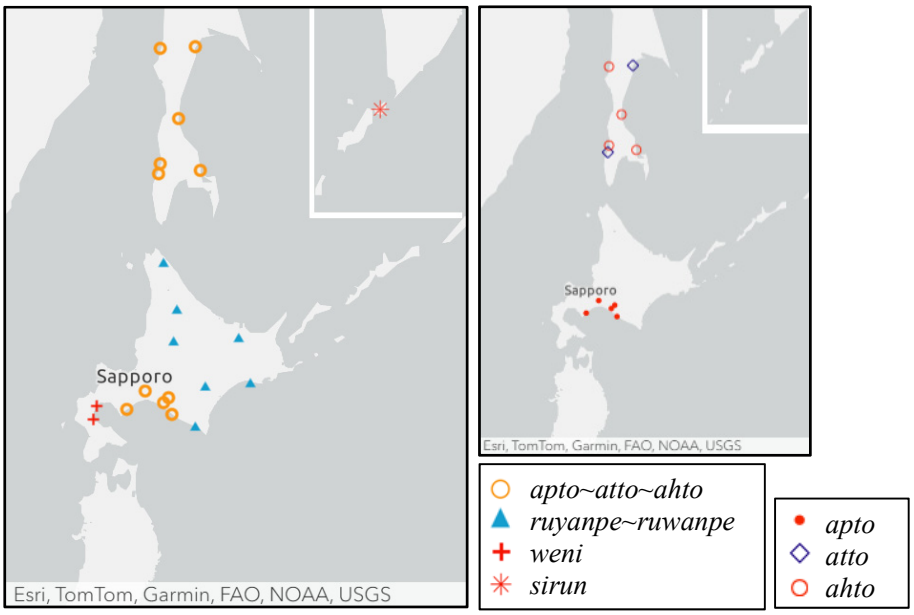


Figure 10: Maps of the lexical forms of 'rain' (No. 76)

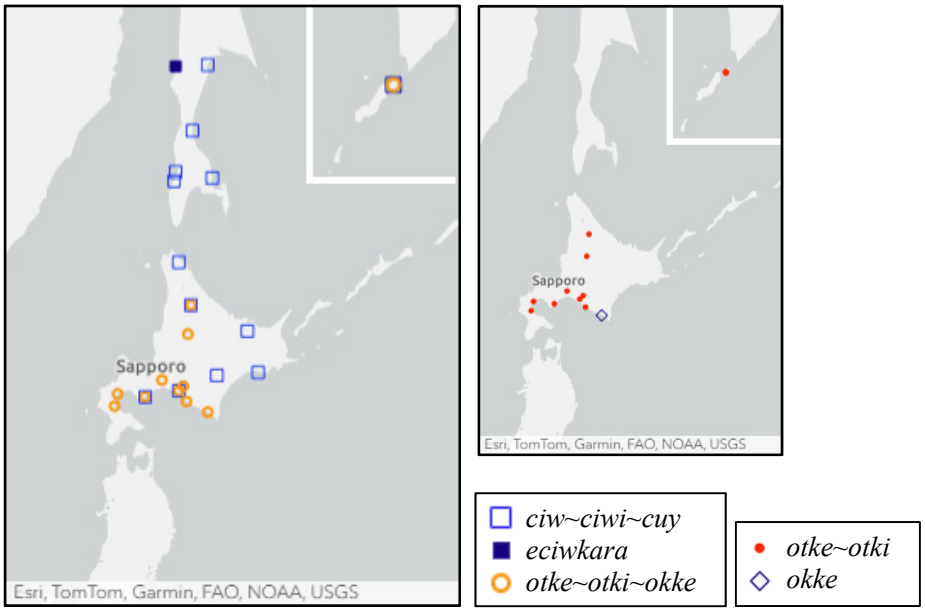


Figure 11: Maps of the lexical forms of 'pierce (stab)' (No. 179)

Figure 11 shows the maps of the lexical forms of 'pierce (stab)' (No. 179), one of which was shown in Figure 8. The C_1C_2 forms, *otke* and *otki*, are found in western Hokkaido and northern Kuril Islands. The C_2C_2 form *okke* is found in the southernmost Hokkaido dialect, Samani.

Figure 12 shows the maps of the lexical forms of 'lie' (No. 67), which also belongs to the consonant regularity, C_1C_2:C_2C_2:hC_2. The C_1C_2 and C_2C_2 forms of {hotke}, *hotke* and *hokke*, are widely distributed in the Hokkaido dialects. The *hokke* form can be found in northeastern Hokkaido dialects, including Asahikawa, Obihiro, Kushiro, and Bihoro. Many dialects use both the forms of various actions. The lexical forms of {hotke} mean 'to lie down,' while the forms of {situri} are used to mean 'to stretch out.' According to Hattori (1964), in a Sakhalin dialect, Raichishika, the hC_2 form of {hotke}, *hohke*, means 'to go to bed and lie down.'

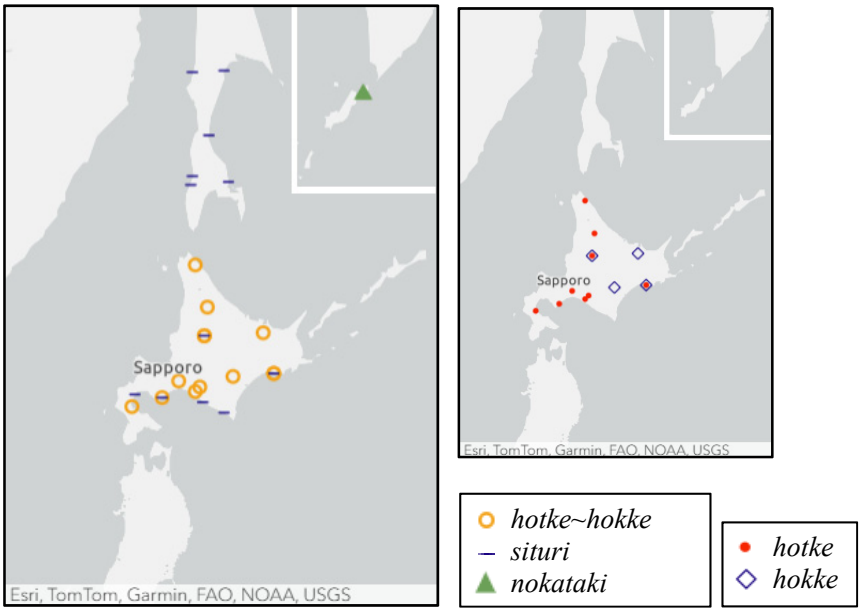


Figure 12: Maps of the lexical forms of 'lie' (No. 67)

Figure 13 shows the regularity map derived from the lexical maps shown in Figures 10, 11, and 12. The C_1C_2 form is found in the Hokkaido and northern Kuril dialects, C_2C_2 form is found in the northeastern Hokkaido and a few Sakhalin dialects, and the hC_2 form is found only in Sakhalin dialects. These dialectal features originate from at least three basic vocabulary items. They compensate for each other.

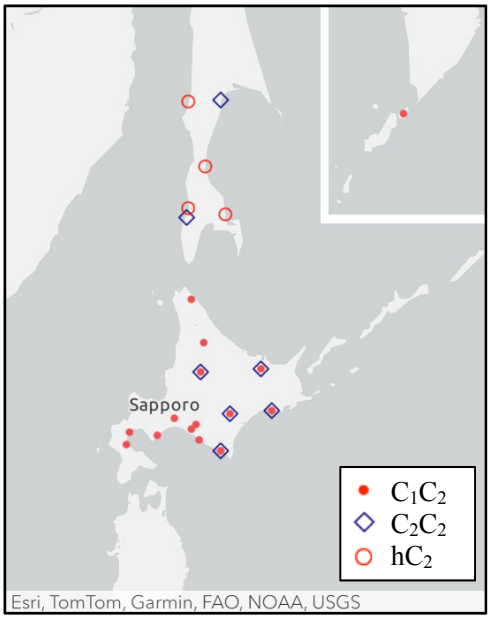


Figure 13: Map of the the consonant regularity (C_1C_2:C_2C_2:hC_2) type

4. Relational map among basic vocabulary items

This section will attempt to draw a relational map by combining the different basic lexical maps with a regularity map. Regularity maps such as those in Figures 9 and 13 exclude lexical information. However, they are quantitatively layered maps, also known as "phonological maps." Lexical maps trace the histories of individual vocabulary items.

This study proposes that the necessary information can be easily obtained from each type of map, although this sometimes requires the exclusion of unnecessary details. Relational maps, or map relations, fill in the missing information from both regularity and lexical maps. For example, when creating a relational map with Figures 7, 8, and 9, the regularity map of the semi-vowel (glide) type, *-iw*:*-uy* (Figure 9), is placed in the center and combined with the lexical maps of Figures 7 and 8. The relational map is shown in Figure 14.

Similarly, a relationship map (Figure 15) can be created by placing the C_1C_2:C_2C_2:hC_2 consonant regularity map (Figure 13) in the center and combining it with the lexical maps shown in Figures 10, 11, and 12. As Figures 8 and 11 show, the

lexical maps of 'pierce (stab)' are related to the two regularity maps. One lexical map can be associated with several regularity maps, thus constituting another relational map.

As noted above, a relational map can show the lexical form that would be inferred if it had not been recorded. For example, the lexical form *hohke* for 'lie' was not recorded in Hattori & Chiri (1960) and Asai (1974), but Hattori (1964) mentioned it in the Raichishika dialect of Sakhalin. Given the relational map between consonant (C_1C_2:C_2C_2:hC_2) regularity and lexical maps, the lack of data would be compensated for. Since the Hokkaido dialects have the *hotke* (C_1C_2) and *hokke* (C_2C_2) forms, one can infer that the Sakhalin dialects have the hC_2 form, *hohke*. Considering which regularity map each lexical form is associated with is useful to trace the history of a lexical item.

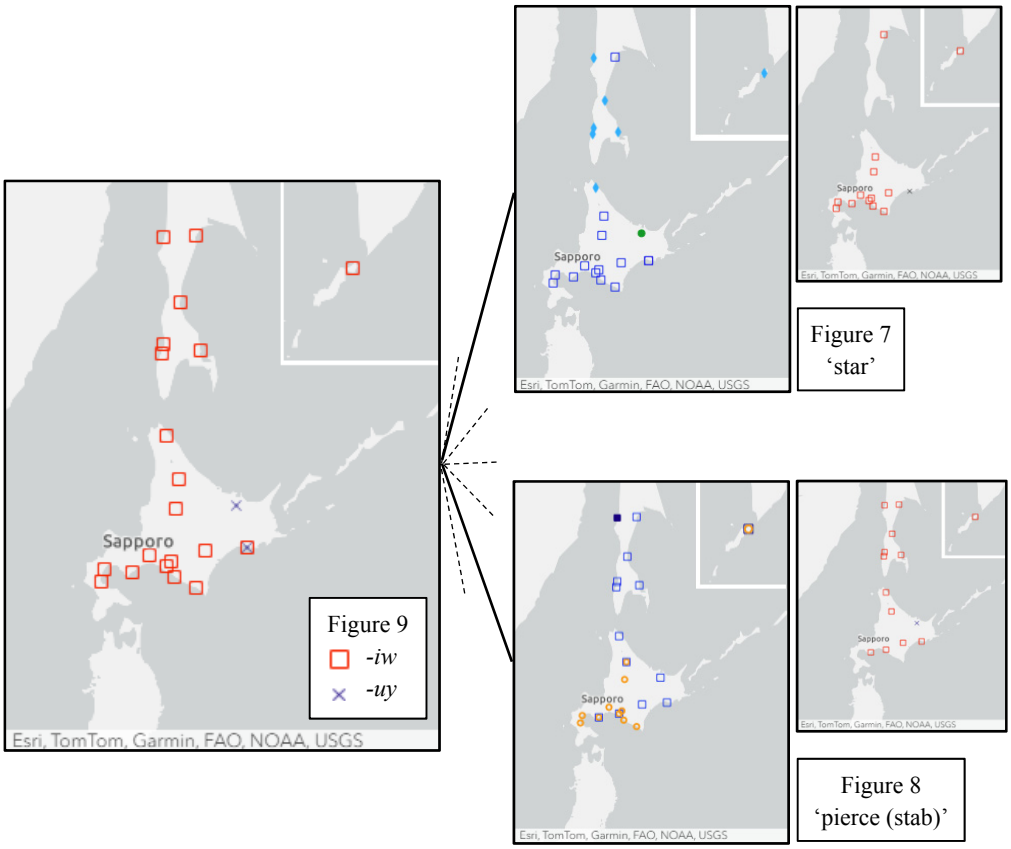


Figure 14: Relational Map of the semivowel (glide) type, *-iw*:*-uy*

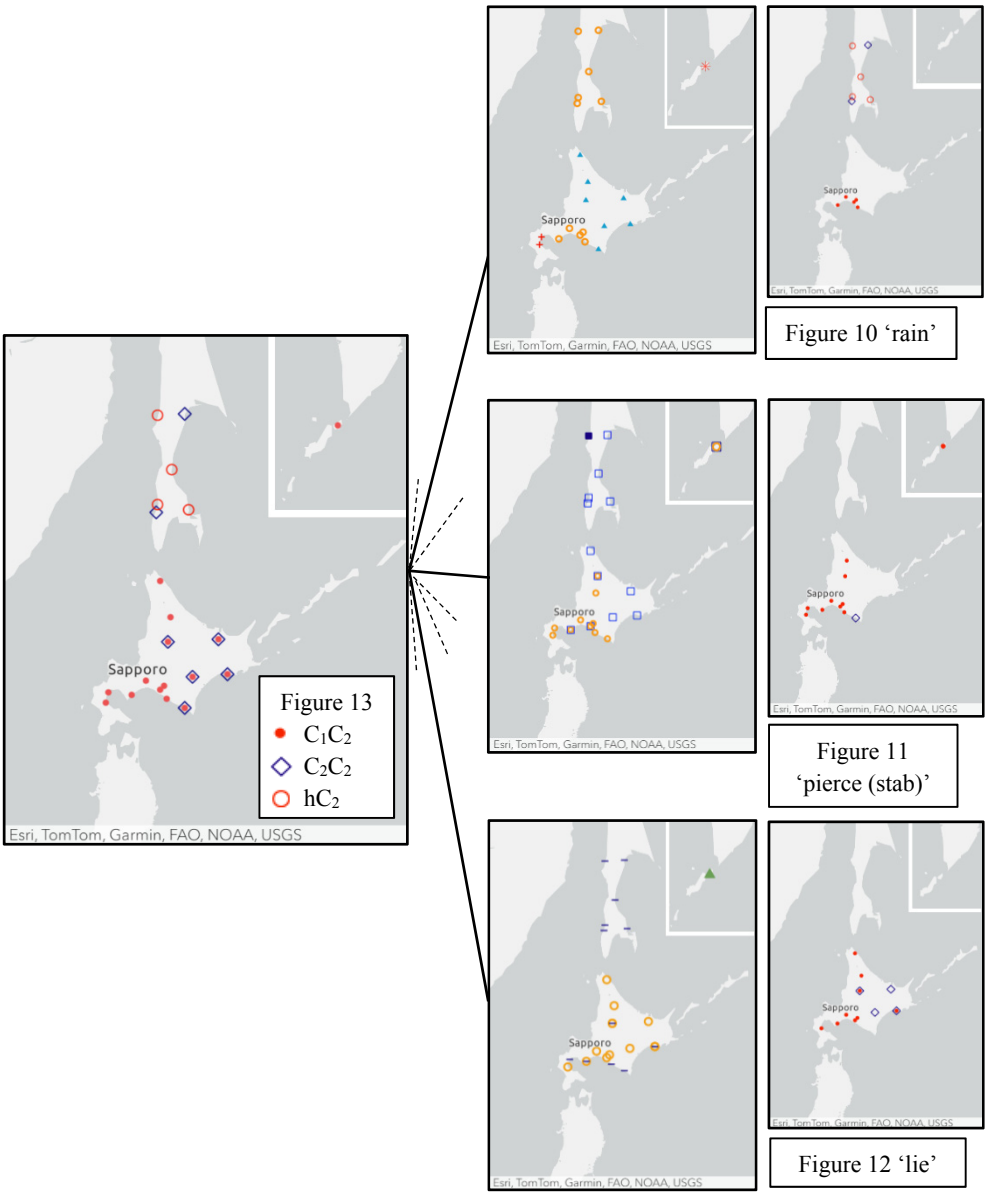


Figure 15: Relational Map of the consonant (C_1C_2:C_2C_2:hC_2) type

5. Concluding remarks

This study proposes the creation of relational maps of basic vocabulary items in Ainu. Fukazawa's (2018) preliminary study started with the first 100 basic vocabulary maps but found that they often had a monotonous distribution. One solution to this issue, as proposed in this study, is to distinguish between regularity and lexical maps and create relational maps. The regularity map used in this study is similar to the "phonological map" in structural geolinguistics. Future studies will also aim to create grammatical and morphological maps as regularity maps, such as those showing how to combine first- and second-person affixes with transitive verbs.

This study also proposes drawing a connection between regularity and lexical maps. A lack of information on a map is acceptable because if the symbols are appropriately assigned and the relations between the maps are distinctly created, they can be more easily interpreted. Relational maps demonstrate the relationships between regularity and lexical maps, and can provide additional information if word recordings are lacking in the linguistic survey.

The histories of vocabulary items in the regularity and lexical maps differ significantly. Lexical maps frequently show more complex distributions because words trace their own histories, including borrowings. Regularity maps may show dialectal categories extracted from quantitative vocabulary items. Relational maps possess the advantages of both types, and are useful for tracing word histories and identifying trends in linguistic regularities.

Abbreviations

1: 1st person, A: transitive subject, C: consonant, EXCL: exclusive (personal affix). PL: plural, V: vowel.

References

Asai, Toru (1974) Classification of dialects: Cluster analysis of Ainu dialects. *Bulletin of the Institute for the Study of North Eurasian Cultures, Hokkaido University* 8: 45–136.

Bugaeva, Anna (2022) Introduction. In: Anna Bugaeva (ed.) *Handbook of the Ainu language*, 1–19. Berlin: Mouton De Gruyter. doi: https://doi.org/10.1515/9781501502859-001

Chiri, Mashiho (1976 [1962]) *Bunrui Ainugo jiten, Dōbutsuhen*『分類アイヌ語辞典　動物篇』*[Classified Ainu dictionary]. Reprinted: Chiri Mashiho chosakushū bekkan* 『知里真

志保著作集　別巻1』[Collected works of Chiri Mashiho Separate volume 1]. Tokyo: Heibonsha.

Fukazawa, Mika [深澤美香] (2017) *Kagake monjo ni okeru Ainugo no bunkengakuteki kenkyū* 『加賀家文書におけるアイヌ語の文献学的研究』[Philological study of the Ainu in Kaga family's archives]. Doctoral dissertation. Chiba: Chiba University. doi: https://doi.org/10.20776/103627 ; URL: https://opac.ll.chiba-u.jp/da/curator/103627/

Fukazawa, Mika (2018) Geographical distribution patterns of basic Ainu vocabulary in Hattori and Chiri (1960). In. Suzuki, Hiroyuki & Mitsuaki, Endo (eds.) *Studies in Asian Geolinguistics, Monograph Series No. 4: Papers from the Fourth International Conference on Asian Geolinguistics*, 91–103. URL: https://publication.aa-ken.jp/papers_4IC_Asian_geolinguistics_2018.pdf

Fukazawa, Mika (2021) 'It rains' in Ainu. In. Endo, Mitsuaki, Makoto Minegishi, Satoko Shirai, Hiroshi Suzuki, and Keita Kurabe (eds.) *Linguistic Atlas of Asia*. 289–290. Tokyo: Hituzi Syobo.

Fukazawa, Mika [深澤美香] (2025) Kokuritsu Ainuminzoku hakubutsukan shozō Chiri Mashiho kinyu no Ainugo kisogoi chosahyō「〔資料紹介〕国立アイヌ民族博物館所蔵知里真志保記入のアイヌ語基礎語彙調査表」[Preliminary survey of the basic Ainu vocabulary: Chiri Mashiho's materials from the National Ainu Museum]. *National Ainu Museum Journal* 3: 136–158. doi: https://doi.org/10.57545/namjournal.2024-06

Fukazawa, Mika [深澤美香] & Yōhei Ono [小野洋平] (2025) Kizogoi niyoru hōgen bunrui no shomondai: Asai (1974) no Ainugo hōgen dēta wo saidaigen katsuyōsuru tameni「基礎語彙による方言分類の諸問題：Asai (1974) のアイヌ語方言データを最大限活用するために」[Challenges of dialect classification in extracting maximum information from the basic vocabulary of Ainu in Asai (1974)]. *Northern Language Studies* 15: 141–163. doi: https://doi.org/10.14943/112975

Hattori, Shirō [服部四郎] (1954) "Gengonendaigaku" sunawachi "goitōkeigaku" no hōhō nitsuite「「言語年代學」即ち「語彙統計學」の方法について：日本祖語の年代」[On the method of Glottochronology and the time-depth of Proto-Japanese]. *Gengo Kenkyu* [Journal of the Linguistic Society of Japan] 26, 27: 29–37. doi: https://doi.org/10.11435/gengo1939.1954.26-27_29

Hattori, Shirō [服部四郎] and Mashiho Chiri [知里真志保] (1960) Ainugo shohōgen no kisogoi tōkeigakuteki kenkyū「アイヌ語諸方言の基礎語彙統計学的研究」[Alexicostatistic study on the Ainu dialects]. *Japanese Journal of Ethnology* 24(4): 307-342. doi: https://doi.org/10.14890/minkennewseries.24.4_307

Murayama, Shichirō [村山七郎] (1971) *Kitachishima Ainugo*『北千島アイヌ語』[Northern Kuril Ainu]. Tokyo: Yoshikawa Kobunkan.

Nakagawa, Hiroshi & Mika Fukazawa (2022) Hokkaido Ainu dialects: Towards aclassification of Ainu dialects. In. Anna Bugaeva (ed.) *Handbook of the Ainu language*, 253–328. Berlin: Mouton De Gruyter. doi: https://doi.org/10.1515/9781501502859-009

Ono, Yōhei [小野洋平] and Mika Fukazawa [深澤美香] (2023) Ainugo shohōgen no gokei no ruiji ni kansuru kiso dētano fukugen「アイヌ語諸方言の語形の類似に関する基礎データの復元：論文に書ききれなかった研究者の判断・思考に迫る」[Reconstruction of

original data on similarity between word forms in Ainu dialects: Approaching the researcher's unwritten judgment and thoughts]. *Northern Language Studies* 13: 213–246. doi: https://doi.org/10.14943/106657

Ono, Yōhei [小野洋平] and Mika Fukazawa [深澤美香] (2024) Hikaku Fukanōdatta Ainugo hōgen bunrui: Tōkeiteki hōgenbunrui wo ruijihandan no ten kara saikōsuru「比較不可能だったアイヌ語方言分類：統計的方言分類を類似判断の点から再考する」[(Un)comparable classifications of Ainu dialects: Reconsidering statistical dialect classification from the similarity judgments]. *Journal of Ainu and Indigenous Studies* 4: 93–126. doi: https://doi.org/10.14943/jais.4.093

Ono, Yōhei & Mika Fukazawa (2025) Statistical approaches to the quantification of regularity in languages and dialects: An exercise in Ainu dialects of Asai (1974). *Northern Language Studies* 15: 179–202. doi: https://doi.org/10.14943/112977

Pinart Alphonse Louis (1872) Vocabulary in Russian and Ainu dialect of the Paramushir and Limushir Islands in the Kuriles. Transcribed and Translated by: Asai, Tōru (1974) Classification of dialects: Cluster analysis of Ainu dialects. *Bulletin of the Institute for the Study of North Eurasian Cultures, Hokkaido University* 8: 101–136.

Swadesh, Morris (1955) Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, 21(2): 121‑137.

Tamura, Suzuko (2000) *The Ainu language*. Tokyo: Sanseido.

Torii, Ryūzō [鳥居龍藏] (1903) *Chishima Ainu* 『千島アイヌ』[Kuril Ainu]. Tokyo: Yoshikawa Kobunkan.

The distribution of /l/ and /n/ variants in the Red River Delta, Vietnam

Trịnh Cẩm Lan (USSH, VNU Hanoi)

Abstract: The phenomenon in which /l/ has two phonetic variants [l] and [n], and /n/ has two variants [n] and [l], results in a state of mutual confusion between these two consonants. This phenomenon is particularly widespread in the Red River Delta region of Northern Vietnam. Approaching this issue from the perspective of geo-linguistics and employing methods such as field surveys and linguistic mapping, this paper aims to map and locate the distribution of variations of /l/ and /n/ in the Red River Delta, Vietnam, and interpret that distribution. The findings of this study include: (1) Identifying the variation models of /l/ and /n/ in the Red River Delta; (2) Mapping the geographical distribution of those variation models; and (3) Explaining the distribution of variations of /l/ and /n/ in space using theoretical models of geo-linguistics.
Key words: variants, /l/ - /n/ confusion, geographical distribution, Red River Delta

1. Introduction

/l/ and /n/ are initials that shared place of articulation in Vietnamese. However, /n/ is a voiced nasal stop, while /l/ is a voiced lateral fricative. /n/ has two phonetic variants: [n] and [l]; similarly, /l/ also has two phonetic variants: [l] and [n]. This is a fairly common pronunciation phenomenon in the rural areas of the Red River Delta (Hoàng Thị Châu 2004; Vũ Kim Bảng 2005; Phạm Văn Hảo 2024). This phenomenon is referred to as *the pronunciation shift of /l/ to [n] and /n/ to [l]* (Vũ Kim Bảng 2005), *the /l/–/n/ confusion* (Hoàng Thị Châu 2004: 137; Hà Quang Năng 2007; Phạm Văn Hảo 2024), *non-standard pronunciation of /l/ and /n/*, or *the "mispronunciation of /l/–/n/"* (Trần Thị Thìn 1979). This represents *a reduction or complete loss of the phonological distinction between the two initial consonants: the voiced lateral fricative /l/ and the voiced nasal stop /n/*" (Đoàn Thiện Thuật 2016).

In the Red River Delta, the confusion occurs in various patterns. In some areas, it is unidirectional (one-way), and in others, it is bidirectional (two-way). Words that originally begin with /n/ can be pronounced with [l] (e.g., nóng > lóng, nồi > lồi).

TRỊNH, Cẩm Lan. 2025. The distribution of /l/ and /n/ variants in the Red River Delta, Vietnam. In Nor Hashimah Jalaluddin, Hiroyuki Suzuki and Mitsuaki Endo (eds.) *Proceedings of the sixth International Conference of Asian Geolinguistics*, 81–98. doi: https://doi.org/10.5281/zenodo.17204612

Conversely, words that originally begin with /l/ may be pronounced with [n] (e.g., lá > ná, lúa > núa). Among these two directions, the shift from /l/ to [n] is considered more common (Trần Thị Thìn 1979; Bùi Đăng Bình 2002; Vũ Kim Bảng 2005).

As a widespread sociolinguistic phenomenon, the /l/–/n/ confusion began to attract scholarly attention in the late 1970s, though interest has remained relatively limited. Some early studies focused on describing the phenomenon and identifying geographic areas where the /l/–/n/ variants appeared, offering preliminary explanations for its origins (Bùi Đăng Bình 2002; Vũ Kim Bảng 2005). Others sought to analyze the sociolinguistic factors influencing this variation (Nguyễn Thị Thanh Bình 2001; Trịnh Cẩm Lan 2017). However, specialists have not yet reached a consensus on the nature of the phenomenon. Some view it as a case of non-standard pronunciation or "mispronunciation" (as a lisp) caused by physiological factors in articulation (Trần Thị Thìn 1979). Others argue that the /l/–/n/ confusion is tied to the evolution of the Vietnamese phonological system - specifically, the merging of lateral and apical initials - and represents a progressive "simplification of articulation" (Phạm V.H., 2024). Still, others consider it merely a "dialectal phenomenon" that has become widespread in Northern Vietnam (Hoàng Thị Châu 2004: 137; Vũ Kim Bảng 2005). Regarding its origin, Hoàng Thị Châu suggests that "*many Chinese dialects in the southwest also exhibit /l/–/n/ confusion and that this phenomenon has spread like a phonological wave from the northwest to the southeast*" (Hoàng Thị Châu 2004: 137). However, she provides no evidence to support this claim. As a result, there remains no unified conclusion about the true nature, geographic distribution, or potential development of this phenomenon.

Although the /l/–/n/ confusion is a dialectal feature, it is treated very differently from other dialectal phenomena (such as the merger of /ʈ/ and /c/, /ʂ/ and /s/, /ʐ/ and /z/ in Northern dialects; the merger of *hỏi* and *ngã* tones in Thanh Hóa or Southern dialects; and many others). It shows signs of being stigmatized, perceived as socially "marked," and often associated with rural populations or individuals with lower education levels, unlike other dialectal features. Recently, this phenomenon has not only been present in Vietnamese pronunciation but has also appeared in Vietnamese speakers' pronunciation of English, creating sometimes humorous or awkward situations. For example: good afternoon > [gʊd ɑːf.təˈluːn]; hello > [həˈnoʊ]; No, I am not > [loʊ aɪ æm lɑːt]; International > [in təˈlæ sən lo]. This reality has drawn significant attention from society, especially within the field of education. One clear example is the active involvement of educational administrators in efforts to "correct mispronunciation" (to correct lisps) on a large scale for both teachers and students in many provinces of the

Red River Delta, particularly in major cities like Hanoi and Hai Phong (Phạm Thịnh 2011).

From the perspective of this paper, the fact that /l/ and /n/ exhibit multiple phonetic variants and become confused with each other should only be considered "mispronunciation" (or a lisp) when it occurs sporadically among a few individuals and does not form a trend. In such cases, the causes are usually: (1) The individual is a young child whose articulatory system is not yet fully developed; (2) The person is an adult with physiological disorders affecting the speech apparatus (e.g., a short or overly long tongue, cleft lip, cleft palate, etc.); (3) The individual has developmental disorders (e.g., intellectual disability, autism), which affect the articulation process. However, when this phenomenon occurs widely across a population and forms a recognizable trend, it cannot be classified as "mispronunciation" (or a lisp). Rather, it must be recognized as a regional pronunciation habit, where speakers do not have any articulation issues and can, in fact, pronounce both /l/ and /n/ correctly. For these reasons, the /l/–/n/ confusion should be viewed solely as a dialectal phenomenon.

By approaching this phenomenon through the perspective of geo-linguistics and employing geolinguistic methods such as field surveys, investigations, and mapping, this paper aims to map and explain the distribution of the variants of /l/ and /n/ across the provinces of the Red River Delta in Northern Vietnam. To accomplish this goal, the paper will address two main objectives:

(1) Survey and map the geographic distribution of /l/ and /n/ variation patterns in the Red River Delta.

(2) Explain the spatial distribution of these variants using theoretical models from geo-linguistics and dialectology.

2. Theoretical foundations

2.1. Phonological foundation

The phonological foundation for the transformation of /l/ > [n] and /n/ > [l] will be examined based on distinctive phonological features related to the place and manner of articulation, airflow, sonority, and voicing in the articulation of /l/ and /n/. Such distinctive phonological features show that /l/ and /n/ share several articulatory similarities.

Homorganicity (Shared Place of Articulation): As previously mentioned, both /l/ and /n/ are alveolar initials, meaning they have the same place of articulation. This homorganic relationship is a key factor in facilitating assimilation. Theoretically,

assimilation tends to occur when a sound changes to another that shares some features, such as place or manner of articulation (Ladefoged & Johnson 2015: 111).

Manner of Articulation and Airflow: In addition to sharing the same place of articulation, /l/ and /n/ also exhibit similarities in their manner of articulation and the direction of airflow. /n/ is a nasal stop: during its articulation, the oral cavity is completely closed due to the contact between the tongue tip and the alveolar ridge, and the airflow is redirected through the nasal cavity as the velum (soft palate) lowers. In contrast, /l/ is a lateral approximant: while the tongue tip also contacts the alveolar ridge, the airflow escapes along the sides of the tongue, and the velum is raised. Although there are some differences in airflow direction — oral versus nasal — both consonants have a special configuration in the way airflow travels during their articulation.

The shared tongue tip — alveolar contact between /l/ and /n/ — along with similar mechanisms of articulatory action in the oral cavity, show that even a small adjustment in the velum's position (raised or lowered) or in the closure of the oral cavity (forming or not forming lateral gaps) can lead to a change from one sound to the other. For instance, if the lateral airflow of /l/ is blocked and the velum is lowered, /l/ can become [n]. Conversely, /n/ can become [l] if the side closures of the oral cavity are relaxed and the velum is raised.

Sonority: Both /l/ and /n/ are classified as resonants or approximants, because their production involves vocal fold vibration and relatively unobstructed airflow (Đoàn Thiện Thuật 2016: 32). As such, they share fundamental characteristics with vowel articulation and exhibit higher sonority than other stops or fricatives.

Voicing: /l/ and /n/ are consistently categorized as voiced consonants in all languages. This is expected, as the nature of being sonorants implies that the proportion of voiced sound in their articulation is significantly higher than the proportion of noise.

Table 1. Distinctive phonological features of /l/ and /n/

Distinctive Features	/l/	/n/
Place of Articulation	Alveolar	Alveolar
Manner of Articulation	Lateral approximant	Nasal stop
Sonority	Sonorant	Sonorant
Voicing	Voiced	Voiced

These phonological similarities contribute to the perceptual similarity between /l/ and /n/. This creates a phonetic pathway for confusion, and ultimately merger, over time as in many dialects at the south of the Yangtze River (Tonghe 2023: 33; Linguistics

Atlas of Chinese Dialects, Phonetics, 057). Such a pathway makes the confusion or merging of the two sounds relatively easy and understandable.

Theoretically, phonetic changes occurring in the history or dialects of a language (when they occur) tend to happen between adjacent phonemes in the system (Hoàng Thị Châu 2004: 100). /l/ and /n/ are such adjacent phonemes, and the gradual erosion of the distinction between them — manifesting in confusion or merger — has a solid phonological foundation.

2.2. Theoretical models for interpreting dialect maps

Language change across geographic regions is not a random process; rather, it is the result of complex processes such as migration, contact, isolation, and various social dynamics. To understand and explain the distributional patterns of linguistic phenomena, linguists have developed a variety of theoretical models. These models not only help describe language variation but also provide analytical frameworks for interpreting dialect maps, thereby illuminating the mechanisms of linguistic diffusion and change (Trịnh, Cẩm Lan & Trần Thị Hồng Hạnh 2025).

The spread of linguistic features across space — also known as spatial diffusion — is a central topic in geolinguistics. Theoretical models help explain different types of diffusion, from contiguous diffusion to leapfrogging between urban centers. The most commonly used spatial diffusion models in dialectology include the Wave Model (Fasold & Wolfram 1974: 76), Gravity Model (Trudgill 1974), Social Network Model (Milroy 2002), Diffusion Models, and Agent-Based Models (ABM) (Bailey, Wikle & Sand 1996), among others.

Interpreting the geographic distribution of linguistic variants is a complex task requiring a combination of field data collection, cartographic mapping, and the application of appropriate theoretical frameworks. In this study, based on the gathered data, the Wave Model and the Social Network Model will be applied to interpret the dialect maps showing the distribution of /l/ and /n/ variants in the Red River Delta.

2.2.1. Wave Model

Proposed by Johannes Schmidt in the 19th century, the Wave Model is one of the most foundational theoretical models. It conceptualizes the spread of linguistic change as ripples radiating from a central point, like waves formed when a stone is thrown into water (Chambers & Trudgill 1998).

The core principle of this model is that linguistic changes originate from a center — typically a culturally or economically prestigious urban hub — and diffuse outward

to surrounding areas in a uniform and continuous manner. The degree of influence diminishes with increasing distance from the center.

On dialect maps, the Wave Model is represented by isoglosses — lines that demarcate linguistic boundaries — shaped like concentric circles or curves radiating from a point. Areas close to the center exhibit linguistic changes more clearly and strongly, while more distant areas show weaker influence. This model effectively explains transitional dialect regions, where linguistic features change gradually over space rather than along abrupt boundaries.

Later, in 1973, the model was further developed by Charler Bailey to better accommodate findings in modern sociolinguistics (Bailey 1973). In Bailey's version, wave models no longer appear strictly as ideal concentric circles. Instead, they account for more complex, directional diffusion patterns, especially when natural or social barriers (e.g., mountain ranges, rivers, political upheavals, or borders) interrupt the spread at certain points. These obstacles can cause the wave to propagate in a single direction. This evolution made Bailey's wave model more dynamic and explanatory (Trịnh Cẩm Lan 2008).

2.2.2. Social Network Model

Developed by Lesley Milroy in her study of Belfast, the Social Network Model shifts the focus from macro-level factors such as population and geographic distance to micro-level interactions among individuals within a community (Milroy 2002).

The key principle of this model is that the spread and maintenance of linguistic variants depend on the social network structure of speakers. There are two main types of networks:

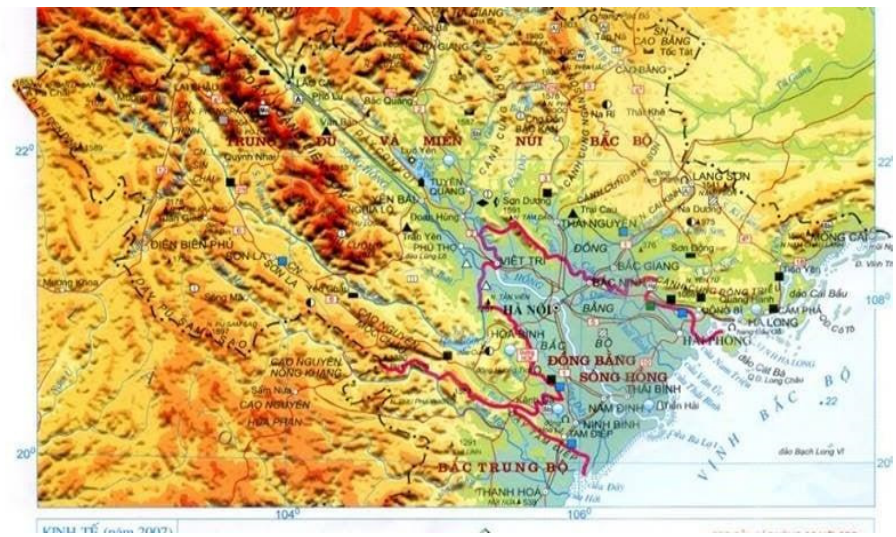
> **A closed network** is one with high density, where members all know each other and have few external ties. Such networks tend to reinforce internal linguistic norms and resist external influence.
>
> **An open network**, by contrast, has lower density and more external connections. These networks serve as effective channels for the introduction and spread of new linguistic variants (Milroy 2002).

On dialect maps, this model helps explain the presence of dialectal "islands" or highly conservative linguistic areas even when they are located near innovative urban centers. For example, a rural community with a tightly-knit social structure (based on dense kinship and neighborhood relations) may preserve traditional dialect features despite being near a major city. Conversely, a social group with an open network — such as university students from various regions — forms an ideal environment for the emergence and diffusion of new linguistic variants, such as internet language or student slang.

3. Data and Research Methods

The subjects of this survey are the phonetic variants of the initials /l/ and /n/ across 11 provinces in the Red River Delta. The investigators, who also served as data providers, were students from the University of Social Sciences and Humanities (majoring in Linguistics) and the University of Education (majoring in Literature Education), both part of Vietnam National University, Hanoi. These students were trained as field investigators to observe the pronunciation habits of residents in the localities where they were born and raised, and to record the findings in a questionnaire.

The geographical scope of the survey includes provinces and cities in the Red River Delta, as well as a few neighboring provinces for comparative purposes. In each province/city, data was collected — where possible — from each district, town, or city under provincial administration, with an average of one sampling point per administrative unit. The total number of surveyed points was 170, covering 14 provinces (including 11 Red River Delta provinces and 3 neighboring provinces — Hòa Bình, Phú Thọ, and Bắc Giang — for comparative analysis).



Map 1: The Red River Delta
(Source: Vietnam Law Library, https://thuvienphapluat.vn/)

In addition to the data provided by student investigators, the author also referred to the results of the *Vietnamese Comprehensive Linguistic Survey Program 1998–2000, Audio Field Recordings*, currently archived at the Institute of Linguistics (Vietnam), and held discussions with the original investigators of the program to verify and refine the research findings. For certain localities where the collected data was incomplete or potentially biased, the researcher personally visited the area to conduct direct observation and data collection.

Data Analysis Methods: The collected data — including linguistic data (phonetic variants of /l/ and /n/ in each locality) and geographical data (longitude and latitude of each survey point) — was encoded and entered into Excel. The dialect maps were then generated using ArcGIS Online, a web-based geographic information system developed by Esri, following the guidelines of Mitsuaki Endo (2016) and Mika Fukazawa (2017).

4. Research Findings

4.1. Patterns of /l/ and /n/ Variation in the Red River Delta

In the Red River Delta, the initial /n/ exhibits two variants. The first is [n], an alveolar nasal stop, which is the standard, officially recognized pronunciation of /n/. The second variant is [l], an alveolar lateral approximant. Similarly, the consonant /l/ also has two variants: the standard [l], and a variant pronounced as [n].

Survey results show that the patterns of /l/–/n/ confusion are quite complex. The realizations of /l/ and /n/ across the surveyed regions present several key features:

* In terms of frequency: For /l/, data indicates that in 84 out of 170 localities, /l/ has two variants, while in 86 out of 170 localities, /l/ has only one variant. For /n/, 61 out of 170 localities show two variants of /n/, whereas 109 localities have only one variant. Thus, overall, the shift /l/ > [n] appears to be more prominent than /n/ > [l].

* In terms of regional variation patterns, the data identifies several models of variation:

- **Model A:** In some localities, both /l/ and /n/ have only one unique variant each, maintaining the phonological contrast between them. There is no confusion.

- **Model B:** /n/ has two variants ([n] and [l]), while /l/ remains unchanged with only the [l] variant. This is a unidirectional confusion model: /n/ > [l], with /l/ remaining stable.

- Model C: /l/ has two variants ([l] and [n]), while /n/ has only one variant ([n]). This is also a unidirectional confusion model, but in the opposite direction: /l/ > [n], with /n/ remaining stable.

- Model D: Both /l/ and /n/ exhibit two variants. This represents **bidirectional confusion**: /n/ > [l] and /l/ > [n].

Table 2. Regional Models of /l/–/n/ Confusion

Model	/n/ Variant(s)	/l/ Variant(s)	Description	No. of Localities
A	[n]	[l]	No confusion	57
B	[n], /n/ > [l]	[l]	Unidirectional confusion: /n/ > [l]	32
C	[n]	[l], /l/ > [n]	Unidirectional confusion: /l/ > [n]	52
D	[n], /n/ > [l]	[l], /l/ > [n]	Bidirectional confusion: /n/ > [l], /l/ > [n]	29

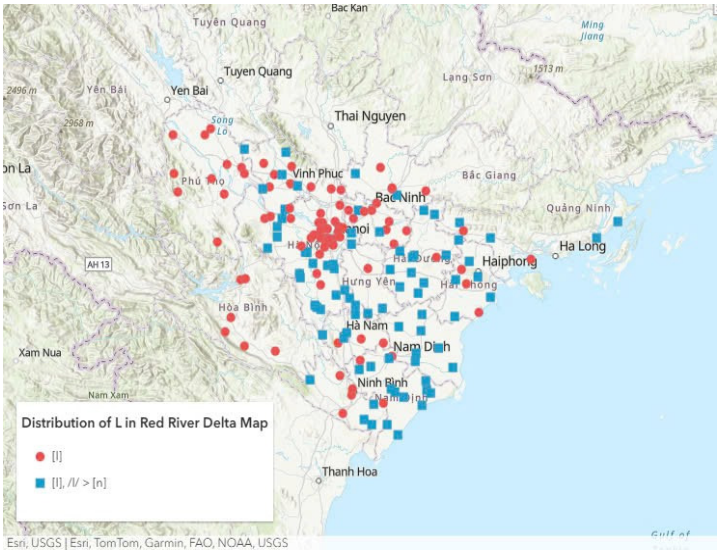
In terms of degree of confusion, the data shows that there is no case of complete confusion (i.e., "everyone in the village mispronounces") for both phonemes in any locality. In all surveyed location, the confusion appears in a patchy or leopard-skin pattern — that is, a portion of the population maintains clear /l/–/n/ distinction, while others exhibit confusion in either one direction (unidirectional) or both (bidirectional).

At the individual level, there are: (1) Speakers who clearly distinguish /l/ and /n/ (no confusion); (2) Speakers who fully confuse them in one direction (e.g., consistently pronouncing both /l/ and /n/ as either [l] or [n]); (3) Speakers who show partial or inconsistent confusion, where some words are affected while others are not.

Preliminary observations suggest that confusion tends to concentrate among specific social groups: the elderly, young children, farmers, individuals with lower levels of education, or those with limited social interaction. Similar patterns have also been observed in dialect communities of Southern China (Ng, Choi Lee Charlie, 2017; Yuyan Zhou, Ying Wu, 2019).

4.2. Distribution of /l/ and /n/ Variants in the Red River Delta

4.2.1. Distribution of /l/ Variants



Map 2. Distribution of /l/ Variants

Regions with a single variant [l]: 86 localities;

Regions with two variants [l] and /l/ > [n]: 84 localities)

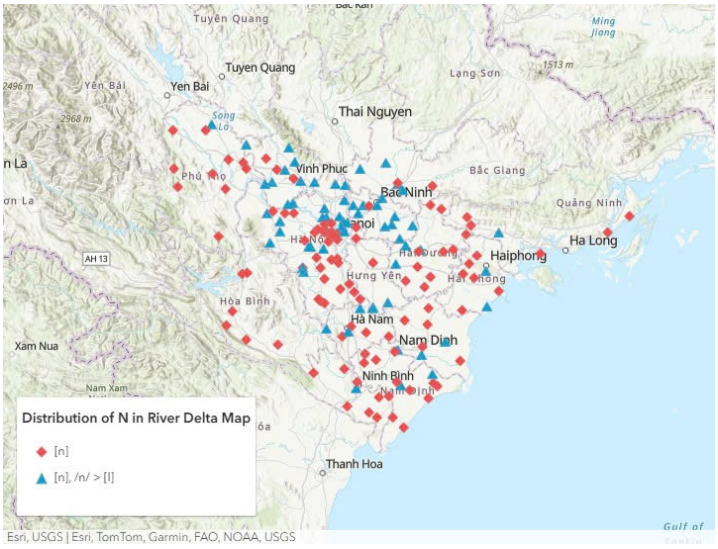
Map 2 shows that regions with only one variant of /l/ tend to be concentrated in the northwestern part of the area, while regions with two variants are mostly found in the southeastern area, particularly in the northern coastal provinces. Survey data reveals several regionally or provincially distinctive centers:

* Regions mainly exhibiting two variants (i.e., regions with confusion): southern districts of Hanoi; provinces of Hải Dương, Hưng Yên, Hải Phòng, Hà Nam, Nam Định, Thái Bình, and Ninh Bình.

* Regions mainly exhibiting a single variant (i.e., no or minimal confusion): urban districts of Hanoi, northern and northwestern districts of Hanoi, Bắc Ninh, and parts of Vĩnh Phúc.

* Bordering provinces of the Red River Delta such as Hòa Bình, Phú Thọ, and Bắc Giang exhibit only one variant, meaning the phenomenon /l/ > [n] does not occur there.

4.2.2. Distribution of /n/ Variants



Map 3. Distribution of /n/ Variants

Regions with a single variant [n]: 109 localities;

Regions with two variants [n] and /n/ > [l]: 61 localities.

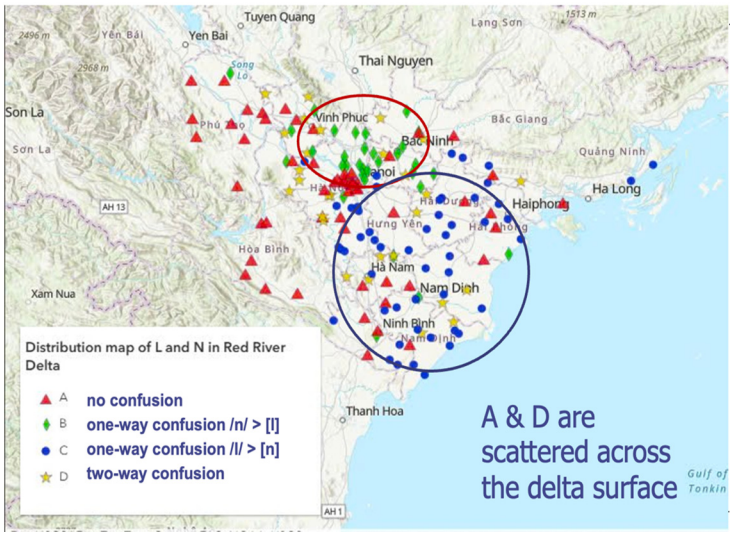
According to Map 3, regions with a single variant of /n/ tend to cluster in the southeastern part of the area, while regions with two variants are more concentrated in the northwest, particularly in provinces bordering hilly terrain, such as Bắc Giang, Hòa Bình, and Phú Thọ. The survey data reveals several regional/provincial centers for /n/ variation:

* Regions with a tendency toward two variants (i.e., with confusion): northwestern districts of Hanoi, Bắc Ninh, and parts of Vĩnh Phúc.

* Regions mainly exhibiting a single variant (i.e., no or minimal confusion): urban districts of Hanoi; southern districts of Hanoi; provinces of Hải Dương, Hưng Yên, Hải Phòng, Nam Định, Thái Bình, and Ninh Bình.

* Bordering provinces such as Hòa Bình, Phú Thọ, and Bắc Giang only exhibit a single variant, indicating no occurrence of the /n/ > [l] phenomenon.

4.2.3. Overall Distribution of /l/ and /n/ Variants



Map 4. Distribution of Variation Patterns by Locality

Map 4 provides an overall picture of /l/–/n/ confusion models across localities:

- Regions where both /l/ and /n/ are pronounced accurately and distinctly (i.e., no confusion) are concentrated in the urban districts of Hanoi, with scattered occurrences elsewhere in the delta. Peripheral areas not technically part of the delta, such as Hòa Bình, Phú Thọ, and Bắc Giang, show virtually no confusion.
- Regions with unidirectional confusion /n/ > [l] are concentrated in the northwest.
- Regions with unidirectional confusion /l/ > [n] are concentrated in the southeast.
- Regions with bidirectional confusion /n/ > [l] and /l/ > [n] are distributed in peripheral areas around Hanoi and scattered throughout the delta.

4.3. Interpretation of the Distribution of /l/ and /n/ Variants

4.3.1. According to the Wave Model

The wave model suggests that linguistic changes spread from a central point outward to surrounding areas, like ripples on water. The farther from the center, the weaker the influence of these 'waves'. Based on the results shown in Maps 2, 3, and 4, several centers and waves of variation can be identified:

Non-variation center (areas without confusion): Map 4 shows that the non-confused areas (Model A, red triangle) are mainly concentrated in the western and

northwestern parts of the delta (Hòa Bình, Phú Thọ, Bắc Giang). These are regions outside the Red River Delta, included in the study for comparative purposes. They maintain a clear phonemic distinction between /l/ and /n/, suggesting that the wave of change has not reached or has been blocked from these areas. In fact, the hilly and mountainous terrain in these regions can be seen as natural barriers that prevent the propagation of linguistic waves from the delta, in accordance with Bailey's directional wave model (Bailey, 1973).

Wave of /l/ > [n] transformation: This is the strongest and most widespread wave. In Map 3, the regions showing unidirectional confusion /l/ > [n] (Model C, blue dots) cover almost the entire eastern and southeastern coastal area of the delta (Hải Phòng, Hải Dương, Hưng Yên, Thái Bình, Nam Định, Ninh Bình). This wave seems to have originated in coastal centers and spread inland. The spread is not abrupt but creates transitional zones, where mixed variants appear—consistent with the idea of waves gradually weakening as they move inward.

Wave of /n/ > [l] transformation: This is a smaller-scale wave represented by unidirectional confusion /n/ > [l] (Model B, green diamonds), forming a cluster in Vĩnh Phúc, Bắc Ninh, and the northern outskirts of Hanoi. It may have originated from a local center in the north of Hanoi (the Kinh Bắc region) and spread outward.

Transition zone (Hanoi and surrounding areas): The area in and around Hanoi presents a complex picture. Maps 2 and 3 show a concentration of standard pronunciation with clear /l/–/n/ distinction. However, Map 4 shows the presence of all variation types, especially bidirectional confusion (Model D, yellow stars). According to the wave model, Hanoi serves as both a center for standardization and a convergence point where transformation waves from different directions collide (i.e., /l/ > [n] from the southeast and /n/ > [l] from the northwest), resulting in a non-homogeneous dialectal area, especially in suburban villages.

Based on previous research results not only in Vietnamese, it can be seen that the above change of /l/ and /n/ is not an isolated phenomenon in the Red River Delta and in Vietnamese. Near Vietnam, the /l/ - /n/ confusion is noted to be very common in dialects at the south of the Yangtze River (China), both in the river valley and in the entire vast South China Plain (Tonghe 2023: 33; Linguistics Atlas of Chinese Dialects, Phonetics, 057). Hoang Thi Chau believes that the /l/ - /n/ confusion may be a linguistic wave spreading across a large area from southern China to northern Vietnam (Hoang Thi Chau 2004: 137). We agree with her that this phenomenon may be related in some way to changes in Chinese in the South China region, but more specific explanations are needed regarding the direction of the spread of linguistic waves. This is related to

direct contact between Vietnamese and Chinese, which must have been a living Chinese language in Giao Chau (Nguyen Tai Can 1979: 38), after Vietnamese separated from Muong language. The fact that no confusion between /l/ and /n/ was found in Muong language shows that the confusion that exists in Vietnamese must have occurred after Vietnamese separated from Muong language and entered the Old Vietnamese period (13th to 16th century). This is the time when history witnessed many waves of Chinese migration into the Red River Delta. The famous commercial centers in the Red River Delta where the Chinese played a central role in commercial activities were the commercial urban area of Van Don (Quang Ninh), the urban area of Pho Hien (Hung Yen), the port urban area of Hai Phong and the urban area of Thang Long - Hanoi... Along with these commercial urban centers are chains of trading routes along the major rivers spreading throughout the provinces in the lower reaches of the Red River such as Hai Duong, Hung Yen, Thai Binh, Nam Dinh... (Chau Thi Hai 2001). This reality completely coincides with the waves that we have explained above. The wave of change /l/ > [n] originated from the coastal centers (Quang Ninh, Hai Phong) and spread deep inland (the entire lower reaches of the Red River) in the Northwest direction. The wave of change /n/ > [l] originated from the north of the urban area of Thang Long - Hanoi and spread in the Southeast direction. The two waves met to create an intersection in the center of Hanoi and the surrounding areas.

4.3.2. According to the Social Network Model

The Social Network Model posits that the maintenance or change of language depends on the structure and density of social relationships within a community. Closed networks (high density) tend to reinforce local pronunciation norms and resist external influences, while open networks (low density) facilitate the adoption of new linguistic variants from outside.

Closed networks and the maintenance of 'local norms': In the region exhibiting /l/ > [n] confusion (Model C) along the southeastern coast, an interesting phenomenon is observed. Instead of resisting the change, communities here appear to have adopted the non-standard /l/ > [n] variant as their own local norm. The consistency across this large region suggests that the rural village-based closed social networks are strong in reinforcing and spreading this variant. The pronunciation of /l/ as [n] has become a community marker, maintained through repeated social interactions.

Similarly, the /n/ > [l] cluster (Model B) in the northwest may also represent a community with a strong enough network to develop and preserve a distinct second local norm, different from both the standardized and coastal patterns.

In fact, Vietnamese dialectology has long noted that the Red River Delta hosts more sub-dialects than any other region (Hoàng Thị Châu 2004: 219). Archaeological evidence suggests this region is the cradle of the Vietnamese people, and where such cradles exist, multiple dialects tend to emerge. In these areas, subsistence-based economies in wet rice agricultural societies fostered village-based settlements, where people lived in near isolation, creating closed social networks with high internal density and few external links. Such networks reinforce and preserve long-standing local norms that diverge from national standards (Hoàng Thị Châu 2004: 221–222).

Open networks and linguistic diversity: This is evident in central Hanoi and its periphery, where bidirectional confusion (Model D) occurs. This is characteristic of open social networks. As a political, economic, and cultural center, Hanoi attracts residents from many regions with diverse linguistic habits. Social relationships here tend to be looser and less dense compared to rural communities, creating a linguistically diverse environment where no single standard prevails. As a result, multiple variants coexist, and bidirectional confusion is observed especially in suburban villages. This reflects linguistic instability and variability among individuals in loose, expansive social networks.

Thus, the combination of the two explanatory models above shows that the phenomenon of confusion of /l/ and /n/ in the Red River Delta is a complex process, reflecting both the history of migration, economic and cultural exchange (Wave Model) and social structure and community identity (Social Network Model).

Conclusion

This study provides a comprehensive overview of the distribution of /l/ and /n/ variants in the Red River Delta, Northern Vietnam. It identifies the phenomenon not only as a widespread pronunciation pattern, but more importantly, as a distinctive dialectal feature of the region.

The key findings are as follows:

1. Solid phonological foundation: The confusion between /l/ and /n/ stems from their articulatory similarities — same place of articulation (alveolar), shared sonority, and voicing — which facilitate assimilation and can lead to eventual phonemic merger over time.

2. Diverse transformation models: The study identifies four main local variation models: i. No confusion (Model A); ii. Unidirectional confusion /n/ > [l] (Model B); iii.

Unidirectional confusion /l/ > [n] (Model C); and iv. Bidirectional confusion (Model D). Among these, Model C (/l/ > [n]) is the most common and dominant.

3. Clear geographical distribution: i. No-confusion areas are concentrated in central Hanoi and border provinces (Hòa Bình, Phú Thọ, Bắc Giang); ii. /n/ > [l] areas are clustered in the northwest; iii. /l/ > [n] areas dominate the east and southeast; and iv. Bidirectional confusion is scattered, especially around Hanoi's outskirts.

4. Wave model insights: The /l/ > [n] transformation wave is the strongest, spreading inland from coastal centers. The /n/ > [l] wave is smaller in scale. Hanoi is a convergence zone where these waves intersect, leading to dialectal diversity.

5. Social network model insights: Rural closed networks reinforce local non-standard norms: e.g., /l/ > [n] in the southeast and /n/ > [l] in the northwest. Urban and peri-urban open networks foster diversity and coexistence of multiple variants, leading to bidirectional confusion.

In summary, the transformation of /l/ and /n/ in the Red River Delta is a complex process, shaped by both historical migration and cultural exchange (explained by the Wave Model), and social structure and community identity (explained by the Social Network Model).

References

Bailey Guy, Wikle Tom, & Sand Lory (1996) The Spatial Diffusion of Linguistic Features in Oklahoma. *Proceedings of the Oklahoma Academy of Science* 77: 1–15.

Bùi, Đăng Bình (2003) *Tìm hiểu thực trạng hiện tượng [l], [n] ở huyện Gia Lâm* [*Study on the current situation of [l], [n] phenomenon in Gia Lam district*]. Tiểu luận tập sự]. Viện Ngôn ngữ học.

Bùi, Minh Yến (2007) Học vấn với việc phát âm [l] - [n] trong tiếng Việt (ở một xã ngoại thành Hà Nội) [*Education with pronunciation of [l] - [n] in Vietnamese (in a suburban commune of Hanoi)*]. Trong *Ngôn ngữ và văn hóa Hà Nội*. Hội Ngôn ngữ học Hà Nội.

Chambers, J. K. & Trudgill Peter (2004) *Dialectology*, Cambridge University Press, Cambridge.

Châu, Thị Hải (2001). *Người Hoa với liên kết khu vực trong bối cảnh toàn cầu hóa* [*Chinese people and regional linkages in the context of globalization*]. Nghiên cứu Đông Nam Á, số 4.

Đoàn, Thiện Thuật (2016) *Ngữ âm tiếng Việt* [*Vietnamese Phonetics*]. Nxb Đại học Quốc gia Hà Nội. Hà Nội.

Endo, Misuaki (2016) *How to make Geolinguistical maps using ArcGIS Online*, β ver. 1.0 (ArcGIS Online β1.0). private edition.

Fukazawa, Mika (2017) *Manual for drawing geolinguistics maps with Arc-GIS online*, Proceedings of the Workshop Geolinguistic Method and Southeast Asian Linguistics, Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies, Tokyo, 16–33.

Hà, Quang Năng (2007) Khảo sát thực trạng cách phát âm lẫn lộn [L] – [N] hiện nay [Survey on the current state of pronunciation confusion [L] – [N]]. Trong *Ngôn ngữ và văn hoá Hà Nội*. Hội Ngôn ngữ học Hà Nội.

Hoàng, Thị Châu (2004) *Phương ngữ học tiếng Việt* [*Vietnamese Dialectology*]. NXB Đại học Quốc gia Hà Nội.

Labov, William (2003) Pursuing the cascade model. In D. Britain & J. Cheshire (eds.) *Social Dialectology: In honour of Peter Trudgill*, 9-21. John Benjamins Publishing Company.

Ladefoged Peter, Johnson Keith (2015) *A Course in Phonetics* (7th ed., tr. XX). Cengage Learning.

Milroy, Lesley (2002) Social Networks. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (eds.) *The Handbook of Language Variation and Change*, 549–572. Blackwell.

Ng, Choi Lee Charlie (2017). *Merger of the syllable-initial [n-] and [l-] in Hong Kong Cantonese* (OAPS). City University of Hong Kong. URL: https://lbms03.cityu.edu.hk/oaps/lt2017-4235-ncl190.pdf

Nguyễn, Tài Cẩn (1979). *Nguồn gốc và quá trình hình thành cách đọc Hán Việt* [*Origin and formation process of Sino-Vietnamese reading*]. NXB Khoa học Xã hội. Hà Nội.

Nguyễn, Thị Thanh Bình (2000) [n] hay [l] ở một làng quê Việt Nam: Một quan sát từ góc độ ngôn ngữ học xã hội [[n] or [l] in a Vietnamese village: An observation from a sociolinguistic perspective]. Trong *Ngôn từ, giới và nhóm xã hội từ thực tiễn tiếng Việt*. NXB Khoa học xã hội.

Phạm, Thịnh (2011) Hơn 46.000 HS và giáo viên Hà Nội phải 'chữa ngọng' [More than 46,000 students and teachers in Hanoi must 'correct lisp']. https://vtcnews.vn/hon-46000-hs-va-giao-vien-ha-noi-phai-chua-ngong-ar58316.html/

Phạm, Văn Hảo (2024) *Ngữ âm địa phương của tiếng Việt* [Local phonetics of Vietnamese]. Cánh Buồm. https://ane.edu.vn/ngu-am-dia-phuong-cua-tieng-viet/

Ralph W. Fasold, Walt Wolfram (1974), *The study of Social Dialects in American English*, Newbury House Publishers & Rowley, Massachusetts.

Tonghe, Dong (2023). *Historical Phonology of Chinese* (1st ed.). Routledge. https://doi.org/10.4324/9781003411444.

Trần, Thị Thìn (1979) Bước đầu tìm hiểu hiện tượng phát âm lệch chuẩn /l/, /n/ [Study the phenomenon of non-standard pronunciation /l/, /n/]. *Tạp chí Ngôn ngữ*, (2).

Trịnh, Cẩm Lan (2008) Lý thuyết làn sóng trong nghiên cứu ngôn ngữ và văn hoá Thăng Long – Hà Nội [Wave theory in the study of Thang Long - Hanoi language and culture]. *Tạp chí Ngôn ngữ,* (1).

Trịnh, Cẩm Lan (2017) *Tiếng Hà Nội - từ hướng tiếp cận phương ngữ học xã hội* [*Hanoi dialect - from a socio-dialectical approach*], NXB Đại học Quốc gia Hà Nội, Hà Nội.

Trịnh, Cẩm Lan &Trần, Thị Hồng Hạnh (2025) Ngôn ngữ học địa lý, Phương ngữ học và Bản đồ phương ngữ: Khái niệm, lịch sử và những vấn đề đang đặt ra [Geolinguistics,

Dialectology and Dialect Maps: Concepts, History and Current Issues], *Tạp chí Khoa học & Công nghệ*, Trường ĐH Công nghiệp Hà Nội, Chuyên san Ngôn ngữ & Văn hoá, ISSN: 1859-3585, số 2. 2025.

Trudgill, Peter (1974) Linguistic Change and Diffusion: Description and Explanation in Sociolinguistic Dialect Geography. *Language in Society* 3(2): 215–246. https://doi.org/10.1017/S0047404500004358

Vũ, Kim Bảng (2005) Nhận xét về sự khác biệt ngữ âm giữa nội thành và hai huyện Gia Lâm, Đông Anh [Comments on the phonetic differences between the inner city and the two districts of Gia Lam and Dong Anh]. *Tạp chí Ngôn ngữ*, (5): 15–26.

Vũ, Thị Thanh Hương (2007) Tiếng Hà Nội khu vực phố cổ [Hanoi accent in the Old Quarter]. *Tạp chí Ngôn ngữ*, (11).

Yuyan Zhou, Ying Wu (2019). An ERP Study for Phonemic Merger in Chinese Dialects. *Modern Linguistics* 现代语言学 7(1): 43–53. Published Online February 2019 in Hans. http://www.hanspub.org/journal/ml https://doi.org/10.12677/ml.2019.71007

Linguistics Atlas of Chinese Dialects. https://zhongguoyuyan.cn/index

Thư viện pháp luật Việt Nam: https://thuvienphapluat.vn/)

Diversity in grammatical voice and noun marking systems in the languages of the Philippines and Indonesia

Atsuko Utsumi (Meisei University)

Abstract: Although Western Malayo-Polynesian (WMP) languages exhibit many typological similarities, inner diversity with regard to noun markers and voice systems is found. Noun markers, often referred to as case markers, play a crucial role in languages with a complex voice system, however, they are not as strongly required in languages with a simple voice system. This paper presents data on the case marking system and voice system in the languages of the Philippines and Indonesia, and discusses the correlation between the two grammatical categories.*

Key words: West Malayo-Polynesian languages, grammatical voice system, case marking, noun marker, Austronesian geolinguistics

1. Introduction: Philippine-type languages and Indonesian-type languages

The Austronesian language family consists of 1,257 languages (Eberhard, Simons & Fennig 2022) and is distributed in a large geographical area; from Taiwan in the north, New Zealand to the south, Easter Island/Rapa Nui in the east, and Madagascar to the west. Western Malayo-Polynesian (WMP) languages, which are one of the sub-families, entail languages spoken in Taiwan, the Philippines, Sulawesi, and all islands to its west, Borneo, Madagascar, and the Palau and Chamorro languages (Himmelmann 2005a:111).

Morphosyntactically speaking, WMP languages display "symmetrical voice" alternations, in which two distinct types of transitive clause are hypothesized. In the majority of languages which exhibiy "asymmetrical voice" alternations, a single transitive construction with the most basic verb form and two (or more) core arguments is posited, as presented in Table 1. When a verb form is marked by a passive or an antipassive morpheme, demotion of one of the arguments occurs, resulting in an intransitive construction.

UTSUMI, Atsuko. 2025. Diversity in grammatical voice and noun marking systems in the languages of the Philippines and Indonesia. In Nor Hashimah Jalaluddin, Hiroyuki Suzuki and Mitsuaki Endo (eds.) *Proceedings of the sixth International Conference of Asian Geolinguistics*, 99–117. doi: https://doi.org/10.5281/zenodo.17204638

* This research is supported by JSPS KAKENHI Grant Number 23K20093.

In "symmetrical voice" languages, on the other hand, Actor voice and Undergoer voice constructions containing a transitive verb exhibit equal transitivity with the same number of arguments, as shown in Table 2. A transitive verb in both constructions is marked by an affix (or two in some cases), which makes it impossible to determine which construction is more basic. Voice marking indicates the syntactic prominence and semantic role of a single core argument, sometimes called the pivot in the literature on Philippine/Indonesian languages (Foley 2008, Brickell 2023). For example, the actor voice marking affix indicates the pivot has the "Actor" semantic role.

Table 1: Commonly attested voice alternation and argument demotion (Adopted from Brickell 2023)

Promotion and Demotion of Core Arguments in Asymmetrical voice alternations			
Clause Construction	Core Argument 1	Core Argument 2	Transitivity
Active/Default (Unmarked)	Yes (pivot)	Yes (non-pivot)	Transitive
Passive/Antipassive (Marked)	Yes(pivot, previously non-pivot)	N/A(previous pivot now oblique)	Intransitive

Table 2: Voice alternation in symmetrical voice languages (Adopted from Brickell 2023)

Promotion and Demotion of Core Arguments in Symmetrical voice alternations			
Clause Construction	Core Argument 1	Core Argument 2	Transitivity
Actor voice	Yes (pivot)	Yes (non-pivot)	Transitive
Patient voice	Yes (pivot, previously non-pivot)	Yes (non-pivot, previous pivot)	Transitive

Within symmetrical voice languages, two major sub-types, the Philippine-type and the Indonesian-type, are often distinguished (Arka 2002; Arka and Ross 2005; Himmelmann 2005a). Philippine-type languages roughly encompass the languages of the Philippines, certain Formosan languages, and some languages of northern Borneo and northern Sulawesi. The Indonesian-type incorporates the rest of the languages in insular Southeast Asia, such as varieties of Malay in the Malay Peninsula and Sumatra, and the languages of Java, Madura, Bali, and Lombok.

The two types exhibit differences in several aspects, three of which will be presented in this paper. First, the number of undergoer voices is limited to one in Indonesian-type languages, whereas Philippine-type languages have at least two, sometimes more. Second, a rigid nominal marking system is typically found in the Philippine-type, whereas the Indonesian-type lacks one or only shows a partial nominal marking system. The third feature relates to the second one: pronouns in Philippine-type languages also exhibit at least three distinct forms with regard to cases, whereas Indonesian-type languages do not. These three characteristics will be examined in detail, and their geographical distribution will be presented below.

The other differences include the absence/presence of the second-position clitics and applicative constructions. Philippine-type languages have pronominal clitics and those which denote negation or Tense-Aspect-Mood (TAM) and which are placed in the second position of the clause, while the Indonesian-type languages lack such clitics. In contrast, applicative constructions are found in Indonesian-type languages, whereas they are absent in Philippine-type languages. Philippine-type languages are assumed to be more conservative and homogeneous, although languages in the southern Philippines, Sulawesi, and Borneo may exhibit intermediate features, and are sometimes called "transitional languages" (Himmelmann and Riesberg 2013; Hemmings 2015). This paper aims to show the geographical distributions of the three features that distinguish the two sub-types of WMP languages, in order to show how transitional features manifest in the intermediate languages of the two established sub-types. The case-marking system and the grammatical voice system will be described in section 2, the distributions of voice-alternation types and pronominal case and noun-marker oppositions will be presented in section 3, followed by the conclusion in section 4.

2. Nominal marking systems and grammatical voice alternations in WMP language

As shown in section 1, Indonesian-type languages exhibit a two-way voice alternation, which includes one Actor voice and one Undergoer voice. The pivot noun in Actor voice sentences denotes ACTOR, EXPERIENCER, or CAUSER, whereas that of Undergoer voice exhibits various semantic roles, including PATIENT, COVEYED THEME, BENEFICIARY, and CAUSEE. Examples (1a) and (2a) are Actor voice constructions from Indonesian, and 1b and 2b are corresponding Undergoer voice constructions. The pivot nouns in (1a) and (2a) are ACTOR, whereas the semantic roles of the pivots in Undergoer constructions differ; that in (1b) is a PATIENT and that in (1b) is a CONVEYED THEME. The prefix *me-* indicates Actor voice, whereas the prefix *di-* indicates Undergoer voice; no TAM meaning is indicated by either of the affixes.

Tables are based on the following style, put as centred. Please take a blank line before a caption and after a table. The content in the table and captions are both in 9pt.

(1) Indonesian

a. *anwar* *me-mukul* *anjing*
Anwar AV-beat dog

b. *anjing* *itu* *di-pukul* *Anwar*
dog that UV-beat Anwar (PATIENT as the subject)
"Anwar beat a dog"

(2) Indonesian

a. *Ibu* *Puji* *mem-bawa* *buku* *ini*
mother Puji AV-bring book this

b. *buku* *ini* *di-bawa* *Ibu* *Puji*
book this UV-bring mother Puji
(CONVEYED THEME as the subject)
"Puji's mother brought this book"

Indonesian-type languages are not equipped with complete (productive?) noun-marking systems nor pronominal case systems, as shown in examples (1) and (2), but instead have productive applicative affixes, as observed in (3), in which the Indonesian applicative affixes *-kan* and *-i* are exemplified (present?). Voice-marking affixes typically exclusively indicate voice, as in examples (1-3), unlike Philippine-type languages which have Tense-Aspect-Mood and Voice complex affixes.

(3) a. *pelayan* *meng-ambil-kan* *tamu* *segelas* *air*
waiter AV-take-APPL guest one.glass water
"The waiter took a glass of water to the guest"(Adapted from *Kyokasho Indonesia-go*, p 121)

b. *dia* *me-naik-i* *gunung* *Merapi*
3sg AV-go.up-APPL mountain Merapi.
"He climbed Mount Merapi"(Adapted from *Kyokasho Indonesia-go*, p 125)

Philippine-type languages, which exhibit two or more Undergoer voices, often display a noun-marking system which plays an important role in a construction. Additionally, a regular alternation between noun-markers and voices is found, as shown in the Tagalog example (4) below. The three pronominal cases are nominative, genitive, and dative/locative, and common nouns are marked by corresponding markers; *ang* for nominative, *ng* for genitive, and *sa* for dative/locative. The pivot noun, which is underscored, is marked by a noun marker *ang* in (4b), (4c), and (4d), or exhibits

nominal case if the pivot is a pronoun, as in 4a. The verb in each sentence includes different voice-marking affixes; the infix *-um-* indicates Actor voice, the suffix *-in* Patient voice, the suffix *-an* Locative voice, and the prefix *i-* Beneficiary voice. In addition, Tagalog exhibits two aspects and two moods for each voice form, resulting in four different forms: non-realis/perfective, non-realis/imperfective, realis/perfective, and realis/imperfective (Himmelmann 2005b).

(4) Tagalog

a. Actor voice
h<um>iram =ka ng libro sa aklatan para sa anak=ko
<AV>borrow= 2s.NOM GEN book DAT library for DAT child=1sg.GEN

b. Patient voice
hiram-in=mo ang libro sa aklatan para sa anak=ko
borrow-PV=2sg.GEN NOM book DAT library for DAT child=1sg.GEN

c. Locative voice
hiram-an=mo ng libro ang aklatan para sa anak=ko
borrow-LV=2sg.GEN GEN book NOM library for DAT child=1sg.GEN

d. Beneficiary voice
i-hiram=mo ng libro sa aklatan ang anak=ko
BV-borrow=2sg.GEN GEN book DAT library NOM child=1sg.GEN

Transitional languages that have mixed features of both Philippine-type and Indonesian-type languages exhibit fewer voice alternations and an imperfect noun-marking system, as in the Talaud examples shown in (5) and the Bantik examples shown in (6). Talaud and Bantik, two languages spoken in north Sulawesi province, Indonesia, have one Actor voice and two Undergoer voices, which are indicated by TAM and voice complex affixes. Their noun-marking system, too, is incomplete when compared with Tagalog. Pronouns, proper names, and singular human nouns exhibit three cases: nominative, genitive, and locative. Other common nouns, on the other hand, do not take a noun marker when in the nominative case, so *daho udde* (these guests) in (5b) lacks any noun marking. In Talaud, many verbs exhibit a three-way voice alternation, but semantically intransitive verb bases show derivational meaning in Undergoer voice constructions. For example, although the same verb base *saraiŋ* (dance) appears in Example (5a),(5b), and (5c), in the Actor voice (5a) it means "to dance", in the Goal voice (5b) "to be entertained by watching dance", and in the Conveyance voice (5c) "use something for dancing". Bantik, unlike typical Philippine-type languages, exhibits applicative constructions, and a three-way voice alternation is only found with applicative verbs, as shown in example (6), as well as with causative

verbs. Conveyance voice is indicated by a zero morpheme, as shown in (6c). All the verbs in examples (5) and (6) are in the non-past tense.

(5) Talaud
a. *i-maŋitou ma-saraiŋŋ u-lamaʔa su-daho udde*
NOM-3pl AV-dance GEN-dish LOC-guests that
"They will dance a dish dance to these guests" Actor voice, non-past
b. *daho udde saraiŋ-an ni-maŋitou riŋan nu-saraiŋŋa lamaʔa*
guest that dance-GV GEN-3pl with GEN-dance dish
"Those guests will be entertained by them with 'dish dance'" Goal voice, non-past
c. *lamaʔa i-saraiŋŋ i-maŋitou*
dish CV-dance GEN-3pl
"Dishes will be used by them in a dance" Conveyance voice, non-past

(6)Bantik
a.*i-remi ma-pa-mandaŋ nu-pisou=ne su-pun nu-teriŋ*
NOM-Remi AV-APPL-test GEN-knife=GEN-3sg LOC-tree GEN-bamboo (AV)
b. *pun nu-teriŋ pa-mandam-en ni-remi nu-pisou=ne*
tree GEN-bamboo APPL-test-GV GEN-Remi GEN-knife=GEN-3sg (GV)
c. *pisou=ne Ø-pa-mandaŋ ni-remi su-pun nu-teriŋ*
knife=GEN-3sg Ø-APPL-test GEN-Remi LOC-tree GEN-bamboo (CV)
"Remi will test his knife on a bamboo trunk."

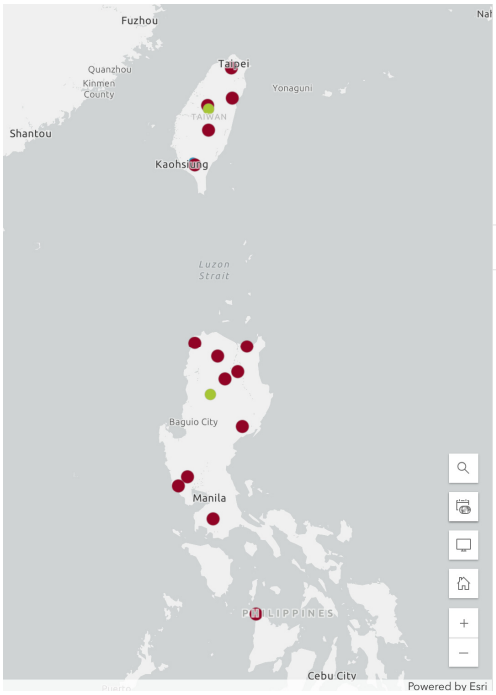
To summarize, Philippine-type languages exhibit one Actor voice and two or more Undergoer voices, altogether more than three. Typical Philippine-type languages, such as Tagalog, show a four-way voice alternation. Also, these languages have rigid noun-marking systems. Indonesian-type languages, on the other hand, have two grammatical voices at most, lack noun markers, and have fewer pronominal cases. In between these two types of symmetrical voice languages, there are languages with transitional features, such as fewer voice alternations and an incomplete noun-marking system.

3. Distribution of voice-alternation types case marking systems

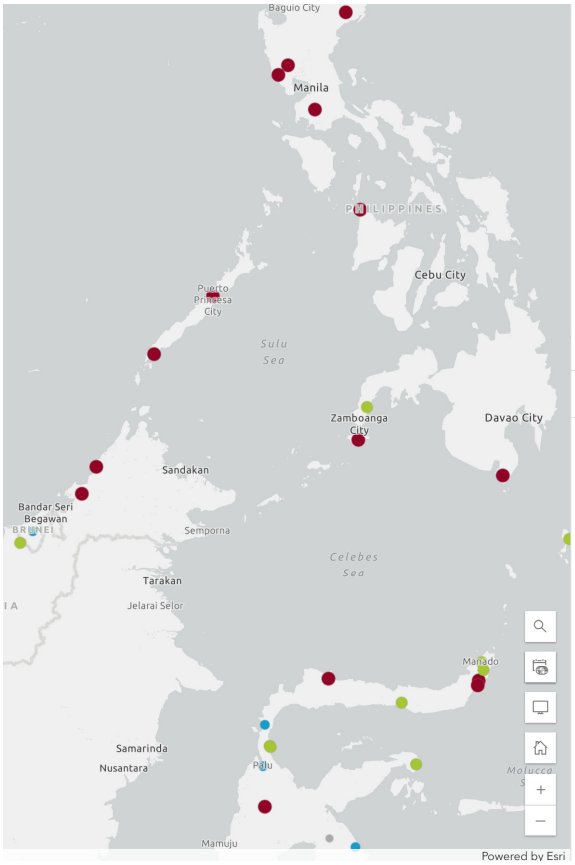
3.1. The number of grammatical voices in WMP languages

WMP languages located in Taiwan, the Philippines, northern Borneo, and northern Sulawesi are categorized as Philippine-type. Map 1 shows the number of voices in Taiwan and northern Philippines, where most of the languages have four-way voice

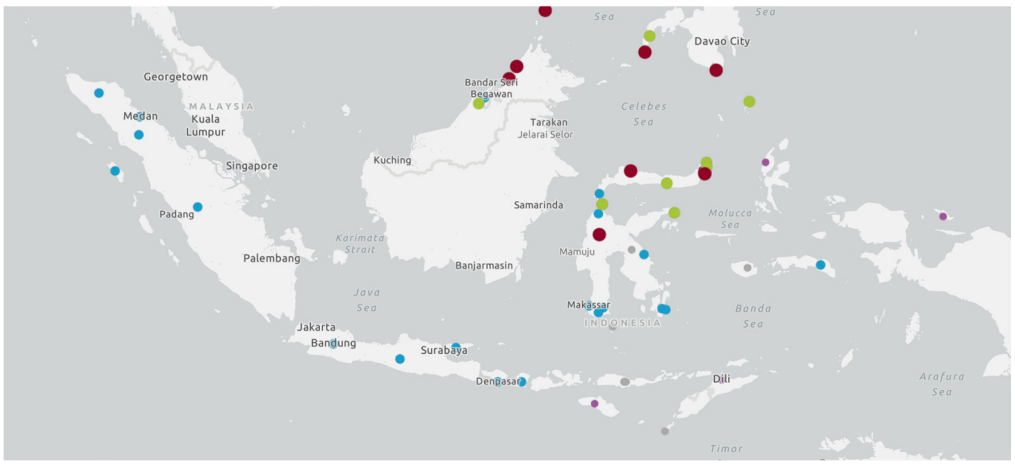
alternations. When we look at the languages further south in Mindanao, northern Borneo, and Sulawesi, the frequency of languages with three-way voice alternations increases, as shown in Map 2. In the languages of western and southern Indonesia, as shown in Map 3, a two-way voice alternation becomes prevalent.



Map 1: The number of grammatical voices in Taiwan and Northern Philippines



Map 2: The number of voices in southern Philippines, northern Borneo, and northern Sulawesi



Map 3: The number of grammatical voices in West and South Indonesia

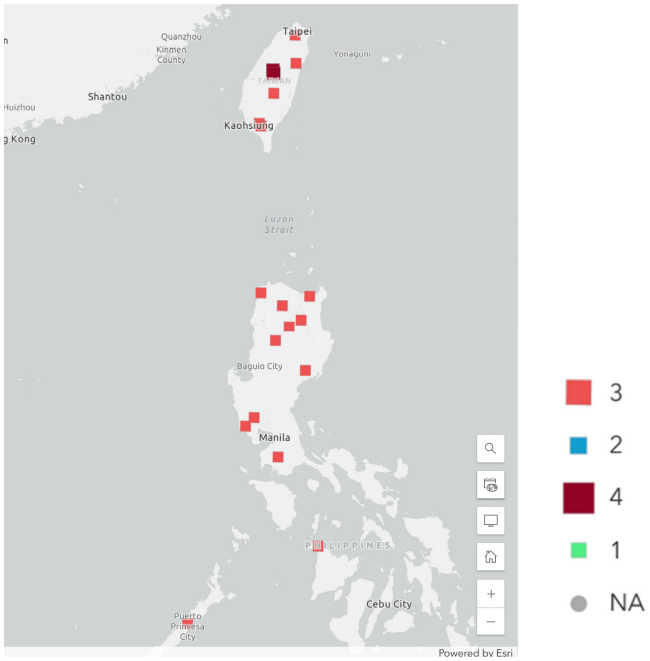
3.2. The number of pronominal cases and noun markers in WMP languages

Philippine-type languages have a pronominal paradigm in which multiple cases are differentiated, mostly three (pivot, non-pivot, locative; or nominative, genitive, dative/locative), at least 2 cases (nominative and genitive/oblique). Noun markers encode corresponding cases on common nouns. In principle, the number of pronominal cases and noun markers coincide, but some languages have fewer noun markers than pronominal cases. Verb-initial word order is prevalent in Philippine-type languages, so two or more nominal arguments are placed next to one another in transitive constructions. Pronominal cases and noun markers indicate grammatical roles, so that a freer (flexible?) word order of arguments is allowed.

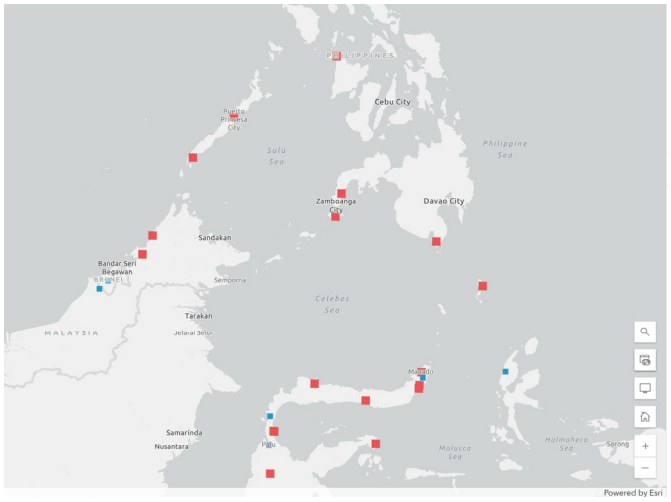
Most Indonesian-type languages differentiate two pronominal cases, nominative and genitive, but in some languages, case distinction is limited to certain items only. Other languages do not exhibit case marking, but rather display other oppositions, such as free forms vs dependent forms. These languages rarely require noun markers. In order to distinguish grammatical roles, stricter word order, mostly subject-verb-object (or Actor-verb-Patient), is required.

Transitional languages exhibit intermediate features; some require verb-initial word order, whereas others require subject-verb-object word order. Three pronominal cases are found in many of them, but the number of noun markers is often less than that of pronominal cases, or they are not strictly required.

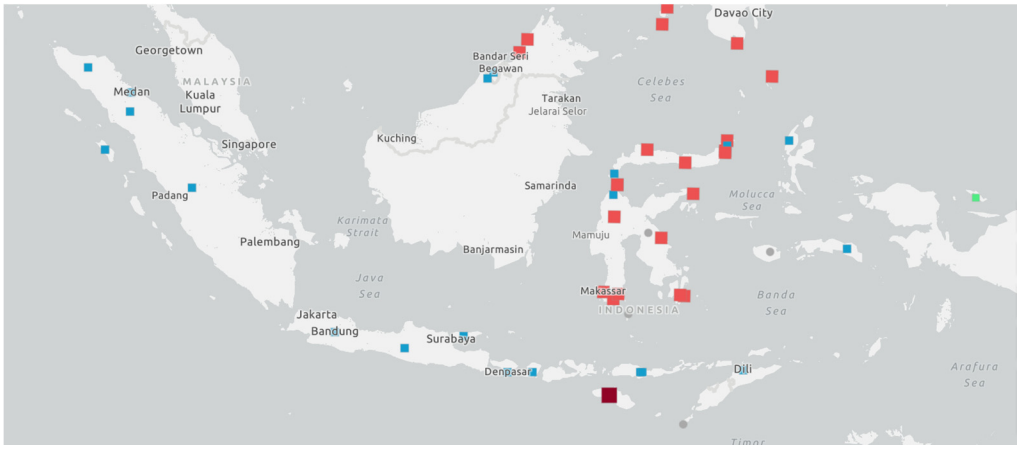
As shown in Maps 4 and 5, in Philippine-type languages and transitional languages located (spoken?) from Taiwan to northern Borneo and Sulawesi, a three-way pronominal case distinction is most common. Map 6 shows that many Indonesian-type languages in Sulawesi also have a three-way distinction, whereas those in other areas mostly show a two-way distinction.



Map 4: Number of pronominal cases in Taiwan and Northern Philippines

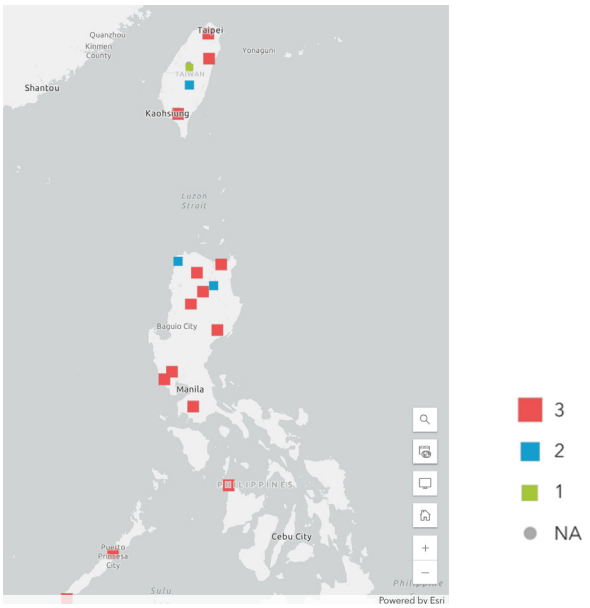


Map 5: Number of pronominal cases in southern Philippines, northern Borneo, and northern Sulawesi

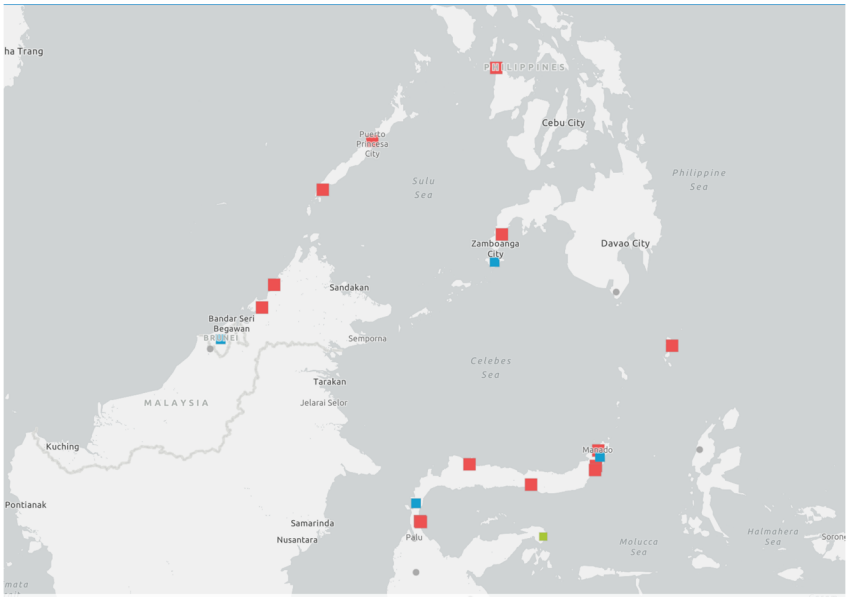


Map 6: Number of pronominal cases in West and South Indonesia

In most Philippine-type languages, three noun markers are found, which distinguish nominative, genitive, and dative/locative cases. However, a few exhibit only two noun markers, as shown in Map 7. Transitional languages, too, exhibit two or three noun markers, as presented in Map 8. In the area where Indonesian-type languages are prevalent, languages with no noun markers are most common, whereas transitional areas find languages with two or three noun markers, as shown in Map 9.



Map 7: Number of noun markers in Taiwan and northern Philippines



Map 8: Number of noun markers in southern Philippines, northern Borneo, and northern Sulawesi

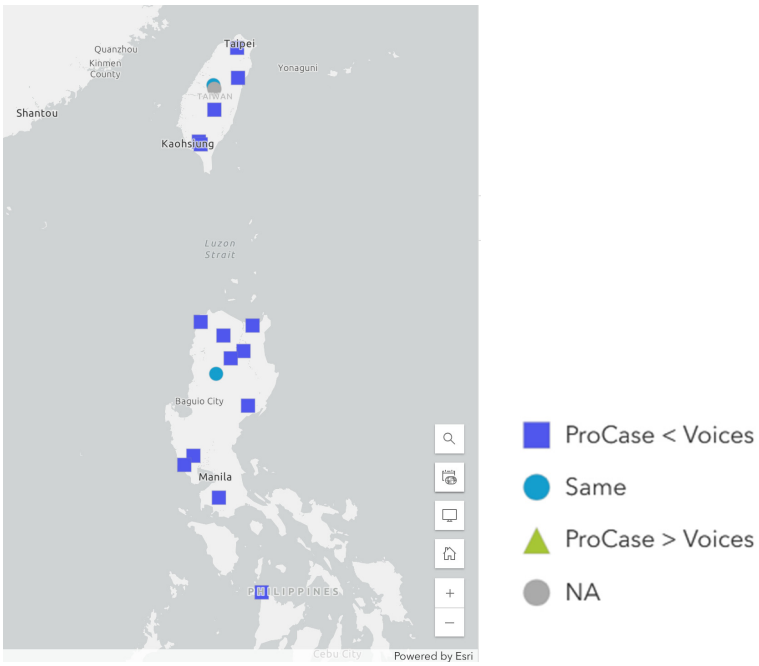
Map 9: Number of noun markers in southern and western Indonesia

3.3. Correlation between the number of grammatical voices and pronominal/ nominal case marking

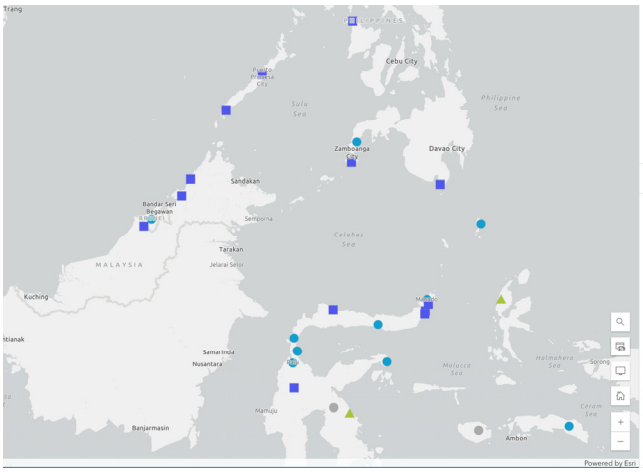
The number of grammatical voices correlates with both the number of pronominal cases and the number of noun markers. Many Formosan languages and northern Philippine languages have four voices or more, transitional languages have three, and (while?) most Indonesian languages have two. The three-way pronominal case distinction is prevalent in Taiwan, the Philippines, and also in Sulawesi, but a two-way distinction is most often found in areas further south and west. Noun marking systems have a similar distributional trend; languages in Taiwan and the Philippines exhibit a three-way noun marker opposition, whereas those in western and southern Indonesia lack noun markers. Two-way distinctions are found in languages in the transitional areas.

In short, languages with four grammatical voices tend to have three pronominal cases and three noun markers. Those with three grammatical voices most commonly have two to three pronominal cases and one to two noun markers, as is often the case with transitional languages. Those with two-way voice alternations seldom have pronominal case opposition or noun markers.

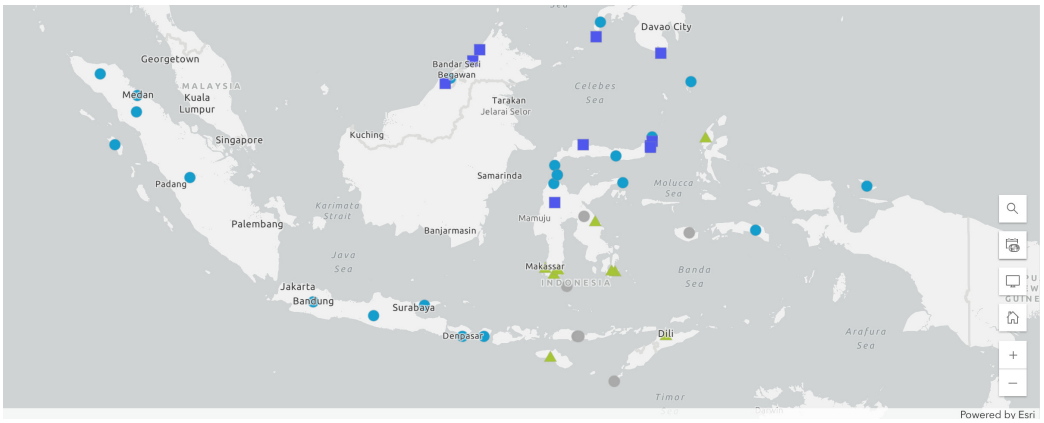
Maps 10, 11, and 12 show the correlation between the number of grammatical voices and pronominal cases. The number of grammatical voices in Philippine-type languages is four, while the pronominal cases are mostly three. As a result, in most languages in Taiwan and northern Philippines, and northern Borneo, there are fewer pronominal cases than the number of grammatical voices, as presented in Map 10. In contrast, some languages in the southern Philippines and northern Sulawesi exhibit the same number of voices and pronominal cases, although others have more voices than pronominal cases, as shown in Map 11. In western and southern Indonesia, most languages have two grammatical voices and two pronominal cases, resulting in the same number of voices and pronominal cases, as in Map 12. No language has more pronominal cases than grammatical voices.



Map 10: Correlation between the number of voices and pronominal cases in Taiwan and northern Philippines

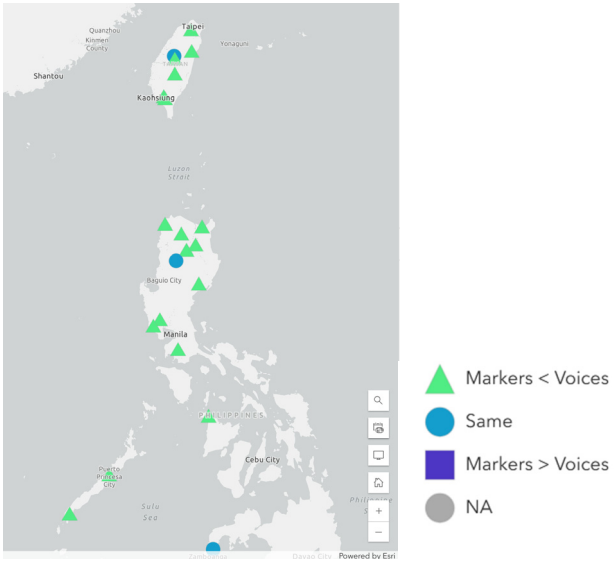


Map 11: Correlation between the number of voices and the number of pronominal cases in southern Philippines, northern Borneo, and northern Sulawesi

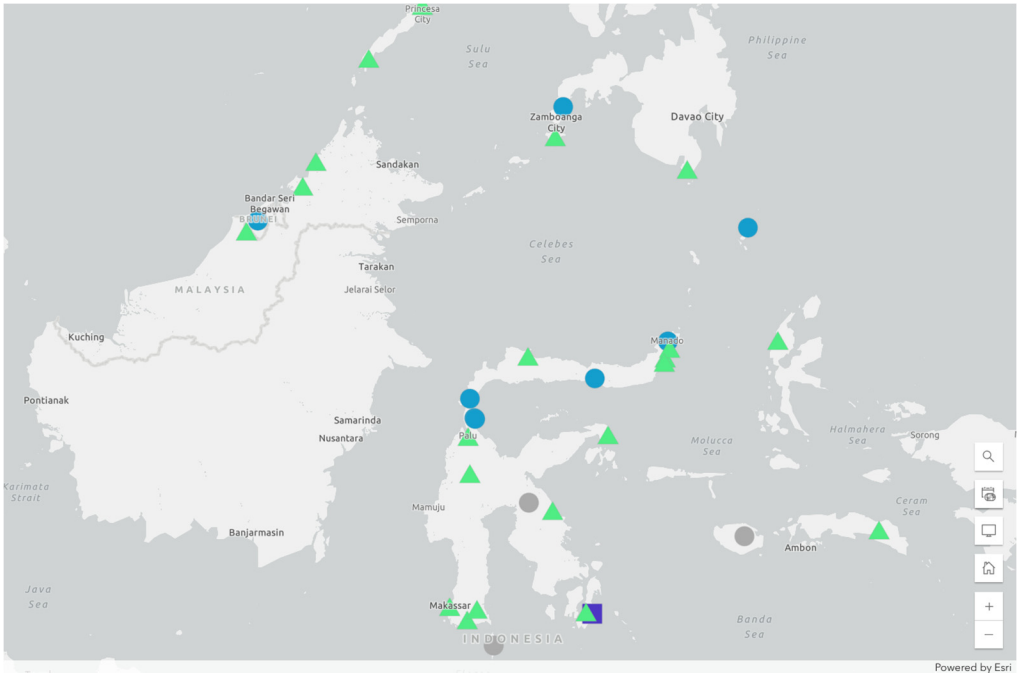


Map 12: Correlation between the number of voices and the number of pronominal cases in western and southern Indonesia

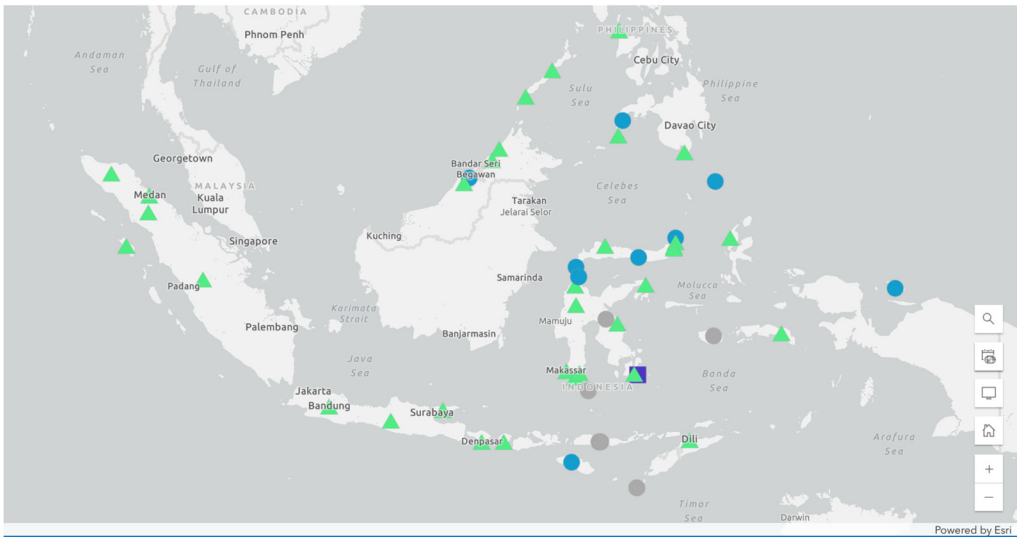
When we look at the correlation between the number of voices and noun markers, a different picture appears. In the majority of WMP languages throughout insular Southeast Asia, the number of noun markers is less than that of grammatical voices irrespective of sub-type, and no languages show a larger number of voices than noun markers. Fewer languages in the Philippines, Borneo, and Sulawesi exhibit the same number of voices and noun markers.



Map 13: Correlation between the number of voices and the number of noun markers in Taiwan and northern Philippines



Map 14: Correlation between the number of voices and the number of noun markers in southern Philippines, northern Borneo, and northern Sulawesi



Map 15: Correlation between the number of voices and the number of noun markers in western and southern Indonesia

4. Conclusion

Western Malayo-Polynesian languages are subcategorized into Philippine-type and Indonesian-type. The former typically exhibit more than three grammatical voices, three-way pronominal case marking, and three types of noun markers. In other words, they have a more complex voice system and a larger number of pronominal cases and noun markers in order to mark the pivot argument and the semantic role of non-pivot arguments in verb-initial clauses. The marking on nouns makes it possible for core arguments to take relatively free word order.

The latter subtype, which encompasses western and southern Indonesia, presents a two-way voice alternation. Pronominal cases mostly number two, and noun markers are rarely present. Languages that exist in the intermediate areas show transitional features; the number of voices and the number of pronominal cases mostly number three, and noun marking systems are not complete (productive?).

The number of grammatical voices and the number of pronominal cases/case markers correlate: languages with a more voices have a larger number of cases on NPs, whereas languages that have a two-way voice alternation exhibit two pronominal cases at best and often lack a noun marking system.

Throughout WMP languages, the number of pronominal cases is the same or greater than noun markers for common nouns. The number of voices is the same or more than that of either pronominal cases or that of noun markers.

Abbreviations

3sg	third person singular
3pl	third person plural
APPL	applicative
AV	active voice
BV	beneficiary voice
CV	conveyance voice
DAT	dative
GEN	genitive
GV	goal voice
LOC	locative
NOM	nominative

References

Arka, I Wayan. 2002. Voice systems in the Austronesian languages of Nusantara: Typology, symmetricality and undergoer orientation. Paper presented at the 10th National Symposium of Indonesia. Bali, Indonesia: Linguistics Society.

Arka, I Wayan & Malcom D. Ross (eds.). 2005. *The many faces of Austronesian voice systems: Some new empirical studies*. Canberra: Pacific Linguistics.

Blust, Robert. 2013. *The Austronesian languages (Revised edition)*. Canberra: Pacific Linguistics. http://hdl.handle.net/1885/10191

Brickell, Timothy. 2022. *Tondano (Toundano): A sketch grammar of an endangered Minahasan language.* London: Routledge.

Brickell, Timothy. 2023. Bound pronouns and constituent order in the Minahasan languages: the significance for western Austronesian typology. A paper presented at the Tokyo University of Foreign Studies.

Cumming, Susanna. 1991. *Functional change: the case of Malay constituent order*. Berlin: De Gruyter.

Donohue, Mark. 2007. Word order in Austronesian from north to south and east to west. *Linguistic Typology* 11: 349–391. https://doi.org/10.1515/LINGTY.2007.026

Eberhard, David M., Simons, Gary F., & Fennig, Charles D. 2022. *Ethnologue: languages of the world*. Twenty-fifth edition. Retrieved from https://www.ethnologue.com/.

Foley, William. 2008. The place of Philippine languages in a typology of voice systems. In Peter K. Autrin & Simon Musgrave (eds.), *Voice and grammatical relations in Austronesian languages*, 22–44. Stanford: CSLI Publications.

Hemmings, Charlotte. 2015. Kelabit voice: Philippine-type, Indonesian-type or something a bit different? *Transactions of the Philological Society* 113(3): 383–405. https://doi.org/10.1111/1467-968X.12071

Himmelmann, Nikolaus P. 2005a. The Austronesian languages of Asia and Madagascar: Typological characteristics. In K.A. Adelaar & N.P. Himmelmann (eds.). *The Austronesian languages of Asia and Madagascar*, 110–181. London: Routledge.

Himmelmann, Nikolaus P. 2005b. Tagalog. In K.A. Adelaar & N.P. Himmelmann (eds.). *The Austronesian languages of Asia and Madagascar*. London: Routledge, 350–376.

Himmelmann, Nikolaus P. & Sonja Riesberg. 2013. Symmetrical voice and applicative alternations: Evidence from Totoli. *Oceanic Linguistics* 52(2): 396–422.

Kroeger, Paul. 1993. *Phrase structure and grammatical relations in Tagalog*. Stanford: CSLI Publications.

Mead, David. 2002. Proto-Celebic focus revisited. In Fay Wouk & Malcom Ross (eds.), *The history and typology of western Austronesian voice systems*, 143–177. Canberra: Pacific Linguistics.

Moriyama, Mikihiro and Akio Kashimura[森山幹弘・柏村彰夫]. 2003. *Kyokasho Indonesia-go* 《教科書インドネシア語》[Textbook of Indonesian]. Tokyo: Mekon.

Riesberg, Sonja. 2014. *Symmetrical voice and linking in Western Austronesian languages*. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9781614518716

Schachter, Paul, & Otanes, Fe T. 1972. *Tagalog reference grammar*. Berkeley: University of California Press.

Sneddon, James N. 1975. *Tondano phonology and grammar*. Canberra: Pacific Linguistics.

Sneddon, James N. 1978. *Proto-Minahasan: Phonology, morphology, and wordlist*. Canberra: Pacific Linguistics.

Terok, R. 1994. *The system of verbal affixes in Tombulu*. Masters dissertation. Melbourne: La Trobe University.

Tryon, Darrell T., ed. 1994. *Comparative Austronesian Dictionary: An introduction to Austronesian studies* (Trends in Linguistics Documentation) Vol 1-5. Berlin: Mouton de Gruyter.

Utsumi, Atsuko [内海敦子]. 2005. Bantik go no kouzou to setsuji no imi/kinou「バンティック語の構造と接辞の意味・機能」[The structure of the Bantik language and meaning and function of its affixes]. Tokyo: University of Tokyo PhD thesis.

van den Berg, René. 1996. The demise of focus and the spread of conjugated verbs in Sulawesi. In Hein Steinhauer (ed.), *Papers in Austronesian linguistics No.3*, 89–114. Canberra: Pacific Linguistics.

Wolff, John U. 1973. Verbal inflection in Proto-Austronesian. In Andrew B. Gonzalez (ed.), *Parangal Kay Cecilio Lopez* [Essays in honor of Cecilio Lopez on his seventy-fifth birthday], 71–94. Manila: Linguistic Society of the Philippines.

Dialect transition along the Perak River

Khairul Ashraaf Saari (University Poly-Tech Malaysia)
Nor Hashimah Jalaluddin (Linguistics Association of Malaysia)
Harishon Radzi (National University of Malaysia)

Abstract: Dialectal transition may occur when there is a gradual change in speech patterns as a result of geographic interaction and social mobility within a community. Phonetic variation, particularly observed along riverine corridors, reflects the dynamic interactions present in such areas. This paper aims to identify the distribution of dialects and analyze the zones of dialect transition occurring along the Perak River corridor. The research instrument focuses on dialectological investigation involving 957 informants from 59 villages situated along the Perak River. This study employs a descriptive geolinguistic approach to examine the causes of dialectal distribution in the study area, with isogloss maps generated using Geographic Information Systems (GIS). The analysis reveals the presence of seven distinct phonetic variants [bantaj], [batɑ], [bantɛ], [bante], [batal], [bantɑ], and [bantal]. Hulu Perak District has been identified as the primary site of dialect convergence and the most significant transition zone, ultimately leading to the emergence of new phonetic variants as a result of dialectal diffusion in the region.

Key words: dialect transition, Perak River, zone transition, lateral, Geographical Information System

1. Introduction

Dialect refers to a linguistic variation that serves as a medium for conveying information used by a specific speech community. According to Mengrui Zhu (2023), a dialect not only functions as a tool for expressing thoughts and emotions but also plays a crucial role in the continuity and preservation of a community's cultural identity. This reflects how language practices often mirror a community's cultural values. Similarly, Asmah Haji Omar (2015) argues that the use of a dialect is inherited through family lineage and passed down from one generation to the next.

Yule (1996) categorizes dialectal variation into two main branches: social dialects (speech influenced by social status) and geographical dialects (speech shaped by

SAARI, Khairul Ashraaf, Nor Hashimah Jalaluddin, and Harishon Radzi. 2025. Dialect transition along the Perak River. In Nor Hashimah Jalaluddin, Hiroyuki Suzuki and Mitsuaki Endo (eds.) *Proceedings of the sixth International Conference of Asian Geolinguistics*, 118–129. doi: https://doi.org/10.5281/zenodo.17204650

regional community). Social dialects relate to linguistic features determined by factors such as social status, education, occupation, age, or gender. Geographical dialects, also referred to as regional dialects, are those spoken by native speakers within a specific area and are closely associated with the speakers' residential zones. This is supported by Matthias Urban (2020), who observed phonological differences (i.e., pronunciation) among communities residing in highland and lowland areas. The analysis of dialect distribution may involve elements such as phonological variation, affixation, and unique expressions employed by specific communities.

According to Trudgill (2011), in sociolinguistic typology, isolated areas tend to exhibit distinct linguistic features when compared to areas with high levels of interaction and communication. This view can be applied to upstream and downstream communities along a river corridor. As highlighted by Sau Heng Leong in J. Kathirithamby-Wells and John Villiers (1990), trade networks in Peninsular Malaysia historically operated along upstream-downstream models, where small-scale farmers in upstream regions transported their goods to collecting centres downstream. Consequently, riverine networks play a critical role in influencing the spread and diversity of dialects in such areas. Hence, this study seeks to identify dialect distribution patterns and analyze dialect transition zones along the Perak River.

2. Perak River

The Perak River is the second-longest river in Peninsular Malaysia, traversing several major districts in the state of Perak, including Hulu Perak, Kuala Kangsar, Kinta, Perak Tengah, and Hilir Perak. Historically, the river has been central to migration patterns and the socio-economic development of local communities, serving as a cradle for early human civilization along its banks. According to Zuliskandar et al. (2015), Kinta was once a major destination for immigrants during the 19th century due to its flourishing tin mining industry. This economic expansion led to the growth of towns and cities such as Ipoh, Kampar, Gopeng, Batu Gajah, Pusing, Pasir Putih, Papan, Lahat, Menglembu, Jelapang, Tasek, Bercham, and Gunung Rapat.

Civilizational progress along the Perak River is further influenced by its geographical proximity to the borders of Kedah, Southern Thailand, Kelantan, and Pahang. Social interactions arising from contact with neighboring states and external traders, such as those from Southern Thailand, have contributed significantly to the presence of diverse dialectal variants along the river. The overlapping of territorial boundaries, geographical features, and historical development in the region provides a

basis for understanding the rich and complex distribution of dialects along the Perak River corridor.

Map 1: Map of the state of Perak

Sumber: https://www.wonderfulmalaysia.com/map-state-perak-malaysia.htm

3. Research Methodology

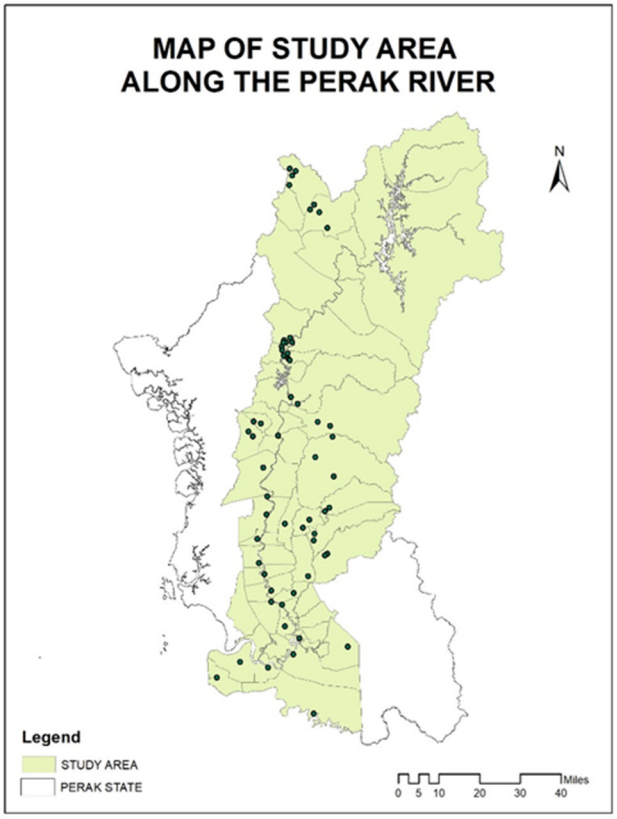
The exploration of dialect transition was conducted across villages situated along the Perak River corridor. A total of 59 villages were selected as research sites using a combination of systematic sampling and random sampling techniques. Systematic sampling refers to the selection of sample sites based on predetermined criteria, including fixed spatial intervals between villages. In contrast, random sampling involves selecting sample sites using a table of random numbers (Choong Wai Cheong, 2004).

A mixed sampling method was adopted to overcome the limitations of each individual approach. Systematic sampling may fail to consider topographical features such as rivers and hills or uneven village distribution, while random sampling may result in clustering of sample sites within certain areas, thereby overlooking broader geographical coverage. To address these issues, a combination of both methods was used, guided by several selection criteria, including:

a) Random selection of study sites.
b) Locations must be homogeneous (readily accessible) and clearly identified using GPS.
c) Sites must not include newly developed settlements (e.g., FELDA, FELCRA) or villages not inhabited predominantly by Malay speakers (e.g., Orang Asli communities).
d) Locations must be situated near the river corridor.

Fieldwork interviews were conducted with 957 informants across the study area, consisting of three demographic groups: elderly, adults, and youth. On average, each village contributed 16 informants. The lexical item /bantal/ (meaning "pillow") was selected as the primary focus for examining dialect distribution. This selection is justified by the presence of lateral consonants /-l/ and /-r/ in coda positions, which have been categorized as sporadic or unstable in the Perak River area (N. Habibah C.H. and Rahim Aman, 2020). A similar phenomenon has been observed along the Batanghari River in Sumatra, where Anderbeck K. R. (2008) noted irregular patterns of lateral consonants /-r/ and /-l/ in dialect variation.

A descriptive geolinguistic approach was employed to analyze dialect distribution and the underlying factors influencing it. This approach links the research database to dialect mapping, as suggested by Khairul Ashraaf Saari (2019). The analysis was carried out descriptively, focusing on mapping dialect distribution patterns and identifying influencing factors using Geographical Information Systems (GIS) to generate isogloss maps based on the selected lexical data.



Map 2: Map of the Study Area along the Perak River

4. Research Analysis

Historically, the Perak River has served as a vital artery for communication and trade since the era of the ancient Perak Sultanate. This river functioned as a principal route for the movement of people, goods, and cultural practices, and it was also one of the earliest sites of Malay settlement. Economic activities such as tin mining, agriculture, and fishing along the river contributed to societal diversification and dialectal interaction. This view is supported by Mior Ahmad Noor (2002), who asserted that any ruling power that managed to control a river's estuary and course could consequently dominate the socio-economic activities of the surrounding region. Additionally, the development of road infrastructure and modern communication systems has facilitated greater social mobility and enhanced intercommunity interaction.

Fieldword enabled classification of seven phonetic variants of the lexical item bantal (pillow) along the Perak River corridor. These variants are as follows: [bɑntaj], [bɑtɑ], [bɑntɛ], [bɑnte], [bɑtɑl], [bɑntɑ], and [bɑntɑl]. Table 1 below presents the frequency distribution of each variant across the respective districts along the Perak River.

Table 1: Frequency Ratio of Phonetic Variant Distribution along the Perak River

NO	DISTRICT	VARIANT OF /bantal/ L1- bɑntaj L2- bɑtɑ L3- bɑntɛ L4- bɑnte L5- bɑtɑl L6- bɑntɑ L7- bɑntɑl							FREQUENCY RATIO
		L1	L2	L3	L4	L5	L6	L7	
1.	Hulu Perak	/	/	/			/	/	71.43%
2.	Kuala Kangsar			/	/		/		42.86%
3.	Kinta	/		/		/		/	57.14%
4.	Perak Tengah	/		/			/	/	57.14%
5.	Hilir Perak			/				/	28.57%

The table demonstrates that the Hulu Perak district exhibits the most diverse dialectal spread, with five variants recorded at a frequency of 71.43%. This is followed by Kinta and Perak Tengah, each with four variants at 57.14%. The Kuala Kangsar district recorded three variants at 42.86%, while Hilir Perak exhibited the least variation with two variants at 28.57%.

4.1. Phonetic Variants Analysis

4.1.1. Phonetic Variant [bɑntaj]

This variant is associated with the Kedah dialect, evident in the phonological change from /-al/ to [-ɑj] at the surface level. As Asmah Haji Omar (2015) explains, in the Kedah dialect, the lateral /l/ typically appears only in prevocalic or intervocalic positions, not in coda. Therefore, the Standard Malay form /-al/ is realized as [-ɑj] in this dialect.

The variant was notably present in Hulu Perak, including Kampung Simpang Pulai, and to a lesser extent in Perak Tengah and Kinta. Its discovery in Kampung Simpang Pulai offers a revision to Nor Hashimah Jalaluddin's (2018) view, which posited that Kedah dialect influence in Perak occurred primarily via riverine and coastal routes. In contrast, this study suggests that dialect diffusion also occurred via land-based routes,

supported by the absence of significant topographical barriers such as highlands or dense forests.

Kinta, as a key destination for immigrants during the 19th-century tin mining era, witnessed rapid urban growth and settlement development, facilitating the inflow of non-local dialects. This supports the argument that modernization and intercultural contact contributed to the integration of the [bɑntaj] variant in the region.

4.1.2. Phonetic Variant [bɑtɑ]

This variant, predominantly found in Hulu Perak, is attributed to the Patani dialect influence. According to Nur Habibah C. H. and Rahim Aman (2020), the variant emerged from the migration of Southern Thai Malay (Patani) communities into the upper Perak River region. Phonologically, this variant shows lateral deletion and nasal deletion in homorganic nasal-obstruent clusters.

Its concentrated presence in Hulu Perak, which borders Southern Thailand, indicates linguistic diffusion due to cross-border interaction. Sociologically, the variant reflects historical migration driven by conflict and oppression in Thailand, resulting in the resettlement of Patani Malays in areas such as Lenggong and Grik (Nor Hashimah Jalaluddin, 2015).

4.1.3. Phonetic Variant [bɑntɛ]

Statistical data show that [bɑntɛ] is widespread along the Perak River and is present in all districts except Kerian. This variant reflects segment coalescence, where the final syllable /-al/ becomes [ɛ]. Zaharani Ahmad (2006) identifies this as a hallmark of the Perak dialect, representing a systematic phonological process.

The variant is especially dominant in Parit and Kuala Kangsar, with the latter considered its primary center of distribution. Historically, Kuala Kangsar functioned as Perak's administrative and trade capital, and its strong local identity facilitated intergenerational retention of this dialect. The variant was institutionalized as part of administrative communication, further reinforcing its linguistic stability.

4.1.4. Phonetic Variant [bɑnte]

Found primarily in Kuala Kangsar, especially in Kampung Rambong and Kampung Laneh, this variant arises in a dialectal transition zone between Kedah and Perak dialects (Asmah Haji Omar, 2015). The district's location between northern and southern Peninsular Malaysia supports its transitional role.

The variant [bɑnte] appears to derive from [bɑntɛ] (Perak) and [bɑntaj] (Kedah). The phonological change from [-ɛ] to [-e] reflects vowel narrowing, where [-e], with

the features [+front, +mid-close], replaces [-ɛ], [+front, +mid-open]. This development is consistent with Schmidt's (1871) wave theory, which describes the gradual spread and interaction of linguistic features across dialect zones.

4.1.5. Phonetic Variant [bɑtɑl]

This variant, minimally present in Hulu Perak, represents another transition form, situated between Standard Malay [bɑntɑl] and the Patani dialect [bɑtɑ]. As Ismail Hussein (1978) notes, these two dialects exert significant influence on the region.

The presence of [bɑtɑl] may be attributed to ethnic convergence and historical events, particularly the Pangkor Treaty of 1874, which brought about British administrative influence. This resulted in the spread of Standard Malay and its interaction with existing dialects, contributing to the emergence of hybrid forms such as [bɑtɑl] (Lizawati Ramli, 2015).

4.1.6. Phonetic Variant [bɑntɑ]

This variant appears sporadically along the Perak River in districts such as Hulu Perak, Kuala Kangsar, Kinta, and Perak Tengah. It represents a blended form of [bɑntɑl] (Standard Malay) and [bɑtɑ] (Patani dialect), consistent with Ismail Hussein's (1978) observation of Perak's linguistic duality: influenced by both northern Patani and southern Johor dialects.

Its phonological features developed along the river's upstream-downstream gradient: [bɑtɑ] in upstream Hulu Perak, transitioning into [bɑntɑ] in midstream areas, and finally into [bɑntɑl] in downstream Hilir Perak.

4.1.7. Phonetic Variant [bɑntɑl]

This is the Standard Malay form, exhibiting no phonological transformation. It is predominantly found in Hilir Perak, located downstream. The spread of this variant is driven primarily by migration from Selangor and Kuala Lumpur, where Standard Malay is the norm.

Mohd Fadzil Abdul Rashid et al. (2012) reported consistent migration into Perak—especially Batang Padang, Perak Tengah, and Hilir Perak—from 1980 to 2000. Smaller-scale migration also affected Manjung, Kerian, and Kuala Kangsar. These movements support the presence of [bɑntɑl] in southern Perak, demonstrating how external migration influences local dialects through linguistic displacement.

5. Research Findings

The distribution of dialects in a given area is closely influenced by social and geographical factors. Social factors encompass migration, urban development, language standardization, and historical events that shape linguistic practices. Geographical factors include physical features such as mountain ranges, rivers, and political boundaries. Together, these factors shape dialectal diffusion along the Perak River.

Based on fieldwork, the researcher identified several villages exhibiting high dialectal diversity (refer to Table 2). This variation arises from dialectal convergence zones, also referred to as transition zones. According to Chambers and Trudgill (1998), a transition zone is a community where neighboring dialects coexist and interact despite geographic barriers. These zones are linguistically dynamic, continually influenced by surrounding dialects. The Perak River study area, bordered by Kedah, Southern Thailand, Kelantan, and Pahang, serves as a strategic linguistic corridor for such interactions.

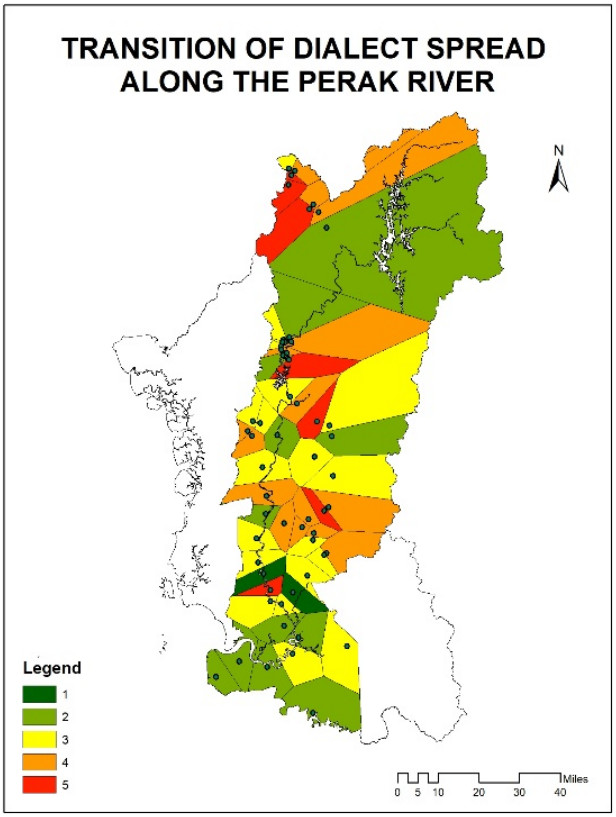
Table 2: Diversity of Phonetic Variant Distribution along the Perak River

District	Village Name	[bantaj]	[bata]	[bantɛ]	[bante]	[batal]	[banta]	[bantal]
HULU PERAK	Kg Pahit Tengah	18%	27%	0	0	18%	18%	18%
	Kg Luat Tengah	0	29%	14%	0	14%	14%	29%
	Kg Simpang Pulai	44%	11%	22%	0	0	11%	11%
	Kg Selarong	17%	17%	0	0	17%	17%	33%
KUALA KANGSAR	Kg Perlop 1	11%	5%	42%	21%	0	0	21%
KINTA	Kg. Tualang Tujuh	22%	0	22%	11%	0	11%	33%
PERAK TENGAH	Kampung Baru Pulau	5%	0	33%	14%	0	10%	38%

The data above confirms the presence of five phonetic variants within single villages, with Hulu Perak demonstrating the highest degree of variation, as seen in Kg Pahit Tengah, Kg Luat Tengah, Kg Simpang Pulai, and Kg Selarong. Kuala Kangsar,

Kinta, and Perak Tengah also show variation in individual villages. Map 3 illustrates these findings visually using color-coded labels .

This pattern further highlights Hulu Perak as a key dialect transition zone, supported by its proximity to Southern Thailand, Kelantan, and Kedah. Ongoing interaction among these border communities has led to linguistic convergence. In addition to geography, migration plays a critical role. Asmah Haji Omar (2015) noted the presence of Patani migrants in Larut Matang and Hulu Perak, who fled regional conflict and settled near river basins and highlands in the Titiwangsa Range (Nor Hashimah Jalaluddin, 2015).



Map 3: Dialect Transition Zones along the Perak River

The spread of dialects along the Perak River demonstrates significant diversity, especially in upstream and midstream areas. Dialect diffusion via language

displacement can lead to overlapping dialect zones and even language replacement. Asmah Haji Omar (2022) notes that unchecked diffusion may result in language shift, and when weaker community languages cannot compete with dominant ones, this may lead to language death.

6. Conclusion

The dialect transition zones identified along the Perak River serve as evidence of the linguistic diversity present in Malaysia. The transitions observed reflect a gradual process of dialectal change, supported by the emergence of new phonetic variants resulting from contact and interaction among different dialects along the river corridor. Both social and geographical factors have significantly shaped the development of dialect variation within the study area. Social factors include migration, modernization, urban expansion, and increased community interaction, while geographical factors encompass riverine networks, district boundaries, and topographical features. These factors jointly contribute to the dynamic nature of dialect distribution. The findings suggest that dialect diversity does not occur randomly but is shaped by historical, social, and geographical influences. The observations from the Perak River region offer insights that may inform broader studies on dialectal variation, especially in areas with high levels of social mobility and interregional contact. In summary, the dialect transitions observed along the Perak River provide a valuable reflection of Malaysia's linguistic richness, highlighting the ongoing evolution of dialects in response to sociocultural and environmental dynamics.

References

Anderbeck K. R. 2008. *Malay Dialects of the Batanghari River Basin (Jambi, Sumatra)*. SIL International (SIL e-Books).

Asmah Haji Omar. 2015. *Susur Galur Bahasa Melayu*. Edisi ke-2 Cetakan ke-2. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Asmah Haji Omar. 2022. *Linguistik Sejarah: Pertumbuhan, Perkembangan dan Penyebaran Bahasa*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Chambers, J. K. & Trudgill, P. 1998. *Dialectology*. England: Cambridge University Press.

Choong, W. C. 2004. *Pencerapan Informasi Bertema dari Teknologi Remote Sensing: Satu Peneguhan Nilai dengan Teknik GIS*. Tesis Sarjana. Universiti Malaya.

Ismail Hussein. 1973. Malay dialects in the Malay Peninsular. *Nusantara* 3.

Khairul Ashraaf Saari. 2019. *Analisis Geolinguistik Variasi Leksikal Bahasa Teras di Negeri Johor*. Tesis Sarjana, Universiti Kebangsaan Malaysia.

Lizawati Ramli. 2015. *Konflik dan Campur Tangan Asing: Raja Abdullah dan Ngah Ibrahim 1860-1875*. Tesis Sarjana Sastera. Universiti Sains Malaysia.

Mior Hamzah, M. A. N. 2001. Hubungan Melayu-Siam: Melihat kepada persoalan sempadan di Kedah. *Jurnal Pembangunan Sosial* 3.

Mohd Fadzil Abdul Rashid, Mohd Roswodi Mat Zin, Norhazlan Haron. 2012. Pola dan penyerakan migrasi di negeri Perak dari 1980–2000 dan penelitian awal implikasinya. Dalam *Proceeding ICITSBE 2012*.

Nofiana S. Putri Malahati. 2021. Comparison of Acehnese Pidie dialect variations between Acehnese Nagan Raya dialects of Acehnese people in Peukan Baro District. *Budapest International Research and Critics Institute (BIRCI-Journal) Humanities and Social Sciences* 4(2).

Nor Hashimah Jalaluddin. 2015. Penyebaran dialek Patani di Perak: Analisis geolinguistik. *Jurnal Antarabangsa Dunia Melayu* 8(2).

Nor Hashimah Jalaluddin. 2018. Dialek Melayu di Perak: Analisis geolinguistik. *International Journal of the Malay World and Civilisation* 6(2): 69–82.

Nur Habibah & Rahim Aman. 2020. Varian Hulu Perak Utara: Fenomena di Gerik dan Pengkalan Hulu. *Jurnal Melayu* 19(2).

Sau Heng Leong. Collecting Centres, Feeders Points, and Entreport in The Malay Peninsula, circa 1000 B.C. – A.D 1400. Dalam J. Kathirithamby Wells & John Villiers, *The Southeast Asian Port and Polity: Rise and Demise*. Singapore University Press.

Schmidt, Johannes. 1871. *Zur Geschichte des indogermanischen Vocalismus, Teil 1*. Weimar: H. Böhlau.

Trudgill, P. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford University Press.

Yule G. (1996) The Study of Language, London Cambridge University,Press.

Zaharani Ahmad & Teoh Boon Seong. 2006. *Fonologi Autosegmental*. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Zaharani Ahmad. 2006. Kepelbagaian dialek dalam Bahasa Melayu: Analisis tatatingkat kekangan. *e-BANGI: Jurnal Sains Sosial dan Kemanusiaan* 1(1): 26.