A Preliminary Survey of Linguistic Areas in East Asia Based on Phonological Features

Ian Joo and Yu-Yin Hsu
The Hong Kong Polytechnic University

Abstract
Previous studies of linguistic areas have often adopted a mainly top-down approach, by first hypothesizing the existence of a linguistic area and then seeking the common linguistic features of that hypothetical area in order to justify its existence. In order to identify linguistic areas in East Asia in a different way, we adopt a mainly bottom-up approach by first investigating the values of the linguistic feature parameters of languages spoken in East Asia and then calculating those values to locate geographical clusters of languages sharing a certain degree of cross-family similarity. Based on 19 phonological features as binary parameters of 52 sample languages of East Asia, we visualize their within-family and cross-family similarities. Many of these similarities confirm the previous theories concerning linguistic areas, such as the Mainland Southeast Asia or the Qinghai-Gansu linguistic area. However, we also demonstrate some similarities that have received less attention thus far, namely between Ryukyuan and southern Sinitic languages.

Key words
Linguistic area, East Asia, phonology

1. Introduction

In this study, we demonstrate the preliminary findings of our project aimed at identifying linguistic areas in East Asia. A linguistic area is an area home to geographically close languages sharing a high proportion of linguistic features not due to genealogical relatedness but due to historical contact. East Asia is defined as the area consisting of China, Japan, Korea, and Mongolia. A calculation of the Simple Matching Coefficient (Sokal and Michener 1958) based on the binary parameters of 19 phonological features reveals the cross-family similarities among languages in East Asia. The patterns of cross-family similarity confirm previous theories regarding Mainland Southeast Asia and Qinghai-Gansu as linguistic areas. The results also point to less studied similarities, such as between Korean and Ainu or between Ryukyuan and southern Sinitic languages.

2. Background

A **linguistic area** (or **sprachbund**) is a geographical area home to multiple languages that share a number of linguistic features due to historical contact and not genealogical relationship (cf. Thomason 2000). Well-known examples of linguistic area include Mainland Southeast Asia (Enfield 2018), Standard Average European (Haspelmath 2001), Mesoamerica (Campbell, Kaufman, and Smith-Stark 1986), and Ethiopia (Bisang 2006).

The languages spoken in East Asia belong to several different linguistic areas. The southwestern Chinese provinces of Yunnan, Guizhou, and Guangxi belong to the Mainland Southeast Asian linguistic area (Enfield 2018) and share many linguistic traits with the Indochinese peninsula. The East Asian region north of the Yellow River belong to the Northeast Asian linguistic area, which extends northward to eastern Siberia and the Russian Far East (Hölzl 2018). The Bodic, Sinitic, and Mongolic languages spoken in Qinghai and Gansu province form together the Qinghai-Gansu linguistic sprachbund (Xu 2017, Ch. 1).

However, many studies on linguistic areas within or overlapping with East Asia have relied on a top-down, theoretical approach, rather than a bottom-up, data-driven approach. In many studies, the existence of a linguistic area was first postulated, and then the features characteristic of that area were sought after in order to justify the existence of that linguistic area. This approach, while convenient, carries the risk of confirmation bias: By postulating the existence of a certain linguistic area, the researcher is tempted to focus on features that are shared by the languages in that area, while neglecting the features that are not.

One way to avoid the confirmation bias would be to first remain agnostic about the presence of areas, investigate the geographical distribution of a given group of features, and then use that data to examine whether a certain geographical space shares a high number of linguistic features, thereby concluding the existence of a linguistic area. This would be the bottom-up, data-driven approach.

Several studies have aimed to justify linguistic areas pertaining to East Asia using a data-driven approach based on the World Atlas of Linguistic Structure (WALS, Dryer and Haspelmath 2013). WALS is a database of 192 linguistic features obtained through surveys of more than a thousand languages across the world. Whitman (2016), based on a number of features present in WALS, suggested that Northeast Asia could be justified as a linguistic area. Comrie (2007), based on 21 features selected from WALS, observes that these features point to common patterns in Mainland Southeast Asian languages, whence he concludes that Mainland Southeast Asia is a coherent linguistic area.

WALS, however, has its limits for investigating the linguistic borders within and across East Asia. One of those limits is that East Asian languages are underrepresented in WALS. Only three

features are investigated with at least 50 East Asian sample languages in WALS: *Order of Subject and Verb*, *Order of Object and Verb*, and *Order of Adjective and Noun*. Not only are these three features limited to the topic of word order, but the corresponding East Asian sample languages are also overrepresented by those spoken in the southern part of East Asia compared those spoken in the northern part.

Yurayong and Szeto (2020) have investigated 40 features of Japonic, Koreanic, Sinitic, and other neighboring languages, thereby demonstrating the typological similarity of the Japonic-Koreanic group. Szeto and Yurayong (2021) have investigated the 30 linguistic features Sinitic, Southeast Asian, and northern East Asian languages, whence they concluded the southern-northern division within the Sinitic branch. Both studies have an impressive number of sample languages and an adequate amount of linguistic features. However, as both studies were focused on the specific topic of Japano-Koreanic (Yurayong and Szeto 2020) and Sintic (Szeto and Yurayong 2021), it would be interesting to approach East Asia as a whole from a bird's-eye view.

Thus, a more balanced sample of East Asian languages, paired with an equally balanced set of linguistic features, is needed in order to investigate the linguistic areas within and overlapping with East Asia.

3. Research question

The research question is as follows: Which regions in East Asia have multiple languages that belong to different language families yet share many linguistic features? In other words, what linguistic areas exist in East Asia?

4. Methodology

As this article reports a preliminary stage of this ongoing study, we will first conduct analysis only based on phonological features. In the future, we plan to include lexico-semantic and morphosyntactic features in our analysis as well.

Fifty-two sample East Asian languages were studied, with genealogical and geographic diversity taken into consideration, as well as data accessibility. Table 1 and Figure 1 list the sample languages.

Table 1. List of sample languages

Family	Language
Ainu	Ainu
Sino-Tibetan	Amdo Tibetan
Austronesian	Atayal
Sino-Tibetan	Bai
Austroasiatic	Bugan
Sino-Tibetan	Cantonese
Sino-Tibetan	Changshanese
Mongolic	Dagur
Tai-Kadai	Dong
Sino-Tibetan	Drung
Japonic	Dunan
Mongolic	East Yugur

Sino-Tibetan	Ersu
Tungusic	Evenki
Japonic	Hachijo
Sino-Tibetan	Hakka
Tai-Kadai	Hlai
Sino-Tibetan	Hohhot
Sino-Tibetan	Hokkien
Sino-Tibetan	Idu
Japonic	Irabu
Hmong-Mien	Iu Mien
Japonic	Japanese
Turkic	Kazakh
Sino-Tibetan	Khams Tibetan
Sino-Tibetan	Kman
Koreanic	Korean
Sino-Tibetan	Lhasa Tibetan
Tungusic	Manchu
Sino-Tibetan	Mandarin
Mongolic	Mongolian
Mongolic	Monguor
Tungusic	Nanai
Sino-Tibetan	Nuosu
Tungusic	Oroqen
Sino-Tibetan	Qiang
Austronesian	Rukai
Turkic	Salar
Indo-European	Sarikoli
Sino-Tibetan	Shanghainese
Austronesian	Tsat
Sino-Tibetan	Tujia
Turkic	Tuvan
Turkic	Uyghur
Austroasiatic	Wa
Sino-Tibetan	Waxiang
Turkic	West Yugur
Hmong-Mien	Xong
Austronesian	Yami
Sino-Tibetan	Yichun
Japonic	Yuwan
Tai-Kadai	Zoulei

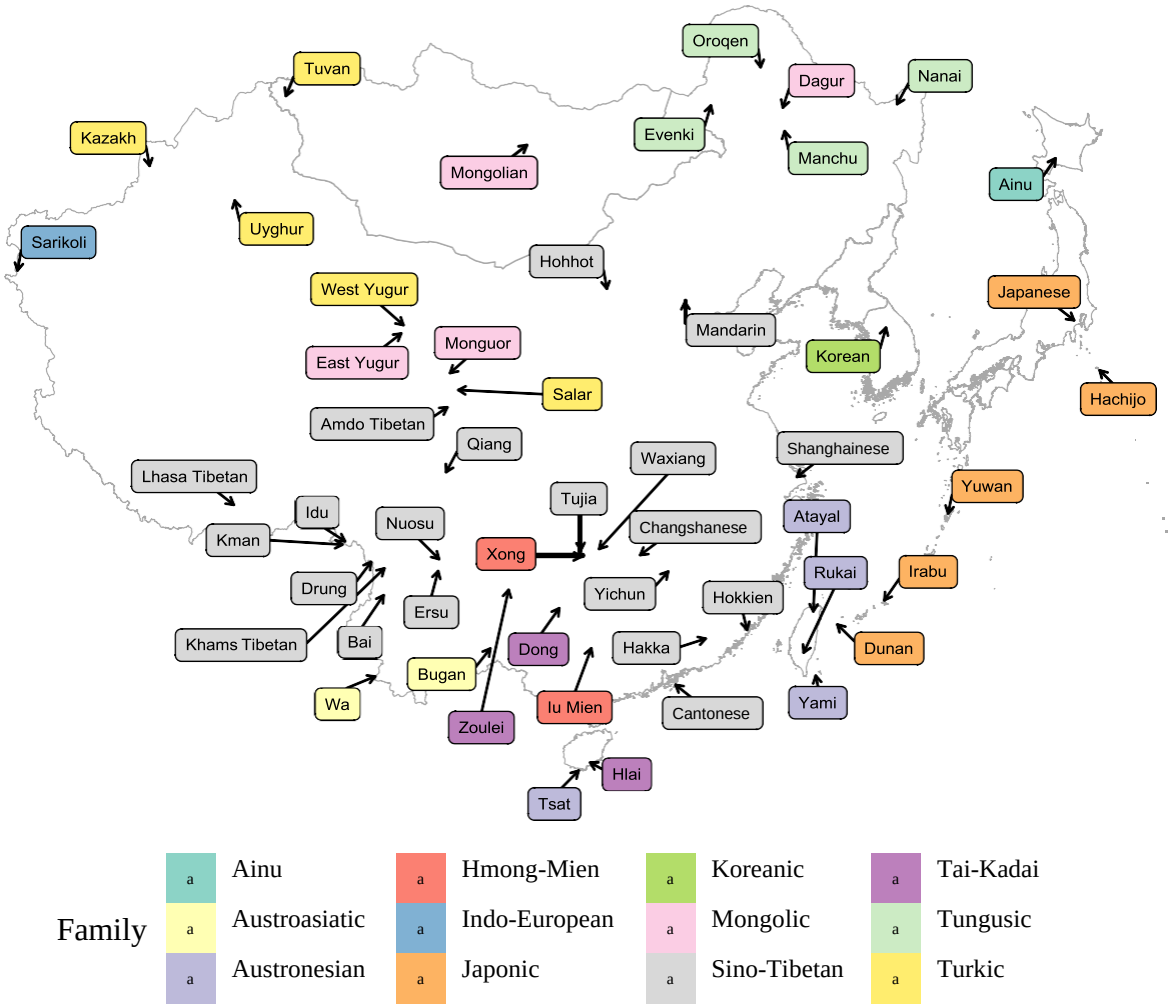

Figure 1: Distribution map of sample languages

We investigated the existence or absence of 19 phonological features in the 52 sample languages. The 19 phonological features and their criteria are listed in Table 2. Five of the features can also be found in WALS (*Consonant Inventories*, *Front Rounded Vowels*, *Uvular Consonants*, *Vowel Nasalization*, and *Vowel Quality Inventories*).

The features were chosen based on their distributive incompleteness within East Asia. A feature would be relevant for distinguishing languages within East Asia only if it were present in some (but not all) East Asian languages. A feature such as *having a rounded vowel* would be irrelevant for our analysis because all East Asian languages have a rounded vowel. A feature such as *having a click consonant* would also be irrelevant, since no East Asian language (to our knowledge) has a click consonant. Thus, it is necessary to first make judgments based on personal knowledge of whether a feature is incompletely distributed among East Asian languages before including it into the sample features. While this part of the methodology is top-down and not bottom-up, it does not contradict to our goal of avoiding confirmation bias, since we only judge whether a feature is incompletely distributed within East Asia and not whether it is concentrated in certain regions of East Asia.

Table 2. List of phonological features

Feature	Criterion
Consonant Clusters	Permits consonant clusters within a syllable?
Consonant Inventories	Has more than the median number of consonant phonemes?
Coronal Sonorants	Has both L ([+lat +cor +son]) and R ([-lat +cor +son -nas])?
Falling Diphthongs	Permits a falling diphthong within a syllable?
Front Rounded vowels	Has a front rounded vowel as a
phoneme? Glottal Stop	Has a glottal stop as a phoneme?
Labiodental Fricatives	Has a labiodental fricative as a phoneme?
Long Vowels	Has phonemic vowel length distinction?
Palatal Nasal	Has a palatal nasal as a phoneme?
Plosive Codas	Allows stops at the coda position?
Retroflex Consonants	Has a retroflex consonant as a
phoneme? Tone	Has phonemic tone?
Uvular Consonants	Has a uvular consonant as a
phoneme? Velar Fricatives	Has a velar fricative as a phoneme?
Velar Nasal Onset	Allows velar nasal at onset
position? Voiced Plosives	Voice distinction in plosives?
Voiceless Glottal Fricative	Has a voiced glottal fricative as a phoneme?
Vowel Nasalization	Has a nasal vowel as a phoneme?
Vowel Quality Inventories	Has more than the median number of vowel phonemes?

For example, we judge that the feature *Consonant Clusters* (whether a language allows consonant clusters in a syllable) is incompletely distributed in East Asia based on personal knowledge that some (but not all) East Asian languages have this feature. But for selecting this feature, we do not take into consideration whether this feature is concentrated in certain regions within East Asia.

All the 19 features are binary, their value being either 0 (absence) or 1 (presence). The features borrowed from WALS that are not binary, such as *Consonant Inventories* (the size of the consonant inventory), were also converted into binary features (having more than the median number of consonants within the sample group or not). We calculated the Simple Matching Coefficient of the 19 features of each pair of languages that are within a reasonable geographical distance. We drew a line between two languages if their geographical coordinates are within 1,500km distance and their 19 binary features show a Simple Matching Coefficient greater than 0.7.

5. Results and discussion

Figure 2 shows the preliminary results. The blue dotted lines represent Simple Matching Coefficient greater than 0.7 between languages belonging to the same family, whereas the red solid lines represent that between languages belonging to different families. The thicker and more opaque a line, the greater the Coefficient.

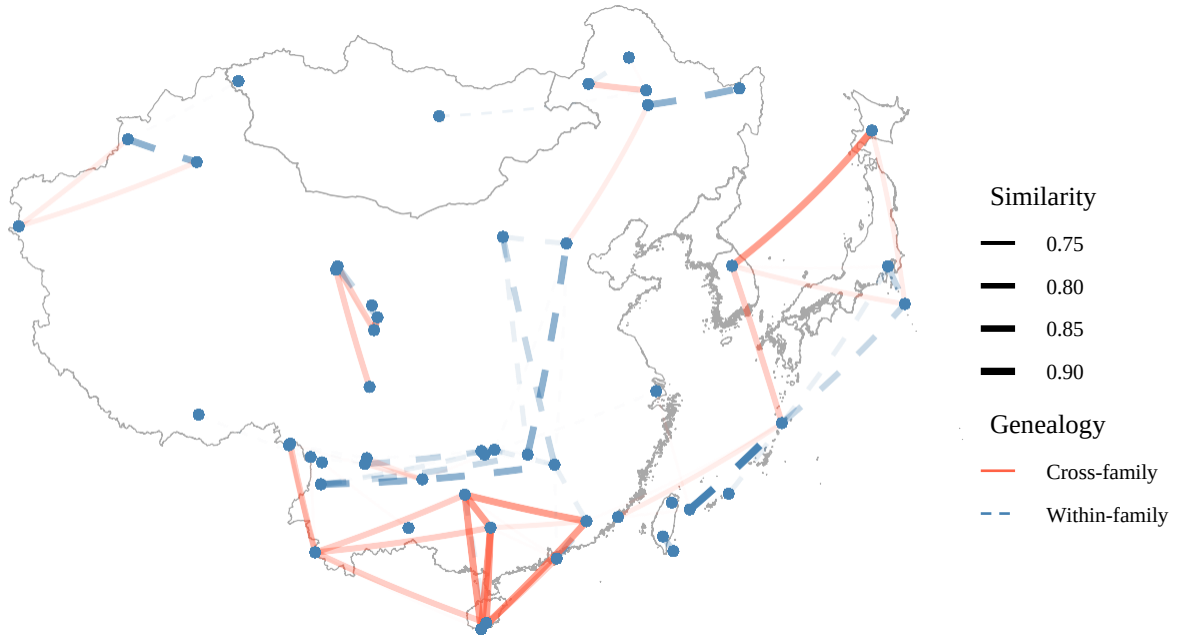

Figure 2: Connections representing phonological similarities between geographically close languages, across or within families

6. Discussion

The languages spoken in the Chinese provinces of Qinghai and Gansu share strong cross-family connections, as predicted by previous studies on the Qinghai-Gansu linguistic area (cf. Xu 2017, Ch. 1). Languages in southwestern China are generally densely connected to each other, supporting the previous theories of the Mainland Southeast Asian linguistic area (cf. Enfield 2018). Manchu is connected to Mandarin, in line with the historical contact between these two languages. Korean is most strongly similar to Ainu, and less so to Japonic languages. Sarikoli, an Indo-European language spoken in northwestern China, shows some connection to Turkic languages (Kazakh and Uyghur) spoken nearby. Formosan languages show no similarity to Ryukyuan languages de- spite their geographical proximity, in line with a genetic study demonstrating no genetic similarity between Taiwanese aboriginals and Ryukyuan islanders (Matsukusa et al. 2010). On the other hand, the Ryukyuan languages show some similarity with some southern Sinitic languages, such as Shanghainese, Hokkien, and Hakka.

7. Conclusion

In this paper, we have presented the first step of our analysis aiming to reveal linguistic areas in East Asia. Even though these observable patterns must be approached with caution given the preliminary stage of the data, they offer a promising outlook for our ongoing project and lead us to believe that, with more features (other than phonological) examined, we will have a clearer view of the linguistic areas within East Asia.

References

Bisang, Walter. (2006) Linguistic areas, language contact and typology: Some implications from the case of Ethiopia as a linguistic area. In Yaron Matras, April McMahon, and Nigel Vincent

(eds.), *Linguistic Areas: Convergence in Historical and Typological Perspective*. 75–98. London, UK: Palgrave Macmillan. doi: 10.1057/9780230287617_4.

Campbell, Lyle, Terrence Kaufman, and Thomas C Smith-Stark. (1986) Meso-America as a linguistic area. *Language* 62, 530–570.

Comrie, Bernard. (2007) Areal typology of mainland Southeast Asia: What we learn from the WALS maps. *Manusya: Journal of Humanities* 10.3, 18–47.

Dryer, Matthew S. and Martin Haspelmath (eds.). (2013) *WALS Online*. URL: https://wals. info/.

Enfield, Nick J. (2018) *Mainland Southeast Asian Languages: A Concise Typological Introduction*. Cambridge, UK: Cambridge University Press.

Haspelmath, Martin. (2001) The European linguistic area: standard average European. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible (eds.), *Language Typology and Language Universals* Vol. 2, 1492–1510. Berlin and Boston: de Gruyter.

Hölzl, Andreas. (2018) *A Typology of Questions in Northeast Asia and beyond: An Ecological Perspective*. Studies in Diversity Linguistics. Berlin: Language Science Press.

Matsukusa, Hirotaka, Hiroki Oota, Kuniaki Haneji, Takashi Toma, Shoji Kawamura, and Hajime Ishida. (2010) A genetic analysis of the Sakishima islanders reveals no relationship with Taiwan aborigines but shared ancestry with Ainu and main-island Japanese. *American Journal of Physical Anthropology* 142.2, 211-223.

Sokal, Robert R and Charles D Michener. (1958) *A statistical method for evaluating systematic relationships*. Lawrence: University of Kansas.

Szeto, Pui Yiu and Chingduang Yurayong. (2021) Sinitic as a typological sandwich: revisiting the notions of Altaicization and Taicization. *Linguistic Typology*. (ahead of print) https://doi.org/10.1515/lingty-2021-2074/html

Thomason, Sarah Grey. (2000) Linguistic areas and language history. *Studies in Slavic and General Linguistics* 28, *Languages in Contact*, 311-327.

Whitman, John (ジョン・ホイットマン). (2016) Tōhoku azia gengo chiiki-no ichizukeni nukete (東北アジア言語地域の位置付けに向けて). [On the Northeast Asia as a Linguistic Area]. 国語研プロジェクトリビュー・ *NINJAL Project Review* 6, 69-82.

Xu, Dan. (2017) *The Tangwang language: An interdisciplinary case study in Northwest China*. Cham: Springer.

Yurayong, Chingduang and Pui Yiu Szeto. (2020) Altaicization and de-Altaicization of Japonic and Koreanic. *International Journal of Eurasian Linguistics* 2.1, 108-148.

Email addresses: ian.joo@connect.polyu.hk

yu-yin.hsu@polyu.ed