CONSTRUCTION OF A COMPARATIVE DICTIONARY OF SINITIC AND SINOXENIC LANGUAGES COGNATES PHONOLOGY

Louis Lecailliez
Graduate School of Informatics, Kyoto University, Japan
louis.lecailliez@outlook.fr

Abstract

About a dozen languages in East-Asia share an important number of cognates because of a common origin (Sinitic family) or extensive borrowings (Sinoxenic languages). This is a useful fact for a speaker who masters one of them and want to learn another. In a bilingual or multilingual dictionary, lexicographic information can be compared but the burden of analysis is placed on the user. This paper describes the construction of a dictionary of comparative phonology of cognates in Sinitic and Sinoxenic languages that targets learners of any of the languages it contains (presently Japanese, Standard Chinese, Taiwanese Southern Min and 6 Hakka dialects). The main dictionary's goal is to make explicit phonological similarities and differences in synchrony between cognates and teach non-obvious phoneme correspondence rules in-between those languages. We expose the theoretical framework and detail the relevant issues and their solutions. In particular, the level of representation (phonetic vs phonemic) and the implication of considering the union set of phonemes of multiple languages are discussed. Practical issues such as dealing with the different scripts and romanizations are also addressed. A comparison algorithm derived from the method of consonant classes from historical comparative linguistic is presented. Finally, we illustrate the planned output with the current prototype of an entry, which make use of the comparison algorithm for displaying data. We conclude on possible future derivate works, enabled by the digital nature of the project, that is fully automated and relies on open-data lexical resources.

Keywords: cognates, learner's dictionary, comparative phonology, multilingual dictionary, language learning

1 Introduction

Learning a language is an activity that can yield numerous benefits on professional and personal levels. In East-Asia, cultural phenomena such as the Chinese literary classics, Japanese animation or Korean popular music are powerful factors that drive people to start learning a language. Migrations and business considerations are other circumstances driving millions to learn an additional language. Moreover, interest in those languages exist in the rest of the world of well.

The task of learning a language is however not a small task and it takes a considerable amount of time and efforts to reach a stage of useful proficiency. Any time and effort spared can be re- invested in advancing to a better proficiency and lower the probability of the learner to give up. In the case of Sinitic and Sinoxenic languages (see the next two sections for a definition), there is an important number of shared cognates, that is "a linguistic form which is historically derived from the same source as another form" (Crystal, 2011). However, sound changes that occurred in each language, as well divergences in their phonology and writing system have obscured their similarity.

A dictionary of cognates would expose the proximity of pronunciation in-between languages and help fostering a multilingual environment both inside one country, and in relation to others, without having to resort to a very distant language such as English. This dictionary would explicit how the pronunciations of cognates relate to each other in different languages, which would help a learner transfer the lexicon he already. This article describes the creation of a comparative dictionary of East-Asian cognates phonology (東亞語言發音對照辭典*, Dōngyǎ yǔyán fāyīn duìzhào cídiǎn*) which aims to support that use case.

More precisely, the dictionary goals are to help learners re-use vocabulary by making explicit sound correspondences between cognates of Sinitic origin, promote multilingualism by including many languages and dialects, and provide a re-usable framework and data for future research supporting the two previously stated goals.

1.1 Sinitic Languages

The Sinitic family of languages (Handel, 2015) is part of the wider Sino-Tibetan family. It regroups a variable number of languages, depending on the linguist describing it. One of these classifications (Kwok, 2018) lists: Mandarin (官話 *guānhuà*[1]), Wu (吳語 *wúyǔ*), Yue (粵語 *yuèyǔ*) also known as Cantonese, Min (閩語 *mǐnyǔ*), Xiang (湘語 *xiāngyǔ*), Hakka (客家話 *kèjiāhuà*), Gan (贛語 *gànyǔ*), Jin (晉語 *jìnyǔ*), Hui (徽語 *huīyǔ*) and Pinghua 平話 (*pínghuà*). All these languages share traits such as being tonal languages and having a common syllable structure (Wee & Li, 2015).

1.2 Sinoxenic Languages

The so-called Sinoxenic (Martin, 1953) languages do not form a single family. Instead, the term designates languages which share the common characteristics of having heavily borrowed vocabulary from Middle Chinese; Late Middle Chinese in the case of Korean (Lee, 1994). The languages in question are Japanese, the larger representant of the Japonic family — its other sub-family being formed by Ryukyu languages —, Korean from Koreanic family that also includes Jeju language, and Vietnamese which is part of the Austra-Asiatic > Mon-Khmer > Viet-Muong family hierarchy (Eberhard, Simons, & Fenning, 2021). It is the important amount and systemic borrowings from Chinese that distinguish Sinoxenic loadwords from sporadic and earlier borrowings (Sybesma et al., 2017). For instance, the word *ume* (梅) in Japanese, coming from Old Chinese **hmay*, is not considered a Sinoxenic borrowing since it was done earlier than the systematic borrowing period and done in isolation.

In Japanese, borrowings that happened during Middle Japanese (Early Middle Japanese: 800-1200, Late Middle Japanese: 1200-1600) from Chinese was so substantial it is qualified a “sinification” of the language by Frellesvig (2010). The number of loanwords was so considerable it brought new phonological phenomena to the language such as palatalization (Labrune, 2016) and bent the existing rules of the language that forbid /r/ at word initial (Labrune, 1993).

The systematic Sinoxenic borrowings include the borrowing of the Chinese writing system and a large corpus of texts, notably the Classics and religious literature (Buddhism). Since the Chinese characters weren’t adapted to write non-Sinitic languages, all the Sinoxenic cultures first used Classical Chinese as the language of written communication, then developed a way to write their vernacular language. Vietnamese used a combination of Chinese characters and characters coined on the model of Sinograms called *chữ nôm* for around a millennia before switching to a script based on the Roman alphabet (Phuong, 1978).

[1] In this paper, words will be glossed in Standard Chinese with *hanyu pinyin* by default, even when the word exist in other languages.

2 Related Work

2.1 Research

2.1.1 Theses on Multilingual Knowledge Transfer

Two recent doctoral theses defended in France, (Labbé, 2018) and (Goudin, 2017) disserted the transfer of knowledge from a known language to another of the same family. Labbé's work dealt with West and South-Western Slavic languages. The section 2-4 is dedicated to underlining the importance of orthographic and phonological equivalence in vocabulary, which stems from historical phonology, where he argues that those can be presented in a "synchronic fashion". This is the approach taken by the dictionary presented here: while historical phonology phenomena are the source of the existing phoneme correspondences in synchrony, making a learner study a reconstructed language and sound change laws to understand current phonological correspondences is adding a huge burden to his learning. The goal of the dictionary is to lower the amount of work for the student, not to double it, so historical reconstructions and the applicable sound changes are explicitly out of the scope of this project.

Goudin's thesis is more directly appliable to the present work since it is a reflection on the use of Sinograms (Chinese characters) as a tool for inter-comprehension between Standard Chinese, Korean and Japanese. Sadly, it is hard to know more about since the thesis isn't available online. The main difference however, is that the Chinese character is the basic unit of analysis, with radicals and pronunciations being the sub-unit of analysis. In the present work, the basic unit being listed is the lemma, with the sub-unit being the syllable.

2.1.2 Contrastive Database of Japanese and Taiwanese Pronunciations

Nakazawa, Iwaki & Koresawa (2013) constructed a comparative table of pronunciation of Chinese characters in Taiwanese Southern Min and Japanese based on the 日台大辞典 (*nittai daijiten*, Japanese-Taiwanese Grand Dictionary) dictionary. Another database was created by Sakai & Nakazawa (2017), which is based on the content of the 台日新辞書 (*tainichi shinjisho,* Taiwanese-Japanese New Dictionary). Both projects have for stated goal to help Taiwanese learners of Japanese and spread the awareness in Japan of the fact that pronunciation of *kango* (漢語, Sino-Japanese words) are more similar to Japanese in Taiwanese than in Standard Chinese. Both databases are available for download as Excel files.

The present dictionary shares the goals expressed in those two papers. The biggest difference lies in the basic unit of comparison, which is the lemma in the cognate dictionary and the Chinese character in the Japanese-Taiwanese comparison table and database. In addition, while Nakazawa et al. (2013) mentions Hakka, Cantonese Vietnamese and Korean as possible future extension of their database, Hakka is integrated from the start in the dictionary presented here and resources have been collected for the three other languages. The technical mean of distribution differs too: Excel file in one hand, a website on the other hand.

2.1.3 Research on Semantic Comparison between Japanese and Chinese

By their prominence in the Japanese language, *kango* have attracted attention of linguists and lexicographers and some works classified the proximity of those words in-between Japanese and Chinese on the semantic level.

Matsushita et al. (2017) developed a database of Japanese-Chinese *kango* comparison. The resulting database is freely accessible on the web. The database lists semantic correspondence patterns such as same, overlapping, or different meaning of the cognate pairs. Xiong & Tamaoka (2014) analyzed the semantic similarity of words made of two characters and found that ~60% of the pairs share the same exact meaning, and an additional ~29% Japanese *kango* have all the Chinese meanings, in addition to Japanese specific ones. On a larger set of 20,000 lexemes, Matsuhita et al. (2017) found a very similar percentage for the noun category: 62.3% of the *kango* and their Chinese counterpart have an identical meaning.

From those research results, it is clear that the difference of meaning in cognates will not be too problematic in the general case and that an important number of cognates are easily transferred on the semantic level. Difference in semantic is thus addressed well in research literature and in the dictionary landscape while phonology isn't. In particular, two comparative dictionaries of Japanese and Chinese have been published, one using words as entries (Wang, Xu & Kodama, 2007) and the other listing Chinese characters (Tang, 1993).

2.2 Dictionaries

2.2.1 Trilateral Cooperation Secretariat Dictionaries

The trilateral Cooperation Secretariat published a set of three dictionaries (one in Japanese, one in Chinese and one in Korean) which list 658 Chinese words. For each entry, the writing in Chinese character is given (Simplified Chinese is used), their pronunciation in *romaji* (Latin letters), *hanyu pinyin* and hangul. At least one meaning is given for an entry, which is accompanied by multiple examples given in the three languages. Each example has the same meaning. However, nothing is done in those dictionaries to explicit the correspondence or divergence of pronunciations of words in-between the three languages.

2.2.2 Proto-Indo-European Lexicon Dictionary

The Proto-Indo-European Lexicon (Pyysalo et al., 2019) has the particularity of not containing directly dictionary entries for the languages it aims to support. In fact, that would be very difficult to do given that 150 to 200 languages are projected to be included. Instead, each language encodes sound change laws with a computer technology (finite-state automaton). Entries in attested languages are generated from the PIE roots by applying successively every sound change rules; when the results divert from the attested form, it is highlighted in red the presentation. The focus is "initially" placed on etymology and more information are provided by linking to existing dictionaries present on the web.

This project, in its technical execution is very similar to the one presented here: both are starting from a small set of third-party data and are encoding linguistic facts as code to make transformations on a set of starting lexicographic data. The data displayed are for the most part computed. Comprehensive lexicographic information (such as meaning) for each language is delegated to existing dictionaries by linking to them. Presentation of data is highly customizable in the interface, albeit not all features are implemented yet.

2.2.3 German-English Etymology Dictionary

Qu (2007) describes an etymology dictionary for Chinese learners of German that have a good command of English already. The stated goal is to allow users to recognize cognates in-between German and English despite the fact "phonological and semantic evolution has concealed much of their formal similarity" and thus allow them to leverage their existing knowledge of English. In contrast to Sinitic and Sinoxenic languages, the sound changes have been more radical in Germanic, leading to cognates that significantly diverge in pronunciation and orthography. Both the Old High German (OHG) and Old English (OE) words are given for a cognate pair, making their relationship more obvious. For example (Qu, 2007): "day (<OE. dæg) – Tag (<OHG. Tag)". In addition to phonology, the dictionary gives semantic information: signposts are used to warn users about important divergence in meaning. The common point of Qu's work and the

present dictionary is that both recognize the importance of phonology of cognates for transferring existing knowledge of a subset of vocabulary to another language.

3 Methodology

Similarly, to the PIE Lexicon project, the dictionary presented here is not a dictionary produced in the traditional fashion: there are no lexicographers or users writing entries. Instead, the content of existing dictionaries is reused, transformed and aggregated to provide new functionality absent from the original dictionaries. The value of the present work lies in aggregating information from disparate sources and the highlighting the similarity and difference in cognates pronunciation in-between different languages.

3.1 Project Overview

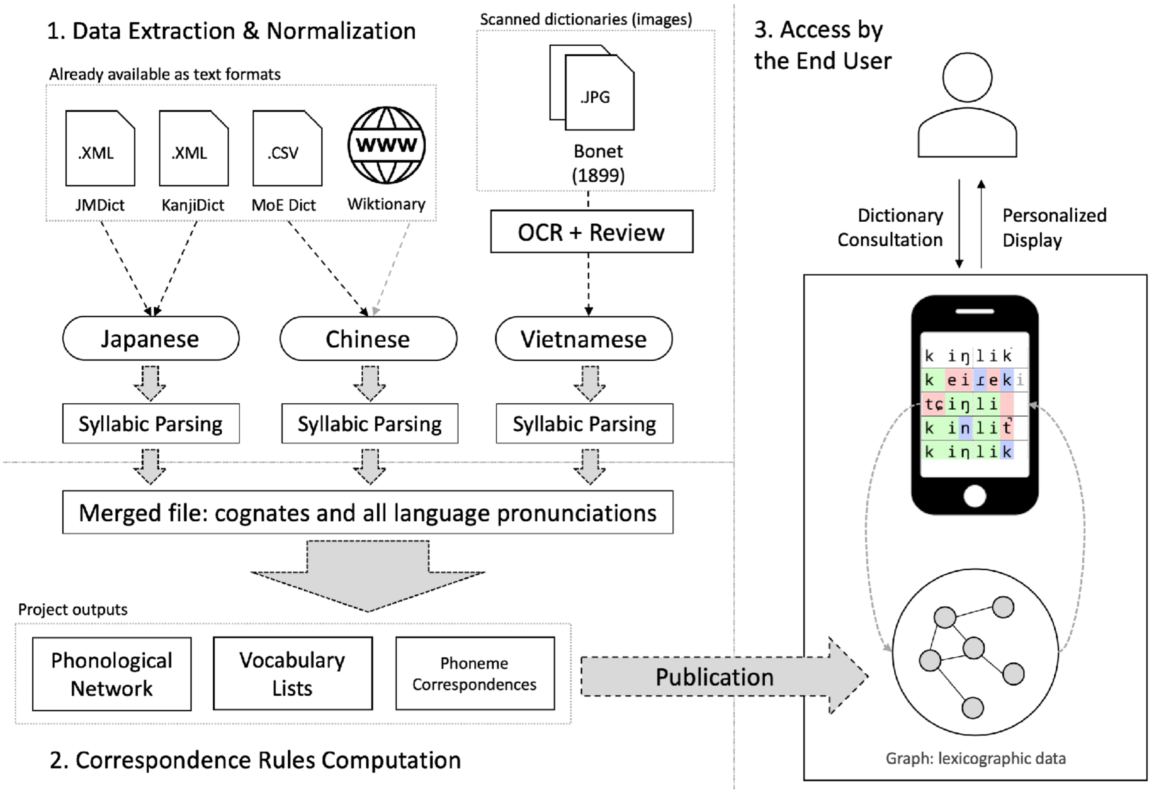


Figure 1: Project technical architecture

The project is structured as collection of data files (see Figure 1, *1. Data Extraction & Normalization*) used as input for a subsequent processing chain (*2. Correspondence Rules Computation*). The data are mainly composed of dictionaries published under open-data licenses such as JMdict (Breen, 2004), but additional resources that have a pedagogical value are used as well. Once data for a language have been collected, they are normalized to fit a common syllable format (see Section 4.3). All normalized word pronunciations are then regrouped under their cognate written in traditional Chinese characters ("Merged file" on Figure 1). The extraction and normalization phase can be arbitrary complex and is done with a collection of programs and scripts written for this purpose.

The merged file is the starting point of the different planned outputs of the project. The main output is a lexical network which is exposed to the public through a website using an existing software platform developed for another dictionary research project (Lecailliez et al., 2020). The website, which will support mobile consultation, is still under development.

An important principle of the project is the requirement that all its output can be recreated from the original data and the transformation chain. This ensure: (1) new and updated data can be retrieved from the original dictionary projects when those get updates, (2) errors introduced by the processing chain can be corrected by fixing the code involved and rebuilding the whole project and (3) the output of the project is parameterizable, which allows for different outputs based on different linguistic modeling.

3.2 Data Sources

Table 1 lists the dictionaries (column 2) used as data sources by the project for each language. The last column indicates how many entries are extracted from the source. When multiple dictionaries were collected for a language, an asterisk (*) marks the dictionary from which entries are extracted. In the case of Sinoxenic languages, the percentage indicates the proportion of extracted entries (for Sinitic languages almost every entry is extracted).

Table 1: Dictionaries used as data sources

Language	Dictionaries	Extracted Entries
Mandarin	重編國語辭典修訂本	160,658
Cantonese	CC-CANTO*, Cantonese Wordnet	105,862
Japanese	JMDict*, KanjiDict	75,351 (~66.7%)
Taiwanese	臺灣閩南語常用詞辭典, 台日大辭典*	56,466
Korean	Kengdic	38,255 (~28.6%)
Hakka	臺灣客家語常用詞辭典	14,484
Vietnamese	Dictionnaire annamite-français, Wiktionary*	5212
Central Okinawan	沖縄語辞典	2,236 (~15,4%)

These dictionaries have been created using different methodologies. Most have been complied be a team of lexicographers or linguists (in particular the ones from Taiwan) while some are crowd-sourced (JMDict (Breen, 2004), KanjiDict, Kengdic). Both CC-Canto and the Cantonese Wordnet (Sio & Morgado da Costa, 2019) employed native speakers to check the pronunciation of words. The sources are thus generally highly trustable, especially since only the pronunciations are extracted, which limit the surface of possible lexicographic issues and the problem of combining dictionaries compiled using different methodologies.

Since the Vietnamese-French dictionary (*Dictionnaire annamite-français*, 大南國音字彙合解 大法國音 *Đại Nam quốc âm tự vị hợp giải Đại Pháp quốc âm* (Bonnet, 1899)) is only available as scanned images it doesn't fit the existing processing chain and entries are extracted yet. Given the complexity of the task (Lecailliez, 2015) this part will likely need to be done manually. Japanese requires the use of a Chinese character dictionary for parsing its words readings unambiguously hence the inclusion of a kanji dictionary (KanjiDict). The Hakka dictionary contains dialects of 6 locations (四縣 *Sìxiàn*, 海陸 *Hǎilù*, 大埔 *Dàbù*, 饒平 *Ráopíng*, 詔安 *Zhàoān*, 南四縣 *Nánsìxiàn*), each of them having them than 13,000, entries save for the Zhaoan dialect which contains only 10,508 words.

Licenses of those dictionaries varies from freely reusable even commercially, to copyright free, passing by allowing reuse without modifications. Most licenses involved are a Creative Commons one. Some are incompatibles with each other or does not allow modifications to be distributed. In particular the 重編國語辭典修訂本 (*zhòng biān guóyǔ cídiǎn xiūdìng běn*, Revised Chinese Dictionary) is available to download and allows reproduction but does not allow redistribution of derivative works. Rights of use will need to be negotiated with copyright holders to make use of the content of those dictionaries.

Three kind of additional information are relevant to the project: semantic comparison, frequency and pedagogical levels. The only file collected so far about semantic comparison is the database created by Matsushita et al. (2017). Wiktionary provides frequency lists for an important number of languages. Pedagogical levels refers to level of standard tests like the JLPT (日本語能力試験, *nihongo nōryoku shiken*), HSK (漢語水平考試, *hànyǔ shuǐpíng kǎoshì*) or TOCFL (華語文能力測驗, *huáyǔwén nénglì* *cèyàn*) when they exist for a language. In some case official vocabulary level lists are available, for other lists have been compiled by netizens. Since they all use different rating, a standardization based on CEFR levels is done. Those data will be used for outputs and features that are outside the scope of this paper.

4 Linguistic Modeling

4.1 Script Normalization

The sources dictionaries make use of 6 different scripts: Chinese characters, katakana, hiragana, hangul, *zhuyin fuhao* (also called bopomofo) and Latin script. Roman alphabet is used for very different romanization schemes: *tâi-lô and peh-ōe-jī* for Taiwanese, *jyutping* for Cantonese and the Vietnamese alphabet. All differ in the value they assign to letters. While it is reasonable to expect the reader to be able read one or two writing systems, it is unrealistic to expect the average user to know the intricacies of a dozen scripts and romanizations. In particular since entries juxtapose pronunciations of a word in multiple languages, confusion in letters' value could arise easily. To solve this issue, the readings of cognates are transformed from their original script to the International Phonetic Alphabet (IPA). Choosing the IPA doesn't solve all problems however: the transcription used could be either phonological or phonetic.

4.2 Phonemic vs Phonetic Transcription

This dictionary lists how cognates are pronounced in the languages it includes. The IPA alphabet is used for that task, but it raises the question of using a phonemic or phonetic transcription. Generally speaking, a phonetic transcription contains more information than a phonemic one. It makes them harder to read for a non-trained user and requires precise information that are present only in specialized dictionaries. The present dictionary thus leans towards phonological transcriptions.

The use of phonological transcription is however problematic in a multilingual context because the phonological system of a language abstract differences that can be meaningful in another language; this occurs particularly with contextual allophones. For instance, Japanese /s/ in front of /i/ is realized as [ɕ] (Labrune, 2006, p. 81). As the same phenomenon applies in Korean (Shin, Kiaer & Sha, 2017, p.70) this is not an issue when comparing words in those two languages. It is however a problem when Japanese is compared to Chinese where both /s/ and /ɕ/ have phonemic status. The same phenomenon applies even if two allophones doesn't exist per se in a language known by a learner but match close ones. For instance, /h/ in Japanese has [h] and [ɸ] as contextual allophones. The phoneme /h/ does not exist in French while [ɸ] would be easily interpreted as /f/.

The way the cognate dictionary handles this problem is to use a phonological transcription that distinguish contextual allophones when relevant (one a case-by-case basis).

4.3 Syllable Structure

Sinitic languages share a common syllable structure made of at most four segments (Wee & Li, 2015). This pattern is commonly referred to as CGVX where C is a consonant, G a glide, V the main vowel and X the coda which can be either a consonant or a vowel. Any segment except the main vowel one is optional. An alternative syllable pattern is that of a single syllabic consonant. The syllable can be described as a tree, for which competing theories exist. For this project the hierarchical model does not yield benefit and a flat model is used instead. In addition, each syllable possesses a tone. An exception to the model exists in Standard Chinese because of the *erhua* (兒化) phenomenon; it is currently not handled by the dictionary and the few entries affected are discarded.

Vietnamese and Korean syllables fit the pattern as well. Japanese exhibits an epenthetic vowels /u/ or /i/ after -/k/ and -/t/ coda. This vowel is discarded for phonological comparison to other languages but is displayed to the user. In the dictionary, diphthongs are split in two parts to ease comparisons between languages, the first part is allocated to the main vowel slot while the remaining part fills the coda slot.

4.4 Comparison Algorithm

4.4.1 Slot Comparison Values

An important part of the project is the similarity algorithm it defines. Phonetic similarity is used in various works pertaining to Chinese Natural Language Processing (NLP); we can cite (Chang et al., 2010) and (Lee et al., 2019) as examples. Metrics created for those works are tailored to the task at hand, and offer limited reusability for a different purpose. Since no existing algorithm fitted our goal, a new one was devised. A measure of similarity between two syllables will allow searching similarly sounding syllable across languages, and provide a numeric value to sort vocabulary, for instance when creating vocabulary lists.

The metric must work across languages, be close to human judgment that is if a human would judge two syllables very similar the score should be very high and it must be computable from the data extracted from dictionaries (i.e. we cannot afford to measure the actual perception in- between all the speakers of the languages involved).

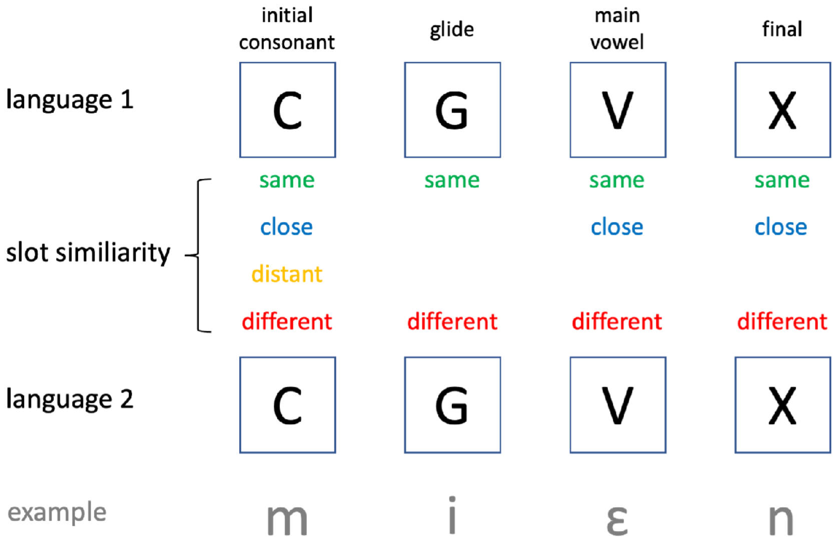


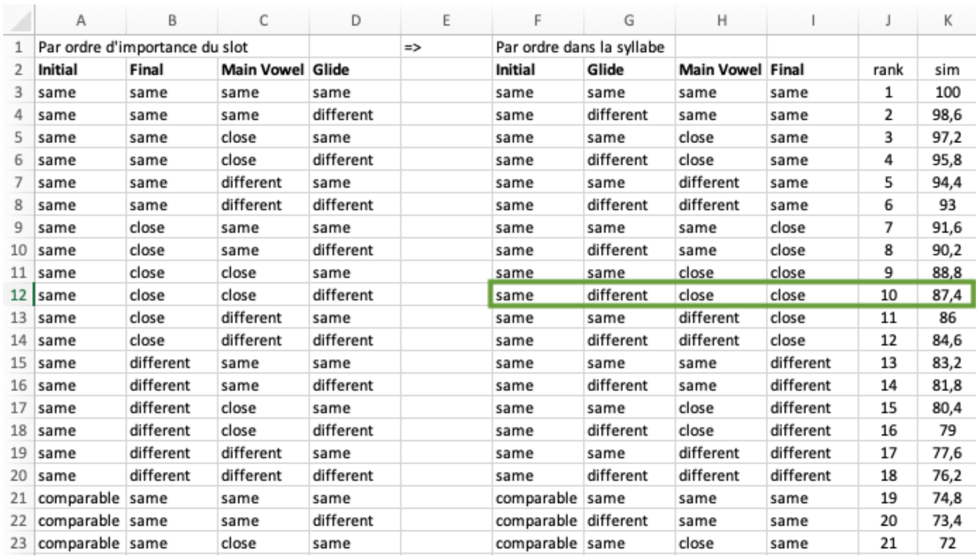
Figure 2: Syllable slots and possible comparison values

The algorithm works by comparing each pair of slots. If the phoneme is identical, the output for the slot is the value “same”. If the phonemes are somewhat close, which is determined on the basis of the user native language and feature geometry (see below), the output for the slot is “close”. Otherwise, the output is “different”. Since the initial consonant is the part of the Sinitic syllable where is the more variety an additional “distant” output value exists. At the syllable level, the number of resulting output combinations is 72 (4*2*3*3).

Intuitively each slot doesn’t participate in the same weight in the similarity between two syllables: for instance, the glide can be absent in one of them without making the syllables too different. More importantly, consonant information is more impactful than vowel one as confirmed to its relative stability over time, and across places and language borrowings (which make the present work feasible in the first place) while vowel information is often highly variable even within dialects of the same language. Those, the algorithm prioritizes consonant information and use the following order of slots: initial, final, main vowel, glide.

4.4.2 Ranking and Similarity

The 72 possible combinations are constructed from the most similar (same, same, same, same) to the most dissimilar (different, different, different, different) and are each effected a rank ranging from 1 to 72. Since metrics usually range from 0 to 1 or 0 to 100, the rank is converted to a measure ranging from 0 to 100 by using the formula *floor(100-(rank-1)*1.4)*.



	A	B	C	D	E	F	G	H	I	J	K
1	Par ordre d'importance du slot				=>	Par ordre dans la syllabe					
2	**Initial**	**Final**	**Main Vowel**	**Glide**		**Initial**	**Glide**	**Main Vowel**	**Final**	rank	sim
3	same	same	same	same		same	same	same	same	1	100
4	same	same	same	different		same	different	same	same	2	98,6
5	same	same	close	same		same	same	close	same	3	97,2
6	same	same	close	different		same	different	close	same	4	95,8
7	same	same	different	same		same	same	different	same	5	94,4
8	same	same	different	different		same	different	different	same	6	93
9	same	close	same	same		same	same	same	close	7	91,6
10	same	close	same	different		same	different	same	close	8	90,2
11	same	close	close	same		same	same	close	close	9	88,8
12	same	close	close	different		same	different	close	close	10	87,4
13	same	close	different	same		same	same	different	close	11	86
14	same	close	different	different		same	different	different	close	12	84,6
15	same	different	same	same		same	same	same	different	13	83,2
16	same	different	same	different		same	different	same	different	14	81,8
17	same	different	close	same		same	same	close	different	15	80,4
18	same	different	close	different		same	different	close	different	16	79
19	same	different	different	same		same	same	different	different	17	77,6
20	same	different	different	different		same	different	different	different	18	76,2
21	comparable	same	same	same		comparable	same	same	same	19	74,8
22	comparable	same	same	different		comparable	different	same	same	20	73,4
23	comparable	same	close	same		comparable	same	close	same	21	72

Figure 3: The first of the 72 possible comparison combinations and their ranks

Figure 3 illustrates how the ranks are computed. On the left the natural progression of ranks is visible (slots are sorted by importance), on the right the slots are re-ordered corresponding to their actual position in the syllable. The two leftmost columns display the rank and the associated similarity. For words, the similarity score is computed using the geometric mean of each syllable similarity. In comparison to the more common arithmetic mean, the geometrical mean is more sensible of important gap in value (e.g. the geometric mean of 1 and 100 is 10).

For example, Japanese 愛 (ai, *love*) and Chinese 愛 (ài, *love*) share the same initial and glide (both empty) as well as the same main vowel and final one. The algorithm gives them a rank of 1, equating a similarity of 100. The Chinese 麵 (*mi*àn, noodle) and Japanese 麵 (*men*, noodle) have the same initial, a different glide, a close main vowel and a close final, leading to a rank of 10 which give a similarity of 87 (see line highlighted in green on Figure 3).

4.4.3 Consonant Comparison with Language Profiles

The comparison of consonants is inspired from the method of consonants classes initiated by Dolgopolsky (1986) and used in comparative-historical linguistics. Examples of such classes can be found in (Kassian et al., 2015). The class of labials (P-class) for instance contains the consonants: p b β ɓ f v… Those classes however are too broad for use in this project.

Another difference is that the data in comparative linguistics are absolute. However, the perception of a phoneme from a foreign language depends on one’s native language.

Table 2: Presence and absence of phonemes /k/, /kʰ/, /g/ in Chinese, Japanese, Taiwanese and French

Phoneme	**/k/**	**/kʰ/**	**/g/**
Chinese	/k/	/kʰ/	—
Taiwanese	/k/	/kʰ/	/g/
Japanese	/k/	—	/g/
French	/k/	—	/g/

One of the common error of speakers (Teramura, 1990) having Chinese as a native language who are learning Japanese as a second language is with the voiced/devoiced characteristic of bilabial plosives (/b/, /p/), alveolar plosives (/d/, /t/) and velar plosives (/g/, /k/) which stems from the voiced series not existing in Chinese. Thus, upon hearing a Japanese word containing a voiced consonant that consonant may be mistaken for its unvoiced counterpart. On the contrary, a native speaker of Taiwanese or French for which the distinction exist will be able to recognize that phoneme correctly. This phenomenon calls for using finer consonant classes, and a different mapping from phonemes to classes that depends on the language of the reader, and on the ability to discriminate phonemes in the second language he is learning.

The output “close” and “distant” is realized in the comparison algorithm by affecting to each phoneme of a language a given class and seeing if the classes match. The association of phonemes to classes is done for every language of the expected readers of the dictionary (this work can be crowd-sourced). For instance, both of the Japanese phonemes /k/ and /g/ are mapped to the class K in the “close” profile language for native speakers of Chinese beginner in Japanese, while /k/ is mapped to K and /g/ to G in the “close” profile of Taiwanese, Japanese, French speaker and advanced learner of Japanese. Moreover, both /k/ and /g/ are associated to class K in “distant” profiles of Taiwanese, Japanese and French. Thus, when comparing 乾 (Chinese *gān*, Japanese *kan*, dry) in Chinese and Japanese the initial will be rated as “close” (since both as K-class) from the point of view of a native Chinese-speaker beginner in Japanese, while being rated only “distant” for a Taiwanese, Japanese, French speaker or advanced learner.

4.5 Correspondences Rules

Besides a visually compelling table of phoneme-to-phoneme comparison, the dictionary aims to include regular correspondences rules between phonemes in language pairs. Despite parallel language evolution, phonemic correspondences still exist in-between the languages included in the dictionary. Some of those correspondences are obvious such as /f/ in Chinese and /h/ in Japanese (方法 *hōhō / fāngf*ǎ, method) while others are less evident; for example Chinese nasal coda -/ŋ/ is usually found as a long vowel in Japanese (e.g. 方 *fang / hō*, direction).

The data and processing tools in the project have for goal to found those correspondences in- between any language pairs, and to compute statistics about their frequency, regularity and their pedagogical potential.

To give an illustration of correspondence rules and their application, let’s observe the pronunciation of three morphemes in Japanese and Vietnamese. For each morpheme, the table 3 give first the pronunciation of the morpheme (in *romaji* for Japanese and *quốc ngữ* for Vietnamese) and then lists a simplified phonemic representation where the glide and vowel information are discarded (symbolized by _). An empty coda is noted ø.

Table 3: Vietnamese and Japanese pronunciation of three morphemes

Morpheme	**Vietnamese**		**Japanese**	
言	ngôn	ŋ _ _ n	gen	g _ _ N
語	ngữ	ŋ _ _ ø	go	g _ _ ø
我	ngã	ŋ _ _ ø	ga	g _ _ ø

From the data listed in Table 3, it is possible to infer the rules listed in Table 4.

With knowledge of these rules, a speaker of Vietnamese will be able to infer that a morpheme pronounced *nguyen* will have the shape g _ _ N in Japanese. Indeed, the rule is true for 元 and 原, both pronounced *nguyên* in Vietnamese and *gen* in Japanese. For those morphemes, the only additional information that a learner has to memorize is the main vowel and glide values. The burden of learning is reduced in comparison to a learner without any prior knowledge.

Table 4: correspondence rules inferable from Table 3

Rule	Representation
Vietnamese voiced velar nasal /ŋ/ at initial is found as voiced velar plosive /g/ in Japanese	ŋ _ _ _ → g _ _ _
Vietnamese -/n/ coda is found as -/N/ coda in Japanese	_ _ _ n → _ _ _ N
Vietnamese empty coda corresponds to an empty coda in Japanese	_ _ _ ø → _ _ _ ø

This is only a small example and rules present in the dictionary will be extracted using all the available data. The important number of cognates pairs collected for most pair of languages (see Table 5) allows to compute how frequent and regular the correspondences are over the lexicon. Correspondence rules will be listed under every entry they are appliable to.

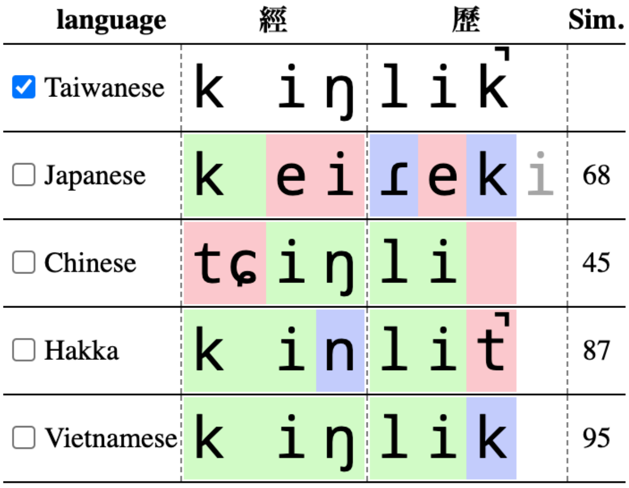
5 Results

5.1 Visualization

The algorithm detailed in Section 4.4 can be used to produce a colored visualization of the difference of pronunciation of a cognate between multiple languages. A graph visualization involving all the language pairs would be hard to read, so the adopted solution is to display comparison data as a dynamic table: one language serves as the basis of comparison and that language can be changed by the user.

The Figure 4 shows the table generated for the cognate (經歷, *jīnglì*), with Taiwanese being used as the basis of comparison. The selected language is put as the first row of the table, and its corresponding checkbox is ticked. In addition, since it's the basis of the comparison, none of its phonemes are colored. The remaining rows of the table contains the other language pronunciations, each with phoneme slots colored based on the output of the comparison algorithm with the top row (see Figure 2 for the meaning of colors).

The content of individual cell is padded with white space so the initials, medials, central vowel and finals are always aligned regardless of their length. A monospace font is used to ensure the alignment is possible. When a slot is missing in every language (the medial of each syllable in the example), the slot is removed from display not to clutter the table with empty columns. The epenthetic vowel present in Japanese is displayed as an addition slot, greyed to indicate the special nature for the vowel in respect to the common syllable structure.

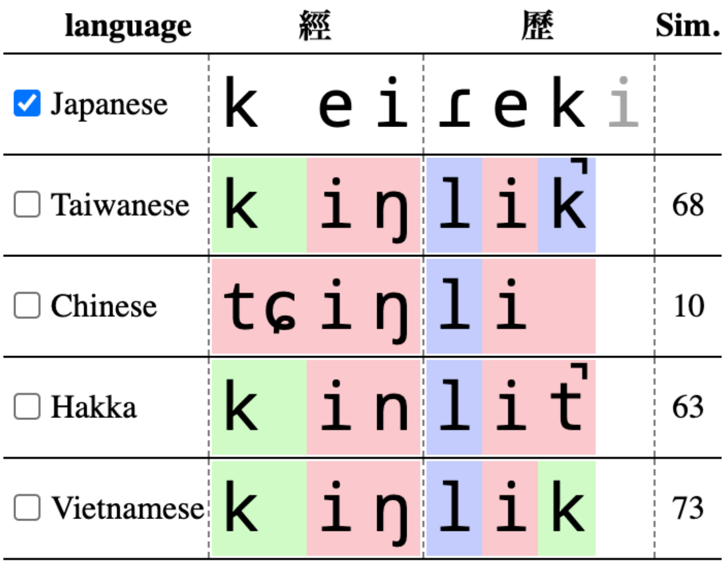


language	經	歷	Sim.
☑ Taiwanese	k iŋ	lik̚	
□ Japanese	k ei	ɾeki	68
□ Chinese	tɕiŋ	li	45
□ Hakka	k in	lit̚	87
□ Vietnamese	k iŋ	lik	95

Figure 4: Entry "經歷" with Taiwanese as comparison basis

This visualization makes very explicit which phonemes are identical in other languages. Moreover, it also gives a quick impression of how the cognate differs in comparison to the other languages: in this case the Taiwanese pronunciation is quite close to most of other languages. In addition, the similarity ("Sim.") column gives the numeric computation of the closeness of pronunciation, which can be used to infirm or confirm the impression given by the coloring scheme.

The base language for comparison is changed by ticking the checkbox corresponding to another language. The Figure 5 displays the same data (經歷 cognate) but uses Japanese as the basis of the comparison. It is immediately clear that the Japanese pronunciation differ greatly from the all languages by the number of red cells present in the table. Besides the initials of the two syllables with is identical or close to most other languages (except Standard Chinese), every other slot, save for the final in Vietnamese, differs.



language	經			歷				Sim.
☑ Japanese	k	e	i	ɾ	e	k	i	
☐ Taiwanese	k	i	ŋ	l	i	k̚		68
☐ Chinese	tɕ	i	ŋ	l	i			10
☐ Hakka	k	i	n	l	i	t̚		63
☐ Vietnamese	k	i	ŋ	l	i	k		73

Figure 5: Entry "經歷" with Japanese as comparison basis

5.2 Shared Vocabulary Between Languages

The dictionary wouldn't be of effective utility if there wasn't a significant number of cognates shared by the languages involved. Since data have been extracted for 7 languages, it is possible to compute the vocabulary common to the possible language pairs (that is, the intersection of their vocabulary).

Table 5 lists the vocabulary in common for the top-6 languages in terms of vocabulary size included in the project. Languages are listed in the first or second column based on the number of entries extracted for that language, the one having the bigger number being put on the first column.

Table 5: Vocabulary common to language pairs

Language 1	Language 2	Shared Cognates
Mandarin	Cantonese	54,024
Mandarin	Taiwanese	19,496
Mandarin	Japanese	18,120
Cantonese	Taiwanese	15,025
Cantonese	Japanese	14,843
Japanese	Korean	11,552
Mandarin	Korean	9,856
Mandarin	Hakka	9,318
Cantonese	Korean	8,630

Japanese	Taiwanese	8,369
Cantonese	Hakka	8,300
Taiwanese	Hakka	6,596
Taiwanese	Korean	4,808
Japanese	Hakka	3,179
Korean	Hakka	1,987

It is notable that 6 combinations of languages share more than 10,000 words and the majority of the pairs share more than 8000 words. While a lot of this vocabulary may be specialized or of very low frequency, this tends to prove that a speaker or learner will be able to reuse a lot of vocabularies by using the dictionary presented here.

It is also possible to compute the vocabularies that are shared by more than two languages at once. In Table 6, the vocabularies present in sets of 4, 5 and 6 languages are computed. It is remarkable that a relatively high number of words (about 5600) are shared by four languages, including a Sinoxenic one.

Table 6: Vocabulary common to 4-6 languages

Languages	**Shared Cognates**
Mandarin, Cantonese, Japanese, Taiwanese	5,599
Mandarin, Cantonese, Japanese, Taiwanese, Korean	2,574
Mandarin, Cantonese, Japanese, Taiwanese, Korean, Hakka	1,001

In addition, there is a set of ~1000 words that are cognates in 6 languages. Example of such words are: 世紀 (*shìjì*, century), 字典 (*zìdiǎn*, dictionary), 完全 (*wánquán*, complete), 將來 (*jiānglái*, future), 人口 (*rénkǒu*, population), 平和 (*pínghé*, peace), 病院 (*bìngyuàn*, hospital), 論文 (*lùnwén*, article), 中央 (*zhōngyāng*, central), which are useful vocabulary for daily life. Other terms bear special cultural interest: 君子 (*jūnzǐ*, a gentleman in Confucianism), 仙人 (*xiānrén*, an immortal in Taoism); those two words are also present in Vietnamese and Central Okinawan, making them existing in at least 8 languages. In some cases, cognates must be accompanied by an explanation if used in a pedagogical context: 三國 (*sānguó*) refers to different periods in different countries (one in China, one in Korea) and other have a particular meaning in one of the language: 風俗 (*fēngsú*) generally means "customs, traditions" but have the additional meaning of "prostitution" in Japanese.

Those words are "high multilingual" and it is arguable that they are of special interest for a learner of multiples languages. Most vocabulary lists and learning materials are created using frequency and/or educators' intuition. The multilingual aspect of lexicon can be an objective metric to use as an additional decision criterion for inclusion into a list of vocabulary.

6 Future Work and Conclusion

6.1 Future Work

The main limitation of the present work regards the tonal information of syllables. This information is always extracted from the script and included in the common syllable format used as output but it is not visible in the comparison tables. Tonal information is currently not normalized and uses a conventional number for each language. In the future, the 5 IPA tone levels will be used, which will allow for automatic comparison of tonal information.

The most obvious future development of the dictionary concerns the languages and dialects it includes. In one hand, more Chinese languages such as Wu and Xiang can be added if data is found. Some of the exploitable dictionaries use Chinese characters as entries. At the moment, the dictionary only contains lemma as entries, so this editing choice might be reconsidered. Second, dialectal variation could be integrated into the dictionary. The source dictionary for Hakka already contains such variations. Both

Vietnamese and Korean feature marked difference in-between their Northern and Southern dialectal groups and are interesting target for inclusion. From a technical point of view, adding additional dialects is no different than adding additional languages to the project, and the process is straightforward since the cognate dictionary have been designed for multilingualism from the start.

On a more distant fashion, the dictionary output could be used to produce pedagogical lists of vocabulary. In comparison to existing lists, the lists generated this way could take two variables in account, that are currently ignored: in one hand the proximity of lexical items pronunciation with the equivalent in the learner native (or known) language. In the other hand, the multilingual aspect of a lemma, that is the number of languages in which it exists. More generally, the dictionary is to support comparative work involving the languages it includes. Since the number of possible pairs included is high (15 pairs when considering 6 languages) some of those work may be the first of their kind.

6.2 Conclusion

This paper presented the on-going effort to create a dictionary of Sinitic and Sinoxenic cognates. Dictionaries with satisfying number of entries have been collected for five languages (Standard Chinese, Cantonese, Japanese, Southern Min, Korean), and smaller scale data exist for three others (Hakka, Vietnamese, Central Okinawan). We presented an overview of the processing toolchain use to extract and compute the content of the cognate dictionary. A major contribution of this paper lies in the algorithm created to compute the similarity of syllables across languages. The output of the algorithm is also used to display clearly the difference in pronunciation between words. The algorithm is adaptable to different native language and proficiency of users, which make is usable for other task such as generating list of word a beginner would likely confound.

Finally, we presented quantitative data on the number of shared cognates between 16 language pairs and found that a significant number of cognates are shared across 6 languages, which confirm the potential usefulness of the dictionary. Further research on language transfer as well as generation of vocabulary lists could be made by leveraging the content of the dictionary.

Acknowledgement

I would like to thanks François Lecailliez for funding the conference registration fees.

References

Bonet, J. (1899). Dictionnaire annamite-français: A-M. Leroux.

Breen, J. (2004). JMDict: a Japanese-multilingual dictionary. In *Proceedings of the workshop on multilingual linguistic resources* (pp. 65-72).

Chang, C. H., Lin, S. Y., Li, S. Y., Tsai, M. F., Liao, H. M., Sun, C. W., & Huang, N. E. (2010). Annotating Phonetic Component of Chinese Characters Using Constrained Optimization and Pronunciation Distribution. *International Journal of Computational Linguistics and Chinese Language Processing*, *15*(2), 145-160.

Crystal, D. (2011). *A dictionary of linguistics and phonetics* (Sixth Edition). Blackwell Publishing.

Dolgopolsky, A. B. (1986). A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. *Typology, relationship and time: a collection of papers on language change and relationship by soviet linguists*, 27-50.

Eberhard, D. M., Simons, G. F., & Fennig C. D. (Eds.). (2021). *Ethnologue: Languages of the World*. Twenty-fourth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.

com.

Frellesvig, B. (2010). *A history of the Japanese language*. Cambridge University Press.

Goudin, Y. (2017). *L'intercompréhension en langues sinogrammiques: théories, représentations, enjeux, et modalités d'une didactique de la variation*. (Doctoral dissertation, Sorbonne Paris Cité).

Handel, Z. (2015). The classification of Chinese. In W. S.-Y. Wang & C. Sun (Eds.), *The Oxford handbook of Chinese linguistics* (pp. 34-44). Oxford University Press.

Kassian, A., Zhivlov, M., & Starostin, G. (2015). Proto-Indo-European-Uralic comparison from the probabilistic point of view. *Journal of Indo-European Studies, 43*(3-4), 301-347.

Kwok, B. C. (2018). *Southern Min: Comparative Phonology and Subgrouping*. Routledge.

Labbé, G. (2018). *Fondements linguistiques et didactiques de l'intercompréhension slave : le cas des langues slaves de l'ouest et du sud-ouest*. (Doctoral dissertation, Sorbonne Paris Cité).

Labrune, L. (2006). *La phonologie du japonais* [Version électronique]. Peeters Publishers.

Labrune, L. (1993). À propos d'un trait typologique du japonais: l'absence de r à l'initiale des mots indépendants de Yamato kotoba. *Ebisu*-Étude*s Japonaises*, *2*(1), 7-21.

Lecailliez, L. (2015). *Approches pour une numérisation de qualit*é *d'un dictionnaire vietnamien-français comprenant des caractères Nôm*. (Master's thesis, Paris Diderot, Paris, France).

Lecailliez, L., Flanagan, B., Chen, M.-R. A., & Ogata, H. (2020). Smart dictionary for e-book reading analytics. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (pp. 89-93). Lee, H. (1994). The Origin of Sino-Korean. *Korean Linguistics, 8*(1), 207-222.

Lee, H. (1994). The Origin of Sino-Korean. *Korean Linguistics, 8*(1), 207-222.

Lee, L.-H., Wu, W.-S., Li, J.-H., Lin, Y.-C., & Tseng, Y.-H. (2019). Building a confused character set for Chinese spell checking. In Proceddings of the 27th International Conference on Computers in Education, (pp. 703-705). Asia-Pacific Society for Computers in Education.

Martin, E. S. (1953). The Phonemes of Ancient Chines. *Journal of the American Oriental Society 73*(2), Supplement 16, 1-46.

Nakazawa, N., Iwaki, H., & Koresawa, N. (2013). Possibility of the Japanese-Taiwanese Fundamental Characters' Contrastive Phonetic Table with the Japanese Language Education. 2nd International Conference on Vietnamese and Taiwanese Studies & 6th International Conference on Taiwanese Romanization. National Cheng Kung University, Taiwan.

Phong, N. P. (1978). À propos du Nôm, écriture démotique vietnamienne. *Cahiers de Linguistique Asie Orientale, 4*(1), 43-55.

Pyysalo, J., Kotiranta, F., Sahala, A., & Hulden, M. (2019). Proto-Indo-European lexicon and the next generation of smart etymological dictionaries: The technical issues of the preparation. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 592-602). Lexical Computing CZ sro. Sin, C. Y.,

Shin, J., Kiaer, J., & Cha, J. (2012). *The sounds of Korean*. Cambridge University Press.

Qu, C. (2017). The Necessity of Compiling a Learners' German-English-Chinese Etymological Dictionary. In Proceedings of *ASIALEX 2017: The 11th International Conference of Asian Association of Lexicography* (pp. 483-499). Guangzhou, China.

Sakai, T. & Nakazawa, N. (2017). Database of Holo-Taiwanese Language in the Process of Making Multi-Linguistic Policies. In proceedings of the 14th Annual Conference of European Association of Taiwan Studies. Ca'Foscari University of Venice, Italy.

Sio, J. U.-S., & Morgado da Costa, L. (2019). Building the Cantonese Wordnet. In Proceedings of the Tenth Global Wordnet Conference (pp. 206-215). Wrocław, Poland.

Sybesma, R. P. E., Behr, W., Gu, Y., Handel, Z. J., Huang, C.-T. J., & Myers, J. (Eds.). (2017). *Encyclopedia of Chinese Language and Linguistics* (Vol 4). Brill.

Wee, L.-H, & and Li, M. (2015). Modern Chinese Phonology. In W. S.-Y. Wang & C. Sun (Eds.), *The Oxford handbook of Chinese linguistics* (pp. 474-489). Oxford University Press.

[Xiong, K., & Tamaoka, K.] 熊可欣 - 玉岡賀津(2014）「雄日中同形二字漢字語の品詞性の対応関係に関 する考察」『ことばの科学 第 27 号 (特集号)，25-51.

[Matsushita, T., Chen, M., Wang, X., & Chen, L.] 松下達彦･陳夢夏･王雪竹･陳林柯（2017）「日中対照漢字 語データベースの開発と応用」『日本語教育会秋季大会予稿集』，336-371.

[Wang, Y., Xu, C. & Kodama, S.] 王永全 - 許昌福 - 小玉新次郎（2007）『日中同形異義語辞典』東方書店. [Tang, L.] 唐磊（1993）『現代日中常用漢字対比詞典』北京出版社.

[Teramura, H.] 寺村秀夫(1990)『外国人学習者の日本語誤用例集』（大阪大学；PDF 版、国立国語研究 所、2011 年）