Annotating Cognates in Phylogenetic Studies of South-East Asian Languages

Mei-Shin Wu[1] and Johann-Mattis List[1]

[1]*Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, DE*

May 19, 2021

Abstract

Compounding and derivation are frequent in South-East Asian languages. Consequently, words in different languages are often only partially cognate, sharing only a few but not all morphemes. While partial cognates do not constitute a problem for the phonological reconstruction of individual morphemes, they are problematic when it comes to phylogenetic reconstruction based on comparative wordlists. Here, we review the current practice of preparing cognate-coded wordlists and develop new approaches that make the process of cognate annotation more transparent. Comparing four methods by which partial cognate judgments can be converted to cognate judgments for whole words on a newly annotated dataset of 19 Chinese dialect varieties, we find that the choice of the conversion method has a large impact on the inferred tree topologies. We conclude that scholars should take cognate judgments in languages in which compounding and derivation are frequent with great care, and recommend to assign cognates always transparently.

Keywords

phylogenetic reconstruction, Chinese dialects, South-East Asian languages, cognate annotation, partial cognates

1 Introduction

Compounding and derivation are important word formation mechanisms in South-East Asian languages. In fact, more than 30% percent of the nouns in Chinese dialects are built from more than two morphemes (List et al., 2016), and similar or larger numbers can be observed for other language families in South-East Asia (such as, for example, Hmong-Mien, see Máo 2004, 184).

When scholars reconstruct South-East Asian proto-languages, compounds are usually separated into parts and only morphemes are reconstructed back to the level of the ancestor language. Attempts to infer which *words* were used to express certain concepts in the proto-languages have rarely been made so far (see Mann 1998 as a rare example). While compounding and derivation can be easily ignored when resorting to root reconstruction, they cause problems when it comes to phylogenetic reconstruction, where the starting point is a comparative wordlist, and the data are coded for *cognate words* which are assumed to go back to the same ancestor form. Although methods for phylogenetic reconstruction have been employed for quite some time now in the fields of historical linguistics, not many attempts have been made to address the problem of coding cognates in language families in which partial cognate relations are frequent due to compounding and derivation, although some studies have discussed the problem of partial cognate relations (see List 2016 for a general discussion).

In this study, we provide some concrete ideas regarding the transparent handling of partial cognate in phylogenetic analyses. After discussing past attempts to convert partial cognates into full cognates which assess the cognacy of two or more words as a whole, we introduce two new methods for the transparent comparison of partial into full cognates along with tests that help to identify which concepts in a comparative wordlist are influenced by partial cognate relations. By applying these new approaches to a dataset of Chinese dialects (Liú et al., 2007), we show that the choice of conversion method has a direct impact on the resulting distances between languages and the phylogenies inferred from the Neighbor-joining algorithm (Saitou and Nei, 1987). While our analyses do not allow us to favor one explicit conversion method over the other methods, we conclude that partial cognate relations should be taken with great care when carrying out phylogenetic analyses of language families in which compounding is frequent.

2 Increasing the Transparency of Cognate Annotation

At the moment, cognate annotation in South-East Asian languages faces two extremes. The one extreme, which is the data model underlying many etymological studies, takes the (unbound) morpheme as a basic unit – ignoring words completely as linguistic units – and assembles cognate sets of morphemes without storing a reference to the words from which these were taken. The alternative extreme can be found in phylogenetic approaches where words are traditionally taken as the basic units of comparison. Here, scholars assemble translational equivalents for a fixed list of basic concepts and then assign these words to cognate sets, without making explicit how partial cognates were handled.

Recent work concentrating on computer-assisted approaches to historical language comparison has shown that the first extreme can be avoided when starting from a careful annotation of partial cognates in comparative wordlists (Wu et al., 2020). Instead of picking cognate morphemes from the literature, the workflows do not only allow researchers to maintain the link between the original words in which the morphemes occur and the morphemes themselves, but even offer convenient ways to inspect sound correspondence patterns (List, 2019) and search for partial colexifications (Hill and List, 2017).

What has *not* been sufficiently solved so far, however, is the question of how to deal with the annotation of cognate sets for the purpose of phylogenetic reconstruction. Here, the main problem is how to derive cognate judgments for full words when words are only partially related. In the following, we will discuss some general ideas regarding the annotation of cognate sets in wordlists for the purpose of phylogenetic reconstruction studies and then share some specific recommendations for concrete issues.

2.1 General Ideas

When assembling comparative wordlists for the purpose of phylogenetic reconstruction, the major problem imposed by language families in which partial cognacy is frequent is that it often becomes very difficult to find clear-cut criteria to assign words to cognate sets. In abstract terms, if one language expresses a concept *X* with a compound word *a-b* another language expresses the same concept with a compound word *a-c*, there are two possibilities: one could either argue that both words are to be judged cognate, given that they have one cognate morpheme *a* in common, or one could argue that they are not cognate, given that they differ due to their respective morphemes *b* and *c* which are not cognate. The complexity increases when more words are brought

to the comparison and can easily lead to cases where the decision to assign all words to the same cognate set which share at least one common morpheme yields situations in which our hypothetical word *a-b* would be cognate with *a-c* and *a-c* would be cognate with *d-c*, but *d-c* would no longer share any common element with *a-b*.

The two most straightforward approaches to assign words to cognate sets when their partial cognate sets are known have been called "strict" and "loose" cognate coding in previous work (List et al., 2016; List, 2016). In the *strict* case, only those words are assigned to the same cognate set which are cognate with respect to all of their morphemes. An example for this coding is the study on Chinese dialect evolution by Hamed and Wang (2006). In the *loose* case, a network of all words is constructed in which words correspond to nodes and links between nodes are drawn whenever two words share at least one cognate morpheme. After the network has been constructed, all words that belong to a *connected component* in the network are assigned to the same cognate set (Hill and List, 2017). An example for this coding procedure can be found in the study by Satterthwaite-Phillips (2011). Both approaches have their advantages and disadvantages. While strict coding may easily increase differences between language varieties, given the wrong impression of a huge amount of linguistic variation in a given language family, the loose coding practice is unsatisfying as it may easily result in cognate sets consisting of word pairs that do not have a single cognate morpheme in common.

Given that we assume that partial cognates have been identified, an additional way to code the data in phylogenetic analyses would consist in ignoring the word level and coding the partial cognate sets directly. This *one-hot encoding* technique, however, would contradict the important criterion of character independence, since individual morpheme cognate sets have not been evolving alone, but together with the words in which they appear. Since character independence is one of the basic criterions upon which phylogenetic models are built, introducing character dependencies may not only impact phylogenetic reconstruction (Felsenstein, 1988, 446), it will also make the results extremely difficult to interpret, since we ultimately want to understand how whole words evolve during language evolution, not how certain morphemes are gained and lost.

In order to avoid counting words as cognate which do not share a single cognate morpheme, Sagart et al. (2019) annotate their cognate sets in such a way that all words assigned to the same cognate set must at least have one morpheme in common. While this coding practice is beyond doubt more principled than the strict or the loose coding practice mentioned before, it has the disadvantage that it cannot be automatically checked. Sagart et al. (2019) make use of alignment analyses in order to make sure

that there is a common morpheme in large cognate sets, but since they do not mark partial cognates in their data, it is not trivial to check all of their codings automatically. As a result, it is possible to check the consistency of their cognate annotation, but it is not easy to do so manually.

It is never trivial to decide if overall cognacy for a set of words should rely on the presence of one single morpheme shared by all words or the presence of several words. As an example, consider the concept "sun", which many Austronesian languages lexify as "eye of the day", with *day* being often equivalent to the original word for "sun" (Starostin, 2013, 121-123). As Starostin, whom we owe this example, rightfully notes, it is important to determine the most likely *processes* by which the words have evolved. As a result, the decision, whether to judge a compound word that literally translates to "eye of the SUN/DAY" to be cognate with a word "SUN/DAY" may well depend on the specific language family in question and can therefore not be resolved by a computational approach that is blind to the specific contexts by which words change in the language family under question.

While it is not possible to design a straightforward algorithm that would make the cognate decisions in our place, it is, however, possible to insist on a more explicit *annotation* of lexical cognacy data that would reflect the individual cognacy decisions taken by individual scholars. The solution we propose for this task is to make use of *morpheme-glosses*, as proposed by Hill and List (2017) and further developed by Schweikhard and List (2020), and to indicate the *salient morphemes* that contributed to a general judgment on the cognacy of words sharing cognate morphemes (see also 2.2).

This idea can be seen as a less restricted variant of the aforementioned *strict* conversion of partial cognates into cognate judgments on whole words. While the strict conversion takes all morphemes in a given word as equally important, our proposal to annotate which morphemes are salient and which are not allows scholars to exclude specific morpheme cognates from the equation. As a result, scholars can, for example, argue that a certain suffix occurs too frequently in a given dataset to be worthwhile to play a significant enough role to decide if one word that has the suffix should be cognate with another word that lacks the suffix.

Morpheme glosses are a free annotation form that serves to describe the *semantic motivation structure* of a given word. The term *motivation* is based on Koch (2001) and is used by Hill and List (2017) and Schweikhard and List (2020) to denote the semantics underlying word formation processes. As an example, consider Mandarin Chinese *shùpí* 树皮 "bark (of tree)", which consists of the two morphemes *shù* 树 "tree" and *pí* 皮 "skin". The semantic motivation underlying the compound is thus the metaphorical

use of "skin" to denote the cover of trees. Hill and List (2017) indicate these motivation structures in their tabular wordlist data with the help of an extra column in which individual morphemes of multi-morphemic words are glossed.

As an example for this annotation practice, consider the example of words denoting "hatchet" in six Mienic varieties (original data taken from (Máo, 2004)) given in Table 1. In this table, we can observe three distinct morphemes from which all six words are built. All words share one morpheme that means "knife" in isolation (colored in red in the table), but in Daping and Dongshan, the reflexes *dziu*^{22} and *ɖu*^{42} appear in the end of the words, while they appear in the beginning in the other four varieties. The first morphemes in Daping and Dongshan, respectively, are reflexes of Proto-Hmong-Mien **dzaŋ*^{A} "firewood" in the reconstruction of Ratliff (2010, p. 254), and the semantic motivation of the words in the two varieties is "firewood + knife", indicating that a hatchet is a specific kind of knife predominantly used for the preparation of firewood. In the remaining four varieties, the morpheme for "knife" appears in the beginning of the word, and the second morpheme can be translated as "bent, crooked" in isolation. Since most Mienic languages place the modifier after the modified, the semantic motivation for "hatchet" is "bent knife", that is, a knife that has a bent form.

Variety	Subgroup	Form	Segments	Morpheme Glosses	Cognates
Daping	Zao Min	hɔŋ^{53}dziu^{22}	h ɔ ŋ ^{53} + dz j u ^{22}	firewood knife	1 2
Dongshan	Biao Mon	tsɑŋ^{31}ɖu^{42}	ts ɑ ŋ ^{31} + ɖ u ^{42}	firewood knife	1 2
Jiangdi	Iu Mien	dzu^{12}ŋau^{33}	dz u ^{12} + ŋ au ^{33}	knife bent	2 3
Liangzi	Kim Mun	ɖu^{22}ŋau^{33}	ɖ u^{22} + ŋ au ^{33}	knife bent	2 3
Luoxiang	Iu Mien	ɖu^{22}ŋau^{35}	ɖ u ^{22} + ŋ au ^{35}	knife bent	2 3
Miaoziyuan	Iu Mien	dzəu^{21}ŋau^{33}	dz əu ^{21} + ŋ au ^{33}	knife bent	2 3

Table 1: Using morpheme glosses to annotate semantic motivation structures for words denoting "hatchet" in six Mienic varieties.

Once morpheme glosses have been added to a dataset, the annotation of *salient morphemes*, that is, morphemes whose cognacy relations with morphemes in other words should be taken into account when deriving cognate sets for whole words from cognate sets annotated for individual morphemes, can be done in a very straightforward way by simply indicating the saliency along with the morpheme glosses. In our concrete annotation, this means that we add an underscore _ in front of each morpheme gloss which we consider as *not* salient. When later converting partial cognates to "full" cognates, we only extract those cognate sets whose morpheme glosses have been annotated as salient and then use the strict conversion procedure on these selected cognate sets.

As an example for this procedure, consider the words for "belly" in five Hmongic languages in Table 2 (Chén, 2012, p. 599). All words show the same basic structure of being composed of a prefix with synchronically intransparent semantics and a main morpheme with the core meaning "belly". As can be seen from our partial cognate annotation (provided in the column *Partial*), we identify three distinct suffixes and two distinct morphemes for "belly", one going back to Proto-Hmong-Mien *chu̯ei*^{A} in the reconstruction of Ratliff (2010), the other of unknown origin to us. When computing strict cognate sets from the partial cognates, all words will be placed into a distinct cognate set, since none of the words coincide in all their morphemes. When using the procedure of loose cognate annotation, all words would be placed into the same cognate set, since they all form one big connected component, in which words containing a reflex of Proto-Hmong-Mien *chu̯ei*^{A}, labeled `belly/A` in our morpheme glosses, are connected to the words with the reflex labeled `belly/B` are connected via the prefix `prefix/A`, shared between Western Baheng and Chuanqiandian. Our procedure of salient cognate coding, on the other hand, deliberately ignores the prefixes – given that their presence or absence provides little evidence for the historical development of the words on which they occur, but rather points to largely language-specific processes of productive prefixation that are not well understood to us by now – and thus divides the five words neatly into two cognate sets, depending on their basic morpheme expressing the meaning of "belly" in the example.

Variety	Segments	Morpheme Glosses	Partial	Strict	Loose	Salient
Western Xiangxi	q o ^{35} + tɕ^{h} i ^{35}	`_prefix/Q belly/A`	1 2	1	1	1
Eastern Xiangxi	k i ^{03} + t^{h} i ^{53}	`_prefix/K belly/A`	3 2	2	1	1
Western Baheng	ʔ a ^{03} + ŋ̊ ŋ ^{31}	`_prefix/A belly/B`	4 5	3	1	2
Numao	n̥ u ŋ ^{13}	`belly/B`	5	4	1	2
Chuanqiandian (NEY)	ʔ a ^{55} + tɕ^{h} a u/w ^{55}	`_prefix/A belly/A`	4 2	5	1	1

Table 2: Using morpheme glosses to derive cognate sets for whole words from partial cognate sets. By marking non-salient morphemes with a preceding underscore _, we can explicitly select only those partial cognate sets relevant for the assignment of word cognates, arriving at a transparent procedure for the annotation of cognate judgments for full words.

2.2 Specific Ideas

The schema presented in the previous section relies entirely on human judgment so far, and it is difficult – at least for the time being – to think of an automated approach

to approximate human judgments. The reason is not the impossibility of finding alternatives to the strict and the loose practice of converting partial to full word cognate sets. As we will show in the previous sections, we can easily implement a method that accounts for the cognate coding practiced by Sagart et al. (2019). The problem is that it is often not clear what should count as the best solution, and that there is no real way to tell so based on the data alone. In the following, we will nevertheless try to provide some general criteria that may help scholars in arriving at decisions in particularly difficult situations.

There are three major caveats when deciding about full-word cognacy in multilingual wordlists. First, when annotating cognates, scholars should try to avoid to code cases as cognates which are highly likely to have evolved as a result of parallel independent evolution (*avoid homoplasy*). Second, one should try to make sure that the characters, that is, the cognate sets, are maximally independent (*minimize character dependency*). Third, one should make sure to identify cases of free or pragmatically conditioned synchronic variation and control for them systematically (*control variation*).

As an example for the first problem, the problem of parallel independent evolution or homoplasy, consider cases of *lexical motivation* in compounding (Koch, 2001). Words for 'tear' in Hmong-Mien languages are a good example for this problem, since as in many South-East Asian languages, 'tear' tends to be expressed with the help of a compound, of which one part in isolation is related to a word that means or originally meant 'water' (consider Mandarin Chinese *lèi-shǔi* 'tears', which can be glossed as 'tear + water'). In the Hmong-Mien languages, the other part of the compound is typically the same as the word for 'eye', and the lexical motivation of 'tear' can thus be described as the 'water' of the 'eye' (Chén, 2012, p. 609). Unlike most Chinese dialects, which tend to place the modifier before the modified in compounds, Hmong-Mien languages typically use the opposite order ('water + eye' instead of 'eye + water'). In Sinitic, there are some exceptions of this rule in the South, which scholars tend to attribute to influence from the Hmong-Mien languages (Vittrant and Watkins, 2019), but we can find the opposite influence in some Hmong-Mien varieties as well. As a result, some Hmong-Mien languages lexify 'tears' as 'eye + water', such as Zao Min *mai*^{53}-*m*^{24} (*mai*^{53} means 'eye' in isolation, going back to Proto-Hmong-Mien **mu̱ɛjH*, and *m*^{24} means 'water', going back to Proto-Hmong Mien **ʔu̱əm*,(see Chén (2012) and Ratliff (2010)), while the majority has a compound 'water eye', such as Western Qiandong *ʔeu*^{44}-*me*^{22} (*ʔeu*^{44} is 'water' and *me*^{23} is 'eye', see Chén (2012)). Note that the morphemes in the words in Zao Min and Western Qiandong both go back to the same proto-forms, even if it is quite likely that the word for 'eye' has been borrowed from

Chinese. While it is trivial (despite the complex sound correspondences) to identify the morphemes in both words as cognate, it is far from trivial to decide on the cognacy of both words. One could assume that Proto-Hmong-Mien once had a compound 'water + eye' and that this compound was inherited by both Zao Min and Western Qiandong, and that the lexical motivation of the compound did not lose its transparency until Zao Min began to revert the order of compound constituents from modified-modifier to modifier-modified, possibly under the influence of Chinese dialects. The reverted word for 'tears' this reflects some global innovation in the language which affected a large part of its lexicon. Another possibility, however, is to assume that the motivation underlying words for 'tears' in the Hmong-Mien languages is so obvious and general that we can easily assume that it could recur independently throughout the history of many languages. As a result, it would be wrong to say that the words as such are cognate, since one would assume that they were coined independently. With the knowledge we have at our disposal, we consider this case as undecidable. As a result, it seems best to ignore items like 'tears' when applying phylogenetic reconstruction methods to the Hmong-Mien language family, in order to make sure that the phylogenetic signal is not contaminated by parallel evolution.

As an example for the problem of character dependence, consider the analytical derivation of plural forms for personal pronouns in many South-East Asian languages. While plural forms for personal pronouns tend to have an independent (suppletive) form in most Indo-European languages (compare German *ich* 'I' vs. *wir* 'we', *du* 'thou' vs. *ihr* 'you (pl.)'), many South-East Asian languages derive plural forms from the singular forms by means of suffixation (Mandarin *wǒ* 我 'I' vs. *wǒ-men* 我們 'we', *nǐ* 你 'thou' vs. *nǐ-men* 你們 'you (pl.)'). As a result, the plural form can be regularly predicted from the singular form for most languages in which the plural is built analytically. Since many questionnaires for phylogenetic reconstruction in linguistics, however, contain concepts for singular and plural personal pronouns, the corresponding characters for 'I', 'thou', 'we', and 'you (pl.)' can no longer be considered to have evolved independently, since singular pronouns are re-used to form the plural pronouns, and all plural pronouns tend to share the same affix that derives the plural meaning.

When encountering these processes across all languages in a given dataset, the only consequent way to deal with the cognate assignments is to code each morpheme only *once*, which would mean that one needs to modify the underlying questionnaire in such a way that only singular forms are used as the base forms, while plural forms of personal pronouns are collapsed into one single 'plural' category. If, however, not all plural forms are constructed analytically, as is the case for the Hmong-Mien lan-

guages, where some varieties have a regular plural suffix, similar to Mandarin Chinese (compare Jongnai, a Hmongic language, *wa*^{31} 'I' vs. *wa*^{31}-*kluŋ*^{53} 'we'; Iu Mien, a Mienic language, *ʑe*^{33} 'I' vs. *ʑe*^{33}-*wo*^{33} 'we'), some also have suppletive forms (Eastern Xiangxi, Hmongic, m^{31} 'thou' vs. *ma*^{53} 'you (pl.)'), we recommend to exclude plural forms directly from the analysis, since the independency of the characters cannot be guaranteed.

As an example for the problem of controlling variation, consider the phenomenon of affixation in the Hmong-Mien language family. In many Hmong-Mien languages one finds a certain number of productive prefixes or suffixes which are typically used to derive nouns from a base form. Some of these derivations are mandatory, while some can be omitted, depending on the context. Thus, the word for 'star' in Xia'ao (Western Xiangxi, Hmongic branch of Hmong-Mien) will typically be elicited as *qa*^{02}-*sin*^{44} (Chén, 2012, p. 145 and 282), consisting of the prefix *qa*^{02}-, which derives inanimate nouns, and the noun *sin*^{44}, an early borrowing from Chinese *xīng* 星, which was pronounced as *seŋ* in the 6th century AD (Baxter, 1992). The use of the prefix, however, is not obligatory: it can be omitted, depending on the context Chén (2012, p. 145). When deriving cognate judgments for similar cases, where free variation can be observed, we recommend first to check and make sure that the variation can be observed in all or most of the languages in a given sample, and if this is the case, then mark only the shorter form as the salient one in a second step.

As we have tried to illustrate throughout this section: it is by no means trivial to deal with theses questions, and we expect that the impact on phylogenies when adopting arbitrary solutions for cognate coding can be rather substantial. In order to address the problems consequently, we suggest that scholars working with languages in which partial cognacy is a frequently recurring problem, resulting from abundant compounding and rich derivational processes, carry out a very close analysis of *language-internal cognacy*. Using morpheme glosses, as they have been originally proposed by Hill and List (2017) and further developed by Schweikhard and List (2020), it is possible to rigorously mark prefixes, suffixes, as well as the lexical motivation structures underlying compounds. Once this analysis has been carried out and partial cognates have been identified across languages as well as language-internally, thus taking both words with the same meaning and words with different meanings into account, scholars can carefully check individual semantic slots and try to identify whether any of the three problems discussed in this section applies. If this turns out to be the case, one should (1) ignore the concepts that are lexified by words that are suspicious of parallel evolution due to frequently recurring patterns of lexical motivation (*avoid homoplasy*), (2) try to identify the phylogenetically important alternations when dealing

with problems of character dependency and recode the data accordingly (*minimize character dependency*), and (3) carefully study how words vary when being used in different contexts in order to handle problems resulting from language-internal variation (*control variation*).

3 A Case Study on Chinese Dialect History

In order to illustrate the problems resulting from cognate coding when working with language families in which compounding and derivation are frequent, we have prepared a case study on Chinese dialect history, based on a dataset which we have coded to the best of our knowledge following the principles discussed in the previous section. In the following, we will first present how the original dataset was lifted from its raw tabular version without cognate judgments to a standardized version in which partial cognates have been identified both across and inside language varieties, and how morpheme glosses were used to characterize the semantics of morphemes (3.1). We will then show, how the standardized version of the data allows us to automatically infer those cases which constitute a problem for phylogenetic analyses (3.2), and finally report the results of this analysis, accompanied by individual examples from the data. Our analyses are all supplemented with this paper and available in the form of a small collection of Python scripts which scholars can use to investigate their own datasets (see Supplementary Material).

3.1 Materials

The dataset was originally published by Liú et al. (2007) and later digitized for this study by typing the data off to text files. The data consists of 201 concepts translated into 19 Sinitic varieties which provide at least one variety as representative for each of the seven major subgroups, proposed by Norman (1988, 181) (Mandarin *guānhuà* 官話, Wú 吴語, Xiāng 湘語, Mǐn 閩語, Yuè 粵語, Gàn (贛語), and Hakka *kèjiā* 客家), as well as one variety for each of the three subgroups which are often additionally proposed (Jìn 晋語, Pínghuà 平話, and Huī 徽語, Yan 2006). In order to guarantee the comparability of our dataset with other datasets, we linked the concept list to the Concepticon reference catalog (`https://concepticon.clld.org`, List et al. 2020), and the language varieties to Glottolog (`https://glottolog.org`, Hammarström et al. 2020).

In the raw data, the translations for each concept in each variety are given in phonetic transcription and in Chinese characters (Liú et al., 2007). The latter are fre-

quently used by Chinese dialectologists in order to mark etymologically related morphemes across different dialects (*běnzì* 本字, literally "original characters", see Mei 1995). Although the Chinese character information on cognacy needs to be taken with some care, it is a good starting point for the annotation of cognate sets both across dialects and inside one and the same dialect.

Phonetic transcriptions in the original dataset were standardized by converting the original transcriptions – which follow specific peculiarities as they are typically found in Chinese dialect descriptions – to the transcriptions proposed by the Cross-Linguistic Data Formats reference catalog (CLTS, `https://clts.clld.org`, List et al. 2019a, see Anderson et al. 2018 for details on the CLTS system). The CLTS system can be seen as a narrower version of the International Phonetic Alphabet in so far as it resolves several of its ambiguities. For the conversion and segmentation of the transcriptions, orthography profiles (Moran and Cysouw, 2018) were used.

Partial cognate sets were first automatically added to the data by employing the Chinese character readings, and later systematically refined, using the interactive web-based EDICTOR tool for the creation of etymological datasets (`https://digling.org/edictor`, List 2017, 2021). Morpheme glosses, following Hill and List (2017) and Schweikhard and List (2020) were manually added for all morphemes, based on the previously inferred partial cognate sets. In order to facilitate the reuse of the data, we use the CLDFBench software package (Forkel and List, 2020) to convert the data to the tabular standards proposed by the Cross-Linguistic Data Formats initiative (CLDF, `https://cldf.clld.org`, Forkel et al. 2018).

ID	Language Name	Pīnyīn	Dialect Group	Chinese Name
Beijing Man	Beijing	Běijīng	Mandarin	北京
Changsha Xia	Changsha	Chángshā	Xiang	长沙
Chengdu Man	Chengdu	Chéngdū	Mandarin	成都
Fuzhou Min	Fuzhou	Fúzhōu	Min	福州
Guangzhou Yue	Guangzhou	Guǎngzhōu	Yue	广州
Guilin Pin	Guilin	Guīlín	Pinghua	桂林
Haerbin Man	Ha_erbin	Hāěrbīn	Mandarin	哈尔滨
Jinan Man	Jinan	Jǐnán	Mandarin	济南
Jixi Hui	Jixi	Jìxī	Hui	绩溪
Loudi Xia	Loudi	Lóudî	Xiang	娄底
Meixian Hak	Meixian	Méixiàn	Hakka	梅县
Nanchang Gan	Nanchang	Nánchāng	Gan	南昌
Nanjing Man	Nanjing	Nánjīng	Mandarin	南京
Rongcheng Man	Rongcheng	Róngchéng	Mandarin	荣成
Suzhou Wu	Suzhou	Sūzhōu	Wu	苏州
Taiyuan Jin	Taiyuan	Tàiyuán	Jin	太原
Wenzhou Wu	Wenzhou	Wēnzhōu	Wu	温州
XiAn Man	Xi_an	Xī’ān	Mandarin	西安
Xiamen Min	Xiamen	Xiàmén	Min	厦门

Table 3: A list of dialects. The *ID* is the labels of dialects in the figures.

3.2 Methods

In the following, we will present a range of techniques that can both be used to detect problems resulting from partial cognacy in phylogenetic reconstruction and to handle problems consistently.

3.2.1 Deriving Full Cognates from Partial Cognates

We have discussed different techniques of converting partial to full cognates in Section 2.1. While the *strict* and the *loose* conversion method are straightforward to implement and have been available as part of the LingPy software package (`https://lingpy.org`, List et al. 2019b) since 2016, the method employed by Sagart et al. (2019) has so far only been manually applied. Notwithstanding certain problems resulting from the proper handling of recurring suffixes, this method can be approximated by a greedy algorithm.

The algorithm we propose proceeds in two stages. In a first stage, we construct *fuzzy clusters* from all words in a given meaning slot by creating one cluster for each distinct morpheme (as indicated by the partial cognate identifier) in the selection. In a second stage, we order the clusters by sizing, starting from the largest cluster, and mark all words which contain the morpheme represented by this cluster as *salient*. We then iterate over the remaining clusters and remove all words which occurred in our first cluster from the remaining clusters.

As an example, consider four languages A, B, C, and D which express one word with two morphemes each, *a-b*, *a-c*, *a-d*, *d-c*. In our first stage, we assign the words to four clusters *a* (A, B, C), *b* (A), *c* (B, D), and *d* (C, D). When iterating over the clusters, we start from cluster *a*, mark all words as salient (***a***-*b*, ***a***-*c*, ***a***-*d*), and remove the words with morpheme *a* from the remaining cluster. As a result, cluster *b* is empty, as it contains only one word with *a*, while *c* looses the word from language B and *d* looses the word from language C. The next cluster in our ordered list is *c*, which contains only one member, the word from language D. Once the morpheme *c* is marked as salient, the word from language D is also removed from cluster *d*, leaving all words assigned exactly one salient morpheme.

The procedure should be taken with some care, since its greediness can easily lead to an overcounting of affixes. In order to preprocess a dataset first and later correctly annotate it manually, however, it has proven useful to us.

3.2.2 Identifying Potential Cases of Homoplasy and Character Dependencies

It is challenging if not impossible for the time being to design algorithms that directly tell apart homoplasy or character dependence. However, we provide two evaluation methods to "flag" the concepts, which may lead to different word cognate sets between different conversion methods, and further influence the subsequent phylogenetic analysis.

The first method is based on the automated comparison of different methods for the conversion of partial to full cognate sets. This method works for all datasets in which partial cognate sets have been identified, regardless of whether partial cognates have been identified within meaning slots or cross-semantically. The approach is extremely straightforward. We first automatically compute strict cognates from the partial cognates in our dataset and then compute loose cognates from the same data. In a second step, strict and loose cognate sets are systematically compared with the help of B-Cubed scores (Amigó et al., 2009), which are typically used to compare how well an automated cognate detection method performs in comparison to a gold standard (List et al., 2017; Hauer and Kondrak, 2011). B-Cubed scores come in the form of

precision, *recall*, and their *harmonic mean*, the *F-scores*, ranging between 0 (completely different clusters) and 1 (identical clusters). List (2013) details the B-Cubed algorithm and the calculation is implemented in the LingPy Python library (List et al., 2019b). By ranking the concepts in a given dataset according to the differences in the F-scores computed for strict and loose cognates, we can identify the extreme cases in which the conversion of partial to full cognates causes trouble. Using strict and loose cognate conversion is specifically useful in this context, since the approaches represent two extremes. Our second evaluation method requires partial cognates to be consistently identified across meaning slots in a given dataset. In contrast to the method based on cluster comparison, it systematically takes language-internal information into account.

The method proceeds in two stages. In a first stage, we iterate over the wordlist and count for each distinct morpheme and each language in our data in how many concepts it recurs. In a second stage, we summarize the *cross-semantic partial cognate statistics* on the word level for each concept by first averaging the number of cross-semantic partial cognates for each individual word and then averaging the individual word scores for an entire meaning slot. The score for individual words starts from 1 (a cognate set occurs one time in the data set for the given language) and has a theoretical maximum of the size of the concept list (a cognate set occurs in all words for a given language). We subtract 1 from this score in order to make sure that the store starts from zero. The resulting score thus ranges between 0 and the length of the concept list minus 1 and allows us to identify those concepts in which most cross-semantic partial cognates occur. Since the identification of cross-semantic partial cognates can be tedious, the method may not be available in the early stages of data curation. Once cross-semantic partial cognates have been identified, however, the method can be very helpful, since it accounts for cases in variation that might not be spotted by the method based on cluster comparison.

3.2.3 Annotating Salient Cognate Sets

Our methodology is oriented towards a *computer-assisted* as opposed to a pure *computer-based* workflow, because we acknowledge the difficulty of identifying full cognates in comparative wordlists automatically. This requires – in addition to providing code that may help to detect inconsistencies in the data – that we also discuss and test options to manually refine a dataset that was computationally preprocessed. We have presented our main idea for the annotation of *salient partial cognate sets* in Section 2.1. While this annotation can be theoretically done in a simple text file or with the help of a spreadsheet editor, we used the web-based EDICTOR tool for the creation and curation of etymological datasets (`https://digling.org/edictor`, List 2017, 2021) which has

recently added a function that allows for an improved handling of morpheme glosses. Once partial cognates and morpheme glosses have been annotated, scholars can toggle the saliency of individual morphemes in a word by right-clicking them with the mouse in the EDICTOR interface. This will add an underscore (which we use as a marker of non-salient cognate sets in our code) to the respective morpheme gloss and also change its visual appearance by increasing the transparency. Once a dataset has been annotated in the form described here, the conversion of partial to full cognates can be done in a rather straightforward way. Our algorithm proceeds in two steps. In a first step, it iterates over all cognate sets and removes all those cognate sets which have been annotated as non-salient. In a second step, we use the remaining cognate sets to compute strict cognate sets, as discussed above.

3.3 Results

We applied the methods described above to the newly compiled dataset for Chinese dialects in order to investigate to which degree an extensive amount of partial cognates could impact upon phylogenetic reconstruction analyses. In the following, we will discuss our experiments in detail. We start from our heuristics for the identification of meaning slots susceptible to high variation due to partial cognacy (3.3.1). We then investigate to which degree highly variable cognate sets may impact the calculation of distance matrices (3.3.2), and conclude by comparing phylogenetic trees reconstructed from distance matrices for all four conversion methods discussed in this study (3.3.3).

3.3.1 Identifying Concepts Susceptible to High Variation due to Partial Cognacy

The upper part of Table 4 shows the 10 concepts with the lowest B-Cubed F-Scores, derived from the comparison of strict and loose partial cognates in the dataset (full table is provided in our Supplementary Material). As can be seen from the table, concepts with high variation mostly comprise certain nouns which tend to have a complex motivation structure in the Chinese dialects (‘knee’, ‘neck’, ‘wing’, etc.) a few complex verbs (‘live’, ‘swim’), as well as demonstrative pronouns (‘here’), which tend to vary greatly among Chinese dialects. The lower part of the table shows 10 out of 100 examples in which F-Scores reach 1.0, indicating that there is nodifference between strictly and loosely converted cognate sets. Here, we find mostly those concepts which are expressed by monosyllabic words in the Chinese dialects, including specifically most adjectives (‘yellow’, ‘wet’), most basic verbs (‘wash’, ‘walk’), and some very basic nouns (‘wind, ‘water’). All in all, these results are not surprising, but they prove the usefulness of our very simple approach to identify those cognate sets which could

cause problems in later phylogenetic analyses.

Concept	Chinese	Pīnyīn	F-Score
knee	膝 | 膝蓋	*xī* | *xī gài*	0.17
child	孩 | 孩子	*hái* | *hái zi*	0.21
neck	脖子 | 頸	*bó zi* | *jǐng*	0.31
here	这里 | 这	*zhè lǐ* | *zhè*	0.33
wing	翅膀 | 翅	*chì bǎng* | *chì*	0.35
live (alive)	活着 | 活的	*huó zhe* | *huó de*	0.37
rope	繩子 | 繩	*shéng zi* | *shéng*	0.40
breasts	奶子 | 乳房	*nǎi zi* | *rǔ fáng*	0.41
night	晚上 | 夜下	*wǎn shàng* | *yè xià*	0.41
nose	鼻子 | 鼻	*bí zǐ* | *bí*	0.43
...	...	...	...
turn	转	*zhuǎn*	1.00
two	二 | 兩	*èr* | *liǎng*	1.00
walk	走 | 行	*zǒu* | *xíng*	1.00
wash	洗	*xǐ*	1.00
water	水	*shuǐ*	1.00
wet	湿 | 潮	*shī* | *cháo*	1.00
white	白	*bái*	1.00
wide	寬 | 闊	*kuān* | *kuò*	1.00
wind	風	*fēng*	1.00
yellow	黄	*huáng*	1.00

Table 4: Upper and lower part of the comparison of B-Cubed F-Scores between loosely and strictly derived cognate sets. Ten concepts with lowest B-Cubed F-Scores are shown in the upper part of the table, ten concepts with highest F-Scores of 1.0 are shown in the lower part of the table. Column *Chinese* shows the three most frequent exemplary reflexes in Chinese for the given concept slot, *Pīnyīn* shows the pronunciation in Mandarin Chinese using Pīnyīn transliteration.

The results of our test on cross-semantic partial cognates are given in table 5, again showing the ten concepts which showed the highest average number of colexifications per word and per concept slot in the upper part of the table and ten concepts (out of 63) for which no colexifications could be identified throughout all words. Among the words which show a high colexification of certain of their parts, we find mostly nouns which share suffixes across the dataset, such as the suffix *zǐ* 子, which originally meant 'son' and often serves for the derivation of nouns. Looking at the cases with no cross-semantic partial cognates, it is difficult to find a clear pattern, apart from a tendency

Concept	Chinese	Pīnyīn	Score
belly	肚子 | 肚	*dù* ***zi*** | *dù*	3.95
seed	种子 | 种	*zhǒng* ***zi*** | *zhǒng*	3.71
rope	繩子 | 繩	*shéng* | *shéng* ***zi***	3.62
guts	腸子 | 腸	*cháng* ***zi*** | *cháng*	3.45
nose	鼻子 | 鼻	*bí* ***zi*** | *bí*	3.27
sand	沙 | 沙子	*shā* | *shā* ***zi***	3.11
claw	爪 | 爪子	*zhǎo* | *zhǎo* ***zi***	2.78
neck	脖子 | 頸	*bó* ***zi*** | *jǐng*	2.77
leaf	叶子 | 叶	*yè* ***zi*** | *yè*	2.75
person	人	*rén*	2.63
...	...	...	...
back	背 | 背骶身	*bèi* | *bèi dǐ shēn*	0
bad	壞 | 痞	*huài* | *pǐ*	0
because	因为 | 庸乎	*yīn wéi* | *yōng hū*	0
bite	咬	*yǎo*	0
blood	血	*xuè*	0
blow	吹	*chuī*	0
burn	烧	*shāo*	0
cloud	云 | 云彩	*yún* | *yún cǎi*	0
count [verb]	數 | 點	*shù* | *diǎn*	0
die	死 | 瓜	*sǐ* | *guā*	0

Table 5: Top 10 concepts with high scores in the test on cross-semantic partial cognate statistics(Overall ranking).

to monosyllabic words, which will naturally decrease the chance of a word of showing at least one part which colexifies across the data selection.

Identifying cases in which high variation results from suffixation is important, since suffixes should definitely be ignored when deriving full cognate judgments from partial cognate judgments, since they constitute an independent, often not very regular process that does not bear much phylogenetic signal. Since our morpheme glosses which we added to the data indicate these cases, we can ignore them from the calculation and only inspect those cases where variation is not due to suffixation. These results are additionally shown in Table 6. As can be seen from this table, the highest scoring concept is 'person', typically expressed as *rén* 人 in Chinese. The word recurs in many words denoting specific kinds of persons, such as 'woman', typically expressed as *nǚ-rén* 女人, or 'man', typically expressed as *nán-rén* 男人. Additional concepts with high potential of being expressed by morphemes that are reused to express other concepts

are 'water' 水, which often recurs in words for 'fruit' (*shǔi-gǔo*, lit. 'water-fruit' 水果), and 'bark' whose lexical motivation is 'tree-skin' (*shù-pí* 树皮) in almost all Chinese dialects.

Concept	Chinese	PinYin	Score
person	人	*rén*	2.58
hit	打 | 拍	*dǎ* | *pāi*	1.95
old	老	*lǎo*	1.85
water	水	*shuǐ*	1.37
bark	树皮	*shù pí*	1.34
woman	女人 | 女的	*nǚ rén* | *nǚ dí*	1.34
man	男人 | 男的	*nán rén* | *nán dí*	1.33
they	他們 | 俚🄵	*tā mén* | *lǐ dǔ*	1.23
husband	老公 | 丈夫	*lǎo gōng* | *zhàng fū*	1.15
we	我們 | 侬家	*wǒ mén* | *nóng jiā*	1.14

Table 6: Top 10 concepts with **high** scores in the test on cross-semantic partial cognate statistics. This table lists the concepts whose reflexes contain no derivation.

All in all the results are not identical with the ones reported in Table 4 above, but they show some similar tendencies with respect to monosyllabicity. This similarity in the rankings of concepts can also be computed. Using Kendall's τ correlation coefficient test, we find a weak negative association between the results of the two rankings (Kendall's τ coefficient: -0.47, p-value < 0.001). The fact that both tests only correlate weakly emphasizes how important it is to use both of them when investigating the potential impact of partial cognates on lexical phylogenies.

3.3.2 Partial Cognates and Language Distances

Having shown that we can identify quite a few concepts in the Sinitic data in which compounding patterns are so complex that they make the conversion of partial into full cognate set difficult, we wanted to analyze to which degree this may influence the computation of lexical distances between languages. We therefore computed distance matrices, following classical lexicostatistical methodology (counting shared cognates per meaning slot) for both strictly and loosely converted cognate sets as well as the two new approaches, the conversion by common morphemes, and the conversion by salient morphemes, which we introduced along with this study. In order to get a better impression on the theoretical impact which partial cognates can have on lexi-

cal distance computation, and the differences between the individual partial cognate conversion schemes, we prepared two distance matrices, one in which only those 66 concepts for which the B-Cubed F-Scores would be 0.8 or less were employed, and one where all data were employed.

In order to compare the four distance matrices which we received from this procedure, we used the traditional Mantel test (Mantel, 1967), which calculates the correlation between distance matrices by means of a permutation method, using 999 permutations per run and the Pearson correlation coefficient as our correlation measure. The correlation scores of the Mantel test fall between -1 and 1, with -1 indicating high negative correlation and 1 indicating high positive correlation, and 0 indicating no correlation.

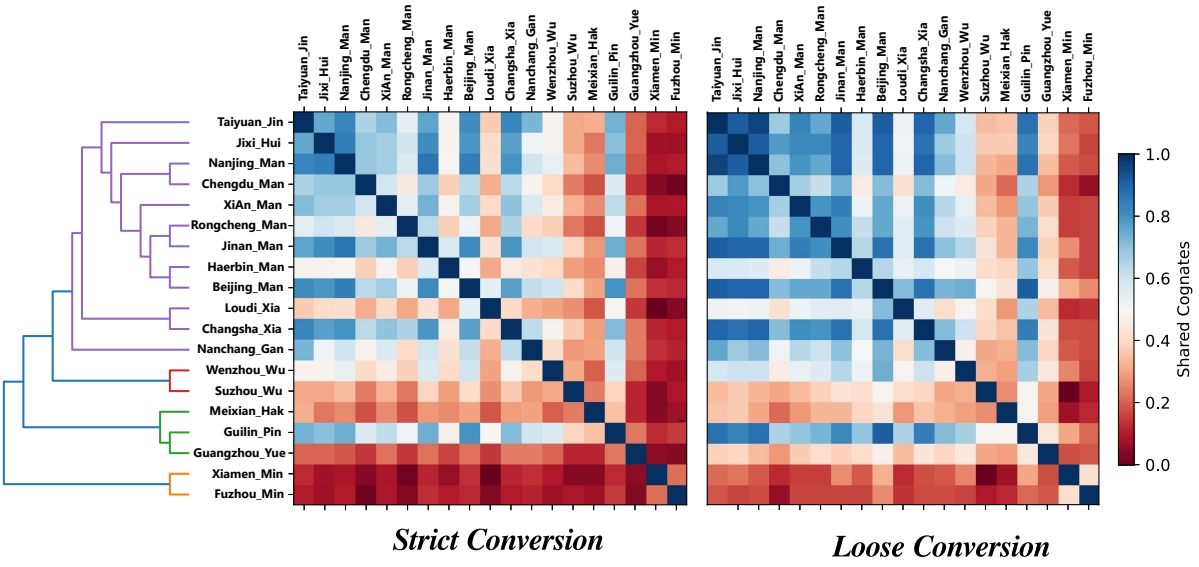

Figure 1: Comparing the pairwise similarities in strictly (left) and loosely (right) converted partial cognate sets for the dialects in our sample. The reference phylogeny is based on the classification by Sagart (2011) for the seven major dialect groups, further extended to include all ten dialect groups and subgrouping inside the groups by List (2015).

Table 7 shows the result of this comparison. While the correlations are extremely high when taking the full datasets into account, we find more fine-grained differences when inspecting only the subsets. The loose and strict conversion schemes show the highest difference, with a (still high) correlation of 0.81. Our salient morpheme conversion (which is based on the hand-curated assignment of salient as opposed to non-salient morphemes in the data) comes second with respect to its difference to the loose coding scheme and a score of 0.87. The highest correlation between distance matrices can be observed for the salient morpheme scheme and the automated common

morpheme coding scheme, with a score of 0.97.

	Subset	Full Dataset
Loose vs. Strict	0.81	0.97
Loose vs. Common morpheme	0.88	0.99
Loose vs. Salient morpheme	0.87	0.99
Strict vs. Common morpheme	0.93	0.97
Strict vs. Salient morpheme	0.97	0.98
Common morpheme v.s. Salient morpheme	0.97	0.99

Table 7: Mantel tests of distance matrices derived from a subset of highly divergent concepts (Subset) and from considering the full data (Full Dataset). Mantel tests were calculated from 999 permutations, using the Person correlation coefficient as the correlation measure. Significance scores are not provided, here, since all permutation tests showed a p-value lower than 0.001, but they are available in the accompanying Supplementary Material.

Although the correlations between the different coding schemes are all high, even for our worst-case subset, the matrix comparison offers us some clearer insights into the specifics of the different conversion schemes. With the strict and the loose conversion schemes representing two extremes, our two new approaches, the automated conversion by common morphemes, and the hand-curated conversion by salient morphemes take places between the two extremes, with the salient morpheme conversion – in the way in which it was practiced by us – coming a bit close to the strict conversion than the common morpheme conversion.

In order to explore the differences between strictly and loosely converted partial cognates, we visualized the results with the help of heatmaps, shown in Figure 1, where we compare pairwise similarities between the dialects (measured by counting shared cognates) for the strictly and loosely converted partial cognates, using the classification of the seven standard dialect groups by Sagart (2011), later adjusted for subgroups and additional dialect groups by List (2015) as our reference tree. As can be seen from this table, we have to deal with a lot of reticulation in this dataset, as reflected in the fact that certain dialects, such as Guīlín (assigned to the Pínghuà group in the source of Liú et al. 2007), or Wēnzhōu (a traditional Wú dialect) show high similarities with the Northern dialects (Mandarin and Jìn) in the sample. We also observe considerably low similarity scores between dialects which are traditionally assigned to the same dialect groups, such as Lǒudì and Chángshā (Xiāng group). Detailed reasons for these skewed similarities need a thorough comparison of the individual cognate

sets which would go beyond the scope of this paper. However, that the history of the Chinese dialects is intertwined and contains many reticulate events, has been observed in many previous studies (Norman, 2003; List et al., 2014) and should not surprise us too much in this context.

The differences between the two matrices in Figure 1 are striking, but difficult to assess from the direct comparison. All in all, and also due to the specific conversion scheme, the loose conversion yields much higher similarity scores than the strict conversion. In Figure 2, we have tried to visualize these by plotting the differences in the observed distances for strict and loose cognate conversion. We can see that specifically the Southern dialects (Mǐn, Yuè, and Hakka), show the largest differences to the other dialects in both conversion methods. The reason for these huge differences which can reach 20% in some extreme cases, can be found in the difference between the word structures in Northern and Southern Chinese dialects. While Northern dialects tend to have more compound words, we find considerably more monosyllabic items in the Southern dialects. Since the dialects still employ the same inherited word material, but differ with respect to compounding, the strictly method will overrate their divergence, while the loose conversion method will give the impression of more similarity.

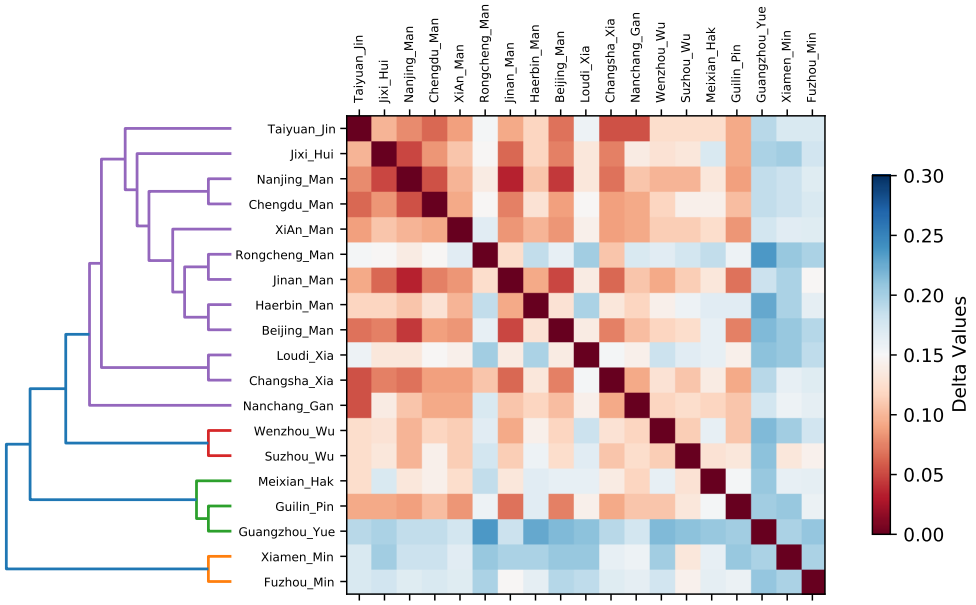

Figure 2: Differences in shared cognate sets between *loosely* and *strictly* converted cognate sets.

3.3.3 Partial Cognates and Language Phylogenies

Having analyzed the differences between the distance matrix retrieved from cognate sets derived from partial cognates using different conversion methods, we find that

there is high correlation between all distance matrices when looking at the dataset as a whole, while these correlations drop (albeit not much) when taking only those concepts into account which we automatically identified as diverse. What remains to be investigated is whether these differences in the distance matrices have a direct impact on the computation of phylogenetic trees. In order to explore this, we computed Neighbor-joining trees from the full dataset containing all concepts. Here, the pairwise language distances were derived from the presence-absence matrix of cognate sets (see Atkinson and Gray 2006) for each conversion method (taking the Hamming distance, which counts the number different items in to vectors of equal length, see Hamming 1950). Computing the presence-absence matrix for all cognate sets allowed us to measure the robustness of the inferred branches by means of the classical bootstrap procedure (Soltis and Soltis, 2003). In a repeated number of trials, the original presence-absence matrix is modified by deleting 30% of the rows and replacing them by duplicating randomly selected items of the remaining rows. The modified matrix is then used to compute a phylogenetic tree using the same approach as originally used, and the splits (branches that divide the tree into two) induced by this new tree are stored for each trial. The overall robustness of the branches is then computed by dividing the number of times a split from the phylogeny computed from all the data by the number of trials.

Figure 3 shows the results of this trial by contrasting the topologies and the branch support for each conversion procedure. All phylogenies were rooted by taking the archaic Mǐn dialects (Fúzhōu and Xiàmén) as outgroup. As can be seen from this figure, all phylogenies suffer from low branch supports, and none of the phylogenies recovers all subgroups perfectly, and the only subgroup which is perfectly identified in all approaches are the Mǐn dialects. While the phylogenies based on the strict and the loose conversion methods recover the Wú dialects, both assign Guǎngzhōu and Tàiyuán to the Mandarin group, and the phylogeny derived from the strict conversion method even places Chángshā and Nánchāng as one subgroup of Mandarin. The conversion by common morphemes identifies Mandarin as one group (apart from Tàiyuán), but places Guǎngzhōu, a dialect of Southern Chinese, which should be much closer to the Mǐn, Wú and Hakka dialects, as the first outgroup of the Mandarin branch. The phylogeny based on the salient morpheme conversion method, finally, sets Guǎngzhōu farther apart from the Mandarin dialects than the rest, but still fails to separate Chángshā and Tàiyuán from the other Mandarin dialects. In sum: none of the approaches can convince here. This should, however, also not surprise us, given that the coding procedures we applied did not control for lexical borrowings. Since our goal was to allow for a direct comparison of different conversion methods for partial cognates, we delib-

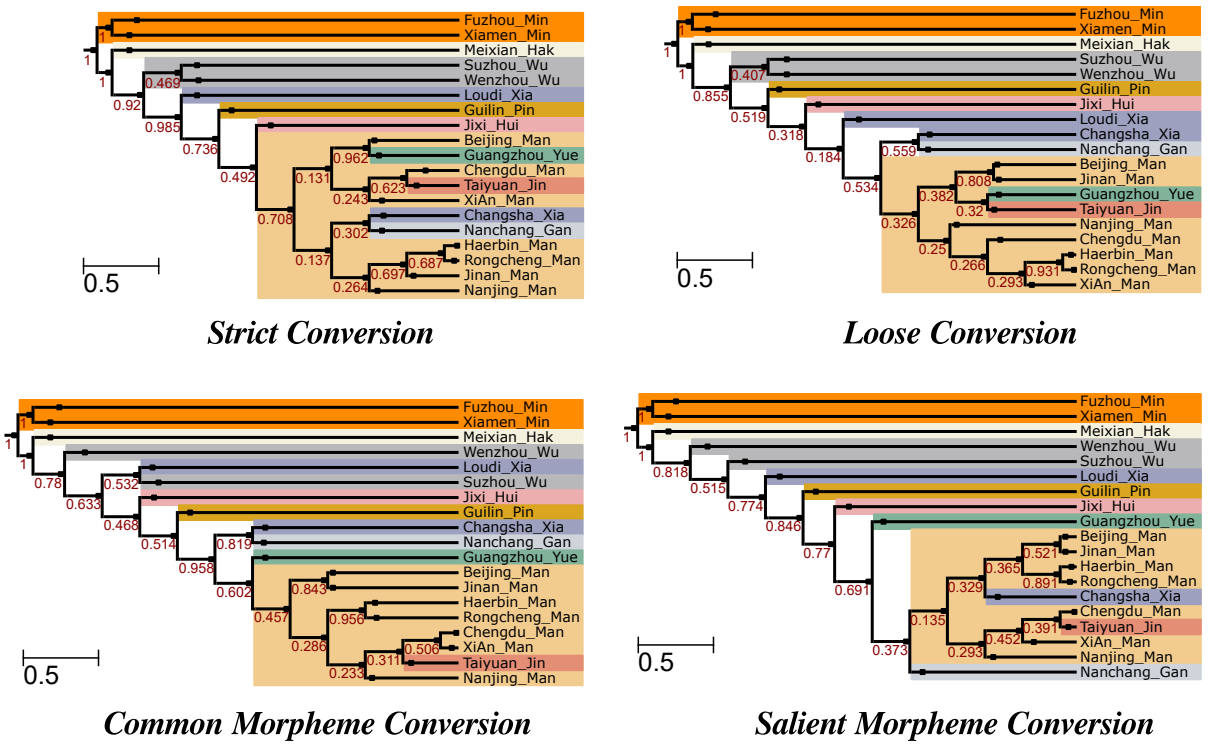

Figure 3: Comparing the tree topologies for the four conversion methods. Major subgroups (Man: Mandarin, Min: Mǐn, Hak: Hakka, Pin: Pínghuà, Hui: Huī, Yue: Yuè, Jin: Jìn, Wu: Wú, Xia: Xiàng, Gan: Gàn) are given the same color.

erately did not modify the data in any additional way, although we know (as can also be seen clearly from the heatmaps in Figure 1 that specifically the data for Guǎngzhōu, but also for Chángshā, Nánchàng, and Wēnzhōu show a large, overproportial amount of shared cognates with Běijīng Chinese, which is the variety closest to Pǔtōnghuà, the standard Chinese variety which has been gaining an enormous prominence in inter-dialectal communication over the past decades. While it is difficult to asses from the topology of the trees alone which conversion method is "better" when it comes to the reconstruction of phylogenetic trees from distance matrices, we can clearly see from this example that despite the rather high correlations between the distance matrices reported in the previous section, the select of the conversion method has an immediate impact on the resulting tree topology and should therefore not be easily ignored when it comes to the preparation of comparative wordlist for the purpose of phylogenetic reconstruction.

Figure 4 shows the unrooted trees with branch lengths for the four conversion methods (Figure 3 only shows the topologies). We can see that the loose conversion method generally leads to much smaller distances between the language varieties, which has the effect that the Mǐn dialects end up to be proportionally more distant from the other

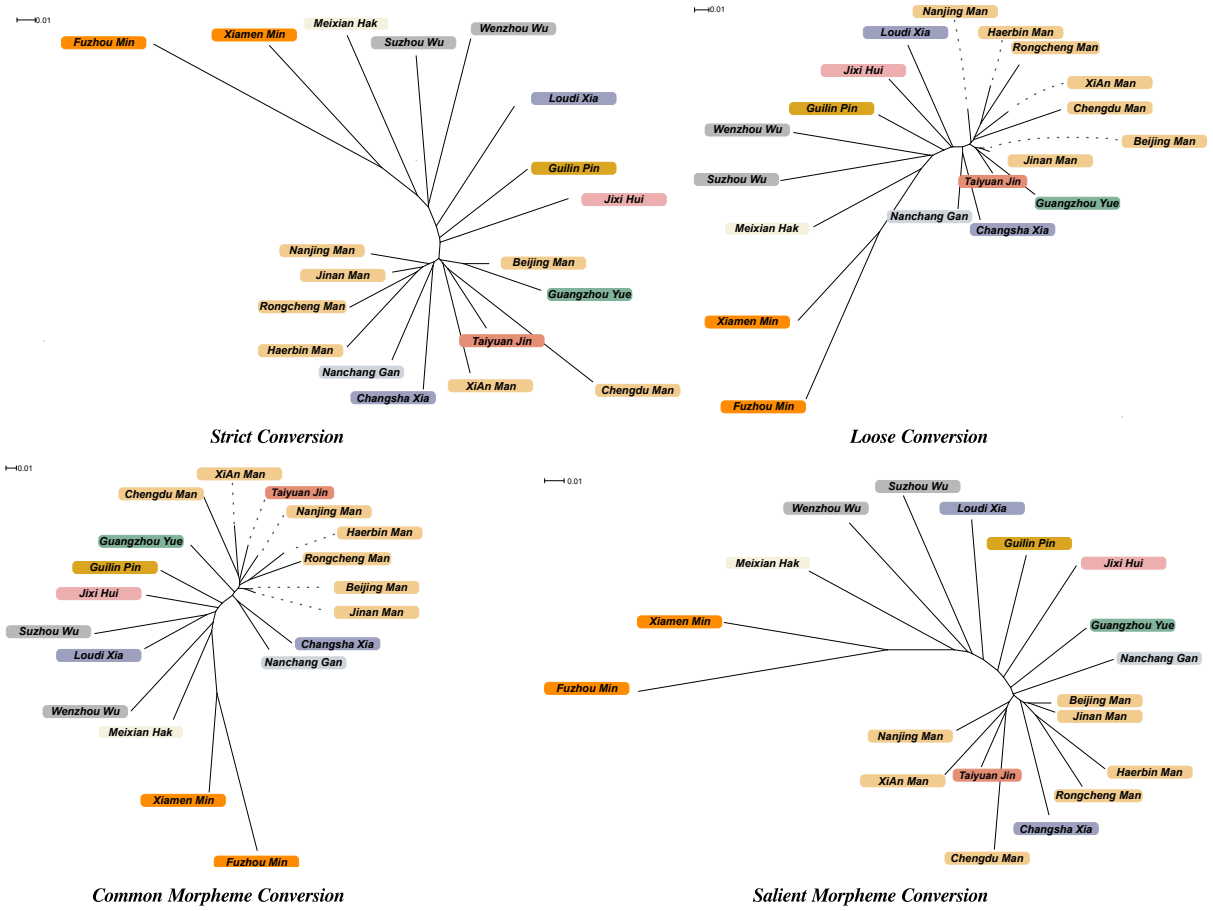

Figure 4: Comparing the unrooted Neighbor-joining trees reconstructed for the four conversion methods.

language varieties. This is not surprising, given that the loose conversion method greedily assigns all words to the same cognate class which share at least one morpheme. Since the Mǐn dialects tend to have many morphemes which cannot be found in the other dialect varieties, their proportional distance increases, because the proportional distance of the other varieties decreases.

In order to assess the overall difference between the tree topologies, we calculated the Quartet Distances (Estabrook et al., 1985; Mailund and Pedersen, 2004) for the topologies inferred from all conversion methods. The results, shown in Table 8 show that the distance between strictly and loosely converted cognate sets is remarkably large (0.44). The same distance can be observed between the loose and the salient morpheme conversion method, which itself shows the lowest (but still remarkable) distance to the strict conversion method (0.25). Since the assessment of salient morphemes is based on our own judgments, it is quite likely that other scholars would come to different solutions. The comparison shows, that we often decided to favor a strict over a loose coding when marking morphemes as salient or not. The auto-

mated common morpheme procedure yields a tree topology which comes closest to our salient morpheme encoding, while being slightly close to the strict than to the loose conversion. That our greedy common morpheme conversion method yields a tree topology which is close to the strict than to the loose conversion procedure may come as a surprise, since the criterion for assigning cognates to the same cognate set is still very lax. It shows, however, that the loose conversion method should generally be taken with great care, since it is well possible that it assigns words to the same cognate set which have no morpheme in common.

	NQD
Loose v.s. Strict Conversion	0.44
Loose v.s. Common Morpheme Conversion	0.40
Loose v.s. Salient Morpheme Conversion	0.44
Strict v.s. Common Morpheme Conversion	0.33
Strict v.s. Salient Morpheme Conversion	0.25
Common Morpheme v.s. Salient Morpheme Conversion	0.33

Table 8: Comparing the differences, using the Normalized Quarted Distance, between the phylogenetic trees inferred from the four conversion methods.

4 Discussion

We have pointed to the general problem of judging whether two words are cognate in the case of compounding. Since compounding is frequent in quite a few language families, notably in South-East Asian languages, we assumed that the methodological uncertainty resulting from words which are only partially cognate may have direct consequences on phylogenetic reconstruction. We reviewed two common approaches to derive "full" cognate judgments from partial cognate judgments, and proposed two more, one automated and one manual approach. In addition, we developed methods that help us to assess which concepts in a comparative wordlist may yield contradictory output depending on the cognate conversion method applied. Applying the methods to a newly compiled dataset of 19 Chinese dialects based on Liú et al. (2007), we could show that although the distance matrices derived from the different conversion methods strongly correlate, they yield quite different tree topologies when analyzing them with the Neighbor-joining algorithm.

Given that none of the four conversion methods could clearly detect all recognized subgroups in the data, it is difficult to assess which method is "good" and which is

"bad" for phylogenetic reconstruction. While we have a strong opinion regarding the loose conversion method, since it may yield cases in which words are assigned to the same cognate set without having any morpheme in common, our study does not make a convincing case for any of the other three conversion methods either. What our study does show, however, is that the differences we observed reflect only the consequences of our decisions regarding the *conversion of partial into full cognate sets*. We did not modify the cognate judgments, and did not compare the cognate judgements of different linguistics, we only modified the way in which partial cognates are handled.

As a result, this means that all analyses in which partial cognates recur frequently (and this includes quite a few language families) should be done with great care. In general, we recommend that scholars should abandon full cognate coding as a common practice for coding cognate sets in comparative wordlist as a rule, and practice partial cognate coding as a first step. Having identified partial cognates, they could then carry out the tests discussed in this study in order to see how much the tree topologies derived from different conversion methods differ. This would of course require much more work in order to code up datasets for phylogenetic reconstruction, but it would also make the process of cognate coding much more transparent than it has been up to now.

5 Outlook

In this study we have tried to show that the problem of cognate coding in languages in which compounding is frequent cannot be easily ignored. We illustrate this with the help of a case study of Chinese dialects which shows that tree topologies can differ drastically, depending on the approaches used to convert partial into full cognates. While we hesitate to recommend any conversion method as the method of choice which should henceforth be used when preparing comparative wordlists for the purpose of phylogenetic reconstruction, we hope that our case study can help to increase awareness among colleagues working in the field of phylogenetic reconstruction that the way in which one derives cognate judgments from comparative data has an immediate impact on the results.

Supplementary Material

The supplementary material contains the source code needed to repeat the analyses described here, and the dataset by Liú et al. (2007), which we used to illustrate the

methods. It has been uploaded to the Open Science Framework where it can be accessed at https://osf.io/2c5m8/?view_only=a3c48d609b18407ab4cf4cfb7564c0a5.

References

E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009.

C. Anderson, T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting*, 4(1):21–53, 2018. doi: https://doi.org/10.2478/yplm-2018-0002. URL https://clts.clld.org.

Q. D. Atkinson and R. D. Gray. How old is the Indo-European language family? Illumination or more moths to the flame? In P. Forster and C. Renfrew, editors, *Phylogenetic methods and the prehistory of languages*, pages 91–109. McDonald Institute for Archaeological Research, Cambridge and Oxford and Oakville, 2006. ISBN 9781902937335.

W. H. Baxter. *A handbook of Old Chinese phonology*. de Gruyter, Berlin, 1992.

Q. Chén. *Miáoyáo yǔwén* 苗瑶语文 *[Miao and Yao language]*. Zhōngyāng Mínzú Dàxué 中央民族大学 [Central Institute of Minorities], Běijīng, 2012. URL https://en.wiktionary.org/wiki/Appendix:Hmong-Mien_comparative_vocabulary_list.

G. F. Estabrook, F. R. McMorris, and C. A. Meacham. Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. *Systematic Biology*, 34 (2):193–200, 06 1985. ISSN 1063-5157. doi: 10.2307/sysbio/34.2.193.

J. Felsenstein. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, 19(1):445–471, 1988. doi: 10.1146/annurev.es.19.110188.002305.

R. Forkel and J.-M. List. Cldfbench: Give your cross-linguistic data a lift. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, page 6997-7004, Luxembourg, 2020. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.864.pdf.

R. Forkel, J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(180205): 1–10, 2018. doi: https://doi.org/10.1038/sdata.2018.205.

M. B. Hamed and F. Wang. Stuck in the forest: Trees, networks and chinese dialects. *Diachronica*, 23(1):29–60, 2006. doi: https://doi.org/10.1075/dia.23.1.04ham.

H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. Glottolog 4.3, 2020. URL `https://glottolog.org/`.

R. W. Hamming. Error detection and error detection codes. *Bell System Technical Journal*, 29(2):147–160, 1950.

B. Hauer and G. Kondrak. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 865–873, 2011.

N. W. Hill and J.-M. List. Challenges of annotation and analysis in computer-assisted language comparison: A case study on burmish languages. *Yearbook of the Poznań Linguistic Meeting*, 3(1):47–76, 2017. doi: https://dx.doi.org/10.1515/yplm-2017-0003.

P. Koch. Lexical typology from a cognitive and linguistic point of view. In *Linguistic typology and language universals*, number 20.2 in Handbook of Linguistics and Communication Science, pages 1142–1178. de Gruyter, Berlin and New York, 2001.

J.-M. List. *Sequence comparison in historical linguistics*. PhD thesis, Heinrich-Heine-Universität Düsseldorf, 2013.

J.-M. List. Network perspectives on chinese dialect history. *Bulletin of Chinese Linguistics*, 8:42–67, 2015. URL `http://booksandjournals.brillMisc.com/content/journals/10.1163/2405478x-00801002`.

J.-M. List. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*, 1(2):119–136, 2016. doi: https://doi.org/10.1093/jole/lzw006.

J.-M. List. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, pages 9–12, Valencia, 2017. Association for Computational Linguistics. URL `http://edictor.digling.org`.

J.-M. List. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics*, 1(45):137–161, 2019. doi: https://doi.org/10.1162/coli_a_00344.

J.-M. List. *EDICTOR. A web-based tool for creating, maintaining, and publishing etymological data*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2021. URL https://digling.org/edictor/.

J.-M. List, S. Nelson-Sathi, W. Martin, and H. Geisler. Using phylogenetic networks to model chinese dialect history. *Language Dynamics and Change*, 4(2):222–252, 2014. doi: https://doi.org/10.1163/22105832-00402008.

J.-M. List, P. Lopez, and E. Bapteste. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*, pages 599–605, Berlin, 2016. URL http://anthology.aclweb.org/P16-2097.

J.-M. List, S. J. Greenhill, and R. D. Gray. The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1):1–18, 2017. doi: https://doi.org/10.1371/journal.pone.0170046.

J.-M. List, C. Anderson, T. Tresoldi, C. Rzymski, S. J. Greenhill, and R. Forkel. Cross-linguistic transcription systems, 2019a. URL https://zenodo.org/record/2633838.

J.-M. List, S. Greenhill, T. Tresoldi, and R. Forkel. Lingpy. a python library for quantitative tasks in historical linguistics, 2019b. URL http://lingpy.org.

J. M. List, C. Rzymski, S. Greenhill, N. Schweikhard, K. Pianykh, A. Tjuka, M.-S. Wu, C. Hundt, T. Tresoldi, and R. Forkel, editors. *Concepticon 2.4.0*. Max Planck Institute for the Science of Human History, Jena, 2020. URL https://concepticon.clld.org/.

L. Liú, H. Wáng, and Y. Bǎi. *Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí* 《现代汉语方言核心词·特征词集》*[Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]*. Fènghuáng 凤凰, Nánjīng 南京, 2007.

T. Mailund and C. N. S. Pedersen. QDist—quartet distance between evolutionary trees. *Bioinformatics*, 20(10):1636–1637, 02 2004. doi: 10.1093/bioinformatics/bth097.

N. W. Mann. *A phonological reconstruction of Proto Northern Burmic*. Phd, The University of Texas, Arlington, 1998.

N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Res*, 27(2):209–220, Feb 1967.

Z. Máo. *Yáozú miǎnyǔ fāngyán yánjiù* 《瑶族勉语方言研究》*[Research on the Mien dialect of the Yao people]*. Mínzú Chūbǎnshè 民族出版社, Běijīng, 2004.

T.-l. Mei. Fāngyán běnzì yánjiū de liǎngzhǒng fāngfǎ〈方言本字研究的两种方法〉. *Wúyǔ Hé Mǐnyǔ de vijiào yánjiū* 《吴语和闽语的比较研究》, 1, 1995.

S. Moran and M. Cysouw. *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press, Berlin, 2018. URL http://langsci-press.org/catalog/book/176.

J. Norman. *Chinese*. Cambridge University Press, Cambridge, 1988.

J. Norman. The sino-tibetan languages. In G. Thurgood and R. J. LaPolla, editors, *The Sino-Tibetan languages*, pages 72–83. Routledge, London and New York, 2003.

M. Ratliff. *Hmong-Mien language history*. Pacific Linguistics, Canberra, 2010.

L. Sagart. Classifying Chinese dialects/Sinitic languages on shared innovations. paperworkshop, 2011. Paper, presented at the Séminaire Sino-Tibétain du CRLAO (2011-03-28).

L. Sagart, G. Jacques, Y. Lai, R. Ryder, V. Thouzeau, S. J. Greenhill, and J.-M. List. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Science of the United States of America*, 116:10317–10322, 2019. doi: https://doi.org/10.1073/pnas.1817972116.

N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987. URL http://mbe.oxfordjournals.org/cgi/reprint/4/4/406.pdf.

D. Satterthwaite-Phillips. *Phylogenetic inference of the Tibeto-Burman languages or on the usefulness of lexicostatistics (and megalo-comparison) for the subgrouping of Tibeto-Burman*. PhD thesis, Stanford University, Stanford, 2011.

N. E. Schweikhard and J.-M. List. Developing an annotation framework for word formation processes in comparative linguistics. *SKASE Journal of Theoretical Linguistics*, 17(1):2–26, 2020. URL http://www.skase.sk/Volumes/JTL43/index.html.

P. S. Soltis and D. E. Soltis. Applying the bootstrap in phylogeny reconstruction. *Statistical Science*, 18(2), 2003.

G. S. Starostin. *Metodologija. Kojsanskie jazyki*, volume 1. Jazyki Russkoj Kulťury, Moscow, 2013.

A. Vittrant and J. Watkins. *The Mainland Southeast Asia Linguistic Area*. De Gruyter Mouton, Berlin and Boston, 2019. doi: 10.1515/9783110401981.

M.-S. Wu, N. E. Schweikhard, T. A. Bodt, N. W. Hill, and J.-M. List. Computer-assisted language comparison. state of the art. *Journal of Open Humanities Data*, 6(2):1–14, 2020. doi: https://doi.org/10.5334/johd.12.

M. M. Yan. *Introduction to Chinese dialectology*. LINCOM Europa, München, 2006.