Annotating Cognates in Phylogenetic Studies of South-East Asian Languages

Mei-Shin Wu and Johann-Mattis List

Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig

May 6, 2022

Abstract

Compounding and derivation are frequent in many language families. As a consequence, 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 an impact on the inferred tree topologies that cannot be ignored. 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

Computational phylogenetic methods in historical linguistics have been gaining popularity of late, and many studies on a diverse range of language families have been published (Gray et al. 2009; Grollemund et al. 2015; Lee and Hasegawa 2011; Sagart et al. 2019). While there were quite a few studies criticizing the new quantitative studies in the beginning (Donohue et al. 2012; Geisler and List 2010; Holm 2007), there have been much less critical accounts recently, although some of the major problems discussed in the earlier literature have not yet been addressed so far. Among these is the problem of *cognate coding*, the representation of cognate words in lexical datasets. Specifically with respect to the coding of *partial cognates*, not many attempts have been made to address the problem, although there are many language families in which partial cognate relations are frequent due to compounding and derivation.

In order to illustrate this problem, consider the cognate judgments by Kolipakam et al. (2018) in Table 1. The authors use strings in the column *Cognate* in order to indicate which word forms they assign to the same cognate set. While this procedure of assigning entire words to cognate sets is common in phylogenetic studies and rarely questioned, a closer investigation of the words assigned to the same cognate set shows that – at least for people who are not experts in Dravidian historical linguistics – is not necessarily easy to understand *where* the words in question are actually cognate. Comparing, for example, word forms like Kota [kanʈiko] with Kurukh [kʰajka], it is obvious that the words are not cognate in their entirety, but since the authors did not provide a morphological analysis, it is not possible for us to see *where* the words are cognate after all, or – more importantly – upon which part of the words the authors base their cognate decisions.

ID	Variety	Form	Cognate
1	Tamil	ularnta	dry-A
2	Telugu	eɳɖu	dry-C
3	Kota	kanʈiko	dry-D
4	Kurukh	kʰajka	dry-D
5	Tamil	kaindadə	dry-D
6	Malto	aːikaː	dry-D
7	Brahui	baːɾun	dry-E
8	Gondi	ʋaʈʈa	dry-E
9	Kannada	battida	dry-E
10	Kannada	oɳagidu	dry-F

Table 1: The word forms of *dry* in a dataset of Dravidian etymologies (Kolipakam 2021).

While the major problem in the Dravidian languages are processes of derivation, which surface in cases where words from different languages share similar roots while derivational suffixes are not necessarily cognate, in other language families, specifically in South-East Asia and South America, the assignment of words to cognate sets is often exacerbated by processes of compounding. Since scholars usually rely on the identification of shared lexical roots in order to assign word forms from different languages to one and the same cognate set, the specific motivation underlying compounds can make it quite challenging to select one part of a compound over the other. In the Chinese dialects, for example, the concept ‘to swim’ can be expressed by different complex forms, such as Xī’ān *fúshŭi* [fu^{24}-fei^{53}] 浮水 (lit. ‘float water’), Chángshā *wánshŭi* [wan^{13}-ɕɥei^{41}] 玩水 (lit. ‘play water’), or Běijīng *yóushŭi* [jou^{35}-ʂwei^{213}] 游水 (lit. ‘wander water’). While all of these verbs share cognate word forms for ‘water’, as well as similar motivations, in so far as they express the concept ‘to swim’ by referring to a concrete action that takes place in water, they differ in the word forms that express the action. From one perspective, one could therefore say that all three word forms are not cognate, since they differ in the main verbs of the phrase, but from another perspective, one might as well argue that the motivation across these varieties is still pretty close, since many languages use a dedicated word form to express the concept ‘to swim’ or they make use of different motivation patterns. No matter how one decides, it becomes clear from this example, that the cognate judgment is not based on the comparison of cognate relations between entire word forms, but rather depends on assumptions regarding the underlying motivation and a – usually – implicit judgment regarding those parts of a morphologically complex word which scholars consider as *representative* or *salient* with respect to the evolutionary process they investigate.

In the concrete practice of phonological reconstruction, scholars often avoid to talk about complex words by shifting the object of comparison from the word to the morpheme. This practice is specifically pervasive in the reconstruction of South-East Asian languages (Mann 1998; Matisoff 2003; Ratliff 2010). In the practice of phylogenetic reconstruction – which typically starts from a list of concepts which are then translated in the target languages before cognate sets inside a given concept slot are identified – complex words cannot be easily ignored. As an example, consider the words for ‘head’ in Tupían languages (South America) in Table 2, taking from the Tupían Lexical Database (Version 0.11, Ferraz Gerardi et al. 2021, see `https://tular.clld.org/parameters/169`). Here, the authors follow Hill and List (2017) and Schweikhard and List (2020) in annotating cognates on the level of the morpheme accompanied by so-called *morpheme glosses* which give hints on the lexical motivation underlying the formation of complex words. As can be seen from the data in the table, there are cases in which ‘head’ is motivated as a compound involving ‘round’ and ‘bone’, but language varieties differ with respect to the details. There are also cases in which ‘head’ is rather interpreted as a simplex word. While assigning cognates on the level of morphemes can again be done in a mostly straightforward manner, it is far from obvious how cognate judgments pertaining to the whole word forms in this example should be derived. Should one assign all words to the same cognate set which show the root glossed as *ROUND* in the example, should one rather insist that words should be cognate with respect to all of their parts, or should one decide on a case-to-case basis?

ID	Variety	Form	Segments	Morpheme	Partial Cognates
1	Akuntsu	anam	a + n a m	ROUND ?	1
2	Amanaye	akɨ	a + k ɨ	ROUND BONE	1 2
3	Amondawa	akaŋ	a + k a ŋ	ROUND BONE	1 2
4	Awetí	ʔaput	ʔ a p + u t	HAIR ?	3 4
5	Arikem	a	a	ROUND	1
6	Cinta-Larga	antar	a n t a r	HEAD	5

Table 2: Partial cognate relations among words for ‘head’ in five Tupían languages.

Given the general importance of handling morphologically complex words in phylogenetic studies in historical linguistics, and the particular pervasiveness of morphologically complex words in South-East Asian language, we have carried out a detailed case study of the impact which different coding practices can have on phylogenies reconstructed from Chinese dialect data. In the following, we will discuss the problem of handling morphologically complex words when assigning words to cognate sets in more detail, proposing ways to increase the transparency of cognate coding (§ 2.1). We will then present the results of a case study on Chinese dialect evolution in which we carry out a detailed comparison of different coding schemes and present simple but efficient data exploration methods that help scholars to identify those parts of their data where morphologically complex words could cause problems.

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 new workflow does not only allow researchers to maintain the link between the original words in which the morphemes occur and the morphemes themselves, but even offers 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* and 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 2016; List et al. 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 criteria upon which phylogenetic models are built, introducing character dependencies may not only impact phylogenetic reconstruction (Felsenstein 1988, p. 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. (ibid.) 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, pp. 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 decisions on cognacy taken by individual scholars. The solution we propose for this task is to make use of *morpheme glosses*, as shown above for the Tupían data in Table 2. Morpheme glosses were first proposed by Hill and List (2017) and further developed by Schweikhard and List (2020). We extend this work by adding one new aspect to the analysis, in so far, as we mark the morpheme or the morphemes which we consider as *salient* with respect to the history of the word in question. Under saliency we understand the potential of one or more morphemes to reflect the major evolutionary processes of the words in which they occur.

As an example, consider the words for ‘head’ in Tupían languages, which can be roughly divided into those words that denote head directly, such as Cinta-Large [antar], words that involve a morpheme for ‘hair’, such as Awetí [ʔap -ut], and words that contain a morpheme that means ‘round’, such as Akuntsu [a-nam] (with [a] glossed as ‘round’). One potential analysis of these partial cognates would be to take ‘round’ as the salient morpheme and to assume that it reflects an innovation in the language family, which was later diversified, leading to various subtypes that can or should be ignored in a phylogenetic analysis. Another possibility would be to say that the specific combination of ‘round’ and ‘bone’ should be treated as the major innovation. In this case, Amanaye [a-kɨ] and Amondawa [a-kaŋ] would reflect one common innovation and therefore treated as one cognate set, while the other words that contain a reflex of ‘round’ but no reflex of ‘bone’ would be kept apart. Table 3 illustrates the consequences of these two decisions regarding the saliency of the morphemes with respect to the evolutionary history of their words.

Variety	Segments	Morpheme	Partial Cognates	# 1	# 2
Akuntsu	a + n a m	ROUND ?	1	1	1
Amanaye	a + k ɨ	ROUND BONE	1 2	1	2
Amondawa	a + k a ŋ	ROUND BONE	1 2	1	2
Awetí	ʔ a p + u t	HAIR ?	3 4	2	3
Arikem	a	ROUND	1	1	4
Cinta-Larga	a n t a r	HEAD	5	3	5

Table 3: Identifying salient morphemes in partial cognates. # 1 and # 2 show two ways to resolve the partial cognate relations to full cognates, the first one taking ROUND to be the sole salient morpheme, while the second one identifies ROUND and BONE as salient morphemes.

This idea of marking those morphemes in the morpheme glosses which one identifies as representative for the word history 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 4. 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}ȡu^{42}	ts ɑ ŋ ^{31} + ȡ 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	ȡu^{22}ŋau^{33}	ȡ u ^{22} + ŋ au ^{33}	knife bent	2 3
Luoxiang	Iu Mien	ȡu^{22}ŋau^{35}	ȡ 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 4: 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 one deems representative for the whole history of the words, 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 5 (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 origin unknown 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 by us 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.

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 following 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.

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	ȵ̥ u ŋ ^{13}	`belly/B`	5	4	1	2
Chuanqiandian (NEY)	ʔ a ^{55} + tɕ^{h} au ^{55}	`_prefix/A belly/A`	4 2	5	1	1

Table 5: 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.

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 'tears' 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-shuǐ* 泪水 'tears', which can be glossed as 'tears + 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 'tears' can thus be described as the 'water' of the 'eye' (Chén 2012, p. 609). Unlike most Chinese dialect varieties, 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 *_mṷɛjH_, and *m*^{24} means 'water', going back to Proto-Hmong Mien *_ʔṷəm_ (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 dialect varieties. The reverted word for 'tears' thus 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 and therefore do not reflect shared innovations in the language family. 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 instances of 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 languages, 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, *ʐe*^{33} ‘I’ vs. *ʐe*^{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 p. 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, to exclude the longer forms from the data.

As we have tried to illustrate throughout this section: it is by no means trivial to deal with these 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 in a straightforward manner, 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, 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 expressed 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 re-code 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, 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 analysis (§ 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 the annotated dataset and 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 Chinese dialect varieties which provide at least one variety as representative for each of the seven major subgroups proposed by Norman (1988, p. 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. 2021b), and the language varieties to Glottolog (`https://glottolog.org`, Hammarström et al. 2021).

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 frequently 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 Sinitic varieties descriptions – to the transcriptions proposed by the Cross-Linguistic Data Formats reference catalog (CLTS, `https://clts.clld.org`, List et al. 2021a, 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 and all individual transcriptions were later manually checked.

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; List 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 used 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). The entire dataset contains a total of 4302 words, including 65.6% of monosyllabic words and 34.4% of polysyllabic words.

Variety	Subgroup	Chinese Name
Běijīng	Mandarin	北京
Chángshā	Xiāng	长沙
Chéngdū	Mandarin	成都
Fúzhōu	Mǐn	福州
Guìlín	Pínghuà	桂林
Guǎngzhōu	Yuè	广州
Hāěrbīn	Mandarin	哈尔滨
Jìxī	Huī	绩溪
Jǐnán	Mandarin	济南
Lóudî	Xiāng	娄底
Méixiàn	Hakka	梅县
Nánchāng	Gàn	南昌
Nánjīng	Mandarin	南京
Róngchéng	Mandarin	荣成
Sūzhōu	Wú	苏州
Tàiyuán	Jìn	太原
Wēnzhōu	Wú	温州
Xī’ ān	Mandarin	西安
Xiàmén	Mǐn	厦门

Table 6: List of Chinese dialect varieties in our sample along with the subgroups they can be assigned to.

The original dataset by Liú et al. (2007) often contains multiple translations for the same concept in the same variety which can easily influence the results of phylogenetic reconstruction approaches. We therefore carefully excluded some of the translations which reflect specific colloquial registers. Following standard practice in phylogenetic reconstruction in historical linguistics, we also made sure to mark known borrowings in the data, relying on our own knowledge of Chinese dialect history as well as cases of borrowings annotated in similar datasets (Sagart et al. 2019). All decisions of the items which were excluded or marked as borrowings are transparently reflected in the data and can be inspected, criticized, and improved in future research.

3.2 Methods

In the following, we present a range of techniques that can be used to detect problems resulting from partial cognacy in phylogenetic reconstruction. Having detected these problems, they can be addressed by refining annotations or excluding concepts with high amounts of variation from an analysis.

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 and Forkel 2021) 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 size, 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 and 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 (Hauer and Kondrak 2011; List et al. 2017). B-Cubed scores come in the form of *precision*, *recall*, and their *harmonic mean*, the *F-scores*, ranging from 0 (completely different clusters) to 1 (identical clusters). List (2013) details the B-Cubed algorithm and the calculation is implemented in the LingPy Python library (List and Forkel 2021). 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 Morphemes

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 morphemes* in 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; List 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 quickly mark whether individual morphemes are considered as "salient" with respect to the history of the languages in question, or not. To classify individual morphemes as salient or not, users just have to right-click the morpheme gloss with the mouse in the EDICTOR interface. This will add or remove an underscore (which we use as a marker of non-salient morphemes 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 dialect varieties in order to investigate to which degree an extensive amount of partial cognates could have an impact on phylogenetic reconstruction analyses. In the following, we will discuss our experiments in detail. We start from our heuristics for the identification of concepts susceptible to high variation due to partial cognacy (§ 3.3.1) and discuss some examples where cognate codings differ, depending on the approach used to make cognate judgments for entire words from partial cognates. We then carry out a systematic comparison of dialect distances resulting from different coding practices (§ 3.3.2) and conclude by investigating how the coding practice influences the results of phylogenetic reconstruction analyses (§ 3.3.3).

3.3.1 Identifying Concepts Susceptible to High Variation

The upper part of Table 7 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 dialect varieties ('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 no difference 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.

The results of our test on cross-semantic partial cognates are given in Table 8, 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 for which no colexifications could be identified throughout all words. 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' (*shuǐ-gǔo*, lit. 'water fruit' 水果), and 'bark' whose lexical motivation is 'tree-skin' (*shù-pí* 树皮) in almost all Chinese dialect varieties. Looking at the cases with no cross-semantic partial cognates, it is difficult to find a clear pattern, apart from a tendency to monosyllabic words, which will naturally decrease the chance of a word of showing at least one part which colexifies across the data under consideration.

All in all the results are not identical with the ones reported in Table 7 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: -0.25, 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.

One can be tempted to assume that our concept of "morpheme saliency" might be replaced by some independent principle, such as, for example, the underlying dependency structure of compound words expressing a given concept. Following this line of argumentation, one could, for example, argue that only heads should be considered as the salient morphemes in a word, or only modifiers. However, due to

Concept	Chinese	Pīnyīn	F-Score
breasts	奶子 | 乳房	*nǎi zi* | *rǔ fáng*	0.35
live (alive)	活着 | 活的	*huó zhe* | *huó de*	0.37
knee	膝蓋 | 膝頭	*xī gài* | *xī tóu*	0.37
here	这里 | 这	*zhè lǐ* | *zhè*	0.39
woman	女人 | 女的	*nǚ rén* | *nǚ de*	0.47
child	孩子 | 孩	*hái zi* | *hái*	0.49
nose	鼻子 | 鼻	*bí zi* | *bí*	0.49
rope	繩子 | 繩	*shéng zi* | *shéng*	0.5
sky	天空 | 天上	*tiān kōng* | *tiān shàng*	0.5
claw	爪子 | 爪	*zhǎo zi* | *zhǎo*	0.51
...	...	...	...
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 7: 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 up to three most frequent exemplary reflexes in Chinese for the given concept slot, *Pīnyīn* shows the pronunciation in Mandarin Chinese using Pīnyīn transliteration.

complexity of lexification processes, head-modifier structures of compounds barely reflect the pathways of lexical motivation. As an example, consider Table 9, where we show how concepts such as 'moon' and 'woman' are expressed in four Chinese dialect varieties in our sample along with the motivation structure underlying the words. The concept 'moon' is expressed as *yuè-liàng* 月亮, literally 'moon-shine', in Mandarin Chinese, with 月 'moon' being the modifier and 亮 'shine' being the head. The concept 'woman' is expressed as *nǚ-rén* 女人, literally 'woman-person', in Mandarin Chinese, with 女 'woman being the modifier and 人 'person' being the head. When comparing how the concepts are reflected across the other varieties, we can quickly see that the archaic varieties in the South of China (Wēnzhōu and Méixiàn) tend to express the concept for 'moon as *yuè-guāng* 月光 'moon-ray', while more innovative Mandarin varieties (Běijīng and Jǐnán) show the Mandarin form 月亮 'moon-shine'. In terms of the motivation underlying this process of lexical change, we therefore find 月, the modifier, as the stable part, while the head of the compound has changed and would therefore be treated as the salient morpheme in our

Concept	Chinese	Pīnyīn	Score
person	人	*rén*	2.47
hit	打 | 拍	*dǎ* | *pāi*	1.95
old	老	*lǎo*	1.6
tree	树 | 树儿	*shù* | *shù ér*	1.53
water	水	*shuǐ*	1.32
bark	树皮	*shù pí*	1.29
woman	女人 | 女的	*nǚ rén* | *nǚ de*	1.17
man	男人 | 男的	*nán rén* | *nán de*	1.16
fight	打架 | 相拍	*dǎ jià* | *xiàng pāi*	1.08
we	我們 | 我竹固哩	*wǒ men* | *wǒ zhú gù lǐ*	1.08
...	...	...	...
back	背 | 背脊	*bèi* | *bèi jǐ*	0
bad	壞 | 否	*huài* | *fǒu*	0
because	因为 | 庸乎	*yīn wéi* | *yōng hū*	0
bird	鳥 | 雀	*niǎo* | *què*	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 [noun]	數	*shù*	0

Table 8: Top 10 concepts with highest scores and lowest scores in the test on cross-semantic partial cognate statistics (Overall ranking).

annotation. Contrasting these cases with the expressions for ‘woman’, we find another situation, with the Mandarin dialects showing the same form, and some Southern dialects showing diverging motivations, like Méixiàn 妹兒人 *mèi-ěr-rén*, ‘sister-suffix-person’ or Wēnzhōu 老娘客 *lǎo-niáng-kè*, ‘old-woman-guest’. While the head stays stable in Méixiàn, we find an innovation with respect to the modifier in both Southern varieties and would therefore annotate the modifier as the salient morpheme. This example shows that the saliency of a morpheme with respect to the history of the word in which the morpheme occurs cannot be determined from the dependency structure alone, although the dependency structure is of crucial importance when it comes to identify the underlying motivation that led to the creation of a compound.

3.3.2 Cognate Coding 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 sets difficult, we wanted to analyze to which degree this may influence the computation of lexical distances between languages. We

Variety	Concept	Segments	Characters	Morphemes
Běijīng	moon	ɥ ɛ ^{51} + l j ɑ ŋ ^{0}	月亮	moon **shine**
Jǐnán	moon	ɥ ɤ ^{21} + l j ɑ ŋ $^{31\ 0}$	月亮	moon **shine**
Wēnzhōu	moon	ɲ y ^{21} + k w ɔ ^{44}	月光	moon **ray**
Méixiàn	moon	ŋ j a t ^{5} + k w o ŋ ^{33}	月光	moon **ray**
Běijīng	woman	n y ^{214} + ʐ ɛ n ^{35}	女人	**female** person
Jǐnán	woman	ɲ y ^{45} + ʐ ẽ ^{53}	女人	**female** person
Wēnzhōu	woman	l ə ^{24} + ɲ j a ŋ ^{341} + k^{h} a ^{41}	老娘客	old **woman** guest
Méixiàn	woman	m oi ^{53} + j e ^{0} + ŋ i n ^{11}	妹兒人	**sister** suffix person

Table 9: The concepts ‘moon’ and ‘woman’ and their inherent motivation structure in four Chinese dialects. Morphemes which we judge as *salient* in this context are marked with bold font.

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 in above. In order to get a better impression on the theoretical impact which partial cognates can have on lexical distance computation, and the differences between the individual partial cognate conversion schemes, we prepared two distance matrices. In one matrix, only those 59 concepts were used for which the B-Cubed F-Scores would be 0.8 or less, and in one matrix all data were used.

In order to compare the two sets of 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, 1 indicating high positive correlation, and 0 indicating no correlation.

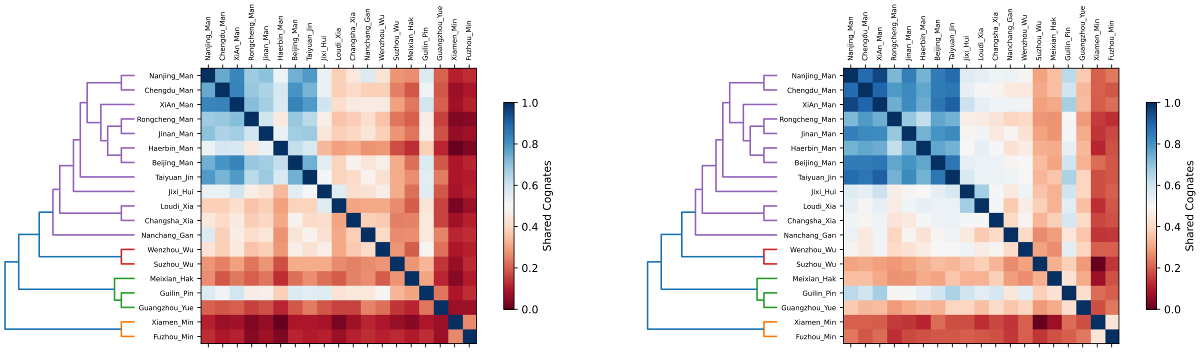
Table 10 shows the result of this comparison. While the correlations are extremely high when taking the full datasets (all 201 concepts) into account, we find more fine-grained differences when inspecting only the subsets. The loose and strict conversion schemes show the highest difference, with a (still high) correlation of 0.71. 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 0.76. The highest correlation between distance matrices can be observed for the salient morpheme scheme and the strict conversion scheme, with a score of 0.96.

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 closer 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

	Subset	Full Dataset
Loose vs. Strict	0.71	0.95
Loose vs. Common morpheme	0.85	0.99
Loose vs. Salient morpheme	0.76	0.97
Strict vs. Common morpheme	0.87	0.96
Strict vs. Salient morpheme	0.96	0.98
Common morpheme vs. Salient morpheme	0.94	0.99

Table 10: 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.



(a) ***Strict conversion*** (b) ***Loose conversion***

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).

this table, we have to deal with a lot of reticulation in this dataset, as reflected in the fact that certain dialects, such as Guìlín (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 Lóudî 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 (List et al. 2014; Norman 2003) 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 and Yuè), show the largest differences compared to the other dialects in both conversion schemes. 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 multisyllabic words with a complex motivation structure, 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 the compositionality of their words, the strict conversion scheme will increase their divergence, while the loose conversion scheme will increase their similarity.

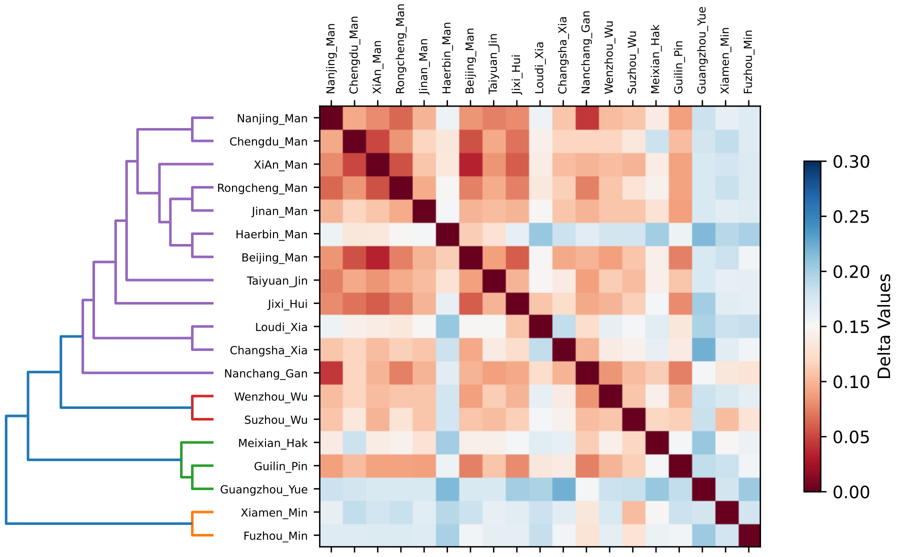
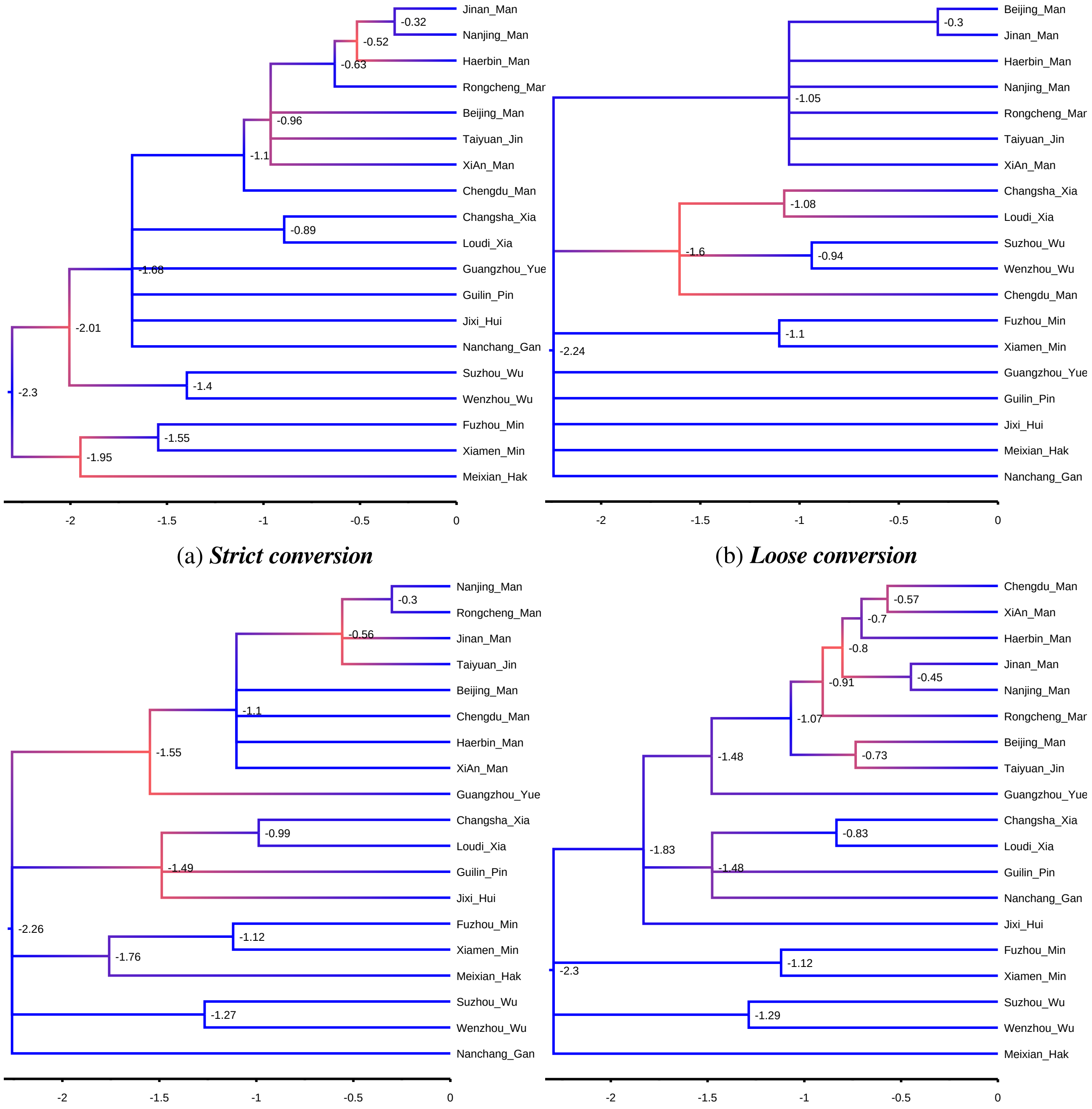


Figure 2: Differences in shared cognate sets between *loosely* and *strictly* converted cognate sets.

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 a high correlation between all distance matrices when looking at the dataset as a whole, while these correlations drop 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 took the cognate sets from the 59 highly diverse concepts and generated four Bayesian phylogenies, one for each of the four conversion schemes, following the standard practice of converting cognate sets to binary presence-absence matrices in which language evolution is modeled as a process of cognate gain and cognate loss (Greenhill et al. 2021).

Bayesian phylogenies have become a standard way of inferring phylogenies from lexical data coded for cognate sets. For our analysis, we used the MrBayes software (Ronquist and Huelsenbeck 2003) and analyzed the data for the four conversion schemes with the help of a fossilized birth-death model (Stadler 2010), commonly used in Bayesian phylogenetic studies applied to linguistic data(Chang et al. 2015; Sagart et al. 2019). In order to make sure we receive comparable results for root ages (also with respect to alternative analyses that have been done on different datasets in the past), we placed the root age between 1500 to 2500 years BP, following a uniform distribution. We had the software generate 20,000,000 different trees in two independent runs from which we sampled every 10,000th tree. Low differences between the trees generated in the independent samples indicated that all four analyses reached convergence. Discarding 10% of the initially generated trees (so-called *burn-in*), we then reconstructed consensus trees from the remaining 1800 trees sampled from each of the two runs.



(a) ***Strict conversion***

(b) ***Loose conversion***

(c) ***Common morpheme conversion***

(d) ***Salient morpheme conversion***

Figure 3: Comparing Bayesian phylogenies (consensus trees) based on our four different conversion schemes. Nodes are annotated with the age of the branching events, branches are colored according to the probabilities, with blue indicating high probabilities and red indicating low probabilities.

Figure 3 displays the consensus phylogenies reconstructed from the different tree samples. As can be seen from the figure, the tree topologies reconstructed from our four conversion schemes vary quite substantially. Thus, while we find that Hakka (Méixiàn) and Mǐn (Xiàmén and Fúzhōu) form a clade in the strict and the common morpheme conversion, they appear in separate groups in the remaining conversion schemes. While the strict conversion phylogeny provides a scenario in which the more archaic dialect groups of Mǐn, Wú, and Hakka – with the exception of Yuè (Guǎngzhōu), which causes problems in all approaches, probably due to the heavy recent contact with Mandarin – split off first, while more innovative groups are established later, this scenario is less supported by the remaining approaches. With the exception of the loose conversion scheme, in which Chéngdū, a Mandarin dialect, is surprisingly clustered with Xiāng and Wú dialects, all schemes basically recover the traditionally proposed dialect subgroups. The only exception is the Jìn group, represented by Tàiyuán, which is heavily disputed among traditional scholars of Chinese dialectology and classified as a Mandarin dialect in alternative proposals, appearing inside the Mandarin group in all four scenarios.

The scenarios also differ quite substantially with respect to the degree to which the trees are *resolved*. While we find a clear binary split at the top of the tree only for the strict conversion scheme, we find star-like top-level branchings of different degree in all other approaches. Here, the loose conversion shows the lowest degree of resolution, failing to resolve 8 branches at the top level, followed by the common morpheme conversion with five branches, and the salient morpheme conversion with four branches.

Given that we fixed the age of the tree, providing divergence dates conforming to traditional assumptions of Chinese dialect diversification, and given that we did not use any internal calibration points, we cannot learn much from the overall tree ages, which are largely the same in all four approaches. However, internal age estimates show some remarkable differences, specifically for the Wú dialect group, where estimates differ by more than 400 years, when comparing the loose conversion estimate of 940 years with the strict conversion estimate of 1400 years. Similarly, the split of the Mǐn varieties of Fúzhōu and Xiàmén is dated at 1550 years in the strict conversion, while the three other conversion methods provide estimates of around 1100 years.

In traditional Chinese historical linguistics, there are different accounts on the overall pattern of Chinese dialect evolution. Norman (2003) assumes that there was a split into three groups, consisting of a Southern group comprising Hakka, Mǐn, and Yuè, a Northern group consisting of the Mandarin dialects (including Jìn), and an intermediate group consisting of Wú, Xiāng, and Gàn dialects. An alternative scenario, specifically propagated by Karlgren (1954) assumes that the Mǐn dialects split off first, and that the other dialects evolved from a *koine* that formed around 600 AD. Sagart (2011) follows Karlgren (and most Chinese dialectologists) in assuming that the Mǐn dialects split off first, but proposes a more complex diversification scenario, in which the other branches split off step by step, starting from Yuè and Hakka, followed by Wú, Gàn, and Xiāng (see List 2015 for details on this scenario).

When comparing these scenarios with the phylogenies based on the four conversion schemes, we can see that all four of them diverge from traditional accounts, most likely due to problems in dealing with the impact of undetected borrowings, large-scale convergence in some of the dialect groups, and due to the fact that the phylogenies were only reconstructed from a small number of concepts susceptible to high variation resulting from lexical compositionality. However, we can also see that the conversion schemes differ regarding the degree to which they diverge from the traditional scenarios. Thus, while the strict conversion scheme conforms in part to the idea of Sagart that Chinese dialect groups split off step by step, the loose conversion scheme proposes a largely star-like diversification of Chinese dialects. While the salient morpheme conversion scheme likewise reflects parts of Sagart's nested scenario in proposing a clade comprising Mandarin, Xiāng, and Gàn (and the highly mixed Pínghuā), the common morpheme comparison only uncovers Mandarin (with Jìn) as a distinct clade, with Gàn as a top-level clade.

4 Discussion

Lexical compositionality creates a considerable problem for the identification of cognate sets in lexicostatistical wordlists. Since processes of derivation and compounding are frequent in the languages of the world and often also include the realm of basic vocabulary, which is predominantly used to reconstruct language phylogenies, we think that it cannot be simply neglected but must be actively taken into account and dealt with if we want to improve current approaches to phylogenetic reconstruction. Given that the problem of lexical compositionality resulting from compounding and derivation is particularly prominent in South-East Asian languages, we conducted an experiment on Chinese dialect evolution by creating a new dataset of Chinese dialects in which partial cognates are annotated in great detail. Assuming that different coding techniques by which cognate judgments for entire words are derived from cognate judgments from cognates annotated for individual morphemes might have a direct impact on phylogenetic reconstruction, we conducted an experiment in which we compared four different coding schemes. Three of these four coding schemes can be automatically derived from data annotated for partial cognates, while one additional coding scheme, which we label “salient morpheme conversion”, requires human assessment. In order to provide guidance in conducting these different forms of data annotation, we developed some basic techniques by which scholars can explore their data in order to identify potential difficulties. Applying the methods to a newly compiled dataset of 19 Chinese dialect varieties, originally collected by Liú et al. (2007), we find that although the distance matrices derived from the different conversion methods strongly correlate, they yield quite different tree topologies when analyzing them with Bayesian methods for phylogenetic reconstruction.

All in all, the differences in the phylogenies allow us to provide a rough ranking of the different approaches to cognate set conversion. We find that the *loose conversion scheme* is performing worst, leading to mostly star-like phylogenies without much resolution, accompanied by clearly wrong groupings of individual varieties, and probably also largely inconsistent age estimates. The reason for these problems lies in the fact that the loose conversion artificially increases similarities between varieties by assigning even words to the same cognate sets which do not share a single cognate morpheme (Hill and List 2017). While the *common morpheme conversion scheme* is to some degree dealing with the problem of low resolution, we find that it yields inconsistent groupings in comparison with traditional accounts. The reason for these problems can be found in the greediness of the approach, which does not further differentiate morphemes with respect to their potential to reflect overall word histories. The *strict* and *salient morpheme conversion schemes* perform best in our opinion, with the strict conversion scheme leading to a higher resolution of the phylogeny, but also to larger divergence estimates for individual subgroups. Specifically in datasets of larger time depths in which diverse language varieties are investigated, the strict conversion scheme might artificially increase the distance among the individual language varieties. As a result, it may be recommendable to code for salient morphemes.

All in all we think that, our study clearly shows that all analyses in which partial cognates recur frequently (and this includes quite a few language families) should be done with great care. Initial cognate annotation should always be done at the morpheme level, ideally including detailed phonetic alignments. Assigning cognate sets to full words should always be based on clear annotation principles. While we know that the conversion of partial cognates to full word cognates is difficult, we think that the techniques for data exploration we provide in this study can definitely help scholars in their concrete annotation practice. Furthermore, by providing a coding techniques that tries to closely reflect how scholars conducted implicit cognate judgments in the past, we hope to contribute to the growing work on *computer-assisted* as opposed to *computer-based* language comparison.

5 Outlook

In this study we have tried to show that the problem of cognate coding in languages in which we find a rich inventory of word formation processes cannot be easily ignored. We illustrated this with the help of a case study of Chinese dialect varieties which shows that tree topologies can differ drastically, depending on the approaches used to convert partial cognates, annotated on the morpheme level, into full cognates, annotated at the word level.

While we hesitate to recommend one particular conversion scheme as the only one to be used in the future, we are convinced that our study shows that certain conversion practices should be taken with great care. Particular practices, like the conversion based on a loose assignment of cognacy (*loose cognate conversion*), or the greedy assignment of words to the same cognate set if they only share at least one common morpheme (*common morpheme conversion*), should be taken with great care. 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

Amigó, E., J. Gonzalo, J. Artiles, and F. Verdejo (2009). "A comparison of extrinsic clustering evaluation metrics based on formal constraints". In: *Information Retrieval* 12.4, pp. 461–486. DOI: 10.1007/s10791-008-9066-8.

Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018). "A Cross-Linguistic Database of Phonetic Transcription Systems". In: *Yearbook of the Poznań Linguistic Meeting* 4.1, pp. 21–53. DOI: https://doi.org/10.2478/yplm-2018-0002. URL: https://clts.clld.org.

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

Chang, W., C. Cathcart, D. Hall, and A. Garrett (2015). "Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis". In: *Language* 91.1, pp. 194–244. DOI: 10.1353/lan.2015.0005.

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

Donohue, M., T. Denham, and S. Oppenheimer (2012). "New methodologies for historical linguistics? Calibrating a lexicon-based methodology for diffusion vs. subgrouping". In: *Diachronica* 29.4, pp. 505–522. DOI: 10.1075/dia.29.4.04don.

Felsenstein, J. (1988). "Phylogenies and quantitative characters". In: *Annual Review of Ecology and Systematics* 19.1, pp. 445–471. DOI: 10.1146/annurev.es.19.110188.002305.

Ferraz Gerardi, F., S. Reichert, C. Aragon, J.-M. List, R. Forkel, and T. Wientzek (2021). *TuLeD: Tupían lexical database. Version 0.11*. Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: 10.5281/zenodo.4629306. URL: https://tular.clld.org/contributions/tuled.

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

Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (2018). "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics". In: *Scientific Data* 5.180205, pp. 1–10. DOI: https://doi.org/10.1038/sdata.2018.205.

Geisler, H. and J.-M. List (2010). "Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics". In: *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik*. Ed. by H. Hettrich. Document has been submitted in 2010 and is still waiting for publication. Wiesbaden: Reichert.

Gray, R. D., A. J. Drummond, and S. J. Greenhill (2009). "Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement". In: *Science* 323.5913, pp. 479–483. DOI: 10.1126/science.1166858.

Greenhill, S. J., P. Heggarty, and R. D. Gray (2021). "Bayesian phylolinguistics". In: *The Handbook of Historical Linguistics. Volume II*. Ed. by R. D. Janda, B. D. Joseph, and B. S. Vance. West Sussex: Blackwell, pp. 226–253.

Grollemund, R., S. Branford, K. Bostoen, A. Meade, C. Venditti, and M. Pagel (2015). "Bantu expansion shows that habitat alters the route and pace of human dispersals". In: *Proceedings of the National Academy of Sciences* 112.43, pp. 13296–13301. DOI: 10.1073/pnas.1503793112. eprint: https://www.pnas.org/content/112/43/13296.full.pdf. URL: https://www.pnas.org/content/112/43/13296.

Hamed, M. B. and F. Wang (2006). "Stuck in the forest: Trees, networks and Chinese dialects". In: *Diachronica* 23.1, pp. 29–60. DOI: https://doi.org/10.1075/dia.23.1.04ham.

Hammarström, H., M. Haspelmath, R. Forkel, and S. Bank (2021). *Glottolog. Version 4.4*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://glottolog.org. URL: https://glottolog.org.

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

Hill, N. W. and J.-M. List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages". In: *Yearbook of the Poznań Linguistic Meeting* 3.1, pp. 47–76. DOI: https://dx.doi.org/10.1515/yplm-2017-0003.

Holm, H. J. (2007). "The new arboretum of Indo-European "trees". Can new algorithms reveal the phylogeny and even prehistory of Indo-European?" In: *Journal of Quantitative Linguistics* 14.2-3, pp. 167–214. DOI: 10.1080/09296170701378916.

Karlgren, B. (1954). "Compendium of phonetics in ancient and archaic Chinese". In: *Bulletin of the Museum of Far Eastern Antiquities* 26, pp. 211–367.

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

Kolipakam, V. (2021). *CLDF dataset derived from Kolipakam et al.'s "DravLex" from 2018*. DOI: 10.5281/zenodo.5121580. URL: https://zenodo.org/record/5121580.

Kolipakam, V., F. M. Jordan, M. Dunn, S. J. Greenhill, R. Bouckaert, R. D. Gray, and A. Verkerk (2018). "A Bayesian phylogenetic study of the Dravidian language family". In: *Royal Society Open Science* 5.171504, pp. 1–17.

Lee, S. and T. Hasegawa (2011). "Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages". In: *Proceedings of the Royal Society B: Biological Sciences* 278.1725, pp. 3662–3669. DOI: 10.1098/rspb.2011.0518.

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

— (2015). "Network perspectives on Chinese dialect history". In: *Bulletin of Chinese Linguistics* 8, pp. 42–67. URL: http://booksandjournals.brillMisc.com/content/journals/10.1163/2405478x-00801002.

— (2016). "Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction". In: *Journal of Language Evolution* 1.2, pp. 119–136. DOI: https://doi.org/10.1093/jole/lzw006.

— (2017). "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*. Valencia: Association for Computational Linguistics, pp. 9–12. URL: http://edictor.digling.org.

— (2019). "Automatic inference of sound correspondence patterns across multiple languages". In: *Computational Linguistics* 1.45, pp. 137–161. DOI: https://doi.org/10.1162/coli_a_00344.

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

List, J.-M., C. Anderson, T. Tresoldi, and R. Forkel (2021a). *CLTS. Cross-Linguistic Transcription Systems*. Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: 10.5281/ZENODO.4705149. URL: https://zenodo.org/record/4705149.

List, J.-M. and R. Forkel (2021). *LingPy. A Python library for quantitative tasks in historical linguistics*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://lingpy.org.

List, J.-M., S. J. Greenhill, and R. D. Gray (2017). "The potential of automatic word comparison for historical linguistics". In: *PLOS ONE* 12.1, pp. 1–18. DOI: https://doi.org/10.1371/journal.pone.0170046.

List, J.-M., P. Lopez, and E. Bapteste (2016). "Using sequence similarity networks to identify partial cognates in multilingual wordlists". In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*. Berlin, pp. 599–605. DOI: 10.18653/v1/P16-2097. URL: http://anthology.aclweb.org/P16-2097.

List, J.-M., S. Nelson-Sathi, W. Martin, and H. Geisler (2014). "Using phylogenetic networks to model Chinese dialect history". In: *Language Dynamics and Change* 4.2, pp. 222–252. DOI: https://doi.org/10.1163/22105832-00402008.

List, J. M., C. Rzymski, S. Greenhill, N. Schweikhard, K. Pianykh, A. Tjuka, C. Hundt, and R. Forkel (2021b). *Concepticon. A resource for the linking of concept lists. Version 2.5.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: 10.5281/zenodo.4911605. URL: https://concepticon.clld.org/. URL: https://concepticon.clld.org/.

Liú, L., H. Wáng, and Y. Bǎi (2007). *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]*. Nánjīng 南京: Fènghuáng 凤凰.

Mann, N. W. (1998). "A phonological reconstruction of Proto Northern Burmic". PhD. Arlington: The University of Texas.

Mantel, N. (1967). "The detection of disease clustering and a generalized regression approach". In: *Cancer Research* 27.2, pp. 209–220.

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

Matisoff, J. A., ed. (2003). *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. University Presses of California, Columbia and Princeton.

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

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

Norman, J. (1988). *Chinese*. Cambridge: Cambridge University Press.

— (2003). "The Sino-Tibetan languages". In: *The Sino-Tibetan languages*. Ed. by G. Thurgood and R. J. LaPolla. London and New York: Routledge, pp. 72–83.

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

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models". In: *Bioinformatics* 19.12, pp. 1572–1574.

Sagart, L. (2011). *Classifying Chinese dialects/Sinitic languages on shared innovations*. Paper presented at the Séminaire Sino-Tibétain du CRLAO (2011-03-28). URL: https://www.academia.edu/19534510/Chinese_dialects_classified_on_shared_innovations.

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

Satterthwaite-Phillips, D. (2011). "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: Stanford University.

Schweikhard, N. E. and J.-M. List (2020). "Developing an annotation framework for word formation processes in comparative linguistics". In: *SKASE Journal of Theoretical Linguistics* 17.1, pp. 2–26. DOI: 10.17613/73w9-x654. URL: http://www.skase.sk/Volumes/JTL43/index.html.

Stadler, T. (2010). "Sampling-through-time in birth–death trees". In: *Journal of Theoretical Biology* 267.3, pp. 396–404. DOI: https://doi.org/10.1016/j.jtbi.2010.09.010. URL: https://www.sciencedirect.com/science/article/pii/S0022519310004765.

Starostin, G. S. (2013). *Metodologija. Kojsanskie jazyki*. Vol. 1. Moscow: Jazyki Russkoj Kul'tury.

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

Wu, M.-S., N. E. Schweikhard, T. A. Bodt, N. W. Hill, and J.-M. List (2020). "Computer-Assisted Language Comparison. State of the Art". In: *Journal of Open Humanities Data* 6.2, pp. 1–14. DOI: https://doi.org/10.5334/johd.12.

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