1 Introduction

There are few disciplines in the humanities that show the impact of quantitative, computer-based methods as strongly as historical linguistics. While individual scholarship and intuition had played a major role for a long time, with only minimal attempts to formalize or automate the painstaking methodology, the last twenty years have seen a rapid increase in quantitative applications. Quantitative approaches are reflected in the proposal of new algorithms that automate what was formerly done by inspection alone [2], in the publication of large cross-linguistic databases that allow for a data-driven investigation of linguistic diversity [3], and in numerous publications in which the new methods are used to tackle concrete questions on the history of the world’s languages (for recent examples, see [4, 5]).

While it is true that – due to increasing amounts of data – the classical methods are reaching their practical limits, it is also true that computer applications are still far from being able to replace experts’ experience and intuition, especially in those cases where data are sparse (as they are still for many language families). If computers cannot replace experts and experts do not have enough time to analyze the massive amounts of data, a new framework is needed, neither completely computer-driven nor ignorant of the assistance computers afford” [1, p. 4]. In this paper, we will illustrate what we consider the current state of the art of computer-assisted language comparison by presenting a workflow that starts with raw data and leads up to a stage where sound correspondence patterns across multiple languages have been identified and can be readily presented, inspected, and discussed. We illustrate this workflow with the help of a newly prepared dataset on Hmong-Mien languages. Our illustration is accompanied by Python code and instructions on how to use additional web-based tools we developed so that users can apply our workflow for their own purposes.

Keywords: computer-assisted; language comparison; historical linguistics; Hmong-Mien language family
introduce what we consider the state of the art in this endeavor, and describe a workflow that starts from raw, cross-linguistic data. These raw data are then consistently lifted to the level of an etymologically annotated dataset, using advanced algorithms for historical language comparison along with interactive tools for data annotation and curation.

2 A workflow for computer-assisted language comparison

Our workflow consists of five stages, as shown in Figure 1. It starts from raw data (tabular data from fieldwork notes or data published in books and articles) which we re-organize and re-format in such a way that the data can be automatically processed (Step 1). Once we have lifted the data to this stage, we can infer sets of etymologically related words (cognate sets) (Step 2). In this first stage, we only infer cognates inside the same meaning slot. That means that all cognate words have the same meaning in their respective languages. Once this has been done, we align all cognate words phonetically (Step 3). Since we only infer cognate words that have the same meaning in Step 2, we now use a new method to infer cognates across meanings by employing the information in the aligned cognate sets (Step 4). Finally, in Step 5, we employ a recently proposed method for the detection of correspondence patterns [7] in order to infer sound correspondences across the languages in our sample.

Our workflow is strictly computer-assisted, and by no means solely computer-based. That means that during each stage of the workflow, the data can be manually checked and modified by experts and then used in this modified form in the next stage of the workflow. Our goal is not to replace human experts, but to increase the efficiency of human analysis by providing assistance especially in those tasks which are time consuming, while at the same time making sure that any manual input is checked for internal consistency.

Our study is accompanied by a short tutorial along with code and data needed to replicate the studies illustrated in the following. The workflow runs on all major operating systems. In addition, we have prepared a Code Ocean Capsule to allow users to test the workflow without installing the software.

3 Illustration of the workflow

3.1 Dataset

The data we use was originally collected by Chén (2012) [8], later added in digital form to the SEALANG project [9], and was then converted to a computer-readable format as part of the CLICS database (https://clics.clld.org) [10]. Chén’s collection comprises 885 concepts translated into 25 Hmong-Mien varieties. Hmong-Mien languages are spoken in China, Thailand, Laos and Vietnam in Southeast Asia. Scholars divide the family into two main branches, Hmong and Mien. The Hmong-Mien languages have been developing in close contact with neighboring languages from different language families (Sino-Tibetan, Tai-Kadai, Austroasiatic, and Austronesian [11, p. 224]). Chén’s study concentrates on Hmong-Mien varieties spoken in China.

In order to make sure that the results can be easily inspected, we decided to reduce the data by taking a subset of 502 concepts of 15 varieties from the dataset. While we selected the languages due to their geographic distribution and their representativeness with respect to the Hmong-Mien language family, we selected the concepts for reasons of comparability with previous linguistic studies. We focus both on concepts that are frequently used in general studies in historical linguistics (reflecting the so-called basic vocabulary [12–15]), and
concepts that have been specifically applied in studies on Southeast Asian languages [4, 16–19]. The 15 varieties are shown in their geographic distribution in Figure 2. While the reduction of the data is done for practical reasons, since smaller datasets can be more easily inspected manually, the workflow can also be applied to the full dataset, and we illustrate in the tutorial how the same analysis can be done with all languages in the original data sample.

3.2 Workflow
3.2.1 From raw data to tokenized data
As a first step, we need to lift the data to a format in which they can be automatically digested. Data should be human- and machine-readable at the same time. Our framework works with data in tabular form, which is usually given in a simple text file in which the first line serves as table header and the following lines provide the content. In order to apply our workflow, each word in a given set of languages must be represented in one row of the data table, and four obligatory values need to be supplied: an identifier (ID), the name of the language variety (DOCULECT), the elicitation gloss for the concept (CONCEPT), and a phonetic transcription of the word form, provided in tokenized form (TOKENS). Additional information can be flexibly added by placing it in additional columns. Table 1 gives a minimal example for four words in Germanic languages.

As can be seen from Table 1, the main reference of our algorithms is the phonetic transcription in its tokenized form as provided by the column TOKENS. Tokenized, in this context, means that the transcription explicitly marks what an algorithm should treat as one sound segment. In Table 1, for example, we have decided to render diphthongs as one sound. We could, of course, also treat them as two sounds each, but since we know that diphthongs often evolve as a single unit, we made this explicit decision with respect to the tokenization.

Transcriptions are usually not provided in tokenized form. The tokenization thus needs to be done prior to analyzing the data further. While one can easily manually tokenize a few words as shown in Table 1, it becomes tedious and error-prone to do so for larger datasets. In order to increase the consistency of this step in the workflow, we recommend using orthography profiles [22]. An orthography profile can be thought of as a simple text file with two columns in which the first column represents the values as one finds them in the data, and the second column allows to convert the exact sequence of characters that one finds in the first column into the desired format. An orthography profile thus allows tokenizing a given transcription into meaningful units. It can further be used to modify the original transcription by replacing tokenized units with new values.

How an orthography profile can be applied is illustrated in more detail in Figure 3.

Our data format can be described as a wide-table format [23–25] and conforms to the strict principle of entering only one value per cell in a given data table. This contrasts with the way in which linguists traditionally code their data, as shown in Table 2, where we contrast the original data from Chén with our normalized representation. To keep track of the original data, we reserve the column VALUE to store the original word forms, including those.

Table 1: A minimal example for four words in four Germanic languages, given in our minimal tabular format. The column VALUE (which is not required) provides the orthographical form of each word [20, 21].

| ID | DOCULECT | CONCEPT | VALUE  | TOKENS |
|----|----------|---------|--------|--------|
| 1  | English  | house   | house  | h aʊ s |
| 2  | German   | house   | Haus   | h au s |
| 3  | Dutch    | house   | huis   | h ʊɪ s |
| 4  | Swedish  | house   | hus    | h ʉː s |

Figure 2: The geographic distribution of the Hmong-Mien languages selected for our sample.
Figure 3: An example to illustrate the usage of orthography profiles to tokenize the phonetic transcriptions.

Table 2: The transformation from raw to machine-readable data. As illustrated in Table 1, the VALUE column displays the raw form. The tokenized forms are added to the TOKENS column.

| English | Chinese | Bana | Numao | Zao Min | Biao Min |
|---------|---------|------|-------|---------|----------|
| moon    | 月亮    | la⁴⁴la³⁵ | to⁴⁴ | lo⁶² | la⁵³gwaŋ³³ |
| sun     | 太陽    | la⁴⁴mi¹³ | ma⁴²ŋaŋ³³ | ra⁵³na⁴⁴ | ni²¹tau³¹ |
| mother  | 母親    | ʔa⁶⁴ŋa³¹³ | mai³³ | ni⁴⁴; ze⁴⁴ | na³¹ |

a) Raw data as given in the digitized version of Chéns (2012) book.

| ID | DOCULECT | SUBGROUP | CONCEPT  | VALUE       | TOKENS       |
|----|-----------|----------|----------|-------------|--------------|
| 1  | Bana      | Hmongic  | moon     | la⁴⁴la³⁵    | 1 a⁰⁴⁴ + 1 a³⁵ |
| 2  | Numao     | Hmongic  | moon     | to⁴⁴        | 1 o⁴⁴        |
| 3  | ZaoMin    | Mienic   | moon     | lo⁶²        | 1 o⁴²        |
| 4  | BiaoMin   | Mienic   | moon     | la⁵³gwaŋ³³  | la⁵³ + g w aŋ |
| 5  | Bana      | Hmongic  | sun      | la⁴⁴ni¹³    | 1 a⁰⁴⁴ + n i¹³ |
| 6  | Numao     | Hmongic  | sun      | ma⁴²ŋaŋ³³   | m a²⁴² + ñ aŋ |
| 7  | ZaoMin    | Mienic   | sun      | ʔa⁵³na⁴⁴   | ʔ a⁵³ + n ai⁴⁴ |
| 8  | BiaoMin   | Mienic   | sun      | ni²¹tau³¹   | ñ i²¹ + t au³¹ |
| 9  | Bana      | Hmongic  | mother   | ʔa⁶⁴ŋa³¹³   | ʔ a⁰⁴⁴ + ñ a³¹³ |
| 10 | Numao     | Hmongic  | mother   | mai³³       | m ai³³       |
| 11 | ZaoMin    | Mienic   | mother   | ni⁴⁴; ze⁴⁴  | n i⁴⁴        |
| 12 | ZaoMin    | Mienic   | mother   | ni⁴⁴; ze⁴⁴  | z e⁴⁴        |
| 13 | BiaoMin   | Mienic   | mother   | na³¹        | n a³¹        |

b) Long-table format in which tokenized forms (TOKENS) have been added, and language names have been normalized.
cases where multiple values are placed in the same cell. The separated forms are placed in the column FORM, which itself is converted into a tokenized transcription with the help of orthography profiles.

In order to make sure that our data is comparable with other datasets, we follow the recommendations by the Cross-Linguistic Data Formats initiative (CLDF, https://cldf.sil.org, [24]) and link our languages to the Glottolog database (https://glottolog.org, [26]), our concepts to the Concepticon (https://concepticon.clld.org, [27]), and follow the transcription standards proposed by the Cross-Linguistic Transcription Systems initiative (CLTS, https://clts.clld.org, [28]).

In the accompanying tutorial, we show how the data can be retrieved from the CLDF format and converted into plain tabular format. We also show how the original data can be tokenized with the help of an orthography profile (TUTORIAL 3.1).

### 3.2.2 From tokenized data to cognate sets

Having transformed the original data into a machine-readable format, we can start to search for words in the data which share a common origin. These etymologically related words (also called cognates) are the first and most crucial step in historical language comparison. The task is not trivial, especially when dealing with languages that diverged a long time ago. A crucial problem is that words are often not entirely cognate across languages [29]. What we find instead is that languages share cognate morphemes (word parts). When languages make frequent use of compounding to coin new words, such as in Southeast Asian languages, partial cognacy is rather the norm than the exception, which is well-known to historical linguists working in this area [30]. We explicitly address partial cognacy by adopting a numerical annotation in which each morpheme, instead of each word form, is assigned to a specific cognate set [31], as shown in Figure 4.

In order to infer partial cognates in our data, we make use of the partial cognate detection algorithm proposed by List et al. [32], which is, so far, the only algorithm available that has been proposed to address this problem. In the tutorial submitted along with this paper, we illustrate in detail how partial cognates can be inferred from the data and how the results can be inspected (TUTORIAL 3.2). In addition, the tutorial quickly explains how the web-based EDICTOR tool (https://digling.org/.tsv/, [33]) can be used to manually correct the partial cognates identified by the algorithm (TUTORIAL 3.2).

#### 3.2.3 From cognate sets to alignments

An alignment analysis is a very general and convenient way to compare sequences of various kinds. The basic idea is to place two sequences into a matrix in such a way that corresponding segments appear in the same column, while placeholder symbols are used to represent those cases where a corresponding segment is lacking (Figure 5) [34]. As the core of historical language comparison lies in the identification of regularly recurring sound correspondences across cognate words in genetically-related languages, it is straightforward to make use of alignment analyses once cognates have been detected in order to find patterns of corresponding sounds. In addition to building the essential step for the identification of sound correspondences, alignment analyses also make it easier for scholars to inspect and correct algorithmic findings.

Automated phonetic alignment analysis has greatly improved during the last 20 years. The most popular alignment algorithms used in the field of historical linguistics today all have their origin in alignment applications developed for biological sequence comparison tasks, which were later adjusted and modified for linguistic purposes [34].

| DOCULECT         | CONCEPT | TOKENS      | COGID | COGIDS |
|------------------|---------|-------------|-------|--------|
| Chuanqiandian    | SUN     | n o 43      | 1     | 1      |
| Numao            | SUN     | m a 42 + n a 33 | 2     | 2 1    |
| ZaoMin           | SUN     | ? a 53 + n ai 44 | 3     | 3 1    |
| EasternBaheng    | SUN     | l a 0/3 + n e 35 | 4     | 4 1    |

**Figure 4:** The comparison of full cognates (COGID) and partial cognate sets (COGIDS). While none of the four words is entirely cognate with each other, they all share a common element. Note that the IDs for full cognates and partial cognates are independent from each other. For reasons of visibility, we have marked the partial cognates shared among all language varieties in red font.
While the currently available alignment algorithms are all very complex, scholars often forget that the same amount of algorithmic complexity is not needed for all languages. Since most Southeast Asian languages have fixed syllable templates, alignments are often predicted by the syllable structure. As a result, one does not need to employ complicated sequence comparison methods in order to find the right matchings between cognate morphemes. All one needs to have is a template-representation of each morpheme in the data.

As an example, consider the typical template for many Southeast Asian languages [35]: syllables consist maximally of an initial consonant (i), a medial glide (m), a nucleus vowel (n), a coda consonant (c), and the tone (t). Individual syllables do not need to have all these positions filled, as can be seen in the following example in Figure 6a.1

Once the templates of all words are annotated, aligning any word with any other word is extremely simple. Instead of aligning the words with each other, we simply align:

**Figure 6:** Illustration of the template-based alignment procedure. a) Representing prosodic structure reflecting syllable templates for each morpheme in the data. b) Aligning tokenized transcriptions to templates, and deleting empty slots.
them to the template, by filling those spots in the template which have no sounds with gap symbols ("-".). We can then place all words that have been aligned to a template in our alignment and only need to delete those columns in which only gaps occur, as illustrated in Figure 6b.

Our accompanying tutorial illustrates how template-based alignments can be computed from the data (TUTORIAL 3.3). In addition, we also show how the alignments can be inspected with the help of the EDICTOR tool (TUTORIAL 3.3).

3.2.4 From alignments to cross-semantic cognates

As in many Southeast Asian languages, most morphologically complex words in Hmong-Mien languages are compounds, as shown in Table 3. The word for ‘fishnet’ in Northeast Yunnan Chuanqianjian, for example, is a combination of the morpheme meaning ‘bed’ [dz³ au³³] and the morpheme meaning ‘fish’ [ª po⁴]. The word for ‘eagle’ in Dongnu is composed of the words [po⁴³] ‘father’ and [tləŋ³³] ‘hawk’. As can be seen from the word for ‘bull’ in the same variety, [po⁴³] [v³³] [po⁴³] can be used to denote male animals, but in the word for ‘eagle’ it is more likely to denote strength [8, p. 328]. As a final example, Younuo lexicalizes the concept ‘tears’ as [ki³² mo³² t³²], with [ki³² mo³²] meaning ‘eye’ and [t³²] meaning ‘water’.

An important consequence of the re-use of word parts in order to form new words in highly isolating languages of Southeast Asia, is that certain words are not only cognate across languages, but also inside one and the same language. However, since our algorithm for partial cognate detection only identifies those word parts as cognate which appear in words denoting the same meaning, we need to find ways to infer the information on cross-semantic cognates in a further step.

As an example, consider the data for ‘son’ and ‘daughter’ in five language varieties of our illustration data. As can be seen immediately, two languages, Chuanqianjian and East Qiandong, show striking partial colexifications for the two concepts. In both cases, one morpheme recurs in the words for the two concepts. In the other cases, we find different words, but if we compare the overall cognacy, we can also see that all five languages share one cognate morpheme for ‘son’ (corresponding to the Proto-Hmong-Mien ‘ʔɛŋ’ in Ratliff’s reconstruction [11]), and three varieties share one cognate morpheme for ‘daughter’ (corresponding to ‘mphje’ in Ratliff’s reconstruction), with the morpheme for ‘son’ occurring also in the words for ‘daughter’ in East Qiandong and Chuanqianjian, as mentioned before.

While a couple of strategies have been proposed to search for cognates across meaning slots [36, 37], none of the existing algorithms is sensitive to partial cognate relations, as shown in Table 4. In order to address this problem in our workflow, we propose a novel approach that is relatively simple, but surprisingly efficient. We start from all aligned cognate sets in our data, and then systematically compare all alignments with each other. Whenever two alignments are compatible, i.e., they have (1) at least one morpheme in one language occurring in both aligned cognate sets, which is identical, and there are (2) no shared morphemes in two alignments which are not identical, we treat them as belonging to one and the same cognate set (see Figure 7). Note that this approach can — by design — only infer strict cognates with different meanings, since not the slightest form of form variation for colexification inside the same language are is allowed. We iterate over all alignments in the data algorithmically, merging the alignments into larger sets in a greedy fashion, and re-assigning cognate sets in the data. The results can be easily inspected with the help of the EDICTOR tool, for example, by inspecting cognate set distributions in the data, as illustrated in detail in the tutorial (TUTORIAL 3.4). When inspecting only those cognate sets that occur in at least 10 language varieties in our sample, the results can be easily inspected with the help of the EDICTOR tool, for example, by inspecting cognate set distributions in the data, as illustrated in detail in the tutorial (TUTORIAL 3.4). When inspecting only those cognate sets that occur in at least 10 language varieties in our sample,

| DOCULECT        | GLOSS     | VALUE   | TOKENS       | MORPHMES       |
|-----------------|-----------|---------|--------------|----------------|
| Northeast-Yunnan-Chuanqianjian | fishnet   | dz³⁶ au³³ mp³³ | dz³⁶ au³³ + p o³³ | fish           |
|                 | fish      | mp³³    | p o³³         | fish           |
|                 | bed       | dz³⁶ au³³ | dz³⁶ au³³     | bed            |
|                 | net       | dz³⁶ o³³  | dz³⁶ o³³      | net            |
| Dongnu          | bull      | po³³ v³³ | p o³³ + v o³³ | father         |
|                 | eagle     | po³³ tləŋ³³| p o³³ + tləŋ³³| father         |
|                 | father    | po³³     | p o³³         | father         |
|                 | bovine    | v³³      | v³³           | cow            |
|                 | hawk      | tləŋ³³   | tləŋ³³        | hawk           |
| Younuo          | tear      | ki³² mo³² 2 ø³⁴ | k i³⁵⁵ + m o³³ + t ø³² | ki-suffix eye water |
|                 | water     | ?ø³⁴     | ?ø³⁴          | water          |
|                 | eye       | ki³² mo³²  | k i³⁵⁵ + m o³³ | ki-suffix eye  |
we already find quite a few interesting cases of cross-semantic cognate sets: morphemes denoting the concept 'one', for example, recur in the words for 'hundred' (indicating that hundred is a compound of 'one' plus 'hundred' in all languages); morphemes recur in 'snake' and 'earthworm' (reflecting that words for 'snake' and 'earthworm' are composed of a morpheme 'worm'); and 'left' and 'right' share a common morpheme (indicating an original meaning of 'side' for this part, such as 'left side' vs. 'right side').

3.2.5 From cross-semantic cognates to sound correspondence patterns

Sound correspondences, and specifically sound correspondence patterns across multiple languages, can be seen as the core objective of the classical comparative method and build the basis of further endeavors such as the reconstruction of proto-forms or the reconstruction of phylogenies. Linguists commonly propose sound correspondence sets, that is, collections of sound correspondences which reconstruct back to a common proto-sound (or sequence of proto-sounds) in the ancestor language, as one of the final stages of historical language comparison. In Hmong-Mien languages, for example, Wang proposed 30 sets [38] and Ratliff reduced the quantity of correspondence sets to 28 [11].

An example for the representation of sound correspondence sets in the classical literature [11] is provided in Table 5. The supposed proto-sound */ntshj/ in
proto-Hmong-Mien is inferred from the initials of four words in 11 contemporary Hmong-Mien languages.

Although this kind of data representation is typical for classical accounts on sound correspondence patterns in historical language comparison, it has several shortcomings. First, the representation shows only morphemes, and we are not informed about the full word forms underlying the patterns. This is unfortunate, since we cannot exclude that compound words were already present in the ancestral language, and it may likewise be possible that processes of compounding left traces in the correspondence patterns themselves. Second, since scholars tend to list sound correspondence patterns merely in an exemplary fashion, with no intent to provide full frequency accounts, it is often not clear how strong the actual evidence is, and whether the pattern at hand is exhaustive, or merely serves to provide an example. Third, we are not being told where a given sound in a given language fits a general pattern less well. Thus, we can find two different reflexes in language 8 in the table, [c] and [dz], but without further information, we cannot tell if the differences result from secondary, conditioned sound changes, or whether they reflect irregularities that the author has not yet resolved.

To overcome these shortcomings, we employ a two-fold strategy. We first make use of a new method for sound correspondence pattern detection [7] in order to identify exhaustively, for each column in each alignment of our data, to which correspondence pattern it belongs. In a second step, we use the EDICTOR tool to closely inspect the patterns identified by the algorithm and to compare them with those patterns proposed in the classical literature.

The method for correspondence pattern identification starts by assembling all alignment sites (all columns) in the aligned cognate sets of the data, and then clusters them into groups of compatible sound correspondence patterns. Compatibility essentially makes sure that no language has more than one reflex sound in all partitioned alignment sites (see [7] for a detailed explanation of this algorithm).

Table 6 provides some statistics regarding the results of the correspondence pattern analysis. The analysis yielded a total of 1392 distinct sound correspondence patterns (with none of the patterns being compatible with any of the other 1392 patterns). While this may seem a lot, we find that 234 patterns only occur once in the data (probably reflecting borrowing events, erroneously coded cognates, or errors in the data). Among the non-singleton patterns, we find 302 corresponding to initials, 74 to medials, 389 to nucleus vowels, 95 to the codas, and 298 to the tone patterns. These numbers may seem surprising, but one should keep in mind that phonological reconstruction will assign several distinct correspondence patterns to the same proto-form and explain the divergence by means of conditioning context in sound change. So far, there are few studies on the numbers of distinct correspondence patterns one should expect, but the results we find for the Hmong-Mien dataset are in line with previous studies on other language families [7]. More studies are needed in order to fully understand what one ought to expect in terms of the numbers of correspondence patterns in datasets of various sizes and types.

While the representation in textbooks usually breaks the unity of morphemes and word forms, our workflow never loses track of the words, although it enables users to look at the morphemes and at the correspondence patterns in isolation. Our accompanying tutorial shows not only how the correspondence patterns can be computed (TUTORIAL 3.5), but also how they can be inspected in the EDICTOR tool (TUTORIAL 3.5), where we can further see that our analysis uncovers the correspondence pattern shown in Table 5 above, as we illustrate in Table 7. Here, we can see that our approach confirms Ratliff’s pattern by clustering initial consonants of cognates for ‘blood’ and ‘fear (be afraid)’ into one correspondence pattern.
4 Discussion

Although our workflow represents what we consider the current state of the art in the field of computational historical linguistics, it is not complete yet, and it is also not perfect. Many more aspects need to be integrated, discussed, and formalized. Based on a quick discussion of the general results of our study, we will discuss three important aspects, namely, (a) the current performance of the existing algorithms in our workflow, (b) possible improvements of the algorithms, and (c) general challenges for all future endeavors in computer-assisted or computational historical linguistics.

4.1 Current performance

Historical language comparison deals with the reconstruction of events that happened in the past and can rarely be directly verified. Our knowledge about a given language family is constantly evolving. At the same time, debate on language history is never free of disagreement among scholars, and this is also the case with the reconstruction of Hmong-Mien.\textsuperscript{30} As a result, it is not easy to provide a direct evaluation of the performance of the computational part of the workflow presented here.

In addition to these theoretical problems, evaluation faces practical problems. First, classical resources on historical language comparison of Hmong-Mien are not available in digital form (and digitizing them would be beyond the scope of this study). Second, and more importantly, even when having recent data on Hmong-Mien reconstruction in digital form, we could not compare them directly with our results due to the difference in the workflows. All current studies merely consist of morphemes that were taken from different sources without giving reference to the original words [31]. Full words, which are the starting point in our study, are not reported and apparently not taken into account. For a true evaluation of our workflow, however, we would need a manually annotated dataset that would show the same completeness in terms of annotation as the one we have automatically produced. Furthermore, since our workflow is explicitly thought of as computer-assisted and not purely computational, the question of algorithmic performance is rather aesthetic than substantial, given that the computational approaches are merely used to ease the labor of the experts.

Nevertheless, to some degree, we can evaluate the algorithms which we assembled for our workflow here, and it is from these evaluations that have been made in the past, that we draw confidence in the overall usefulness of our workflow. Partial cognate detection, as outlined in Section 3.2, for example, has been substantially evaluated with results ranging between 90% (Chinese dialects) and 94% (Bai dialects) compared to expert judgments. The alignment procedure we propose is supposed to work as good as an expert, provided that experts agree on the prosodic structure we assign to all morphemes. For the cross-semantic cognate set detection procedure we propose, we do not yet have substantial evaluations, since we lack sufficient test data. The correspondence pattern detection algorithm has, finally, been indirectly evaluated by testing how well so far unobserved cognate words could be predicted (see also [39]), showing an accuracy between 59% (Burmish languages) and 81% (Polynesian languages) for trials in which 25% of the data was artificially deleted and later predicted.

As another quick way to check if the automated aspects of our workflow are going in the right direction, we can compute a phylogeny based on shared cross-semantic

| Language                  | ‘blood’ | ‘fear’ |
|---------------------------|---------|--------|
| Numao                     | ts\textsuperscript{h} an\textsuperscript{13} | ts\textsuperscript{h} ei\textsuperscript{33} |
| Western Luobuohe          | ts\textsuperscript{h} en\textsuperscript{44} | ts\textsuperscript{h} e\textsuperscript{35} |
| Biao Min                  | s a n\textsuperscript{35} | Ø |
| Zao Min                   | z a m\textsuperscript{24} | z a\textsuperscript{42} |
| Younuo                    | ts\textsuperscript{h} un\textsuperscript{33} | ts\textsuperscript{h} i\textsuperscript{44} |
| Western Xiangxi           | ts\textsuperscript{h} i\textsuperscript{44} | ts\textsuperscript{h} a\textsuperscript{53} |
| Eastern Luobuohe          | ts\textsuperscript{h} en\textsuperscript{44} | ts\textsuperscript{h} e\textsuperscript{24} |
| Bana                      | Ø       | dz i\textsuperscript{13} |
| Eastern Xiangxi           | ts\textsuperscript{h} i\textsuperscript{55} | Ø |
| Western Qiangdong         | c\textsuperscript{h} e\textsuperscript{13} | c\textsuperscript{h} e\textsuperscript{44} |
| Eastern Baheng            | ts\textsuperscript{h} e\textsuperscript{313} | Ø |
| Chuanqianqian             | ts\textsuperscript{h} a n\textsuperscript{55} | ts\textsuperscript{h} al\textsuperscript{44} |
| Western Baheng            | Ø       | Ø |
| Central Guizhou Chuanqian | n\textsuperscript{s} o\textsuperscript{13} | n\textsuperscript{s} e\textsuperscript{42} |
| Eastern Qiangdong         | c a n\textsuperscript{33} | c a\textsuperscript{24} |

Table 7: Cells shaded in blue indicate the initial consonants belonging to a common correspondence pattern, with missing reflexes indicated by a Ø.

\textsuperscript{30} Wu et al: Computer-Assisted Language Comparison
cognates between all language pairs and see if the phylogeny matches with those proposed in the literature. This analysis, which can be inspected in detail in the accompanying tutorial (TUTORIAL 4.2), shows that the automated workflow yields a tree that correctly separates not only Hmongic from Mienic languages but also identifies all smaller subgroups commonly recognized.

4.2 Possible improvements
The major desideratum in terms of possible improvements is the inclusion of further integration of our preliminary attempts for semi-automated reconstruction, starting from already identified sound correspondence patterns. Experiments are ongoing in this regard, but we have not yet had time to integrate them fully. In general, our workflow also needs a clearer integration of automatic and manual approaches, ideally accompanied by extensive tutorials that would allow users to start with the tools independently. This study can be seen as a first step in this direction, but much more work will be needed in the future.

4.3 General challenges
General challenges include the full-fledged lexical reconstruction of words, i.e., a reconstruction that would potentially also provide compounds in etymological dictionaries. This might help to overcome a huge problem in historical language comparison in the Southeast Asian area, where scholars tend to reconstruct only morphemes, and rarely attempt at the reconstruction of real word forms in the ancestral languages [31]. Furthermore, we will need a convincing annotation of sound change that would ideally allow us to even check which sounds changed at which time during language history.

5 Outlook
This article provides a detailed account on what we consider the current state of the art in computer-assisted language comparison. Starting from raw data, we have shown how these can be successively lifted to higher levels of annotation. While our five-step workflow is intended to be applied in a computer-assisted fashion, we have shown that even with a purely automatic approach, one can already achieve insightful results that compare favorably to results obtained in a purely manual approach. In the future, we hope to further enhance the workflow and make it more accessible to a wider audience.

Notes
1 By “state of the art”, we refer to approaches that have been developed during the past two decades and are available in the form of free software packages that can be used on all major computing platforms and have shown to outperform alternative proposals in extensive tests. These approaches themselves build on both qualitative and quantitative considerations that have been made in the field of historical linguistics during the past two centuries (for early quantitative and formal approaches, compare, for example, Hoenigswald [40] and Kay [41]).

2 The permanent link of the Code Ocean Capsule is: https://codeocean.com/capsule/8178287/tree/v2.
3 Orthography profiles proceed in a greedy fashion, converting grapheme sequences in the reverse order of their length, thus starting from the longest grapheme sequence.
4 Linguistic terms which are further explained in our glossary, submitted as part of the supplementary information, are marked in bold font the first time they are introduced.
5 Note that this template of (initial) m(edian) n(ucleus) c(oda) and t(one) in general is generally sufficient to represent all syllables in the Hmong-Mien data we consider here. Seemingly complex cases, such as *ntsæn²² “clear”, for example, can be handled by treating nts as one (initial) sound, resulting in a phonetic transcription of ["ts æ n²²].
6 We are aware of the fact that the transcriptions by Chén are not entirely “phonetic”, but since they are much less phonologically abstract than, for example, the transcriptions provided by Ratliff [11], we prefer to place them in phonetic rather than phonological brackets.
7 In cases of very intensive language contact, one would expect to find recurring correspondence patterns that include borrowings, but in the case of sporadic borrowings, they will surface as exceptions.
8 How this step of identifying conditioning context can be done in concrete is not yet entirely clear to us. Computational linguists often use n-gram representations in order to handle context of preceding and following sounds, but this would not allow us to handle situations of remote context.
9 The other two cognate sets in Ratliff’s data could not be confirmed, because they do not occur in our sample.
10 Compare, for example, the debate about regular epenthesis in Proto-Hmong-Mien among Ratliff [42] and Ostapirat [43].
11 A specific problem in semi-automated reconstruction consists in the importance of handling conditioning context in sound change. To our knowledge, no approaches that would sufficiently deal with this problem have been proposed so far. This reflects one apparent problem of common alignment approaches, as they cannot handle cases of structural equivalence which require information on conditioning context [44].

Supplementary information and material
The appendix that is submitted along with this study consists of two parts. First, there is a glossary explaining the most important terms that were used throughout this study. Second, there is a tutorial explaining the steps of the workflow in detail. In addition to this supplementary information, we provide supplementary material in the form of data and code. The data used in this study is archived on Zenodo (DOI: 10.5281/zenodo.3741500) and curated on GitHub (Version 2.1.0, https://github.com/lexibank/chenhmongmien). The code, along with the tutorial, has also been archived on Zenodo (DOI:
Acknowledgements
This research was funded by the ERC Starting Grant 715618 “Computer-Assisted Language Comparison” (CALC, http://calc.digling.org). MSW, NES, JML, the ERC Synergy Grant 609823 “Beyond Boundaries: Religion, Region, Language and the State” (ASIA, NWH), and the Grant of P2BEP1_181779 “Reconstruction of Proto-Western Khobi-Bwa” of the Swiss National Science Foundation (TAB). The workflow was presented in the workshop “Recent Advances in Comparative Linguistic Reconstruction” in SOAS, London. We thank the workshop participants for giving valuable feedback regarding several aspects of the workflow in their studies. In addition, we thank Christoph Rzymski and Tiago Tesoldi who provided technical support on setting up our Code Ocean Capsule.

Competing Interests
The authors have no competing interests to declare.

Author Contributions
MSW, NWH, and JML initiated the study. MSW, NWH, JML, and TAB drafted the workflow. MSW and JML implemented the workflow. NES wrote the glossary. TAB, NWH and NES tested the workflow on different datasets. MSW and JML wrote the accompanying tutorial. MSW and JML wrote the first manuscript. NES, NWH and TAB helped in revising the manuscript. All authors agree with the final version of the manuscript.

References
1. List J-M. Computer-assisted language comparison: Reconciling computational and classical approaches in historical linguistics [Internet]. Jena: Max Planck Institute for the Science of Human History. 2016. Available from: https://hcommons.org/deposits/item/hc:25045/.
2. List J-M, Greenhill SJ, Gray RD. The potential of automatic word comparison for historical linguistics. PLOS ONE. 2017; 12(1): 1–18. DOI: https://doi.org/10.1371/journal.pone.0170046.
3. Dellert J, Daneyko T, Münch A, Ladygina A, Buch A, Clarinus N, Grorjew I, Balabel M, Boga HI, Baysarova Z, Mühlenbernd R, Wahle J, Jäger G. NorthEuraLex: A wide-coverage lexical database of Northern Eurasia. Language Resources and Evaluation. 2020; 54(1): 273–301. DOI: https://doi.org/10.1007/s10579-019-09480-6
4. Sagart L, Jacques G, Lai Y, Ryder R, Thouzeau V, Greenhill SJ, List JM. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proceedings of the National Academy of Science of the United States of America. 2019; 116(21): 10317–10322. DOI: https://doi.org/10.1073/pnas.1817972116
5. Kolipakam V, Jordan FM, Dunn M, Greenhill SJ, Bouckaert R, Gray RD, et al. A Bayesian phylogenetic study of the Dravidian language family. Royal Society Open Science. 2018; 5(171504): 1–17. DOI: https://doi.org/10.1098/rsos.171504
6. Barrachina S, Bender O, Casacuberta F, Civera J, Cubel E, Khadivi S, Ligarda A, Ney H, Tomás J, Vidal E, Vilar J-M. Statistical approaches to computer-assisted translation. Computational Linguistics. 2008; 35(1): 3–28. DOI: https://doi.org/10.1162/coli.2008.07-055-R2-06-29
7. List J-M. Automatic inference of sound correspondence patterns across multiple languages. Computational Linguistics. 2019; 1(45): 137–161. DOI: https://doi.org/10.1162/coli_a_00344
8. Chén Q. Miàoyáo yúwén 苗瑶语文 [Mao and Yao Language]. Běijīng 北京: Zhòngyāng Miànzú Dàxué 中央民族大学出版社 [Central Institute of Minorities]. 2012. Available from: https://en.wiktionary.org/wiki/Appendix:Hmong-Mien_comparative_vocabulary_list.
9. Cooper D. Data Warehouse, Bronze, Gold, STEC, Software. In: Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages. 2014; 91–99.
10. Rzymski C, Tesoldi T, Greenhill SJ, Wu M-S, Schweikhard NE, Koptjevskaja-Tamm M, Gast V, Bodt TA, Hantgan A, Kaiping GA, Chang S, Lai Y, Morozova N, Arjava H, Hübner N, Koile E, Pepper S, Proos M, Epps B, Blanco I, Hundt C, Monakhov S, Pianykh K, Ramesh S, Gray RD, Forkel R, List J-M. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemy. Scientific Data. 2020; 7(13): 1–12. DOI: https://doi.org/10.1038/s41597-019-0341-x
11. Ratliff M. Hmong-Mien Language History. Canberra: Pacific Linguistics; 2010.
12. Swadesh M. Lexico-statistic dating of prehistoric ethnic contacts: With special book to North American Indians and Eskimos. Proceedings of the American Philosophical Society. 1952; 96(4): 452–463.
13. Swadesh M. Towards greater accuracy in lexico-statistical dating. International Journal of American Linguistics. 1955; 21(2): 121–137. DOI: https://doi.org/10.1086/464321
14. Comrie B, Smith N. Linguistic Descriptive Series: Questionnaire. Lingua. 1977; 42: 1–72. DOI: https://doi.org/10.1016/0024-3841(77)90063-8
15. Liú L, Wáng H, Bǎi Y. Xiàndài Hànyǔ fāngyán héxíncí, tèzhèng cíjì 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: Fēnghuáng 凤凰, 2007.
16. So-Hartmann H. Notes on the Southern Chin languages. Linguistics of the Tibeto-Burman Area. 1988; 11(2): 98–119.
17. Matisoff JA. Variational semantics in Tibeto-Burman. The “organic” approach to linguistic comparison. Institute for the Study of Human Issues; 1978.
18. Blust R. Variation in retention rate among Austronesian languages. Unpublished paper presented at the Third International Conference on Austronesian Linguistics, Bali, January 1981.

19. Běijīng Dàxué. Hànyǔ fānyǔ cíhuà 汉语方言词汇 [Chinese dialect vocabularies]. Běijīng: Wénzhì Gāigé 文字改革, 1964.

20. Baayen RH, Piepenbrock R, Gulikers L. (eds.). The CELEX Lexical Database. Philadelphia: University of Pennsylvania; Linguistic Data Consortium; CD-ROM; 1995.

21. PONS.Eu Online-Wörterbuch. Stuttgart: Pons GmbH; [Accessed 2019 October 24].

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

23. Wickham H, others. Tidy data. Journal of Statistical Software. 2014; 59(10): 1–23. DOI: https://doi.org/10.18637/jss.v059.i10

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

25. Broman KW, Woo KH. Data organization in spreadsheets. The American Statistician. 2018; 72(1): 2–10. DOI: https://doi.org/10.1080/00031305.2017.1375989

26. Hammarström H, Haspelmath M, Forkel R. Glottolog. Version 4.0. Jena: Max Planck Institute for the Science of Human History; 2019. Available from: https://glottolog.org.

27. List JM, Rzymski C, Greenhill S, Schweikhard N, Pianykh K, Tjuka A, Tjuka A, Wu M-S, Forkel R. Concepticon: A resource for the linking of concept lists (Version 2.3.0) [Internet]. Jena: Max Planck Institute for the Science of Human History; 2020. Available from: https://concepticon.clld.org/.

28. List J-M, Anderson C, Tresoldi T, Rzymski C, Greenhill S, Forkel R. Cross-linguistic transcription systems (Version 1.3.0). Jena: Max Planck Institute for the Science of Human History; 2019. Available from: https://cts.clld.org/.

29. List J-M. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. Journal of Language Evolution. 2016; 1(2): 119–136. DOI: https://doi.org/10.1093/jole/jzw006

30. Matisoff JA. On the uselessness of glottochronology for the subgrouping of Tibeto-Burman. In: Renfrew C, McMahon A, Trask L. (eds.), Time depth in historical linguistics. 2000: 333–371. Cambridge: McDonald Institute for Archaeological Research.

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

32. List J-M, Lopez P, Bapteste E. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In: Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers) [Internet]. Berlin: Association of Computational Linguistics; 2016. 599–605. DOI: https://doi.org/10.18653/v1/P16-2097

33. List J-M. 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 [Internet]. Valencia: Association for Computational Linguistics; 2017. 9–12. Available from: https://digl.org/edictor/.

34. List J-M, Walworth M, Greenhill SJ, Tresoldi T, Forkel R. Sequence comparison in computational historical linguistics. Journal of Language Evolution. 2018; 3(2): 130–44. DOI: https://doi.org/10.1093/jole/jzy006

35. Wang WS-Y. Linguistic diversity and language relationships. In: Huang C-T J. (ed.) New horizons in Chinese linguistics. Dordrecht: Kluwer; 1996. 235–267. (Studies in natural language and linguistic theory). DOI: https://doi.org/10.1007/978-94-009-1608-1_8

36. Arnaud AS, Beck D, Kondrak G. Identifying cognate sets across dictionaries of related languages. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2509–2518. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/D17-1267

37. Wahle J. An approach to cross-concept cognacy identification. In: Bentz C, Jäger G, Yanovich I. (eds.) Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. Tübingen: Eberhard-Karls University; 2016. DOI: https://doi.org/10.105496/publikation-10060

38. Wang F. Miàoyǔ gǔyùn gōuní 苗语古音构拟 [Reconstruction of the sound system of Proto-Miao]. Tokay: Institute for the Study of languages; Cultures of Asia; Africa; 1994.

39. Bodt TA, List J-M. Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa languages. Papers in Historical Phonology. 2019; 4(1): 22–44. DOI: https://doi.org/10.2218/piphp.4.2019.3037

40. Höningwald HM. Phonetic similarity in internal reconstruction. Language. 1960; 36(2): 191–192. DOI: https://doi.org/10.2307/410982

41. Kay M. The logic of cognate recognition in historical linguistics. Santa Monica: The RAND Corporation; 1964.

42. Ratliff M. Against a regular epenthesis rule for Hmong-Mien. Papers in Historical Phonology. 2018 Dec; 3. DOI: https://doi.org/10.2218/piphp.3.2018.2877

43. Ostapirat W. Issues in the reconstruction and affiliation of Proto-Miao-Yao. Language and Linguistics. 2016; 17(1): 133–145. DOI: https://doi.org/10.1177/1606822X15614522

44. List J-M. Beyond edit distances: Comparing linguistic reconstruction systems. Theoretical Linguistics. 2019; 45(3–4): 1–10. DOI: https://doi.org/10.1515/tl-2019-0016
