DE GRUYTER OPEN

Yearbook of the Poznań Linguistic Meeting 3 (2017), pp. 47–76
DOI: 10.1515/yplm-2017-0003

Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages

Nathan W. Hill[a] and Johann-Mattis List[b]
[a]*SOAS, London*
[b]*Max-Planck-Institute for the Science of Human History, Jena*
[b]*mattis.list@shh.mpg.de*

Abstract

The use of computational methods in comparative linguistics is growing in popularity. The increasing deployment of such methods draws into focus those areas in which they remain inadequate as well as those areas where classical approaches to language comparison are untransparent and inconsistent. In this paper we illustrate specific challenges which both computational and classical approaches encounter when studying South-East Asian languages. With the help of data from the Burmish language family we point to the challenges resulting from missing annotation standards and insufficient methods for analysis and we illustrate how to tackle these problems within a computer-assisted framework in which computational approaches are used to pre-analyse the data while linguists attend to the detailed analyses.

Keywords: historical linguistics, linguistic reconstruction, Burmish languages, annotation, analysis, computer-assisted language comparison

1. Introduction

The quantitative turn in historical linguistics created a gap between "new and innovative" quantitative methods and classical approaches. Classical linguists are often skeptical of the new approaches, partly because the results do not seem to coincide with those of classical methods (Holm 2007), partly because they only confirm well established findings (Campbell 2013: 485f). Computational linguists, on the other hand, complain about inconsistencies in the application of the classical methods (McMahon and McMahon 2005: 26–29).

Both classical and computational approaches have strong and weak points. Steeped in philological learning, classical linguists enjoy extensive knowledge of, and refined intuitions about both common and language-specific processes of language change. Basing their analyses on multiple types of evidence, classical linguists can work out probable solutions even in situations where data are sparse. Their disadvantage is that they have difficulties coping with large amounts of data. The advantage of computational methods is their efficiency and consistency, and thus their ability to handle large amounts of data. The weakness of computational linguists is their tendency to ignore language-specific idiosyncrasies, being accustomed to deal only with homogeneous evidence. For this reason, computational approaches function poorly with sparse data. Since most of the data in historical linguistics are sparse and heterogenous (Sturtevant 1920: 11; Makaev 1977: 88), it is no wonder that the triumphs of computational analyses still lag behind those of classical approaches.

In the following, we concentrate on two specific challenges which both computational and classical historical linguists encounter when working with South-East Asian and specifically Sino-Tibetan (Trans-Himalayan) languages.[1] In particular, we focus on the Burmish languages, a small Sino-Tibetan sub-branch, but the analogous challenges are encountered in South-East Asian languages of other language families. We concentrate on processes of lexical change, pointing to specific challenges of *annotation* (Section 2) and *analysis* (Section 3). We then turn to addressing these problems in the Burmish Etymological Database (BED, https://dighl.github.io/burmish), where we use improved annotation and analysis techniques in order to create an etymological dictionary of the Burmish languages which is amenable to both qualitative and quantitative analyses.

2. Challenges of annotation

In historical linguistics we look back at a tradition of over 200 years of research on language families from around the world. Given this long tradition,

[1] By the term "Sino-Tibetan" we mean that language family of which Chinese, Tibetan, and Burmese are members. We use this term agnostically with regard to the shape of the Stammbaum of this family. Specifically, we see no reason to posit a branch of this family that contains Tibetan and Burmese but not Chinese.

it is surprising that our field still lacks common *annotation guidelines*: a general set of best practices stating how particular findings should be presented. By this, we do not mean the use of certain characters, like the asterisk to indicate that a word is reconstructed and not attested in written or spoken sources (see Koerner 1976 on the history of this practice), but rather a standardized way of how the fundamental findings, such as regular sound correspondences, convincing cognate sets, or shared innovations, are not only presented to the readers in publications, but also handled as data points amenable to statistical analyses. Historical linguistics has always been a data-driven discipline, even in pre-computer times, scholars would develop their individual practice of arranging their data with the help of index cards (see, for example, the detailed description in Gabelentz 1891, as well as his questionnaire for foreign language documentation from 1892, which is discussed in detail in Kürschner 2014) or punch cards (Swadesh 1963). Unfortunately, scholars rarely shared or discussed their practice but instead expected neophytes to learn by doing (Schwink 1994: 29).

The lack of annotation guidelines has immediate consequences both for classical and computational approaches. Computational approaches suffer from ambiguously annotated data which may confuse the algorithms, bound as they are by strict assumptions about the major processes of lexical change. Classical approaches suffer from a lack of transparency in data annotation when it comes to assessing the work of colleagues, especially vis-à-vis proposed regular sound correspondences and cognate sets. Since arguments on cognates and sound correspondences are often presented in an idiosyncratic way that varies not only from subfield to subfield but also among scholars working on the same language family, it is extremely difficult to base discussions on data and conclusions alone. This may be one of the reasons why debates often become personal in historical linguistics: since it is often not entirely clear where two scholars exactly differ, debates drift into polemics with scholars accusing each other of deliberately disregarding major facts.

In the following we quickly point to two major problems of annotation when analysing South-East Asian languages: cognates and sound correspondences. While the former constitutes primarily a problem for computational approaches to phylogenetic reconstruction, the latter is a major drawback for the discussion and evaluation of proposals in classical historical linguistics.

2.1. Partial cognate annotation

Cognacy is not a binary relation and cannot be reduced to a simple yes-no question. Instead, judging whether two words are cognate is both a question of perspective and degree. For example, one can distinguish "root" cognates from "stem" cognates. An example of *root cognates* is French *donner* 'to give' compared to Italian *dare* 'to give'. Both words descend from Proto-Indo-European *deh_3- 'to give', the French indirectly, via a verbalized *no*-stem (PIE *deh_3-no*- 'that which is given' > Latin *dōnāre* 'to give as present'), the Italian directly (PIE *deh_3 > Latin *dare* 'to give', Meiser 1998). An example of *stem cognates* is the comparison of Italian *dare* and Spanish *dar* 'to give', which both descend directly from Latin *dare*. The relativity of perspective and degree inherent in the notion of cognancy is comparable to the relation of *homology* in evolutionary biology, which denotes a relation of *commen descent* (Koonin 2005: 311). While we can say, for example, that wings in birds and wings in bats are deeply homologous, in so far as both represent the upper limbs of tetrapods, we can also say that they are homoplastic (i.e., independent innovations), in so far as their specific function, allowing tetrapods to fly, has evolved independently (Butler 2000, Morrison 2015).

Even more problematic than the vagaries of root etymology versus stem etymology are cases of *partial cognacy* (List 2015: 42; List 2016). Partial cognacy reflects a situation where words share cognate material only in part, such as French *aujourd'**hui***, which can be seen as partially cognate with Latin *hodiē*, itself a compound of Latin *hic* 'this' and *dies* 'day' (Vaan 2008: 287), of which the latter is again cognate with Ancient Greek Ζεύς [dzeus] (Meier-Bruegger 2002: L303). While partial cognacy generally holds for all root cognates reflected in words with different stems, including the case of French *donner* and Italian *dare*, mentioned above, partial cognacy is most frequently met in languages in which compounding is a frequent and productive process of word formation, such as South-East Asian languages.

As an example from the Burmish languages, consider the translational equivalents for 'yesterday' in Bola, Lashi, Rangoon Burmese, and Xiandao, given in Figure 1. As we have indicated with the aid of font colors, four languages have at least one morpheme in common (Bola [nɛʔ^{31}], Lashi [nap^{31}], Rangoon [ne^{53}] and Xiandao [n̩^{31}] all meaning 'day' in isolation), but only Bola and Lashi share the same compound structure. If we were forced to make a binary cognate decision out of this example, as we must when prepar-

Language	**Form**	**Strict**	**Loose**	**Exact**
Bola	a^{31} ŋji^{35} nɛʔ^{31}	1	1	1 2 3
Lashi	a^{31} ŋjei^{55} nap^{31}	1	1	1 2 3
Rangoon	mɑ^{53} ne^{53} kɑ^{53}	2	1	0 3 0
Xiandao	n̥^{31} m̥an^{35}	3	1	3 4
Achang	man^{35}	4	1	4

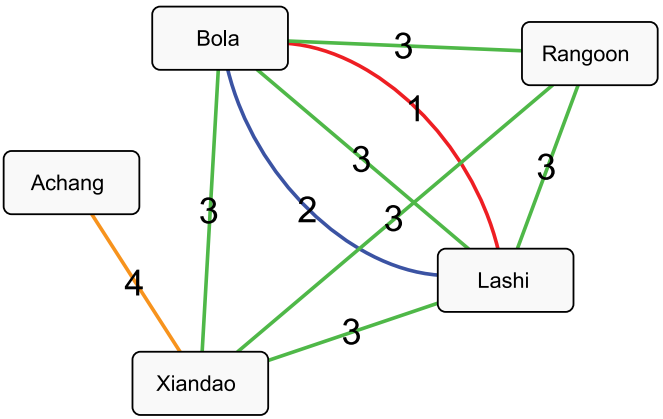

Figure 1. Annotation of cognate relations for words for ‘yesterday’ in five Burmish languages. Four languages share one morpheme, originally meaning ‘day’, marked in green in the table. But while Bola and Lashi show an identical compound structure, Rangoon and Xiandao show different structures, and the mono-morphemic word in Achang could have easily resulted from the loss of the first element of the cognate word in Xiandao. Coding these relationships in a strict fashion (column Strict) will ignore the similarity among all word forms in the morphemes they share, while coding in a loose fashion leads to an exaggeration of the similarities, rendering all words cognate. The same problems are further illustrated in the network on the right, where each edge represents one shared cognate morpheme across the five languages, based on the data in the table on the left. While all words form a connected component in this network, not all connections are equally strong.

ing cognate-coded datasets for the purpose of phylogenetic reconstruction analyses (Atkinson and Gray 2006), we would have a hard time deciding where to draw the boundaries in our cognate judgments. Are only Bola [a^{31} ŋji^{35} nɛʔ^{31}] and Lashi [a^{31} ŋjei^{55} nap^{31}] truely cognate, or should we say that all words are cognate, given that they form a connected component in a network, as illustrated in Figure 1? These decisions are reflected in what List (2016) calls *strict* and *loose partial cognate coding*. In strict cognate coding, only words which share the same compound structure and are cognate in all their parts are assigned to the same cognate set. In loose coding, one shared element is sufficient to assign two words to the same cognate set. For lexicostatistical datasets and phylogenetic reconstruction loose cognate coding necessarily masks important processes of *lexical replacement*: the fact that four of the five Burmish languages have a cognate morpheme in the word for 'yesterday' does not provide any important information for subgrouping. On the other hand, the case of Achang [man^{35}] and Xiandao [n̥^{31} m̥an^{35}] can be easily explained by assuming a recent loss of the first element in Achang, which is further confirmed by the overall closeness of the two languages. These examples illustrate that we should not blindly follow a strict cognate coding, as we may easily loose information relevant for subgrouping.

It seems that the best way to treat partial cognacy would be to follow an exact cognate coding of partial cognates, by annotating the cognacy of each morpheme in each word rather than for each word form. Unfortunately, available tools are not up to the task. Computational methods for automatic cognate detection, which could be used to pre-parse the data for the linguists, usually assume that words are morphologically simple (Steiner et al. 2011; List et al. 2017) and automatic partial cognate detection is still in its infancy (List et al. 2016).

Manual handling of partial cognacy is extremely tedious, since we lack consistent standards and tools for partial cognate annotation. As a result, studies which make use of manually annotated cognate sets usually ignore the problem of partial cognacy, as can be seen when inspecting the current practice of cognate coding in large lexicostatistic databases such as the Austronesian Basic Vocabulary Database (ABVD, Greenhill et al. 2008) or the Indo-European Lexical Cognacy Database (IELex, Dunn et al. 2012). In classical studies, scholars usually content themselves with the extraction of morphemes to establish sound correspondences or etymologies (Mann 1998),

and often even omit the information that the data from which their examples were drawn originally were morphologically complex words (Nishi 1999).

2.2. Sound correspondence annotation

Processes of sound change can be incredibly complex, especially when they involve suprasegmental developments, such as tone change or tono-genesis, which is often triggered by segmental features like the phonation of syllable-initial consonants, or the presence or absence of syllable-final plosives. For scholars who are unfamiliar with a particular language family, it is often impossible to say which sounds correspond when looking at a particular set of cognate words.

But even when ignoring complex sound correspondences, it may be extremely difficult for non-experts to see where two or more cognate sets display correspondences. As an example, consider two words for the comparison concept 'grease/fat', taking from the ABVD (Greenhill et al. 2008), namely Central Amis *simar* vs. Thao *lhimash*. The two words are labelled as cognates in the databases, but for non-experts, it is difficult to see which sounds correspond in the word forms. While it is straightforward to assume non-trivial sound correspondences between Central Amis *s-* and Thao *lh-*, as well as *-r* and *-sh*, it is still impossible for non-experts to assess whether this comparison makes sense or not, as we do not know how regular these correspondences are. Whether the sounds actually correspond or not, is not important for the sake of our example. What *is* important is the fact that we cannot transparently see what the people who annotated the words as being cognate were basing their opinion on.

3. Challenges of analysis

In the preceding section, we mentioned challenges of *annotation*, pointing to cases in South-East Asian languages where both computational and classical approaches have a hard time in achieving transparency. In the following, we show that similar problems arise in *analysing* the processes which pose a challenge for annotation. Having discussed the challenge of partial cognate annotation and sound correspondence annotation above, we here turn to the

problem of the reconstruction of compounds (Section 3.1) and the identification of irregular cognates (Section 3.2).

3.1. Reconstruction of compounds

Compounding is a frequent and vivid process in many languages and language families, not only in South-East Asia, but the world over. Given the prevalence of compounding in some Sino-Tibetan branches like Burmish or Sinitic, it is implausible to assume that the ancestors of the relevant languages had only monomorphemic words. Suprisingly, however, scholars have rarely tried to reconstruct concrete compounds in ancestral languages. Reconstruction systems of Proto-Burmish, for example, only give collections of morphemes with tentative semantic reconstructions (Burling 1967; Nishi 1999), and even where scholars provide reconstructions for tentative compounds in the proto-language (Mann 1998), they fail to provide a transparent account of how they arrived at these conclusions, that is, how they *analysed* the data.

That reconstructions and etymological dictionaries neglect the lexeme level is a general South-East Asian problem, found in etymological analyses of Hmong-Mien (Ratliff 2010), for Austro-Asiatic (Jenny and Sidwell 2015), and Tai-Kadai (Norquest 2007). Furthermore, the problem of treating compound structures consistently in etymological analysis is not unique to South-East Asian linguistics. In 1954, Malkiel criticized the lack of typological investigations on derivation and composition in historical linguistics. What he said by then, namely, that "[one] finds fleeting allusions and casual hints at certain varieties of derivational and compositional hierarchy, but surely no attempt at organized typology" (Malkiel 1954: 266) still holds today.

It is obvious that reconstruction at the lexeme level is more challenging than reconstruction at the morpheme level. True lexical reconstruction may at times even be impossible due to the incompleteness of available data and the complexity of compounding processes. However, scholars often do not even attempt to address these questions and there is little awareness of the inadequacies of the current "morphemes-first" approaches in South East Asian historical linguistics. If we want to advance our knowledge of language change, we cannot stop with sound change but need to try to find regularities and tendencies throughout all levels of language, including processes of word formation.

3.2. Identifying irregular cognate sets

If language contact can be excluded, sound change is a predominantly regular process that affects the whole lexicon of a language (Blevins 2004: 260–268; Kiparsky 1988; Labov 1981). Morphological processes, like suffixation, compounding, or analogy, however, are predominantly *sporadic*. Such morphological processes can mask the regularity of sound change and obstruct the identification of regular sound correspondences.

While the regularity of correspondence is still the major criterion to identify cognate words in different languages, it is by no means the only criterion employed by scholars applying the comparative method. As an example, consider German *fünf* 'five' vs. French *cinq* 'five'. While both words go back to the same Proto-Indo-European root **pénk^{w}e* 'five' (see Meier-Brügger 2002: 265), their phonetic development is highly irregular. While **pénk^{w}e* became *quinque* [k^{w}ink^{w}e] in Latin as a result of an assimilation process replacing the original **p* with **k^{w}* (Meiser 1998), a similar process happened in Proto-Germanic, where the word is reconstructed as **fimfe* (<**pimpe*), reflecting a sporadic change that replaced the **k^{w}* with **p,* which then became *f in Proto-Germanic (Kroonen 2013: 140). Without forms like Classical Greek *πέντε* [pénte] 'five' (with t <**k^{w}*) and Sanskrit *páñca* 'id.' (*c* < **k^{w}*), it is unlikely that we could identify the French and the German forms as true cognates going back to the same Indo-European root. It is the *cumulative evidence* drawn from regular sound correspondences among Greek, Sanskrit, Latin, and Proto-Germanic that allows us to first identify the Germanic and the Latin forms as irregular and then resolve this irregularity relying on our general knowledge of language-specific and general processes of sound change.

To operationalize such language specific developments when working on concrete language data is difficult. Regularities, at least in shallow language families, can usually be reliably detected when following the general protocol of the comparative method. Even automatic methods for cognate detection are getting more and more reliable and yield convincing results for shallow language families like Germanic or Romance (List et al. 2017). With their help, linguists could preparse the data, and quickly identify the major sound correspondences after manual correction. Finding the irregularities, however, is a much more difficult task, since it not only requires the knowledge of the regularities, but also a general strategy of how to identify cognate material which behaves irregularly in terms of the sound corre-

spondences. Up to today, no heuristics has been proposed for this task, neither in classical nor in computational historical linguistics.

4. Improving annotation and analysis in the Burmish etymological database

Our concerns with annotation and analysis in historical linguistics result from our own efforts in creating an etymological database of the Burmish language family. In this Burmish Etymological Database project (BED, http://dighl.github.io/burmish/), we aim to establish a new type of etymological database which provides data in both human- and machine-readable form, serving both for manual inspection and computational analysis. In the following, we briefly show how we address the aforementioned problems. Since the major part of our endeavour is still a work-in-progress, we are unable to present full-fledged solutions for all the problems mentioned, but we hope that our initial ideas serve future discussions in the field and may inspire new approaches.

4.1. Materials

4.1.1. The Burmish language family

The Burmish languages comprise a small and neatly identifiable group of languages spoken in Southwest China and Northeast Burma. The major languages of the Burmish Family include Burmese, Achang, Xiandao, Maru, Atsi (Zaiwa), Bola, and Lashi, as indicated in the map in the top panel of Figure 2. As can be seen, four of the varieties were recorded in the same city (Máng City 芒市 in China, formerly called Lùxī 路西). When comparing the languages their close proximity must be borne in mind, as we should expect intensive language contact among them. Characteristics of the languages in this family include a generally isolating morphological structure, the use of lexical tone, and tense or creaky phonation.

Nishi (1999: 68) distinguishes two subbranches, Maruic and Burmic, the latter comprising Burmese, Achang, and Xiandao. His classification rests on the observation that the Burmic languages lost tense phonation, replacing it with aspiration of the initial. However, this development does not allow the

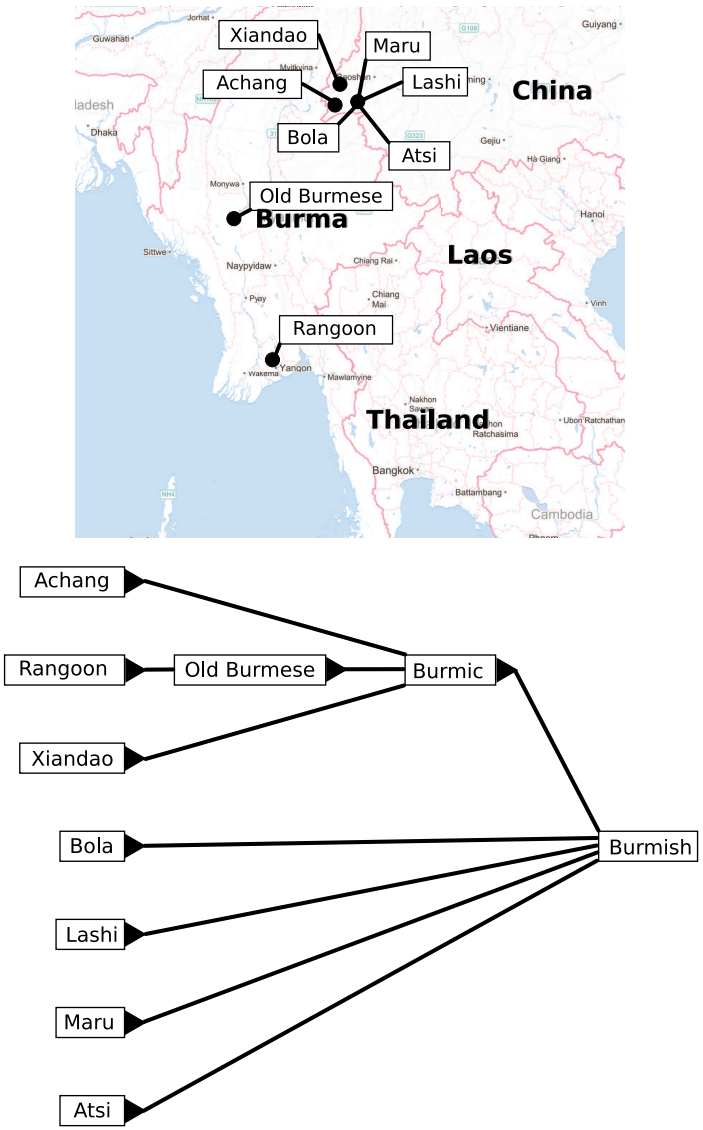

Figure 2. The top panel shows the geographic location of the Burmish varieties in our database (Rangoon is the prestige dialect of modern Burmese), with the location of Old Burmese at Pagan, the capital of the first Burmese dynasty. The bottom panel shows a tentative phylogeny based on sound changes identified as shared innovations, using multi-furcations to indicate uncertainty.

identification of Maruic as a sub-branch, since by keeping tense phonation the languages in question share a retention rather than an innovation. Thus, we propose the preliminary genetic classification seen in the bottom panel of Figure 2, with uncertainties indicated using polytomic (multifurcating) splits. Note that this classification deviates from the one provided in Glottolog (Hammarström et al. 2017), which follows the classification of Mann (1998), one that is not sufficiently substantiated with linguistic evidence.

4.1.2. The Burmish Etymological Database

The Burmish Etymological Database (BED) currently provides data for a basic word list of 240 items translated into the 8 varieties (including Rangoon as the modern prestige dialect of Burmese) given in Figure 2. The data were taken from Huáng et al. (1992) in the digital version provided by the Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT) project (Matisoff 2015), to which we added Old Burmese on the basis of Okell (1971), Luce (1985) and Nishi (1999). The etymologies we arrived at independently of the STEDT project, and the degree of annotation was, as will be further illustrated below, considerably refined.

4.1.3. Availability of data, tools, and code

All data which we used for the following illustrations along with the source code of the software we applied are available in the supplementary material accompanying this paper. In addition to our analyses, we provide explicit links for the languages in the data to Glottolog (Version 3.0, Hammarström et al. 2017), and the concepts in the data to the CLLD Concepticon (Version 1.0, List et al. 2016). All words are further linked to the STEDT database, apart from those for Old Burmese which was not taken from STEDT.

4.2. Methods and tools for annotation and analysis

In order to address the problems mentioned above, several methods and tools were developed, which are presented in more detail below. Computationally

intensive methods for automatic analyses were generally written as plugins for LingPy, a Python software library for quantitative tasks in historical linguistics (Version 2.5.1, http://lingpy.org, List and Forkel 2016), and are available in the supplementary material accompanying this paper. Tools for manual annotation and inspection were implemented as part of the Etymological Dictionary Editor (EDICTOR, http://edictor.digling.org, List 2017), a web-based interactive tool for creating, inspecting, and editing etymological datasets, and are already implemented in the most recent online version of the tool.

4.2.1. Partial cognate annotation

As mentioned above, the manual annotation of partial cognates is tedious. In order to ease the task, a partial cognate editor was included in the most recent version of the EDICTOR tool, which greatly facilitiates the annotation task. All that is required is that the data are morphologically segmented by the user. Once this is done, users can load their data into the EDICTOR tool and indicate which morphemes in a set of pre-defined words (usually translations of the same comparison concept) are cognate. Since this can be done in a simple drag-and-drop fashion, by which the user selects and deselects the words which are grouped into one partial cognate set, the annotation can be carried out quickly and is also less prone to error than the use of spreadsheet software not designed for this task.

In order to identify partial cognates in the BED projects, we first analysed the data automatically, using the algorithm recently proposed by List et al. (2016) for the automatic detection of partial cognates, and then manually corrected the errors in the automatic analysis.

4.2.2. Using alignments for sound-correspondence annotation

To detect regularly recurring sound correspondences linguists usually rely on *alignment analyses* (Prokić et al. 2009; List 2014). Alignments are a formal way to compare sequences. In an alignment analsyis, two or more strings of segments are arranged in a matrix in such a way that corresponding segments are placed in the same column, while placeholders (so-called *gaps*, usually represented by the symbol "-") mark segments lacking a counterpart. In addi-

Figure 3. Partial cognate annotation with the EDICTOR tool. Annotation of partial cognates is essentially drag and drop. The user first selects morphemes by clicking on them in order to assign them to a common cognate set in a second step. The figure shows how we cluster translations of the comparison concept 'the feather' in the BED.

tion to identifying partial cognates in the Burmish language data, we also aligned the data, using a computer-assisted work-flow in which we first aligned the partial cognate sets automatically using the SCA algorithm (List 2012) available in the LingPy software package, and then refined them manually, using the alignment module of the EDICTOR tool. An example alignment analysis is illustrated in Figure 4 for translations of the comparison concept 'the man (male human)'.

DOCULECTS	CONCEPTS	ID: 446					ID: 448			
Achang_Longchuan	the man (male human)	-	i	-	-	31	tɕ	i	-	55
Atsi	the man (male human)	j	u	-	ʔ	21	k	e	-	51
Bola	the man (male human)	j	a	u	ʔ	31	k	a	i	55
Lashi	the man (male human)	j	-	u	ʔ	55	k	ɛ	-	31
Maru	the man (male human)	j	a	u	k	31	k	a	i	31
Xiandao	the man (male human)	j	-	u	ʔ	31	ɕ	ɛ	-	55
Rangoon	the man (male human)	j	ɑ	u	ʔ	4	tɕ	ɑ	-	55

Figure 4. Example for the tentative alignment of words for the comparison concept 'the man (male human)' in seven of the eight Burmish languages in our sample.

The use of alignments to annotate sound correspondences is an old technique that goes at least back to the early 20th century (Dixon and Kroeber 1919), long before automatic alignment algorithms were proposed (Covington 1996, Kondrak 2000). Unfortunately, alignments have only sporadically been employed so far (Haas 1969; Fox 1995: 67; Payne 1991). Scholars often consider alignments as too simple to represent the complex relations they see when looking at cognate words. This, however, is not a convincing ground for the rejection of alignments. If alignments are indeed too simple to reflect sound correspondences in all their complexity, scholars should work on enhanced ways to transparently annotate their judgments.

4.2.3. Compound analysis and word family detection

List (2016) presents an initial approach to reconstructing processes of word compounding with the help of a reference phylogeny and ancestral state re-

construction based on weighted parsimony. Given that our data are available in a similar form, we could use the same technique to analyse compound processes in the Burmish languages. However, since this approach requires a good idea of the general phylogeny of the languages, whereas the phylogeny of the Burmish languages remains rather unclear, we base our initial compound analysis on a semi-automated approach that helps to identify the *motivation structure* underlying the formation of specific compounds. Our core idea is to follow Urban (2011) in searching for *partial colexifications* across the words in our data, and to represent them as *bipartite networks*. Following François (2008), we see colexification as a term to cover cases in which a word form is used to denote more than one concept, without distinguishing between homophony or polysemy. Partial colexification therefore points to cases where a specific morpheme is shared across two words denoting distinct concepts.

Given that each syllable usually corresponds to one morpheme in the Burmish languages, it is easy to write a computer application to search for these patterns in our data. In contrast to approaches that are solely interested in the relations between different concepts (List et al. 2013), we wish to investigate both the actual word forms in our data and the concepts which they denote. *Bipartite* networks, which are increasingly used to investigate molecular datasets in evolutionary biology (Corel et al. 2016), provide an intuitive and simple structure for such a computer-assisted investigation. Bipartite networks are networks consisting of two types of nodes. Edges in these networks are only allowed to be drawn from nodes of one type to nodes of another type. In our case the first node type are the *concepts* in the concept list and the second node type are the *word forms* in a given language. We create our network by linking all individual morphemes in our data to the concepts denoted by the words in which they occur. This yields a large graph, which is almost completely connected, but sparse enough to allow interactive search for interesting structures using graph-visualization software, such as Cytoscape (Smoot et al. 2011), and without applying heavy algorithmic machinery. In our supplementary material, we provide the full network created from our data along with the source code as an interactive web-application that works in most web browsers.

In addition, and in order to complement this computational analysis, the EDICTOR tool contains a *morpheme annotation module* that allows one to inspect automatically created bipartite networks for individual languages and to annotate compounds in a meaningful way. The general idea behind this

compound structure analysis is to annotate compounds in a way similar to how linguists annotate sentences in inter-linear glossed text. For each word in the data, we provide a language-internal analysis that reveals the motivation of compound formation. Essentially, this yields a language-internal *word family analysis*, as it allow us to identify cognates within the same language.

ID	COGID	CONCEPT	MORPHEMES	TOKENS
3368	400	the river	water + mo-suffix	v u ɪ 51 + m o 55
3535	425	the sea	water + sea	v u ɪ 51 + m ɪ ŋ 21
4868	409	the water	water	v u ɪ 51
598	619	to buy	buy	v u ɪ 51

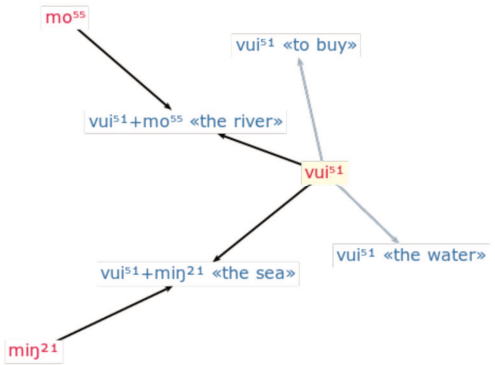

Figure 5. Compound analysis (language-internally) with the help of partial colexification networks. The example shows four words in Atsi (Zaiwa), of which three constitute a word family. The table shows the morpheme analysis and the raw data, while the network below shows the bipartite graph which is automatically created by the EDICTOR tool.

As an example, consider Atsi [vui^{51} mo^{55}] 'river', [vui^{51} miŋ^{21}] 'sea' and [vui^{51}] 'water'. When inspecting these words, it is obvious, that [vui^{51}] 'water' recurs in the words for 'sea' and 'river', and it is also easy to identify [mo^{55}] as a suffix, as it recurs in a few other words , such as [lo^{21} mo^{55}] 'tiger'

and [vam^{51} k^{h}ui^{21} mo^{55}] 'wolf'.[2] The language-internal bipartite networks drawn from partial colexifications available in the EDICTOR drastically facilitate this task. Scholars can first automatically search for potential word families and then annotate them step by step, eventually distinguishing coincidental cases of homophony, such as Atsi [vui^{51}] 'to buy', from the reuse of an etymon in distinct lexemes. Figure 5 shows the user-annotated data and the automatically reconstructed partial colexification network for this example.

5. Results

In the following, we present the results of the analyses described above. We should add that most of these results are anecdotal and not quantitative. There are two reasons for this: first, our general intention in the BED project is to pursue a computer-assisted rather than a computer-based approach to language comparison. This means that we use quantitative analyses to do the bulk of the heavy lifting while we inspect the data manually to find those patterns which cannot be explained with algorithms alone. Second, our methodology comprises preliminary work that to our knowledge has so far not yet been tested on other language families. By pointing to some of our initial findings, we hope we can advertise the tools and approaches discussed here. In this way, we hope that the preliminary approaches presented in this study may in the future bear further fruits, be it in our own work or that of our colleagues working on language families that present similar difficulties.

5.1. Comparison with STEDT

As we assigned the cognate sets independently of the cognates provided by the STEDT project (Matisoff 2015), one can compare the differences between our analysis of the Burmish languages and the analysis provided by the STEDT project. The 240 concepts and 7 languages which were originally taken from STEDT's digitalized version of Huáng et al. (1992) consists of 1611 distinct words and 1002 distinct morphemes. 743 (46%) of the words

[2] By 'wolf' we understand 'dhole' (*Cuon alpinus*). The grey wolf (*Canis lupus*) is not endemic to the relevant parts of Asia.

are annotated in STEDT, i.e., they are given etymologies; 828 (83%) of the morphemes are assigned to cognate sets in STEDT. Having excluded 23 out of the 743 words for which we found no link between our data and the data in STEDT, we compared the similarity in cognate judgments for the remaining 720 words, using B-Cubed Scores (Bagga and Baldwin 1998) to estimate the differences. These scores are usually measured in *precision*, *recall*, and *harmonic mean* (F-Score), by comparing the results of a cluster analysis A with a cluster analysis B. Precision indicates how often clusters proposed by analysis B are also found in analysis A, recall indicates how often clusters proposed in analysis A are also found in analysis B, and the harmonic mean provides a summary of the two scores. All scores are measured in terms of floating points between 0 and 1, with 1 indicating complete identity and 0 indicating complete difference.

The comparison of our BED analysis with the analysis provided by STEDT (assuming that BED is analysis A and STEDT is analysis B) yielded a precision of 0.88, a recall of 1.0, and an F-Score of 0.94. These results are remarkable, given that the analyses were carried out independently. The high recall means that whenever BED says that two words are cognate, STEDT will also do so. The low precision shows that our analysis is more conservative, having the tendency to refuse cognate judgments rather than to propose them, and as a result, if BED refuses cognacy, STEDT may in quite a few cases still tend to propose it.

5.2. Proving cognacy despite irregularities

Thanks to the alignment analyses carried out on our data, we are able to determine quickly whether the sound correspondence patterns inherent in a given cognate set are regular or not. For convenience, the EDICTOR offers a module in which sound correspondences are automatically counted for each pair of languages in the data. Ideally, this should likewise be offered for the major patterns across all of the languages in the data, but at the moment, this is not feasible, as no algorithms for the detection of general correspondence patterns have been proposed so far.

In order to identify potential cognates independently of regular sound correspondences, we can employ our bipartite partial colexification networks. As an example for this idea, compare the words for ‘good’ across seven Burmish varieties given in Table 3. At first sight, the words all look quite

similar, and no linguist would immediately rule out the possibility that they could be cognate. Based on the sound correspondences we identified, however, the forms in Achang and Xiandao are not regular, as the correspondence among [tɕ] in Achang, [ɕ] in Xiandao and [k] in the other Burmish varieties is only attested in the words for 'good' and the word for 'man', also given in Table 3.

Despite the irregularity of the sound correspondences between Achang and the other varieties, it is still justifiable to regard all words as cognate (except for Rangoon Burmese [kãu^{55}] 'good' and Lashi [kɛː^{31}] which has an unpredicted long vowel). We reconstruct the word 'man' in Proto-Burmish as a compound of 'person' and 'good', supported by the fact that the first morpheme of the words for 'man' occurs in the words for 'who' in Bola and Maru (as shown in the same table), and that – except for in Rangoon Burmese – the second morpheme in the words for 'man' is cognate in all languages in the table (we suspect that the vowel length in Lashi is a secondary phenomenon, probably resulting from loss of syllable weight in compounds).

Table 3. Irregular sound correspondences among Achang and Xiandao and five other Burmish languages: Achang [tɕ] and Xiandao [ɕ] in the word for 'good' exhibits an irregular correspondence with [k] in the other Burmish languages. The fact that the compound word 'man' has the word for 'good' as its second part in all Burmish languages apart from Rangoon, and the peculiarity of the motivation of this compound justify assuming cognacy despite irregularity. As a result, we label cognacy among the morphemes in the table by assigning the same color to cognate morphemes, leaving black as the color for words we cannot relate to any other word.

Language	'man'	'good'	'who?'
Achang	i^{31} tɕi^{55}	tɕi^{55}	xau^{55}
Atsi	juʔ^{21} ke^{51}	ke^{51}	o^{55}
Bola	jauʔ^{31} kai^{55}	kai^{55}	kʰak^{55} jauʔ^{31}
Lashi	juʔ^{55} kɛ^{31}	kɛː^{31}	xaŋ^{55}
Maru	jauk^{31} kai^{31}	kai^{31}	kʰə̆^{31} jauk^{31}
Rangoon (Burmese)	jɑuʔ^{4} tɕɑ^{55}	kãu^{55}	bɛ^{22} tθu^{22}
Xiandao	juʔ^{31} ɕɛ^{55}	ɕɛ^{55}	xau^{55}

Since this compound is semantically and syntactically peculiar and uniquely occurs in the Burmish languages (we found no similar motivation in the more than 40 other Sino-Tibetan languages in Huáng et al. 1992), it is very likely that this word originated only once in the history of the Burmish languages. No matter what the explanation for the irregular sound correspondences in Achang and Xiandao will be (if it can ever be found), given the overwhelming similarity in the *motivation structure* of the compound for 'man' in the Burmish languages, one cannot resist the conclusion that these words are indeed cognate, and we mark them accordingly in Table 3.

5.3. Compound structure and subgrouping

Compound structure can provide us with initial hints regarding subgrouping. We must be careful, however, since it is obvious that words can easily be borrowed among languages, and closely related languages will also allow for the borrowing of full compounds, as we can see in numerous examples from the Chinese dialects (compare, for example, List et al. 2014). Nevertheless, when such cases can be excluded, compound structure may serve as a proxy for the identification of shared traits between languages and thus help us to identify potential innovations that provide us evidence for subgrouping.

As an example, consider Table 4 which gives words for 'mountain', 'dog', 'thunder', 'wolf', and 'bear (n.)' in the modern languages in our sample along with our comparative analysis of the motivation structure of these words, derived from the bipartite partial colexification networks. First, we find four different motivations for 'wolf' in the sample. Except for the Rangoon word form, all are derived from the word for 'dog', but the first part of the compound differs, and we find 'bear' + 'dog' in Atsi and Lashi, 'thunder' + 'dog' in Bola and Maru, and 'mountain' + 'dog' in Achang and Xiandao. Achang and Xiandao further show the same motivation structure for 'thunder', which can be seen as a further argument that both varieties form a subbranch of the Burmic branch of Burmish.

The situation with Lashi, Bola, and Maru is more complicated and requires further explanation. We find that Maru shares the same motivation structure for 'thunder' with Lashi ('sky' + 'thunderB'), while it also shares the motivation structure for 'wolf' with Bola ('thunder' + 'dog'). Note that our analysis of Maru [mjaŋ^{31} k^{h}a^{35}] as 'thunder' + 'dog' is based only on the similarity with Bola, as the word for 'thunder' in Maru does not contain [mjaŋ^{31}].

Table 4. Compound motivation patterns across the modern Burmish languages. Items with identical color in the annotation of the motivation structure are presumed to be cognate across and inside the four varieties. Black is reserved for items which are not related to any other item in the data.

Language	'mountain'	'dog'	'thunder'	'wolf'	'bear (n.)'
Atsi	pum^{51}	k^{h}ui^{21}	mau^{21} mjiŋ^{51}	vam^{51} k^{h}ui^{21} mo^{55}	vam^{51}
	mountain	dog	sky + thunder	bear + dog + *m-suff.*	bear
Bola	pam^{55}	k^{h}ui^{35}	mau^{31} mjaŋ^{55}	mjaŋ^{55} k^{h}ui^{35}	vɛ̃^{55}
	mountain	dog	sky + thunder	thunder + dog	bear
Lashi	pɔm^{31}	k^{h}ui^{55}	mou^{33} kɔm^{33}	wɔm^{31} k^{h}ui^{55}	wɔm^{31}
	mountain	dog	sky + thunderB	bear + dog	bear
Maru	pam^{31}	lă^{31} k^{h}a^{35}	muk^{55} kum^{31}	mjaŋ^{31} k^{h}a^{35}	vɛ̃^{31}
	mountain	? + dog	sky + thunderB	thunder + dog	bear
Achang	pum^{55}	xui^{31}	mau^{31} ʐau^{31}	pum^{55} xui^{31}	ɔm^{55}
	mountain	dog	sky + thunderC	mountain + dog	bear
Xiandao	pum^{55}	fui^{31}	mau^{31} cau^{31}	pum^{55} fui^{31}	om^{55}
	mountain	dog	sky + thunderC	mountain + dog	bear
Rangoon	tãu^{22}	k^{h}we^{55}	mo^{55} tɕ^{h}ẽ^{55}	wũ^{22}pu^{53}lwe^{22}	wũ^{22}
	mountain2	dog	sky + thunderD	bear + ? + ?	bear

Given that the data for Maru, Lashi, Bola, and Atsi were collected in the same area, and close contact among the varieties is therefore expected, we may suspect that the divergence in compound structures results from language contact. Given that 'bear' occurs in the word for 'wolf' in Atsi, Lashi, Achang, Xiandao and particularly in the otherwise untransparent Rangoon Burmese form , we suspect that the 'thunder-dog' in Maru and Bola is a later innovation rather than a retention. This suspicion however gives rise to a further complication. If Maru and Bola together innovated the structure 'thunder-dog' then the Maru word for 'thunder' should be cognate with the form of the word 'thunder' that occurs in the Maru word for 'wolf', which it is not. To explain the Maru word for 'thunder' one can suggest that Maru has borrowed it from Lashi. This proposal is not only confirmed by the irregular vowel correspondence between the two varieties, but also by alternative data in Clerk (1911: 163), who gives *muk myang* as the word for 'thunder' in a Maru variety spoken in the Myitkina area of Burma, far away from Máng

City, where the Maru variety we considered for our database is spoken. The Myitkina form appears to preserve the inherited etymon as opposed to the Máng City form, which is borrowed from Lashi. This explanation is yet further buttressed by the fact that Wannemacher (2011: 37) gives /mou⁴ gøm⁴/ as translation for 'thunder' in a Lashi variety spoken in the Waimaw area of the Kachin State in Burma, again far away from the Lashi variety we considered in our study. The obvious cognancy of the Lashi forms from distinct regions of Burma points to the fact that Lashi here retains an inheritance. In other words, the Lashi word is geographically stable whereas the Maru word is not.

It would go beyond the scope of this paper to resolve the phylogeny of the Burmish languages by listing potential shared innovations or even using phylogenetic methods to arrive at a subgrouping of the language family. We think, however, that our small analysis of the words in Table 4 has shown that compound motivation structures bears substantial potential for linguistic subgrouping, provided they are analysed with care, and borrowing are thoroughly identified. Both the analysis of compound motivation structures and the identification of borrowings cannot be done automatically. Our methods for the reconstruction of bipartite partial colexification networks, however, provide great help for a detailed computer-assisted analysis.

5.4. Compound structure and semantic reconstruction

A compound motivaton structure analysis derived from bipartite partial colexification networks can also serve as a starting point for semantic reconstruction, both from a semasiological perspective, seeking the original meaning of a given morpheme, and form an onomasiological perspective, seeking to identify how a given concept was pronounced in ancestral languages. As an illustration, consider the colexification newtork given in Figure 6. In this example, we find three major semantic complexes: the verbs 'to shoot (arrow)' and 'to throw', the verb 'to hunt', and several concepts denoting body parts ('hair', 'tail', 'bone', etc.). These semantic groups are connected by two form groups, the first one pointing to Proto-Burmish **pak⁴* and the second pointing to **ʃa²* (both in the reconstructions of Mann 1998). The verb 'to shoot' is expressed by single morphemes (reflexes of **pak⁴*) in Atsi, Bola, Maru, and Xiandao, while the verb for 'to hunt' is expressed by two morphemes, the former colexifying with the forms for 'to shoot', and the latter,

reflexes of Mann's **ʃa²*, occurring as one of the elements in the numerous body part terms in our third semantic cluster. Given these patterns, we find it straightforward to reconstruct the rough semantics of Proto-Burmish **pak⁴* as 'to throw/to shoot', and the semantics of **ʃa²* as 'body/flesh', since these meanings (which are admittedly not extremely precise at this stage of the analysis) allow best to explain why reflexes of **ʃa²* occur in compounds denoting body parts, and as the object of verb-object compounds meaning 'to hunt' (lit. 'shoot meat' or 'shoot bodies') in the Burmish varieties.

The pattern in Figure 6 is but a small example of a computer-assisted procedure, but it illustrates the main idea of computer-assisted approaches : the analytical work is still carried out by the linguists who interpret the data and draw their conclusions, but an advanced computational modeling of linguistic problems helps the linguists in identifying patterns deserving explanation. No doubt one could identify the pattern in Figure 6 by simply inspecting the data in a book. The representation as bipartite networks of partial colexifications, however, drastically speeds up this process.

6. Conclusion

With more than 7000 languages currently spoken and numerous other languages now lost, existing in philological records, historical linguistics faces the tremendous task of charting the evolution of these languages into their current shape. Computational approaches offer quick solutions to analyze large amounts of digitally available data. However, they face specific difficulties, resulting from their lack of flexibility which makes them vulnerable in situations of sparse data. Classical approaches handle data sparseness well, but they face efficiency and transparency problems. A combined framework can cope with the shortcomings of both disciplines while at the same time preserving their specific advantages.

In this paper, we have tried to illustrate how computational and classical approaches can be combined, concentrating on specific challenges of annotation and analysis in the Burmish language family. With the help of computational methods and interactive tools for the correction of errors, we consistently annotated partial cognates and regular sound correspondences for eight Burmish varieties. With the help of bipartite partial colexification networks, we further annotated compound motivation structures for a large part of the words in our data. We illustrated the benefit of these new approaches to an-

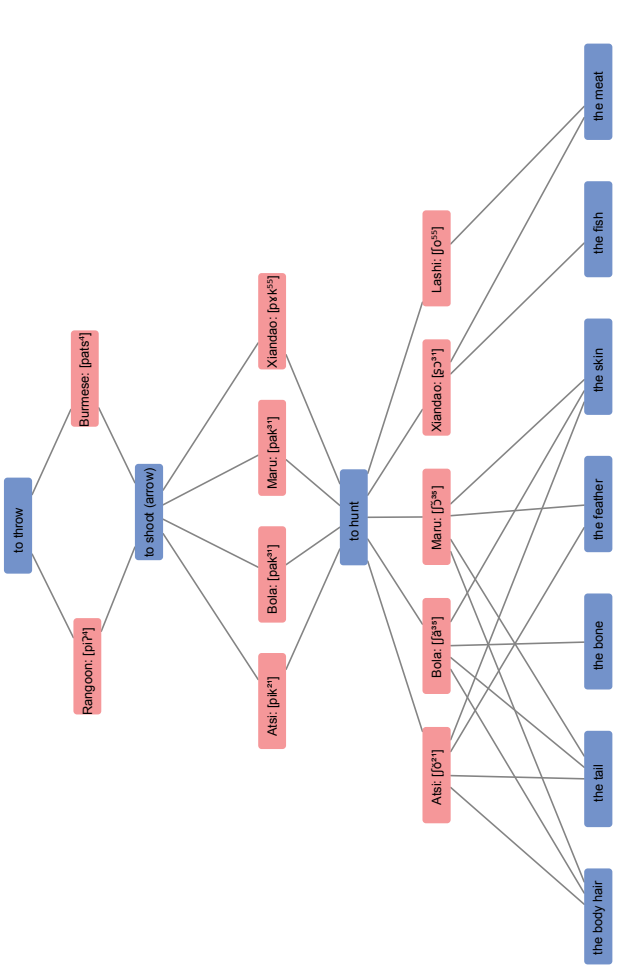

Figure 6. A bipartite word family network indicating the relations between words for 'to shoot' and 'to hunt' in the Burmish language. The bipartite graph constructed from partial colexifications shows which concepts (blue nodes) are expressed with similar morphemes (red nodes) in the Burmish languages in our sample. As can be seen from the network, the word for 'to hunt' is expressed by morphemes in most languages: one which also means 'to shoot' in isolation, and one which occurs as a further element in words like 'body hair', 'bone', 'skin', and 'meat'. This leads us to conclude that the Proto-Burmish word for 'to hunt' was a compound motivated as {shoot} + {body/flesh}. Our evidence is two-fold, drawing from information regarding the regularity of the sound correspondences in the languages under investigation as well as the structural information exhibited in the bipartite word family network.

notation and analysis, by showing how cognate words can be identified even when sound correspondences are irregular, how shared innovations can be detected by searching for similar compound structures, and how compound structure comparison allows us to make initial steps towards semantic reconstruction. The proposed methods and techniques are preliminary and need to be further developed. We are, however, confident that they provide new insights not only into the Burmish languages but also into South-East Asian languages in general, since they offer not only a more complete perspective on linguistic reconstruction, but also deliver additional evidence for subgrouping, hidden cognates, and semantic reconstruction.

7. Acknowledgements

This research would not have been possible without the LFK Young Scholars Symposium (University of Washington, Seattle, 2013), generously hosted by the Li Fang-Kuei Society for Chinese Linguistics (http://lfksociety.org/), during which both authors made first acquaintance and began their collaboration. We would like to acknowledge the generous support of the European Research Council for supporting this research under the auspices of 'Beyond Boundaries: Religion, Region, Language and the State' (ERC Synergy Project 609823 ASIA, NWH) and 'Computer-Assisted Language Comparison' (ERC Starting Grant, JML), and the German Research Foundation (DFG) for supporting JML from 2015 to 2016 with a research scholarship on 'Vertical and Lateral Aspects of Chinese Dialect History' (Grant No. 261553824). We would further like to thank Guillaume Jacques and Harald Hammarström for providing helpful comments on an earlier version of this paper, as well as Doug Cooper and Mark Miyake for providing invaluable help with linguistic data.

8. Supplementary material

The supplementary material accompanying this paper contains the source code with which we created our sample bipartite network application, as well as the Burmish data set which we analyzed and prepared with help of the Etymological Dictionary Editor (http://edictor.digling.org). All material can be downloaded from https://zenodo.org/record/886179. The code was curated

on GitHub at http://github.com/digling/challenges-of-annotation-paper. The data is additionally shared in CLDF (http://cldf.clld.org), following the most recent specifications.

References

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

Bagga, A. and B. Baldwin. 1998. "Entity-based cross-document coreferencing using the vector space model". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Association of Computational Linguistics. 79–85.

Blevins, J. 2004. *Evolutionary phonology*. The emergence of sound patterns. Cambridge: Cambridge University Press.

Burling, R. 1967. *Proto-Lolo-Burmese*. Bloomington: Indiana University Press.

Butler, A. and W. Saidel. 2000. "Defining sameness: Historical, biological, and generative homology". *BioEssays* 22. 846–853.

Campbell, L. 2013. *Historical linguistics*. Edinburgh: Edinburgh University Press.

Clerk, F. 1911. *A manual of the Lawngwaw or Maru language, containing: the grammatical principles of the language, glossaries of special terms, colloquial exercises, and Maru–English and English–Maru vocabularies*. Rangoon: American Baptist mission Press.

Corel, E., P. Lopez, R. Méheust and E. Bapteste. 2016. "Network-thinking: Graphs to analyze microbial complexity and evolution". *Trends in Microbiology* 24(3). 224–237.

Covington, M. 1996. "An algorithm to align words for historical comparison". *Computational Linguistics* 22(4). 481–496.

Dixon, R. and A. Kroeber. 1919. *Linguistic families of California*. Berkeley: University of California Press.

Dunn, M. (ed.). 2012. Indo-European lexical cognacy database (IELex). http://ielex.mpi.nl/.

Fox, A. 1995. *Linguistic reconstruction. An introduction to theory and method.* Oxford; Oxford University Press.

François, A. 2008. "Semantic maps and the typology of colexification: Intertwining polysemous networks across languages". In: Vanhove, M. (ed.), *From polysemy to semantic change*.Amsterdam: Benjamins. 163–215.

Gabelentz, G. v. d. 1891. *Die Sprachwissenschaft. Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: T. O. Weigel.

Gabelentz, G. v. d. 1892. *Handbuch zur Aufnahme fremder Sprachen* [Handbook for the description of foreign languages]. Berlin: Ernst Siegfried Mittler & Sohn.

Greenhill, S., R. Blust and R. Gray. 2008. “The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics”. *Evolutionary Bioinformatics* 4. 271–283.

Haas, M. 1969. *The prehistory of languages*. Mouton: The Hague and Paris.

Hammarström, H., R. Forkel and M. Haspelmath. 2017. *Glottolog*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Holm, H. 2007. “The new arboretum of Indo-European ‘trees’. Can new algorithms reveal the phylogeny and even prehistory of Indo-European?” *Journal of Quantitative Linguistics* 14(2–3). 167–214.

Huáng Bùfán 黃布凡 .1992. *Zàngmiǎn yǔzú yǔyán cíhuì* [A Tibeto-Burman lexicon]. Zhōngyāng Mínzú Dàxué 中央民族大学 [Central Institute of Minorities]: Běijīng 北京.

Jenny, M. and P. Sidwell (eds.). 2015. *The handbook of Austroasiatic languages*. Leiden and Boston: Brill.

Kiparsky, P. 1988. “Phonological change”. In: Newmeyer, F. (ed.), *The Cambridge Survey of Linguistics* (vol. 1). Cambridge: Cambridge University Press. 363–415.

Koerner, E. 1976. “Zu Ursprung und Geschichte der Besternung in der historischen Sprachwissenschaft. Eine historiographische Notiz”. *Zeitschrift für vergleichende Sprachforschung* 89(2). 185–190.

Kondrak, G. 2000. “A new algorithm for the alignment of phonetic sequences”. In: *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. 288–295.

Koonin, E. 2005. “Orthologs, paralogs, and evolutionary genomics”. *Annual Review of Genetics* 39. 309–338.

Kroonen, G. 2013. *Etymological dictionary of Proto-Germanic*. Leiden and Boston: Brill.

Kürschner, W. 2014. “Georg von der Gabelentz’ *Handbuch zur Aufnahme fremder Sprachen* (1892). Entstehung, Ziele, Arbeitsweise, Wirkung“. In: Ezawa, K., F. Hundsnurscher and A. Vogel (eds.), *Beiträge zur Gabelentz-Forschung*. Tübingen: Narr. 239–259.

Labov, W. 1981. “Resolving the Neogrammarian Controversy”. *Language* 57(2). 267–308.

List, J.-M. 2012. “LexStat. Automatic detection of cognates in multilingual wordlists”. In: *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*. 117–125.

List, J.-M., A. Terhalle and M. Urban. 2013. “Using network approaches to enhance the analysis of cross-linguistic polysemies”. In: *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*. Association for Computational Linguistics. 347–353.

List, J.-M., S. Nelson-Sathi, W. Martin and H. Geisler. 2014. “Using phylogenetic networks to model Chinese dialect history”. *Language Dynamics and Change* 4(2). 222–252.

List, J.-M. 2014. *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

List, J.-M. 2015. "Network perspectives on Chinese dialect history". *Bulletin of Chinese Linguistics* 8. 42–67.
List, J.-M., M. Cysouw and R. Forkel. 2016. "Concepticon. A resource for the linking of concept lists". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. 2393–2400.
List, J.-M. and R. Forkel. 2016. *LingPy. A Python library for historical linguistics*. Jena: Max Planck Institute for the Science of Human History.
List, J.-M. 2016. "Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction". *Journal of Language Evolution* 1(2). 119–136.
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*.) Association of Computational Linguistics. 599–605.
List, J.-M., S. Greenhill and R. Gray. 2017. "The potential of automatic word comparison for historical linguistics". *PLOS ONE* 12(1). 1–18.
List, J.-M. 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*. 9-12.
Luce, G.H. 1985. *Phases of Pre-Pagán Burma: Languages and history*. Oxford: Oxford University Press.
Makaev, E. 1977. *Obščaja teorija sravnitel'nogo jazykoznanija* [General theory of comparative linguistics]. Moscow: Nauka.
Malkiel, Y. 1954. "Etymology and the structure of word families". *Word* 10(2–3). 265–274.
Mann, N. 1998. A phonological reconstruction of Proto Northern Burmic. (MA thesis, the University of Texas at Arlington.)
Matisoff, J. 2015. *The Sino-Tibetan Etymological Dictionary and Thesaurus project*. Berkeley: University of California.
McMahon, A. and R. McMahon. 2005. *Language classification by numbers*. Oxford: Oxford University Press.
Meier-Brügger, M. 2002. *Indogermanische Sprachwissenschaft*. Berlin: de Gruyter.
Meiser, G. 1998. *Historische Laut- und Formenlehre der lateinischen Sprache*. Darmstadt: Wissenschaftliche Buchgesellschaft.
Morrison, D. 2015. "Molecular homology and multiple-sequence alignment: an analysis of concepts and practice". *Australian Systematic Botany* 28. 46–62.
Nishi, Y. 1999. *Four papers on Burmese: Toward the history of Burmese (the Myanmar language)*. Tokyo: Institute for the study of languages and cultures of Asia and Africa, Tokyo University of Foreign Studies.
Norquest, P. 2007. A phonological reconstruction of Proto-Hlai. (PhD dissertation, The University of Arizona.)
Okell, J. 1971. "K Clusters in Proto-Burmese". Paper presented at the Sino-Tibetan Conference, October 8–9, 1971. Bloomington, IN.

Payne, D. 1991. “A classification of Maipuran (Arawakan) languages based on shared lexical retentions”. In: Derbyshire, D. and G. Pullum (eds.), *Handbook of Amazonian languages* (vol. 3). Berlin: Mouton de Gruyter. 355–499.

Prokić, J., M. Wieling and J. Nerbonne. 2009. “Multiple sequence alignments in linguistics”. In: *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. 18–25.

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

Schwink, F. 1994. *Linguistic typology, universality and the realism of reconstruction*. Washington: Institute for the Study of Man.

Smoot, M., K. Ono, J. Ruscheinski, P. Wang and T. Ideker. 2011. “Cytoscape 2.8. New features for data integration and network visualization”. *Bioinformatics* 27(3). 431–432.

Steiner, L., P. Stadler and M. Cysouw. 2011. “A pipeline for computational historical linguistics”. *Language Dynamics and Change* 1(1). 89–127.

Sturtevant, E. 1920. *The pronunciation of Greek and Latin*. Chicago: University of Chicago Press.

Swadesh, M. 1963. “A punchcard system of cognate hunting”. *International Journal of American Linguistics* 29(3). 283–288.

Urban, M. 2011. “Asymmetries in overt marking and directionality in semantic change”. *Journal of Historical Linguistics* 1(1). 3–47.

Vaan, M. 2008. *Etymological dictionary of Latin and the other Italic languages*. Leiden: Brill.

Wannemacher, M. 2011. *A phonological overview of the Lacid language*. Chiang Mai: Linguistics Institute, Payap University.

Address for correspondence:
Johann-Mattis List
Kahlaische Str. 10
07754 Jena
Germany
mattis.list@shh.mpg.de