Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa languages

TIMOTHEUS A. BODT
SOAS London

JOHANN-MATTIS LIST
MPI-SHH, Jena

Abstract

Although it is well-known to most historical linguists that the comparative method could in principle be used to predict hitherto unobserved words in genetically related languages, the task of word prediction is rarely discussed in the linguistic literature. Here, we introduce "reflex retrodiction" as a new task for historical linguistics and report an ongoing experiment in which we use a computer-assisted workflow to retrodict reflexes for so far unobserved words in eight varieties of Western Kho-Bwa (a subgroup of Sino-Tibetan). Since, at the time of writing this report, the experiment is still ongoing, we do not report concrete results, but instead provide an estimate of our expectations by testing the performance of the computational part of our workflow on existing language data. Our results suggest that reflex retrodiction has the potential of becoming a useful tool for historically oriented fieldwork.

1 Introduction

It is well known that the comparative method cannot only be used to reconstruct languages no longer reflected in writing systems, but that it can also be used to *predict* structures or words in that have not yet been investigated or observed. Thus, when based on comparative and internal evidence, Saussure (1879) proposed the existence of *coefficients sonantiques* in the system of the Indo-European proto-language he predicted that – if ever a language was found that retained these elements – these new sounds would surface as segmental elements in certain cognate sets of the so far undetected language. These sounds are nowadays known as *laryngeals* (*h_1, *h_2, *h_3, see Meier-Brügger 2002), and when Hittite was identified as an Indo-European language (Hrozný 1915), one of the two sounds prognosticated by Saussure could indeed be identified

in several word forms, thus providing evidence for Saussure's original 'prediction'.

Saussure's prediction was not planned as such, and it is unlikely that Saussure even thought of his theory in this way. That prediction in this sense, which is more appropriately called *retrodiction*, is possible in our discipline, however, is well-known, even if it less frequently discussed as such in the literature. When dealing with linguistic retrodiction, linguists try to infer the structure of so far unobserved datapoints based on the data available to them at a given point in time. Classical examples for linguistic retrodiction are the universals of grammar proposed by Greenberg (1963). As these universals are usually stated in the form of implications, we can – provided the universal holds – infer the presence of one structural feature if we know the feature that implies it (see also Blevins 2004 on predictions in the field of historical phonology). Another example is the common practice in historical linguistics to retrodict missing reflexes of cognate sets when searching for etymologies in a given language (see, for example, Michael et al. 2015: 196). Even among speakers living in contact areas, we can at times observe how they learn to guess how a word unknown to them would sound in the target language (Branner 2006: 215).

In order to test the predictive strength and the usefulness of prediction studies in historical linguistics, we are currently carrying out an experiment on missing words in Western Kho-Bwa language data. Western Kho-Bwa is a branch of the Sino-Tibetan (or Tibeto-Burman, or Trans-Himalayan) language family, that has not been thoroughly investigated so far. The main idea of this ongoing experiment is to use a computer-assisted workflow by which missing reflexes in an etymological dataset of eight Western Kho-Bwa language varieties are predicted, using computational techniques which are later refined manually. These missing reflexes can then be directly tested in fieldwork, by comparing predictions against attested reflexes.

In the following sections, we will introduce *reflex retrodiction* as a new explicit task of historical linguistics and point to existing automatic solutions (Section 2). We will then present our experiment in detail, providing information on its background, on the language varieties involved, and how we plan to evaluate the results (Section 3). Given that – at the time of writing this paper – the experiment is still ongoing – we will then provide a succinct outlook on our expectations, by testing the performance of the algorithm on existing Western Kho-Bwa language data (Section 4).

2 Reflex retrodiction as a new task for historical linguistics

Although prediction is rarely mentioned and mostly implicitly practised in historical linguistics, we consider it a vital aspect of the comparative method, and we think that a more explicit discussion of prediction techniques could play a vital role for the future of our discipline. While our linguistic knowledge derived from the techniques for historical language comparison could be used for a wide range of predictions targeting different linguistic domains, we think that the task of *reflex retrodiction* deserves more attention in particular. Reflex retrodiction is hereby understood as the task by which a linguist tries to predict the form of the *reflex* of a given proto-form or a cognate set attested in different languages.

Linguists apply reflex retrodiction routinely when searching for thus far unattested cognates in a specific language. Cognate sets are often spotty, showing reflexes only in a small sample of all languages under investigation, especially when initial research only considers cognate sets that share the same meaning across all languages. Hence, the actual search for the missing words in other regions of the lexicon can turn out to be very tedious and time consuming. In order to ease this search, scholars intuitively predict missing forms, based on known sound laws or known patterns of sound correspondences. When asking informants or sifting through dictionaries, they search for forms that match their guess, and if they find these (even in partly deviated form), they can directly add them to their list of attested reflexes for a given cognate set.

In the following sections, we will quickly discuss how retroflex retrodiction is carried out traditionally, and which automatic methods have been proposed so far. We will conclude by discussing the potential of more formalized approaches to reflex retrodiction when dealing with unstudied or understudied language families and linguistic sub-groups.

2.1 Classical approaches for reflex retrodiction

In principle, there are two basic ways how reflex retrodiction can be carried out: *top-down* or *pattern-based*. Top down approaches start from a given proto-form and an ordered list of sound laws, which researchers apply step by step, until the form in the language missing the reflex has been derived. Applying this technique successfully requires both a very good knowledge of the sound change processes (both the individual sound changes as well as the diachronic order in which they occurred) of all the languages under investigation and a reliable proto-form. Given the complexity of sound-law-based derivations, which require a very de-

tailed knowledge of the sound change processes that lead to the diversification of a given language family, top-down approaches are only applicable when dealing with very well-attested and deeply investigated language families, such as Indo-European.

Pattern-based approaches make use of the knowledge of observed *correspondence patterns* to fill gaps where reflexes are missing in cognate sets. There are two basic ways in which this can be done. Firstly, one can use pairwise sound correspondences to try to predict a word form unknown in one language from a word form known in another language. An example for this would be to use the German *Dorf* 'village' to predict the English counterpart *thorp*, which nowadays is only attested in village names. The disadvantage of pairwise sound correspondences is that we often face complex correspondences by which one sound in one language may have two or more counterparts in the other language. Although phonetic conditions can at times give us hints as to the choice of the correct sound, this is not necessarily given in all cases, specifically also because conditions for mergers or splits in sound change can easily be lost during language change.

To circumvent the problem of missing information when trying to predict words from one language into another, one can, secondly, predict reflexes from sound correspondence patterns across multiple languages. In the linguistic literature, we often find examples of recurring correspondences across more than one language, which are usually used to illustrate how certain proto-sounds are reflected in the languages under investigation (Clackson 2007: 37). Correspondence patterns across multiple languages have, of course, a greater predictive strength, given that evidence lost in the majority of languages may still be present somewhere. An example is the vocalism of Indo-European, which can barely be reconstructed without resorting to Ancient Greek (Meier-Brügger 2002). The disadvantage of correspondence patterns, however, is, that they are difficult to formalise. Furthermore, it is unlikely that linguists can remember the complexity that correspondence patterns across multiple languages can show in reality with enough detail.

Given that the distinction between pairwise and multiple language comparison is essentially arbitrary, it is obvious that linguists pursuing reflex retrodiction in practise will resort to an intuitive weighting of evidence. If linguists know that for a given unattested reflex a specific language provides the clue information, such as the vowel in Indo-European, they will naturally try to start with the language that provides the crucial information. In cases where the situation is less clear, they will successively increase the number of witnesses in order to come up with the

form that, in their opinion, best matches the evidence. It is also obvious that intuitive correspondence-based retrodiction is very hard to formalise for computational applications, given that the weightings will usually be language-specific and that humans are very flexible in taking different kinds of evidence into account. It would probably even be incorrect to say that a given form predicted by a linguist has been solely arrived at by correspondence-based retrodiction alone, given that linguists who study a language family closely usually also have at least a rough idea about the major sound changes that took place in the past in order to produce the patterns we observe at present.

2.2 Automatic methods for reflex retrodiction

While the task of reflex retrodiction is not strictly divided into different strategies and the distinction between pairwise sound correspondences and correspondence patterns across multiple languages is usually being not made in practice, automatic approaches that have been proposed so far tend to follow one of these two major strategies. Since, as we have emphasised in the previous section, reflex retrodiction by means of sound laws is usually only applicable to well-studied language families whose history is already well-understood by experts, we won't discuss automatic approaches for this task in detail here. For readers interested in this topic, we recommend the very detailed survey on the broader task of 'computerised forwards reconstruction' by Sims-Williams (2018).

Although not necessarily labelled as such, automatic approaches for reflex retrodiction based on pairwise sound correspondences have been used for quite some time. An example is the early work by Chen (1997) on mutual dialect intelligibility, in which the author proposes an automatic measure to assess how well speakers can understand words from different dialects, assuming that these words are, in fact, cognate. If we turn this idea around, and ask, how well speakers could predict the pronunciation of a word, taking the potential knowledge of pairwise sound correspondences into account, we would have a first idea to develop an automatic method for reflex retrodiction based on language pairs.

In times of growing popularity of machine learning, in particular neural network approaches, as a powerful tool for multiple different purposes, it is not surprising that scholars have already tested the power of these tools for the purpose of reflex retrodiction. Thus, Dekker (2018) uses the data provided by the NorthEuralex project of Dellert & Jäger (2017) along with methods for automatic cognate detection as provided by the LingPy software package of List, Greenhill, et al. (2018) to test the power

of different neural network approaches and settings to handle the task of pairwise word prediction across different languages. The results are generally promising, showing at times rather low differences between predicted and attested words. A drawback is that the method does not use phonetic transcriptions as input, but instead converts the data to the reduced sound class system proposed by the ASJP project (Wichmann et al. 2016), the so-called ASJP code (Brown et al. 2008), which consists of only 40 symbols instead of the much richer inventory offered by the International Phonetic Alphabet (*IPA Handbook* 1999). Thus, while very interesting as a pilot study, the approach is less feasible for people interested in practical applications, although we hope that the author will find time in the future to increase the flexibility of the work flow, allowing scholars to use the method for their own work.

In contrast to the pairwise approach proposed by Dekker, List (2019) uses sound correspondence patterns across multiple languages for the task of reflex retrodiction. The basic goal of the algorithm proposed by List is not to predict missing reflexes across different languages, but rather to identify sound correspondence patterns in multilingual datasets containing many 'gappy' or 'patchy' cognate sets. If cognate sets show reflexes in only a few languages, it is often not clear which of the observed sound correspondences, derived from phonetically aligned cognate sets, should belong to the same *correspondence pattern*.

As an example, consider data from four Western Kho-Bwa varieties in Table 1. In this example, we can easily spot two clear-cut correspondence patterns for the initial consonant, showing exactly the same set of reflexes in 'push' and 'human being'[1] and another set of reflexes in 'know' and 'poison'. Based on these data only, this division is straightforward. For the concepts 'burn, roast' and 'scratch' there are missing values for one variety each. Based on the data provided here, we would predict the initial of Jerigaon 'burn, roast' as [r] and that of Jerigaon 'scratch' as [d]. In the case of 'fireplace, hearth',[2] however, we would have a hard time guessing the correct initial, let alone the entire segment, based on the available data, since the correspondences of the initials could – due to the two gaps in Khoina and Jerigaon – be assigned to both the pattern of 'push' and 'human being', and to the pattern of 'know' and 'poison'. Whenever we assign this cognate set as a whole to one of the patterns, we make an implicit,

[1] In all varieties, the word expressing this concept consists of a prefix denoting human beings and the root.

[2] In all varieties, the word expressing this concept consists of the morpheme for 'fire' (or a 'fire'-prefix) and the root.

Concept	**Khispi**	**Duhumbi**	**Khoina**	**Jerigaon**
'push'	d u	d u	r yː	r yː
'human being'	bu + d u n	bu + d u n	dʑ ə + r i ŋ	dʑ ə + r i ŋ
'burn, roast'	d i	d i	r i ː	Ø
'hearth, fireplace'	b u + d u	b ɛj + d u p	Ø	Ø
'scratch'	d ɔ k	d ɔ k	d ɔ k	Ø
'know'	d ɛ n	d ɛ n	d ɛ n	d ɛ n
'poison'	d u k	d u k	d y k	d y k

Table 1: From sound correspondence to correspondence patterns: Four words are missing (indicated by Ø) in our data. While, on basis of the data presented here, we can probably safely assume that the correspondence pattern for 'burn, roast' is the same as that for 'push' and the correspondence pattern for 'scratch' is the same as that for 'know', the case for 'fireplace, hearth' is ambiguous.

testable prediction for the corresponding initial of the missing forms. In this case, this would either be [r] or [d].

The algorithm proposed by List (2019) essentially infers correspondence patterns from aligned cognate sets, by treating columns (also called *sites*) in all aligned cognate sets in the data as the *nodes* of a network, with *edges* displaying those sites which are *compatible* with each other. By modelling the data as a network, the cluster problem can then be treated as the well-known *minimum clique cover problem* in graph theory (Bhasker & Samad 1991), for which approximate solutions exist (Welsh & Powell 1967). Given that the algorithm assigns all alignment sites in a given dataset to unique correspondence patterns, all 'gappy' sites which are assigned to a given correspondence pattern contain inherently a prediction regarding the form of missing reflexes. Thus, when assigning the initial site of the alignment from the cognate set for 'fireplace, hearth' to the pattern of 'push' in Table 1, we would predict that the reflexes for the two missing words should start with [r]. If we assigned it to the pattern 'know', we would predict it to be [d].

The predictive strength of the algorithm for sound correspondence pattern inference was tested as part of the initial study and revealed a rather high accuracy of automated reflex retrodictions, with accuracy scores ranging between 50% and 80% for varying data sets. The reasons for the difference in the accuracy scores is still not fully understood, but it seems clear that they are not only related to the genetic diversity of the languages under question, but also to phonotactic aspects, such as the size of the phoneme inventories of the target languages. Despite these

aspects remaining unclear for the moment, we consider the results as interesting enough to justify a further testing of the method.

3 A prediction experiment on Western Kho-Bwa languages

In order to test the predictive strength of the comparative method, we designed a prediction experiment for hitherto unobserved words in the Western Kho-Bwa languages. The main idea of this ongoing experiment is to use the information provided by regular sound correspondences across a set of eight Western Kho-Bwa varieties to retrodict pronunciations for reflexes (words and morphemes) that have so far not yet been elicited in field work. Once predicted and registered with the Open Science Framework (https://osf.io), follow-up field work will allow us to verify the retrodictions that were made before, thereby testing not only the current knowledge of sound correspondences, but also the general predictive power of the comparative method.

3.1 Background on the experiment

The starting point of this experiment was an initial etymological data set, assembled by Bodt during fieldwork conducted in Arunachal Pradesh between 2012 and 2017. These data were initially only available in a non-standardised form, namely a manually prepared Word table. The data set was first converted to a spreadsheet with standardised notations. During one week of intensive work, we then normalised the data to a level where it could not only be automatically processed with the help of different software tools provided by the LingPy Python package (List, Greenhill, et al. 2018), but also sufficiently post-edited and corrected with help of the web-based EDICTOR tool (List 2017). The initial goal was to prepare the data to such an extent that we could annotate the data and pursue the work flow of computer-assisted language comparison which Hill and List developed as part of a long-term project aimed at the reconstruction of Proto-Burmish (Hill & List 2017).

While this work was on-going, List finished his article on the automatic inference of sound correspondence patterns across multiple languages, mentioned above (List 2019). In order to evaluate the performance of the method, he designed an experiment in which the data was split into parts of different sizes, and the information present in the inferred correspondence patterns was used to retrodict the most likely pronunciation of word forms that were artificially deleted from the data. As mentioned before, this experiment turned out to be surprisingly success-

Variety	Items	Ratio
Duhumbi	632	0.95
Jerigaon	460	0.69
Khispi	526	0.79
Khoina	504	0.76
Khoitam	521	0.79
Rahung	519	0.78
Rupa	534	0.81
Shergaon	437	0.66
TOTAL	662	0.78

Table 2: Status of data coverage in Bodt's Western Kho-Bwa data after initial fieldwork between 2012 and 2017.

ful, reaching accuracy levels between 50% (Burmish languages) and 80% (Polynesian languages).

The Western Kho-Bwa etymological data set was based on initial fieldwork aimed at obtaining a first understanding of the possible genetic relatedness of these languages. Since not all concepts were elicited for all the varieties, there was a considerable number of missing words in the data, ranging between 5% (Duhumbi) to 34% (Shergaon) of the words, with an average of 22% (see Table 3.1). Given that Bodt's data was missing potentially crucial witnesses for a historical reconstruction of the subgroup, and List just finished his draft on the algorithm that could be used as an automatic method for reflex retrodiction, it was clear that the Western Kho-Bwa language data would be a very good test case to check both how well the algorithm performed on the reflex retrodiction task and how well expert predictions on word forms would perform in general. Additionally, such a test would also allow us to get a clearer picture about the power of the comparative method and the regularity of sound change. Given that unattested word prediction – be it based on human assessment or automatic methods – heavily relies on the classical assumption that sound change is regular, the accuracy of word prediction also gives us direct insights into the regularity of sound change within a given language family.

3.2 Background on Western Kho-Bwa languages

In 1952, Stonor, basing himself on local sources, reported that the two languages 'Sulung' or 'Puroik' and 'Bugun' are mutually intelligible (Stonor

1952). It was not until the last two decades of the 20th century that the first linguistic materials on Bugun/Khowa, Puroik/Sulung, Sherdukpen and Sartang/Boot/Butpa Monpa became available: the works of the Indian research and language officers Tayeng (Tayeng 1990) and Dondrup (Dondrup 1988, 1990, 2004). On the Chinese side of the border, the first Puroik data were published as part of the large-scale survey Tibeto-Burman Phonology and Lexicon (Sūn 1991). Based on these materials and his own field work data, Jackson Sun (Sun 1992, 1993) was the first to suggest that Puroik, Bugun, Sherdukpen and 'Lishpa-Butpa' are not just a random residue when all other major languages are subtracted, but that they might belong together and form a coherent linguistic group.

Other researchers after him either adopted his view or independently reached the same conclusion (Burling 2003, Rutgers 1999). van Driem (2001) dubbed the group 'Kho-Bwa cluster' in his handbook *Languages of the Himalayas*, by combining his provisional reconstructions for 'water' and 'fire' in the subgroup. More recent publications include the Puroik description from China by Lǐ (2004), the Sherdukpen description by Jacquesson (2015) and the elicited wordlists of different varieties in the unpublished report by Abraham (2005). Blench & Post (2014) and Post & Burling (2017) expressed scepticism about Puroik being part of this proposed group of languages. Nonetheless, all commonly consulted handbooks (Burling 2003, Genetti 2016, Post & Burling 2017) and the online language encyclopedias Ethnologue (Lewis & Fennig 2013) and Glottolog (Hammarström et al. 2018) mention 'Kho-Bwa' as a (potential) branch of Tibeto-Burman in western Arunachal Pradesh. Although the exact phonological shape of the reconstructions *kho* WATER and *bwa* FIRE needs to be established, we follow Lieberherr & Bodt (2017) and others before them in using 'Kho-Bwa' as a label for these languages. Besides the fact that this terminology is already established to some extent, it has the advantage of not being biased toward one language like 'Bugunish' (Sun 1993), or a region like 'Kamengic' (Blench & Post 2014, Post & Burling 2017). Furthermore, 'Kho-Bwa' offers an exhaustive definition of the group: Any language of western Arunachal Pradesh in which the word for 'water' starts with *k* and the word for 'fire' starts with *b* is a 'Kho-Bwa' language.

The Western Kho-Bwa languages (Bodt 2014) are eight distinct linguistic varieties spoken in the western part of the Kho-Bwa area: the valleys of the Gongri and Tenga rivers. The languages belonging to this sub-group are Khispi (Lishpa), Duhumbi (Chugpa), Sartang and Sherdukpen. Sartang has four distinct speech varieties, whereas Sherdukpen has two. The number of speakers of these linguistic varieties combined is around 8,500, and considering the low speaker population and the rapid

socio-economic and cultural changes in this area, all varieties must be considered endangered.

3.3 Linguistic data on Kho-Bwa used in our study

Our linguistic data reflects eight distinct Western Kho-Bwa varieties. The data in its current form is presented in form of a spreadsheet file that can be directly imported by the LingPy software (http://lingpy.org, List, Greenhill, et al. 2018); by the LingRex package (List 2018), which provides the code for automatic reflex retrodiction as presented in List (2019); and by the EDICTOR interface (http://edictor.digling.org), a web-based tool that allows for a quick manual correction of automatic analyses (List 2017).

The basic structure of this data format is a header in the first row, which indicates the content of the cells in each column, and one word per language and per row. In addition to basic columns, such as a unique identifier (ID), the name of the language variety (DOCULECT), or an elicitation gloss for the concept (CONCEPT), the original data entry for the given word (VALUE) and a semi-automatically segmented form (TOKENS), the data contains very detailed manually corrected analyses on *cognate relations* (CROSSIDS), expressed in the form of *partial, cross-semantic cognates* (List 2016), morphological glosses (MORPHEMES), the prosodic structure of each entry (STRUCTURE), and a phonetic alignment analysis of the data (ALIGNMENT).[3]

ID	DOCULECT	CONCEPT	VALUE	TOKENS	MORPHEMES	CROSSIDS	ALIGNMENT	STRUCTURE
617	Khispi	burn, roast (vt)	di	d i	BURNVT	195	d i	i n
618	Duhumbi	burn, roast (vt)	di	d i	BURNVT	195	d i	i n
619	Khoina	burn, roast (vt)	ri:	r i:	BURNVT	195	r i:	i n
621	Khoitam	burn, roast (vt)	ri:	r i:	BURNVT	195	r i:	i n
623	Rupa	burn, roast (vt)	ri:	r i:	BURNVT	195	r i:	i n

Table 3: Short example of the data and format employed in our approach.

An example for the data format employed in our approach is provided in Table 3, where reflexes for the concept 'burn, roast' are given across five varieties (with three entries missing so far). The column with the prosodic structure (STRUCTURE) plays an important role in clustering the alignment sites into correspondence patterns, since – as a rule – the algorithm will only try to cluster those alignment columns into the same partition

[3] Compare List, Walworth, et al. (2018) for more information on the tabular format employed by LingPy and related tools.

that are identical with respect to their prosodic label. This also reflects the classical practice of distinguishing between sound correspondences of the initials and the rhymes in comparative analyses of Sino-Tibetan languages, and South-East-Asian languages in general.

Our format comes very close to the specifications required by the Cross-Linguistic Data Formats initiative (https://cldf.clld.org, Forkel et al. 2018), which seeks to increase the overall comparability of linguistic data by encouraging scholars to adhere to general standards by linking their data to reference catalogues, such as Glottolog for languages (Hammarström et al. 2018), Conception for concepts (List et al. 2016), and the transcription system advocated by the Cross-Linguistic Transcription Systems initiative (CLTS, https://clts.clld.org, Anderson et al. 2018). When registering the experiment, our data was still missing the links to the Concepticon and the CLTS transcription system, but in the meantime, we have prepared the data in CLDF format, and it can be found on GitHub (https://github.com/lexibank/bodtkhobwa) and Zenodo (Version 1.0.0, +++Link2BAdded+++).

3.4 Computer-assisted reflex retrodiction

Our computer-assisted workflow for word prediction consists of two parts, an automatic and a manual task. In the automatic task, we employ the automatic correspondence pattern recognition method by List (2019) in order to predict the missing words in the data. The results of this analysis is a table of morphemes as predicted by the algorithm. These come along in three variants, as shown in Table 4. The difference between these three variants is that they display different degrees of uncertainty. At times, an alignment site could be assigned to different correspondence patterns, as we have seen for the concept 'hearth, fireplace' in our example in Table 1. If we have to decide between two or more correspondence patterns, the algorithm orders these patterns in decreasing order of alignment sites supporting a given pattern. The variant shown as *Word1* in the table only picks the first value for a given language, while *Word2* picks the first two values (if more than one are found), and displays them in a single segment slot, separated by a pipe (|) symbol. If no correspondence pattern can be found for a given alignment site (which may happen if the sites do not occur regularly in the data) the algorithm displays this by using ∅, as our symbol for missing data. That a certain prediction suffers from uncertainty in one of the alignment sites is further displayed in the Column *Qu.* by a question mark.

Hence, in Table 4, we see predictions for the same three concepts with missing values in Table 1: 'burn, roast', 'scratch' and 'fireplace, hearth'. As we predicted earlier, the 'best fit' for the initial for the concept 'burn, roast' in Jerigaon is indeed an [r], and the 'best fit' for the concept 'scratch' in Jerigaon is indeed a [d]. The automatic analysis already comes up with predictions for the initial for the concept 'fireplace, hearth' in Khoina and Jerigaon because, unlike the data presented in Table 1, the concept actually has attested reflexes in Khoitam, Rahung, Rupa and Shergaon on basis of which the reflexes can be assigned to the correspondence set 'push', rather than to the correspondence set 'know'. That the concept 'fireplace, hearth' nonetheless has a question mark in the Column *Qu.* is because the algorithm cannot assign a value to the alignment sites of the rhymes (i.e. the nucleus in the case of the prefix and the coda in case of the root) of the predicted word.

No.	Qu.	Cogn.	Language	Concept	Morpheme	Word1	Word2	Word3
361		195	Jerigaon	burn, roast (vt)	BURNVT	r iː	r|tʰ iː	r|tʰ iː
362		195	Rahung	burn, roast (vt)	BURNVT	r iː	r iː	r iː
363		195	Shergaon	burn, roast (vt)	BURNVT	r iː	r|t iː|i	r|t iː|i
465		251	Jerigaon	scratch	SCRATCH	d ɔ k	d ɔ|a k|-	d ɔ|a|u k|-
466		251	Khoitam	scratch	SCRATCH	d ɔ k	d|dʑ ɔ|a k|-	d|dʑ ɔ|a|yː k|-
467		251	Rahung	scratch	SCRATCH	d ɔ k	d|dʑ ɔ|øː k|-	d|dʑ ɔ|øː k|-
1020	?	515	Jerigaon	hearth, fireplace	hearthfireplace	b Ø -	b Ø -	b Ø -
1021	?	515	Khoina	hearth, fireplace	hearthfireplace	b Ø -	b Ø -	b Ø -
1022	?	516	Jerigaon	hearth, fireplace	hearthfireplace	r ɛ Ø	r|tʰ ɛ Ø	r|tʰ ɛ Ø
1023	?	516	Khoina	hearth, fireplace	hearthfireplace	r ø Ø	r ø Ø	r ø Ø

Table 4: Example output format of our automatic reflex retrodiction experiment.

Given that our algorithm predicts all missing words mechanically regardless of whether this makes sense in terms of lexical considerations, it is clear that the selection of items to be explicitly elicited does not need to contain all items for which predictions could be made. The automatic transcriptions provided in the supplementary material[4] may thus contain predictions for which we already know they are unlikely to exist. This may be due to lexical innovations, borrowings, or because no concepts exist for a given elicitation gloss. For this reason, Bodt made a *manual analysis*, extracting those predictions that on basis of his experience would be most promising and interesting. This list comprises a list of 630 detailed predictions (including full words with prefix and main root), as well as an informed guess by Bodt that at times overrides the automatic prediction. Essentially, this allows us to compare two different kinds of predictions:

[4] File `<predictions-automatic.tsv>` in our registration.

the fully automated ones, and the ones corrected by the expert, which is also shared in the supplementary material.[5]

So, to continue with the example in Table 4 above, the manually adjusted prediction for the concept 'fireplace, hearth' in Khoina ([bɾ ɔ p]) and Jerigaon ([bɾ ɔ p]) is based on the evidence from the other Sartang and Sherdukpen varieties, including contraction of the fire-prefix to the root.

3.5 Status and time line of the experiment

As mentioned before, the initial fieldwork was carried out by Bodt between 2012 and 2017, with the major part of the data on the Western Kho-Bwa varieties collected in 2014. A first overview of the results, including a Word table with provisional proto-forms and the underlying sound correspondences, was presented at the South East Asian Linguistic Seminar in Padang, Indonesia, in May 2017. In June 2018, the spreadsheet was sufficiently enhanced by converting it into formats that can be computationally processed. On August 20, 2018, List carried out the experiment on automatic word prediction. On October 3, 2018, Bodt used the automatic predictions to come up with a list of sensible manual predictions to be checked during his follow-up field work. This list contained 630 different word forms in total, and about 65 words on average per variety.[6]

On October 5, 2018, List registered the experiment, including both the code, the data, and the automatically and the manually corrected reflex retrodictions with the Open Science Framework (`https://osf.io`) at `https://osf.io/evcbp/`. The basic idea of registering an experiment is to deposit a hypothesis prior to testing it with some provider, in order to make sure that the hypothesis was not created after the scholars inspected data and results. In the case of our word prediction experiment, the hypothesis consists of the 630 predictions we have come up with. That means, we do not provide a single hypothesis to be tested, but a rather long list of predictions that can all be tested individually. The field work to check the reflex retrodictions against the real word forms was carried out in October and November 2018. We are now in the process of comparing the accuracy of the word predictions and share our results in form of a publication and talks starting with the International Historical Linguistics Meeting in Canberra 2019.

---

[5] The corrected data was provided as file `<predictions-manual.tsv>` in our registration

[6] Given that the degree by which varieties were missing data was skewed, there was a high variation as to how many word forms were actually predicted per variety

3.6 Future evaluation of the experiment

Evaluating the accuracy of word predictions can be done in a very straightforward way by comparing the predicted word form with the attested word form segment by segment. A metric would then score all those cases in which a predicted segment differs from an attested one, and yield the average accuracy of a predicted word by dividing the number of correctly predicted sound segments by the number of incorrectly predicted sound segments. This procedure is already implemented in the LingRex software we used for this study, and likewise described and illustrated in List (2019). Thus, the reflex retrodiction task on controlled data sets can be easily automated and tested in those cases where data is artificially distorted and we know in advance that each missing word indeed *has* a counterpart in a cognate set of the given language.

When working with real-language data, and words that are really unattested at the time of the prediction, however, it is also possible that the predicted word does not exist in the target language, but has been lost due to lexical replacement. As a result, any metric that wants to judge the accuracy of a prediction experiment as we have conducted it on Western Kho-Bwa language data needs to assess first if the words that are attested for a given semantic slot are indeed cognate with the cognate set which was used in order to predict the unknown word form. If it turns out that the word is indeed not cognate with the words used for the prediction, this should not count as a failure of the method, but should instead be ignored when comparing the accuracy of the prediction experiment.

We are currently still discussing and evaluating the most useful metrics for evaluating the accuracy of both the automatic and the manually corrected predictions. Ideally, we would have addressed this problem even *before* registering the experiment. However, as we consider our research as a pilot study on the task of reflex retrodiction, we hope that our colleagues will understand that we were not able to *predict* completely how this could be carried out in an optimal manner.

4 Testing automatic predictions on Western Kho-Bwa

Before conducting an experiment of this kind, it is useful to compute the rate of accuracy we might expect from a random sampling of the data alone. For this purpose, we randomly deleted words from the existing data and then used the distorted data set to predict the deleted words. The accuracy is then computed for each word form by counting how many times the algorithm proposes the correct word form and how many times

it fails. This can be represented in a percentages score, our accuracy score. After 100 trials, documented in the supplementary material, the accuracy of the prediction experiment on the data reached 59% (0.5854), with an average proportion of 61% of the data being retained. Comparing this score with other data sets, as reported in List (2019), we can see that the Western Kho-Bwa language varieties are less easy to predict than Polynesian languages or Chinese dialects, but rather seem to be as challenging as the Burmish languages in the sample of Hill & List (2017). The fact that the prediction did not reach higher scores may also result from the fact that the original data is already sparse with respect to mutual coverage.

5 Conclusion

With languages disappearing at rates never experienced before, reflex retrodiction could become more and more important as a practical tool for historically oriented linguistic fieldwork. As the speakers of these languages are becoming fewer in number and older, there is a genuine risk that words that are important from a historical-comparative view point may be lost before they are recorded. Retrodicting these words may help to render the search for cognate forms in these languages less time-consuming and more efficient. For example, rather than having to ask how to say 'to carry by hand', 'to carry on the shoulder', or 'to carry on the back' in a given language, one could directly retrodict the missing forms and ask whether they exist in the target language, along with their exact semantic interpretation. Additionally, reflex retrodiction will make it easier to elicit cognates of certain words that contain rare segments in a given variety, such as marginally occurring distinctive onsets or rhymes. These actually attested forms can then be used to strengthen purported sound correspondences between linguistic varieties and reconstruct proto-forms.

In this paper, we have described an initial attempt to test how reflex retrodiction could be used in actual field work. Our experiment proposes a first workflow that illustrates how similar experiments could be carried out by colleagues working on other language families. There is no need to follow our workflow completely: scholars could just use their intuition before going back to the field to make lists of forms they think they should check again, which may already be a regular – though largely unreported – practice among field linguists. By sharing these predictions with the public through registering experiments with the Open Science Framework, scholars can not only share their current state of knowledge with the community, but also test it against the data they observe. Ideally, this can help to strengthen specific hypotheses, and it can also help to in-

crease the awareness that sound change is – in reality – to a large extend proceeding along regular pathways.

Readers may ask themselves why we report this experiment here in a stage where the major work of checking how well the reflex retrodiction works in the end has not yet been carried out. We decided to report this study already at this stage, since we hope to get some feedback from our colleagues. We are not only interested to receive suggestions for enhancement of our current study, but we would also like to hear how field workers dealing with historical language comparison of hitherto poorly investigated languages are making or have made use of reflex retrodiction.

Acknowledgements

This research was funded by the Swiss National Science Foundation Postdoc Mobility grant number P2BEP1_181779 (TB), the DFG research fellowship grant 261553824 "Vertical and lateral aspects of Chinese dialect history" (JML, 2015-2016), and by the the ERC Starting Grant 715618 "Computer-Assisted Language Comparison" (JML, http://calc.digling.org, 2017-2022). We thank Nathan W. Hill for his incredible help in developing and optimizing the workflow for computer-assisted language comparison. This research would not have been possible without the patient cooperation of the main language consultants in Arunachal Pradesh: Dorji Choijom, Rincin Buti, Phuntso Tsering and Palden Norbu (Chug); Norbu Dema, Dawa Lhamu, Nima Tsering and Rincin Dema (Lish); Tshering Dolma Nethungji, Geshi Tamu Yamchodu and Phinje Nasidu (Khoina); Sena Phinju Nathongji, Veena Rockpudu and Pema Choijom Yamnojee (Jerigaon); Nima Lhamu Chanadok and Kezang Rokpu (Khoitam); Karma Tsering Ngoimu, Dolma Sarmu and Chomu Sarmu (Rahung); late Dorji Dema Thungdok, Rincin Khandru Karma, Pema Sinchaji and Tashi Sinchaji (Rupa); Prem Khandu Thungon and Dombu Tsering Thongon Lama (Shergaon).

Associated material

The supplementary material accompanying this paper contains the code, the data, and all instructions needed to replicate the analyses described here. It can be downloaded from the Open Science Framework (https://osf.io/evcbp/), or from Zenodo (https://zenodo.org/record/1451176). The project is also hosted and curated with GitHub at https://github.com/lingpy/predict-khobwa/. The underlying data is available in CLDF

format from GitHub at https://github.com/lexibank/bodtkhobwa and from Zenodo at +++Link2Badded+++.

Author's contact details

List the postal and e-mail addresses for all authors using the following format:

Tim Bodt
School for Oriental and African Studies
University of London
Thornhaugh Street, Russell Square
London WC1H 0XG
United Kingdom

tb47@soas.ac.uk

Johann-Mattis List
Max Planck Institute for the Science of Human History
Kahlaische Str. 10
Jena, 07743
Germany

list@shh.mpg.de

References

Abraham, Binny et al. 2005. *A sociolinguistic research among selected groups in Western Arunachal Pradesh highlighting Monpa*. Unpublished manuscript.

Anderson, Cormac, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel & Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting* 3. 1–23.

Bhasker, J. & Tariq Samad. 1991. The clique-partitioning problem. *Computers & Mathematics with Applications* 22(6). 1–11.

Blench, Roger & Mark W. Post. 2014. Rethinking sino-tibetan phylogeny from the perspective of northeast indian languages. In Thomas Owen-Smith & Nathan W. Hill (eds.), *Trans-himalayan-linguistics*, 71–104. Berlin: de Gruyter.

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

Bodt, Timotheus Adrianus. 2014a. Ethnolinguistic survey of westernmost arunachal pradesh. a fieldworker's impressions. *Linguistics of the Tibeto-Burman Area* 37(2). 198–239.

Bodt, Timotheus Adrianus. 2014b. Notes on the settlement of the gongri river valley of western arunachal pradesh. In Anna Balikci Denjongpa & Jenny Bentley (eds.), *The dragon and the hidden land: social and historical studies on sikkim and bhutan. Proceedings of the bhutan-sikkim panel at the 13th seminar of the international association for tibetan studies*, 153–190. Ulaanbataar: International Association for Tibetan Studies.

Branner, David Prager. 2006. Some composite phonological systems in Chinese. In David Prager Branner (ed.), *The Chinese rime tables. Linguistic phylosophy and historical-comparative phonology*, 209–232. Amsterdam: Benjamins.

Brown, Cecil H., Eric W. Holman, Søren Wichmann, Viveka Velupillai & Michael Cysouw. 2008. Automated classification of the world's languages: A description of the method and preliminary results. *Sprachtypologie und Universalienforschung* 61(4). 285–308.

Burling, Robbins. 2003. The tibeto-burman languages of northeastern india. In Graham Thurgood & Randy J. LaPolla (eds.), *The sino-tibetan languages*, 169–191. New York: Routledge.

Chen, Chin-Chuan. 1997. Measuring relationship among dialects: doc and related resources. *Computational Linguistics and Chinese Language Processing* 2(1). 41–72.

Clackson, James. 2007. *Indo-European linguistics*. Cambridge: Cambridge University Press.

Dekker, Peter. 2018. *Reconstructing language ancestry by performing word prediction with neural networks*. Amsterdam: University of Amsterdam Master.

Dellert, Johannes & Gerhard Jäger. 2017. *NorthEuraLex (Version 0.9)*. Tübingen: Eberhard-Karls University Tübingen.

Dondrup, Rinchin. 1988. *A handbook on sherdukpen language*. Itanagar: Government of Arunachal Pradesh.

Dondrup, Rinchin. 1990. *Bugun language guide*. Itanagar: Government of Arunachal Pradesh.

Dondrup, Rinchin. 2004. *An introduction to boot monpa language*. Itanagar: Government of Arunachal Pradesh.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin

Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5(180205). 1–10.

Genetti, Carol. 2016. The tibeto-burman languages of south asia: the languages, histories, and genetic classification. In Hans Heinrich Hock & Elena Bashir (eds.), *The languages and linguistics of south asia: a comprehensive guide*. Berlin: Mouton de Gruyter.

Greenberg, Joseph Harold. 1963. *The Languages of Africa*. Bloomington: Indiana University Press.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2018. *Glottolog*. Version 3.3. Leipzig. http://glottolog.org.

Hill, Nathan W. & Johann-Mattis List. 2017. Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting* 3(1). 47–76.

Hrozný, Bedřich. 1915. Die Lösung des hethitischen Problems. *Mitteilungen der Deutschen Orient-Gesellschaft* 56. 17–50.

IPA Handbook. 1999. *Handbook of the International Phonetic Association: A guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press.

Jacquesson, François. 2015. *An introduction to sherdukpen*. Bochum: Brockmeyer.

Lewis, M. Paul & Charles D. Fennig (eds.). 2013. *Ethnologue: Languages of the world*. Dallas.

Lǐ, Dàqín. 2004. *Sūlóngyǔ yánjiū [research on puroik]*. Běijīng: Mínzú chūbǎn shè.

Lieberherr, Ismael & Timotheus Adrianus Bodt. 2017. Sub-grouping Kho-Bwa based on shared core vocabulary. *Himalayan Linguistics* 16(2). 25–63.

List, Johann-Mattis. 2016. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1(2). 119–136.

List, Johann-Mattis. 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. Valencia: Association for Computational Linguistics.

List, Johann-Mattis. 2018. *LingRex: Linguistic Reconstruction with LingPy*. Jena. https://doi.org/10.5281/zenodo.1544944.

List, Johann-Mattis. 2019. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* 1(45). 1–24.

List, Johann-Mattis, Michael Cysouw & Robert Forkel. 2016. Concepticon. A resource for the linking of concept lists. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2393–2400. European Language Resources Association (ELRA).

List, Johann-Mattis, Simon Greenhill, Tiago Tresoldi & Robert Forkel. 2018. *LingPy. A Python library for quantitative tasks in historical linguistics*. Jena. http://lingpy.org.

List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi & Robert Forkel. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution* 3(2). 130–144.

Meier-Brügger, Michael. 2002. *Indogermanische Sprachwissenschaft*. In collab. with Matthias Fritz & Manfred Mayrhofer. 8th edn. Berlin & New York: de Gruyter.

Michael, Lev, Natalia Chousou-Polydouri, Keith Bartolomei, Erin Donnelly, Vivian Wauters, Sérgio Meira & Zachary O'Hagan. 2015. A Bayesian phylogenetic classification of Tupí-Guaraní. *LIAMES* 15(2). 193–221.

Post, Mark W. & Robbins Burling. 2017. The tibeto-burman languages of northeastern india. 213–233.

Rutgers, Leopold Roland. 1999. *Puroik or Sulung of Arunachal Pradesh*. Paper presented at the 5th Himalayan Languages Symposium. Kathmandu.

Saussure, Ferdinand de. 1879. *Mémoire sur le système primitif des voyelles dans les langues indo- européennes*. Leipzig: Teubner.

Sims-Williams, Patrick. 2018. Mechanising historical phonology. *Transactions of the Philological Society* 116(3). 555–573.

Stonor, Charles Robert. 1952. The sulung tribe of the assam himalayas. *Anthropos* 47. 947–962.

Sūn, Hóngkāi. 1991. *Zàngmiǎnyǔ yǔyīn hé cíhuì* [*tibeto-burman phonology and lexicon]*. Běijīng: Zhōngguó Shèhuì Kēxué.

Sun, Tianshin Jackson. 1992. Review of zangmianyu yuyin he cihui (tibeto-burman phonology and lexicon). *Linguistics of the Tibeto-Burman Area* 15. 73–113.

Sun, Tianshin Jackson. 1993. *A historical-comparative study of the tani (mirish) branch in tibeto-burman*. Berkeley: Department of Linguistics, University of California PhD.

Tayeng, Aduk. 1990. *Sulung language guide*. Itanagar: Directorate of Research, Government of Arunachal Pradesh.

van Driem, George. 2001. *Languages of the himalayas - an ethnolinguistic handbook of the greater himalayan region*. Leiden: Brill.

Welsh, D. J. A. & M. B. Powell. 1967. An upper bound for the chromatic number of a graph and its application to timetabling problems. *The Computer Journal* 10(1). 85–86.

Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2016. *The ASJP database*. Jena: Max Planck Institute for the Science of Human History. http://asjp.clld.org.