Please quote this paper as follows:

Bodt, Timotheus A. and List, Johann-Mattis (forthcoming): Benefits of reflex prediction. A case study of Western Kho-Bwa. To appear in *Diachronica*.

Benefits of reflex prediction: A case study of Western Kho-Bwa[1]

Timotheus A. Bodt[1] and Johann-Mattis List[2]

[1] School of Oriental and African Sciences, University of London, London | [2] Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena

Abstract

While analysing lexical data of Western Kho-Bwa languages of the Sino-Tibetan or Trans-Himalayan family with the help of a computer-assisted approach for historical language comparison, we observed gaps in the data where one or more varieties lacked forms for certain concepts. We employed a new workflow, combining manual and automated steps, to predict the most likely phonetic realisations of the missing forms in our data, by making systematic use of the information on sound correspondences in words that were potentially cognate with the missing forms. This procedure yielded a list of hypothetical reflexes of previously identified cognate sets, which we first preregistered as an experiment on the prediction of unattested word forms and then compared with actual word forms elicited during secondary fieldwork. In this study we first describe the workflow which we used to predict hypothetical reflexes and the process of elicitation of actual word forms during field work. We then present the results of our reflex prediction experiment. Based on the experience we made during this experiment, we identify four general benefits of reflex prediction in historical language comparison. These comprise (1) an increased transparency of linguistic research, (2) an increased efficiency of field and source work, (3) an educational aspect which offers teachers and learners a wide plethora of linguistic phenomena, including the regularity of sound change, and (4) the possibility to kindle speakers' interest in their own linguistic heritage.

Keywords: prediction; word prediction; comparative method; regularity of sound change; computer-assisted language comparison; Western Kho-Bwa; preregistered research

1 An abridged popular scientific version of this paper appeared in Babel: the language magazine (Bodt and List 2020).

1. Introduction

The comparative method can be used to reconstruct proto-phonemes and proto-forms in languages no longer spoken or written. The method can also be used to predict phonemes, words or grammatical structures that have not yet been investigated or observed in a specific language, using techniques such as the one which Watkins calls FORWARD RECONSTRUCTION (Watkins 1962: 5, quoted after Sims-Williams 2018: 11). Field linguists, when eliciting data, rely on predictions to ease their work, whether it concerns minimal pairs for distinctive phonemes in the phonology, contrastive morphemes with distinct grammatical functions, additional lexemes with meanings similar or related to already elicited lexemes, or distinctive syntactic constructions. Hence, 'prediction' is an integral but hitherto largely undocumented part of linguistics. Exceptions include, for example, Grimm's work on Germanic, where he mentions the possibility to predict word forms based on his comparative analysis (Grimm 1822: 589), Greenberg's universals of grammar (Greenberg 1963), Blevin's predictions of possible and impossible sound patterns according to the theory of historical phonology (Blevins 2004: 3-24), the prediction of missing reflexes of cognate sets when searching for etymologies in a given language (Michael et al. 2015: 196), Amery's use of predictions and comparative linguistics to fill gaps in the vocabulary observed during the reclamation efforts on the Kaurna language (Amery 2016: 36), and Branner's description of word prediction in language contact situations (Branner 2006: 215).

However, as is the case with many aspects of the classical techniques of historical language comparison, including the identification of cognates and the proposal of proto-phonemes, prediction methods have, at least to our knowledge, never been explicitly proposed or discussed. Nonetheless, judging from conversations with actual practitioners of the comparative method, predictions are an indispensable tool in the field. We, therefore, think that a more explicit discussion of prediction techniques could play a vital role for the future of our discipline.

While the linguistic knowledge derived from the techniques for historical language comparison could be used for a wide range of predictions targeting different linguistic domains (see Bodt & List 2019: 24-27, for a recent overview on computational and manual techniques), we think that the task of 'reflex prediction'[2] deserves more attention in particular. Reflex prediction 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 given cognate set attested in different languages .

In order to test the predictive force and the usefulness of prediction studies for hypothesis testing, data validation, and cognate discovery in historical linguistics, we carried out an experiment on missing words in Western Kho-Bwa language data. Western Kho-Bwa is a sub-group of the Sino-Tibetan (or Tibeto-Burman, or Trans-Himalayan) language family that has thus far not been thoroughly investigated. Recent studies have, however, convincingly shown that the Western Kho-Bwa linguistic varieties form a coherent sub-group (Lieberherr & Bodt 2017, Bodt 2019, Bodt forthcoming). Our current paper does not aim to present further evidence for the internal coherence of the Western Kho-Bwa group. Rather, it presumes a priori that the

2 Strictly speaking, we are not predicting the pronunciation of words and word forms here, since the pronunciation already exists in the languages. A more adequate term might be 'retrodiction', a term occasionally used in the German literature, which denotes statements on possible past events. However, since the current pronunciation of a word neither belongs to the past nor the present, and because the term retrodiction is used slightly differently in the English literature, we decided to use the term 'prediction' throughout this study. In addition, as one of the anonymous reviewers pointed out, we predict phoneme sequences and the most likely phonemic realisation, and not the exact phonetic surface realisations by speakers. This reviewer also mentioned that what we are doing could be termed as 'hindcasting': doing prediction from a starting point in the past. However, we want to stress that our starting point lies at present, with actually attested forms in actually attested varieties.

eight Western Kho-Bwa varieties are, in fact, genetically related.

We used a computer-assisted workflow for predicting the most likely phonological shapes of a set of missing morpheme reflexes in an etymological dataset of eight Western Kho-Bwa language varieties. These predicted values were then manually refined by combining these morphemes into lexeme reflexes, or actually verifiable words, and evaluated by comparing them to the attested reflexes observed during subsequent fieldwork.

In the following sections, we will succinctly describe the background of the experiment and the way in which the predictions were made and evaluated (Section 2). We will then present the results (Section 3) and the benefits of predictions (Section 4), followed by a conclusion and outlook for future applications and research (Section 5).

2. Predicting unelicited words in Western Kho-Bwa

2.1 The Western Kho-Bwa languages

The Indian state of Arunachal Pradesh is located in one of the ethno-linguistically most diverse regions of the world. The difficult topography and the geopolitical location of the state, being governed by India but claimed by China, has for long restricted research. Hence, descriptions of the languages of Arunachal only started appearing during the last two decades of the previous century. A concise overview of the works from both the Indian and the Chinese side of the border is presented in Lieberherr & Bodt (2017) and Bodt & List (2019). Based on these descriptions, all commonly consulted linguistic handbooks such as Genetti (2016) and Post & Burling (2017) and reference catalogues on languages, such as Ethnologue (https://www.ethnologue.com; Eberhard, Simons & Fennig 2019) and Glottolog (http://glottolog.org; Hammarström, Forkel & Haspelmath 2020), mention a cluster named ‘Kho-Bwa’ (van Driem 2001) as a (potential) branch of Tibeto-Burman in western Arunachal Pradesh.

The languages hypothesised to belong to this cluster are Puroik, Bugun, Sherdukpen, Sartang, Khispi (Lishpa) and Duhumbi (Chugpa). The latter four languages comprise a total of eight distinct varieties: Khispi; Duhumbi; the four varieties of Sartang (Khoina, Khoitam, Jerigaon and Rahung); and two varieties of Sherdukpen (Rupa and Shergaon). These linguistic varieties, spoken in the valleys of the Gongri and Tenga rivers in the western part of the Kho-Bwa speech area, form a coherent sub-group within the Kho-Bwa cluster: Western Kho-Bwa (Bodt 2014a, Bodt 2014b). Considering the low speaker population (between 400 for the Jerigaon variety of Sartang and 3,000 for the Rupa variety of Sherdukpen) and the rapid socio-economic and cultural changes in this area, all these varieties must be considered endangered.

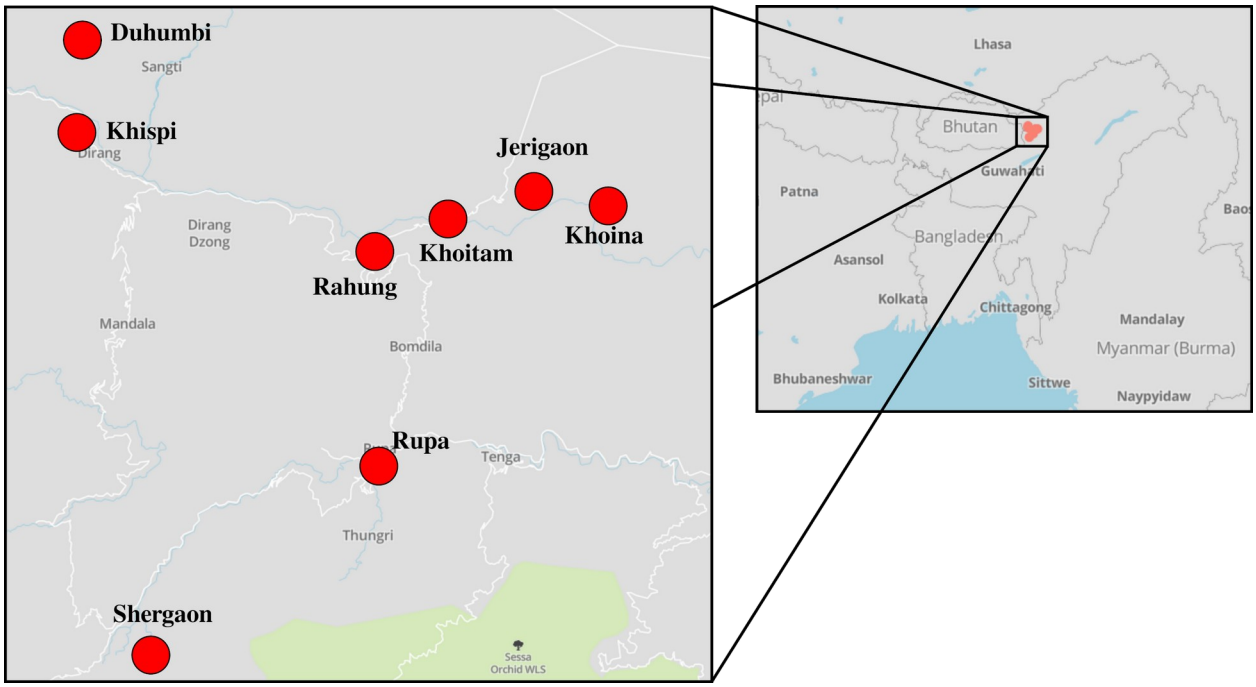

Figure 1. Approximate location of the Western Kho-Bwa varieties.

One salient morphological characteristic of the Western Kho-Bwa languages has had a considerable influence on the way in which the data were analysed. The Western Kho-Bwa languages have a rich system of affixes that define parts of speech and lexico-semantic categories of nouns. Many of the initial predictions were of such affixes that form concepts in combination with roots. But neither roots nor affixes could be elicited in isolation: They had to be combined to create meaningful predictions. So, in addition to 'morphological predictions' of sequences of phonemes in individual morphemes, we made 'lexical predictions' in which we combined morphemes to form concepts that could be elicited in the field.

The starting point of our experiment was an etymological dataset, reflecting all eight distinct Western Kho-Bwa varieties and assembled during fieldwork on Duhumbi, conducted in Arunachal Pradesh between 2012 and 2017. The same 550 concepts from a single wordlist were elicited from at least two speakers, one male and one female, from each variety, with an ad-hoc collection of additional items as they came up during elicitation. These data were used for a lexicostatistical analysis of the Kho-Bwa languages (Lieberherr & Bodt 2017), the reconstruction of Proto-Western Kho-Bwa (Bodt 2019 and Bodt forthcoming) and have subsequently been stored on Zenodo (https://zenodo.org). Links to these data, including the elicitation wordlist and all the original and cut sound files, can be found in Appendix A1.

2.2 Background of the Study

While analysing the data both quantitatively and qualitatively, we observed that there were gaps, where certain varieties lacked the forms for certain concepts. These gaps occurred because of oversights or confusion during elicitation, because informants indicated they did not know or remember the form of the concept, or because informants stated that a certain concept did not exist in their variety. Given that, shortly before we started our analysis, a new automated method had been developed that allows to infer sound correspondence patterns across multiple languages

and predict how unknown reflexes of a given cognate set would sound (List 2019), we decided to take the gaps in the data as an opportunity to test how well unknown word forms can be predicted for Western Kho-Bwa languages.[3]

After having set up the computer-assisted workflow that would allow us to predict the missing word forms in our data, we made a preregistration of the predicted word forms via the Open Science Framework in order to ensure that an immutable version of our hypotheses was openly available prior to verification.[4] In addition, we wrote a working paper in which we introduced our experiment in more detail, along with technical details on the computer-assisted workflow and our plans for the verification of the results (Bodt & List 2019). After carrying out the fieldwork during which the predictions were verified, we analysed our results and presented these to colleagues at the 24th International Conference for Historical Linguistics (Canberra, Australia, July 2019). Subsequently, we committed our findings to writing, both in an abridged form for a popular science journal with a focus on language (Bodt & List 2020) and as the current study.

2.3 Workflow for reflex prediction

In order to predict a sufficiently large number of words that we could use to conduct our prediction experiment, we designed a computer-assisted workflow that would help us to 1) fill gaps in our data more systematically and 2) make sure that the data would be machine- and human-readable at the same time.

Our workflow consists of seven steps. All seven steps can, theoretically, be done by the linguistic expert manually. However, some of these steps can make use of existing computational solutions, which greatly increase the efficiency of the experiment. In addition, for all steps, tools exist that support the annotation process. Hence, we refer to our workflow as a 'computer-assisted' (as opposed to both a fully 'computer-based' and an entirely 'manual') workflow. We schematically present our seven steps in Figure 2.

In the first step, we normalized the data in such a way that they would be amenable for computational treatment (1, normalization, see also Appendix A2). In the second step, partial cognates in the data were manually identified and annotated (2, partial cognate identification), and then automatically aligned (3, partial cognate alignment). Once cognate sets were aligned, correspondence patterns - regular sound correspondences between cognate forms in the different varieties - were automatically identified (4, correspondence pattern identification). These correspondence patterns were then used to automatically predict individual morphemes wherever the original data lacked a form for a given concept in a certain language variety (5, morpheme prediction). Not all the gaps in the original dataset were due to missing data. Some had been deliberately excluded before, since they were obvious borrowings that would not be useful for the reconstruction of the Western Kho-Bwa proto-language, which was the original purpose of the data collection.[5] Therefore, only those morphemes that were judged to be suitable for the experiment were manually selected in the sixth step (6, reflex selection). In the final step, these predicted morphemes were manually inspected, if deemed necessary corrected, and assembled to

3 For readers interested in a more detailed overview of the background of this experiment, we suggest the supplement to this publication and Bodt & List (2019), which describes the underlying data, the set-up of the dataset, and the way in which we came to the predictions. The predictions themselves were registered online at an Open Science Platform online registration at https://osf.io/evcbp/ (Bodt, Hill & List 2018).

4 This kind of 'planned research' is now common in the social sciences and psychology (Nosek et al. 2019), but we do not know of any applications in historical linguistics and language documentation so far.

5 These loans were later added to the database, filling up earlier gaps.

form potential words expressing the missing concepts in the targeted language varieties (7, lexeme prediction).

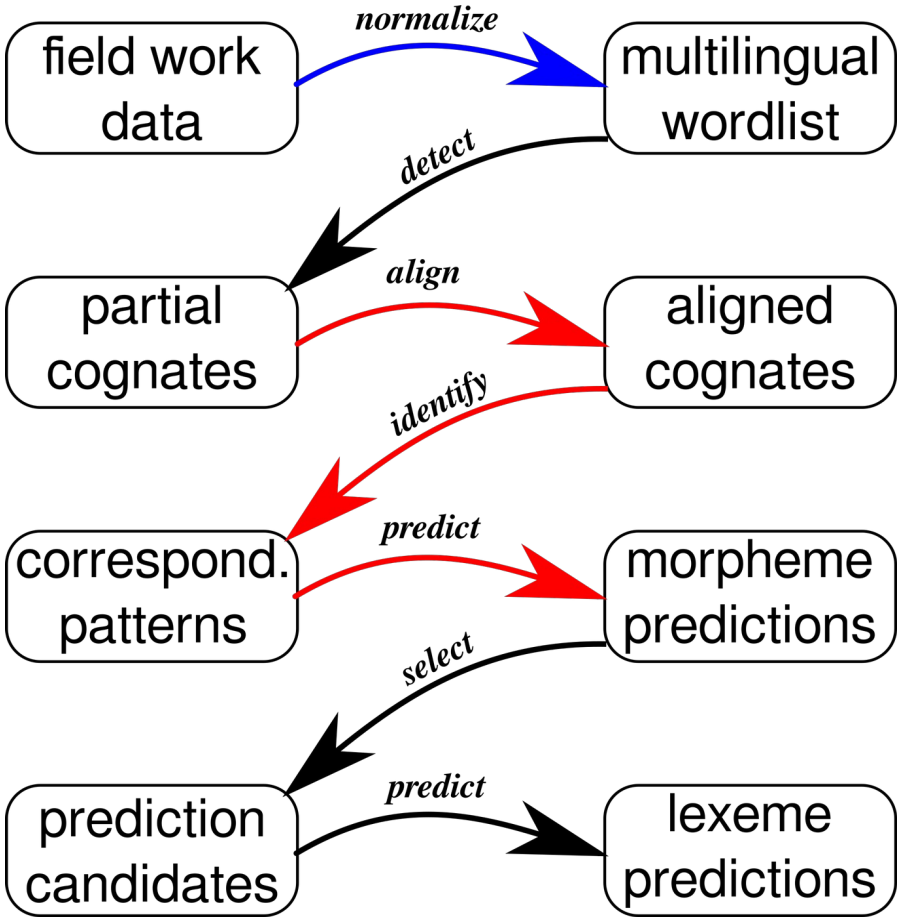

Figure 2. Workflow for the prediction experiment. Red arrows indicate fully automated approaches, black arrows indicate fully manual approaches, and blue arrows indicate semi-automated approaches.

The first three steps of our workflow, the normalization, the semi-automated initial assignment of partial cognates, and the automated alignment of partial cognates, have been presented in both our study introducing the experiment prior to conducting it (Bodt & List 2019), and in an extended tutorial in presenting the application of the workflow to Hmong-Mien language data (Wu et al. 2020). For this reason, we will not detail these three steps here, and instead refer the reader to our previous studies as well as to the appendices accompanying this paper (Appendix A2 and A3), where major aspects are summarized.

In order to retrieve sound correspondence patterns from the aligned cognate sets in our data, we used the method proposed by List (2019), which infers them from aligned cognate sets with the help of a network-based procedure. To illustrate this procedure, consider the data for three exemplary concepts and four exemplary varieties of Western Kho-Bwa as shown in Table 1, which is representative for the way in which all of our data (4721 words distributed across 662 concepts and 8 varieties) was annotated during the application of our computer-assisted

workflow. In this long-table format (Forkel et al. 2018), every word is displayed in its own row. Aligned word forms can be found in the column 'Aligned Form' (called ALIGNMENT in our machine-readable format), with sounds being separated by a space and morphemes being separated by a plus character. The column Glosses (called MORPHEMES in our machine-readable format) provides explanations of the lexical structure in glossed form, with glosses written in capital letters pointing to lexical morphemes and glosses in lower-case pointing to grammatical morphemes (prefixes, suffixes, etc.), following the suggestion by Schweikhard & List (2020). Cognates are annotated for each morpheme, not for entire lexemes, by assigning the same numeric identifier to all morphemes which are considered to be cognate (regardless of their original meaning).

Table 1. Exemplary data of aligned partial cognates.

Variety	Concept	Aligned Form	Glosses	Cognates
Duhumbi	spittle, spit	h i n + t u s	hna-prefix SPITTLE	1 2
Jerigaon	spittle, spit	t ɛː -	SPITTLE	2
Khispi	spittle, spit	h i n + t u s	hna-prefix SPITTLE	1 2
Khoitam	spittle, spit	t ɛː -	SPITTLE	2
Duhumbi	throw	t ɔ s	THROW	3
Jerigaon	throw	tʰ øˀ -	THROW	3
Khispi	throw	t ɔ s	THROW	3
Khoitam	throw	tʰ eˀ -	THROW	3
Duhumbi	down	b e	DOWN	4
Jerigaon	down	b uː + t ɛ n	DOWN allative	4 0
Khispi	down	b e	DOWN	4
Khoitam	down	b uː + r ɔ	DOWN ablative	4 0

Even without the aid of a computer, we can easily derive the sound correspondences from this example by simply considering each alignment in separation and tabulating the sounds which we find in this alignment in each particular column, as shown in Table 2 for all sounds found in the cognate sets in Table 1.[6] Due to the small number of examples in these Tables, most of the correspondence patterns observed occur only once, but when comparing across the entire dataset, we find enough evidence to support each of them by at least two more examples.

6 While our example can be easily digested manually, it is important to note, as also shown in the study by List (2019), that the inference of correspondence patterns can become very complex, especially when the number of languages one compares at the same time increases.

Table 2. Deriving sound correspondence patterns from aligned cognate sets. Column Count is based on the cognate sets in Table 1.

Position	Count	Duhumbi	Jerigaon	Khispi	Khoina	Cognates
initial	1	h	Ø	h	Ø	1
initial	1	t	t	t	t	2
initial	1	t	tʰ	t	tʰ	3
initial	1	b	b	b	b	4
nucleus	1	i	Ø	i	Ø	1
nucleus	1	u	ɛː	u	ɛː	2
nucleus	1	ɔ	øˀ	ɔ	eˀ	3
nucleus	1	e	uː	e	uː	4
coda	1	n	Ø	n	Ø	1
coda	2	s	-	s	-	2, 3

When inspecting the table, three aspects are important to consider for the automated part of our prediction procedure. First, right from the start, we distinguish the sounds in an alignment according to their basic positions (initial, nucleus, coda). Second, we may face situations in which a correspondence pattern is not filled, due to a lack of data. This is, for example, the case for the patterns of the cognate set 1, where we indicate missing data with the symbol Ø. Third, we will inevitably find mergers and splits in our correspondence patterns, reflected in the same sound value in one language variety which shows different correspondences in other language varieties, as in the case of the initials in cognates 2 and 3, showing mergers in Duhumbi and Khispi, or splits in the other varieties, depending on the perspective.

When predicting lexemes for words missing in our data, we start by predicting missing morphemes based on the aligned cognate sets identified before. This is stage 5 in our workflow and follows a very schematic procedure: For a given aligned cognate set in which a reflex for a particular language variety is missing, we look at each column in our alignment and compare it with our list of correspondence patterns. Take, for example, the concept "curcuma", for which we have [b ɔ s] as form in Duhumbi and [b eˀ] in Khoitam, and no attested forms in Khispi and Jerigaon. To align both word forms with each other, we would add a gap symbol to the Khoitam form to indicate that this form lacks a coda: [b eˀ -]. The alignment along with the missing forms is shown in Table 3.

To predict the forms, we start from the initial column of the alignment, which shows [b, ?, ?, b] as reflexes ("?" marks the sounds in Jerigaon and Khispi, which we want to predict), and compare it with our correspondence pattern table, Table 2. Here, in the fourth row, we find the pattern [b, b, b, b] and therefore conclude that the initial sound in both Jerigaon and Khispi should be [b]. Proceeding in this way with the other columns of the alignment for "curcuma",

[ɔ, ?, ?, eˀ] and [s, ?, ?, -], we find [ɔ, øˀ, ɔ, eˀ] in the seventh row, and [s, -, s, -] in the final row. Hence, we predict [b ɔ s] for Khispi and [b øˀ] for Jerigaon “curcuma”.

Table 3. Alignment of the two word forms for "curcuma" in Duhumbi and Khoitam, with question marks indicating those sounds that our prediction method needs to predict.

Language	**Alignment**		
	Initial	**Nucleus**	**Coda**
Duhumbi	b	ɔ	s
Khispi	?	?	?
Jerigaon	?	?	?
Khoitam	b	eˀ	-

Note, that this procedure may also yield ambiguous cases, in which a given column in an alignment is compatible with more than one correspondence pattern. In order to display such potential ambiguity, it is possible to list the potential candidates in the order of their frequency of occurrence in the dataset. Hence, the automatic approach selects correspondence pattern candidates according to the frequency in which they recur in the data. When applying this procedure, the automated procedure yields the form [b ɔ s|ɛ] for "curcuma" in Khispi, since across the whole dataset, we find 8 examples for the pattern [s, -, s, -] and three examples for an alternative pattern [s, -, ɛ, s]. When computing our automated predictions, we computed three different versions, one where no fuzzy sound candidates were allowed, one with up to two fuzzy candidates per sound, and one with up to three candidates. Since our workflow for reflex prediction is explicitly computer-assisted, and not computer-based, all automatically proposed predictions for individual morphemes were later manually refined, taking additional knowledge about conditioning context into account.

In the case of "curcuma" the predicted morpheme is identical with the predicted lexeme since all Western Kho-Bwa languages have a mono-morphemic noun for "curcuma". However, this does not hold for all cases, and often the lexeme which we want to predict may consist of multiple morphemes. Thus, the word for "deity, ghost" in Duhumbi and Khispi is [l a]. In Khoitam, we find [m ə + l ɔ:], composed of the prefix [m ə] which recurs in many nouns in this variety, and [l ɔ:], which is cognate with [l a] in the other varieties. While it is straightforward to predict a morpheme [l ɔ:] for Jerigaon, we could not find a way to decide algorithmically if we should also propose a prefix [m ə] for the lexeme, since this requires a lot of circumstantial knowledge about the language varieties in question which we cannot formalize in a straightforward manner. For this reason, the last stage of our workflow, the prediction of full word forms based on the previously selected morpheme candidates, was carried out by our language expert in an exclusively qualitative manner.

All in all, the workflow yielded as many as 2106 morpheme predictions of which 630 candidates were selected for the experiment. From these 630 candidates, 519 word forms were qualitatively composed and refined. To make sure that the prediction candidates were publicly available before they could be verified in fieldwork, the experiment was registered with the Open

Science Framework (https://osf.io/evcbp) on October 5th, 2018 (Bodt, Hill & List 2018), and described in detail in a working paper published in early 2019 (List & Bodt 2019).

2.4 Elicitation

The elicitation sessions for verification of the predictions took place in October and November 2018 and were conducted with a single speaker of each variety. The entire elicitation session was recorded, and the concepts were written down in IPA. Every predicted form that was reflected in a variety was triple recorded separately, to allow closer scrutiny of the phonetic form later on. The recordings of most of the individual forms and triple repetitions were cut, named, and saved as WAVE files. They are publicly available as part of the supplementary material accompanying this paper.

There are two main reasons why the prediction of a lexical reflex for a hitherto unelicited word form may not match the actually attested form. One major reason are erroneous predictions for individual sound segments that result from the workflow by which the individual morphemes were predicted (stage 5 in our workflow). Another major reason is lexical change. The word form expressing the concept in question may have been replaced, or it may have been an innovation in the languages where its cognate counterparts have been attested. Since lexical change processes are extremely hard (if not impossible) to predict, a failure of lexeme predictions due to lexical change cannot be directly controlled and needs to be distinguished from a failure resulting from the prediction based on sound correspondences. Typically, these two major sources of error can be distinguished rather easily. In the case of lexical replacement, the attested word form would diverge greatly from the predicted word form. In the case of the erroneous selection of correspondence patterns, the attested and the predicted word form would show a certain phonetic similarity, but not be completely identical with respect to all sound segments. Although lexical change happens frequently, the original word forms are often not completely lost from the language variety but have rather shifted their meaning. They can still be elicited, but elicitation with the help of the expected meaning is not possible. In order to account for the problems introduced by lexical change, we used the two-stage elicitation process described below.

During the elicitation sessions, in all cases except Khoina[7], a standard procedure was followed. The respondent would be explained the purpose and goal of the elicitation session and asked for consent to the recording and its subsequent storage, usage, and dissemination. The respondent would be asked the concept, commonly in Hindi, Tshangla or English.[8] For example, the Jerigaon respondent was asked “How do you say *nīce utarnā* "to descend"?”. If a respondent would provide a form the same as, or similar to, the predicted form, this attested form would be noted. Since the Jerigaon respondent gave the form [j y:] "to descend" (ID 267 in our wordlist), this is the same as the predicted form [j y:] (ID 265) and it matches both the lexeme prediction as well as the individual sounds given in the morpheme prediction. If the respondents would state a different form, they would be asked its general meaning, which would be noted. For example, the prediction for "to wait" in Rahung was [l a ŋ] (ID 2132). However, the response to the question “How do you say *pratīkśā karnā* or *ruknā* "to wait"?” was [tʰ u ŋ] (ID 2133). Inquiring if there were more forms for "to wait" also did not uncover a cognate form. Here, the lexeme prediction clearly failed, because the attested form [tʰ u ŋ] does not even approximately match

7 The literate Khoina respondent took the elicitation list a day beforehand and wrote her answers on the sheet, these were then discussed and recorded the next day.

8 The link to the concept list in English and Hindi is provided in Appendix A1.

the predicted form [l a ŋ]. In such a case, where the attested form was not considered ‘cognate’ with the prediction, the respondent would be asked whether there are any other words that describe the concept that was elicited: In some cases, based on background knowledge, hints would be given. Sometimes, this resulted in a cognate form: for example, after being asked “How do you say hearth or fireplace?”, a respondent may have first provided the name of the trivet used for placing a cooking pot above the fireplace, but asking “Is there another word that can refer to the hearth or fireplace?” may provide the form for "hearth or fireplace" itself. This was then noted as 'full match'.

In some cases, only one morpheme of a polymorphemic prediction was cognate with a morpheme in an attested form. This was especially the case with prefixed concepts, where different varieties had different prefixes or even lost them. These cognate morphemes were then listed as 'partial matches'.

If directly asking for the concept did not yield a form that could be considered cognate, the prediction itself would be asked. This would sometimes result in a cognate form as well, as this method of elicitation encourages respondents to think beyond the box, to dig in their memory, and also captures words that may have undergone semantic change or lexical compounding. These forms were noted down as ‘semantically shifted matches’. For example, when the Jerigaon respondent was asked how to say *kharāb*, a loose Hindi translation of the English adjective "bad", she replied [a + n uː] (ID 147). Indeed, this form has cognates in the Rahung, Rupa and Khoitam forms for "bad", but it is not phonetically similar to the predicted form, which was [a + z ɐ̃ː] (ID 146). The respondent also said that [a + n uː] is the only word they have for "bad". Then, the Jerigaon respondent was asked “Does your language have a word that sounds like [a + z ɐ̃ː], and what does it mean?”. In this case, the respondent replied that there is a word [a + z ɑ̃ː] (ID 150), and that it means "white", which is not semantically equivalent to "bad". So, the conclusion was that Jerigaon does not have a word that sounds like [a + z ɐ̃ː] and means "bad". In a comparative perspective, it was found there are two reconstructed forms for the concept "bad". The form *a-z^{jw}an (in segmented notation [a + z^{jw} a n]) has reflexes in Khispi, Duhumbi, Khoina, Khoitam, Rahung and Rupa, the form *a-na ([a + n a]) has reflexes in Khoina, Jerigaon, Khoitam, Rahung, Rupa and Shergaon. The distinction in languages that have reflexes of both forms, i.e. Khoitam, Rahung and Rupa, is a semantic one: reflexes of *a-z^{jw}an refer to "poor (antonym of "rich", or "poor (in quality)")" whereas reflexes of *a-na refer to "bad (of character)". This reflects the multiple semantic contexts in which one can use the concept "bad" in English ("a bad person", "a bad day", "a bad mark", "a bad car") and *kharāb* in Hindi (as meaning "bad, inferior (of quality or character)", "destroyed", "dysfunctional (of character or a machine or tool)").

Elicitation of the prediction could also yield a positive response, for example, in the case of the verb "to cover" which was predicted for Rahung as [t^h ɛ ŋ] (ID 1820). When the respondent was asked “How do you say *dhãknā* "to cover"”, she replied [b y k]. The respondent said there is no other word with a meaning "to cover". When asked whether there is a word like [t^h ɛ ŋ] with a meaning like "to cover", the respondent replied there is the word [k^h a n + t^h ɛ ŋ] with the meaning *dhakkan* "cover, lid" (ID 1821). So, whereas a form [t^h ɛ ŋ] did not survive as the verb "to cover" in Rahung, it survived in a semantically related compound "cover, lid".[9] Cases such as this were noted as 'partial matches' where a lexical compound was attested that could nonetheless be considered (partially) cognate with the prediction.

9 In the Western Kho-Bwa languages, the verb "to cover" has reflexes of two inherited roots, with semantic distinctions in those varieties that have reflexes of both roots. There are also two non-cognate words, one of which is likely a loan.

2.5 Evaluation

Based on the recordings of the elicitation sessions, all elicited forms were transcribed into a spreadsheet and later added to a comparative wordlist, containing predicted and attested forms. In the comparative wordlist, which is available in the form of a spreadsheet that can be browsed and edited with the help of the EDICTOR application (List 2017), we use the same annotation practices that we used for our comparative database to compare predicted with attested forms. Examples for this practice are given in Table 4.

The format shown in Table 4 allows for a very convenient evaluation of the prediction experiment, since it makes explicit if (1) a prediction can be verified at all, and if this is the case, (2) how well the predicted morpheme corresponds to the attested one. First, if the cognate IDs in the Cognates column of the (automatically or expert-) predicted morphemes and the cognate IDs in the Cognates column of the attested morphemes for a certain concept do not show any overlap, the predicted form cannot be verified against the attested form, since according to our expert judgments, the attested form is not 'cognate' with the predicted form due to various processes of lexical change. We can then note a 'mismatch' for every lexeme where the prediction projects different morphemes than we actually observed. If not all of the attested morphemes match with morphemes in our predicted word form, we note a 'partial match' for that specific morpheme. For example, in Jerigaon "bad", the prefix with cognate ID 99 is correctly predicted, but the main morpheme or root has cognate ID 220 in the (automatic and expert) predicted form, but cognate ID 102 in the attested form. Hence, there was lexical change that caused an incorrect prediction for this morpheme. Second, for all predicted morphemes whose cognate identifiers in the Cognates column have a counterpart among the attested word forms, such as in the case of the concept "split" in Table 4, we can VERIFY TO WHICH DEGREE THE PREDICTED MORPHEME RESEMBLES THE ATTESTED MORPHEME by measuring the phonetic similarity of the aligned predicted and attested forms, which gives us insights in the phonetic accuracy of the predictions we made. Any dissimilarities between the predicted and attested forms of full matches can only be attributed to human failure.

To score the prediction accuracy of an individual pair of predicted and attested word forms, in the case of full or partial matches, we align both forms with each other and simply count how many times each predicted sound segment is identical with the attested form and divide the number of matches by the overall length of the alignment. In the case of Rahung "split" in Table 4, for example, we find that both the automated and the expert prediction [j ɔ] differ in the nucleus from the attested form [j oˀ]. We thus find one match and divide this by the length of the alignment and hence arrive at a score of 1 / 2 = 0.5. In the case of Shergaon "ask", which was predicted as [dʑ i k] (ID 2438) and attested as [z i t] (ID 2439), only one segment is identical in the predicted and the attested reflex, and we thus calculate the score as 0.333..., dividing 1 by 3. In this way, we can calculate the prediction accuracy for all pairs of predicted and attested words in our sample. In order to calculate general scores for prediction accuracy, we take the average of all the pairs in our sample for which an attested form could be elicited.

In the case of the automated prediction allowing for 'fuzziness' (with up to three candidates per predicted sound), the algorithm yielded the form [tɕʰ ũ:|ɔ|a ŋ] for Rahung "above, top" (cognate set 58). The attested form is [tɕʰ ũ: ŋ]. Since the fuzzy predictions show a preference order, with the first candidate being the supposedly best one, reflected in the majority of the sound correspondence patterns, we treat this as a perfect match, since we have two direct matches, and the first candidate of the fuzzy proposal matches the attested sound as well ([ũ:|ɔ|a]

vs. [ũ:]). Had the order been differently, with the correct sound as the second item in the fuzzy proposal ([ɔ|ũ:|a]), we would score the prediction as 0.833…, rewarding the fact that the second candidate matches with 0.5 points, and calculating 1 + 0.5 + 1 = 2.5, divided by the number of segments (3). Had the correct sound been the third of three proposed sounds ([ɔ|a|ũ:], we would score the prediction as 0.777…, rewarding the fact that the third out of three proposals matches with one third (0.333…), counting 1 + 0.333… + 1 = 2.333…, divided by 3. By evaluating fuzzy matches in this way, we account for their ordered nature as well as for the fact that the non- or less fuzzy predictions are directly derived from the fuzzier ones, following the preference order of proposed sound segments.

Table 4. Annotation of predicted and attested forms in our comparative wordlist.

ID	Language	Concept	Prediction	Aligned Form	Glosses	Cognates
1819	Rahung	cover (v)	Automatic	Ø ɛ ŋ	COVER	505
1820	Rahung	cover (v)	Expert	tʰ ɛ ŋ	COVER	505
3114	Rahung	cover (v)	Attested	b y k	COVER-2	144
1821	Rahung	cover (n)	Attested	kʰ a n + tʰ ɛ ŋ	khan COVER	0 505
145	Jerigaon	bad	Automatic	z ɐ̃ː	BAD-1	220
146	Jerigaon	bad	Expert	a + z ɐ̃ː	a-pref. BAD-1	99 220
147	Jerigaon	bad	Attested	a + n uː	a-pref. BAD-2	99 102
2089	Rahung	split	Automatic	j ɔ	SPLIT-2	153
2090	Rahung	split	Expert	j ɔ	SPLIT-2	153
2091	Rahung	split	Attested	j oˀ	SPLIT-2	153

3. Results

We separate the discussion of our results in two sections. Section 3.1 presents the quantitative evaluation of our prediction experiment, discussing the categorisation of the predictions and the evaluation of both the automated and the manually adjusted predictions. Section 3.2 presents the qualitative evaluation of our experiment, discussing possible reasons for discrepancies between the predicted and the attested forms, including examples of how these discrepancies resulted in the discovery of previously unknown sound correspondences.

3.1 General results

As mentioned above, a total of 519 predictions were made. Of these, 454 could be elicited, that is, for 454 items, a response was obtained from the informants. In 65 cases, no response could be obtained. Either the informant did not understand the concept and the concept could not be correctly explained, or the respondent did not have any response. Of the 454 elicited predictions, we obtained 'full matches' for 235 cases. This means that the attested word form was a true reflex

of all the cognate sets that were used to predict it. In 48 cases, no full matches could be found, but 'partial matches'. This means that not all morphemes of the attested form were true reflexes of the cognate sets we used to predict the word form. In 44 cases, the attested form neither fully nor partially matched with the predicted form, but we uncovered a semantically shifted reflex through the second step in the elicitation process, the elicitation of the predicted forms themselves, i.e. 'semantically shifted matches'. In 127 cases there were no matching forms: Neither full, nor partial, nor any forms displaying semantic shift. In total, this means that 72% of the elicited predictions (235 'full direct matches', 44 'partial matches', and 48 'semantically shifted matches' out of 454 successfully elicited forms) could also be VERIFIED: Since we provided an explicit prediction in the form of a concrete sound sequence, we can now compare, how well our prediction compares to the attested sound sequence.

Table 5. General results of the experiment on word prediction.

Variety	Predicted	Elicited	Full Match	Partial Match	Semantically Shifted	No Match	Proportion
Duhumbi	19	19	3	1	7	8	0.58
Jerigaon	109	80	53	3	6	18	0.78
Khispi	39	37	18	3	5	11	0.70
Khoina	72	66	30	4	4	28	0.58
Khoitam	53	49	26	8	5	10	0.80
Rahung	65	56	28	11	6	11	0.80
Rupa	46	40	15	6	4	15	0.63
Shergaon	116	107	62	12	7	26	0.76
TOTAL	519	454	235	48	44	127	0.72

The results of this first comparison of predicted and verifiable forms, which have been automatically derived from our comparative wordlists, are given in Table 5, with details for each language variety (data and code are available from the supplementary material accompanying this paper and described in Appendix A4). Among the predictions that could not be elicited, a few, such as "horsefly" and "present marker", could not be elicited in any of the varieties. As an example for a full match, consider the concept "hanging bridge" which was predicted as [ɛ a m] (ID 782) in Khispi and for which the elicitation yielded the form [ɛ a m] (ID 783). In the concepts "sambar deer" and "pubic hair", we find examples of a partial match. "Sambar deer" was predicted for Shergaon as [s ə + z u k] (ID 2582) but had as attested form [z u k] (ID 2583) because of the LOSS of the prefix. The concept of "pubic hair" was predicted as [m y ŋ] in Khoitam (ID 1634), but had as attested form [a + m i ŋ] (ID 1635) because, compared to the forms on which the prediction was based, this variety had ADDED the a-prefix for body parts. In both cases, only the predicted root (i.e. the second morpheme) could be evaluated for accuracy, not the prefix. There are also several examples of semantically shifted matches, such as the concept of "fence" which was predicted in Khoina as [g u ŋ] (ID1100). The elicited response for "fence", however, was [s + t^h ɑ:] in Khoina (ID 1104), which is a loan. When eliciting the actual prediction, the respondent indicated that the word [g u ŋ] refers to "a small, moveable, temporary

bamboo enclosure that is used to separate the calves from milking cows at night": The inherited reflex had undergone semantic change.

The vast majority of the items for which neither a partial, nor a semantically shifted match could be obtained, were those where the proto-language had two semantically closely related roots. Some descendant varieties have reflexes of one root, other varieties have reflexes of another root, and some varieties may have reflexes of both roots, with the original or a different semantic distinction preserved. An example are the predictions based on Duhumbi [d ɔ ŋ], Khispi [d ɔ ŋ], Khoina [r u ŋ] "to bind" (IDs 369, 370 and 371). During initial elicitation of the concept, it was found that all the other Sartang and Sherdukpen varieties have the word [h a k] for "to bind" (IDs 159, 1455, 1779, 2169 and 2481). In subsequent elicitation of the predicted form, it was found that Khoitam has a form [r u ŋ] "to assemble (people); to pile up (things)" (ID 1458) and Rahung has a word [r u ŋ] "to cut" (ID 1782) which may be considered a semantically related antonym (e.g. to "bind / tie (a rope)" vs. "to cut (a rope)"). However, these semantics are considered too feeble to consider these forms as cognate, especially considering that a conservative approach was adopted concerning cognate decisions. Ultimately, it was considered that "to bind" had two roots in Proto-Western Kho-Bwa, *hak ([h a k]) and *zruŋ ([zr u ŋ]), whereby Khispi, Duhumbi and Khoina reflect the latter root, and the other varieties reflect the former root. The exact semantic distinction between the two roots for “to bind” are unclear. A second reason why sometimes there was no match between the predicted and the attested form was due to lexical replacement through borrowing. In the case of the concept “pumpkin”, the prediction for Shergaon was [m a + pʰl u ŋ] (ID 2876), but the attested form was [br u m + ɐ a] (ID 2877), which is a direct loan from Tshangla *brumɐa* "pumpkin". A final reason for mismatches are clear lexical innovations, such as the verb "to flow", which was predicted for Jerigaon as [h ɔ:] (ID 326), but where the attested form was [kʰ ɔ: + a ŋ] (ID 327), which is a noun-verb compound of [kʰ ɔ:] "water" and [a ŋ] "to go".

Table 6. Performance of the expert predictions.

Variety	Words	Morphemes	Perfect	Proportion	Score
Duhumbi	11	14	10	0.7143	0.869
Jerigaon	62	83	51	0.6145	0.7992
Khispi	26	33	19	0.5758	0.7828
Khoina	38	48	20	0.4167	0.6875
Khoitam	39	54	28	0.5185	0.7685
Rahung	45	53	29	0.5472	0.7453
Rupa	25	33	15	0.4545	0.6616
Shergaon	81	99	49	0.4949	0.734
TOTAL	327	417	221	0.53	0.756

Having identified those items where our prediction can be verified directly, we can proceed to calculate how well these predictions conform to the attested forms. The results for the expert predictions are given in Table 6. Here, the 327 verifiable predicted word forms correspond to a total of 417 verifiable morphemes. 221 of these (or 53%) were perfectly predicted. While this

may seem a bit low, from our detailed evaluation scores based on the segment-wise count of correctly and incorrectly predicted sounds per morpheme we can see that the predictions were correct in 76% of all cases.

In order to add more context to these results, it is useful to compare them with those predictions which we retrieved by the strictly automated procedure. These are shown in Table 7. As can be seen from this table, the automated procedure yielded fewer morphemes than the expert predictions, which emphasizes the importance of detailed background knowledge on a languages' morphology and lexical structures, which were not accessible to the automated approach. When comparing the quality of the individual predictions, we can also see rather drastic differences, both in the proportion of perfectly predicted morphemes (45% in the automated approach vs. 53% in the computer-assisted approach) and the more detailed accuracy scores (69% to 71% vs. 76% of overall accuracy with respect to predicted sounds per morpheme).

Table 7. Performance of the automated predictions. Scores F2 and F3 reflect the scores for the fuzzy prediction that allowed to predict 2 (F2) and 3 (F3) sound candidates per sound segment.

Variety	Words	Morphemes	Perfect	Proportion	Score	Score F2	Score F3
Duhumbi	11	13	6	0.4615	0.6923	0.6923	0.7179
Jerigaon	62	73	34	0.4658	0.6963	0.7169	0.7192
Khispi	26	31	13	0.4194	0.7097	0.7151	0.7151
Khoina	38	45	16	0.3556	0.6593	0.6667	0.6728
Khoitam	39	47	23	0.4894	0.734	0.7447	0.7482
Rahung	45	48	24	0.5	0.7153	0.7292	0.7292
Rupa	25	31	13	0.4194	0.6505	0.6559	0.6649
Shergaon	81	91	40	0.4396	0.6923	0.7051	0.7088
TOTAL	327	379	169	0.4459	0.6937	0.7032	0.7095

The concrete reasons for the failure or success of individual predictions for individual language varieties are difficult to assess. We assume that prediction quality should depend on different factors, such as (1) the amount of data that was already present at the time when we conducted the computer-assisted prediction experiment; (2) the expert knowledge for individual language varieties that would have helped our expert in the correction of the computed predictions; and (3) the number of informants consulted.

If the amount of initial data would have influenced the result of the prediction experiment, we would expect the initially more data-deficient varieties, Shergaon and Jerigaon, to have less accurate predictions than the other varieties. This would especially hold in case of the automated predictions, which were purely based on the sound correspondences derived from this initial dataset. However, both from the analysis of the automatic and from analysis of the expert's performance it becomes clear that the accuracy of the Shergaon and Jerigaon predictions was not lower than that of Rupa and Khoina and only marginally lower than that of Rahung. On the other hand, the accuracy of the automatic predictions of Duhumbi, which was by far the most completely covered variety in the initial dataset, does not outperform the accuracy of the

automatic predictions of five of the seven other varieties, including Shergaon and Jerigaon. Hence, it seems that the level of initial coverage of concepts in the database seems to have had little or no direct impact on the accuracy of the predictions that were based on it.

We do, however, observe that the expert prediction outperforms the pure computational ones in all varieties. This is not surprising, given the additional knowledge that experts have at their disposal. Since our expert worked actively on Duhumbi, it was expected that the prediction results would be higher for this variety than for the other varieties in the sample. More surprising is the high percentage of accurate expert predictions for Jerigaon, which is not only a variety that the expert does not know well, but also had the lowest percentage of positive responses to the elicitation. Since Khoina is the least well-described variety as well as the most aberrant variety phonologically, perhaps as a result of contact language influence, the fact that the predictions for this variety are least correct was expected. On the other hand, the low accuracy for both the expert and the automated predictions for Rupa was unexpected: This is hypothesized to reflect a high level of intergenerational variability and ongoing linguistic change in Rupa, the most modernised and exposed Western Kho-Bwa speech community. Nonetheless, we can carefully conclude from these results that expert knowledge has a definite impact on prediction quality.

During the elicitation sessions, it was noted that in cases where, in addition to the main informant, other speakers were also present (either permanently or occasionally), more concepts could be successfully elicited. Similarly, informants who decided to consult other speakers, either in person or through phone or social media, would achieve a higher coverage of concepts. We are not sure to what extent multiple inputs also improved the accuracy of the prediction: It could be surmised that more attestations would level out individual speaker's speech characteristics. At least it improved the number of predictions that could be successfully elicited.

A final result was our observation that the automated prediction improves when allowing for more uncertainty, as can be seen when comparing the results for the automated prediction which did not allow for fuzzy sound proposals (69%) vs. those predictions allowing for up to two sound candidates (70%) and up to three candidates (71%). Although the increase is not huge, the accuracy scores of the predictions increase slightly for all varieties. This is not a surprising result: Introducing more optional phonemes means that the chance for a correct prediction increases. However, even with the highest number of options, the accuracy of the automated predictions never outdid the expert-adjusted predictions. We expect that the expert score would also be slightly higher, if our expert had been allowed to include uncertainty, reflected in multiple solutions for individual predictions.

Our elicitation sessions generated a large number of observations regarding the elicitation and evaluation process itself that cannot all be addressed here. In the next section, we will therefore concentrate on a couple of selected points that are most interesting with respect to the general task of reflex prediction in historical linguistics.

3.2 Specific results

Our elicitation sessions and the subsequent analysis revealed several phonetic discrepancies between the predicted and the attested forms. Why, in many cases, did the prediction not exactly predict the form that was attested? We identified four main reasons for these discrepancies: 1. the specific word structure in the Western Kho-Bwa languages; 2. elicited concepts that turned out to be loans; 3. adjustments made to the phonetic transcriptions in individual varieties; and 4. previously unacknowledged sound correspondences. The vast majority of phonetic discrepancies can be explained through these reasons, and we discuss each of these reasons in more detail.

A first reason for discrepancies between the predicted morphemes and the attested morphemes is related to the word structure in the Western Kho-Bwa languages. As explained before, most Western Kho-Bwa parts of speech, such adjectives, adverbs, and demonstratives, are characterised by prefixes that identify parts of speech as well as lexico-semantic categories in nouns. The phonetic form of these prefixes in individual varieties is, in fact, by and large regular and almost entirely predictable based on phonotactic conditions. For example, vowels in prefixes may harmonise with vowels in the roots they modify; onsets of prefixes may harmonise in voicing or aspiration with the onsets of the roots they modify; and epenthetic nasal codas may be added to prefixes harmonising in point of articulation with the onset of the root. However, such intricate, variety-specific conditioning factors were not modelled in the semi-automatic method, and not perfectly understood by the expert at the time of making the predictions. For example, the predictions for the Sartang and Sherdukpen varieties commonly have prefixes with a phonetically reduced vowel (i.e. a schwa *ə*). Therefore, a prediction like "Bugun" in Jerigaon was predicted as [s ə + l u ŋ] (ID 182), but the attested form was [s u + l u ŋ] (ID 183) with vowel harmony between the vowel in the prefix and the vowel in the root.

In a few cases, the attested forms did not match the predicted forms because, contrary to expectation, the elicited concepts turned out to be loans. For example, the predicted form for Rupa "story", based on the available evidence from Khispi, Duhumbi, Khoitam and Rahung, was [kʰ a n + t a ŋ]. But the actually attested form was [kʰ a r + t a m]. The unexpected rhymes can be explained because this form is a direct loan from Tibetan *mkhar-tam* "story of the mansion", through the regionally popular Tshangla riddles also called *kʰartam*.

A third reason for discrepancies can be found in adjustments that were made to the transcriptions during the course of the prediction experiment. For example, based on the original dataset, the computer algorithm predicted Shergaon "to defeat" as [pʰ ɔ̃ ŋ] (ID 2587), which was changed by the expert to [pʰ ɔ̃ŋ] (ID 2588) because in Shergaon, long vowels only occur in open syllables and not in closed syllables, something not 'realised' by the computer algorithm, as phonotactic conditioning factors were not modelled. However, the actually attested form was [pʰ ɔ̃] (ID 2589 'form'), with a long nasalised vowel in open syllable. Although the attested form does not correspond exactly to the predicted form, this discrepancy should not be seen as an incorrect prediction. Rather, this is due to sub-phonemic idiolectal variation whereby some speakers still realise the nasal coda in addition to a nasalised vowel, and other speakers only realise a long nasalised open vowel. This is not 'irregular phonological change': It merely reflects the often-observed and in Shergaon currently on-going change from closed syllables with rhymes with a nasal coda to open syllables with nasalised vowel rhymes. However, to stay true to the nature of our prediction experiment, where the algorithm cannot be expected to factors such as idiolectal variation and phonotactic conditioning into consideration and the evaluation was fully automated, we transcribed the attested form as [pʰ ɔ̃ ŋ] (ID 2589), favouring the original predicted form by the algorithm over the expert's adjusted form.

During the collection and the subsequent analysis of the prediction experiment, the expert made minor adjustments to the phonological inventories of the Western Kho-Bwa varieties based on new insights uncovered through the additional lexemes that were elicited: For example, the transcription of Jerigaon nasalised vowel [ɐ̃] was changed to [ɑ̃]. Whereas in the original draft these adjustments were incorporated in the transcription of the attested forms, they were not included in the second evaluation of our prediction experiment in order to maintain consistency and comparability with the predicted and online registered forms.

A final reason for the discrepancies is at the same time one of the great benefits of the method. Through our analysis of the predictions, we were able to reveal new sound

correspondences that had missed our attention earlier. In some cases, the attested forms for certain concepts in the original dataset were insufficient to find a specific sound correspondence among all or most of the varieties. In other cases, a marginal sound correspondence was identified that had too few attestations (typically in less than three cognate sets) to be considered a solid sound correspondence. Our prediction experiment provided additional evidence of such sound correspondences that elevated them from the level of 'unknown' or 'marginal' correspondences to solid sound correspondences based on sufficient evidence.

Table 8. Comparing correspondence patterns for predicted initials for "to release" based on initials in "to meet" (predicted forms in cells shaded in grey) with revised correspondence patterns for attested initials. Attested forms in brackets are not cognate with the other forms and therefore excluded from the pattern. Digits following each form correspond to the ID in our table of predicted and attested items (see Appendix A3) and in the case of non-predicted forms, the ID used in our original dataset.

Variety	**Correspondence Pattern for Predicted Initials**				**Correspondence Pattern for Finals**		**Correspondence Pattern with Attested Initials**			
	"to release"		**"to meet"**		**"load"**		**"to release"**		**"to fly"**	
Khispi	ɕ ɔ ŋ	3105	ɕ u	2473	j ɔ ŋ	2337	ɕ ɔ ŋ	3105	ɕ ɛ l	1561
Duhumbi	ɕ ɔ ŋ	3106	ɕ u	2474	j ɔ ŋ	2338	ɕ ɔ ŋ	3106	ɕ ɛ r	1562
Khoina	ʂ u ŋ	1280	ʂ y j	2475	j u ŋ	2339	(tsʰ y:)	1281	tsʰ ɛ n	1563
Jerigaon	s u ŋ	548	s y:	423	j u ŋ	405	tɕʰ ɔ ŋ	549	tɕʰ ɛ n	1564
Khoitam	s u ŋ	1652	s y:	2477	j u ŋ	2341	tɕʰ ɔ ŋ	1653	tɕʰ a n	1564
Rahung	s u ŋ	2006	s y:	1923	j u ŋ	2342	tɕʰ ɔ ŋ	2007	tɕʰ ɛ m	1566
Rupa	s u ŋ	2360	s y:	2479	j u ŋ	2343	tsʰ ɔ ŋ	2361	tɕʰ a n	1567
Shergaon	s u ŋ	2888	s i:	2781	j u ŋ	2757	(pr ɔ:)	2889	tɕʰ a n	1568
PWKB	*sʲoŋ		*sʲu		*joŋ		*bʲoŋ		*bʲar	

There is one particular sound correspondence in the Western Kho-Bwa languages that had not been proposed when the predictions were set up. This is the correspondence between Khispi and Duhumbi fricative onset *ɕ-* and the Sartang and Sherdukpen affricate onsets *tsʰ-* ~ *tɕʰ-*. An example of this correspondence from the dataset is presented in Table 8. Because the sound correspondence had not yet been identified at the time of making the predictions, it is clear that the predictions were assigned to the most likely available sound correspondence, namely Khispi and Duhumbi *ɕ-*, Khoina *ʂ-*, other Sartang and Sherdukpen *s-*. The verb "to release" occurs as [ɕ ɔ ŋ] in the noun-verb compound "to quarrel" in Khispi and Duhumbi, however, it was elicited in other noun-verb compounds (such as "to drive a car" or "to shoot a bullet from a gun") in other varieties. Based on this Duhumbi and Khispi form [ɕ ɔ ŋ] (ID 305 and 306), predictions were made as given in Table 8. No manual adjustment was made to these predictions: As examples "to meet" and "load" show, both the predicted onset and the predicted rhyme are regular. However, the attested forms were slightly different from the predicted forms. With the exception of the Shergaon and Khoina reflexes, all others are considered as cognate, based on a sound

correspondence of initials also reflected in forms such as "to fly" and proposed to derive from a palatalised onset *bʲ-. This palatalised onset is also thought to condition the irregular rhyme reflexes in the Sartang and Sherdukpen varieties (*-ɔŋ* [ɔ ŋ] not *-uŋ* [u ŋ] as exemplified by "load"). The expected form for Shergaon is [tɕʰ ɔ ŋ], for Khoina [tsʰ ɔ ŋ]: Whereas the Shergaon form is probably a loan, the unexpected rhyme in the Khoina form could not yet been explained, and the form may not be cognate with the other Western Kho-Bwa forms. Rupa has variation among younger and older speakers between realisation of the affricate onset: /tɕʰ/ for younger speakers and /tsʰ/ for older speakers, hence older speakers will realise "to fly" as [tsʰ a n].

Table 9. Finding new sound correspondences and positing new proto-phonemes through predictions. The pattern for *qr-.

Variety	**Correspondence Pattern for Predicted Initials**				**Corr. Pattern for Finals**		**Corr. Pattern for Attested Initials**			
	"alive, healthy"		**"new"**		**"you"**		**"alive, healthy"**		**"red"**	
Khispi	ɔ + h a ŋ	97	ɔ + h a n	2721	n a ŋ	25	ɔ + h a ŋ	97	ɔ + h ɛ k	3177
Duhumbi	u + kʰ a ŋ	98	ɔ + kʰ ɔ n	2722	n a ŋ	26	u + kʰ a ŋ	98	ɔ + kʰʲ ɛ k	3178
Khoina	a + f a ŋ	1004	a + f ɛ n	2723	n a ŋ	27	a + x a ŋ	1005	a + x aj k	3179
Jerigaon	a + h a ŋ	122	ə + h ɛ n	2724	n a ŋ	28	a + h a ŋ	123	ə + h ɛ k	3180
Khoitam	a + h a ŋ	101	a + f a n	2725	n a ŋ	29	a + h a ŋ	101	ə + h ɛ k	3181
Rahung	a + h a ŋ	102	a + h ɛ n	2726	n a ŋ	30	a + h a ŋ	102	ə + h ɛ k	3182
Rupa	a + h a ŋ + b a	103	a + f a n	2727	n a ŋ	31	a + h a ŋ + b a	103	ə + h ɛ k	3183
Shergaon	a + h a ŋ	104	u + f a n	2728	n a ŋ	32	a + h a ŋ	104	ə + h ɛ k	3184
PWKB	*a-qʰeŋ		*a-qʰen		*naŋ		*a-qraŋ		*a-qrek	

A second example is the Khoina and Jerigaon prediction for "alive, healthy", also "strong" and the verb "to be healthy", in Table 9. The algorithm and researcher made the prediction for Khoina and Jerigaon based on the Rahung, Khispi and Duhumbi evidence and the correspondence set with as examples the rare cognate set "new" for the onset, disregarding the Khoitam, Rupa and Shergaon evidence, and the cognate set "you (thou)" for the rhyme. It was primarily the attested value for Khoina that pointed to another, extremely rare sound correspondence, namely the one also represented by the cognate set "red". These forms surface in Duhumbi with an aspirated uvular stop onset as an allophone of the aspirated velar stop onset in intervocalic position, i.e. *ukhang* [u + kʰ a ŋ], also realised as [u + qʰ a ŋ] "healthy, strong" and *okhek* [ɔ + kʰʲ ɛ k], also realised as [ɔ + qʰʲ ɛ k] "red". If we consider the conservative realisations [u + qʰ a ŋ] as the regular form for "healthy, strong" and [ɔ + qʰʲ ɛ k] as the regular form for "red", it could be argued that they form (near-)minimal pairs with [kʰ a ŋ] "carry" and [kʰʲ ɛ k] "ice", respectively. Realisations [u + kʰ a ŋ] and [ɔ + kʰʲ ɛ k] are typical of younger, educated speakers that realise

the underlying uvulars as velars. Moreover, even among those speakers who realise the uvulars, they do not occur as onset in monomorphemic words, but only when preceded by a prefix, hence occurring intervocalically in "healthy, strong" and "red". Therefore, the Duhumbi uvular stops are considered as allophones of the velar stops and not as distinctive phonemes.

Although the reconstructed uvular onset in "new" also surfaces in Duhumbi with allophonic variation in intervocalic position as *okhon* [ɔ + kʰ ɔ n] or [ɔ + qʰ ɔ n] "new" contrasting with, for example, [kʰ ɔ n + z u m] "crown of the head", the Khoina onset [x] not [f] and Khoitam and Sherdukpen onsets [h] not [f] indicate that these reflexes derive from a distinct onset *qr-, not *qʰ-. The positioning of several uvular onsets and uvular onset clusters based on correspondence sets that did not fit in with any of the onsets or onset clusters that had already been identified is perhaps one of the most significant outcomes of the prediction experiment for the reconstruction of Proto-Western Kho-Bwa.

4. Benefits of reflex prediction

Conducting and evaluating our reflex prediction experiment has greatly benefited the historical-comparative reconstruction of the Western Kho-Bwa languages. In addition, we were able to understand how predictions could assist descriptive work by field linguists, and even our respondents themselves pointed to a benefit that word prediction had for them.

As historical linguists and field linguists, we make predictions all the time. However, we do not commonly keep a record of these predictions vis-à-vis the actual elicited data in our own field notes and work, nor do we communicate these predictions to the scientific community. Explicitly stating our predictions will enhance the rigour of our own research, forcing us to think about what we predict, and come to more structured predictions. When we register these predictions and publish them, it will enable other researchers to cross-check our data and results and also allow cross-checking of our data with other's data, greatly increasing the transparency of our research. This is of particular importance to relatively subjective research methods, such as cognate decisions. A different set of criteria for determining whether data are cognate may greatly influence the outcome of the research. Making predictions will enable us to replicate research with a different set of decision-making factors and see how this affects the conclusions we attach to it.

When an automated prediction in our experiment had several choices, or when there were gaps in the prediction, it was easier to make a manual prediction based on whatever information the automated prediction gave. Similarly, even when the final manual prediction may not have been exactly correct, it made it easier to elicit the actual form, because an approximate phonetic form combined with an approximate semantic content may result in the respondents coming up with a cognate form. What this means for the applicability of the methodology is that predictions can make both elicitation in the field and finding cognates in published work more effective and efficient. This is crucial in light of both increasing language death and endangerment as well as funding limitations. We simply may not have the time to elicit the vocabulary and the grammatical structure of a language or to browse through lexical lists. Linguistic information essential to, for example, language reclamation, language revival or historical reconstruction may need to be uncovered sooner, rather than later. In both elicitation and finding cognates in lexical lists, predictions made on the basis of better known varieties will enable us to ask or search for what we think we need to know in related but poorly described varieties. Having a prediction for what a form in a given linguistic variety may look like already reduces the options from the phonetic point of view, and having to search among, for example, words with a few onsets while

knowing the approximate semantics will greatly facilitate the search. Both the linguist relying on secondary data and browsing through lexical lists, and the respondents working with the field linguist are more likely to recognise or remember a lexeme with similar form and perhaps similar meaning.

Prediction experiments can serve as useful tools for teaching purposes. They can provide students in linguistics – both descriptive linguistics and comparative linguistics – with a practical, hand-on test of both the linguistic theory and their own knowledge and skills. For example, by letting students predict cognate forms in a language based on regular sound correspondences they have identified in a subset of data, and then letting them test their predictions, either in the existing data sets, or by actually eliciting new data from respondents. Prediction experiments, both in their regularity and their deviation from regularity, can show students the basic tenets of sound change and the importance of factors such as cognate decisions, complementary distributions, semantic change and innovations, and loans and borrowing.

Finally, we often consider the benefits of certain elicitation techniques, studies, or analyses for the scientific community: the researchers, the academics, the students. But one important benefit of predictions that came up during the elicitation sessions for this study was directly noted by the respondents. In many communities, especially younger speakers who have not lived in a rural setting in their own speech community for extended periods of time tend to forget a considerable part of the vocabulary of their language. Not only asking the concepts themselves but also asking the predicted form made them remember their own language. They were also encouraged to either ask speakers nearby or even use social media like WhatsApp to find out more about a concept, to try to find out whether a form like 'x' or 'y' with a certain meaning existed or not. This kind of a renewed interest in their own language is very important for the possible survival of endangered languages.

5. Conclusion

Predictions attest to the regularity of sound change and hence to the validity of the comparative method. Predicting missing values based on regular sound correspondences that follow from regular sound changes results in valid predictions, hence the sound changes that they are based on must be regular. If the automated predictions would be largely incorrect and the accuracy of the automated predictions would be low, this could mean that a substantial part of the sound correspondences that were largely automatically identified and manually adjusted were not correct and hence that there is no regularity in sound change, contradicting the basic tenet of the comparative method. Moreover, we identified several compelling reasons explaining the inaccurate morpheme predictions, that were not linked to the sound correspondences upon which these predictions were based.

Computer-assisted prediction, where the values predicted by the algorithm are manually adjusted before being elicited in the field, are more accurate than automated predictions in all but one variety of Western Kho-Bwa. Moreover, the difference between computed predictions and expert-refined predictions increases with the expert's knowledge about the languages in question. Therefore, manual cross-checking and adjustment of automated predictions further increases the accuracy. In addition, our automated procedure yielded fewer morphemes for comparison than the expert predictions. This also emphasizes the importance of detailed background knowledge on a languages' morphology and lexical structures and makes a strong case for a workflow that combines automated approaches with the linguistic expert's inputs.

It is our hope that reporting this research and making the procedure, the data, the analysis and the results openly available will encourage other linguists not only to more consciously employ predictions, but also register, report, and discuss their individual results.

Supplementary material

The supplementary material accompanying this paper contains the code, the data, and all instructions needed to replicate the analyses described here. It has been curated on GitHub (https://github.com/lingpy/prediction-study) and archived with Zenodo (https://zenodo.org/badge/latestdoi/298542681).

Funding

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 ERC Starting Grant 715618 "ComputerAssisted Language Comparison" (JML, https://digling.org/calc, 2017-2022).

Acknowledgements

We thank two anonymous reviewers for helpful and constructive comments. We thank Nathan W. Hill for his incredible help in developing and optimizing the workflow for computer-assisted language comparison. We also thank Guillaume Jacques for comments on a later draft of this study. 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).

References

Amery, Rob. 2016. *Warraparna Kaurna!* University of Adelaide Press.

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.

Bodt, Timotheus Adrianus. 2019. The Duhumbi perspective on Proto-Western Kho-Bwa rhymes. *Die Sprache* 52 (2016 / 2017) 2. 141–176.

Bodt, Timotheus Adrianus. Forthcoming. The Duhumbi perspective on Proto-Western Kho-Bwa onsets. *Historical Linguistics.*

Bodt, Timotheus A. & Johann-Mattis List. 2019. Testing the Predictive Strength of the Comparative Method: An Ongoing Experiment on Unattested Words in Western Kho- Bwa Languages. *Papers in Historical Phonology* 4 (1). 22–44.

Bodt, Timotheus A. & Johann-Mattis List. 2020. The multiple benefits of making predictions in linguistics. *Babel, The Language Magazine* 31: 8-12.

Bodt, Timotheus A., Nathan W. Hill and Johann-Mattis List. 2018. Prediction experiment for missing words in Kho-Bwa language data. *Open Science Framework Preregistrations* October 5. https://osf.io/evcbp/

Branner, David Prager. 2006. Some Composite Phonological Systems in Chinese. In David Prager Branner (ed.), *The Chinese Rime Tables. Linguistic Philosophy and Historical-Comparative Phonology*, 209–32. Amsterdam: Benjamins.

van Driem, George. 2001. *Languages of the Himalayas - An Ethnolinguistic Handbook of the Greater Himalayan Region. 2*. Leiden: Brill.

Eberhard, David M., Gary F. Simons & Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the World. Twenty-second edition. Dallas, Texas: SIL International. https://www.ethnologue.com

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 H. 1963. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Joseph H. Greenberg*, Universals of Human Language*, 73–113. Cambridge, Mass: MIT Press.

Grimm, Jacob. 1822. *Deutsche Grammatik.* Erster Theil. Göttingen: Dieterich.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2020. Glottolog. Version 4.2.1. Jena, Max Planck Institute for the Science of Human History. http://glottolog.org

Lieberherr, Ismael & Timotheus Adrianus Bodt. 2017. Sub-Grouping Kho-Bwa Based on Shared Core Vocabulary. *Himalayan Linguistics* 16 (2). 25–63.

List, Johann-Mattis. 2019. Automatic Inference of Sound Correspondence Patterns Across Multiple Languages. *Computational Linguistics* 1 (45). 137–61. https://doi.org/10.1162/coli_a_00344.

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.

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.

Nosek, Brian, Emorie D. Beck, Lorne Campell, Jessica K. Flake, Tom E. Hardwicke, David T. Mellor, Anna E. van 't Veer and Simine Vazire. 2019. Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences* 23(10): 815-818.

Post, Mark W. & Robbins Burling. 2016. The Tibeto-Burman languages of Northeastern India. In Graham Thurgood & Randy J. LaPolla (eds.), *The Sino-Tibetan Languages,* 213-233.

Abingdon: Routledge.
Schweikhard, N. and J.-M. List. 2020. Developing an annotation framework for word formation processes in comparative linguistics. *SKASE Journal of Theoretical Linguistics* 17.1. 2-26.
Sims-Williams, P. 2018. Mechanising historical phonology. *Transactions of the Philological Society*. 116.3. 555-573.
Watkins, C. 1962. *Indo-European origins of the Celtic verb. Volume I. The sigmatic aorist*. Dublin: Dublin Institute for Advanced Studies.
Wu, M.-S., N. Schweikhard, T. Bodt, N. Hill, and J.-M. List. 2020. Computer-Assisted Language Comparison. State of the Art. *Journal of Open Humanities Data* 6.2. 1-14.

Appendix

A1 Information on language varieties, informants, and data

The data collection on which the initial database and the analysis and predictions themselves are based relied on elicitation sessions with at least two speakers, one male and one female, from each of the eight Western Kho-Bwa varieties recorded between 2012 and 2017. This dataset was based on the 550-item regionally relevant word list in English and Hindi that can be found on http://doi.org/10.5281/zenodo.3608408.

The metadata of the speakers of this original dataset can be found on http://doi.org/10.5281/zenodo.1210131 (Sartang), http://doi.org/10.5281/zenodo.1213719 (Sherdukpen), http://doi.org/10.5281/zenodo.1406887 (Khispi) and http://doi.org/10.5281/zenodo.1291599 (Duhumbi). The sound files of the recordings of this original dataset will be made available once the complete reconstruction of Proto-Western Kho-Bwa is published.

For the "lifting" of the original data into a form where the dataset could be used for our prediction experiment, see A2.

The metadata of the speakers and the evaluation of the prediction can be found on http://doi.org/10.5281/zenodo.2632141. The sound files and their transcriptions can be found on http://doi.org/10.5281/zenodo.2529727. The predictions were verified during a one-month fieldwork period in October and November 2018 during which several other tasks had to be completed as well. Although care was taken to find knowledgeable people, the elicitation involved eight different linguistic varieties in eight different locations located at travel distances of one hour to a full day from each other, depending on transport availability. Not all the informants consulted earlier were available this time around. Local festivals, community meetings, agricultural occupations, family events, day labour, court cases – a multitude of reasons made finding the 'right' consultant within a limited period of time sometimes difficult or even impossible. Eliciting and triple-recording the predictions took one or two days per consultant. In the single case of Jerigaon, the local circumstances meant we had to rely on an informant who was perhaps not the most suitable candidate. Unfortunately, in linguistic fieldwork we can never create the 'perfect' conditions of laboratory settings. If more time had been available, we could have searched longer, or waited for other people to become available, as well as cross-check the data with a second or even third speaker. We acknowledge this lapse in our methodology, but also want to state that, from a purely scientific point of view, this has not affected our results. If anything, the number of possible cognate terms that a more prolific speaker would have identified, and hence the number of correct predictions, would have improved our final result: At least for Jerigaon, the figure is an understatement, and not an overstatement of the correctness of the predictions.

Multilingualism is extreme in some cases of older speakers of Western Kho-Bwa varieties, who, depending on the exact location and their personal history, may understand and often also speak Indo-Aryan languages such as Hindi, Assamese and Nepali, Bodish Tibeto-Burman languages such as Tawang Monpa, Brokpa and Tibetan, and other Tibeto-Burman languages such as Tshangla, Miji and Hruso Aka. In addition, older speakers often have a fair knowledge of one or more of the other varieties but will still be able to distinguish them from each other, and actually point out the differences themselves. On the other hand, younger speakers usually only speak their own language and Hindi. Beyond these age-related generalisations, the exact proficiency of speakers, and hence individual informants, in contact languages is difficult to

assess.

The influence of loans from contact languages on the outcome of the prediction experiment is negligible. If a consultant provided a loan from a contact language rather than the inherited form in the native variety, and that loan was identified as such, either because of self-admission by the informant, or because of the language expert's knowledge of several of the contact languages, this would be recorded as non-cognate with the predicted form, and hence the accuracy of the prediction was not considered. Although a native, inherited, cognate form may exist, it was not recorded. If it would have been recorded, and hence the prediction could be tested for accuracy, this could have resulted in either an increase or a decrease of the overall predictive capacity of the experiment, but because the prediction was not assessed in this experiment in the first place, the fact that the response was a loan did not influence the accuracy.

In some cases, the respondent / informant simply said: "I don't know", where they would understand the concept, but were unable to recall whether it had a name in their language, and if so, how, it was called. In other cases, the respondent said: "I don't understand what you mean", and even after trying to explain the concept, the respondent did not understand the concept. This was, for example, the case with animals like "pangolin", that are now so rare in the area that even after explaining or showing pictures the respondent did not know the species and its name. These non-verifiable predictions, where there was a zero response, are fundamentally distinct from responses that differed from our predictions. The vast majority (30) of such non-verifiable predictions was from Jerigaon, where, by her own admission, the young female respondent had spent considerable time in an urban setting and was therefore not as fluent in her language as would be required for this purpose. Although it was, sometimes successfully, attempted to recover the missing concepts, including through telephone calls or WhatsApp, neighbours, and passers-by or a second meeting the next day, this did not result in these 30 gaps being filled up.

A2 Preparing the data for automated treatment

Since software solutions that model certain aspects of the comparative method need to be based on the assumption of standards, in order to make sure they are applicable for a wide range of languages from a wide range of language families, any application of automated methods requires to 'normalize' or 'lift' the data beforehand.

Concrete steps of lifting address each aspect of the typical triples of language, form, and concept, in which most lexical datasets can be represented in a straightforward manner (Forkel et al. 2018). In order to make sure that our data is comparable with other datasets, we link all languages in our sample to Glottolog (Hammarström, Forkel & Haspelmath 2020). To make sure that the glosses eliciting the meanings of the data conform to current standards, they are all linked to the concept sets proposed by the Concepticon project (List et al. 2020, https://concepticon.clld.org, see List, Cysow & Forkel 2016 for an overview). Finally, to guarantee that the methods we use for the automated reflex prediction can be applied to the word forms in the data, we converted all phonetic transcriptions to the standards proposed by the Cross-Linguistic Transcription Systems initiative (CLTS, List et al. 2019b, https://clts.clld.org; see Anderson et al. 2018 for details).

The 'lifting' of the original data was carried out with the help of CLDFBench (Forkel and List 2020), a Python package which eases the conversion of lexical and structural datasets to the formats proposed by the Cross-Linguistic Data Formats (CLDF) initiative (Forkel et al. 2018, https://cldf.clld.org). CLDF requires that specific aspects of the data are consistently 'linked' to so-called reference catalogues, that is, meta-data collections that make it easier to compare a given resource with other resources. The full dataset was curated on GitHub (https://github.com/lexibank/bodtkhobwa/) and later archived with Zenodo (https://doi.org/10.5281/zenodo.3537604). Specific aspects that are interesting in this context are:

- The list of languages, also linked to Glottolog (with geocoordinates): https://github.com/lexibank/bodtkhobwa/blob/v2.0/etc/languages.csv,
- the extended concept list, as it was submitted to the Concepticon project: https://concepticon.clld.org/contributions/Bodt-2019-664, and
- the database file, containing the data we used to make the predictions: https://github.com/lexibank/bodtkhobwa/blob/v2.0/raw/bodt-khobwa-cleaned.tsv.

The code and data for the preparation of the predictions were again curated on GitHub (https://github.com/lingpy/predict-khobwa) and archived with Zenodo (https://zenodo.org/badge/latestdoi/146638428) and additionally submitted to the Open Science Framework in the form of a research registration (https://osf.io/evcbp/).

A3 Additional Notes to the Prediction Workflow

A3.1 Notes on Data and Code for the Prediction Workflow

While we presented the prediction workflow in due detail in Section 2.3 of this study, more technical aspects of the procedure have not been discussed in this context. These are, however, available from previous studies, which include first a short working paper which we wrote prior to conducting our prediction experiment (Bodt and List 2019) and which was also mentioned in the main manuscript, but also in the form of the code repository which we submitted as part of the registration process. Instead of repeating what we wrote before, we point to the most relevant resources which help those interested in the details to learn more about the approaches we carried out in concrete.

- The working paper can be found online (open access) at https://doi.org/10.2218/pihph.4.2019.3037.
- The code which we submitted for the registration procedure is available from GitHub: (https://github.com/lingpy/predict-khobwa (or the OSF: https://osf.io/evcbp/).
- Wu et al. (2020), furthermore (article available at https://doi.org/10.5334/johd.12) provide another very detailed discussion of the workflow that is used to infer correspondence patterns (also available open access).

A3.2 Manual refinement

In the sixth stage of the workflow, we selected suitable morphemes for the experiment. Here, we used the following four major criteria to reduce the list and to guarantee the feasibility of the experiment:

(a) The list of predictions contained morphemes that were already known not to be used in a given variety, for example, because this variety had uniformly lost certain prefixes or suffixes, or because a variety used a different prefix or suffix than the one predicted based on the evidence from the other varieties. For example, because the Sartang and Sherdukpen varieties have an *s*-prefix in the word for "pillow", the algorithm made a prediction for the s-prefix in Khispi and Duhumbi "pillow" as well, even though these two varieties do not have this prefix in their reflex of "pillow", but another prefix. The algorithm saw the missing reflex of the prefix as a gap in the data. The predicted form of the s-prefix for "pillow" was not adopted in the final list of concepts to be elicited for Khispi and Duhumbi, as the form that was already in the dataset was known to be 'complete'.

(b) The list of predictions also included predicted forms that were based on gaps in the original data that should not have been there, because the given variety uses a borrowed form or lexical innovation that was initially excluded from the database because it is not cognate with the form in the majority of varieties. For example, Khispi and Duhumbi have loans for concepts like "otter", "banana" and "oil" and a lexical innovation for a concept like "snake". These loans and lexical innovations had not been included in the original database and hence the algorithm came up with predictions for these concepts in Khispi and Duhumbi based on the forms in the other varieties and the established sound correspondences. These predictions were subsequently excluded from verification.

(c) In the case of concepts that derive from a polymorphemic root in the ancestor language, several morphemes were predicted, but not all these morphemes would be used in the reflex of the concept in a given single variety. Two examples are the concepts "sparrow" and "bat", which

have five, respectively four distinct morphemes across the eight varieties, occurring in sets of at least two, to a maximum of three morphemes in each individual variety. The algorithm made predictions for all distinct morphemes for each variety based on the forms in the other varieties and the established sound correspondences, considering some morphemes as missing, even though none of the varieties used all these morphemes and the concept was already present in the data for each variety. In general, such polymorphemic predictions were excluded from verification when they were already present for each variety.

(d) In some cases, the available evidence was insufficient for the algorithm to come to a prediction. This was mostly the case when there were too few attested forms in the data and the number of missing forms was too high, for example, when only one or two varieties had data for a concept. In other instances, the attested forms in the dataset were simply too diverse to derive a sensible prediction. And finally, in some cases, the sound correspondences that were derived by the algorithm and manually adjusted were inconclusive as to which phonemes would occur in a given position in a given morpheme. If there were sufficient grounds to presume that a form could actually be elicited in the field, this 'prediction' would still be included in the final list, but if not, it was ignored.

In the seventh stage of the experiment, predicted morphemes were combined to concepts that could actually be elicited and also manually modified where expert knowledge of the varieties could help to spot errors in the automatically created results. Knowing that inherited mammal names in the Western Kho-Bwa varieties almost invariable consist of a reflex of an *s*-prefix and the root, the prediction for the s-prefix morpheme 'meat3' and the prediction for the root morpheme 'PANGOLIN' was merged to the concept "pangolin" for the varieties in which there were data gaps (IDs 37, 475, 883, 2317, 2815). It would have been futile to simply elicit a morpheme 'meat3', and similarly, elicitation of the morpheme 'PANGOLIN' would most likely have included a prefix deriving from a morpheme 'meat3'. Sometimes, on the other hand, a respondent would provide a certain concept without the predicted prefix, because the prefix was lost in that particular variety. In that case, only the root, and not the missing prefix, was evaluated.

While initially none of the forms predicted by the algorithm were changed, manual refinements were made to these predictions after having identified the major lexemes, using the automated predictions as a baseline from which we derived expert predictions guided by the broader knowledge of the sound laws observed for the language family. These expert assessments would at times follow the algorithm completely, but at times, they would also diverge from it. For example, the algorithm predicted a form [m y:] for the morpheme 'PUBEHAIR' in the varieties Khoina, Jerigaon, Rahung, Khoitam and Rupa (IDs 1267, 529, 1993, 1633 and 2341), and [m u] in Shergaon (ID 2863) based on the Duhumbi form [m u r] (ID 4797 in the original database, ID 36 in the predictions database), Khispi form [m u l] (ID 4796 in the original database) and the regular sound correspondence between vowel /u/ in Khispi and Duhumbi, long vowel /y:/ in the Sartang varieties in Rupa and vowel /u/ in Sherdukpen. However, this correspondence only holds in open syllables. In closed syllables, Shergaon, Khispi and Duhumbi vowel /u/ corresponds regularly to short vowel /y/ in the Sartang varieties and Rupa; in addition, the final *-r* in the uncommon rhyme *-ur* in Duhumbi was thought to correspond to final *-ŋ* in the other varieties. Hence, the predictions for the concept "pubic hair" were adjusted to [m y ŋ] in the Sartang varieties and Rupa (IDs 1268, 530, 1994, 1634, 2341) and [m u ŋ] in Shergaon (ID 2864) before elicitation in the field.[10]

10 Note that the actually attested reflexes of the concept "pubic hair" resulted in the distinct sound correspondence between Khispi -ul [m u l], Duhumbi -ur [m u r] (ID 36), Khoitam -iŋ [a + m i ŋ] (ID 1635) and Shergaon -in [a + m

In addition, in those cases where the algorithm could somehow not make a prediction, background knowledge of the phonology and morphology of the Western Kho-Bwa varieties was used to make a manual prediction. For example, the morpheme "year" had an automatic prediction for Jerigaon and Shergaon with a missing coda, [tʰ a Ø] (IDs 763 and 3109). Based on the knowledge that a rhyme *-ar* in Khispi and Duhumbi corresponds to rhyme *-am* in Jerigaon and Shergaon, the morpheme was predicted to be [tʰ a m]. In addition, a suffixed morpheme *-pu* in Khispi and Duhumbi, included in the concept [tʰ a r + p u] "this year" (IDs 4769 and 4770 in the original database) and also attested as morpheme *-bu* in the cognate forms in Khoina, Khoitam, Rahung and Rupa (IDs 4771, 4773, 4774, 4775 in the original database), is known to correspond to a morpheme *-bu* in Jerigaon and Shergaon. The combination resulted in the predicted guesstimate [tʰ a m + b u] for the concept "this year" for both Jerigaon and Shergaon (IDs 764 and 3110).

i n] (ID 2865), regularly thought to derive from a proto-rhyme *-ur, i.e. *a.mur. This testifies both to the usefulness of prediction for finding new sound correspondences, as well as to the need to distinguish a morpheme structure of onsets and rhymes rather than of initials, nuclei and coda.

A4 Evaluation Procedure

For the evaluation procedure, we wrote original Python code which would help us to evaluate the findings properly (the code itself made use of LingPy, List et al. 2019a, http://lingpy.org and LingRex, List 2018, https://github.com/lingpy/lingrex, to handle sound correspondences). In order to compare the predicted with the attested forms, we made use of the EDICTOR tool (List 2017, https://digling.org/edictor) for the creation and manipulation of etymological dictionaries. Here, all data were manually annotated, but the data themselves were created with the help of the code we wrote. This code is curated on GitHub (https://github.com/lingpy/prediction-study) and was submitted as a Zip folder to the Open Science Framework for the purpose of peer review, where it can be accessed at https://osf.io/dv8mh/?view_only=ab33edcb080540c1aadf2e123a1aaed1. After having been notified of the acceptance of the study, we also archived the data with Zenodo (https://zenodo.org/badge/latestdoi/298542681). The repository contains further detailed instructions on how the code can be run and how the data was annotated.

A5 Additional References Cited in Appendix

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* 4 (1). 21–53.

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

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. 2018. LingRex: Linguistic Reconstruction with LingPy. Version 0.1.1. Max Planck Institute for the Science of Human History: Jena. https://doi.org/10.5281/zenodo.1544944.

List, J.-M., S. Greenhill, T. Tresoldi, and R. Forkel. 2019a. LingPy. A Python library for quantitative tasks in historical linguistics. Version 2.6.5. Max Planck Institute for the Science of Human History: Jena. http://lingpy.org.

List, J.-M., C. Anderson, T. Tresoldi, C. Rzymski, S. Greenhill, and R. Forkel. 2019b. Cross-Linguistic Transcription Systems. Version 1.3.0. Max Planck Institute for the Science of Human History: Jena. https://clts.clld.org.

List, J., C. Rzymski, S. Greenhill, N. Schweikhard, K. Pianykh, A. Tjuka, M. Wu, and R. Forkel. 2020. Concepticon. A resource for the linking of concept lists (Version 2.3.0). Version 2.3.0. Max Planck Institute for the Science of Human History: Jena. https://concepticon.clld.org/.

Résumé

Durant l'analyse d'un ensemble de données lexicales du Kho-Bwa occidental (sino-tibétain/trans-himalayan) au moyen d'une approche assistée par ordinateur de la comparaison historique des langues, nous avons observé des lacunes dans les données, où une ou plusieurs variétés ne disposaient pas d'une forme attestée pour un certain concept. Nous avons appliqué un nouveau flux de travail dans lequel nous avons combiné les étapes manuelles traditionnelles avec des approches automatisées pour prédire la forme phonétique la plus probable des mots manquants dans notre ensemble de données (utilisant l'information des correspondances régulières). Le résultat de ce flux de travail était une liste des mots réelement vérificables, que nous avons ensuite pré-enregistré comme expérience pour comparer la liste avec les réflexes découverts ultérieurement lors de travaux de terrain. Dans cette étude, nous décrivons notre processus de travail pour la prédiction des mots hypothétiques et le processus d'élicitation lors de travaux de terrain, et présentons ensuite les résultats de notre expérience pour la prédiction de réflexes. Sur la base de l'expérience que nous avons faite au cours de cette expérience, nous identifions quatre avantages généraux de la prédiction des mots dans la comparaison linguistique historique. Ce genre de travail peut (1) renforcer la transparence de la recherche linguistique ; 2) accroître l'efficacité des méthodes de recherche linguistique historique, tant sur le terrain qu'à partir de sources secondaires ; 3) fournir aux enseignants et aux apprenants des exemples pratiques d'une large pléthore de phénomènes linguistiques, y compris la régularité du changement des sons ; et 4) susciter l'intérêt et l'engagement des locuteurs pour leur propre patrimoine linguistique.

Zusammenfassung

Bei der Analyse lexikalischer Daten von westlichen Kho-Bwa-Sprachen aus der sinotibetischen oder transhimalayanischen Sprachfamilie mit Hilfe eines computergestützten Ansatzes zum historischen Sprachvergleich stießen wir auf Lücken in den Daten, in denen eine oder mehrere Varietäten keine attestierte Form für bestimmte Konzepte hatten. Wir verwendeten daraufhin einen neuen Workflow, in dem wir manuelle mit automatisierten Arbeitsschritten kombinierten, um die wahrscheinlichsten phonetischen Realisierungen der fehlenden Formen in unseren Daten vorherzusagen, wobei systematisch auf die Information von Lautkorrespondenzen mit möglicherweise kognaten Wörtern zurückgriffen wurde. Dieses Verfahren lieferte uns eine Liste hypothetischer Reflexe von zuvor als kognat identifizierten Wörtern, die wir als Experiment zur Vorhersage bisher nicht observierter Wörter zunächst präregistrierten, um sie dann im Rahmen einer erweiterten Feldforschung mit den tatsächlich attestierten Wortformen zu vergleichen. In dieser Studie beschreiben wir zunächst den Workflow, mit dem hypothetische Reflexe vorhergesagt werden können, sowie den Prozess der Elizitierung von aktuellen Wortformen im Rahmen der Feldforschung, und präsentieren dann die Ergebnisse unseres Experiments zur Reflexvorhersage. Basierend auf der Erfahrung, die wir mit diesem Experiment gemacht haben, identifizieren wir dann grundlegende Vorteile, welche die aktive Vorhersage unbekannter Wortformen für den historischen Sprachvergleich bietet. Zu diesen Vorteilen gehören insbesondere (1) die erhöhte Transparenz der linguistischen Forschung, (2) die erhöhte Effizienz von Feldforschung und Quellenarbeit, (3) der edukative Aspekt, der Lehrenden wie Lernenden eine Vielzahl von Beispielen für linguistische Phänomene, wie zum Beispiel die Regelmäßigkeit des Lautwandels, liefert, und (4) die Möglichkeit, das Interesse von Sprecherinnen und Sprechern zu wecken, sich mit ihrem linguistischen Erbe aktiv auseinanderzusetzen.

Address for correspondence

Timotheus A. Bodt
Postdoctoral researcher
Department of East Asian Languages and Cultures
School of Oriental and African Studies
University of London
10 Thornhaugh Street, Russell Square
London WC1H 0XG, United Kingdom
tb47@soas.ac.uk

Johann-Mattis List
Faculty member
Linguistic and Cultural Evolution Department
Max Planck Institute for the Science of Human History
Kahlaische Str. 10
07745 Jena, Germany
list@shh.mpg.de