Preregistered with the Open Science Framework: https://osf.io/evcbp/

Cite as: Bodt, Hill and List (2018): Prediction experiment for missing words in Kho-Bwa language data. Open Science Framework Preregistration. October, 5, https://osf.io/evcbp/.

Prediction experiment for missing words in Western Kho-Bwa language data

Timotheus A. Bodt, Nathan W. Hill, and Johann-Mattis List
October, 2018

Introduction

This is an experiment on the prediction of words that have thus far not been recorded in field work. The basic idea is that, given an existing dataset of cognate words, we can use the correspondence pattern detection algorithm by List (2018) to infer (or "predict", or "retrodict") those words in any given doculect that do not yet have a recorded reflex in a given cognate set. During his initial field work on the Western Kho-Bwa languages, Tim Bodt was not able to record a value for every concept for each of the eight varieties yet. Hence, we can now use the existing dataset to predict how potential cognates would have sounded. When Tim Bodt goes back to the field in November, he will try to elicit those missing words. This will give an idea as to how well the prediction algorithm works, but also how well prediction works in historical linguistics in general.

Installing and running the code

To install the source code of this package, you need lingpy, sinopy, and lingrex, three Python libraries. You can install them with help of PIP, provided you have a fresh Python3 installation on your computer:

```
$ pip install -r pip-requirements.txt
```

To run the code, simply type:

```
$ python predict.py
```

The output will be written to the file `predictions.tsv`.

Short introduction to the Western Kho-Bwa languages

In 1952, Stonor, basing himself on local sources, reported that the two languages 'Puroik' and 'Bugun' are mutually intelligible (Stonor 1952). It was not until the last two decades of the previous century that the first linguistic materials on Bugun/Khowa, Puroik/Sulung, Sherdukpen

and Sartang/Boot/Butpa Monpa became available: the works of the Indian research/language officers Deuri (Deuri 1983), Dondrup (Dondrup 1988), Dondrup (Dondrup 1990), and Dondrup (Dondrup 2004). On the Chinese side, the first Puroik data were published as part of the large-scale survey Tibeto-Burman Phonology and Lexicon (Sūn 1991). Based on these materials and own data, Jackson Sun (Sun 1992, 1993) was the first to suggest that Puroik, Bugun, Sherdukpen and 'Lishpa-Butpa' are not just a random residue when all other major languages are subtracted, but that they might belong together and form a coherent linguistic group.

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

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

Note on the data used for this study

The data taht form the basis of this analysis were elicited through a 441-gloss lexical wordlist, collected by Tim Bodt between 2012 and 2017. Before re-arranging the data with help of computer-assisted tools like EDICTOR (List 2017) and LingPy (List and Moran 2013), the data was already arranged and partly analyzed with the goal of building an etymological dictionary of Western Kho-Bwa. For most of the data points, audio recordings are also available, as well as extensive materials collected during Bodt's fieldwork (see the overview on Zenodo). In the

current form, the data is arranged in the format required by LingPy and EDICTOR, namely, a tab-separated CSV-file (see `data/bodt-khobwa-cleaned.tsv`) which contains cross-semantic, partial cognate sets that are further annotated with help of alignments. Each column in the file format contains data of a specific type. For details on this format, we recommend to consult List, Walworth, et al. (2018). Notable differences to previous analyses are:

- the column `TOKENS` shows morphological and sound segmentation, with sound segments being segmented by a space, and morphemes being further segmented by a + symbol,
- the column `crossids` lists cross-semantic cognate sets, thus, one cognate set may span multiple concepts, as is the case for most of the prefixes in the data,
- the column `structure` lists the prosodic structure of the entries in the data, represented by abbreviations i (initial), n (nucleus), and c (coda),
- the column `morphemes` follows the idea presented in Hill and List (2017), to annotate morphological structures with help of glosses, but it expands the format by making a strict distinction between free and bound morphemes (free morphemes are written in uppercase, and bound morphemes are prefixed by a _ underscore)

For the time being, the concepts in the data are not yet linked to the Concepticon, but we plan to do this when publishing the results of this study. By then, we will publish the underlying data officially in format proposed by the CLDF initiative (Forkel et al., forthcoming), to offer simplified re-use of the data in other projects, and to increase its comparability.

Basic statistics of the data

In its current form, the data consists of 8 language varieties and 662 elicitation glosses. The following table lists the basic coverage of the data, by comparing for how many of the 662 eliciation glosses a translation equivalent could be found in the current data.

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

As we can see from the table, the basic is rather low, with an average of 78% of the elicitation glosses being covered. This justifies our prediction experiment, as it means that there are quite a few concepts that Bodt could try to elicit during fieldwork in order to find out how well our automatic prediction method works.

How the predictions are presented

The predictions are given in tabular form in the file `predictions.tsv`. The table header indicates the content of the cells in each column, and each row presents one prediction. The predictions, however are not given for entire words, but only for specific *morphemes*, given the agglutinative, multi-morphemic character of these languages in which prefixation, suffixation and word derivation patterns are rather frequent. To allow for a precise identification of the semantics of a given morpheme, all morphemes in the data were annotated using the EDICTOR's (List 2017) morpheme annotation functionalities. Predictions are further given in three differen forms, reflecting different degrees of uncertainty. The first form provides the most likely sounds for the word, ignoring any uncertainty. The second word form provides up to two different sounds per alignment site, if there are multiple possibilities. The third word form provides up to three different sounds per site. In all cases where multiple sounds are provided, this is displayed by separating the sounds in the same slot by the pipe symbol `|`. If multiple possibilities are encountered, the order of the sounds reflects the number of occurrences (following the rough logic that the more frequently observed correspondence patterns should be preferred in prediction over less frequently observed ones).

No.	Qu.	Cogn.	Language	Concept	Morpheme	Word1	Word2	Word3
577	?	320	Khoina	dig	DIG	Ø ɔ k	Ø ɔ|ɔː k	Ø ɔ|ɔː k
578	?	320	Khoitam	dig	DIG	Ø ɔ k	Ø ɔ k	Ø ɔ k
579	?	320	Shergaon	dig	DIG	Ø ɔ k	Ø ɔ k|t	Ø ɔ k|t
580		321	Duhumbi	dirty	DIRTY	ʑ ɛ t	ʑ|ɕ ɛ|ɛˀ t	ʑ|ɕ|z ɛ|ɛˀ t
581		321	Khispi	dirty	DIRTY	ʑ ɛ t	ʑ|z ɛ|ɛː t	ʑ|z|ɕ ɛ|ɛː t
582		322	Duhumbi	soft / smooth	SMOOTH	d ɔ ŋ	d|t ɔ|a ŋ|m	d|t ɔ|a|u ŋ|m|-
583		322	Khispi	soft / smooth	SMOOTH	d ɔ ŋ	d|t ɔ|a ŋ|m	d|t ɔ|a|u ŋ|m|l
584	?	324	Duhumbi	disease	DISEASE	Ø ɔ ŋ	Ø ɔ|a ŋ|m	Ø ɔ|a|u ŋ|m|-
585	?	324	Khispi	disease	DISEASE	Ø ɔ ŋ	Ø ɔ|a ŋ|m	Ø ɔ|a|u ŋ|m|l
586	?	324	Shergaon	disease	DISEASE	Ø u ŋ	Ø u ŋ|(ŋ)	Ø u ŋ|(ŋ)
587		326	Duhumbi	pull	PULL2	g i	g|h i|ɛj	g|h|kʰ i|ɛj|u
588		326	Khispi	pull	PULL2	g i	g|s i|uj	g|s|kʰ i|uj|u
589		326	Rupa	pull	PULL2	g iː	g iː|yː	g iː|yː
590		326	Shergaon	pull	PULL2	g iː	g|k iː|i	g|k iː|i

Given that our algorithm predicts all missing words regardless of whether this makes sense (including grammatical markers, prefixes and suffixes), it is clear that the selection of items to be explicitly elicited does not need to contain all items for which predictions could be made. The

automatic transcriptons provided in the file `<predictions-automatic.tsv>` may contain predictions for which we already know they are unlikely to exist. This may be due to lexical innovations, borrowings, or because no concepts exist for a given elicitation gloss. For this reason, Bodt made a manual analysis, extracting those predictions that on basis of his experience would be most promising and interesting. This list comprises a list of 630 detailed predictions (including full words with prefix and main root), as well as an informed guess by Bodt that at times may override the automatic prediction. Essentially, this allows us to compare two different kinds of predictions: the fully automated ones, and the ones corrected by the expert. The corrected data is provided as file `<predictions-manual.tsv>`.

Testing the predictions during field work

To verify the predictions, we will proceed as follows:

1. During the upcoming field work, Bodt will elicit the 630 potentially missing cognate words, in addition to other concepts deemed worthwhile for historical reconstruction.
2. All the tested words will be directly tracked with their concept, their variety, and their form.
3. Later, we will compare the number of those words that have been correctly proposed by the algorithm, those correctly predicted by Bodt's corrections, and those words where both these predictions failed.
4. The results will be reported.

There are various reasons why a prediction may fail, the two most important ones of which are:

1. that the algorithm (or the expert) proposes a wrong form (or no form at all), due to problems in the algorithm settings, sparseness in the data, erroneous annotations of cognate sets or alignments;
2. that the word under question is no longer reflected in the language, since it was lost, replaced by a loan or an innovation.

While both reasons seem easy to evaluate, they are in fact not easy to disentangle, as they require a linguistic analysis to be applied to the words that were elicited. We can distinguish three different situations here, best illustrated with hypothetical examples:

1. A word form could be elicited which is cognate with the other words in the cognate set for which a prediction was made, and it can be directly compared with the predicted word form. For example, the predicted form is *zɛ̃ʔ* 'wring, squeeze', the attested form for 'wring, squeeze' is *zɛ̃ːn*, which gives us the possibility to evaluate the predictive capacity.
2. The predicted word form could be elicited, but is not cognate with the other words in the cognate set for which a prediction was made. Instead, the word form has obtained a different meaning. If both meanings (of the original concept and the new meaning) can still be somehow related, they can still be considered cognate, with the shift in meaning the result of semantic change. For example, the predicted form is *zɛ̃ʔ* 'wring, squeeze' and the elicited form for 'wring, squeeze' is *kʰuŋ*, which is clearly not cognate. However, when asking whether there is a word in the language called *zɛ̃ʔ*, the response is that it means 'to milk a cow', which, given that a cow's teat needs to be wrung or squeezed in order to milk

it, renders the predicted and the attested form cognate, and hence enables a meaningful comparison.

3. No cognate word form could be elicited, so the prediction fails since the word could not be found to be still reflected in the language under question. For example, the predicted form is *zɛ̃ʔ* 'wring, squeeze', the elicited form for 'wring, squeeze' is *kʰuŋ* and either there is no existing form resembling *zɛ̃ʔ* or a form resembling *zɛ̃ʔ* means something not cognate like 'mud'.

Strictly speaking, since our algorithm does not predict the *likelihood* of lexical replacment and the loss of words as illustrated above, we can only evaluate those predictions where a cognate word can indeed be identified. For this reason, it is deemed useful not restrict our sample of predicted words too drastically. The more we restrict our sample, the fewer datapoints will be left for comparison. To make clear, however, that a word for which we made a prediction could not be found, we will report all these attempts, and the final evaluation scores will consist of the following comparisons:

1. elicited words vs. words which could be assigned to an existing cognate set; and
2. correctly predicted words vs. incorrectly predicted words

One of the shortcomings in our test is that it is not clear in advance how many of our predictions can, in fact, be tested. However, this is also due to the very nature of "prediction" or "retrodiction" in historical linguistics.

Testing the prediction on the current data

Before conducting an experiment of this kind, it is useful to compute the rate of accuracy we might expect from a random sampling of the data alone. For this purpose, we randomly deleted words from the existing data and then used the distorted dataset to predict the deleted words. The accuracy is then computed for each word form by counting how many times the algorithm proposes the correct word form and how many times it fails. This can be represented in a percentages score, our *accuracy score*. In addition, we report two more values in these experiments, namely:

- *proportion*: the average proportion of words that were excluded from the data; and
- *density*: the cognate density of the wordlist (a score computed with help of the LingPy library).

To test the predictive force of the current algorithmic settings, we ran 100 trials of the algorithm, by running the script `test-prediction.py`. Below are the results:

accuracy	**proportion**	**density**
0.5854	0.6103	0.3452

What they indicate, in brief, is that we can predict a fair amount of the data, namely 58% on average, which is specifically remarkable when considering that the dataset is already quite gappy, with notably missing items.

Timeline

- since 2016: Johann-Mattis List and Nathan W. Hill developed the main framework for correspondence pattern detection and word prediction which is an on-going process,
- July 30 – August 3, 2018: Tim Bodt visited the CALC group, and we started working on the data, aligning words, assigning cross-semantic cognate sets, and annotating morphemes.
- August 30, 2018: Johann-Mattis List carried out the prediction experiment using the lingrex software package, on the datafile `bodt-khobwa-cleaned`.
- October 04, 2018: Tim Bodt sent his final data, a careful selection of words to be tested for the prediction experiment to Johann-Mattis List, with 630 word forms in total, and about 65 on average per doculect, and full annotations on both the automatically predicted form and the manually corrected one
- November 2018: Tim Bodt will check the data during field work and compare with the predictons registered in this experiment before.

References

Abraham, Binny et al. 2005. "A Sociolinguistic Research Among Selected Groups in Western Arunachal Pradesh Highlighting Monpa." Unpublished manuscript.

Blench, Roger, and Mark W. Post. 2014. "Rethinking Sino-Tibetan Phylogeny from the Perspective of Northeast Indian Languages." In *Trans-Himalayan-Linguistics*, edited by Thomas Owen-Smith and Nathan W. Hill, 71–104. Berlin: de Gruyter.

Burling, Robbins. 2003. "The Tibeto-Burman Languages of Northeastern India." In *The Sino-Tibetan Languages*, edited by Graham Thurgood and Randy J. LaPolla, 169–91. New York: Routledge.

Deuri, R. K. 1983. *The Sulungs*. Shillong: Research Department, Government of Arunachal Pradesh.

Dondrup, Rinchin. 1988. *A Handbook on Sherdukpen Language*. Itanagar: Government of Arunachal Pradesh.

———. 1990. *Bugun Language Guide*. Itanagar: Government of Arunachal Pradesh.

———. 2004. *An Introduction to Boot Monpa Language*. Itanagar: Government of Arunachal Pradesh.

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

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A Kaiping, and Russell D. Gray. forthcoming. "Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics." *Scientific Data*.

Genetti, Carol. 2016. "The Tibeto-Burman Languages of South Asia: The Languages, Histories, and Genetic Classification." In *The Languages and Linguistics of South Asia: A Comprehensive Guide*, edited by Hans Heinrich Hock and Elena Bashir. Berlin: Mouton de Gruyter.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath. 2018. "Glottolog." Leipzig: Max Planck Institute for Evolutionary Anthropology. 2018. http://glottolog.org.

Hill, Nathan W., and Johann-Mattis List. 2017. "Challenges of Annotation and Analysis in Computer-Assisted Language Comparison: A Case Study on Burmish Languages." *Yearbook of the Poznań Linguistic Meeting* 3 (1): 47–76.

Jacquesson, François. 2015. *An Introduction to Sherdukpen*. Bochum: Brockmeyer.

Lǐ Dàqín. 2004. Sūlóngyǔ yánjiū [Research on Puroik]. Běijīng: National Minority Publisher.

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

List, Johann-Mattis. 2017. "A Web-Based Interactive Tool for Creating, Inspecting, Editing, and Publishing Etymological Datasets." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, 9–12. Valencia: Association for Computational Linguistics.

———. 2018. "Automatic Inference of Sound Correspondence Patterns Across Multiple Languages." *bioRxiv*. https://doi.org/10.1101/434621.

List, Johann-Mattis, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. 2018. "CLICS². An Improved Database of Cross-Linguistic Colexifications Assembling Lexical Data with Help of Cross-Linguistic Data Formats." *Linguistic Typology* 22 (2): 277–306.

List, Johann-Mattis, and Steven Moran. 2013. "An Open Source Toolkit for Quantitative Historical Linguistics." In *Proceedings of the ACL 2013 System Demonstrations*, 13–18. Stroudsburg: Association for Computational Linguistics.

List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi, and Robert Forkel. 2018. "Sequence Comparison in Computational Historical Linguistics." *Journal of Language Evolution* 3 (2): 130–44.

Post, Mark W., and Robbins Burling. 2017. "The Tibeto-Burman Languages of Northeastern India." Edited by Graham Thurgood and Randy J. LaPolla. London: Routledge, 213–33.

Rama, Taraka, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. "Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics?" In *Proceedings of the North American Chapter of the Association of Computational Linguistics*, 393–400.

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

Stonor, Charles Robert. 1952. "The Sulung Tribe of the Assam Himalayas." *Anthropos* 47: 947–62.

Sun, Tianshin Jackson. 1992. "Review of Zangmianyu Yuyin He Cihui (Tibeto-Burman Phonology and Lexicon)." *Linguistics of the Tibeto-Burman Area* 15: 73–113.

———. 1993. "A Historical-Comparative Study of the Tani (Mirish) Branch in Tibeto-Burman." PhD, Berkeley: Department of Linguistics, University of California.

Sūn Hóngkāi, ed. 1991. *Zàngmiǎnyǔ yǔyīn hé cíhuì* [Tibeto-Burman Phonology and Lexicon]. Běijīng: Chinese Social Science Press.