Author's version, please cite as:

Bodt, Timotheus Adrianus and Johann-Mattis List. 2020. The multiple benefits of making predictions in linguistics. *Babel: The Language Magazine* 31(2): 8-12.

The multiple benefits of making predictions in linguistics

Timotheus A. Bodt and Johann-Mattis List

Through an experiment on a Western Kho-Bwa linguistic dataset, ***Timotheus A. Bodt*** *and* ***Johann-Mattis List*** *provide evidence for the regularity of sound change.*

Predictions are an integral part of our lives. We listen to the weather report to plan our next day's trip and we observe economic forecasts to plan our spending and investment. We are relieved when the weather is as predicted and annoyed when it is not. When economic trends follow the forecasts, we are glad we have been able to make the best decisions.

In science, predictions play an equally important role as in real life. Weather reports are built on predictions made by meteorologists and economic forecasts rely on studies by economists. In the field of linguistics, too, we make predictions on a regular basis. Descriptive linguists working on languages in the field make predictions about phonemes, morphemes and syntactic constructions and their functions when eliciting data from speakers. Historical linguists make predictions about how words would have looked in a proto-language based on evidence from contemporary and historically attested languages. And language learners make predictions of grammatically acceptable sentences in a foreign language, sentences which they hope that native speakers will understand correctly. A major difference with meteorologists and economists is, that linguists hardly make their predictions *explicit*. Linguists publish the final results, the most likely outcomes of the various predictions they make, but the individual predictions remain restricted to the individual thought processes and notes on paper or in electronic form.

The experiment

Between September 2018 and November 2019, we conducted an experiment, to see what benefits may be obtained from more explicitly formulating and communicating predictions in historical linguistic research. The background for the experiment was a dataset of eight Western Kho-Bwa varieties. Western Kho-Bwa is a small sub-group of the Kho-Bwa cluster, belonging to the Tibeto-Burman (Sino-Tibetan / Trans-Himalayan) language family and spoken in the Indian state of Arunachal Pradesh. Mattis normalised and uploaded this dataset to the software programme *EDICTOR* to facilitate analysis of the data and make this analysis more transparent and insightful.

The **comparative method in historical linguistics** is based on establishing regular, exceptionless **sound correspondences** between attested (spoken or written) language varieties by comparing **cognates**. Cognates are lexical roots or grammatical morphemes with a common etymological origin. From the sound correspondences between these cognates, sequences of **sound changes** are derived that are used to reconstruct the **proto-language**. Proto-languages provide insights into which languages are related and that these related languages descend from a single common language. This proto-language can also lend evidence to the habitat, culture and livelihood of the people that once spoke it.

In the subsequent analysis, Mattis first automatically identified **cognates** and then Tim manually adjusted those. An example of cognates from better known languages are the English noun *man* : Dutch noun *man* : German noun *Mann*. These forms are most commonly pronounced as English [mæn] : Dutch [mɑn] : German [man], which is how they are noted in the International Phonetic Alphabet. Mattis then used a new algorithm to automatically identify **regular sound correspondences** between the eight varieties. Using the same example, the initial *m-* corresponds regularly in all varieties, not just in the example for ‘man’, but also in the English adverb *more* [**m**ɔː(r)] : Dutch adverb *meer* [**m**eːr] : German adverb *mehr* [**m**eːɐ̯] and in the English noun *milk* [**m**ɪlk] : Dutch noun *melk* [**m**ɛlk] : German noun *Milch* [**m**ɪlç]. Similarly, the intermediate vowel *-a-* and the final *-n* in ‘man’ also correspond in

the English conjunction *than* [ð**æn**] : Dutch adverb and conjunction *dan* [d**ɑn**] 'then / than' : German adverb *dann* [d**an**] 'then' and in the first person present form of English verb *can* [k**æn**] : Dutch verb *kan* [k**ɑn**] : German verb *kann* [k**an**].

During this analysis, we found that there were certain **gaps** (or *blanks*, or *missing values*) in the data, where certain varieties did not have an elicited form for a certain concept. Based on the sound correspondences we had established, Mattis made automatic predictions for the phonemes in the variety for which a concept had a gap. In other words, he predicted the *reflexes*: what we expect that a certain concept would sound like in a certain variety, based on the available data from the other varieties. To use the example above, if we had English *man* [mæn] and German *Mann* [man] 'man', but no form in Dutch, we would predict the Dutch form to be *man* [mɑn] based on the regular sound correspondence English *m-* : Dutch *m-* : German *m-*; English *-æ-* : Dutch *-ɑ-* : German *-a-*; English *-n* : Dutch *-n* : German *-n*.

The predictions

This generated a list of in total 2108 automatically generated predictions. For each concept, Mattis made three automated predictions, which introduced increasing uncertainty, in other words, more possible phonemes for each segment within a predicted form based on the sound correspondences we had identified. Tim, as the expert on the Western Kho-Bwa varieties, then further refined these predictions manually, selecting 631 morphemes that were combined to 519 concepts that could be verified. Some predicted forms – such as prefixes or suffixes, or morphemes of words that consisted of more than one syllable – had to be combined with others, because they expressed a single concept. For other predicted forms, Tim already knew that they did not exist in a certain variety, for example, because that variety had borrowed a form from another language. In some cases, the algorithm made a prediction, but the evidence on which this prediction was based was too limited, too diverse, or inconclusive. After these manual refinements, we *registered* these predictions online. Registration of hypotheses is now common in psychology and related disciplines to ensure that scientists do not create hypotheses *after* having conducted experiments. Creating

hypotheses after the experiment has been conducted is considered to be statistically problematic. To register our linguistic prediction experiment, we uploaded the original data to the Open Science Framework, accompanied by instructions on how to replicate the automated part of the process of word prediction from the original data and created a time-stamped version that could no longer be modified. We also published a working paper explaining the procedure we followed.

The evaluation

Subsequently, Tim went back to the field, where he elicited as many of the 519 predicted concepts as possible. He initially used the technique of *direct elicitation*, asking for the concept in a given variety. So, for example, he would have asked a Dutch informant "How do you say 'man' in Dutch?", to which the respondent may say [mɑn] or [mɛn], which would both have been noted as a *direct cognate*. The value provided by the informant could also not have been cognate with the form we predicted, for example, the Dutch respondent may have said *mens* [mɛns], which means 'human' and has an additional final *-s*. In those cases, Tim asked whether there perhaps was another word for 'man'. If the response was still not cognate, he asked for the predicted form itself, for example, "Is there a word called [mɑn] in your language, and what does it mean?". Sometimes, this also resulted in a positive cognate decision, because there had been semantic change between the original meaning of the word, and the present meaning of the word. These forms were then noted as *indirect cognates*.

Tim was able to elicit a total of 452 predicted concepts: Depending on the specific variety, this ranged between 72.5% and 100%, with an average of 87.1%. 66 predictions could not be verified, because the respondent did not understand the concept and the concept could not be correctly explained, or because the respondent did not have any response. In addition to the 66 predictions that Tim could not verify, there were 132 attested forms that he did not consider as being cognate to the predicted forms, because of the limited knowledge of certain varieties and their contact languages; because of different roots; because of loans from contact languages; or because of lexical innovations.

Tim adjudged 319 attested forms as cognate to the predicted form: This varied between 52.6% and 80.0% depending on the variety, with an average of 70.1%.

We then evaluated the reflex predictions that had attested cognate forms, by comparing every segment of the predicted form with every segment of the attested form. Mattis calculated **accuracy scores** by dividing the number of correctly predicted segments in a prediction by the total number of segments. To use the example above: Since the Dutch word for 'man' is [mɑn], an attested form [mɑn] would obtain a score of 1.0 (three correct segments / phonemes divided by a total of three segments), whereas an attested form [mɛn] would obtain a score of 0.66… (two correct segments / phonemes divided by a total of three segments).

The results

The most conservative automated prediction was based on the most likely sound correspondence for each segment and hence the lowest level of uncertainty. This automated prediction has an average accuracy score of 0.71, ranging between 0.64 and 0.78. When introducing more uncertainty by adding optional phonemes for each segment, the average accuracy score increased to 0.73, ranging between 0.66 and 0.79. The manually adjusted predictions Tim made had an average accuracy score of 0.77, ranging between 0.66 and 0.89.

We observed that Tim's expert predictions were better than the automated predictions in six out of eight varieties, and the same or marginally worse in the two remaining varieties. This is expected, as an expert will always have more knowledge than a computer algorithm. Furthermore, we could clearly see that the predictions for those varieties that Tim knew best, Duhumbi and Khispi, had higher accuracy scores than the varieties that he knew least well, and that the predictions for the variety that was phonologically the most aberrant because of contact language influence, Khoina, were least accurate. We could also conclude that introducing more uncertainty in the automated predictions improved their accuracy scores.

Benefits of predictions

One of the main reasons for a lower accuracy of the predictions is at the same time one of the greatest benefits of conducting our prediction experiment. The existing dataset with Western

Kho-Bwa concepts had not been completely analysed before the prediction experiment was set up. Although the main sound correspondences were identified and added to the dataset, there were still several sound correspondences that were not included in the analysis. Some had not been automatically detected by the algorithm, and although Tim had identified them he had not manually added them yet (i.e. omissions); others were automatically detected but occurred in such low frequencies that they were ignored; and some were neither automatically, nor manually detected yet (i.e. unidentified sound correspondences). In several cases, discrepancies between the predicted value and the attested value forced Tim to address the latter issue, in which he was able to set up a new, hitherto unidentified sound correspondence. In this way, the prediction experiment greatly benefited the reconstruction of the linguistic history of the Western Kho-Bwa languages and the reconstruction of its ancestor language, Proto-Western Kho-Bwa.

But we realised several other benefits to conducting the prediction experiment and to making predictions in linguistics in general. If historical linguists and field linguists explicitly state their predictions and communicate these predictions to the scientific community, this will enhance the rigour of their own research, forcing them to think about what they predict, come to more structured predictions, enable other researchers to cross-check their data and results and allow cross-checking of their data with other's data. This will greatly increase the transparency of linguistic research.

Predictions made on the basis of better known linguistic varieties can make both elicitation in the field and finding cognates in published work more effective and efficient, because they will enable us to ask or search for what we think we need to know in related but poorly described varieties. This is especially important in view of language death and funding limitations.

Prediction experiments, both in their regularity and their deviation from regularity, can show students in linguistics 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.

The respondents themselves noted that asking concepts and predicted forms made them remember their own language and encouraged them to either ask speakers nearby or even

use social media like WhatsApp, kindling the renewed interest in their own language so important for the possible survival of endangered languages.

And, last but not least, our experiment has shown, that predicting missing values based on regular sound correspondences that follow from regular sound changes results in valid predictions, hence that the sound changes that they are based on must be regular. Because the accuracy of the automated predictions was high, a substantial part of the sound correspondences that were largely automatically identified and manually adjusted must have been correct. Therefore, there is regularity in sound change, which confirms the basic tenet of the comparative method in historical linguistics.

We describe the full results of our prediction experiment in an article submitted for review to the journal Diachronica.

Find out more:

Articles

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

Online

- Database of languages of the world: http://glottolog.org
- Resource linking different linguistic concept lists: https://concepticon.clld.org
- Database of phonetic Cross-Linguistic Transcription Systems: https://clts.clld.org
- Database of Cross-Linguistic Colexifications: https://clics.clld.org
- The Open Science Framework: https://osf.io
- The International Phonetic Alphabet: https://www.internationalphoneticassociation.org/

About the authors

Tim Bodt is a postdoctoral researcher affiliated with the University of Bern in Switzerland and the School for Oriental and African Studies in London. He has worked intensively on languages of the Himalayan region, including Tshangla, the Western Kho-Bwa languages and Kusunda.

Johann-Mattis List is a senior scientist working on computer-assisted approaches in historical linguistics and linguistic typology at the Max Planck Institute for the Science of Human History in Jena. His work includes the development of automated methods for historical language comparison, most prominently represented by the LingPy software package (http://lingpy.org), and the curation of large cross-linguistic datasets, such as the Database of Cross-Linguistic Colexifications. He reflects on the interdisciplinary setting of his research in monthly contributions to https://phylonetworks.blogspot.com and explains the importance of linguistic research for our daily life in his German blog at https://wub.hypotheses.org.