Predicting historical phonetic features using deep neural networks: A case study of the phonetic system of Proto-Indo-European

Frederik Hartmann
University of Konstanz
frederik.hartmann@uni-konstanz.de

Abstract

Traditional historical linguistics lacks the possibility to empirically assess its assumptions regarding the phonetic systems of past languages and language stages beyond traditional methods such as comparative tools to gain insights into phonetic features of sounds in proto- or ancestor languages. The paper at hand presents a computational method based on deep neural networks to predict phonetic features of historical sounds where the exact quality is unknown and to test the overall coherence of reconstructed historical phonetic features. The method utilizes the principles of coarticulation, local predictability and statistical phonological constraints to predict phonetic features by the features of their immediate phonetic environment. The validity of this method will be assessed using New High German phonetic data and its specific application to diachronic linguistics will be demonstrated in a case study of the phonetic system Proto-Indo-European.

1 Introduction

Since the beginning of historical linguistics, one of the main aims of historical phonology and phonetics has been to reveal phonetic features of now lost sounds and phonological systems of past languages. The study of the phonetic system of earlier stages of languages is a crucial prerequisite to uncover sound change and effects on sound change precisely. The methods, however, are limited for every language whose speakers cannot be invited to a phonetics lab for detailed testing. The temporal scope of inquiries into language change would be fairly limited if we could only examine language change as far back as voice recording and experimental testing methods were present. If we want to study language change over thousands of years, we must rely on robust techniques to approximate historical phonetic features as well as possible. The most prominent methods in historical linguistics so far to achieve this goal are based on comparative approaches (cf. Campbell, 2013; Beekes and Vaan, 2011; Meier-Brügger et al., 2010). Especially for the reconstruction of proto-languages, historical phonologists use the comparative method to estimate the approximate quality of sounds by investigating their outcomes and effects in the descendant languages. However, approaching historical phonetics by comparative means bears the disadvantage that the more the daughter languages disagree in certain respects, the less precise are the estimates scholars can make for the respective proto-sounds. For some problems for which comparative techniques yield imprecise results, there is a need for alternative methods to tackle these issues. Moreover, there is also no alternative method for cross-checking assumptions obtained through the traditional methods as such an alternative would need to operate on a basis different from diachronic comparison. Thus the method proposed in this paper makes use of synchronic structures and features of a language's phonology and feeds this data into a deep neural network to predict the phonetic features of unknown sounds. The data the network can draw upon is the direct phonetic environment of each sound with the goal to predict its features only by the features of its environment.

The reason for the predictability of sound features in the context of their environment is due to coarticulatory effects, statistical constraints and local predictability. Coarticulation refers to the observation that sounds tend to both influence and be influenced by their environment phonetically (see e.g. Kühnert and Nolan, 1999; Ohala, 1993a; Hardcastle and Hewlett, 2006; Fowler, 1980). This reciprocal influence can be detected synchronically which makes it a possible alterna-

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 98–108
Florence, Italy, August 2, 2019.

tive to be used for historical phonology if applied to historical language stages or proto-languages: In theory, sounds constantly influence their environment and are affected by it at the same time so that a tight net of interlaced dependencies between sounds and their environment arises. There are indications that sound changes which better fit into this phonetic structure in their initial stage are more likely to become widely adapted (Donegan and Nathan, 2015; Blevins, 2015; Ohala, 1993a,b; Hale, 2003).

Similarly, and partially originating from coarticulatory processes, we find certain types of phonological constraints in languages, be it syllable composition constraints or the prevention of certain consonant clusters which make up a language's phonotactics. These constraints can be both *absolute* and *statistical*, whereby absolute constraints are rules which are never violated, whereas statistical constraints constitute a strong dominance of one phonological shape over others. The network can utilize a language's phonotactics, constraints and coarticulatory effects to predict the phonetic features of a target sound. Feature predictions from environmental properties have already been studied in quantitative phonetics and proven to be possible to some degree due to local predictability effects (see e.g. Priva, 2015; Van Son and Van Santen, 2005; Raymond et al., 2006).

It is important to keep in mind that local predictability on the basis of the phonetic environment is, in fact, *not* contradictory to the observation that different sounds can occur in the same environments which can be demonstrated using minimal pairs. Predictability in this context does not mean that a certain environment of a given sound *always* yields certain phonetic properties, it is rather a probabilistic observation that environments *tend* to occur paired with certain phonetic features and that this tendency of forming patterns is what can be predicted using probabilistic models and machine learning algorithms.

2 The deep neural network approach

Using machine learning algorithms is not new to the field of linguistics, though it is one of the more recent methods.[1] While these approaches are found in an increasing number of studies in linguistics in general, in historical linguistics in particular the method is less used although some studies have been published in this or adjacent fields such as cladistics (Jäger et al., 2017; Jäger and Sofroniev, 2016). Since this approach of predicting sound features by the features in the phonetic environment only works synchronically, the deep neural network used for this needs to be trained on better known phonological features as the basis for predicting unknown features.

The data fed to the network must therefore contain a dataset where the phonetic environment serves as the input that is mapped on the target sound. To achieve this, the lexical corpus data needs to be split into trigrams or pentagrams of phonetic segments which are then categorized with regard to their phonetic features. Afterwards, the middle or target sound is removed and the remaining environment passed through the network with the respective target sound features as labels. Doing this trains the model to detect the correct phonetic features for the target sound given its environment. If the network has successfully trained, the environments of unknown sounds can be passed to the model which will, in turn, predict the features of the sounds on the basis of its weights and biases obtained in the training process. When the network performs well on the training data, we have little reason for it performing worse on the prediction of unknown sounds. Deep neural networks are especially suited for this task since other methods such as random forests or support vector machines have performed worse on this classification in preliminary tests I conducted beforehand. These three approaches, Deep neural networks, random forests and support vector machines, are entirely different approaches to machine learning classification tasks: While random forest classificators aim at finding the best decision tree by partitioning the data in subgroups, support vector machines establish the best splitting function, a hyperplane, to classify new samples according to their position in the multi-dimensional space. Deep neural networks on the other hand aim at optimizing the decision function through means of building abstract representation of the data and 'learning' the occurrence patterns of data features. It is not always possible to determine why some algorithms perform worse on some datasets and better on others. In the task at hand we can merely state that deep neural networks

[1] See e.g. Chollet (2018); Nielsen (2015) for a general introduction to deep learning.

seem to find the global minimum, or a better local minimum, of the decision function well while other algorithms do not perform on the same level, presumably due to their characteristics not being ideal for this particular case. In the following section, a case study on Proto-Indo-European shall function as an example study that can be conducted using neural networks.

3 Case study: The phonetic system of Proto-Indo-European

The phonetic system of Proto-Indo-European (PIE) is an ideal field to demonstrate the capabilities of this neural network approach for several reasons: (1) while the phonetic inventory of PIE, along with its phonotactics, has been reasonably well investigated (Clackson, 2007, 64-71; Meier-Brügger et al., 2010, 272-275; Byrd, 2015; Ringe, 2017, 13-17; Fortson IV, 2011, 62-64), there are still unknown aspects that lead to scholarly discussions and diverging theories such as the Glottalic theory.[2] (2) three sounds of PIE, the so-called laryngeals, are still a matter of debate since they are only scarcely attested in PIE's daughter languages and sometimes only through their effects on neighbouring sounds. The case study will therefore aim to propose an attempt to predict the laryngeals and to uncover possible inconsistencies in the phonetic system of PIE. The three laryngeals (h_1, h_2, and h_3) are reconstructed sounds in PIE whose exact phonetic value is unknown. Apart from some direct evidence of laryngeal reflexes in the Anatolian languages, most of our knowledge of those sounds stems from structural and phonetic patterns the laryngeals induced in the daughter languages before they faded altogether. Previous research interprets the laryngeals h_1 : h_2 : h_3 as [ʔ]/[h] : [χ]/[x]/[ɣ]/[ʕ] : [ɣ^w]/[ʕ^w]/[ʁ] .(Rasmussen, 1994; Kümmel, 2007; Meier-Brügger et al., 2010; Beekes, 1994; Bomhard, 2004; Gippert, 1994; Weiss, 2016; Mayrhofer and Cowgill, 1986)

3.1 The data

One of the best resources to obtain reconstructed word data that is already digital is the English version of Wiktionary.[3] Its validity as a repository of data for linguistic research has been assessed by multiple studies and many other studies have already used its database for linguistic inquiry (e.g. Chiarcos et al., 2013; Navarro et al., 2009; de Melo, 2015; Zesch et al., 2008; Meyer and Gurevych, 2012). Especially regarding reconstructed language data, Wiktionary has the decisive advantage that the reconstructions follow certain guidelines (see Wiktionary contributors) unlike data collected from various different traditional dictionaries.

For this study, I extracted all PIE reconstructions found in page headings from the English Wiktionary .xml dump on 20.10.2018. Such a dump file contains all English Wiktionary pages including page and edit histories. The lemmas that were extracted were subsequently split into segments of trigrams: preceding sound, target sound and following sound with a final trigram count of 7782. Where a trigram contained a root ending, '-' was used as following sound to encode the root ending, cases of word-final or word-initial were added as 'zero' in the preceding or following sound slot, respectively. Each sound was ultimately classified according to its place and manner of its articulation according to the reconstructed phonetic inventory of PIE most scholars agree on (e.g. Clackson, 2007, 34; Beekes and Vaan, 2011, 119; Ringe, 2017, 8) without considering the glottalic theory.[4]

3.2 Approaches to verify the method

Before we are able to apply any machine learning techniques to the data, we need to establish whether coarticulatory and statistical constraint effects exist in PIE and that the method is actually feasible for predicting sound features in general. Although there have been studies suggesting the existence of such effects as mentioned above, a preliminary analysis needs to be conducted to *demonstrate* the data shows these effects and that a deep neural network can indeed 'learn' them and make correct predictions on the basis of the observed patterns.

For this reason, I set up a generalized linear logistic regression model as an example to determine the phonetic effects on the occurrence of the feature *aspirated* in PIE. The model was fit for best AIC through both top-down and bottom-up fitting. Before fitting, aliases were removed as

[2] See Byrd (2015); Beekes and Vaan (2011); Clackson (2007) for a comprehensive overview of the scientific debate.

[3] https://en.wiktionary.org, accessed: 2019-03-13

[4] For the full list of features used in this study, please refer to the appendix.

well as collinear predictors up to a cutoff-point of Variance Inflation Factor (VIF) greater than 4.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.302	0.091	-25.312	0.000
labial preceding	-1.504	0.240	-6.269	0.000
sibilant preceding	-2.310	0.587	-3.938	0.000
liquid preceding	-0.843	0.245	-3.446	0.001
syllabic cons. preceding	1.137	0.380	2.994	0.003
back vowel preceding	-1.131	0.267	-4.232	0.000
mid vowel preceding	-1.022	0.178	-5.730	0.000
close vowel preceding	-0.947	0.468	-2.021	0.043
h_1 preceding	-1.519	0.518	-2.933	0.003
h_2 preceding	-2.106	0.513	-4.103	0.000
h_3 preceding	-1.185	0.598	-1.981	0.048
word boundary following	1.439	0.154	9.336	0.000
voiceless cons. following	-2.279	0.390	-5.848	0.000
nasal following	-1.691	0.512	-3.304	0.001
liquid following	0.444	0.175	2.543	0.011
syllabic cons. following	0.381	0.185	2.055	0.040
velar following	1.487	0.509	2.919	0.004
back vowel following	0.526	0.170	3.091	0.002
plosive following	-1.041	0.415	-2.510	0.012
h_2 following	-2.494	1.006	-2.481	0.013

Table 1: Generalized linear logistic regression for the occurrence of the feature *aspirated*

It can be observed in table 1 that several predictors were significant. E.g. preceding sibilant reduces the probability of the target sound being *aspirated* whereas following velar increases this probability. As suggested by this model, the data contains information on coarticulatory and statistical constraint effects the neural network can draw upon.

As a second approach to ensure that the presented method and data is suitable for predicting sound features, I conducted a preliminary study using the same method to predict the features of New High German sounds. For this analysis, I utilized the German phonology lemma data from CELEX2 (Baayen et al., 1995) in the syllabified phonetic lemma transcription with stress in the DISC character set (*PhonStrsDISC*). After extraction from the CELEX2 file, the data were prepared using the same process as for the PIE data with a final sample size of 441236 German trigrams. The method was simultaneously tested with a dataset in which each lemma was oversampled proportional to its frequency of occurrence in the 'Mannheimer Korpus' provided by CELEX2 (*Mann_Freq*) (see Gulikers et al., 1995). While this approach would ideally proportion the dataset more realistically and could, in theory, improve model training, it did not enhance the performance of the network and was therefore discarded.

Each sound of those trigrams was classified according to 38 phonetic features (e.g. *consonant*, *nasal*, *plosive*) where 0 and 1 indicate the absence/presence of a particular feature, respectively.[5] Note, that these 38 features contain some redundancies (e.g. *vowels* are entirely contained in the feature *continuant*). This is due to the fact that a deep neural network performs best on as many input features as possible since there might be some relevant signal in a seemingly redundant or unimportant feature vector. Accordingly, specifying two complementary features like e.g. *voiced* and *voiceless* can increase the network's performance since the two categories only apply to consonants. Otherwise, a single binary feature [+voice] would not only encode voiced consonants but also all vowels and therefore decrease the ability of the network to detect voiced consonants specifically. Redundancy itself is also not a problem as redundant or irrelevant information in the data is weighted less important during training while the network focuses on those features that have predictive power.

Also, only basic features (13 features in total for consonants and 10 features for vowels) such as e.g. *consonant*, *velar* and *labial* were used as target features for the prediction of German sound features. The reason for this decision was that the more fine-grained the distinctions become, the fewer occurrences of the feature there are on which the network can train. Therefore, although the feature *liquid* containing German *r* and *l* was further divided into *rhotic* and *lateral* as features contained in the classification of the phonetic environments, only *liquid* was tested as a target feature. If *rhotic* were tested as target feature on a sound with unknown features, the network would train only on the sound *r* and therefore not necessarily train on the feature *rhotic* but rather learn to discriminate *r* from all other sounds which has in turn little explanatory power when predicting the *rhotic* feature for other sounds.

The method was tested on the German sounds *p, r, ɛː, aː* as an arbitrary preliminary selection that ideally is representative of all other sounds in the New High German phonetic inventory. Therefore, four datasets were prepared, where the respective sound was removed as target sound and its presence in any phonetic environment was indicated by adding a new feature only for this sound. For example when the phonetic environment in a par-

[5] For the full list of features used in this study, please refer to the appendix.

ticular trigram contained *r* while *r* was the sound to be later predicted by the network, *r* was classified in a dummy feature category that only encodes presence/absence of this particular sound. This procedure is necessary since removing all instances of the particular sound, *r* in this case, in the phonetic environment would reduce the number of environments and therefore distort the data.

After data preparation, a single network was set up for each feature and trained one feature at a time with a binary output to predict the presence or absence of the feature. I.e. this binary network was trained to detect a particular feature and to predict its presence or absence for unseen sound environment data. After the entire data were shuffled and the test and validation data were separated from the training sets using the Stratified ShuffleSplit cross-validator included in the python package *scikit-learn* (Pedregosa et al., 2011), the training sets were over-sampled before each run to counter class imbalance with the SMOTE algorithm (Chawla et al., 2002) implemented in the 'Imbalanced-learn' (Lemaître et al., 2017) python package. The network was trained for 30 epochs using the optimizer Adam with a learning rate of 0.01 with a batch size of 250 samples with the layer configuration displayed in table 2.

Layer	Layer size	Activation
Dense layer 1	256	ReLU
Dense layer 2	128	ReLU
Dense layer 3	64	ReLU
Dense layer 4	32	ReLU
Output layer	2	softmax

Table 2: Network architecture for the German feature prediction task

For the subsequent evaluation of the model performance, weights and biases were used form the epoch at which the network performed best on the validation data during training using the Keras callback *ModelCheckpoint* (Chollet et al., 2015). This procedure minimizes the risk of the model being stuck at a local minimum in the search space at the time training stops after an arbitrarily chosen number of epochs. It has been established in preliminary tests that the model performance was enhanced when training on an all-consonant or all-vowel subset of the data: First, a model was trained to predict the feature [± consonant] and after the prediction, the main model was trained on consonant or vowel data according to the prediction of the preliminary model. After each training, the network performance was evaluated and subsequently tasked with predicting the particular feature for the respective test sound. The results are presented in tables 3, 4, 5, and 6 which show which number of samples in the test sets were classified correctly or incorrectly. I.e. 24656 consonant samples in the column *TP* means that 24656 samples of all positive samples in the test set were correctly classified as positive. Similarly, in table 3 in the first row, 7211 samples in *prediction: feature present* denote that 7211 of all tested instances of *p* were classified as [+consonant].

Note that model accuracy metrics such as F1 score, precision, or recall are not given here since these measures only evaluate a classifier's performance on a mixed dataset. Because the method proposed here aims at performing well on determining whether a sound shows a given feature and since this feature is either present in all samples of this sound or absent in all samples, the main goal is that the deep network yields more true positives than false negatives and more true negatives than false positives. Applied to the example in table 3 this means that since German *p* is [+consonant], ideally the majority of classified samples will be classified as such. If after model evaluation the number of false negatives were higher than the number of true positives, the model would likely not be able to classify the majority of samples correctly. More samples would end up being incorrectly labeled as negatives as a result of the poor model training yielding more false negatives than true positives. Therefore, a high false positive or false negative count is not a concern in itself as long as the ratio of true positives to false negatives and true negatives to false positives is always in favor of true positives or true negatives, respectively.

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
consonant	24656	2184	859	15541	7211	1627
nasal	3885	1553	4255	17148	2609	6229
plosive	5417	1984	4341	15099	4860	3978
affricate	732	172	6750	19187	2352	6486
fricative	7156	3627	4272	11786	2394	6444
liquid	4698	1615	5361	15167	1670	7168
sibilant	2148	1072	5676	17945	1634	7204
voiced	11560	3681	3442	8158	3582	5256
labial	3447	864	7656	14874	5507	3331
dental/alveolar	8747	4093	3834	10167	4155	4683
palatal	1019	270	4497	21055	2373	6465
velar/uvular	4896	3035	4200	14710	1972	6866
glottal	428	43	5481	20889	1856	6982

Table 3: Network evaluations and predictions for German *p*

The results show that all 13 tested features of *p* are predicted correctly, *r* is correctly predicted to be a voiced liquid, yet regarding place of articulation, which in German r-allophones is ranging

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
consonant	21986	1739	1311	15089	39071	920
nasal	3722	1716	2702	15585	16238	23753
plosive	5524	2761	3082	12358	6333	33658
affricate	732	172	6339	16482	7972	32019
fricative	4881	1903	4314	12627	11352	28639
liquid	1604	710	5259	16152	22942	17049
sibilant	2172	1048	4683	15822	8997	30994
voiced	8006	3236	3568	8915	30068	9923
labial	3907	1288	6205	12325	12071	27920
dental/alveolar	8974	3865	2844	8042	28021	11970
palatal	1006	283	3138	19298	3895	36096
velar/uvular	2916	1015	4937	14857	11004	28987
glottal	432	39	4728	18526	8695	31296

Table 4: Network evaluations and predictions for German *r*

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
consonant	25884	1840	1111	15105	204	1638
front vowel	3658	1963	2714	7881	589	1253
central vowel	4546	1667	2529	7474	858	984
back vowel	1920	1032	3391	9873	831	1011
round	1395	653	3845	10323	898	944
close	3054	1729	2132	9301	493	1349
mid	5790	1670	2525	6231	585	1257
open	2582	1391	3110	9133	1219	623
diphthong	1097	333	2776	12010	464	1378
long	6595	1544	2365	5712	1248	594

Table 5: Network evaluations and predictions for German *ɛː*

from alveolar to uvular (cf. Meinhold and Stock, 1982, 131-133), only dental/alveolar is predicted which makes a total of 11 out of 13 features. The German vowels were less well detected, with a total of 8 out of 10 for ɛː and 6 out of 10 for aː. Although the model performs on some sounds and features better than on others, it performs better than expected by chance. Since these results stem from a selected set of sounds in a preliminary study, specific questions as to which features are detected better than others and why some features are incorrectly predicted for certain kinds of sounds need to be established in further research.

3.3 The deep learning method applied to Proto-Indo-European

To prepare the PIE data for training, the data were randomly shuffled and split into training and test set using the Stratified ShuffleSplit cross-validator included in the python package *scikit-learn* (Pedregosa et al., 2011). Afterwards, the training set was first oversampled with the SMOTE algorithm and subsequently under-sampled by removing Tomek links using SMOTETomek (Batista et al., 2003) implemented in the 'Imbalanced-learn' (Lemaître et al., 2017) python package to counter class imbalance in the dataset. Yet the SMOTE over-sampling process performed on the minority group increases the dataset's variation, so to cope with this variation and to make sure that findings were not due to random biases during oversampling or stratification, I ran each network 100 times to have a representative number of slightly varying model outputs. Each of these runs yields a confusion matrix with the count of true positive, false negative, false positive and true negative predictions of the test samples. To determine whether the model performs significantly better than expected by a random class assignment, all confusion matrices were compared using Wilcoxon signed rank tests with continuity correction. For each model, I performed this test on the output of the 100 runs of true positives vs. false negatives to determine whether the network can clearly find a present feature and a second test on the 100 runs of false positives vs. true negatives to determine whether the network can clearly find the absence of a feature. When the Wilcoxon signed rank test is significant, the tested groups are 'non-identical' populations.

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
consonant	25694	2030	1102	14445	597	7936
front vowel	3864	1941	2457	7285	4691	3842
central vowel	3996	1364	2302	7885	2760	5773
back vowel	1877	1075	2846	9749	2720	5813
round	1377	671	3550	9949	3848	4685
close	3177	1606	2282	8482	2403	6130
mid	5953	1691	2303	5600	5151	3382
open	2070	1050	2827	9600	3104	5429
diphthong	1164	266	2810	11307	2178	6355
long	5834	1636	2121	5956	4839	3694

Table 6: Network evaluations and predictions for German *aː*

3.4 Example 1: The phonetic quality of the PIE laryngeals

In the following stage, a deep neural network can be set up to learn to detect the feature *aspirated* and to subsequently predict whether the laryngeals had this feature.

The network was trained for 50 epochs using the optimizer Adam with a learning rate of 0.01 and a batch size of 64 samples with the layer configuration displayed in table 7.

Layer	Layer size	Activation
Dense layer 1	128	ReLU
Dropout layer 1	0.25 dropout rate	
Dense layer 2	64	ReLU
Dropout layer 2	0.25 dropout rate	
Dense layer 3	32	ReLU
Output layer	2	softmax

Table 7: Network architecture for the feature *aspirated*

The dropout layers in this network architecture were implemented to reduce the effect of over-fitting due to the limited amount of training samples. Analogous to the training on the mod-

ern German dataset above, only weights and biases form the epoch at which the network performed best on the validation data during training were used. As mentioned above, the network was trained and evaluated 100 times in order to further minimize the effect of accidental findings in single runs. The results are listed in table 8 which is a summary of all test set prediction confusion matrices obtained in the 100 runs.[6]

	True positives	False negatives	False positives	True negatives
Mean	58.32	19.68	219.5	602.5
Median	58	20	221	601
Std. dv.	**2.044**		**6.920**	

Table 8: Statistics of the confusion matrices from 100 runs for classifying the feature *aspirated*

Subsequently, a Wilcoxon signed rank tests with continuity correction with the alternative hypothesis H_1: True positives greater than false negatives gives W = 10000.00 p < 0.00001. A second Wilcoxon signed rank tests with continuity correction with the alternative hypothesis H_1: True negatives greater than false positives gives W = 10000.00 p < 0.00001. These test statistics show that in these 100 runs, the network was able to detect the feature *aspirated* reliably and, most importantly, when presented with an unseen dataset which either contains sounds that have the feature *aspirated* or sounds that do not, the network will correctly identify over 70 percent of the samples. The variance in the prediction accuracy in table 8 can be explained by, as previously addressed, noise in the data and variation in the partitioning and subsequent oversampling of the training set. Having established the functioning network, the model can be used to predict the target feature for sounds with unknown qualities. Since the laryngeals cannot be assigned a phonetic value by means of the comparative method, their properties can be predicted. To achieve this, the phonetic environment was passed through the networks after training at the end of each of the 100 runs. The output of every prediction is a classification matrix for each of the three laryngeals. Table 9 shows the summary of these classification matrices.

To determine the significance of these findings, Wilcoxon signed rank tests with continuity correction were applied to the predictions. Table 10

[6]The figures in the tables provided here and below represent the number of classified samples from the test set. E.g. a mean of 58.32 in *true positives* means that from all positive samples in the test set, an average of 58.32 samples were classified correctly as positive.

	h_1		**h_2**		**h_3**	
	Positives	Negatives	Positives	Negatives	Positives	Negatives
Mean	113.8	85.2	160.5	187.5	54.06	51.94
Median	115	84	163	187	53	53
Std. dv.	**4.141**		**4.079**		**1.879**	

Table 9: Prediction results by the trained model for the laryngeal feature *aspirated*

shows the test results. The networks trained on detecting the feature *aspirated* clearly predict the aspirated feature for h_1. For h_2, the model clearly rejects the feature *aspirated*. In the case of h_3, the statistical tests indicate that the laryngeal possessed the feature *aspirated*, however because of the thin difference in the number of predicted samples, we still need to treat this finding with caution, since the feature is not as clearly predicted for h_3 as it is for h_1. It is likely that the aspiration present in h_3 is weaker than or different from that of h_1.

	h_1		**h_2**		**h_3**	
H_1	W	p-value	W	p-value	W	p-value
P greater N	9991.00	< 0.00001	0	1	7477.00	< 0.00001
N greater P	9.00	1	10000	< 0.00001	2523.00	1

Table 10: Results of Wilcoxon signed rank test with continuity correction applied to the predictions for each laryngeal with H_1: positives greater than negatives and H_1: negatives greater than positives

3.5 Example 2: The internal coherence of PIE nasals

Besides predicting phonetic features of unknown sounds, the deep neural networks can moreover detect inconsistencies or idiosyncrasies in PIE. One such example is the feature *nasal* which is present in both PIE non-syllabic (*m, *n) and syllabic nasals (*m̥, *n̥). While both are regarded to be phonetically identical and only differing in their syllabicity (Clackson, 2007, 35), an investigation using the deep neural network approach gives some insights into their relationship to one another: To analyze this feature, a deep neural network was set up with the architecture displayed in table 11.

Layer	Layer size	Activation
Dense layer 1	128	ReLU
Dropout layer 1	0.25 dropout rate	
Dense layer 2	64	ReLU
Dropout layer 2	0.25 dropout rate	
Dense layer 3	32	ReLU
Output layer	2	softmax

Table 11: Network architecture for the feature *nasal*

The method used in this case is equal to the training and evaluation procedure of the model

used in 3.4. The resulting confusion matrices obtained after each evaluation of the 100 training runs are summarized in table 12.

	True positives	False negatives	False positives	True negatives
Mean	49.32	53.68	135.5	661.5
Median	48	55	118.5	678.5
Std. dv.	**6.689**		**38.951**	

Table 12: Statistics of the confusion matrices from 100 runs for classifying the feature *nasal*

As this summary shows, the neural network had more difficulties learning the properties of *nasal* than it had learning the feature *aspirated*. The classifier only detects the feature less than 50 percent of the time it is presented with nasal sounds, which is approximately what could be expected by randomly classifying the rest samples. Moreover, a Wilcoxon signed rank test with continuity correction with the alternative hypothesis H_1: True positives greater than false negatives gives W = 3144 p = 1. As a result, it was not possible to successfully train the network on this feature. Given the large discrepancy in performance between this and the previous network and the fact that both models were optimized using the same methods, the problem must be data inherent. This finding raises the question of why exactly this series differs from the other features. This leaves three possible explanations: (1) The data containing the nasals is noisier compared to the other phonetic features so that the classifier cannot train on a consistent set of properties. Although data can be varying degrees of noisy, it is unlikely that this feature is overly affected by noise. (2) The nasal feature was weakly articulated in PIE and thus it had little effect on its environment. An effect so small that it did not leave stable traces the classifier could detect. (3) The third explanation is that the nasal series does not possess internal coherence. This reason is arguably the most probable given that the nasals consist of two different sets of nasals that contrast in their syllabicity, especially since syllabic and non-syllabic resonants are also allophones and are therefore in complementary distribution (cf. Schindler, 1977). Yet since the model was trained on detecting nasality – not syllabicity – while there were other syllabic consonants in the non-nasal group, it is also possible that the model is not *solely* misled by the difference in syllabicity and their complementary distribution. There might also be a difference in nasality itself which results in the feature not forming a consistent, classifiable group. In other words, the syllabic and non-syllabic nasals might additionally have also differed in their nasality (i.e. nasality being differently articulated in both cases), yet this observation needs to be further investigated before one can make more substantiated claims.

4 Conclusion

As has been demonstrated in this paper, using deep neural networks in historical phonetics is a viable method to predict unknown features and to uncover previously unnoticed inconsistencies within a language's phonetic system. The tool is specifically powerful for historical linguistics since it does not rely on diachronic methods such as the comparative method to analyze and determine phonetic features but can draw upon synchronic phonetic patterns arising from coarticulation and statistical constraints. The results obtained through the machine learning technique presented in this paper are moreover reproducible and empirical, and can therefore be seen as complementary to previous results obtained by other empirical approaches such as the comparative method. However, the specific strengths and weaknesses of this method need to be further investigated in future research.

References

R Baayen, Richard Piepenbrock, and Léon Gulikers. 1995. Celex2 ldc96l14. *Web Download. Philadelphia: Linguistic Data Consortium*.

Gustavo EAPA Batista, Ana LC Bazzan, and Maria Carolina Monard. 2003. Balancing training data for automated annotation of keywords: a case study. In *WOB*, pages 10–18.

R. S. P. Beekes. 1994. Who were the laryngeals. In Jens Elmegård Rasmussen, editor, *In honorem Holger Pedersen*, pages 450–454. Reichert, Wiesbaden.

R. S. P. Beekes and Michiel de Vaan. 2011. *Comparative Indo-European linguistics: An introduction*, 2. ed. edition. Benjamins, Amsterdam.

Juliette Blevins. 2015. Evolutionary phonology: A holistic approach to sound change typology. In Patrick Honeybone and Joseph Curtis Salmons, editors, *The Oxford handbook of historical phonology*, Oxford handbooks in linguistics, pages 485–500. Oxford University Press, Oxford.

Allan R. Bomhard. 2004. The proto-indo-european laryngeals. In Adam Hyllested and Jens Elmegård

Rasmussen, editors, *Per aspera ad asteriscos*, Innsbrucker Beiträge zur Sprachwissenschaft, pages 69–80. Institut für Sprachen und Literaturen der Universität Innsbruck, Innsbruck.

Andrew Byrd. 2015. *The Indo-European Syllable*. Brill, Leiden.

Lyle Campbell. 2013. *Historical linguistics*. Edinburgh University Press, Oxford.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. *Towards Open Data for Linguistics: Linguistic Linked Data*, pages 7–25. Springer Berlin Heidelberg, Berlin, Heidelberg.

François Chollet. 2018. *Deep learning with Python*. Manning Publications Company, Shelter Island.

François Chollet et al. 2015. Keras. `https://keras.io`.

James Clackson. 2007. *Indo-European linguistics: An introduction / James Clackson*. Cambridge textbooks in linguistics. Cambridge University Press, Cambridge.

Patricia J. Donegan and Geoffrey S. Nathan. 2015. Natural phonology and sound change. In Patrick Honeybone and Joseph Curtis Salmons, editors, *The Oxford handbook of historical phonology*, Oxford handbooks in linguistics, pages 431–449. Oxford University Press, Oxford.

Benjamin W Fortson IV. 2011. *Indo-European language and culture: An introduction*, volume 30. John Wiley & Sons, Hoboken.

Carol A Fowler. 1980. Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 8(1):113–133.

Jost Gippert. 1994. Zur phonetik der laryngale. In Jens Elmegård Rasmussen, editor, *In honorem Holger Pedersen*, pages 455–466. Reichert, Wiesbaden.

Léon Gulikers, Gilbert Rattink, and Richard Piepenbrock. 1995. German linguistic guide. *The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, Philadelphia, PA*.

Mark Hale. 2003. Neogrammarian sound change. In Brian D. Joseph, editor, *The handbook of historical linguistics*, Blackwell handbooks in linguistics, pages 343–368. Blackwell, Malden, MA.

William J. Hardcastle and Nigel Hewlett. 2006. *Coarticulation: Theory, data and techniques*. Cambridge University Press, Cambridge.

Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of Conference: Volume 1: Long Papers,ACL*, pages 1204–1226.

Gerhard Jäger and Pavel Sofroniev. 2016. Automatic cognate classification with a support vector machine. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, Bochumer Linguistische Arbeitsberichte, pages 128–134.

Barbara Kühnert and Francis Nolan. 1999. The origin of coarticulation. In Nigel Hewlett and William J. Hardcastle, editors, *Coarticulation*, Cambridge studies in speech science and communication, pages 7–30. Cambridge University Press, Cambridge.

Martin Joachim Kümmel. 2007. *Konsonantenwandel: Bausteine zu einer Typologie des Lautwandels und ihre Konsequenzen für die vergleichende Rekonstruktion: Teilw. zugl.: Freiburg, Univ., Habil.-Schr., 2005*. Reichert, Wiesbaden.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Manfred Mayrhofer and Warren Cowgill. 1986. *Indogermanische Grammatik. Bd 1*. Winter, Heidelberg.

Michael Meier-Brügger, Matthias Fritz, and Manfred Mayrhofer. 2010. *Indogermanische Sprachwissenschaft*, 9., durchgesehene und ergänzte auflage 2010 edition. De Gruyter Studium. De Gruyter, Berlin.

Gottfried Meinhold and Eberhard Stock. 1982. *Phonologie der deutschen Gegenwartssprache*, second edition edition. Bibliographisches Institut, Leipzig.

Gerard de Melo. 2015. Wiktionary-based word embeddings. *Proceedings of MT Summit XV*, pages 346–359.

Christian M. Meyer and Iryna Gurevych. 2012. Wiktionary: A new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography. In Magali Gragner, Sylviana;Paquot, editor, *Electronic Lexicography*. Oxford University Press, Oxford.

Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, Hsieh ShuKai, Kuo Tzu-Yi, Pierre Magistry, and Huang Chu-Ren. 2009. Wiktionary and nlp: Improving synonymy networks. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic*

Resources, People's Web '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael A. Nielsen. 2015. *Neural Networks and Deep Learning*. Determination Press.

J. J. Ohala. 1993a. Coarticulation and phonology. *Language and speech*, 36 (Pt 2-3):155–170.

J. J. Ohala. 1993b. The phonetics of sound change. In Charles Jones, editor, *Historical linguistics*, Longman linguistics library, pages 237–278. Longman, London.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Uriel Cohen Priva. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, 6(2):243–278.

Jens Elmegård Rasmussen, editor. 1994. *In honorem Holger Pedersen: Kolloquium der Indogermanischen Gesellschaft vom 25. bis 28. März 1993 in Kopenhagen*. Reichert, Wiesbaden.

William Raymond, Robin Dautricourt, and Elizabeth Hume. 2006. Word-medial /t,d/ deletion in spontaneous speech: Modeling the effects of extralinguistic, lexical, and phonological factors. *Language Variation and Change*, 18:55—-97.

Donald A. Ringe. 2017. *From Proto-Indo-European to Proto-Germanic*, second edition edition, volume volume I of *A linguistic history of English*. Oxford University Press, Oxford.

Jochem Schindler. 1977. Notizen zum sieversschen gesetz. *Die Sprache*, 23(1):56–65.

Rob JJH Van Son and Jan PH Van Santen. 2005. Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, 47(1-2):100–123.

Michael Weiss. 2016. The proto-indo-european laryngeals and the name of cilicia in the iron age. In Andrew Miles Byrd, Jessica DeLisi, and Mark Wenthe, editors, *Tavet tat satyam*, pages 331–340. Beech Stave Press, Ann Arbor and New York.

Wiktionary contributors. Wiktionary:about proto-indo-european — Wiktionary, the free dictionary. Accessed: 2019-03-13.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Using wiktionary for computing semantic relatedness. In *AAAI*, volume 8, pages 861–866.

A Appendices

word boundary	zero
nasal	m, n, ŋ
plosive	p, t, d, k, g, b
affricate	pf, ts, tʃ
fricative	f, v, s, z, ʃ, ʒ, x/ç, h
liquid	l, r
rhotic	r
lateral	l
sibilant	s, z, ʃ, ʒ
voiced	m, n, ŋ, d, g, ʒ, v, l, r, z, b
voiceless	p, t, k, pf, tʃ, ʃ, s, f, x/ç, h
labial	m, p, pf, f, v, b
bilabial	m, p, b
labiodental	pf, f, v
dental/alveolar	n, t, d, ts, s, z, l, r
palatal	ʃ, ʒ, tʃ, r
velar/uvular	ŋ, k, g, x/ç, r
glottal	h
obstruent	p, b, t, d, k, g, pf, ts, tʃ, f, v, s, z, ʃ, ʒ, x/ç, h, r
sonorant	m, n, ŋ, ɪ, i, ɛ, e, y, ʏ, ø, œ, æ, ə, a, ɑ, u, ʊ, o, ɔ, a͡ɪ, a͡ʊ, ɔ͡ʏ
occlusive	p, b, t, d, k, g, m, n, ŋ, pf, ts, tʃ
continuant	f, v, s, z, ʃ, ʒ, x/ç, h, r, ɪ, i, ɛ, e, a͡ɪ, ɔ͡ʏ, y, ʏ, ø, œ, æ, ə, a, ɑ, u, ʊ, o, ɔ
consonant	m, n, ŋ, d, g, ʒ, v, l, r, b, p, t, k, pf, tʃ, ʃ, s, z, f, x/ç, h
front vowel	ɪ, i, ɛ, e, y, ʏ, ø, œ, æ
central vowel	ə, a, ɑ
back vowel	u, ʊ, o, ɔ
close	i, ɪ, u, ʊ, y, ʏ
mid	ɛ, e, ə, o, ɔ, ø, œ, æ
open	a, ɑ
diphthong	a͡ɪ, a͡ʊ, ɔ͡ʏ
open diphthong	a͡ɪ, ɔ͡ʏ
mid diphthong	a͡ʊ
front diphthong	a͡ɪ, a͡ʊ
back diphthong	ɔ͡ʏ
round	o, ɔ, y, ʏ, ø, œ
unround	ɪ, i, ɛ, e, ə, a, ɑ, u, ʊ, æ
long	i, e, a, u, o, æ, œ, y
short	ɪ, ɛ, ʊ, ɑ, ʏ, ø, ə

Table 13: Phonetic feature assignment of each considered New High German sound

root ending	-
word boundary final/initial	zero
voiced	*bʰ, *dʰ, *ǵʰ, *gʰ, *gʷ, *b, *d, *ǵ, *g, *gʷ, *m, *ṃ, *n, *r, *l, *l̥, *r̥, *y, *w
voiceless	*p, *t, *s, *ḱ, *k, *kʷ
nasal	*m, *ṃ, *n, *n̥
aspirated	*bʰ, *dʰ, *ǵʰ, *gʰ, *gʷʰ
labial/labialized	*m, *ṃ, *p, *b, *bʰ, *w, *kʷ, *gʷ, *gʷʰ
sibilant	*s
liquid	*r, *r̥, *l̥, *l
syllabic	*r̥, *l̥, *ṃ, *n̥, *i, *u, *ū
coronal	*n, *n̥, *t, *d, *dʰ, *s, *r, *l, *l̥, *r̥
postvelar	*k, *g, *gʰ, *kʷ, *gʷ, *gʷʰ
velar	*ḱ, *ǵ, *ǵʰ
palatal	*y
front vowel	*e, *ē, *i
back vowel	*o, *ō, *u, *ū
center vowel	*a, *ā
short vowel	*e, *o, *u, *a, *i
long vowel	*ē, *ō, *ū, *ā
open vowel	*a, *ā
close vowel	*u, *ū, *i
laryngeal 1	*h_1
laryngeal 2	*h_2
laryngeal 3	*h_3
unspecified laryngeal	*H
consonant	*bʰ, *dʰ, *ǵʰ, *gʰ, *gʷ, *b, *d, *ǵ, *g, *gʷ, *m, *ṃ, *n, *r, *l, *l̥, *r̥, *y, *w, *p, *t, *s, *ḱ, *k, *kʷ
back consonant	*k, *g, *gʰ, *kʷ, *gʷ, *gʷʰ, *ḱ, *ǵ, *ǵʰ
front consonant	*m, *ṃ, *p, *b, *bʰ, *w, *s, *r, *r̥, *l̥, *l, *n, *n̥, *t, *d, *dʰ
stop	*ḱ,*b, *bʰ, *p, *ǵ, *ǵʰ, *k, *g, *gʰ, *kʷ, *gʷ, *gʷʰ, *t, *d, *dʰ
obstruent	*ḱ, *p, *b, *bʰ, *ǵ, *ǵʰ, *k, *g, *gʰ, *kʷ, *gʷ, *gʷʰ, *t, *d, *dʰ, *s
sonorant	*m, *ṃ, *n, *n̥, *r, *r̥, *y, *w, *e, *o, *u, *a, *i, *ē, *ō, *ū, *ā
occlusive	*ḱ, *p, *b, *bʰ, *ǵ, *ǵʰ, *k, *g, *gʰ, *kʷ, *gʷ, *gʷʰ, *t, *d, *dʰ, *m, *ṃ, *n, *n̥
continuant	*s, *y, *w, *e, *o, *u, *a, *i, *ē, *ō, *ū, *ā

Table 14: Phonetic feature assignment of each considered PIE sound