BRILL

INDO-EUROPEAN LINGUISTICS 9 (2021) 26–84

brill.com/ieul

The phonetic value of the Proto-Indo-European laryngeals

A computational study using deep neural networks

Frederik Hartmann | ORCID: 0000-0002-6649-5371
University of Konstanz, Konstanz, Germany
frederik.hartmann@uni-konstanz.de

Abstract

Discussion of the exact phonetic value of the so-called 'laryngeals' in Proto-Indo-European has been ongoing ever since their discovery, and no uniform consensus has yet been reached. This paper aims at introducing a new method to determine the quality of the laryngeals that differs substantially from traditional techniques previously applied to this problem, by making use of deep neural networks as part of the larger field of machine learning algorithms. Phonetic environment data serves as the basis for training the networks, enabling the algorithm to determine sound features solely by their immediate phonetic neighbors. It proves possible to assess the phonetic features of the laryngeals computationally and to propose a quantitatively founded interpretation.

Konstanzer Online-Publikations-System (KOPS)
URL: http://nbn-resolving.de/urn:nbn:de:bsz:352-2-1h7u0x447h8wy8

Keywords

Proto-Indo-European – laryngeals – machine learning – deep neural networks

1 Introduction

Ever since Ferdinand de Saussure (1879) proposed to reconstruct a series of additional sounds for Proto-Indo-European (PIE) on the basis of indirect reflexes in the daughter languages, the phonetic nature of those sounds has remained a puzzle of Indo-European linguistics. What has come to be known as the "laryngeal theory" is almost universally accepted today, and the scientific community has identified three 'laryngeals', written h_1, h_2, and h_3, as well

as their phonological properties and effects on other sounds and ablaut patterns in the daughter languages of PIE.[1] Originally coined by early proponents of the theory (e.g., Møller 1917), the term 'laryngeal' is in fact a misnomer, as their phonetic quality is disputed to this day and it has only been possible to show that they were consonants (Cuny 1912; Fortson 2010: 62–64). This is especially remarkable given that we understand their phonological and morphophonological functions in great detail. Despite their prominence, there is still no consensus on the details of the laryngeals' phonetic properties.

Previous investigations have predominantly focused on comparing the outcomes in different daughter languages and made arguments based on typological considerations. While the comparative approach has had great success in identifying their phonotactics and many properties of the laryngeals, findings obtained by computational methods have not yet entered the scholarly debate on this topic. In an attempt to provide the discussion with such a computational perspective, the study at hand aims to employ computational and statistical means to obtain an approximation of the phonetic[2] properties of the laryngeals while also being repeatable. The methodological foundation of this paper is deep neural networks, a sub-domain of machine learning algorithms which are best known for their application in bioinformatics, natural language processing, and speech and image recognition. Among the many methods currently in use in computational linguistics, deep neural networks are ideal for the task of approximating the laryngeal phonetic features since they are able to detect ('learn') complex patterns and properties of a given dataset and to make predictions on the basis of these patterns afterwards.

It needs to be stressed that this method of investigation is not intended to replace traditional methods of comparative reconstruction and phonotactic analysis. Instead, it aims at presenting and conducting a novel approach which is able to provide a different perspective on the phonetics of the laryngeals. This perspective is meant to complement the various traditional approaches, not least as without those approaches, the data and theoretical background necessary to conduct this study would not exist. Ideally, this investigation will cap-

1 For an extensive discussion of the literature on laryngeals, see Kümmel (2007: 327–36).

2 The use of the term *phonetic* in this context calls for comment. The input values to the models used in this study are binary to enable computational processing and thus follow a phonological classification system. However, they also contain redundant features and yield results which require phonetic interpretation. Moreover, the theoretical background of this study makes use of phonetic principles and concepts such as coarticulation and local predictability. Thus, although the term *phonetic* is used here, one must bear in mind that this study is neither purely phonetic nor exclusively phonological.

ture previously undiscovered connections and patterns about certain aspects of laryngeal phonology which can in turn provide a basis for future inquiry into the laryngeals with traditional means.

The discussions and hypotheses about each of the laryngeals can be briefly summarized as follows:

1.1 *h_1

The first laryngeal seems to be the most elusive of the series insofar as it is not securely continued as a consonantal segment in any PIE daughter language.[3] However, if the assumption is correct that the laryngeals themselves were vocalized in later stages, *h_1 does have a vocalic reflex in some daughter languages (see also section 1.4). Additionally, it seems to have had a rather weak effect on its environment. Thus, it has been argued that *h_1 was a somewhat 'neutral' laryngeal since it left adjacent *e unaffected, as opposed to the coloring effects of *h_2 and *h_3, apparent especially in Greek (cf. Kümmel 2007: 333–34; Beekes & de Vaan 2011: 147). It is not clear, however, if the laryngeal was aspirated, mainly because there is some disagreement on the possibility of adjacent aspiration (cf. Rasmussen 1994: 436–37; Kümmel 2007: 333). Most researchers posit either a glottal stop ([ʔ]) (e.g., Beekes 1994; Beekes & de Vaan 2011; Bomhard 2004; Gippert 1994) or a voiceless aspirate ([h]) (e.g., Kümmel 2007; Meier-Brügger, Fritz, & Mayrhofer 2010; Rasmussen 1994) as the phonetic realization of *h_1.

1.2 *h_2

The second laryngeal *h_2 is directly attested in Anatolian, e.g., as Hittite $ḫ$, -$ḫḫ$-, which was either pharyngeal, velar/uvular, or glottal. Its most prominent feature is causing a-coloring of an adjacent *e in the IE daughter languages (e.g., Greek). The realization in Anatolian was likely either pharyngeal, velar/uvular, or glottal.[4] In addition, *h_2 could aspirate a preceding voiceless or voiced stop. Its phonetic realization is assumed to be a voiceless velar/uvular fricative [[x/χ]] (cf. Rasmussen 1994; Kümmel 2007; Meier-Brügger, Fritz, & Mayrhofer 2010) or a pharyngeal [ʕ] (cf. Beekes 1994; Bomhard 2004), with more scholars favoring the former interpretation.

3 Pace Kloekhorst (2004; 2006) for a different view on the realization of *h_1 in Hittite and Luwian.

4 See Kümmel (2007) and Weiss (2016) for an overview of discussion on the realization of the Anatolian reflex.

1.3 **h₃*

**h₃* is directly attested only in Anatolian, e.g., as Hittite *ḫ-*, and had an *o*-coloring effect on adjacent **e*. This fact, among other reasons, has contributed to the often suggested interpretation as a sound with lip-rounding. Due to its observed sonorizing effect in the present stem **pi-ph₃-e/o-* > **pibe/o-* 'drink', it is often considered to be the voiced counterpart of voiceless **h₂* (cf. Kümmel 2007: 332; Rasmussen 1994: 435). It is therefore believed to be realized as either a voiced labialized velar fricative [ɣʷ] (Rasmussen 1994), a labialized pharyngeal [ʕʷ] (Beekes 1994), or a uvular fricative [ʁ] (Kümmel 2007).

1.4 *Vocalization of laryngeals*

All of the laryngeals could be vocalized in some positions (see Cowgill & Mayrhofer 1986: 121–50; Beekes & de Vaan 2011), as can be observed in the Greek triple reflex of **H* in #_C positions where the three laryngeals are preserved as three different vowels, e.g., PIE **h₁rud^{h}rós* > Gr. ἐρυθρός 'red', PIE **h₂stér-* > Gr. ἀστέρ 'star', PIE **h₃neb^{h}-* > Gr. ὀμφαλός 'navel'. However, it is unclear to what degree vocalization was an inherent property of the laryngeals. Some scholars have gone as far as positing a vocalized set of laryngeals alongside the consonantal reflexes (e.g., Rasmussen 1994: 440). More probable seems to be a vocalization of **H* in the daughter languages by insertion of a vowel which was itself not present in PIE (see e.g., Meier-Brügger, Fritz, & Mayrhofer 2010: 236–54).

2 Method and prerequisites

2.1 *Local predictability through phonetic constraints*

As mentioned above, previous investigations into the phonetics of the PIE laryngeals have relied on their reflexes in the daughter languages, above all Anatolian. There has also been some work building upon findings about pharyngeals/uvulars in Arabic, taken as a proxy for the potential situation in PIE, to investigate the PIE laryngeals (see e.g., Sanker 2016). But what sort of method would enable us to infer the phonetics of the laryngeals directly from PIE? It would require a technique which operates synchronically on the basis of the reconstructed proto-language. Of course, any analysis of PIE sounds will be ultimately based on previous research using the comparative method on the daughter languages. It is not the goal of this study to claim independence from comparative results, but rather to begin the analysis at the level of reconstructed PIE directly.

Studies in quantitative phonetics have demonstrated the existence of local predictability, i.e., the predictability of sounds based on features of immedi-

ately surrounding sounds (see e.g., Cohen Priva 2015; Van Son & Van Santen 2005; Raymond, Dautricourt, & Hume 2006; K.C. Hall et al. 2018). This predictability is reflected in both *absolute* and *statistical* constraints on sound patterns and co-occurrences. It is important to differentiate between these *absolute* and *statistical* constraints: Absolute constraints are, e.g., syllable composition constraints or the prevention of certain consonant clusters, which make up a language's phonotactics, whereas statistical constraints involve a strong dominance of one phonological form or pattern.

One of the origins of such statistical constraints, which as just mentioned are more subtle *tendencies* and sound patterns than absolute constraints, is coarticulation. In modern phonetics, the phenomenon of coarticulation has mostly received attention in synchronic studies of various languages (see, e.g., Kühnert & Nolan 1999; Ohala 1993a; Hardcastle & Hewlett 2006; Fowler 1980). The term "coarticulation" refers to the observation that adjacent sounds influence each other's articulation to some extent. This results in a complex system of codependent articulation where different sounds exhibit varying susceptibility depending on the phonetic environment. Two variants of coarticulation are commonly distinguished: anticipatory and carryover (see Bybee 2015). The former term refers to coarticulation in which a sound receives its articulatory features from the following sound, whereas the latter describes a sound which receives its articulatory features from the preceding sound. Over time, coarticulation leads to an interlaced system of mutual articulatory influence among the sounds of a language dependent on the respective environments. The assumption is that over long time periods, these coarticulatory influences become inherent features of the respective sound or at the very least favor sound changes in accordance with the respective environments (compare Donegan & Nathan 2015; Blevins 2015; Ohala 1993a; Ohala 1993b; Hale 2003). Note however that coarticulation is only one factor yielding phonotactic patterns and constraints.

Several of these phonotactic and co-occurrence effects have already been identified in PIE, mainly in the form of numerous root and syllabic constraints as well as constraints on the possible combinations of adjacent sounds (compare Clackson 2007: 64–71; Meier-Brügger, Fritz, & Mayrhofer 2010: 272–75; Byrd 2015; Ringe 2017: 13–17; Fortson 2010: 62–64). Many studies have already investigated phonotactic constraints, co-occurrences of sounds, and environmental effects in PIE, both statistical and absolute. Some researchers take a quantitative approach, such as Cooper (2009), who analyzes the concept of similarity avoidance as proposed in Frisch, Pierrehumbert, & Broe (2004). Cooper identifies several co-occurrence constraints of PIE sounds on the root level, including constraints against laryngeal pairings. Other studies have taken

a more general approach with quantitative elements such as Iverson & Salmons (1992). Regarding the laryngeals in particular, previous research has established that the laryngeals occur in environments associated with fricatives (see, e.g., Byrd 2017).

For unknown sound qualities of unattested languages such as the PIE laryngeals, local predictability could help determine their quality. Therefore, once it can be established that local predictability and statistical constraint effects exist in PIE, the phonetic properties of the unknown sounds can be predicted from the environment using machine learning techniques. It has to be noted, however, that the formation of these phonotactic patterns might have taken place at some point before the latest reconstructible stage of PIE. This is relevant to this study insofar as the data is drawn from the latest reconstructible stage of PIE, whereas some phonotactic patterns might reflect an earlier stage of this language. The results therefore have to be further interpreted to allow for this possibility.

2.2 *The data*

The data was extracted from the English Wiktionary .xml dump on 20.10.2018, which involves a complete history of all English Wiktionary articles alongside their edit history and discussion sites. Only those PIE reconstructions were used which appeared in the headings of existing pages, so as to include only established reconstructions that have their own entry.

The use of Wiktionary as a data source calls for some comment. There are three main benefits to using Wiktionary for historical linguistic data, especially for reconstructed languages. First, the online dictionary is a collection of known reconstructed lexical items of PIE not limited to those reconstructions that a traditional dictionary provides and thus has a broader scope. Secondly, the reconstructions are regularly maintained, checked, and updated according to Wiktionary reconstruction guidelines that adhere to the latest scientific research.[5] Therefore, it is up-to-date and internally consistent, in contrast to word lists that one could alternatively craft by merging reconstructions from different traditional dictionaries. Lastly, the data are digital and therefore easy to parse and to include in a corpus.

Regarding its reliability, the database has already been used and assessed as a valid foundation for quantitative linguistic research on languages other than PIE (e.g., Zesch, Müller, & Gurevych 2008; Navarro et al. 2009; Meyer &

5 https://en.wiktionary.org/wiki/Wiktionary:About_Proto-Indo-European, accessed: 13 March 2019.

Gurevych 2012; Chiarcos et al. 2013; De Melo 2015). To check the validity of the data, 10 percent of the entries were randomly selected beforehand and compared with the current state of reconstruction (e.g., Mallory & Adams 2006; Ringe 2017). No systematic discrepancies in the reconstructions were detected. In those entries where individual discrepancies were found, this was due to different reconstruction traditions or "schools". For example, the entry **albʰós* 'white' also contains the note that an alternative reconstruction of this word is **h₂elbʰós*. In these ambiguous cases, I selected the main entry for consistency. Regarding the accuracy of Wiktionary as a data source beyond the reconstructions checked here, no evidence was found that PIE reconstructions from the Wiktionary dump used in this study are affected by any systematic error or malicious tampering.

Note that these entries mainly reflect the underlying form of the reconstruction, with the exception of the syllabic surface forms of /j/ and /w/, following the standardized notation Wiktionary provides. Therefore, allophonic variations such as simplification of **ss* or thorn clusters are not represented in the data. For this study, the data were left unedited to account for this lack of allophonic information, since these are generally effects specific to particular phonemes or classes of phonemes. In the worst case, a neural network would interpret them as a general pattern and its subsequent predictions would be inaccurate.

To outline the particular problem with surface forms, assume the two following hypothetical trigrams: (1) V T V and (2) V D V where V is a vowel, D is a voiced consonant and T stands for a voiceless consonant. If we further assume that T possesses a voiced allophone D in intervocalic position, the underlying trigram (1) /V T V/ would result in a surface representation [V D V], whereas (2) would also result in [V D V]. If we further trained the network to determine the feature [voice] from the environment using only the surface forms, it would generalize this pattern, which in itself is valid in this case. However, if we then used the trained network to predict the feature [voice] for an unknown sound *H* of which we are only certain that it also occurs intervocalically, we would not be able to discriminate whether H was underlyingly voiced or only voiced in this environment due to the phonological rule. Using underlying representations yields a model basing its predictions more on phonotactic patterns in such cases.

Although this outlines a specific phenomenon which does not apply in the current dataset, replacing the underlying reconstructions with reconstructions including such surface alternations risks increased noise in the data and inaccuracies in the predictions due to such phenomena. Although this risk is lower for surface-transparent patterns such as assimilatory voicing or laryngeal vowel

coloring, the entries were left unedited so as not to introduce a dataset bias by including partial surface representations. However, it has to be acknowledged that since this decision was made beforehand and is based on the theoretical assumptions outlined above, it cannot be ruled out that including the surface forms would have yielded different results with regard to the prediction of laryngeal features.

Furthermore, it must be noted that the Wiktionary dataset consists of both roots and inflected forms. To account for the difference between roots and inflected words, a segmental feature denoting root ending was added whenever a segment shows a root ending in its right periphery (see below). This prevents the networks from treating word-final and root-final segments equally.

After extracting the lemmas from the dump, I split each lemma into segments of three sounds: preceding sound, target sound and following sound. Where the trigram contained a root ending, '-' was used as following sound to encode the root ending. Word-final or word-initial position was added as 'zero' in the slot for preceding or following sound, respectively. The total count of entries extracted from Wiktionary was 1483.

The decision to favor trigrams over a wider scope such as five-grams was made on the basis of preliminary testing: in the early stages of the study, different models were tested which were fed different n-grams. While the different n-gram sizes tested did not alter the results for the application of the model to German (see section 3.1), the Indo-European models suffered loss of accuracy if five-grams were fed instead of trigrams. The reasons for this might be (1) added noise due to the doubling of the input signal for each token in relation to the small size of the PIE dataset, and (2) the proportional increase in heterosyllabic and long-distance effects over adjacency and coarticulatory effects. The latter might have led to a shift in focus during the training process and yielded inaccurate predictions more frequently. This issue could have been resolved by the network during training if the corpus size for PIE had been similar to that of the German corpus used in the test on German. Given that the model evaluation for the networks trained on trigrams was successful and the approach was cross-validated on models trained on German trigrams, one can be confident that the results obtained from trigram data for PIE will be reliable. A more detailed investigation into the phonological long-distance effects of PIE and their influence on deep neural network training would need to be the subject of a future study.

Thereafter, each of these trigrams was labeled with an ID number specifying the lemma from which it was extracted. Each sound was then classified according to its place and manner of articulation following the traditional inventory matrices of the field (e.g., Clackson 2007: 34; Beekes & de Vaan 2011: 119; Ringe

2017: 8).[6] In accordance with the consensus, the sounds of PIE were assigned to the categories shown in Table 39, included in the Appendix.

There is some consensus that the PIE 'palatals' were in fact plain, unmarked velars, while the PIE 'velars' were pronounced further back as uvulars or postvelars and the 'labiovelars' might have represented the labialized form of the 'velars' (see Kümmel 2007: 310–27; Ringe 2017: 9).[7] In the classification of PIE sounds, I decided to adhere to this consensus and to move away from the traditional nomenclature. Therefore in this paper, the term 'velar' will only refer to the series **ḱ, *ǵ, *ǵʰ* and the term 'postvelar' will refer to the two series **k, *g, *gʰ*, and **kʷ, *gʷ, *gʷʰ*. Following Kümmel (2007: 318–19), I assume that the 'postvelars' reflect either backed velars or uvulars.

Contrary to the conventions of distinctive feature notation, the features "voiced" and "voiceless" were encoded as *voiceless* for [-voice] and *voiced* for [+voice] only in consonants and vocalized/syllabic consonants. This is due to the set-up of the analysis where this feature is only intended to apply to consonants, since otherwise the feature [+voice] would not only encode voiced consonants but also all vocalics, which would decrease the discriminatory power of the algorithm to specifically differentiate between voiced and voiceless consonants.

All vocalics were encoded without regard to whether they are stressed or unstressed in the particular word from which they were extracted. Moreover, the category *syllabic* contains only syllabic consonants so as to create a category which makes it possible to train the model to specifically detect this phonotactic subgroup. The selection of tested features was compiled exclusively from known articulation manners and places of PIE sounds. It should be noted that the features [±fricative], [±sibilant], and [±palatal] could not be tested since each feature is represented by only one sound, and the nature of this task requires features to be found in at least two different sounds in order for the network to draw on the common properties of these features. Therefore, if the laryngeals were tested for the feature [±fricative] and since there is only one fricative (PIE **s*) in the inventory except, potentially, the laryngeals themselves, such a test would not determine whether the laryngeals were fricatives but rather whether they are identical with the tested sound **s*.

6 The glottalic theory was not considered for this approach.

7 Note that it is still debated whether the 'labiovelars' were in fact labialized counterparts of the 'velars'. An extensive discussion of this matter is provided by Kümmel (2007: 319–24). As this debate cannot be reviewed here, I have opted for the phonetic interpretation of the dorsal series as velar - postvelar - labiopostvelar (as also found in Ringe 2017: 9), in part because traditional reconstructions posit a closer relationship between the 'velars' and 'labiovelars'.

Regarding the selection of feature categories, it might seem counter-intuitive to include the three laryngeals along with the unknown laryngeal character **H*, considering that their phonetic features are unknown. Yet in the deep neural network approach, the goal will be to predict *which* of these categories found in the phonetic environment of a sound require said sound to be of a certain quality, whereby the actual phonetic realisation is not important. For example, if in a hypothetical environment **H_xSe* the sound *S* is always palatal, we do not need to know the phonetic quality of the laryngeal to predict a palatal for this environment.

Each category was then added as phonetic feature for each sound slot and encoded as *1* or *0*, so that each sound had 23 columns denoting each individual feature and was automatically classified as *1* when it possesses the particular feature and *0* when it does not. The total count of trigrams was 6236 with non-laryngeals as target sound, 199 with **h_1* as target sound, 348 with **h_2* as target sound, and 106 with **h_3* as target sound.

2.3 *Local predictability and statistical constraint effects in* PIE

To demonstrate that local predictability and statistical constraint effects applied to PIE can still be recovered so that a deep neural network analysis can be conducted in the first place, I calculated the distances among the PIE sounds in the data on the basis of their phonetic environment (preceding sound and following sound) using Spearman's rho. The output of such a calculation is a matrix of the distance between each pair of target sounds on the grounds of how similar they are with regard to their phonetic environment. Thereafter, I applied multidimensional scaling (MDS) to plot the distances in two-dimensional space, which can be seen in Fig. 1. For this analysis, I used the algorithms *distanceMatrix* and *cmdscale* from the R-packages *classDiscovery* (Coombes 2018) and *stats* (R Core Team 2013), respectively.

As we can observe from Fig. 1, the method clustered phonetically similar sounds together in almost all cases. We find a large distance between vocalics on the right-hand side and consonants on the left. The vocalics are furthermore distinguished on a vertical axis from syllabic sonorants, with **i*, **$\underset{\circ}{r}$*, and **$\underset{\circ}{l}$* on the upper end and 'true' vowels below. The consonants on the left side can be partitioned into stops on the lower left side, with mostly aspirated stops at the bottom. Sonority and friction increases on a center left to top right trajectory, with sonorants and semi-vowels at the upper end. At the rightmost end of the consonants are the laryngeals, vertically aligned with **s* and the sonorants and horizontally aligned with the non-aspirated stop series. The fact that even this trivial distance-based approach projected the sounds rather homogeneously on a 2-D plane is remarkable given that the actual features of each target sound

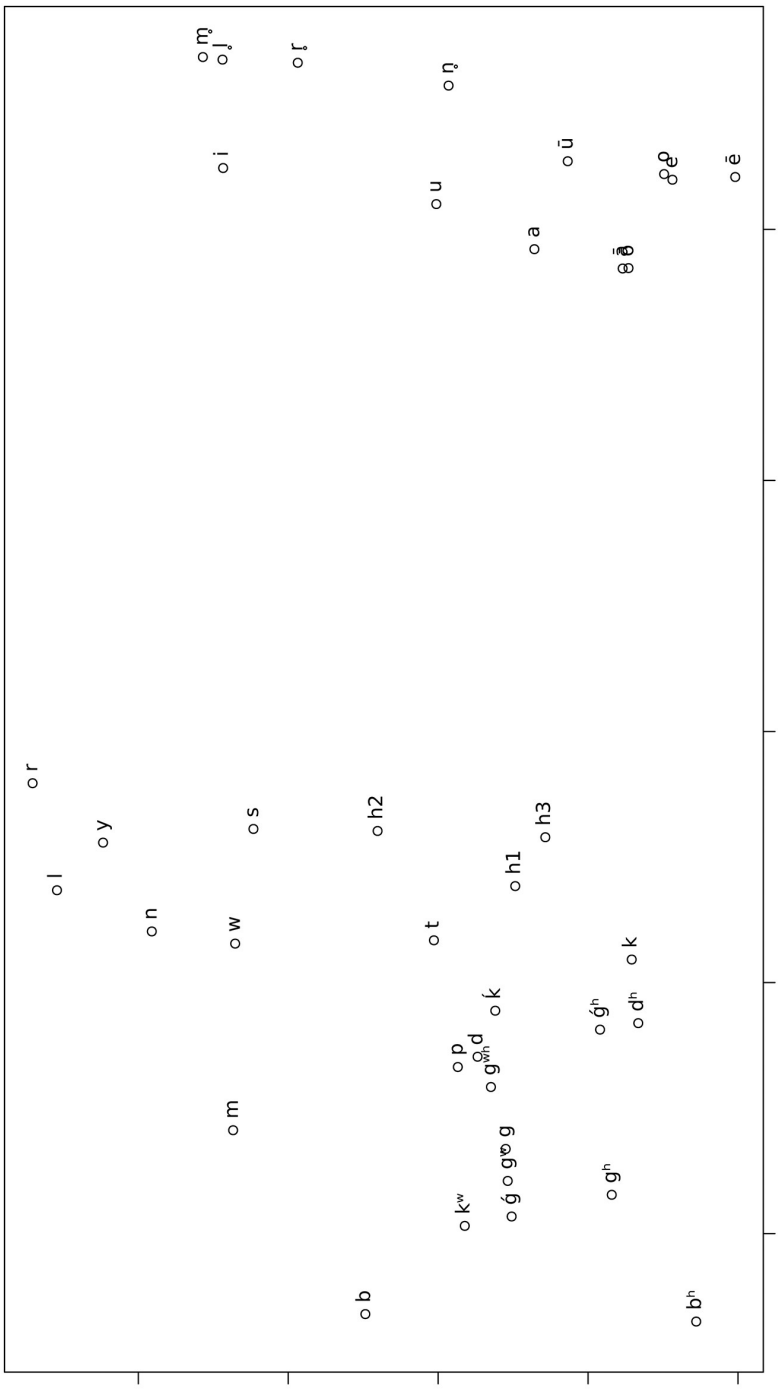

FIGURE 1 2-D scaled positions of PIE sounds according to the distances of their phonetic environments

were not used to compute these distances. It is a strong indicator of the existence of coarticulatory and statistical constraint effects that it is possible to computationally group sounds into correct subsets based only on the phonetic features of adjacent sounds. If there were no such effects in PIE at all, we would see a much more random distribution among the sounds in the network. To illustrate this, I randomized the sound-environment assignments in the data, computed the distances, and applied multidimensional scaling. The result can be seen in Fig. 2. As expected, the distribution of the sounds is much more random, with no clearly distinguishable cluster patterns.

In addition to MDS, I further analyzed the data using hierarchical clustering methods. The aforementioned distance data were clustered hierarchically with the clustering method Ward's criterion as implemented as *ward.D2* in the R-package *pvclust* (Suzuki & Shimodaira 2015). The number of bootstrap replications was 100,000. This procedure yields a cluster dendrogram including the Approximately Unbiased (AU) p-value and BP (Bootstrap Probability) for each cluster. This means that the higher the p-value for a given cluster, the better the cluster is supported by the data and the more confidently one can assume the cluster not to be a random finding. The results of this clustering are displayed in Fig. 3.

Those clades which exhibit an AU value higher than 0.95 are highlighted. From this plot we can observe that seven clusters are strongly supported by the data. It must be noted that those clusters which are not significant do not need to be actual clusters. For example, although **k* and **g^h* are grouped together in the dendrogram, since this clade is not significant, those two sounds likely do not form a clade and this grouping must be discarded. The two major clusters split the sounds into vocalics and "true" consonants. Within the vocalics, we also find a cluster consisting of **e*, **o*, **$\bar{e}$*, and **$\bar{o}$*. This finding is unsurprising, given these vowels are the most common in PIE and regularly participate in patterns of alternation or ablaut (see, e.g., Clackson 2007: 71–75; Ringe 2017: 12–13). The consonants are subdivided into three strongly supported clusters: one encompassing all stops and **m*, one containing the laryngeals and **s*, and one including all sonorants except **m*. It is worth noting that the stop cluster cannot confidently be subdivided further, which is indicative of a lack of clear-cut hierarchies among those sounds. This does not mean that there are no further subclusters to be found, but rather that none of the possible subdivisions are strongly supported by the data as to be statistically significant. We can expect a rather coarse method such as hierarchical clustering to correctly identify the major clusters (e.g., vowels, stops, liquids) while being less reliable for sound clusters which are less different from one another (e.g., voiced stops, velars, coronals).

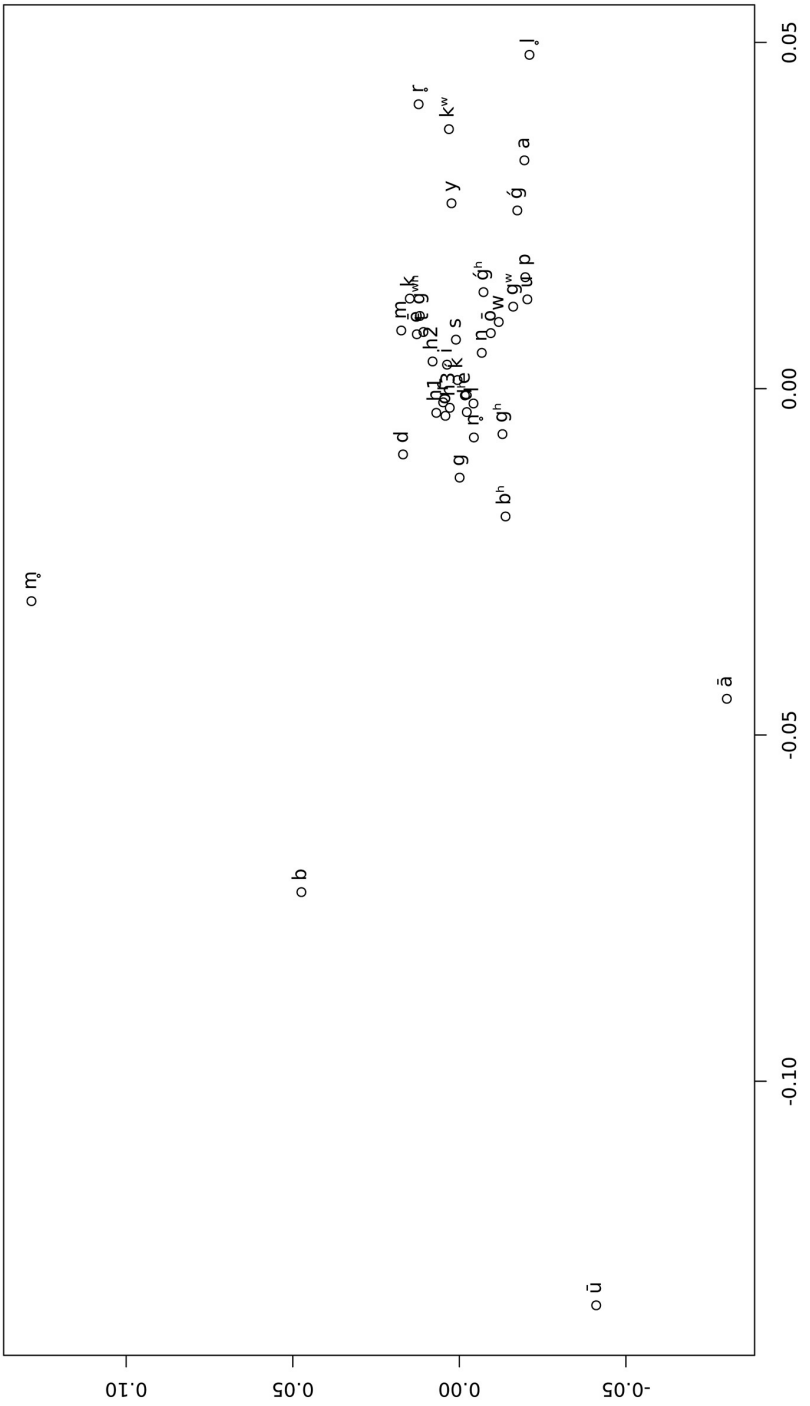

FIGURE 2 2-D scaled positions of randomly assigned PIE sounds according to the distances of their phonetic environments

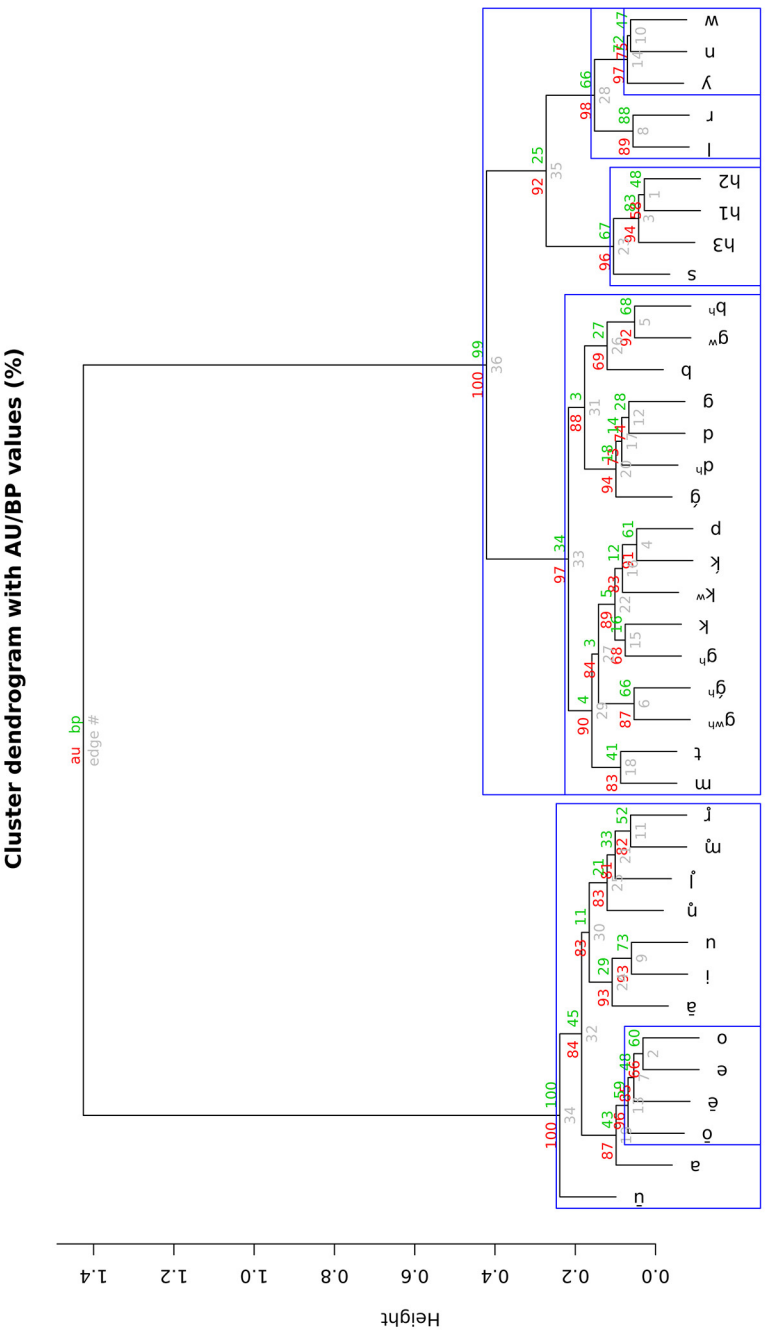

FIGURE 3 Hierarchical clustering result with highlighted clusters above $p > 0.95$ confidence

The one surprising finding in this clade is **m*, which would be expected to be more strongly associated with the sonorants rather than with the stops. Since the probability of this being a random outlier is relatively low given that the clade is strongly supported by the data, this clustering might be an indication that **m* phonotactically differs from other sonorants. An inconsistency within the PIE nasal series was already detected in Hartmann (2019). Although this is an accidental finding, it ties in with recent discussions on the sonority of PIE */m/ in the context of the PIE sonority hierarchy.[8]

The third clade containing the laryngeals and **s* shows that the laryngeals appear in environments similar to those where **s* is found. This observation is hardly new (see, e.g., Byrd 2017: 2064); however, it replicates the findings displayed in Fig. 1 that the laryngeals can be associated with increased friction. The third consonantal clade, the sonorants, are subdivided into the nasal **n* plus semi-vowels on the one hand and the liquids **l* and **r* on the other.

As an approach to demonstrating coarticulatory and statistical constraint effects on a feature level, I conducted a generalized linear logistic regression analysis contained in the R-package *lme4* (Bates et al. 2015). The goal was to explore the statistically observable influences of phonetic features of the environment on the target sound and vice versa. To achieve this, the generalized linear logistic regression analysis was set up with the *postvelar* feature of the target sound as dependent variable and the environmental features as independent variables in the form of binary vectors (the feature *postvelar* was only chosen as an example; all other features are equally suited to demonstrate these effects as well). Before the model was fitted, the laryngeal trigrams were excluded to only have sounds of known quality in the dataset in order to avoid potential interference from laryngeal environments. In the full model, aliases and multicollinearity effects were detected, and the affected variables were removed. Collinear variables were removed up to a cutoff-point of Variance Inflation Factor (VIF) greater than 4. The best fit model was chosen by AIC comparison through contrasting top-down and bottom-up comparative model fitting. The coefficients of this final model are listed in Table 1.

The interpretation of the effects of the single independent variables is as follows. *Estimate* gives the log odds of the feature *postvelar* occurring in the target sound when the given feature is present, e.g., if a central vowel follows, the target sound is by log 1.671 percent (= 5.317 percent) more likely to exhibit the feature *postvelar*. Accordingly, the feature *postvelar* is by log –2.052 percent (= –7.783 percent) more likely (i.e., less likely) to occur in the target sound if a sound

8 See Cooper (2013; 2015) in favor of a lower sonority of */m/, recently reassessed by Zair (2018).

TABLE 1 Generalized linear logistic regression for the occurrence of the feature *postvelar*

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−2.130	0.074	−28.821	0.000
aspirated preceding	−1.602	0.589	−2.720	0.007
labial preceding	−0.666	0.214	−3.106	0.002
liquid preceding	−1.136	0.347	−3.272	0.001
palatal preceding	0.551	0.244	2.260	0.024
postvelar preceding	−2.616	1.007	−2.598	0.009
mid vowel preceding	−0.324	0.140	−2.324	0.020
*h_1 preceding	−2.570	1.007	−2.553	0.011
*h_2 preceding	−1.244	0.421	−2.955	0.003
*h_3 preceding	−2.054	1.010	−2.034	0.042
nasal following	−0.866	0.393	−2.202	0.028
labial following	−2.052	0.588	−3.487	0.000
sibilant following	−0.764	0.257	−2.969	0.003
central vowel following	1.671	0.293	5.698	0.000
plosive following	−1.957	0.343	−5.707	0.000
*h_1 following	−2.138	1.009	−2.119	0.034
*h_2 following	−1.110	0.420	−2.643	0.008

with the feature *labial* follows. Although this analysis was only preliminary and only conducted on one feature as dependent variable, the high number of significant coefficients is nevertheless indicative of the strong mutual influences of sound features and their environmental features in PIE. Moreover, these effects are not random coincidences of statistical co-ocurrences among sounds: the coefficients show clearly that the feature *postvelar* occurs predominantly in palatal and central vowel environments. Preceding or following laryngeals, however, reduce the likelihood of the feature *postvelar* being present in the target sound. Since the p-value of all coefficients is smaller than 0.05, the findings are highly unlikely to be due to chance. This is additional evidence for coarticulatory and statistical constraint effects in PIE.

2.4 *Environmental constraints and coarticulatory effects with laryngeal target sounds*

The method above can be applied to the laryngeals to gain deeper insight into the coarticulatory and statistical constraint effects governing these sounds. For

TABLE 2 Generalized linear logistic regression for the occurrence of *h_1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−3.502	0.100	−34.942	0.000
labial preceding	−1.412	0.420	−3.359	0.001
sibilant preceding	−2.195	1.006	−2.181	0.029
syllabic preceding	0.906	0.323	2.807	0.005
back vowel preceding	−1.007	0.415	−2.429	0.015
mid vowel preceding	0.655	0.164	3.995	0.000
word boundary following	−1.322	0.433	−3.053	0.002
nasal following	0.819	0.228	3.591	0.000
postvelar following	−2.054	1.007	−2.040	0.041

this purpose, the laryngeal data that were excluded earlier were reintroduced into the above dataset. To enable binary logistic regression, for each of the laryngeals a binary vector was added with *1* if the target sound was the particular laryngeal in question and *0* if it was any other sound. Then for each of the laryngeals, generalized linear logistic regression analyses were conducted with the binary laryngeal vector as dependent variable and the environmental features as independent variables. Similar to the model above, aliases and multicollinearity effects were detected in the full model. The affected variables were subsequently removed. To counter the effects of collinearity, variables were removed up to a cutoff-point of Variance Inflation Factor (VIF) greater than 4. Tables 2, 3, and 4 give insights into the question of which environmental features affect the occurrence of the laryngeals.

The most salient observation from Table 2 is that a vocalic and nasal environment is the most influential factor that increases the likelihood of *h_1 occurring in PIE. In sibilant or postvelar environments, however, *h_1 seems to be less likely to surface. Note that this does not mean that *h_1 does not occur under other circumstances (which it doubtless does!), only that these environments are either neutral with respect to the likelihood of *h_1 occurring, in which case they are not listed as effects, or even decrease the likelihood, as is the case with features that exhibit a negative effect.

As for *h_2, the influences on its likelihood of appearance are different (see Table 3). Many other features such as aspirated environments decrease the likelihood, while preceding syllabic consonants and following plosives or sonorants favor the occurrence of *h_2. Overall, this laryngeal seems to have more constraints than positive factors, which means that the phonological occur-

TABLE 3 Generalized linear logistic regression for the occurrence of $*h_2$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−3.305	0.137	−24.052	0.000
aspirated preceding	−2.431	1.006	−2.417	0.016
labial preceding	−0.975	0.281	−3.471	0.001
sibilant preceding	−0.772	0.390	−1.977	0.048
syllabic preceding	0.555	0.265	2.097	0.036
back vowel preceding	−1.771	0.323	−5.481	0.000
mid vowel preceding	0.880	0.135	6.540	0.000
word boundary following	0.655	0.212	3.087	0.002
aspirated following	−1.220	0.529	−2.308	0.021
syllabic following	−0.737	0.271	−2.724	0.006
postvelar following	−0.906	0.441	−2.052	0.040
plosive following	0.598	0.188	3.187	0.001
sonorant following	0.512	0.151	3.399	0.001

rence of $*h_2$ is mainly governed by wide-ranging constraints on which environments it can appear in.

$*h_3$ exhibits only a few significant predictors which increase or decrease its likelihood (see Table 4). The odds are increased by following nasals, while preceding labial consonants and following word boundaries greatly disfavor $*h_3$.

The overall situation of the environments of laryngeal target sounds shows that only a few features increase the probability of laryngeals occurring ($*h_1$ and $*h_2$) or many environmental predictors are not significant ($*h_3$). This underlines the special status of the laryngeals within the phonetic system of PIE and indicates that they are mainly defined by the environments in which they do not appear rather than by the circumstances under which they appear.

Although the statistical methods in the previous analyses were able to demonstrate various effects of the phonetic environment on a particular sound and vice versa, this does not lead to any deeper insights into laryngeal phonetics. After all, identification of the factors contributing to the likelihood of appearance has little explanatory power regarding actual phonetic features. For this reason, the main goal of this study is to take the computational tools provided by artificial intelligence and to exploit their capabilities to determine the phonetic values of the PIE laryngeals from their environments.

TABLE 4 Generalized linear logistic regression for the occurrence of $*h_3$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−4.095	0.111	−37.031	0.000
labial preceding	−1.588	0.587	−2.703	0.007
word boundary following	−2.437	1.007	−2.421	0.015
nasal following	1.283	0.255	5.036	0.000

3 The neural network problem

Only recently have methods in cladistics that involve machine learning been applied (see, e.g., Jäger, List, & Sofroniev 2017; Jäger & Sofroniev 2016), but most researchers focus on the theoretical underpinnings of the application of machine learning to linguistics. The general term *machine learning* refers to a set of artificial intelligence algorithms that are programmed to "learn" patterns and properties of a given input with the goal of achieving a good approximation of the output data. Often, these algorithms are used as classifiers to first learn the properties of the input objects, then to categorize them into given classes. The overall appeal of these methods in computer science is the possibility for artificial intelligence to learn, for instance, the distinction between classes on its own without further specification. This is especially useful for cases where the relation between the input's features and the expected classes is not clear, only that such a relation exists. These algorithms are highly specialized and usually perform exceptionally well on large datasets.

One subfield of machine learning is *deep learning*, which uses *deep neural networks* on datasets. Neural networks are the internal structure of a deep learning algorithm and are modelled analogically to a simple, idealized network of biological brain cells. To briefly illustrate the functional mechanisms of such a neural network, I will use the problem of hand-written digit recognition, an example often found in introductory textbooks (e.g., Nielsen 2015). In this example, the input is a dataset with images of handwritten digits along with their correct labels (i.e., what digit the particular image shows). After preprocessing the data to make it fit the neural network structure, the data is split into training and test data in order to train the network on one dataset and test its performance on a dataset unknown to the network to assure general validity. The image data is now passed through two or more layers of so-called neurons, with each neuron being connected to each neuron in the previous and next layer. Especially for smaller and simpler tasks, the size and number of these

layers is chosen by an exploratory search to find which configuration yields the best results. The network initially sets random values, so-called weights and biases, as strengths for these connections. Once the input data has passed through all layers and has reached the output layer, the prediction for the class is then checked with the actual class the image belongs to. With each example, the network adjusts the weights and biases of its connections such that the difference between the prediction and the actual class is minimized with each new sample. In this way, the network enhances itself to make better predictions until, ideally, new and unknown digits can be classified correctly based solely on the features of the input. In a second step, it is possible to present the network with test data to calculate its performance on previously unknown data. The predictions it makes for the test data are checked against the actual labels, and the more test examples it predicts correctly, the better the network performs in general.

Regarding the laryngeal data, I have set up a neural network to learn the patterns and associations between the environmental features and the features of the target sounds. I have trained it to correctly predict the phonetic features of the known PIE sounds based on the particular environmental features. In a second step, it is then possible to feed in the data for the laryngeals to predict the phonetic qualities of each laryngeal, hence to approximate their actual phonetics.

3.1 *Testing the method on modern German*[9]

As a second approach to ensure that the presented method and data are 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, Piepenbrock, & Gulikers 1995) in the syllabified phonetic lemma transcription with stress in the DISC character set (*PhonStrsDISC*). Using CELEX2 as a corpus has the advantage that it consists of different inflected forms and thus approximates the PIE data provided by Wiktionary. 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 in proportion to its frequency of occurrence in the 'Mannheimer Korpus' provided by CELEX2 (*Mann_Freq*) (see Gulikers, Rattink, & Piepenbrock 1995). While this approach would ideally proportion the dataset more realistically and could, in theory,

9 This section is an extended version of a section in Hartmann (2019).

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*)[10] where *0* and *1* indicate the absence/presence of a particular feature, respectively.[11] 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 during training while the network focuses on those features that have predictive power.

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 focus 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 in turn has little explanatory power in 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 is ideally 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 particular 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

10 The features chosen for this study follow the definitions in Hall (2007), filtered by whether they describe sounds present in PIE or German, respectively.

11 For the full list of features used in this study, please refer to the Appendix.

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. That is, 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 dataset was 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 oversampled before each run to counter class imbalances with the SMOTE algorithm (Chawla et al. 2002) implemented in the 'Imbalanced-learn' (Lemaître, Nogueira, & Aridas 2017) Python package. The network was trained for 30 epochs using the optimizer Adam with a learning rate of 0.01 on a batch size of 250 samples, with the layer configuration displayed in Table 24 in the Appendix.

For the subsequent evaluation of the model performance, weights and biases were used from the epoch at which the network performed best on the validation data during training using the Keras callback *ModelCheckpoint* (Chollet 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 25, 26, 27, and 28 (see the Appendix) which show which number of samples in the test sets were classified correctly or incorrectly. For instance, 24656 consonant samples in the column *TP* in the first row of Table 25 means that 24656 samples of all positive samples in the test set were correctly classified as positive. Similarly, 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 for the deep network to yield more true positives than false negatives and more true negatives than false positives. Applied to the example in Table 25, 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, and more samples would end up being incorrectly labeled as negatives. 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.

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 ranges from alveolar to uvular (cf. Meinhold & Stock 1982: 131–33), 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 better on some sounds and features than on others, it performs better than expected by chance.

To assess whether the neural networks perform similarly on data of the same size as the PIE dataset, this test has been conducted a second time with a random subset of the data of size 6236. The count of correct feature predictions was 7 out of 10 for *aː*, 8 out of 10 for *ɛː*, 12 out of 13 for *p*, and 10 out of 13 for *r*.[12] As the performance exhibits no significant deviation from the results obtained by the model using the larger dataset, the difference in data size is unlikely to influence the accuracy of the model to a significant degree. It was observed, however, that the prediction confidence, in this case the difference between positive and negative labeled samples, is decreased in comparison with the model trained on the full dataset.

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 addressed in further research.

Due to the possibility that the PIE laryngeals might have a place of articulation (POA) different from all other PIE sounds that are used as reference here (e.g., glottal), it is necessary to show how the network behaves when given environmental data of sounds that have a POA unknown to the network. For this

12 The reason for the *liquid* feature not being predicted correctly might be that the only other sound with the feature *liquid* in the dataset was *l* and the network was unable to generalize the patterns associated with this feature based solely on one sound.

purpose, I trained five networks to detect the POA features *bilabial* ([m], [p], [b]), *labiodental* ([p͡f], [f], [v]), *alveolar* ([n], [t], [d], [t͡s], [s], [z], [l]), *postalveolar* ([t͡ʃ], [ʃ], [ʒ]), and *velar/uvular* ([ŋ], [k], [g], [x], [ʁ]). For the training of each network, one of these features was withheld, and it was tested what the respective network predicts for the sounds of unknown POA.[13]

For instance, to investigate how the network predicts the POA for bilabial sounds when the feature *bilabial* is unknown, the network was first trained on a dataset excluding the feature *bilabial* and the bilabial sounds [m], [p], and [b]. Afterwards, the network was tasked with predicting the POA of the bilabials based their environmental information. In theory, a good network correctly predicts in the majority of cases that the bilabials do not exhibit any other POA features or predict locally adjacent features. Note that the feature *palatal* was not trained, as it is only represented by one sound ([j]) which does not yield useful results as outlined in section 2.2. The results show that for those five tested places of articulation, 15 out of 20 features were correctly predicted. In two instances (*labiodental*, *postalveolar*), the network predicted the adjacent POA. Therefore, 17 out of 20 predictions gave either the correct or the directly adjacent POA to the feature that was withheld. Although the networks struggled with two features of the velars and one feature of the alveolars, the method seems to yield correct results in the majority of cases.

3.2 *The specifications and training of the networks for PIE*

The dataset for predicting PIE sound features based on their phonetic environment came with certain idiosyncrasies and biases that needed to be dealt with in order to make the task manageable for the deep learning algorithms. One of the problems was data-inherent and did not arise due to lack of observations or data collection problems, namely the high imbalance of some sample classes. In reconstructed PIE, sounds with certain features appear relatively infrequently compared to other features. Training networks on imbalanced datasets can cause the network to always favour the majority group. Therefore, the data samples were first stratified using StratifiedShuffleSplit (Pedregosa et al. 2011) and distributed as evenly as possible over the training and test data. The test data sample size was set to 20 percent of the whole dataset. To deal with the class imbalance, the training set was oversampled with the SMOTE algorithm and subsequently under-sampled by removing Tomek links using SMOTETomek (Lemaître, Nogueira, & Aridas 2017). Note that oversampling minimizes the influence of frequency of individual tokens during training,

13 Please refer to Tables 33 – 37 in the Appendix for the detailed results.

which means that the higher frequency with which individual samples occur does not influence the prediction accuracy. Since the samples in the data were randomly assigned to the train and test sets each time, the quality of the training varied to some degree from run to run depending on the severity of the difference. Additionally, the SMOTE oversampling process performed on the minority group enhances this variation. To cope with this variation, I ran each network 100 times to obtain 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. This means that in a run testing a hypothetical feature A, the network correctly classifies *n* samples as having feature A (true positives), while *n* samples with feature A are misclassified as not having feature A (false negatives). Equally, *n* samples are incorrectly assigned the feature A although they do not have the feature (false positives) and *n* samples are correctly classified as not having feature A (true negatives).

To determine whether the model performs significantly better than expected by a random class assignment, all confusion matrices were compared using Wilcoxon-Mann-Whitney tests. For each model, I performed this test on 100 runs of true positives vs. false negatives to determine whether the network can clearly find a present feature, and a second test on 100 runs of false positives vs. true negatives to determine whether the network can clearly find the absence of a feature. When the Wilcoxon-Mann-Whitney test is significant, the tested groups are non-identical populations. If, for example, a network performs well on the given data, the Wilcoxon-Mann-Whitney test will find a significant difference between the 100 true positive and the 100 false negative values, since most samples containing the feature will be classified as positives. Similarly, there will be a significant difference between the 100 false positive and the 100 true negative values, since most samples where the feature in question is absent will be classified as negatives. In a case in which the test is not significant, the model could not detect the presence or absence of a feature.[14]

After each training run, the environmental features of the three laryngeals were passed to the network with the network's predictions as output. Therefore, at the end of the process, the samples of each laryngeal were classified as positive (has the feature) or negative (does not have the feature) in a total of 100 runs. By doing this, it can be made sure that the feature predictions stem from the 100 training runs, thereby assuring that the networks perform well.

14 The dependence requirement of this test is fulfilled since true positives and false negatives, and also false positives and true negatives, are dependent with a correlation value of *1* by the nature of the experiment.

Whether a particular laryngeal has the tested feature can once again be determined using a Wilcoxon-Mann-Whitney test.

3.3 *The results of the model*

To access the model architectures and network training specifications, please refer to the Appendix. Only those statistics are given in the paper which are of immediate relevance to the analysis of the results. All U-values of the Wilcoxon-Mann-Whitney tests displayed in the evaluation tables refer to TP˜FN or TN˜FP.

3.3.1 [±consonant]

TABLE 5 Statistics of the confusion matrices from 100 runs for classifying the feature [±consonant]

	True positives	**False negatives**	**False positives**	**True negatives**
Mean	837.20	63.80	28.58	318.42
Median	837	64	28	319
Standard deviation	**3.58**		**2.73**	
Wilcoxon-Mann-Whitney U	**10000.00**		**10000.00**	
p-value	**< 0.001**		**< 0.001**	

TABLE 6 Prediction results by the trained model for the laryngeal feature [±consonant]

	***h*$_1$**		***h*$_2$**		***h*$_3$**	
	Positives	**Negatives**	**Positives**	**Negatives**	**Positives**	**Negatives**
Mean	184.56	14.44	319.14	28.86	100.01	5.99
Median	186	13.0	320	28	100	6
Standard deviation	**3.27**		**4.04**		**1.55**	
Wilcoxon-Mann-Whitney U	**10000.00**		**10000.00**		**10000.00**	
p-value	**< 0.001**		**< 0.001**		**< 0.001**	

3.3.2 [±velar]

TABLE 7 Statistics of the confusion matrices from 100 runs for classifying the feature [±velar]

	True positives	False negatives	False positives	True negatives
Mean	38.25	18.75	221.63	621.37
Median	38	19	222	621
Standard deviation	**1.22**		**4.24**	
Wilcoxon-Mann-Whitney U	**10000.00**		**10000.00**	
p-value	**< 0.001**		**< 0.001**	

TABLE 8 Prediction results by the trained model for the laryngeal feature [±velar]

	*h_1		*h_2		*h_3	
	Positives	Negatives	Positives	Negatives	Positives	Negatives
Mean	112.46	86.54	162.03	185.97	64.30	41.7
Median	113	86	161.5	186.5	65	41
Standard deviation	**4.24**		**5.16**		**5.08**	
Wilcoxon-Mann-Whitney U	**9971.00**		**0.50**		**9946.00**	
p-value	**< 0.001**		**< 0.001**		**< 0.001**	

3.3.3 [±postvelar]

TABLE 9 Statistics of the confusion matrices from 100 runs for classifying the feature [±postvelar]

	True positives	**False negatives**	**False positives**	**True negatives**
Mean	53.66	22.34	237.85	586.15
Median	54	22	239	585
Standard deviation	**2.14**		**8.95**	
Wilcoxon-Mann-Whitney U	10000.00		10000.00	
p-value	< 0.001		< 0.001	

TABLE 10 Prediction results by the trained model for the laryngeal feature [±postvelar]

	h_1		***h_2***		***h_3***	
	Positives	**Negatives**	**Positives**	**Negatives**	**Positives**	**Negatives**
Mean	100.11	98.89	152.90	195.1	49.80	56.2
Median	99	100	153	195	50	56
Standard deviation	**5.18**		**9.38**		**2.98**	
Wilcoxon-Mann-Whitney U	**5140.00**		0		77.00	
p-value	**0.73**		< 0.001		< 0.001	

3.3.4 [±labial]

TABLE 11 Statistics of the confusion matrices from 100 runs for classifying the feature [±labial]

	True positives	**False negatives**	**False positives**	**True negatives**
Mean	143.96	60.04	225.11	470.89
Median	145	59	227	469
Standard deviation	**7.58**		**13.34**	
Wilcoxon-Mann-Whitney U	**10000.00**		**10000.00**	
p-value	**< 0.001**		**< 0.001**	

TABLE 12 Prediction results by the trained model for the laryngeal feature [±labial]

	***h_1**		***h_2**		***h_3**	
	Positives	**Negatives**	**Positives**	**Negatives**	**Positives**	**Negatives**
Mean	115.67	83.33	179.22	168.78	60.11	45.89
Median	116	83	180	168	60	46
Standard deviation	**6.38**		**9.45**		**4.57**	
Wilcoxon-Mann-Whitney U	**9998.00**		**8126.00**		**9803.50**	
p-value	**< 0.001**		**< 0.001**		**< 0.001**	

3.3.5 [±aspirated]

TABLE 13 Statistics of the confusion matrices from 100 runs for classifying the feature [±aspirated]

	True positives	**False negatives**	**False positives**	**True negatives**
Mean	54.55	23.45	209.58	612.42
Median	55	23	210.5	611.5
Standard deviation	**1.53**		**7.21**	
Wilcoxon-Mann-Whitney U	10000.00		10000.00	
p-value	< 0.001		< 0.001	

TABLE 14 Prediction results by the trained model for the laryngeal feature [±aspirated]

	***h*$_1$**		***h*$_2$**		***h*$_3$**	
	Positives	**Negatives**	**Positives**	**Negatives**	**Positives**	**Negatives**
Mean	112.16	86.84	155.36	192.64	54.65	51.35
Median	113.5	85.5	157	191	55	51
Standard deviation	**4.96**		**4.55**		**2.2**	
Wilcoxon-Mann-Whitney U	**9959.00**		0		8718.50	
p-value	< 0.001		< 0.001		< 0.001	

3.3.6 [±voice]

TABLE 15 Statistics of the confusion matrices from 100 runs for classifying the feature [±voice]

	True positives	False negatives	False positives	True negatives
Mean	406.67	85.33	136.08	271.92
Median	407.5	84.5	137	271
Standard deviation	**13.45**		**13.17**	
Wilcoxon-Mann-Whitney U	10000.00		10000.00	
p-value	< 0.001		< 0.001	

TABLE 16 Prediction results by the trained model for the laryngeal feature [±voice]

	$*h_1$		$*h_2$		$*h_3$	
	Positives	Negatives	Positives	Negatives	Positives	Negatives
Mean	125.67	73.33	246.51	101.49	62.41	43.59
Median	126	73	247	101	63	43
Standard deviation	**6.4**		**10.43**		**2.92**	
Wilcoxon-Mann-Whitney U	10000.00		10000.00		9998.00	
p-value	< 0.001		< 0.001		< 0.001	

3.3.7 [±syllabic]

TABLE 17 Statistics of the confusion matrices from 100 runs for classifying the feature [±syllabic]

	True positives	False negatives	False positives	True negatives
Mean	89.13	12.87	25.36	772.64
Median	89	13	24	774
Standard deviation	**2.23**		**4.34**	
Wilcoxon-Mann-Whitney U	10000.00		10000.00	
p-value	< 0.001		< 0.001	

TABLE 18 Prediction results by the trained model for the laryngeal feature [±syllabic]

	$*h_1$		$*h_2$		$*h_3$	
	Positives	Negatives	Positives	Negatives	Positives	Negatives
Mean	15.54	183.46	29.77	318.23	9.79	96.21
Median	15.5	183.5	29.5	318.5	10	96
Standard deviation	**4.94**		**4.6**		**2.05**	
Wilcoxon-Mann-Whitney U	0		0		0	
p-value	< 0.001		< 0.001		< 0.001	

3.3.8 [±coronal]

TABLE 19 Statistics of the confusion matrices from 100 runs for classifying the feature [±coronal]

	True positives	False negatives	False positives	True negatives
Mean	290.44	170.56	103.34	335.66
Median	292	169	103	336
Standard deviation	**12.71**		**12.19**	
Wilcoxon-Mann-Whitney U	**10000.00**		**10000.00**	
p-value	**< 0.001**		**< 0.001**	

TABLE 20 Prediction results by the trained model for the laryngeal feature [±coronal]

	$*h_1$		$*h_2$		$*h_3$	
	Positives	Negatives	Positives	Negatives	Positives	Negatives
Mean	48.66	150.34	117.63	230.37	30.8	75.2
Median	50	149	118.5	229.5	32	74
Standard deviation	**7.76**		**9.99**		**4.91**	
Wilcoxon-Mann-Whitney U	**0**		**0**		**0**	
p-value	**< 0.001**		**< 0.001**		**< 0.001**	

3.3.9 [±nasal]

While the network optimized for detecting the other features did not yield significant results due to possible inconsistencies (cf. Hartmann 2019), a network optimized for this feature performed better on this task.

TABLE 21 Statistics of the confusion matrices from 100 runs for classifying the feature [±nasal]

	True positives	**False negatives**	**False positives**	**True negatives**
Mean	68.30	34.7	277.51	519.49
Median	69	34	280	517
Standard deviation	**5.77**		**38.82**	
Wilcoxon-Mann-Whitney U	10000.00		10000.00	
p-value	< 0.001		< 0.001	

TABLE 22 Prediction results by the trained model for the laryngeal feature [±nasal]

	***h*₁**		***h*₂**		***h*₃**	
	Positives	**Negatives**	**Positives**	**Negatives**	**Positives**	**Negatives**
Mean	69.9	129.1	157.63	190.37	40.45	65.55
Median	68	131	165.5	182.5	41.5	64.5
Standard deviation	**14.56**		**29.63**		**8.03**	
Wilcoxon-Mann-Whitney U	0		**2496.5**		0	
p-value	< 0.001		< 0.001		< 0.001	

3.4 *Investigating the model's decisions*

Generally, trained deep neural networks are difficult to examine in terms of the decision boundaries for classification tasks. More concretely, the important question 'which inputs lead to a positive/negative decision for one of the output classes?' (i.e., why does the model make the decisions it does?) is notoriously difficult to answer for neural networks. Its complex inner mechanics, which are the reason for this difficulty, are simultaneously the model's best asset. Such models operate in a multi-dimensional space, which is optimal for

recognizing complex patterns but increasingly disadvantageous for human reasoning to understand. Nevertheless, a large field in current AI research is dedicated to developing methods for analyzing the decisions of machine learning networks. Although much work is still in progress, I have conducted an input feature visualization technique which utilizes expected gradients for approximating *SHAP* (*Sh*apley *A*dditive Ex*p*lanations) values (Lundberg & Lee 2017) as implemented in the Python package *iNNvestigate* (Alber et al. 2019). Figure 4 below shows the corresponding visualization plot.[15]

This figure shows the average approximated *SHAP* values per present input feature on the basis of the correctly predicted test data of the positive class. In other words, this figure attempts to show which input features contribute positively or negatively towards classifying the sample as possessing a certain feature. The x-axis gives the input features (i.e., the environmental features), whereas the y-axis displays the tested features. A *0* or a *1* after the name of a tested feature indicates the effect of the input when the input feature was present or not. Blue features indicate a negative influence on the outcome, whereas red features represent positive contributions. In practice, the plot can be read as follows: When the model is tasked with determining whether a certain sound is [nasal], different inputs lead to different decisions by the model on whether or not to classify the sound as [nasal]. In the rows indicated by *nasal* on the y-axis, we find that if a sound is preceded by a liquid (*liquid prec.*), the probability of predicting [nasal] for this sound is reduced (indicated by the blue cell). If, however, the model is preceded by a syllabic consonant, the likelihood of predicting [nasal] increases. Regarding missing features, if on the other hand a word boundary (*boundary*) is *not* present, the model is somewhat more likely to predict the feature [nasal] for the sound in question (indicated by a light red value in the corresponding cell in the column *nasal 0*).

The disadvantage of this visualization is that it can only display linear influences assuming independence among the input values. However, since neural networks operate in complex multi-dimensional spaces taking feature interactions into account, this method only provides limited insight into the actual decision process of the networks. Nevertheless, it can be a useful tool to determine some important decision boundaries the model draws upon. In this context, it can serve as an illustration of what the model perceives to be input values which decrease/increase the probability of a certain outcome.

There are a few noteworthy findings in this figure. For example, aspiration is less likely to be predicted for a sound which follows $*h_3$. Likewise, if a sound

15 The plot was generated using the Python library *matplotlib* (Hunter 2007).

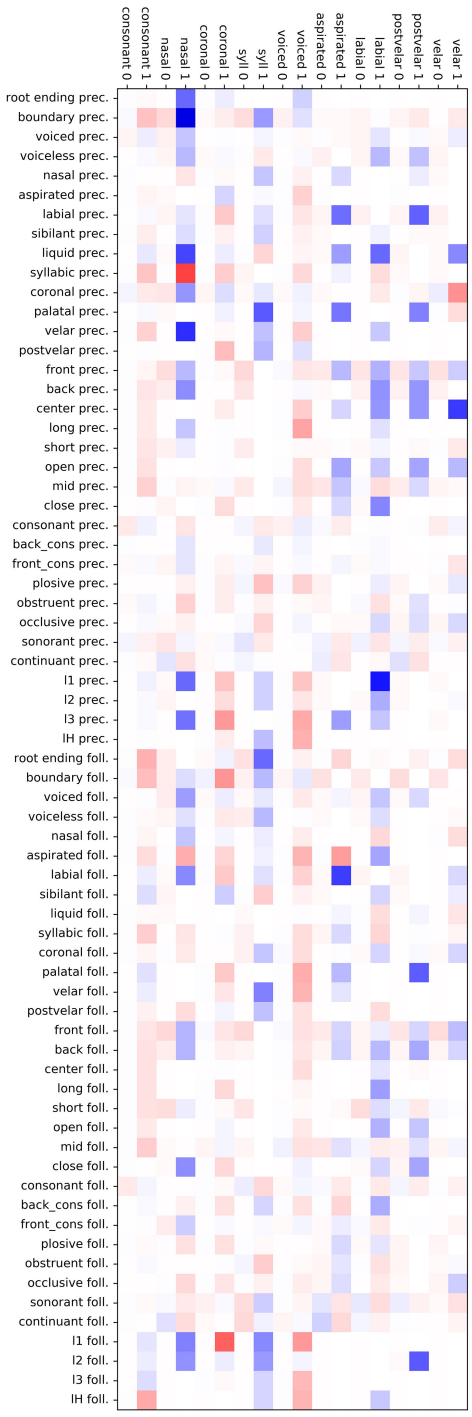

FIGURE 4
Visualization of approximated *SHAP* values

follows *h_1, it is less likely to be predicted [labial]. The feature [nasal] is generally less likely to be predicted in a laryngeal environment, and the prediction of [postvelar] is discouraged in environments with preceding labials, vowels, palatals and following *h_2. These findings can be interesting beyond the context of this study, as they might motivate the investigation of other phonotactic phenomena in PIE.

As it might also be insightful to investigate what about the environment of each laryngeal impacted the model's prediction of certain features, the same visualization technique was applied to those model runs as well. The corresponding Figs. 5, 6, and 7 may be found in the Appendix. These plots were generated analogously to the plot in Fig. 4 with the only exception that the predicted features of each laryngeal were used for estimating the *SHAP* values. This means that since, e.g., *h_1 is predicted to be voiced but not coronal, the corresponding rows indicate which input features impacted the prediction of [voice] as present in *h_1 and [coronal] as absent. Thus, the prefixes *pos* and *neg* indicate when the *SHAP* values are calculated for the positive or the negative class. Applied to, e.g., Fig. 5, this means that when there were preceding palatals in the phonetic environment of *h_1, the model's probability of predicting [-postvelar] was increased.

A few observations following this analysis can be briefly outlined. *h_1 was partially predicted [velar] because of following voiceless consonants, preceding and following sonorants and continuants, and a lack of following word boundaries. *h_2 was partially predicted [-postvelar] because of preceding liquids and labials, following word boundaries and following palatals. *h_3 was partially predicted as voiced because of following aspirated and labial sounds and vowels.

4 Discussion and interpretation

Table 23 summarizes the results of the three laryngeals combined for better comparability. Those features where the predictions were not significant are indicated by a question mark (?). Some of the laryngeal predictions, while being statistically significant, exhibit a large standard deviation in addition to the values for positives and negatives being close together. Their occurrence therefore is a warning sign, especially when the network performed smoothly on the test set as part of model evaluation. This issue arises when a network is not able to consistently predict the samples of the laryngeal in question. The cause of this is that the sample data are different from the training data in such a way that the trained network is unable to apply its decision function uniformly.

TABLE 23 Summary of detected phonetic features

Laryngeal	consonant	velar	postvelar	labial	aspirated	voice	syllabic	coronal	nasal
*h_1	+	+	?	+	+	+	–	–	-
*h_2	+	–	–	(+)	–	+	–	–	(-)
*h_3	+	+	–	+	(+)	+	–	–	–

The results of such predictions therefore need to be treated with caution since the likelihood of the data being unsuited for the prediction of that particular feature is high. Those instances where this is the case are given in parentheses.

According to the predictions of the networks, all laryngeals exhibit voicing. This finding has to be contextualized differently for *h_1 and *h_2. Regarding *h_2, voicing is in line with those interpretations favoring a pharyngeal. In the case of *h_1, however, voicing is uncommon in previous reconstructions. Two possible explanations for this can be offered. One could attribute the voicing to an earlier state by positing that *h_1 had lost voicing by late PIE, while the environment still reflected the earlier voiced feature.[16] Such a devoicing could have applied to the glottal fricatives, whereas supralaryngeal fricatives remained unchanged. On the other hand, voicing could have been less prominent and therefore not fully realized in all environments. However, as the current model design does not allow for secure predictions of which environments could have given rise to this allophonic variation, the issue must remain uncertain.

The place of articulation of *h_1 and *h_3 is clearly predicted as velar, hence this prediction runs contrary to the view that the laryngeals were pharyngeals, epiglottals, or glottals (e.g., Beekes 1994; Beekes & de Vaan 2011; Bomhard 2004). If this were the case, the models would not have predicted velar as place of articulation. Only if a clear postvelar feature had been detected could it be argued to be indicative of a back or even debuccalized place of articulation. The fact that the models uniformly suggest a rather central place of articulation, much like the PIE velar series *$\acute{k}$, *$\acute{g}$, *$\acute{g}^h$, excludes any more backed realization than postvelar.

16 Recall that due to some phonotactic patterns being potentially formed before others in the time period leading up to the last reconstructible stage of PIE, the neural networks might detect patterns which reflect features of an older stage of the particular sound in question (see section 2.1).

The predicted consonantal properties of the laryngeals, even though expected, do not exclude conditioned syllabification and are therefore compatible with the previously observed syllabic reflexes (see Kümmel 2007; Meier-Brügger, Fritz, & Mayrhofer 2010: 236–55; Cowgill & Mayrhofer 1986: 121–50). Yet the predictions nevertheless suggest a general consonantal character of the laryngeals which was already assumed by researchers a century ago (e.g., Cuny 1912; Møller 1917). Moreover, the preliminary analyses shown in Figs. 1 and 3 suggest that the laryngeals had properties closer to continuants and the strident *s, which decreases the possibility that they were stops.

Concerning the specific features of each laryngeal, the model suggests the following:

*h_1 is predicted as a voiced labialized velar consonant with aspiration. The likely articulation as a fricative and the detected aspiration favor the realization as an aspirate, if aspiration is taken as an indication of a spread glottis articulation. The [+velar] feature imposes a problem here since an aspirate can only be [+velar] if approximant velar co-articulation is assumed, comparable with the voiced labiovelar approximant [w]. This suggests [ɦʸʷ], a glottal fricative articulated with narrowed oral cavity and lip rounding, as the realization closest to the model's predictions. However, such a phonetic value would be uncommonly complex and typologically difficult to justify. It would be more reasonable to assume that the data show blending of two anachronistic features of *h_1: it is plausible that *h_1 tended to be reduced in certain environments already in PIE and thus yielded a deletion in certain environments (cf. Fritz 1996; Kümmel 2007: 334–35).

The velar feature detected for *h_1 is surprising, as the current literature posits this laryngeal as glottal. There are, however, some grounds to argue that this prediction reflects a feature *h_1 once had but lost before the final stage of PIE. In this case, the environmental properties typical for a velar, which was recognized by the model, indicate that at an older stage *h_1 might indeed have been velar. Kümmel (2007: 336), for example, proposes a change *velar* > *glottal* for the history of *h_1 which might provide an explanation for the model predicting *h_1 to be velar. A change of the form [ɣ$^{(w)}$]/[x$^{(w)}$] > [ɦʷ] would best approximate the model results. However, it is not possible to ascertain whether *h_1 and *h_3 were identical at an older stage, as *h_3 might have also had an earlier value different from its reconstructible form; such speculations are purely hypothetical. The model itself only provides grounds for considerations about an earlier stage of *h_1. This means that while in this instance there are reasons to suggest [velar] as an earlier feature of *h_1, with fossilized traces in its environmental patterns, we are unable to speculate about any previous stage of *h_3. Although it would raise a number of logical and phonological problems if *h_1 and *h_3 were in fact

very similar or identical in previous stages, we have to acknowledge that the model as it stands does not offer insights into this matter.

The model's prediction of a labialized articulation of *h_1 does not conflict with the observation that adjacent *e is not colored by this laryngeal, as it is possible that lip rounding in *h_1 was either less prominent as, for example, in *h_3 or might have been reduced before it could fully color adjacent *e. Moreover, a marginal surface coloring of adjacent *e, which is too marginal to be reanalyzed as a rounded vowel underlyingly, is also possible as a result of a weaker lip rounding. However, these are not the only possibilities for why *h_1 does not show coloring of adjacent *e. The confluence of different properties of *h_1, such as its being a glottal aspirate, might have led to a different outcome in the daughter languages compared to *h_3. Ultimately, this matter cannot be decided here, as the model itself only provides the information that it detects *h_1 predominantly in environments where we would expect to find a labial or labialized sound.

For *h_2, the features *voice* and *consonant* are predicted. Since the networks could not be trained to detect fricatives, this feature has to be inferred from the results in the preliminary analyses (see section 2.3) and the outcomes in the daughter languages. By doing so, it emerges that the most likely realization of *h_2 was a fricative with a place of articulation that was neither velar nor postvelar. The most likely candidates for the place of articulation are therefore uvulars and pharyngeals. Uvulars are only valid candidates if we assume the 'postvelars' to be backed velars ([k̠] : [g̠] : [g̠ʰ] – [k̠ʷ] : [g̠ʷ] : [g̠ʷʰ]). There is some room for debate considering *h_2: the networks predict that *h_2 was articulated at a place *not* equal to that of the postvelar series, but likely further back. The alternative laryngeal realization therefore could also have been [ɣ(ʷ)]/[x(ʷ)] > [ɦʷ] : [ʁ] : [ɣʷ]. It could be argued in favor of this interpretation that a uvular interpretation of *h_2 would be more in line with velar *h_3 which is often assumed to be a labialized counterpart of *h_2. Yet given the discussion about the likely place of articulation of the postvelars outlined above and the fact that those studies arguing in favor of a voiced fricative interpretation of h_2 often also assume pharyngeal articulation (see, e.g., Beekes 1994; Bomhard 2004), it is reasonable to assume a pharyngeal articulation here as well. Since this matter is not definitively decided, I prefer the interpretation of the postvelar series as uvulars since further considerations assume and at times require them to be uvulars (e.g., Kümmel 2007: 310–27). Furthermore they would phonologically contrast with the velar series more strongly, which is preferable: if PIE had contrasting sets *velars* : *backed velars*, the two series would be prone to neutralization and increased confusability due to phonetic surface variation. *h_2 could therefore have been realized as either uvular [ʁ] or pharyngeal [ʕ], the latter being preferred here.

This interpretation of *h_2 as voiced differs somewhat from previous interpretations, since [+voice] was never assumed in combination with a uvular or velar interpretation. Only those researchers proposing a pharyngeal realization of this laryngeal assumed voice as a property of *h_2 (e.g., Beekes 1994), but even if one rejects pharyngeal *h_2, evidence from Anatolian and the fact that it had a vocalizing effect do not stand in the way of its interpretation as a voiced consonant.

In contrast to what is sometimes assumed (e.g., Rasmussen 1994; Beekes 1994; Weiss 2016), *h_3 is *not* predicted to have been the labialized counterpart of *h_2. Some scholars have already argued against this interpretation (e.g., Gippert 1994; Kümmel 2007).[17] The most likely interpretation of *h_3 according to the model is [ɣʷ], which is consistent given that the most salient effect of this labialization in the daughter languages is the *o*-coloring on adjacent **e*. The interpretation of *h_3 as velar was already proposed by Rasmussen (1994: 435–36), who drew his conclusion in part on loss of *h_3 before Celtic /kʷ/ and assimilation of the sequence *h_3w to [gʷ] in Germanic (cf. Ringe 2017: 86–88). Although this interpretation of *h_3 coincides with Rasmussen (1994) in this respect, scholars who favor a velar interpretation also tend to favor a velar realization of *h_2. Therefore, the possibility needs to be considered that the model has detected a dorsal component of *h_3, which is predicted to be more fronted than *h_2. It is possible that *h_3, although predicted to be *velar*, could be an adjacent [back] dorsal, a phenomenon observed, albeit less frequently, when testing the model on the German dataset (see section 3.1). This would result in the interpretation of *h_3 as *uvular*, a notion also voiced by Kümmel (2007: 336), which would result in *h_3 being [ʁʷ].

This predicted difference between *h_2 and *h_3 despite their similar behavior in the IE daughter languages calls for comment. In terms of phonetic co-occurrence patterns they are rather different, as the analyses shown in Tables 3 and 4 have already suggested. The major difference between them and *h_1 lies in the fact that *h_1 is predicted to be aspirate [ɦʷ], whereas *h_2 and *h_3 are supralaryngeal fricatives. It is likely that this property is what set them apart from both *h_1 and other back stops. Moreover, the common property of the laryngeals as a whole might have been that they were back (including glottal) fricatives. This common feature was likely detected already by the MDS projection displayed in Fig. 1. There, the three laryngeals are projected to the right of the stops in the positive x-direction and on the same level as fricative **s*, glides, and liquids. They are differentiated by frication from stop consonants and by

17 The latter assumes labialization only as an earlier feature of *h_3.

their backness and low sonority from *s and the glides. It makes them unique insofar as they are, according to this interpretation, the only phonemic fricatives beside *s in PIE. The further development of the laryngeals in the daughter languages, such as loss following either direct vocalization or vowel insertion, can be reconciled with this interpretation insofar as frication in combination with backness is the unifying feature set which differentiates the laryngeals from other consonants. Any further changes of either direct vocalization or anaptyxis would thus apply to all sounds included in this set.

5 Conclusion

To conclude the analysis, the most likely values of the three laryngeals are [ɣ(ʷ)]/[x(ʷ)] > [ɦʷ] : [ʕ] : [ɣʷ]/[ʁʷ]. These results are based on the interpretation of the computational model, although previous research was taken into account to cover the aspects that could not be predicted by the deep neural networks, namely the fricative features of $*h_2$ and $*h_3$.

If these findings are on the right track, the common property of $*h_2$ and $*h_3$ is that they are supralaryngeal fricatives. This sets them apart from the glottal aspirate $*h_1$ and other supralaryngeal consonants such as the stop series. The similar behavior of *s as the only other fricative (as also indicated by the results of the hierarchical clustering in Fig. 3) supports the assumption that frication could be the most important feature setting them apart from most other sounds in the PIE inventory.

These findings are to a large extent compatible with previous results and validate the findings of studies such as Kümmel (2007), Bomhard (2004), and Rasmussen (1994). However, it also yielded predictions of features which were not suggested by previous research. Because the neural networks implemented in this study correctly detected every tested feature during training, this is a strong argument for the validity of the feature predictions obtained for the laryngeals, though it has to be acknowledged that these predictions were not always unequivocal. Thus, this study presents a new approach and a different perspective which can provide the basis for further research into the laryngeals and the phonetic inventory of Proto-Indo-European as a whole.

References

Alber, Maximilian, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, & Pieter-Jan Kindermans (2019). iNNvestigate neural networks. *Journal of machine learning research* 20: 1–8.

Baayen, Harald, Richard Piepenbrock, & Léon Gulikers (1995). CELEX2 LDC96L14. In *Web Download*, https://catalog.ldc.upenn.edu/LDC96L14. Philadelphia: Linguistic Data Consortium.

Bates, Douglas, Martin Mächler, Ben Bolker, & Steve Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software* 67: 1–48.

Beekes, Robert S.P. (1994). Who were the laryngeals? In *In honorem Holger Pedersen*, ed. Jens Elmegård Rasmussen, 450–54. Wiesbaden: Reichert.

Beekes, Robert S.P. & Michiel de Vaan (2011). *Comparative Indo-European linguistics: An introduction*. Amsterdam: Benjamins.

Blevins, Juliette (2015). Evolutionary phonology: A holistic approach to sound change typology. In *The Oxford handbook of historical phonology*, ed. Patrick Honeybone & Joseph C. Salmons, 485–500. Oxford: Oxford University Press.

Bomhard, Allan R. (2004). The Proto-Indo-European laryngeals. In *Per aspera ad asteriscos*, ed. Adam Hyllested & Jens Elmegård Rasmussen, 69–80. Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck.

Bybee, Joan (2015). Articulatory processing and frequency of use in sound change. In *The Oxford handbook of historical phonology*, ed. Patrick Honeybone & Joseph Salmons, 467–84. Oxford: Oxford University Press.

Byrd, Andrew M. (2015). *The Indo-European syllable*. Leiden: Brill.

Byrd, Andrew M. (2017). The phonology of Proto-Indo-European. In *Handbook of comparative and historical Indo-European linguistics*, ed. Jared Klein, Brian Joseph, & Matthias Fritz, 2056–79. Berlin: De Gruyter Mouton.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, & W. Philip Kegelmeyer (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16: 321–57.

Chiarcos, Christian, John McCrae, Philipp Cimiano, & Christiane Fellbaum (2013). Towards open data for linguistics: Linguistic linked data. In *New trends of research in ontologies and lexical resources: Ideas, projects, systems*, ed. Alessandro Oltramari, Piek Vossen, Lu Qin, & Eduard Hovy, 7–25. Berlin: Springer.

Chollet, François (2015). Keras. Documentation available at https://keras.io.

Clackson, James (2007). *Indo-European linguistics: An introduction*. Cambridge: Cambridge University Press.

Cohen Priva, Uriel (2015). Informativity affects consonant duration and deletion rates. *Laboratory phonology* 6: 243–78.

Coombes, Kevin R. (2018). ClassDiscovery: Classes and methods for "Class Discovery" with microarrays or proteomics. https://CRAN.R-project.org/package=ClassDiscovery (accessed 5 March 2020).

Cooper, Adam (2009). Similarity avoidance in the Proto-Indo-European root. *University of Pennsylvania working papers in linguistics* 15: 55–64.

Cooper, Adam (2013). The typology of PIE syllabic sonorants. *Indo-European linguistics* 1: 3–67.

Cooper, Adam (2015). *Reconciling Indo-European syllabification*. Leiden: Brill.

Cowgill, Warren, & Manfred Mayrhofer (1986). *Indogermanische Grammatik*. Vol. I. Heidelberg: Winter.

Cuny, Albert (1912). Indo-européen et sémitique. *Revue de phonétique* 2: 101–03.

De Melo, Gerard (2015). Wiktionary-based word embeddings. In *Proceedings of MT Summit XV*, ed. Yaser Al-Onaizan & Will Lewis, 346–59. Miami: Association for Machine Translation in the Americas.

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

Donegan, Patricia J., & Geoffrey S. Nathan (2015). Natural phonology and sound change. In *The Oxford handbook of historical phonology*, ed. Patrick Honeybone & Joseph Salmons, 431–49. Oxford: Oxford University Press.

Fortson, Benjamin W. IV. (2010). *Indo-European language and culture: An introduction*[2]. Chichester: Wiley-Blackwell.

Fowler, Carol A. (1980). Coarticulation and theories of extrinsic timing. *Journal of phonetics* 8: 113–33.

Frisch, Stefan A., Janet B. Pierrehumbert, & Michael B. Broe (2004). Similarity avoidance and the OCP. *Natural language & linguistic theory* 22: 179–228.

Fritz, Matthias (1996). Das urindogermanische Wort für 'Nase' und das grundsprachliche Lautgesetz *RHV > *RV. *Historische Sprachforschung* 109: 1–20.

Gippert, Jost (1994). Zur Phonetik der Laryngale. In *In honorem Holger Pedersen*, ed. Jens Elmegård Rasmussen, 455–66. Wiesbaden: Reichert.

Gulikers, Léon, Gilbert Rattink, & Richard Piepenbrock (1995). German linguistic guide. In *The CELEX Lexical Database* (*CD-ROM*), https://catalog.ldc.upenn.edu/LDC96L14 (accessed 13 March 2019). Philadelphia: Linguistic Data Consortium.

Hale, Mark (2003). Neogrammarian sound change. In *The handbook of historical linguistics*, ed. Brian D. Joseph, 343–68. Malden: Blackwell.

Hall, Kathleen C., Elizabeth Hume, Tim F. Jaeger, & Andrew Wedel (2018). The role of predictability in shaping phonological patterns. *Linguistics vanguard* 4 (S2).

Hall, Tracy A. (2007). Segmental features. In *The Cambridge handbook of phonology*, ed. Paul de Lacy, 311–34. Cambridge: Cambridge University Press.

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

Hartmann, Frederik (2019). Predicting historical phonetic features using deep neural networks: A case study of the phonetic system of Proto-Indo-European. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 98–108. Florence: Association for Computational Linguistics.

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering* 9: 90–95.

Iverson, Gregory K., & Joseph C. Salmons (1992). The phonology of the Proto-Indo-European root structure constraints. *Lingua* 87: 293–320.

Jäger, Gerhard, Johann-Mattis List, & 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 the 15th conference of the European chapter of the Association for Computational Linguistics*, 1205–16. Valencia: Association for Computational Linguistics.

Jäger, Gerhard, & Pavel Sofroniev (2016). Automatic cognate classification with a support vector machine. In *Proceedings of the 13th conference on Natural Language Processing*, ed. Stefanie Dipper, Friedrich Neubarth, & Heike Zinsmeister, 128–34. Bochumer Linguistische Arbeitsberichte.

Kloekhorst, Alwin (2004). The preservation of $*h_1$ in hieroglyphic Luwian: two separate *a*-signs. *Historische Sprachforschung* 117: 26–49.

Kloekhorst, Alwin (2006). Initial laryngeals in Anatolian. *Historische Sprachforschung* 119: 77–108.

Kühnert, Barbara, & Francis Nolan (1999). The origin of coarticulation. In *Coarticulation*, ed. Nigel Hewlett & William J. Hardcastle, 7–30. Cambridge: Cambridge University Press.

Kümmel, Martin J. (2007). *Konsonantenwandel: Bausteine zu einer Typologie des Lautwandels und ihre Konsequenzen für die vergleichende Rekonstruktion*. Wiesbaden: Reichert.

Lemaître, Guillaume, Fernando Nogueira, & 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: 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

Lundberg, Scott M., & Su-In Lee (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems 30*, ed. Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S.V.N. Vishwanathan, & Roman Garnett, 4765–74. Long Beach: Neural Information Processing Systems.

Mallory, James P., & Douglas Q. Adams (2006). *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world*. Oxford: Oxford University Press.

Meier-Brügger, Michael, Matthias Fritz, & Manfred Mayrhofer (2010). *Indogermanische Sprachwissenschaft*. Berlin: de Gruyter.

Meinhold, Gottfried, & Eberhard Stock (1982). *Phonologie der deutschen Gegenwartssprache*. Leipzig: Bibliographisches Institut.

Meyer, Christian M., & Iryna Gurevych (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In *Electronic lexicography*, ed. Sylviana Gragner & Magali Paquot, 259–91. Oxford: Oxford University Press.

Møller, Hermann (1917). *Die semitisch-vorindogermanischen laryngalen Konsonanten*. København: Andr. Fred. Høst & Søn.

Navarro, Emmanuel, Franck Sajous, Bruno Gaume, Laurent Prévot, Hsieh Shukai, Kuo Tzu-Yi, Pierre Magistry, & 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*, ed. Iryna Gurevych & Torsten Zesch, 19–27. Singapore: Association for Computational Linguistics.

Nielsen, Michael A. (2015). *Neural networks and deep learning*. Downloadable at http://neuralnetworksanddeeplearning.com/ (accessed 5 March 2020).

Ohala, John J. (1993a). Coarticulation and phonology. *Language and speech* 36: 155–70.

Ohala, John J. (1993b). The phonetics of sound change. In *Historical linguistics*, ed. Charles Jones, 237–78. London: Longman.

Pedregosa, Fabian, Gaël Varoquaux, Gramfort Gramfort, Vincent Michel, Bertrand Thirion, Grisel Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, & Édouard Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12: 2825–30.

R Core Team (2013). R: A language and environment for statistical computing. http://www.R-project.org/ (accessed 5 March 2020). Vienna: R Foundation for Statistical Computing.

Rasmussen, Jens Elmegård (1994). On the phonetics of Indo-European laryngeals. In *In honorem Holger Pedersen*, ed. Jens Elmegård Rasmussen, 434–47. Wiesbaden: Reichert.

Raymond, William, Robin Dautricourt, & Elizabeth Hume (2006). Word-medial /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language variation and change* 18: 55–97.

Ringe, Donald A. (2017). *From Proto-Indo-European to Proto-Germanic*[2]. Oxford: Oxford University Press.

Sanker, Chelsea (2016). Phonetic features of the PIE 'laryngeals': Evidence from misperception data of modern postvelars. In *Proceedings of the 27th annual UCLA Indo-European Conference*, ed. David M. Goldstein, Stephanie W. Jamison, & Brent Vine, 163–81. Bremen: Hempen.

Suzuki, Ryota, & Hidetoshi Shimodaira (2015). Pvclust: Hierarchical clustering with

p-values via multiscale bootstrap resampling. https://CRAN.R-project.org/package=pvclust (accessed 5 March 2020).

van Son, Rob J.J.H., & Jan P.H. van Santen (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech communication* 47: 100–23.

Weiss, Michael (2016). The Proto-Indo-European laryngeals and the name of Cilicia in the Iron Age. In *Tavet tat satyam: Studies in honor of Jared S. Klein on the occasion of his seventieth birthday*, ed. Andrew M. Byrd, Jessica DeLisi, & Mark Wenthe, 331–40. Ann Arbor: Beech Stave Press.

Zair, Nicholas (2018). On the relative sonority of PIE /m/. *Indo-European linguistics* 6: 271–303.

Zesch, Torsten, Christof Müller, & Iryna Gurevych (2008). Using Wiktionary for computing semantic relatedness. In *Proceedings of the twenty-third AAAI conference on artificial intelligence*. Vol. 8, 861–66. Menlo Park: AAAI Press.

A Appendix

TABLE 24 Network architecture for the German feature prediction task

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 25 Network evaluations and predictions for German *p*

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 26 Network evaluations and predictions for German *r*

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

TABLE 26 Network evaluations and predictions for German *r* (*cont.*)

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
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 27 Network evaluations and predictions for German *ɛ:*

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 28 Network evaluations and predictions for German *a:*

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

TABLE 28 Network evaluations and predictions for German *a:* (*cont.*)

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
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 29 Network evaluations and predictions for German *p* with training data size 6236

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
consonant	360	32	14	218	7222	1616
nasal	43	32	52	265	1720	7118
plosive	77	30	66	219	4661	4177
affricate	6	9	15	362	823	8015
fricative	65	38	61	228	2680	6158
liquid	65	26	75	226	2324	6514
sibilant	18	32	34	308	514	8324
voiced	176	51	53	112	4140	4698
labial	25	40	62	265	2635	6203
dental/alveolar	96	89	34	173	3230	5608
palatal	13	7	27	345	1127	7711
velar/uvular	70	46	71	205	2678	6160
glottal	2	4	32	354	938	7900

TABLE 30 Network evaluations and predictions for German *r* with training data size 6236

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
consonant	343	26	24	231	38626	1365
nasal	57	28	42	243	11453	28538
plosive	90	43	74	163	10614	29377
affricate	7	7	44	312	2497	37494
fricative	81	22	79	188	11227	28764

TABLE 30 Network evaluations and predictions for German *r* (*cont.*)

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
liquid	7	27	25	311	7757	32234
sibilant	16	33	42	279	3839	36152
voiced	149	33	100	88	34624	5367
labial	57	25	103	185	13715	26276
dental/alveolar	120	79	43	128	26155	13836
palatal	12	8	32	318	1030	38961
velar/uvular	39	22	59	250	7557	32434
glottal	4	4	38	324	4719	35272

TABLE 31 Network evaluations and predictions for German *ɛː* with training data size 6236

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
consonant	373	25	24	202	254	1588
front vowel	47	33	30	119	544	1298
central vowel	62	28	29	110	397	1445
back vowel	22	18	17	172	78	1764
round	14	14	36	165	255	1587
close	42	23	50	114	712	1130
mid	78	29	35	87	576	1266
open	22	35	44	128	820	1022
diphthong	9	10	24	186	454	1388
long	93	21	39	76	976	866

TABLE 32 Network evaluations and predictions for German *aː* with training data size 6236

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
consonant	371	35	11	207	926	7607
front vowel	47	33	39	96	4528	4005
central vowel	53	22	42	98	3290	5243
back vowel	20	20	27	148	2592	5941
round	17	13	32	153	2653	5880
close	34	32	19	130	2085	6448

TABLE 32 Network evaluations and predictions for German *a:* (*cont.*)

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
mid	83	23	33	76	3736	4797
open	21	22	27	145	1914	6619
diphthong	14	6	20	175	1556	6977
long	85	18	41	71	5296	3237

TABLE 33 Network evaluations and POA predictions for German *bilabial* sounds when POA *bilabial* is excluded

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
labiodental	1689	450	5725	16804	14932	15630
alveolar	9184	3655	3659	8170	14382	16180
postalveolar	1031	197	3754	19686	12865	17697
velar	5514	2416	4371	12367	13151	17411

TABLE 34 Network evaluations and POA predictions for German *velar* sounds when POA *velar* is excluded

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
bilabial	2254	802	5192	11546	38009	41295
labiodental	1697	442	5102	12553	41221	38083
alveolar	8799	4041	1602	5352	48279	31025
postalveolar	957	271	2057	16509	38117	41187

TABLE 35 Network evaluations and POA predictions for German *labiodental* sounds when POA *labiodental* is excluded

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
bilabial	2387	669	7368	15162	14179	7210
alveolar	8759	4081	3557	9189	8767	12622
postalveolar	1028	200	3823	20535	4201	17188

TABLE 35 Network evaluations and POA predictions for German *labiodental* sounds (*cont.*)

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
velar	5206	2725	4409	13246	4964	16425

TABLE 36 Network evaluations and POA predictions for German *postalveolar* sounds when POA *postalveolar* is excluded

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
bilabial	2277	779	7507	15934	3568	8708
labiodental	1718	421	6839	17519	5189	7087
alveolar	8709	4131	3951	9706	6914	5362
velar	5410	2521	4705	13861	4089	8187

TABLE 37 Network evaluations and POA predictions for German *alveolar* sounds when POA *alveolar* is excluded

Feature	TP	FN	FP	TN	Pred: feat. present	Pred: feat. absent
bilabial	2375	681	3760	8069	71761	56632
labiodental	1703	436	3506	9240	63527	64866
postalveolar	980	248	2161	11496	61779	66614
velar	5790	2141	1483	5471	59845	68548

TABLE 38 Phonetic feature assignment for each New High German sound

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

TABLE 38 Phonetic feature assignment for each New High German sound (*cont.*)

sibilant	s, z, ʃ, ʒ
voiced	m, n, ŋ, b, d, g, ʒ, z, v, l, r, j
voiceless	p, t, k, pf, tʃ, ʃ, ts, 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, j
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
sonorant	m, n, ŋ, l, r, j, ɪ, i, ɛ, e, y, ʏ, ø, œ, æ, ə, a, ɑ, u, ʊ, o, ɔ, a̯ɪ, a̯ʊ, ɔ̯ʏ
occlusive	p, b, t, d, k, g, m, n, ŋ, pf, ts, tʃ
continuant	f, v, s, z, ʃ, ʒ, x/ç, h, l, r, j, ɪ, i, ɛ, e, a̯ɪ, ɔ̯ʏ, 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, ts, 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	a̯ɪ, a̯ʊ, ɔ̯ʏ
open diphthong	a̯ɪ, ɔ̯ʏ
mid diphthong	a̯ʊ
front diphthong	a̯ɪ, a̯ʊ
back diphthong	ɔ̯ʏ
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 39 Phonetic feature assignment for each PIE sound

root ending	-
word boundary final/initial	zero

TABLE 39 Phonetic feature assignment for each PIE sound (*cont.*)

voiced	*bʰ, *dʰ, *ǵʰ, *gʰ, *gʷ, *b, *d, *ǵ, *g, *gʷ, *m, *m̥, *n, *n̥, *r, *r̥, *l, *l̥, *y, *w
voiceless	*p, *t, *s, *ḱ, *k, *kʷ
nasal	*m, *m̥, *n, *n̥
aspirated	*bʰ, *dʰ, *ǵʰ, *gʰ, *gʷʰ
labial/labialized	*m, *m̥, *p, *b, *bʰ, *w, *kʷ, *gʷ, *gʷʰ
sibilant	*s
liquid	*r, *r̥, *l, *l̥
syllabic consonant	*r̥, *l̥, *m̥, *n̥, *i, *u, *ū
coronal	*n, *n̥, *t, *d, *dʰ, *s, *r, *r̥, *l, *l̥
postvelar	*k, *g, *gʰ, *kʷ, *gʷ, *gʷʰ
velar	*ḱ, *ǵ, *ǵʰ
palatal	*y
front vowel	*e, *ē, *i
back vowel	*o, *ō, *u, *ū
central 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, *m̥, *n, *n̥, *r, *r̥, *l, *l̥, *y, *w, *p, *t, *s, *ḱ, *k, *kʷ
back consonant	*k, *g, *gʰ, *kʷ, *gʷ, *gʷʰ, *ḱ, *ǵ, *ǵʰ
front consonant	*m, *m̥, *p, *b, *bʰ, *w, *s, *r, *r̥, *l, *l̥, *n, *n̥, *t, *d, *dʰ
stop	*p, *b, *bʰ, *ḱ, *ǵ, *ǵʰ, *k, *g, *gʰ, *kʷ, *gʷ, *gʷʰ, *t, *d, *dʰ
obstruent	*p, *b, *bʰ, *ḱ, *ǵ, *ǵʰ, *k, *g, *gʰ, *kʷ, *gʷ, *gʷʰ, *t, *d, *dʰ, *s
sonorant	*m, *m̥, *n, *n̥, *r, *r̥, *l, *l̥, *y, *w, *e, *o, *u, *a, *i, *ē, *ō, *ū, *ā
occlusive	*p, *b, *bʰ, *ḱ, *ǵ, *ǵʰ, *k, *g, *gʰ, *kʷ, *gʷ, *gʷʰ, *t, *d, *dʰ, *m, *m̥, *n, *n̥
continuant	*s, *y, *w, *r, *r̥, *l, *l̥, *e, *o, *u, *a, *i, *ē, *ō, *ū, *ā

Note that *$\bar{\imath}$ is not featured in Table 39 as words containing it were not included in the dataset as provided by Wiktionary.

A.1 *Standard model architecture*

For all networks, except for the [±nasal] detection task, the following model specifications were used. During network optimization, different architectures were tested for individual tasks, none of which enhanced the performance reliably as they either increased false positives or false negatives. The model presented here represents the best possible model.

TABLE 40 Standard model architecture

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

As the optimizer, *Adam* was used with a learning rate of 0.01 and a batch size of 25.

A.2 *Model architecture for* [± *nasal*]

TABLE 41 Network architecture for the feature *nasal*

Layer	Layer size	Activation
Dense layer 1	364	ReLU
Dropout layer 1	0.25 dropout rate	
Dense layer 2	32	ReLU
Dropout layer 2	0.25 dropout rate	
Dense layer 3	364	ReLU
Output layer	2	softmax

As the optimizer, *Adam* was used with a learning rate of 0.01 and a batch size of 15.

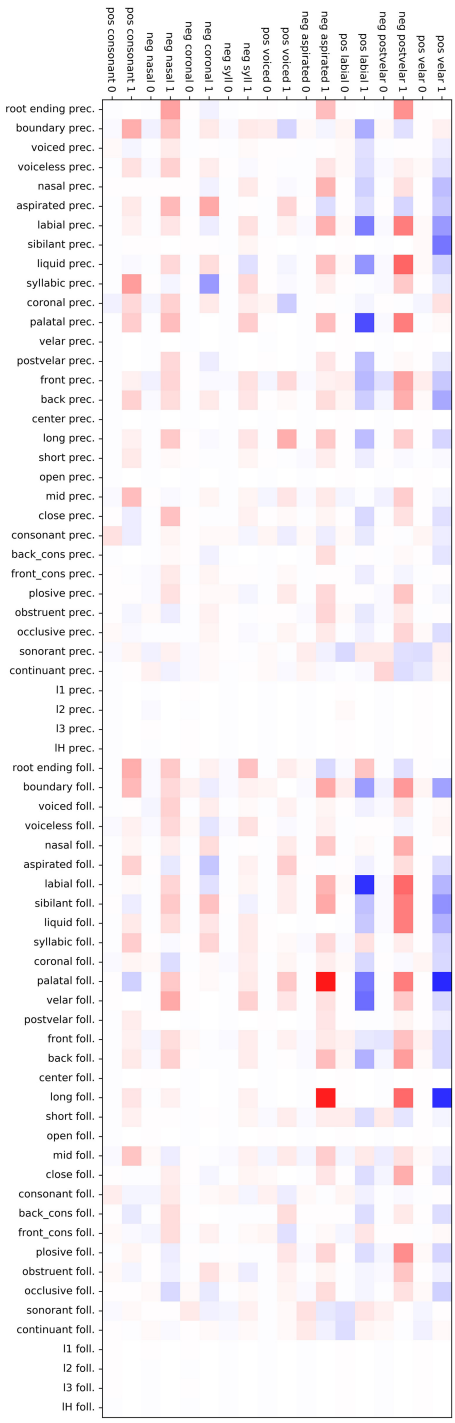

FIGURE 5
Visualization of approximated *SHAP* values for the prediction of $*h_1$

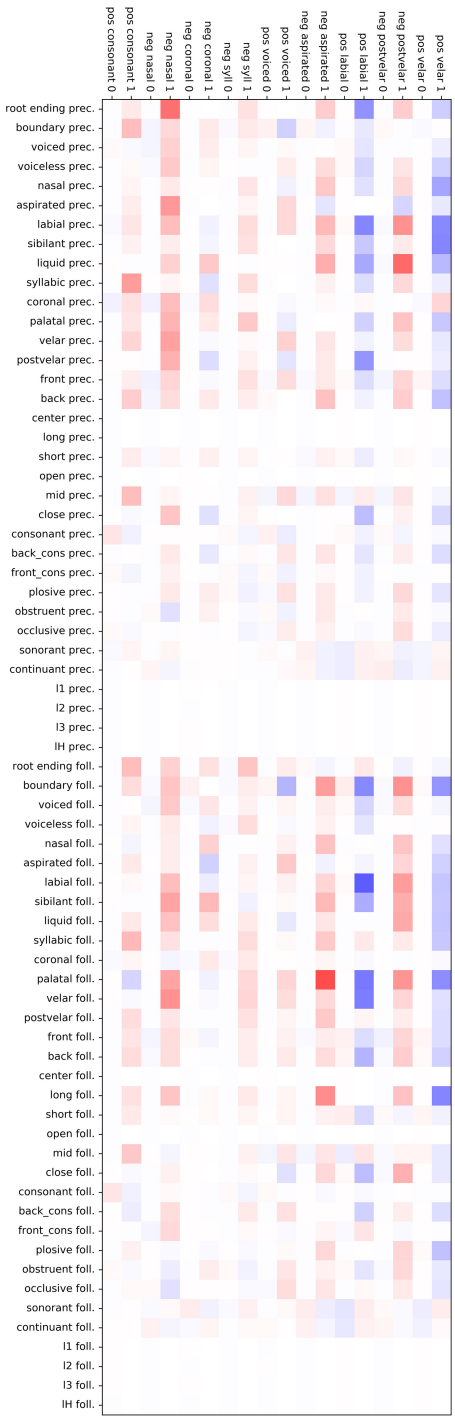

FIGURE 6
Visualization of approximated *SHAP* values for the prediction of $*h_2$

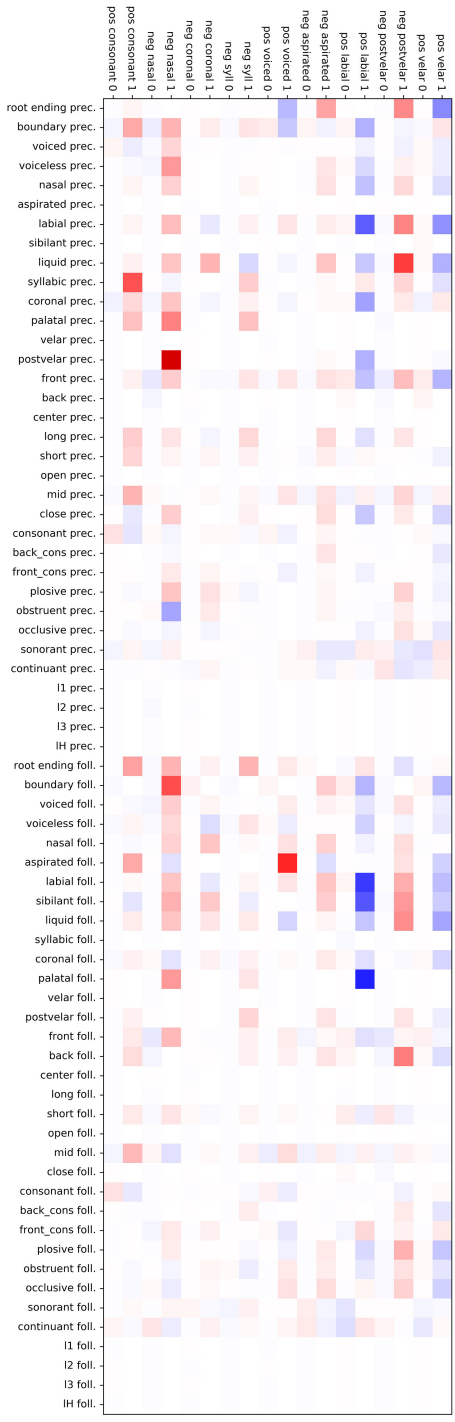

FIGURE 7
Visualization of approximated *SHAP* values for the prediction of $*h_3$