The organization of sound inventories: A study on obstruent gaps

Sheng-Fu Wang
Dept. of Linguistics
New York University
10 Washington Place
New York, NY 10003
shengfu.wang@nyu.edu

Abstract

This study explores the organizing principles of sound inventories by examining attested one-segment gaps in obstruent inventories. Models based on different theories of inventory organization are built and compared in a computational task where models make a binary decision to identify gaps and attested sounds. Results show that segment markedness, defined either in terms of grounded phonetic properties or typological frequencies, is a good predictor of whether a segment is likely to be gapped in an inventory. On the other hand, whether an attested segment, compared to a gapped segment, makes the feature representation more symmetric or economical, is not a good predictor of whether a segment is gapped. Finally, artificial neural networks that take inventories and segments as bags of feature values outperform all aforementioned models, demonstrating the extent to which the task of gap identification is learnable from distributional properties in the data.

1 Introduction

Sound inventories of human languages are not made up of random bags of possible speech sounds. Proposals on inventory shapes such as feature economy and symmetry (Clements, 2003a) predict that a stop inventory like /p, t, b, d/ is more likely than /p, t, b/ or /p, t, b, g/. Proposals on markedness and implicational universals (Greenberg, 1970; Gamkrelidze, 1975) also make predictions on inventory shapes. For example, when a language lacks a voiced stop, it is more likely to be /g/ than /d/. When it comes to inventory shapes, such proposals predict that /p, t, b, d/ is more likely than /p, t, b, g/.

This study compares different models on the organizing principles of inventory structure by examining the distribution of gaps in the inventories of obstruents across languages. A gap is referred to as the absence of an [α voice] stop/fricative in a certain place of articulation when a [$-\alpha$ voice] counterpart exists in the inventory. The computational task is a binary choice task, adapoted from the cloze task in Cotterell and Eisner (2017): between two sounds, the models have to decide which is the 'gap', and which is the 'foil', an attested stop/fricative that shares the [α voice] feature with the gap, with a different place of articulation. A model is more successful if it more frequently identifies the gap correctly. Examples of gaps and corresponding foils in Wogeo (Exter, 2003) is shown in (1)

(1) Inventory of obstruents in Wogeo

b	d	g
–	t	k
v	–	
f	s	

gap: /p/ foil: /t, k/
gap: /z/ foil: /s/

This task can be seen as an approximation of the task of identifying phoneme categories by a human learner: The process of acquiring a phonemic inventory can be modeled as decisions to construct abstract categories based on input that has certain distributional property in an available articulatory, acoustic, and/or perceptual space (e.g., Dillon et al., 2013; McMurray et al., 2009; Vallabha et al., 2007). Take word-initial stops for example, it has been found that sounds that differ only in place of articulation differ on a continuum of time-varying spectral properties 20-40 ms after the stop release (Kewley-Port et al., 1983; Kewley-Port, 1983). A human learner has to form categories along the such dimensions, based on a mixture of bottom-up and top-down information. In forming categories, the learner makes latent decisions on whether some given types of sounds be-

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 195-204.
New York City, New York, January 3-6, 2019

long to a certain emergent or emerging category, and whether sounds of certain types make up a distinctive category on their own.

Two major types of models are tested: The MARKEDNESS and the FEATURE-SYSTEMIC models. The MARKEDNESS models predict that the gap is always more marked than the foils. There are two variants of the MARKEDNESS model in this study: The *grounded markedness* model ranks the markedness of obstruents based on their constriction site (Gamkrelidze, 1975): A voiced obstruent is more marked when the constriction site is further back in the vocal tract. Conversely, a voiceless obstruent is more marked if it is fronter. The *typological markedness* model uses the frequency of segments and places of articulation across inventories for ranking markedness: the less frequent sounds are more marked.

For FEATURE-SYSTEMIC models, the gap is the sound whose presence in an inventory would decrease the overall goodness of the inventory based on some feature-based metrics. Three metrics are used in this study: *feature entropy* (Mukherjee et al., 2007), *local feature symmetry*, and *global feature symmetry* (Dunbar and Dupoux, 2016). *Feature entropy* is taken as a measure of feature economy: the feature representation of a system is more economical if it can be expressed by fewer bits. *Local feature symmetry* measures the number of pairs of sounds in an inventory that differ only in one feature, and an inventory is more locally symmetric if it has more such pairs. *Global feature symmetry* measures the difference in the size of the plus and the minus feature values in a feature system: an inventory is more globally symmetric if the difference is smaller.

To further investigate to what extent the place of articulation of gaps can be learned from the distribution of segments across inventories, artificial neural networks are trained to do the gap-prediction task. The input contains inventories represented as bags of segments, which in turn are represented as bags of feature values. The training objective to either choose a gap from two sounds given the knowledge of the inventory at issue (*Inventory* model), the gap and the foil (*Segment* model), or both (*Inventory+Segment* model).

Results show that MARKEDNESS models can accounts for 65%-59% of the gapping patterns. In other words, the gap is often more marked than the foils (the attested sounds) either in terms of their places on markedness scales defined with reference to speech production, or in terms of typological segment frequencies. On the other hand, the FEATURE-SYSTEMIC models have worse performance on average, showing that the decision between a foil and a gap is not actively governed by principles on the optimality of feature representation. To locate a potential domain where these FEATURE-SYSTEMATIC models may be active, a supplement experiment is done and the results show that some of these Feature-Systematic models are able to differentiate inventories with and without gaps. Finally, artificial neural networks are able to be trained to perform better than other types of models in this task, and the architecture that utilizes information both on the inventory as a whole and on gaps and foils performs the best.

The rest of the paper is organized as follows: Section 2 reviews theories on inventory structure in a greater detail. Section 3 describes the data, task, and models in this study. Section 4 reports results and analyses. Section 5 discusses the findings and concludes the paper.

2 Theories on Inventory Structure

2.1 Segment Markedness

Gamkrelidze (1975) explicitly discusses how the markedness of sounds affects the organization of sound inventories. He proposes scales of markedness at the segmental level, shown in (2), and discusses how the scales account for attested and unattested types of stop and fricative inventories. The presentation of gaps in the inventory serves as an important way to demonstrate his main points. He proposes that voiced stops and voiceless stops have the opposite markedness scales: whereas labial stops are marked when they are voiceless, they are unmarked when they are voiced, as shown in (2). He also states that the markedness scales with respect to voicing are the same for fricatives and stops. Within voiceless stops, aspirated and unaspirated ones share the same scale.

(2) The markedness scales in (Gamkrelidze, 1975)

marked	↔	unmarked
p	t	k
g	d	b
f	s	x
ɣ	z	v

These markedness scales are used to predict attested and unattested inventory types. The assumption is that the presence of a marked element predicts the presence of all less marked elements. On the other hand, inventories where the certain sounds exist while the less marked sounds are absent are predicted to be unattested. Examples are shown shown in (3).

(3) Inventory shapes that are predicted to be attested and unattested in Gamkrelidze (1975)

a. Attested inventory shapes

b	d	g	b	d	–
–	t	k	p	t	k

b. Unattested inventory shapes

b	d	g	–	d	g
p	t	–	p	t	k

The markedness scales are described to be derived from frequency counts of segments both within and across inventories. However, these scales can also be interpreted as being motivated from the aerodynamics of voicing and voicelessness in different places of articulation, especially concerning stops (Greenberg, 1970; Smith, 1975; Ohala, 1983). For voiced stops, a constriction site that is further back in the vocal tract makes the stop less optimal, since it is difficult to maintain voicing due to smaller space for air-pressure buildup behind the constriction site. This motivates why velar stops are more marked than bilabial stops. For voiceless stops, When the constriction site is further back in the oral tract, the smaller volume of cavity behind the constriction site makes it easier to build up air pressure, resulting in a stronger amplitude for the burst of the release, making the voiceless stop more perceptually salient. This motivates why bilabial voiceless stops are more optimal than velar voiceless stops.

It should be noted that even though the tendency to miss /p/ and /g/ both have aerodynamic motivations, the extent to which these motivations play a role in shaping inventory patterns has been questioned. Maddieson (2013) shows that inventories that miss /p/ have geographically concentrated distribution around the Sahara desert, where the major language families are Niger-Congo, Nilo-Saharan, and Afro-Asiatic. He argues that such clustering of these languages suggests that the aero-dynamic motivation for inventories to miss /p/ may not be valid, and areal factors may play a better role in explaining the occurrence of such inventories. Even though the present study does not seek to address the issue of areal factors in inventory shapes, the question that Maddieson raises is still relevant, as it suggests that the explanatory power of grounded markedness, thus defined, may be weaker for voiceless stops.

2.2 Feature system and inventory structure

The notion of feature economy, according to Clements (2003b), can be dated back to de Groot (1941) and Martinet (1955). It is proposed as a basic principle of sound system organization. It refers to a tendency to maximize the number of segments that can be represented per feature dimension, as shown in (4), where E refers to the economy index, S refers to the number of segments in an inventory, and F refers to the number of feature dimension necessary for representing all segments in an inventory.

(4) E = S/F

There have been studies that examine feature economy in attested inventories. Mackie and Mielke (2011) finds that attested inventories in P-base (Mielke, 2008) are more economical than randomly generated ones, based on four different kinds of economy metrics. Dunbar and Dupoux (2016) also have similar findings with similar simulation-based methodology.

Feature symmetry is another feature-based principle of inventory structure. It states a preference for sounds in an inventory to have symmetric distribution along feature dimensions. Thus it prefers an inventory /p, t, b, d/ over an inventory such as /p, t, b/. In other words, it can also be restated as a dispreferrence for having gaps. Clements (2003b) states that feature symmetry may be conceptualized as a tendency for languages to avoid having gaps in their inventory. This notion of symmetry is further developed and tested by Dunbar and Dupoux (2016), where they propose two non-equivalent but related symmetry metrics: local symmetry and global symmetry.

Local symmetry in an inventory refers to the notion that the number of 'oppositions' in the inventory is relatively low or high. An opposition refers to a pair of sounds that differ only in one feature. An inventory is more locally symmetric if the number of oppositions is high. Global symmetry, on the other hand, refers to whether an inventory is well-balanced among feature dimensions.

It measures whether the inventory has an imbalanced number of [+] and [−] values along all feature dimensions. It can be calculated by taking each non-redundant feature, calculating the difference between number of sounds with [+] and with [−], and summing the values across different features, divided by number of features. A lower value indicates greater global feature symmetry.

2.3 Segment Co-occurrence

Mukherjee et al. (2007) and a series of studies (Choudhury et al., 2006; Mukherjee et al., 2009) approach the issue of principles in inventory shapes by modeling the co-occurrences of segments in network models. In these models, each node represents a segment, and the weight of a node is the number of languages with that segment. The weight of the edge between two nodes is the number of languages with both segments. With actual consonant inventories from UPSID (Maddieson and Disner, 1984), the model is able to group consonants into communities that contain homogeneous sounds such as dental, retroflex, and laryngealized sounds. In addition, they also measure 'feature entropy' within each community, which calculates how many bits are needed to transfer the feature representations segments in group of sounds. They compare the feature entropy of communities formed from attested and randomly generated inventories, and find that network drawn and weighted from attested inventories have communities with lower feature entropy. The finding shows that there are some regularities in the co-occurrence pattern of sounds across inventories.

3 Method

3.1 Data

This study uses the PHOIBLE database of phoneme inventories (Moran et al., 2014), which contains 2155 inventories, where the segments are described by a phonetically detailed feature set. Only [−soronant] sounds are used in this study, which limits the scope of the study to a natural class of sounds that are more homogeneous.

After filtering out repetitive inventories, 1874 obstruent inventories remain. From these obstruent inventories, 'gaps' are identified by examining if an inventory lacks [α voice] stops and fricatives in certain places of articulation when the [$-\alpha$ voice] counterpart exists. The corresponding 'foils' are identified, which refer to attested sounds in an inventory that share the same [α voice] feature with the gap but with different places of articulation.

Example data points from Wogeo are shown in (5), along with the obstruent inventory. Three data points are generated from this inventory.

(5) Inventory of obstruents and data points in Wogeo

b	d	g
–	t	k
v	–	
f	s	

gap: /p/ foil: /t, k/
gap: /z/ foil: /s/

data point I: gap-/p/, foil-/t/
data point II: gap-/p/, foil-/k/
data point III: gap-/z/, foil-/v/

The computational task is an adaptation of the cloze task in Cotterell and Eisner (2017): a model sees each pair of sounds and labels one of the sounds as a gap and the other as a foil based on different strategies that each model employs. The success of the models is measured by how often the labels given by the model fit the actual data.

Models that require training, including the *typological markedness* model and the neural network models, are trained on data points from 70% of the inventories (training set) and tested on data points from 20% of the inventories (test set). The neural network models' hyperparameters are tunes with data points from 10% of the inventories (development set).

3.2 Models

3.2.1 Markedness Models

Two Markedness models are included in this study. The first one is *grounded markedness*, which compares segments based bases on their positions on predefined markedness scales. In the current study, the scales are expanded from the ones described in Gamkrelidze (1975), where stops and fricatives share the same scales, and voiced and voiceless sounds have inverse scales. The scale for voiced stops and fricatives are shown in (6).

(6) Markedness scale for places of articulation in voiced obstruents, presented in constraint ranking.

*pharyngeal ≫ *uvular ≫ *velar ≫ *labial-velar ≫ *palatal ≫ *retroflex ≫ *post-alveolar ≫ *alveolo-palatal ≫ *alveolar ≫ *dental ≫ *labiodental ≫ *bilabial

In the computational task, the model decides that the more marked segment on the corresponding scale is the gap for a data point. For example, for data point III, /z/-/v/, from Wogeo, shown in (5), the *grounded markedness* labels /z/, a voiced alveolar sound, as a gap, since the avoidance of an alveolar voiced sound (*alveolar) ranks higher than the avoidance of a labiodental sound (*labiodental).

The other markedness model, the *typological markedness* model, only takes into account the frequencies of segments or places of articulation across inventories in the task. In the computational task, the model labels the typologically less frequent sound as the gap. This model has two further variants: the frequency can either be calculated based on the typological frequency of segments or of places of articulation.

3.2.2 Feature-Systemic Models

Three models fall into this category: *global feature symmetry*, *local feature symmetry*, and *feature entropy*. All these models rely on finding a minimal feature set that is required to represent an inventory, thus the algorithm to arrive at the minimal feature set is crucial. In this study, this is done by first ranking the entropy of each feature in the training/development set. In the process of shrinking the feature set, features with lower entropy are removed first, until the point where the removal of any feature would prevent all segments in an inventory to be unique represented. As a result, the resulting minimal feature sets are more likely to contain features with high entropy, which are features with a more balanced use of [+] and [−] values in the data set.

For each of the FEATURE-SYSTEMIC models, at each data point, the algorithm to find a minimal feature set is applied twice: once to the attested inventory, and once to the 'alternative' inventory, where the foil is replaced by the gap in the attested inventory. For example, for the data point /t-p/ from the inventory of Wogeo, the algorithm will find the minimal feature set for /b, d, g, **t**, k, v, f, x/ (the attested inventory) and for /b, d, g, **p**, k, v, f, x/.

The three FEATURE-SYSTEMIC models can then be seen as three different metrics to score the attested and alternative inventories at a data point. When the score for a metric favors the attested inventory, the model 'scores' at this data point.

The metric for *local feature symmetry* is the number of pairs of sounds that differ only in one feature in a system. As mentioned earlier, such a pair is referred to as an 'opposition' The metric for *global feature symmetry*, on the other hand, is the averaged absolute difference between the number of sounds with the plus and the minus values in all feature dimensions.

Feature entropy uses the metric proposed in Mukherjee et al. (2007), which measures the homogeneity of segment 'communities' in a network model. It can be seen as an information-theoretic measure of the minimal bits required to convey the feature representation of a set of sounds. The metric is calculated as follows: for an inventory with N segment types that are represented with the feature set F, which contains multiple features f, where the number of segments with feature value p is referred to as p_f and the number of segments with feature value q is referred to as q_f, feature entropy is $\sum_{f \in F}(-\frac{p_f}{N}\log_2\frac{p_f}{N} - \frac{q_f}{N}\log_2\frac{q_f}{N})$. The interpretation of this metric is that an inventory is more economical if it has a more skewed distribution of [+] and [−] values or/and it has fewer required features in the feature system.

3.3 Artificial Neural Network

Artificial neural network models are built and tested to see whether the shape of obstruent inventories is governed by some underlying co-occurrence principle between features and segments.

The training objective approximates the task for non-neural models: each item in the input to the model is a pair of inventories. One of them is an attested inventory, and the other is the attested inventory with a foil being replaced by a gap. The task is for the model to decide which of these inventories is the attested one. It is implemented as a binary decision task. Each data point appears twice in two different orders.

Segments are represented as bags of feature values. Each feature value (e.g., [+voice], [−voice], [+labial], etc) is represented by a 10-dimensional vector that is randomly initialized before training and updated through back propagation during

training. The representation for a segment, also a 10-dimensional vector, is derived in the following way: The vector representations for its feature values are first averaged. Then, the averaged vector passes through a perceptron layer with a rectified linear unit (ReLU) to obtain the vector representation for the segment. This gives the model to capture more complicated association between feature representations and segment representation. The output representation for each segment is then summed element-wise to obtain the representation for the inventory, which is also a 10-dimensional vector.

To investigate the role of inventory-level and segment-level information in performing this task, the are three variant architectures. In the *Inventory* architecture, the representations for all segments in the attested and the alternative inventories are summed before being concatenated and passed through two layers before a softmax layer for the classification task, as shown in Figure 1. In the *Segment* architecture, the 10D representations for the gap and the foil are concatenated, while in the *Inventory+Segment* architecture, the inventory representation is still the sum of segment representations, but the 10D representation of the gap and the foil, in stead of being part of the sum of the representation for the alternative and the attested inventory, are concatenated to the inventory representation (thus yielding a 40D representation before being passed into the hidden layers). This ensures that the model has explicit access to the property of the segments that are crucial for the task.

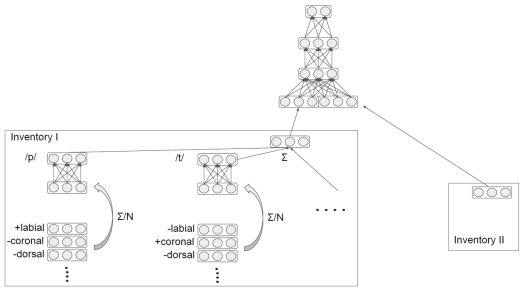

Figure 1: The *Inventory* architecture of the neural network model

For all three model architectures, hyperparameter tuning was done by a random search through 100 combinations of size of embedding dimension (between 10 and 50), learning rate (between 0.001 and 0.005) and L2 regularization weight decay (between 1×10^{-5} and 9×10^{-4}). During training, training and development set accuracies are evaluated every 100 steps. Training stops when the development accuracy does not improve after 500 steps. The best model in terms of development set performance for each architecture is used to report the result on the test set.

All the aforementioned training and testing procedures are repeated ten times over ten different random divisions of the data into training, development, and test sets. For comparison across different models, the performance of the MARKEDNESS and the FEATURE-SYSTEMIC models are also broken down into three different sets, with the test set being used for comparison.

4 Results

The overall results are shown in Table 1. Among the non-neural network models, MARKEDNESS models generally outperform the FEATURE-SYSTEMIC models. Among the former, the *typological markedness* model is the best performing one. In other words, the results shows that two third of the time the gap is less frequent than the foil. *Grounded markedness* ranks the second, showing that 62.13% of the time, the gap is more marked on the markedness scales than the foil. Another variant of the *typological markedness* model, based on the frequencies of places of articulation, ranks the third just behind grounded markedness.

The best performing FEATURE-SYSTEMIC model is the *global symmetry* model, achieving a mean accuracy of 55.91% in the test set. The *local symmetry* model is almost at chance level, while the *feature entropy* model is below chance, having a mean accuracy of 46.36%.

The neural network models are the best performing ones. The *Inventory+Segment* model, which concatenates the inventory-level summed embeddings with the embeddings of the gaps/foils for the classification task, performs the best, achieving a mean accuracy of 82.75% in the test set. The *Segment* model that only uses the embeddings of gaps and foils perform worse, with a mean accuracy of 74.89%, but is better than the performance of the *Inventory* model, which only takes the inventory-level summed embeddings. The fact that the *inventory* model can perform relatively well shows that this task can be solvable to a certain extent by information pro-

Table 1: Overall Results in average accuracy across ten different data splits. Numbers in the parentheses show standard deviations. Cells with gray numbers show that the corresponding models do not have a training component and the numbers merely indicate results of calculation in train/dev sets

Model	train	dev	test
grounded markedness	62.32 (0.57)	62.29 (2.61)	62.13 (2.65)
typological markedness – segment	66.13 (2.99)	66.91 (2.35)	65.47 (1.62)
typological markedness – place of articulation	59.49 (0.62)	60.94 (2.58)	59.36 (1.76)
feature entropy	48.18 (2.99)	46.21 (2.83)	46.36 (1.62)
local symmetry	50.29 (0.39)	48.61 (4.07)	50.38 (0.92)
global symmetry	54.95 (0.53)	54.42 (1.88)	55.91 (2.08)
NN: inv	72.21 (3.38)	73.75 (1.77)	72.16 (2.19)
NN: inv+seg	83.75 (2.35)	83.64 (2.38)	82.75 (2.51)
NN: seg	75.73 (0.60)	78.43 (2.14)	74.89 (1.09)

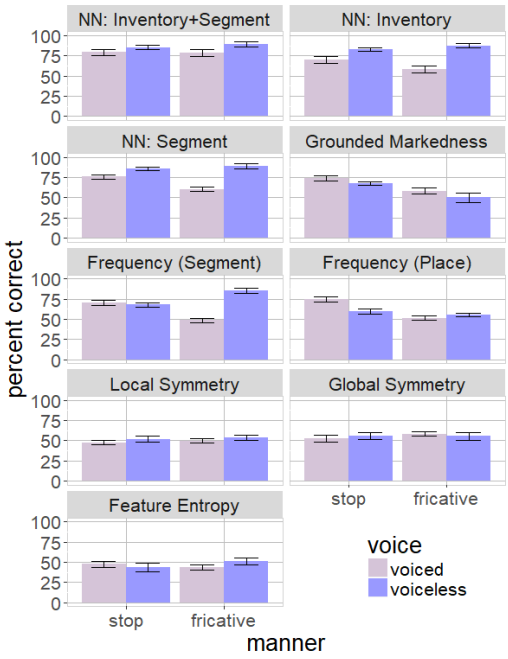

Figure 2: Model performance as a function of sound types. Note that 'Frequency (Segment)' and 'Frequency (Place)' refers to the *typological markedness* model's variants based on segments and places of articulation.

vided by the inventory as a whole. The low performance of the FEATURE-SYSTEMIC models do not really show that inventory-level information is not important for the task. It only shows that the right kind of information is not extracted via the feature-systemic measures.

Figure 2 shows the test-set results for all models, broken down in terms of segment types. Three major trends are worth noting. First, the *grounded markedness* model performs better in stops, especially for voiced stops. This is not surprising given that the grounded markedness scales are better motivated in terms of aerodynamic for stops than fricatives, and the pressure to maintain voicing in stops is considered to have stronger correlation with constriction site in the vocal tract.

Second, the *typological markedness* models have comparable performances as the *grounded markedness* models in stops, but the *typological markedness* model vastly outperforms the *grounded markedness* model in the subset with voiceless fricatives. There is a noticeable difference between the segment-based and the place-based *typological markedness* models in voiceless stops. The major reason is that since labiodental sounds are overall less frequent when compared with bilabial, alveolar, and velar sounds, the place-based *typological markedness* model does not account for the fact that /f/ is a frequent fricative. Being a frequent sound among fricatives, /f/ frequently serves as a foil, and these data points can scored by simply taking into account segment-level typological frequencies. On the other hand, these *typological markedness* models struggle with voiced stops, presumably because of cases that involve /z/: /z/ is a frequent segment, and alveolar is a frequent place of articulation. However, since /s/ is a lot more frequent than /z/, it is very frequent for an inventory to have /s/ but not /z/; when that happens and when the inventory has a voicing distinction elsewhere in its fricatives, /z/ would be a gap, and the *typological markedness* models would struggle in such cases because the high-frequency /z/ should actually be a gap.

Finally, the result of neural network models are similar to frequency-based models, especially in the discrepancy of performance between voiceless and voiceless fricatives. The *Inventory+Segment* model is able to narrow down the discrepancy,

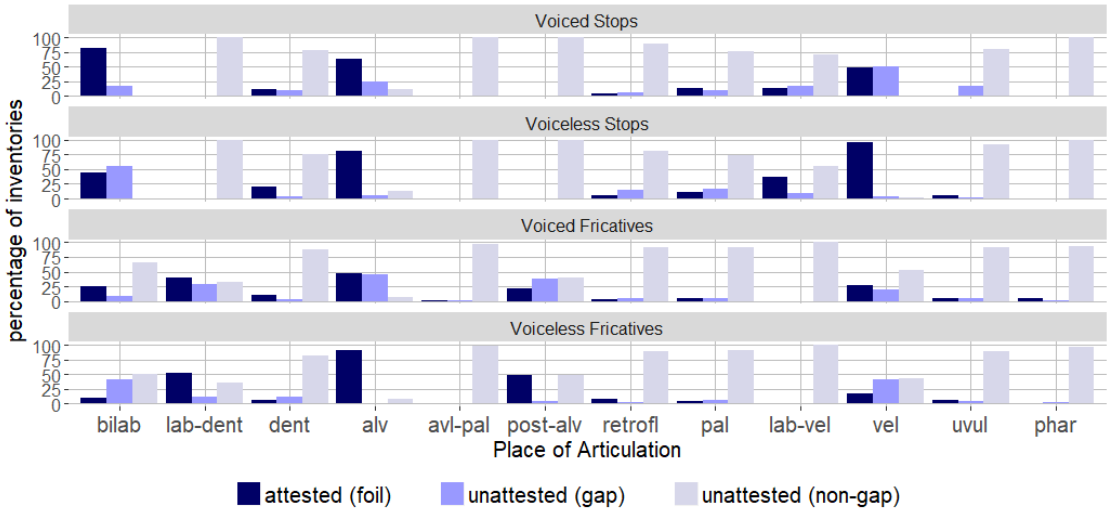

Figure 3: Distribution of attested sounds and gaps across places of articulation in the inventories in the data set

showing that it is possible to solve this problem by taking sophisticated statistical patterning at the level of both the inventory and thekey segments (the gap and the foil).

Figure 3 shows in each segment category, how often an inventory with a particular kind of gap has a gap and a foil in a particular place of articulation. For example, the dark purple bar in the 'bilab' column in the 'Voiced Stops' panel shows the following information: within inventories that have at least a gap in voiced stops, how often they have an attested bilabial voiced stop. The light purple bar shows how often do these inventories lack a bilabial voiced stop, which is defined as a gap in this study. Finally, the palest white bar, which is almost nonexistent for the 'bilab' column, shows the percentage of inventories that do not have a bilabial stop but the absence does not constitute a gap (i.e., the inventory also does not have a bilabial voiceless stop).

For voiced stops, foils are well-attested in the fronter regions of the vocal tract. Constriction in these regions supposedly makes maintenance of voicing easier. On the other hand, a more frequent distribution of gaps in the back region of the vocal tract would be advantageous for the *grounded markedness* model, which is true to a certain extent.

As for voiceless stops, the *grounded markedness* model prefers a distribution where gaps are in the front region of the vocal tract. This is again true to a certain extent, as seen in the second panel in Figure 3. For fricatives, the distributional patterns are not advantageous for the *grounded markedness* model: The gaps in voiced fricatives occur in the fronter region in the vocal tract, contradictory to how the markedness scales expect gaps to occur in the backer region. As for voiceless fricatives, the high frequency of velar gaps also goes against a preference for having gaps in the fronter region. On the other hand, as mentioned earlier, the concentration of attestedness in certain places of articulation for voiceless fricatives show why the *typological markedness* model has an advantage in this subset of data.

Due to the sub-par performance of feature-systematic models, a supplement experiment is conducted to test the following hypothesis: The feature-systematic measures do not account for where an inventory may have a gap. They simply measure a preference for inventories to not have a gap. To examine this, I compare the feature-systematic measures of obstruent inventories with and without the gaps that are investigated in this study: gaps in stops and fricatives.

The results are shown in Figure 4. To put all three measures in the same graph, the feature-systemic scores are z-transformed. Statistical tests show that the global symmetry metric is significantly lower in gapless inventories in the data set [$t(1871.9) = 5.91$, $p < .0001$], suggesting a better global symmetry for gapless inventories. Feature entropy is also shown to be significantly lower in gapless inventories [$t(1860.6) = 7.46$, $p < .0001$], suggesting a better feature economy for gapless inventories. These results may seem contradictory,

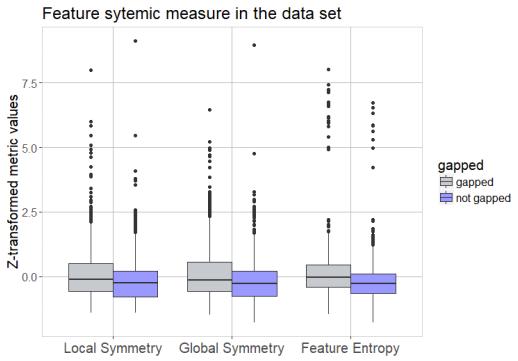

Figure 4: Comparison of gapped and non-gapped inventories in feature-systematic measures. The upper regions in the graphs indicate better symmetry and economy.

as feature entropy prefers the distribution of feature values to be skewed, while global symmetry prefers otherwise. Further inspection shows that gapless inventories requires a larger feature set [$t(1857) = 3.15$, $p < .01$], and this potentially explains why feature entropy and global symmetry show the same direction: it is possible that feature entropy for gapless inventories is brought down by requiring fewer features in the system. Local symmetry, on the other hand, shows an effect that is opposited as expected: gapped inventories have a significantly better local symmetry value than gapless ones [$t(1871) = 5.42$, $p < .0001$].

5 Discussion & Conclusion

This study shows whether a segment is more marked, either in terms of aerodynamics or typological frequencies, plays a role in deciding whether a segment is attested or gapped in an inventory. Crucially, the effectiveness of markedness in this task correlates with whether a particular markedness scale has a strong motivation in phonetic grounding.

Even though the problem of identifying gaps is more directly related to the larger question of inventory shapes, models that implement theories that directly address the nature or tendency of human sound inventories, such as feature economy or feature symmetry, do not perform well in this task. This may suggest that the drive towards a more efficient use of feature system, regardless of how it is construed, is not active in identifying what sounds should be phonemes in an inventory. However, it is possible that these theories account for inventory preference at a larger level, such as a preference for attested inventories over randomly generated sound inventories (Dunbar and Dupoux, 2016; Mukherjee et al., 2007). It remains to be seen whether the effectiveness of these metrics can be found at a level that is more accessible and related to phonological learning. As a starting point, the supplement experiment in this study has shown that these measures can differentiate inventories with and without gaps.

The performance of neural networks with different architectures complement the findings for MARKEDNESS and FEATURE-SYSTEMIC models. The good performance of the model that only looks at the key segments (i.e., the gap and the foil) shows that this task can be performed well to a certain extent by doing exactly what the MARKEDNESS models are doing, and with a more powerful statistical learner the result can be better. The performance of the Inventory model, which looks at inventories as a whole, also has good performance, showing that considering the inventory as a whole, without paying specific attention to the key segments, is also a viable way to perform this task. Finally, the *inventory+segment* model shows that the performance can be further improved when key segments are highlighted while also taking the whole inventories into account. It points to the fact that there are indeed usable information in the inventories at issue for deciding whether a segment is likely to be gapped in those particular inventories. The information is just not utilized by the FEATURE-SYSTEMIC models.

To conclude, this study shows that segmental properties and statistical patterns both play a role in shaping inventories in specific ways. In terms of methodology, this study makes two contributions: First, it demonstrates how a large database of phonemic inventories can be informative in answering phonological questions that may have ramifications for phonological learning. Seconds, it shows that models that are theoretically informed and models that only use statistical information can join force to show corroborating results. Finally, if the computational task in this study can be considered analogous to the task of phoneme identification for a human learner, this study suggests the potentially active role of segment markedness for stops, especially voiced ones, as well as the distributional patterns of feature values and segments, in learning phonemic inventories.

References

Monojit Choudhury, Animesh Mukherjee, Anupam Basu, and Niloy Ganguly. 2006. Analysis and synthesis of the distribution of consonants over languages: A complex network approach. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 128–135. Association for Computational Linguistics.

George N Clements. 2003a. Feature economy in sound systems. *Phonology*, 20(3):287–333.

Georges N Clements. 2003b. Feature economy as a phonological universal. In *15th International Congress of Phonetic Sciences, Barcelona, Spain*.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 37–46.

Brian Dillon, Ewan Dunbar, and William Idsardi. 2013. A single-stage approach to learning phonological categories: Insights from inuktitut. *Cognitive science*, 37(2):344–377.

Ewan Dunbar and Emmanuel Dupoux. 2016. Geometric constraints on human speech sound inventories. *Frontiers in psychology*, 7.

Mats Exter. 2003. Phonetik und phonologie des wogeo. Master's thesis, Institut fr Sprachwissenschaft Universitt zu Kln.

Thomas V Gamkrelidze. 1975. On the correlation of stops and fricatives in a phonological system. *Lingua*, 35(3-4):231–261.

Joseph H Greenberg. 1970. Some generalizations concerning glottalic consonants, especially implosives. *International Journal of American Linguistics*, 36(2):123–145.

Albert Willem de Groot. 1941. *Structural linguistics and phonetic law*. Noord-Hollandsche Uitgeversmaatschappij.

Diane Kewley-Port. 1983. Time-varying features as correlates of place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, 73(1):322–335.

Diane Kewley-Port, David B Pisoni, and Michael Studdert-Kennedy. 1983. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *The Journal of the Acoustical Society of America*, 73(5):1779–1793.

Scott Mackie and Jeff Mielke. 2011. Feature economy in natural, random, and synthetic inventories. *Where do phonological features come from*, pages 43–63.

Ian Maddieson. 2013. Voicing and gaps in plosive systems. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ian Maddieson and Sandra Ferrari Disner. 1984. *Patterns of sounds*. Cambridge university press.

André Martinet. 1955. Economie des changements phonétiques. *Berne: Francke*.

Bob McMurray, Richard N Aslin, and Joseph C Toscano. 2009. Statistical learning of phonetic categories: insights from a computational approach. *Developmental science*, 12(3):369–378.

Jeff Mielke. 2008. *The emergence of distinctive features*. Oxford University Press.

Steven Moran, Daniel McCloy, and Richard Wright. 2014. Phoible online. *Leipzig: Max Planck Institute for Evolutionary Anthropology*.

Animesh Mukherjee, Monojit Choudhury, Anupam Basu, and Niloy Ganguly. 2007. Modeling the cooccurrence principles of the consonant inventories: A complex network approach. *International Journal of Modern Physics C*, 18(02):281–295.

Animesh Mukherjee, Monojit Choudhury, Anupam Basu, and Niloy Ganguly. 2009. Self-organization of the sound inventories: Analysis and synthesis of the occurrence and co-occurrence networks of consonants. *Journal of Quantitative Linguistics*, 16(2):157–184.

John J Ohala. 1983. The origin of sound patterns in vocal tract constraints. In *The production of speech*, pages 189–216. Springer.

BL Smith. 1975. Effects of vocalic context, place of articulation, and speaker's sex on voiced stop consonant production. *The Journal of the Acoustical Society of America*, 58(S1):S61–S61.

Gautam K Vallabha, James L McClelland, Ferran Pons, Janet F Werker, and Shigeaki Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273–13278.