Phone2Vec

Michael Hammond
U. of Arizona

August 25, 2025

Abstract

In this paper, we investigate whether a generic word embedding model can be applied to phonetic segments to discover phonological features. The data come from English and we consider the question both from the direction of the classes that emerge from the model and from the perspective of the phonology: what classes do we expect to come from the model?

1 Overview

Here we investigate whether natural classes of segments can be deduced from the contexts they occur in. In technical terms we ask whether representing phones as compressed vectors (Mikolov et al., 2013a,b,c) can capture phonological categories based solely on distributional evidence.

In less technical terms, we know that phonotactic restrictions can be expressed in featural terms. For example, while an obstruent consonant can follow a sonorant consonant at the end of an English word, the opposite cannot occur. Thus:

(1)

Possible		Impossible
stamp	[stæmp]	*[...pm]
tent	[tʰɛnt]	*[...tn]
tank	[tʰæŋk]	*[...kŋ]
barf	[barf]	*[...fr]
tilt	[tʰɪlt]	*[...tl]

The question we address here is whether such phonotactic distributions are *sufficient* to learn the featural distinctions of a language like English.

When we consider the range of featural distinctions in English, we expect that those feature oppositions supported by phonotactics can be extracted, but distinctions only supported by alternations or typological evidence cannot. For example, as described above, the distinction between sonorant and obstruent consonants is supported by the distribution of word-final consonants (Hammond, 1999b) and thus should be

learnable. On the other hand, distinctions involving vowel height and backness are not supported by the phonotactics of English and thus should not be learnable. This focus distinguishes the work here from a number of previous investigations of similar techniques.[1]

Determining what phonological distinctions are learnable from compressed vectors is addressed by building vector models from the CMU pronouncing dictionary (Weide, 1998) and then testing them with *vector similarity* and *vector dissimilarity* tasks. We describe these in more depth below, but both techniques involve assessing the similarities and dissimilarities among the vectors that represent the phones of the language.

For vector similarity, we expect featurally similar sounds to have similar vectors. The specific task is to take a set of featurally similar sounds and ask what other sounds have similar vectors. Vector dissimilarity is a related task where a set featurally similar sounds and one outlier are given and we ask if the outlier can be identified based on its vector.[2]

The organization of this paper is as follows. First we introduce compressed vector models as they have been used with words, i.e. *word embeddings*. We then sketch out the phonological preliminaries, what the general feature space of English is. We then lay out how we apply compressed vectors to phones and the structure of our experiments. We then present our results, how features are and are not supported by distributional data. We then compare our results to previous work in this area. Finally, we conclude.

2 Embeddings

"You shall know a word by the company it keeps" (Firth, 1957, p.11).

Word embeddings are a particular language-modeling technique in natural language processing. The basic idea is to represent the semantics of a word in terms of a numerical vector that captures the distributional properties of that word. Thus if two words exhibit similar distributions, they will have similar vectors.

A particularly dramatic example of the utility of such representations comes from word embedding analogies (Mikolov et al., 2013c). The basic idea is that if you take the vector for a word like *king* in English, subtract the vector for *man* and add the vector for *woman*, you'll essentially end up with the vector for *queen*. Similarly, if you take the vector for *apples*, subtract the vector for *apple* and add the vector for *orange*, you end up with the vector for *oranges*.

There are a number of ways to create these vectors, but we will use the *skipgram* architecture of Mikolov et al. (2013a), Mikolov et al. (2013b), and Mikolov et al. (2013c).

[1] These are reviewed in Section 6.

[2] Note that our investigation proceeds from phonotactic restrictions and does not consider acoustic or articulatory similarity. See ... for an investigation that proceeds using only those.

This is an efficient and simple system that is implemented in the python `gensim` module (Řehůřek and Sojka, 2010).

The basic idea is to collect information on the distribution of words from corpus resources and then use that information to train a neural net. Then weights that have been trained in the net are extracted and used to generate vector representations for each word.

Let's understand this with a simplified schematic system. Imagine we have a language with only three words: $\{w_1, w_2, w_3\}$. We first represent each word using "1-hot" encoding. Each word is encoded as a sequence of zeros and ones where the length of that sequence is equal to the size of the vocabulary, in this case 3. The identity of a word can be determined by seeing which column contains a 1; all other columns will contain 0.

(2)

Word	Representation		
w_1	1	0	0
w_2	0	1	0
w_3	0	0	1

These encoded 1-hot vectors are used as input to a neural net which is trained to produce the probabilities of words that occur in the context of the word in question. Let's imagine we want to create a 2nd-order skipgram model with the set of words above. We would then use our neural net to predict 12 numbers from any input. These 12 numbers represent the probabilities for every word occurring as one of the two words preceding or following the input word:

(3) $w_i \Longrightarrow w_{i-2}\ w_{i-1}\ w_{i+1}\ w_{i+2}$

For example:

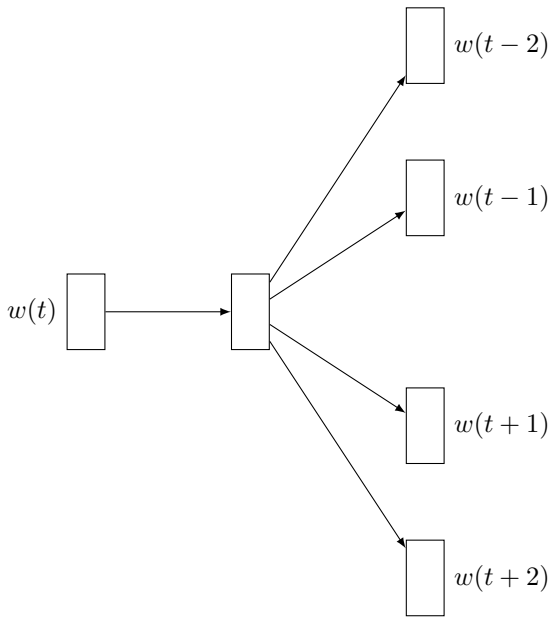
(4)

	Input	$\Longrightarrow$	Output	
w_i	0		$.2$	w_{i-2}
	1		$.1$	
	0		$.7$	
			$.5$	w_{i-1}
			$.2$	
			$.3$	
			$.1$	w_{i+1}
			$.7$	
			$.2$	
			$.8$	w_{i+2}
			$.1$	
			$.1$	

In the example above, we are using the word w_2 to predict the probabilities of the two preceding words and the two following words. For example, the probability of w_3 as the immediately preceding word is .3.

The interest of this process is in the structure of the neural net we use to accomplish it. This is schematized below. Here there are 3 input nodes corresponding to the 1-hot encoding of any word in the system. There are 12 output nodes corresponding to the probabilities of all words in the two preceding and two following positions (3×4), as exemplified above.

(5)



The network contains a hidden intermediate layer as well and it is here that the embedding is created. Typically, this hidden layer contains fewer nodes than the input layer. Thus, in this case, we might imagine that it contains two nodes.[3]

Once the network architecture has been settled on, the network is trained by presenting it with input words and the required output probabilities. Using a training regime like stochastic gradient descent, weights are adjusted in small increments until the best fit is obtained. The output layer is then removed and the weights connecting the input layer to the hidden layer remain and can be used to "look up" the

[3]Our schematic network is a bit unrealistic as we are usually scaling down from thousands of input nodes to a hundred or so hidden nodes.

hidden layer activation patterns generated from different inputs. These patterns are the word embeddings we are actually interested in.

The trained network thus associates each word of the vocabulary with a vector of numbers. These numbers are part of the calculation of the probabilities of context words and thus "represent" that context.

Following our schematic example a bit further, let's replace $\{w_1, w_2, w_3\}$ with $\{a, b, c\}$. We might have the following set of "sentences" in our training data:

(6)
a b c c
b b c c b a
a a a a
b a c b
c c b b a a
b b b b c c b a a

Using the python *gensim* module (Řehůřek and Sojka, 2010), we train a model with only two nodes in the hidden layer. This produces the following embeddings for each of our "words":

(7)

Word	Embedding	
a	0.255	0.45
b	-0.027	0.012
c	-0.465	-0.356

These vectors, or embeddings, can be taken as proxies for the meanings of our three words. We can assess the similarity between embeddings using *cosine similarity*, defined as follows:

(8)
$$\frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Cosine similarity values range from 1 to -1 with very similar vectors having a value close to 1 and dissimilar vectors having a value close to -1. For the three words above, we get the similarity calculations below:

(9)

Pair		Similarity
a	b	-0.101
a	c	-0.92
b	c	0.482

We will use cosine similarity to assess the similarity and dissimilarity of vector pairs or larger sets of vectors.

3 English features & phonotactics

In this section we review the features that distinguish English segments. Our intent is *not* to defend some specific featural representation and so our characterization of the features will be fairly general.[4]

The ultimate goal is to separate the featural distinctions into those that are supported by the phonotactics and those that are not. We do this because, as described above, we expect that our vectors will only distinguish featural classes that can be inferred from phonotactics.

Our experiments are based on the CMU Pronouncing dictionary (`cmudict`) (Weide, 1998). This resource provides a transcription of 127069 words. The transcription is described in the appendix and is fairly broad rhotic American. Predictable properties like aspiration, flapping, and vowel nasalization are not marked. Stress is marked such that any vowel is marked as having primary or secondary stress, or as having no stress. The transcription does distinguish [a] as in *cot* from [ɔ] as in *caught*. Including stress, the total number of distinct segments in `cmudict` is 69 and all segments are listed in the appendix.

Given this resource, we only consider features that distinguish between different segments there. We consider the following broad featural categories:

1. vowels vs. consonants
2. stress: primary, secondary, stressless
3. major class: obstruent, nasal, liquid, glide
4. manner: stop, fricative, affricate
5. voicing: voiced, voiceless
6. place: labial, coronal, dorsal
7. vowel length (tenseness)
8. vowel height: high, mid, low
9. vowel backness: front, back

For each of these, we demonstrate in Section 5 below whether the distinction is inferable from the distribution of segments.

4 The task and clustering

In this section, we describe our basic models and do some basic clustering with them.

We build an embedding model using the transcription from the `cmudict` dictionary (Weide, 1998). Specifically, we build a model where the vectors are sensitive to a window that includes 2 segments on each side. We set the length of each vector to 8 values. Given the relatively small amount of training data, model structure is sensitive to random initial parameters, so our comparisons will be over multiple runs, specifically 100 models.

[4] For comparison, a more specific featural decomposition is given in the appendix.

Recall that the question isn't whether Phone2Vec can learn all the features used in English, but whether it can learn the phonologically relevant ones.

Just to get a sense of what the model produces, we can look at individual vectors; here is the vector for [d] for a single model:

(10)

0.250	-0.468	-0.396	0.425
0.217	-0.270	-0.613	-0.420

Looking at individual vectors is not very informative of course.

As a first test, let's consider what happens when we simply cluster the vectors with k-means clustering. This is a statistical tool for finding some specific number of optimal clusters for vector representations.

The basic method is as follows (Lloyd, 1982). First we specify some number of clusters n. We then randomly choose n points in our vector space as the cluster centers or *centroids*. From those centroids, we determine which vectors belong to which clusters by proximity: a vector belongs to the cluster with the closest centroid. We then recalculate the position of the centroids based on the vectors in each cluster. We iterate this process a specific number of times or until the locations of centroids settle to some specific degree.

Note that k-means finds optimal clusters, but it does not find the optimal *number* of clusters. To do this we make use of the "elbow method" (Thorndike, 1953).[5] Here the basic idea is to apply k-means with increasing numbers of clusters and see at what point the relationship between the number of clusters and the tightness of clusters begins to loosen up. Specifically, we plot the within-cluster sum of squares (WCSS) for each number of clusters and look for the "elbow", the point at which WCSS starts to plateau.

Here, x refers to the vectors of each cluster and c refers to the cluster centroid.

(11) $$\text{WCSS} = \sum_{i=1}^{m}(x_i - c_i)^2$$

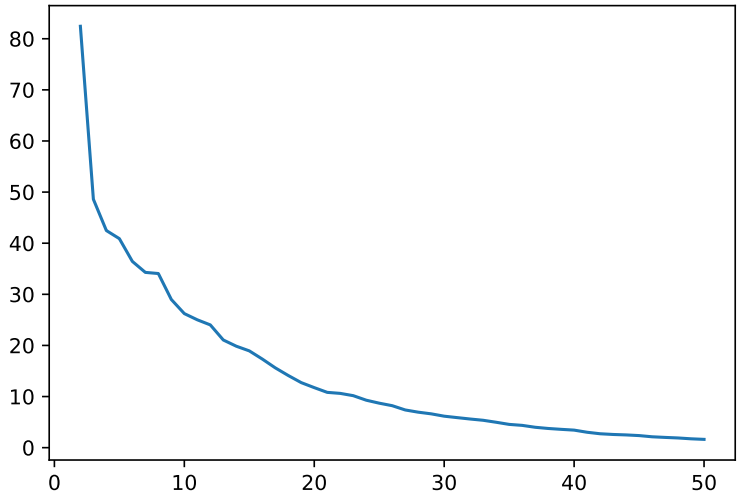
We plot these values for one of our models in Figure 1. Notice that the "elbow" occurs at around 10 clusters.

Let's now look at the clusters predicted by that model when there are 10 or fewer clusters. The fundamental observation is that, while some of the clusters align with familiar featural classes, others do not. Basically, large class distinctions are fairly clear, but smaller cross-classifying features are far less apparent.[6]

[5] There are a number of alternative techniques that can be used here, but this one is quick and intuitive.

[6] Note that individual vector models are subject to some variability. In addition, k-means is subject to variation as well.

Figure 1: Sum of squared distances to centroids



For two clusters, we get the following. Here, vowels without primary stress are distinguished from everything else.

1. ɛ̀, ə, ɔj, ɾ̩, æ, ɪ, ì, i, ò, ì, ɛ, ɔ̀, a, à, o, e, æ, àw, è, u, ʌ̀, ù, ɔ, àj, ʊ̀, aj, ɨ̀, ɔj, ʊ, aw
2. k, s, l, m, é, š, n, p, t, ó, z, w, d, ʌ́, b, ɛ́, v, í, á, r, áj, ǽ, ɔ́, ŋ, g, θ, f, ð, í, h, ʊ́, ɔ́j, č, ú, áw, ǰ, j, ž, ɾ́

For three clusters, things are a little murkier. Primary stressed vowels comprise one class. All vowels with secondary and some vowels with no stress fall into another, and all consonants and other vowels with no stress fall into a third.

Strikingly, the stressless vowels that fall in with the vowels with secondary stress are precisely those that classical phonological theory (Chomsky and Halle, 1968) says cannot occur without stress, i.e. [ɛ, a, e, æ, ɔ, aj, ʊ, aw]. In other words, these are vowels that it's reasonable to say are mistranscribed in `cmudict`.

1. k, s, l, ə, m, š, n, p, t, z, w, d, b, v, r, ɾ̩, ŋ, g, ɪ, θ, f, ð, h, i, č, o, u, ǰ, j, ž
2. é, ó, ʌ́, ɛ́, í, á, áj, ǽ, ɔ́, í, ʊ́, ɔ́j, ú, áw, ɾ́
3. ɛ̀, ɔj, æ, ì, ò, ì, ɛ, ɔ̀, a, à, e, æ, àw, è, ʌ̀, ù, ɔ, àj, ʊ̀, aj, ɨ̀, ɔj, ʊ, aw

For four clusters, consonants form one cluster, and then vowels are roughly divided by stress: primary, secondary, stressless.

1. ə, r, ɾ̩, ɪ, i, ɛ, a, o, e, u, ʊ
2. é, ó, ʌ́, ɛ́, í, á, áj, ǽ, ɔ́, í, ʊ́, ɔ́j, ú, áw, ɾ́
3. ɛ̀, ɔj, æ, ì, ò, ì, ɔ̀, à, æ, àw, è, ʌ̀, ù, ɔ, àj, ʊ̀, aj, ɨ̀, ɔj, aw

4. k, s, l, m, š, n, p, t, z, w, d, b, v, ŋ, g, θ, f, ð, h, č, ǰ, j, ž

With five clusters, we have consonants in one category, vowels divide across three of them—primary, secondary, stressless—and a fourth category which just comprises a subset of the vowels with primary stress.

1. ə, r, ɾ, ɪ, i, ɛ, a, o, e, u, ʊ
2. ʌ́, áj, ǽ, áw
3. ɛ̀, ɔj, æ̀, ì, ò, ì, ɔ̀, à, æ, àw, è, ʌ̀, ù, ɔ, àj, ʊ̀, aj, ɾ̀, ɔj, aw
4. k, s, l, m, š, n, p, t, z, w, d, b, v, ŋ, g, θ, f, ð, h, č, ǰ, j, ž
5. é, ó, ɛ́, í, á, ɔ́, í, ʊ́, ɔ́j, ú, ɾ́

With six clusters, we usually two primary stressed vowel classes, vowels with secondary stress, two stressless vowel classes, and consonants. On some runs, we get a fricative class.

1. i, ɛ, a, o, e, u, ɔ, ʊ
2. ʌ́, áj, ǽ, áw
3. ɛ̀, ɔj, æ̀, ì, ò, ì, ɔ̀, à, æ, àw, è, ʌ̀, ù, àj, ʊ̀, aj, ɾ̀, ɔj, aw
4. k, s, l, m, š, n, p, t, z, w, d, b, v, ŋ, g, θ, f, ð, h, č, ǰ, j, ž
5. é, ó, ɛ́, í, á, ɔ́, í, ʊ́, ɔ́j, ú, ɾ́
6. ə, r, ɾ, ɪ

With seven classes, consonants are now partitioned into two classes, with no obvious basis.

1. i, ɛ, a, o, e, u, ɔ, ʊ
2. ʌ́, áj, ǽ, áw
3. ɛ̀, ɔj, æ̀, ì, ò, ì, ɔ̀, à, æ, àw, è, ʌ̀, ù, àj, ʊ̀, aj, ɾ̀, ɔj, aw
4. š, z, w, d, b, f, ð, j, ž
5. é, ó, ɛ́, í, á, ɔ́, í, ʊ́, ɔ́j, ú, ɾ́
6. ə, r, ɾ, ɪ
7. k, s, l, m, n, p, t, v, ŋ, g, θ, h, č, ǰ

We give the remaining divisions below for eight, nine, and ten clusters. In these cases, we again see the role of stress, stop vs. fricative, obstruent vs. sonorant, and smaller divisions showing up. It would be fair to say that as the classes get smaller, the correspondence with feature theory diminishes.

Eight classes:

1. ə, r, ɾ, ɪ, i, o, u, j

2. ʌ́, áj, ǽ, áw
3. ɛ̀, ɔj, æ̀, ì, ò, ì, ɔ̀, à, æ, àw, è, ʌ̀, ù, àj, ʊ̀, aj, ɨ̀, ɔj, aw
4. ð
5. k, s, l, n, t, ŋ, g, θ, h, č, ǰ
6. é, ó, ɛ́, í, á, ɔ́, í, ʊ́, ɔj, ú, ɨ́
7. m, š, p, z, w, d, b, v, f, ž
8. ɛ, a, e, ɔ, ʊ

Nine classes:

1. ə, r, ɻ, ɪ, i, o, u
2. ʌ́, áj, ǽ, áw
3. ɛ̀, ɔj, æ̀, ì, ò, ì, ɔ̀, à, æ, àw, è, ʌ̀, ù, àj, ʊ̀, aj, ɨ̀, ɔj, aw
4. ð
5. k, s, l, n, t, ŋ, g, θ, h, č, ǰ
6. é, ó, ɛ́, í, á, ɔ́, í, ʊ́, ɔj, ú, ɨ́
7. m, š, p, z, w, d, b, v, f, ž
8. ɛ, a, e, ɔ, ʊ
9. j

Ten classes:

1. ə, r, ɻ, i, o, u
2. ʌ́, áj, ǽ, áw
3. ɛ̀, ɔj, æ̀, ì, ò, ì, ɔ̀, à, æ, àw, è, ʌ̀, ù, àj, ʊ̀, aj, ɨ̀, ɔj, aw
4. ð
5. k, s, l, n, t, ŋ, g, θ, h, č, ǰ
6. é, ó, ɛ́, í, á, ɔ́, í, ʊ́, ɔj, ú, ɨ́
7. m, š, p, z, w, d, b, v, f, ž
8. ɛ, a, e, ɔ, ʊ
9. j
10. ɪ

The general conclusion is that with just a few classes, there is a clear correspondence to the major category distinction between consonants and vowels. There is also a fairly clear correspondence to degrees of stress. As the number of categories increases though, featural correspondences are more difficult to discern.

5 Specific features

Let's now look more closely at the models in the other direction, with respect to specific featural relationships in English. Here we consider the various featural classes we identified in Section 3 above in light of what we know of the phonology of English.[7]

5.1 Vowels vs. consonants

The distinction between vowels and consonants should be available from phonotactics. We know, for example, that words can begin with only certain sequences of consonants. For example, a word can begin with the consonant sequence [br], as in *break* [brek] or *brew* [bru]. Any segment that follows [br] must therefore be a vowel. The same observation holds for any maximal word onset cluster, e.g. [str, kr, j, šr, č, ǰ, etc.].

Working from the other direction, we know that short vowels like [ɪ, ɛ, æ, ʊ], as in *bit* [bɪt], *bet* [bɛt], *bat* [bæt], and *could* [kʰʊd], must be followed by a consonant.

Let's test this in two ways. First, let's consider just similarity; we give the system a set of vowels (with varying stress) and ask what segments are most similar to them across. Here and following we do this across *all* 100 models. Here, we give the system [á, è, i, ó, ù] and ask for the five most similar segments and their average cosine distance from those segments.

(12)

ì	0.861
è	0.842
í	0.841
à	0.828
é	0.817

Indeed, the most similar segments are other vowels.

We can do this in the other direction as well. We give the system five consonants, [p, m, s, b, l], and ask for the five most similar other segments across all models:

(13)

d	0.963
t	0.962
n	0.958
z	0.940
f	0.929

Here, the most similar segments are other consonants.

[7] Our review here draws heavily on Hammond (1999b).

We can also test dissimilarity. Here, we give the system a set of segments that include all vowels and one consonant or all consonants and one vowel and ask which one is the odd man out. We do this for all the models and report the number of times each segment is chosen.[8] First, we do this for [p, m, s, a].

(14) a 100

Now we do the same thing in the other direction with [a, o, e, s].

(15) s 100

In both cases, the task successfully distinguishes vowels from consonants.

5.2 Stress

Stress should be learnable based on phonotactic restrictions, but there are some complications.

First, as we've already noted above, there are a fair number of mismarked items in this source.[9] These typically involve vowels that should or should not be marked for secondary stress. Here are a few examples:

(16)

spelling	cmudict/IPA	should be
abduct	AE0 B D AH1 K T	
	[æbdʌ̀kt]	[æ̀bdʌ́kt]
backdoor	B AE1 K D AO2 R	
	[bǽkdɔ̀r]	[bæ̀kdɔ́r]
veto	V IY1 T OW0	
	[víto]	[vítò]
abstruse	AH0 B S T R UW1 S	
	[æbstrús]	[æ̀bstrús]
klingon	K L IH1 NG D AH0 N	
	[klɪŋ̀dan]	[klíŋ̀ən]
rebound	R IY0 B AW1 N D	
	[ribáwnd]	[rìbáwnd]
defeatism	D AH0 F IY1 T IH2 Z AH0 M	
	[dəfítìzəm]	[dəfítɪzəm]
defects	D IY1 F EH0 K T S	
	[dífɛkts]	[dífɛ̀kts]

[8] Note that we do not make use of the analogy task discussed in the introduction. We do not use that task because it effectively tests two feature classes at the same time. Our use of similarity and dissimilarity tests only allows us to focus in on one feature class at a time.

[9] Some of these may reflect different dialects and/or analysis.

Another consideration is that generalizations about stress can involve dependencies that are a fair distance from the syllable or vowel in question. For example, the primary stress in an English noun must fall on one of the last three syllables of the word (or even further to the left with certain suffixes). This means that the possibility of a primary stressed vowel can depend on segments far to the right.

To keep things as simple as possible, we have limited our models to an extremely small window, 2 segments to each side. The downside of this move is that the system might have difficulty in learning stress distinctions.

With these caveats in mind, are differences in stress inferable from "local" phonotactics? Yes. First, all (content) words have a primary stress.

The distribution of stressless syllables is also restricted. For example, while a word can begin with a single stressless syllable, it may not begin with two. Thus, we can have words like *anaconda* [ænəkándə], but not [ənəkándə].

Secondary stresses are also restricted. Except with certain affixes, a word cannot end with two secondary stresses. Thus we can have *formaldehyde* [fòrmældəhàjd], but not [fórmældəhàjd].

Let's first consider similarity with primary stress. If we give the system the five primary stressed vowels [í, ǽ, ɔ́j, ʊ́, áj], we get back only vowels with primary stress:

(17)

ɛ́	0.997
ó	0.987
í	0.986
á	0.976
é	0.960

If we do the same for the secondary stressed vowels [ì, æ̀, ɔ̀j, ʊ̀, àj], the picture is murkier. We only get vowels, but the list includes [aw] which is marked as stressless. Interestingly, this is a vowel that classical phonological theory (Chomsky and Halle, 1968) says cannot be stressless.

(18)

aw	0.972
ɛ̀	0.968
ɔ̀	0.967
ò	0.959
à	0.951

Turning to stressless vowels, we now do the same thing with [i, ə, ʊ]. We specifically select vowels that should be unambiguously stressless. Again, however, the picture is murky and we see vowels with secondary stress in the list [ì], as well as vowels that should not be possible without stress: [ɛ, aj, æ].

(19)

ɛ	0.928
o	0.906
aj	0.870
ì	0.866
æ	0.859

Turning now to dissimilarity tests, we give the results for primary stress [í, æ̀, ɔ́j, ʊ́, áj] below:

(20)

æ̀	100

The correct vowel with secondary stress is selected above.

Here's the opposite test: [ì, ǽ, ɔ̀j, ʊ̀, àj] where we want to see if the vowel with secondary stress is picked out of a group of vowels with primary stress.

(21)

ǽ	100

In both cases, the correct outlier is identified.

Next we test for the difference between vowels with primary stress and stressless vowels with [í, æ, ɔ́j, ʊ́, áj] and [i, ǽ, ɔj, ʊ, aj] respectively.

(22)

æ	100

(23)

ǽ	100

Again, the correct outlier is picked out in both cases.

We now turn to the difference between vowels with secondary stress and stressless vowels using [ì, o, ɔ̀j, ʊ̀, àj] and [i, æ̀, o, ʊ, ə]. Here we carefully select stressless vowels that can truly be stressless.

(24)

o	100

(25)

æ̀	99
ə	1

Unlike with the similarity test, the correct outlier is consistently picked out in both cases.

Summarizing, the differences involving primary stress are robust, but tests involving stressless vowels are less clear. This may be due to analytic differences involving which vowels are stressed or simply mistranscriptions in the source.

5.3 Major class

From a traditional phonological perspective, major class is quite transparent based on phonotactic data. By major class, we mean the distinctions between obstruents [p, f, č, etc.], nasals [m, n, ŋ], liquids [l, r], and glides [w, y]. It's not clear how robust the relevant generalizations are statistically. The fact that a generalization is categorical does not entail that it is sufficiently in evidence for the vector model to be sensitive to it. There are some wrinkles to the patterns that may contribute as well.

The evidence here comes from syllable-based restrictions involving sonority. For example, obstruents can occur after nasals word-finally, but not vice versa, e.g. *dump* [dʌmp], *tent* [tʰɛnt], *stink* [stɪŋk], but not [dʌpm], [tʰɛtn], [stɪkŋ].

The distinction between nasals and liquids can also be seen in word-final position, e.g. *harm* [harm], *torn* [tʰorn], *helm* [hɛlm], but not *[hamr], *[tʰonr], or *[hɛml].

Notice that the pattern here is not fully symmetric; the velar nasal cannot occur after liquids. Hence we might expect the status of [ŋ] to be less clear. Similarly, [l] only occurs before [m] and neither of the other nasals, so its status may also be less clear.

Glides are also distinct from liquids, e.g. *pyre* [pʰajr], *file* [fajl], *hour* [awr], *owl* [awl], and not *[pʰarj], *[falj], *[arw], or *[alw].

The glide class also breaks down with [w] exhibiting a relatively free distribution after word-initial non-labial consonants. For [j], if there is a word-initial consonant, it cannot be coronal and the following vowel must be [u] (Davis and Hammond, 1995).

(26)

	w	j
p	—	puce [pʰjus]
t	tweed [tʰwid]	—
k	quick [kʰwɪk]	cube [kʰjub]

Let's now examine these classes with our similarity and dissimilarity tests. First, similarity for obstruants using [p, d, g, s, č] usually returns mostly obstruents.

(27)

t	0.980
m	0.947
k	0.945
ǰ	0.943
b	0.941

Testing for similarity with nasals is a bit trickier. The first problem is that there are only three nasals, so this only makes sense by selecting two of them. The second

issue is that the nasals are rather varied in their behavior, so the results from [m, n], [m, ŋ], and [n, ŋ] differ.

First [m, n]:

(28)

s	0.933
z	0.923
d	0.920
p	0.919
t	0.907

Now [m, ŋ]:

(29)

l	0.877
s	0.869
k	0.861
n	0.852
v	0.829

Now [n, ŋ]:

(30)

l	0.896
s	0.866
m	0.831
k	0.829
v	0.784

The results are mixed. With [m,n], we do not get [ŋ]. With [m, ŋ], we sometimes get [n], but not as the most similar. Similarly, with [n, ŋ], we get [m], but not as the first choice. The upshot is that the nasals do not fare well on the similarity test. It would probably be fair to say that [ŋ] is so different in its behavior that this disrupts the pattern.

Turning now to liquids, we are faced again with the problem that there are only two elements in the class. First, we start with [l]:

(31)

s	0.959
n	0.918
m	0.874
r	0.873
t	0.850

Now with [r]:

(32)

l	0.873
s	0.774
n	0.722
o	0.682
w	0.682

Starting with [l], we usually get [r], but often not as the first choice. When we start with [r], we get [l] and we sometimes also get [ɽ].[10]

Finally, let's consider similarity with the glides. Again, there are only two members in the class. First, starting with [j]:

(33)

š	0.700
ž	0.691
č	0.691
ǰ	0.624
d	0.558

Now starting with [w]:

(34)

f	0.744
ɽ	0.742
ə	0.741
z	0.717
d	0.689

In neither case is the other glide usually among the segments returned. We conclude, perhaps as expected from our discussion of the behavior of glides in English generally, that the glide class is not supported (by similarity test).

We now turn to the dissimilarity tests. First, we test if nasals can be distinguished from obstruents with [p, s, d, v, m] and this is not the case:

(35)

v	91
s	8
d	1

Similarly, liquids are not distinguished most of the time: [p, s, d, v, l].

(36)

v	90
l	10

[10] Recall that the transcription does not include [ɻ], so its absence is not meaningful here.

Glides, however, are distinguished: [p, s, d, v, j].

(37) j 100

Let's now turn to the nasals. Obstruents are not distinguished from nasals: [m, n, ŋ, v].

(38) ŋ 100

Liquids are also not distinguished from nasals: [m, n, ŋ, l].

(39) ŋ 100

Glides, however, are usually distinguished: [m, n, ŋ, j].

(40) j 94
ŋ 6

For the nasal cases, these results are surely a function of the very asymmetric behavior of [ŋ].

Turning to liquids, obstruents are usually distinguished: [l, r, z].

(41) z 83
r 17

Nasals are only occasionally distinguished: [l, r, n].

(42) r 80
n 20

Glides are consistently distinguished: [l, r, j].

(43) j 100

Turning to glides, obstruents are never distinguished: [w, j, z].

(44) j 99
w 1

Nasals are never distinguished: [w, j, n].

(45)

j	98
w	2

Liquids are nearly never distinguished: [w, j, l].

(46)

w	26
j	73
l	1

The dissimilarity results are best understood in a chart. Here we plot how many times the relevant distinction was made in each comparison. It's clear that glides are fairly consistently picked out, but the other categories are not.

(47)

	obstruents	nasals	liquids	glides
obstruents	NA	0	10	100
nasals	0	NA	0	94
liquids	83	20	NA	100
glides	0	0	1	NA

5.4 Manner

By manner, we intend the differences between stops [p, b, t, d, k, g], fricatives [f, v, s, z, š, ž, θ, ð], and affricates [č, ǰ]. These differences are *not* robustly supported by the phonotactics in English.

When we consider word-final clusters, these segments have essentially the same distribution. They can all occur after nasals, liquids, and glides.

(48)

	nasals	liquids	glides
stops	lint	wilt	out
	[lɪnt]	[wɪlt]	[awt]
fricatives	fence	pulse	mouse
	[fɛns]	[pʰʌls]	[maws]
affricates	pinch	arch	grouch
	[pʰɪnč]	[arč]	[grawč]

Word-initially, stops and voiceless fricatives pattern together in allowing a following liquid or [w], but voiced fricatives and affricates do not.

(49)

	alone	liquids	[w]
stops	tee	tree	tweed
	[tʰi]	[tʰri]	[tʰwid]
voiceless fricatives	thigh	thrive	thwack
	[θaj]	[θrajv]	[θwæk]
voiced fricatives	vie		
	[vaj]		
affricates	chin		
	[čɪn]		

The fricative class breaks down with [s] which has a much freer distribution than the other fricatives. For example, a word can begin with [s] followed by a voiceless stop or fricative, but other fricatives cannot do this.

(50)

	alone	stop	fricative
s	sear	steer	sphere
	[sir]	[stir]	[sfir]
z	xere		
	[zir]		
f	fear		
	[fir]		
v	veer		
	[vir]		
θ	thigh		
	[θaj]		
ð	thy		
	[ðaj]		
š	shy		
	[šaj]		
ž	genre		
	[žanr@]		

Let's now turn to the similarity tests. First, we see that the stops fail this; the set [p, t, b, d, g] should give us [k], but does not.

(51)

m	0.945
θ	0.941
z	0.933
f	0.931
h	0.926

The fricatives also fail; the set [f, v, s, θ, ð, š, ž] should give us [z], but does not.

(52)

d	0.966
ǰ	0.956
t	0.955
b	0.954
m	0.945

The affricates do succeed; [č] gives us [ǰ].

(53)

ǰ	0.954
g	0.904
k	0.878
t	0.876
ž	0.871

Turning to the dissimilarity tests, we see that fricatives are not picked out from stops; the set [p, t, d, g, f] should give us [f], but only does so about half the time.

(54)

g	48
f	51
d	1

On the other hand, affricates are picked out of the stop glass with [p, t, d, g, č] consistently returning [č].

(55)

č	100

Fricatives are not separated from stops; [f, v, s, ð, t] does not return [t].

(56)

ð	99
v	1

Affricates are only picked out from fricatives [f, v, s, θ, č] about 10% of the time.

(57)

č	10
v	79
f	9
s	2

Stops are separated from affricates; [k] is picked out of [č, ǰ, k] most of the time.

(58)

k	95
ǰ	1
č	4

Finally, fricatives are picked out of affricates with [č, ǰ, v] giving us [v].

(59)

v	95
č	5

To summarize, stops and fricatives failed the similarity test, but affricates succeeded. The dissimilarity tests were also clearer for affricates.

(60)

	stops	fricatives	affricates
stops	NA	51	100
fricatives	0	NA	10
affricates	95	95	NA

5.5 Voicing

Sonorant consonants and vowels are all voiced, but obstruents contrast in voicing in English. The contrast is robustly supported by past tense and plural/third singular allomorphy.

(61)

	Past		3sing/plural	
vowel	rowed	[rod]	rows	[roz]
voiced	bagged	[bægd]	bags	[bægz]
voiceless	looked	[lʊkt]	looks	[lʊks]
“identical”	sated	[setəd]	fusses	[fʌsəz]

The embedding model does not make use of alternations, so these patterns are not available. We expect the distinction to not be available to our vector models.

For similarity, we see that [p, t, s, θ, š] does not return [k].

(62)

d	0.984
b	0.948
m	0.943
z	0.938
g	0.933

On the other hand, [b, d, z, ð, ž] does return [g].

(63)

t	0.950
p	0.938
ǰ	0.927
g	0.917
m	0.916

Turning to dissimilarity, [g] is only picked out of [p, t, s, θ, g] once.

(64)

s	77
θ	15
p	7
g	1

The segment [k] is only picked out of [b, d, z, ð, k] twice.

(65)

ð	98
k	2

As expected then, our tests show that voicing is not learned by the models.

5.6 Place

Turning to place of articulation, we intend the difference between labial [p, b, m, f, v], coronal [t, d, n, s, z, θ, ð], and dorsal [k, g, ŋ] consonants. Other segments also have a place of articulation, but the set of contrasts is not as complete as for the obstruents and nasals. For example, there are only coronal liquids [l, r]. We therefore focus on obstruents and nasals.

Phonotactic evidence for these differences in English comes from place assimilation; nasal consonants generally agree in place with a following obstruent. The following chart shows the distribution word-finally in monomorphemes:

(66)

	labial	coronal	dorsal
m	hemp [hɛmp]	—	—
n	—	sand [sænd]	—
ŋ	—	—	sink [sɪŋk]

This pattern breaks down medially or across morpheme boundaries.

(67)

	labial	coronal	dorsal
m	camper	hamster	kumquat
	[kʰæmpɻ]	[hæmstɻ]	[kʰʌmkʰwat]
n	input	gander	vanguard
	[ɪnpʰʌt]	[gændɻ]	[vængard]
ŋ	gangplank	wrongdoer	trinket
	[gæŋpʰlæŋk]	[raŋduɻ]	[tʰrɪŋkət]

We therefore do not expect this distinction to be well supported.

In terms of similarity, the class [p, m, f, v] should return [b] and does.

(68)

b	0.968
d	0.933
n	0.932
t	0.927
k	0.916

In terms of similarity, the class [t, n, s, z, θ, ð] should return [d] and does.

(69)

d	0.982
p	0.947
b	0.947
m	0.937
g	0.932

Finally, the class [k, ŋ] should return [g], but does not. This is, of course, a smaller class and that may contribute. There is also, again, the exceptional behavior of [ŋ].

(70)

l	0.834
s	0.820
m	0.782
ǰ	0.774
n	0.772

Turning to dissimilarity, [p, m, f, v, t] should return [t] but does so only three times.

(71)

v	89
f	8
t	3

The class [p, m, f, v, k] should return [k], but does so only 20 times.

(72)

k	20
v	48
f	31
m	1

For coronals, the class [t, n, s, z, θ, ð, p] should return [p], but does not.

(73)

ð	100

For coronals, the class [t, n, s, z, θ, ð, k] should return [k], but also does not.

(74)

ð	100

Turning to dorsals, the class [k, g, ŋ, m] should return [m], but does not.

(75)

ŋ	100

The class [k, g, ŋ, n] should return [n], but does not.

(76)

ŋ	100

Summarizing, labials and coronals pass the similarity test, but dorsals do not. In terms of dissimilarity, none of the place specifications are robust.

(77)

	labial	coronal	dorsal
labial	NA	3	20
coronal	0	NA	0
dorsal	0	0	NA

5.7 Vowel length

By vowel length, we intend the contrast between tense or long vowels and diphthongs [i, e, u, o, a, aw, aj, ɔj, ju] and lax or short vowels [ɪ, ɛ, æ, ʊ, ʌ, ə].

This difference is well supported phonotactically in several ways. First, only the long vowels can occur stressed at the end of a word:

(78)

bee	[bí]	bay	[bé]
boo	[bú]	bo	[bó]
spa	[spá]	cow	[káw]
bye	[báj]	boy	[bɔ́j]

Second, the tense vowels are disallowed before certain consonant clusters (word-finally). For example:

(79)

	nasal + [pk]	liquid + [pkbg]
ɪ	limp [límp]	milk [mílk]
ɛ	hemp [hɛ́mp]	help [hɛ́lp]
æ	rank [rǽŋk]	talc [tʰǽlk]
ʌ	bunk [bʌ́ŋk]	gulp [gʌ́lp]

The vowels [ə, ʊ] are more restricted in their distribution. The vowel [ə] only occurs stressless, so cannot occur in lexical monosyllables as above. In addition, in function words it is the only lax vowel that can occur word-finally, e.g. in *the* [ðə]. The vowel [ʊ] is quite restricted in terms of following consonants; it's not clear that it can ever occur before a final cluster unless it is morphologically complex, e.g. *woods* [wʊdz], *wolves* [wʊlvz], *hooves* [hʊvz]. We leave these two vowels out of the tests below.

The length distinction crosscuts the stress distinction, so our comparisons below are limited to vowels with primary stress.

For the similarity tests, [í, é, ú, á] should give us [ó] and does.

(80)

ó	0.982
ɛ́	0.976
ɔ́j	0.962
í	0.948
ǽ	0.932

For lax vowels, the set [í, ɛ́, ʌ́] should give us [ǽ], and does.

(81)

ǽ	0.989
áw	0.974
ɔ́j	0.973
áj	0.970
ó	0.960

For the dissimilarity tests, [í, é, ú, ó, ǽ] should give us [ǽ], but does not.

(82)

ú	100

Finally, [í, ɛ́, ʌ́, ó] should give us [ó], but does not.

(83)

ʌ́	100

Summarizing, when we restrict ourselves to vowels with primary stress, both groups pass the similarity test, but neither class passes the dissimilarity test.

5.8 Vowel height

By vowel height, we mean the difference between high vowels [i, ɪ, u, ʊ], mid vowels [e, ɛ, o, ʌ, ə], and low vowels [æ, a, ɔ]. While this distinction is amply supported by the vowel shift (Chomsky and Halle, 1968), it is not supported by the phonotactics. Hence, we do not expect our vector models to support it.

Again, these classes intersect with stress distinctions, so we focus only on vowels with primary stress.

For similarity, the class [í, ɪ́, ʊ́] should give [ú], but does not.

(84)

ɛ́	0.988
á	0.981
ó	0.978
ɔ́j	0.963
ɔ́	0.957

The class [é, ɛ́, ʌ́] should give [ó], but does so only as the 5th choice.

(85)

ǽ	0.988
í	0.987
ɔ́j	0.976
áw	0.971
ó	0.969

Finally, the class [ǽ, ɔ́] should return [á], but does not.

(86)

í	0.991
ɛ́	0.987
áj	0.969
áw	0.967
ɔ́j	0.964

Turning to the dissimilarity tests, [í, ɪ́, ú, ʊ́, é] should return [é], but does so only three times.

(87)

ú	24
í	72
é	3
ʊ́	1

The set [í, ɪ́, ú, ʊ́, á] should return [á], but does not.

(88)

ú	93
ʊ́	5
í	2

The set [é, ɛ́, ó, ʌ́, í] should return [í] and only does so twice.

(89)

ʌ́	96
í	2
é	2

The set [é, ɛ́, ó, ʌ́, æ] should return [æ], but does not.

(90)

ʌ́	57
é	41
ó	2

Considering now low vowels, the set [ǽ, á, ɔ́, í] should return [í], but does so only twice.

(91)

ǽ	70
ɔ́	28
í	2

Finally, the set [ǽ, á, ɔ́, é] should return [é] and suprisingly does so 39 times.

(92)

ɔ́	54
é	39
ǽ	5
á	2

Summarizing, even when we restrict ourselves to vowels with primary stress, high and low vowels fail the similarity test, and mid vowels give marginal support, except for the low–mid dissimilarity test just above.

(93)

	high	mid	low
high	NA	3	0
mid	2	NA	0
low	2	39	NA

5.9 Vowel backness

By vowel backness, we intend the difference between the front vowels [i, ɪ, e, ɛ, æ] and the back vowels [u, ʊ, o, ʌ, ə, a, ɔ]. This difference is not supported phonotactically in English, so we do not expect to see our vector models substantiate it. This distinction crosscuts stress differences, so as we did in the previous section, we confine our tests to vowels with primary stress.

We first consider similarity with respect to the front vowels. The set [í, ɪ́, é, ǽ] should return [ɛ́] and does so.

(94)

ɛ́	0.992
ó	0.988
ɔ́j	0.984
á	0.969
ɹ́	0.959

For similarity with back vowels, the set [ú, ʊ́, ó, ʌ́, á] should return [ɔ́], but does not.

(95)

ɛ́	0.985
ɪ́	0.972
ɔ́j	0.962
é	0.959
ǽ	0.954

In terms of dissimilarity, [ó] should be picked out of the otherwise front vowel set [ɪ́, ɪ́, é, ɛ́, ǽ, ó], but is not.

(96)

ɪ́	69
é	31

The vowel [ɛ́] should be picked out of the otherwise back vowel set [ú, ʊ́, ó, ʌ́, á, ɔ́, ɛ́], but is not.

(97)

ú	69
ʌ́	31

Summarizing, the front vowels meet the similarity test, but the back vowels do not. In terms of dissimilarity, both sets fail.

5.10 Feature summary

Summarizing the patterns, we find the following:

Vowels vs. consonants We expected this to be supported and it is.

Stress We expected primary stress to be supported and it is. The difference between secondary and stressless vowels is not unambiguously supported, but this is perhaps not surprising as this distinction is controversial for some vowels in the data source.

Major class (obstruents, nasals, liquids, glides) We expected some of this to be supported and this was the case. Obstruents and glides are sporadically distinguished from the other classes.

Manner (stops, fricatives, affricates) We expected this to be supported, but it was not. Affricates are the only clear class here.

Voicing We did not expect this distinction to be supported and it was not.

Place (labial, coronal, dorsal) We did not expect this distinction to be supported and this is indeed the case with the dissimilarity tests. On the other hand, there was evidence for labial and coronal with the similarity tests.

Vowel length (tense/long, lax/short) We expected this distinction to be supported, but it was only supported in the similarity tests, not the dissimilarity tests.

Vowel height (high, mid, low) We did not expect this distinction to be supported and it was not.

Vowel backness (front, back) We did not expect this distinction to be supported and it was not.

In general, our predictions with respect to larger distinctions were borne out, but some of our predictions with respect to smaller distinctions were not. We suspect there are several reasons for this.

First, as already noted, embedding models are sensitive to phonotactics only and featural distinctions that depend on alternations cannot be made.

Second, as reviewed above, some distinctions are not clear even in the linguistic literature. The features do not work cleanly, even with classical linguistic data.

Third, linguistic generalizations like feature classes are made on the basis of *available* data, not *frequent* data. The embedding model, on the other hand, is sensitive only to patterns that are very frequent.[11]

Finally, the embedding model does not do very well with cross-classification. That is, as we saw in the k-means discussion, classes that are nested seem to be discoverable. Cross-classifying classes seem to be less robust.

[11] See Hammond (1999a) for an ancient discussion of this issue. See Finley et al. (2017) for discussion in the context of word embeddings.

6 Previous work

The general idea of learning feature classes from statistical distributions has come up before. In fact, the specific issue of whether a word embedding model can be profitably applied to segmental classes has also been addressed. In this section, we discuss the most relevant previous work.

Goldsmith and Xanthos (2009) use a number of techniques to investigate several questions: "i) Given a sample of data ..., can we infer which segments are vowels and which are consonants? ii) can we infer on the basis of such data whether the language in question possesses a system of vowel harmony, and if so, what the patterns of vowel harmony are in the language? iii) can we draw inferences about the organization of segments into syllabic structure?" (p.4).

The work is more extensive than the current project in its focus on issues beyond feature classes and in the language data considered: English, French, and Finnish. On the other hand, the only feature distinction treated is the vowel vs. consonant distinction and feature distinctions are not treated in light of the phonological properties of the languages in question. Goldsmith and Xanthos show that the vowel vs. consonant distinction is learnable—to varying degrees—with several techniques.

Silfverberg et al. (2018) investigate directly the question of whether vector models can learn feature classes. Their work differs from the current project in several ways.

First, the data they use are incomplete morphological paradigms from the reinflection shared task (Cotterell et al., 2017). The results of the shared task were extremely interesting and bear directly on questions of whether morphology and phonological alternations can be learned computationally, but these paradigms are not clearly representative of the sort of data a child might be exposed to for learning phonological feature classes.

Second, they investigate vector models with several different languages, working from orthographic representations in languages with fairly transparent orthographies, i.e. Finnish, Turkish, and Spanish. They do not treat English.

Third, their focus is on alternative architectures for building embedding models. They compare the Word2Vec model we use with an RNN encoder-decoder model and with embeddings constructed directly from truncated Singular Value Decomposition (SVD) on a matrix of positive point-wise mutual information (PPMI) values (Bullinaria and Levy, 2007).

Finally, they do not offer a detailed discussion of what feature classes are discovered and how that might relate to the phonological facts of the language, our focus here.

Mayer (2018) investigates vector models in several languages. He does not use the Word2Vec system specifically, but employs similar statistical tools. Specifically, he starts with context vectors, normalizes, reduces the vectors with principal components analysis, then does clustering with k-means. He does this with four real languages, Samoan, English, French, and Finnish, and with a constructed language.

The constructed language is potentially quite interesting as it allows for the possibility of manipulating the phonology and distributional regularities of the source language to see how the computational model fares. Mayer does this, but only with respect to how much noise the data presents.

Mayer offers very interesting discussion of how the classes determined by his model correlate or not with the phonologies of the languages he considers, but not at the level of detail offered here. His discussion is also from the direction of the classes his model discovers, rather than from the classes the phonology might predict (as we do here).

Kolachina and Magyar (2019) use a series of artificial languages to explore how the CBOW version of Word2Vec fares in finding feature classes. The artificial languages vary in the complexity of the phonologies they exhibit. Kolachina and Magyar argue that their vector models perform better on vowel patterns, rather than consonant patterns. Kolachina and Magyar also investigate the effect on vector models of varying the number of sounds in a language.

To summarize, the work here differs from previous work in a number of ways, including the type of models used, the language focus, and the theoretical goals.

7 Conclusion

In conclusion, we have shown that some featural classes are learnable with phone embedding models using the generic Word2Vec skipgram architecture, specifically, consonants vs. vowels and primary stress. All other distinctions are either partially supported or not supported at all.

We provided a detailed examination of each contrast in the context of the phonology of English and saw that we do not expect all featural distinctions to be supported because the phonotactics of English do not support all featural distinctions.

There were some distinctions that were not made by the models that we might have expected the models to make, e.g. major class, manner, vowel length. We suspect there are three reasons for this.

First, as already noted, some of these distinctions are complicated that exceptional behavior that presumably makes it difficult for the models to capture the class behavior. For example, [ŋ] and [ʊ] exhibit quite exceptional distributions that complicate the major class and vowel length classes.

A second factor is that, as linguists, we focus on how robust a generalization is in terms of whether it allows exceptions. The embedding models, on the other hand, are quite sensitive to how often a pattern shows up, regardless of whether there are counterexamples.

Finally, it's not clear how well embedding models handle cross-classifying class behavior. As seen in Section 4, the classes we observe with k-means are subset classes, rather than cross-classifying classes.

There are three obvious next steps for this research.

First, are the classes that the models discover, e.g. in Section 4, real or artifactual? One way to investigate this might be to feed these classes into a downstream model like Maxent (Hayes and Wilson, 2008) and see how they perform compared to traditional phonological classes.

Second, the other side of that coin would be to investigate whether the classes that the models did *not* discover are real or artifactual. In other words, maybe the model didn't discover some classes because they really aren't classes. In this light, one would want to pursue the line of research in Mielke (2004).

Finally, it may be that computational models with a different structure would fare better with these other classes. As discussed in Section 6, there are several papers that address this already, e.g. Silfverberg et al. (2018) and Mayer (2018).

8 Acknowledgments

Thanks to.... All errors are my own.

References

Bengio, Y., Ducharme, R., and Vincent, P. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word cooccurrence statistics: A computational study. *Behavior research methods*, 39:510–526.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row, New York.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., andJason Eisner, D. Y., and Hulden, M. (2017). CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In Hulden, M., editor, *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30. Association for Computational Linguistics, Vancouver.

Davis, S. and Hammond, M. (1995). Onglides in American English. *Phonology*, 12:159–182.

Finley, G., Farmer, S., and Pakhomov, S. (2017). What analogies reveal about word vectors and their compositionality. In *Proceedings of the 6th joint conference on lexical and computational semantics*, pages 1–11.

Firth, J. R. (1957). *A synopsis of linguistic theory 1930-1955*. Oxford University Press, Oxford. Special Volume of the Philological Society.

Goldsmith, J. and Xanthos, A. (2009). Learning phonological categories. *Language*, 85:4–38.

Hammond, M. (1999a). Is phonology irrelevant? *Literary and Linguistic Computing*, 13:165–175.

Hammond, M. (1999b). *The Phonology of English*. Oxford University Press, Oxford.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440.

Kolachina, S. and Magyar, L. (2019). What do phone embeddings learn about phonology? In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 160–169.

Lloyd, S. P. (1982). Least square quantization in PCM. *IEEE Trans. Inform. Theor.*, 18. Bell Telephone Laboratories Paper, 1957, published in journal much later.

Mayer, C. (2018). An algorithm for learning phonological classes from distributional similarity. UCLA MA thesis.

Mielke, J. (2004). *The Emergence of Distinctive Features*. PhD thesis, Ohio State U.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. pages 3111–3119.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Pater, J. (2019). Generative linguistics and neural networks at 60: foundation, friction, and fusion. *Language*, 95:41–74.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Silfverberg, M. P., Mao, L., and Hulden, M. (2018). Sound analogies with phoneme embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Thorndike, R. L. (1953). Who belongs in the family. *Psychometrika*, 18:267–276.

Weide, R. L. (1998). The CMU pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

Appendix

CMU Transcription

Vowels:

IY	i			UW	u
IH	ɪ			UH	ʊ
EY	e			OW	o
EH	ɛ	AH	ʌ/ə	AO	ɔ
AE	æ			AA	a

Diphthongs:

ER	ɻ	AW	aw
AY	aj	OY	ɔj

Notice that syllabic consonants are not marked consistently. Compare:

ER	ɻ	butter
AH L	l̩	battle
AH M	m̩	bottom
AH N	n̩	button

All vowels and diphthongs are also marked for stress, either as stressless, primary stress, or secondary stress, with a following integer: 0, 1, or 2 respectively. For example: IY0, IY1, IY2.

Consonants:

P	p			T	t			K	k
B	b			D	d			G	g
F	f	TH	θ	S	s	SH	š	HH	h
V	v	DH	ð	Z	z	ZH	ž		
						CH	č		
						JH	ǰ		
M	m			N	n			NG	ŋ
				L	l				
				R	r				
W	w					Y	j		

Examples

Vowels:

IY	i	these	DH IY1 Z	ðíz
UW	u	view	V Y UW1	vjú
IH	ɪ	tic	T IH1 K	tɪk
UH	ʊ	goods	G UH1 D Z	gʊ́dz
EY	e	save	S EY1 V	sév
OW	o	alone	AH0 L OW1 N	əlón
EH	ɛ	bellow	B EH1 L OW0	bɛ́lo
AH	ʌ/ə	buzzes	B AH1 Z AH0 Z	bʌ́zəz
AO	ɔ	all	AO1 L	ɔ́l
AE	æ	abby	AE1 B IY0	ǽbi
AA	a	bomber	B AA1 M ER0	bámɻ

Diphthongs:

ER	ɻ	burp	B ER1 P	bɻ́p
AW	aw	county	K AW1 N T IY0	káwnti
AY	aj	dial	D AY1 AH0 L	dájəl
OY	ɔj	coins	K OY1 N Z	kɔ́jnz

Consonants

P	p	spiffy	S P IH1 F IY0	spífi
T	t	taiphoon	T AY2 F UW1 N	tàjfún
K	k	weaken	W IY1 K AH0 N	wíkən
B	b	buzzard	B AH1 Z ER0 D	bʌ́zɻd
D	d	driver	D R AY1 V ER0	drájvɻ
G	g	grasp	G R AE1 S P	grǽsp
F	f	fable	F EY1 B AH0 L	fébəl
TH	θ	thank	TH AE1 NG K	θǽŋk
S	s	salary	S AE1 L ER0 IY0	sǽlɻi
SH	š	shaggy	SH AE1 G IY0	šǽgi
HH	h	Ahab	EY1 HH AE2 B	éhæb
V	v	avail	AH0 V EY1 L	əvél
DH	ð	another	AH0 N AH1 DH ER0	ənʌ́ðɻ
Z	z	fuzzy	F AH1 Z IY0	fʌ́zi
ZH	ž	fusion	F Y UW1 ZH AH0 N	fjúžən
CH	č	champ	CH AE1 M P	čǽmp
JH	ǰ	agent	EY1 JH AH0 N T	éǰənt
M	m	ample	AE1 M P AH0 L	ǽmpəl
N	n	nail	N EY1 L	nél
NG	ŋ	bang	B AE1 NG	bǽŋ
L	l	lazy	L EY1 Z IY0	lézi
R	r	rabid	R AE1 B AH0 D	rǽbəd
W	w	cobweb	K AA1 B W EH2 B	kábwɛ̀b
Y	j	yankee	Y AE1 NG K IY0	jǽŋki

Features

Vowel features:

	IY	UW	IH	UH	EY	OW	EH	AH	AO	AE	AA
	i	u	ɪ	ʊ	e	o	ɛ	ʌ/ə	ɔ	æ	a
hi	+	+	+	+	-	-	-	-	-	-	-
lo	-	-	-	-	-	-	-	-	+	+	+
bk	-	+	-	+	-	+	-	+	+	-	+
rnd	-	+	-	+	-	+	-	-	-	-	-
tns	+	+	-	-	+	+	-	-	+	-	+

Consonant place specifications:

LAB		COR		LAB/DOR		COR/DOR		DOR		∅	
P	p	T	t	W	w	SH	š	K	k	HH	h
B	b	D	d			ZH	ž	G	g		
F	f	TH	θ			CH	č	NG	ŋ		
V	v	S	s			JH	ǰ				
M	m	Z	z			Y	j				
		DH	ð								
		N	n								
		L	l								
		R	r								

Other consonant features:

		vcd	cont	nasal	cons	son
P	p	-	-	-	+	-
T	t	-	-	-	+	-
K	k	-	-	-	+	-
B	b	+	-	-	+	-
D	d	+	-	-	+	-
G	g	+	-	-	+	-
M	m	+	-	+	+	+
N	n	+	-	+	+	+
NG	ŋ	+	-	+	+	+
F	f	-	+	-	+	-
TH	θ	-	+	-	+	-
S	s	-	+	-	+	-
SH	š	-	+	-	+	-
V	v	+	+	-	+	-
DH	ð	+	+	-	+	-
Z	z	+	+	-	+	-
ZH	ž	+	+	-	+	-
CH	č	-	-/+	-	+	-
JH	ǰ	+	-/+	-	+	-
W	w	+	+	-	-	+
Y	j	+	+	-	-	+
L	l	+	+	-	+	+
R	r	+	+	-	+	+
HH	h	-	+	-	-	+