Phone2Vec

Michael Hammond

U. of Arizona

Abstract

In this paper, we investigate whether a generic word embedding model can be applied to phonetic segments to discover phonological classes. 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? Our general prediction is that classes we know are supported by the phonotactics should emerge, but classes that are only evidenced by alternations will not.

1 Overview

In this paper we investigate whether natural classes of segments can be deduced solely from the contexts those segments 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. We find that a vector space model learns major phonetic dimensions, but exhibits marginal performance with other dimensions.

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]

Are 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, 1999) 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 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*

[1] These are reviewed in Section 7.

Proceedings of *AMP 2023/2024*

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 outline our methods including the neural architecture, training procedure, and testing techniques. We then review English phonotactic structure at a high level to make general predictions about the outcomes of our tests. We then present the results of those tests. Last, we discuss these results and contextualize our work with respect to previous efforts.

2 Methods

We build vector models from the CMU pronouncing dictionary (Weide, 1998) and then test them with *vector similarity* and *vector dissimilarity* tasks.

We use the *skipgram* architecture of Mikolov et al. (2013a), Mikolov et al. (2013b), and Mikolov et al. (2013c). This is an older, efficient, and simple system that is implemented in the python `gensim` module (Řehůřek & Sojka, 2010). The skipgram logic is that we train a neural net to predict adjacent sounds in a word from each sound. This procedure creates an internal numerical representation of each sound that reflects the contexts it occurs in. In our experiments, we set the number of adjacent sounds to two on each side and pad each word with two dummy symbols on each side. Thus for a word like *hats* [hæts], we pad as ##hæts##. This results in the following input-output pairs for training:

(2)

input	output
h	##-æt
æ	#h-ts
t	hæ-s#
s	æt-##

Let's understand this with a simplified schematic system. Imagine we have a language with only three sounds: $\{s_1, s_2, s_3\}$. We first represent each sound using "1-hot" encoding. Each sound is encoded as a sequence of zeros and ones where the length of that sequence is equal to the total number of distinct sounds, in this case 3. The identity of a sound can be determined by seeing which column contains a 1; all other columns will contain 0.

(3)

Sound	Encoded		
s_1	1	0	0
s_2	0	1	0
s_3	0	0	1

We give the schematic neural architecture in Figure 1. From each sound we attempt to predict the two sounds on each side of it. In terms of our hypothetical example, we would predict an output of twelve numbers from an input of three numbers.

The neural architecture has at least one hidden layer between the input and the output. Embeddings are created at that bottleneck hidden layer. We decide how big those embeddings are. For example, in our hypothetical example, we might decide to have only two numbers at the bottleneck layer. Thus, we'd map from 3 to 2 to 12 numbers.

[2] Note that our investigation proceeds from phonotactic restrictions and does not consider acoustic or articulatory similarity.

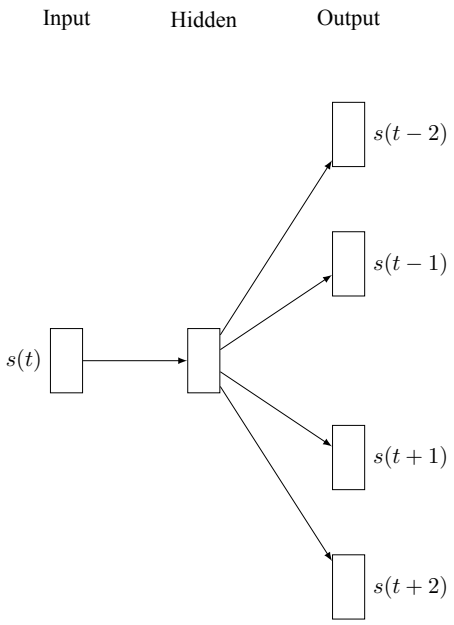

Figure 1: Network architecture

In our actual experiments, the numbers are of course different. There are 69 different sounds in the `cmudict` dictionary (plus our padding symbol), and our context is two sounds on each side. We set the size of our embeddings to 8, so we map from 70 numbers to 8 numbers to $4 \times 70 = 280$ numbers.

Finally, neural nets start with random initial values and results vary depending on these. We therefore run our model 100 times to control for this.

3 Tests

Once the network has been trained, we can use it to map from input sounds to the hidden layer and collect the embedding representation for each sound there. Thus, in our case, we have a vector of 8 numbers for each of the 69 sounds in `cmudict`. Because of the way the network is trained, these embeddings encode the contexts each sound occurs in.

We can visualize the vectors in two dimensions using the t-SNE dimension reduction algorithm implemented in `scikit-learn`. This reduces the eight-dimensional space to two dimensions. This is an approximation and does *not* capture the full model. We do this for our first model in Figure 2. Notice that there are clear apparent clusters in the plot.

With the 100 models we create, we test *vector similarity* and *vector dissimilarity* for contrastive phonetic dimensions of English.

For vector similarity, we take a set of sounds and ask what other sound is closest to them in vector space. We do this by averaging distance across all 100 models. Distance is measured in terms of cosine similarity, which basically measures the angular difference for any pair of points calculated from the origin.

$$\frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4}$$

For example, with the set of vowels [á, è, i, ó, ù], across all models, the segment most similar is [ì] (highest

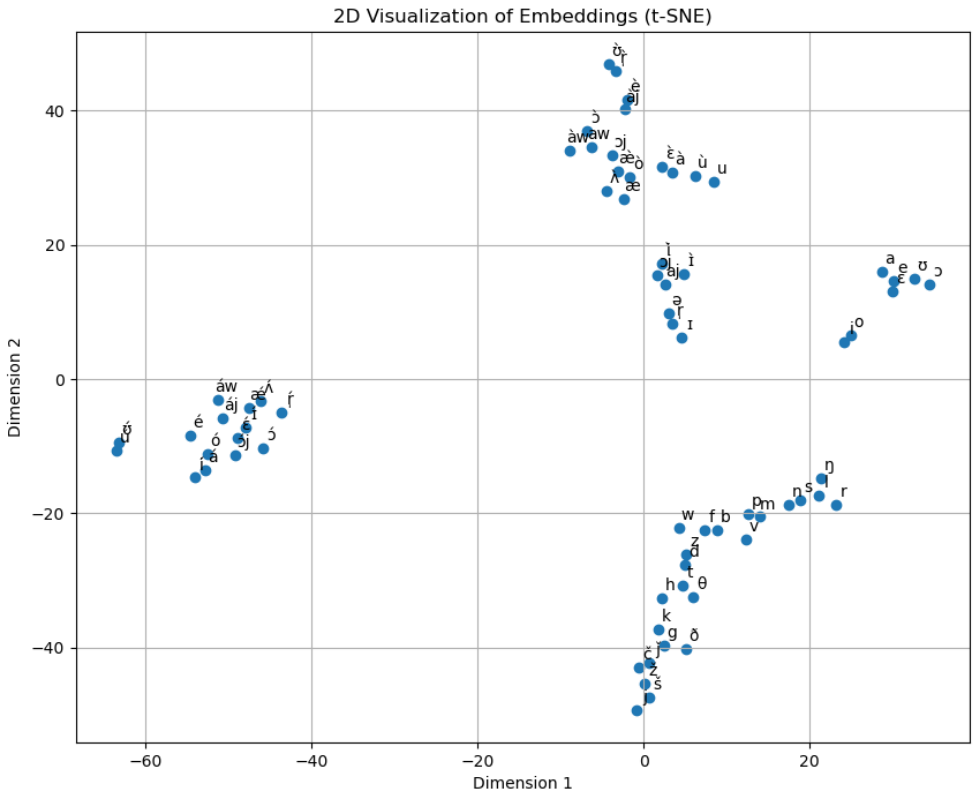

Figure 2: t-SNE visualization of vectors from initial model

average cosine similarity across all models). This would be taken as a correct prediction, since the segment selected is also a vowel.

We also test for *dissimilarity*. The basic idea is that we take a set of sounds with an "odd man out", a single sound that is dissimilar from the rest. We plot the set in vector space, calculate the center of the cluster and ask which sound is furthest from that center. What we're looking for is whether the odd man out is that furthest sound.

For example, if we give the set [p, m, s, a] with three consonants and one vowel, the least similar is [a] in all 100 models. This also would be taken as a correct prediction.

4 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.

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 $127,069$ words. The transcription 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.

Given this resource, we only consider phonetic dimensions 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

What we expect is that the first seven dimensions should emerge since they are all supported by the phonotactics of English (Hammond, 1999). The last two dimenions are supported by alternations, but not phonotactics, so we do not expect them to emerge.

5 Results

Let's first consider dissimilarity test results. Table 1 gives dissimilarity results for major class, manner, and voice. Recall that this is an odd-man-out test. The first column gives the general class followed by the category of the odd man out. The second column gives the actual set of sounds used. The third column gives the sound chosen most often. The fourth column indicates how many models this happened with.

If the test produced an unexpected result, it is marked with a question mark. These are cases where the relevant distinction is apparently not learned by the model. One row is marked with an exclamation mark; here the correct outlier is chosen, but in only 51 of the models.

The general pattern here is that stress distinctions are learned and some major categories are learned, but not all. That the latter are not learned is not predicted.

In Table 2 we give the dissimilarity results for place of articulation and for vowel height and backness. The structure of the table is the same as Table 1. Place of articulation is supported by the phonotactics and should

Category/category	Class	Best	Models	
consonants/vowel	[p, m, s, a]	a	100	
vowels/consonant	[a, o, e, s]	s	100	
primaries/secondary	[í, æ̀, ɔ́j, ʊ́, áj]	æ̀	100	
secondaries/primary	[ì, ǽ, ɔ̀j, ʊ̀, àj]	ǽ	100	
secondaries/stressless	[ì, o, ɔ̀j, ʊ̀, àj]	o	100	
stressless/secondary	[i, æ̀, o, ʊ, ə]	æ̀	99	
obstruants/nasal	[p, s, d, v, m]	v	91	?
obstruants/liquid	[p, s, d, v, l]	v	90	?
obstruants/glide	[p, s, d, v, j]	j	100	
nasals/obstruant	[m, n, ŋ, v]	ŋ	100	?
nasals/liquid	[m, n, ŋ, l]	ŋ	100	?
nasals/glide	[m, n, ŋ, j]	j	94	
liquids/obstruant	[l, r, z]	z	83	
liquids/nasal	[l, r, n]	r	80	?
liquids/glide	[l, r, j]	j	100	
glides/obstruant	[w, j, z]	j	99	?
glides/nasal	[w, j, n]	j	98	?
glides/liquid	[w, j, l]	j	73	?
stops/fricative	[p, t, d, g, f]	f	51	!
stops/affricate	[p, t, d, g, č]	č	100	
fricatives/stop	[f, v, s, ð, t]	ð	99	?
fricatives/affricate	[f, v, s, θ, č]	v	79	?
affricates/stop	[č, ǰ, k]	k	95	
affricates/fricative	[č, ǰ, v]	v	95	
voiceless/voiced	[p, t, s, θ, g]	s	77	?
voiced/voiceless	[b, d, z, ð, k]	ð	98	?

Table 1: Dissimilarity results for major class, manner, and voice

Category/category	Class	Best	Models	
labial/velar	[p, m, f, v, k]	v	48	?
coronal/labial	[t, n, s, z, θ, ð, p]	ð	100	?
coronal/velar	[t, n, s, z, θ, ð, k]	ð	100	?
dorsal/labial	[k, g, ŋ, m]	ŋ	100	?
dorsal/coronal	[k, g, ŋ, n]	ŋ	100	?
tense/lax	[í, é, ú, ó, ǽ]	ú	100	?
lax/tense	[í, ɛ́, ʌ́, ó]	ʌ́	100	?
high/mid	[í, í, ú,ʊ́, é]	ú	24	?
high/low	[í, í, ú, ʊ́, á]	ú	93	?
mid/high	[é, ɛ́, ó, ʌ́, í]	ʌ́	96	?
mid/low	[é, ɛ́, ó, ʌ́, ǽ]	ʌ́	57	?
low/high	[ǽ, á, ɔ́, í]	ǽ	70	?
low/mid	[ǽ, á, ɔ́, é]	ɔ́	54	?
front/back	[í, í, é, ɛ́, ǽ, ó]	í	69	?
back/front	[ú, ʊ́, ó, ʌ́, á, ɔ́, ɛ́]	ú	69	?

Table 2: Dissimilarity results for place and vowel height and backness

be learned. None of the vowel distinctions are learned, as predicted.

We now turn to the similarity results. Recall that we start with a set of sounds and ask what *other* sound is most similar to this set. In Table 3, we give the similarity results for major class, manner and voicing. The structure of the table is a bit different than for the dissimilarity results. Here the first column is the name of the class. The second column gives the set of sounds used. The third column gives the best additional sound across all 100 models. If a wrong item is selected, it is marked with a question mark.

Two items are marked with exclamation points, in these cases, the results reflect a transcription anomoly in `cmudict`. The transcription system there includes stressless [aw] and [ɛ], neither of which are possible in most treatments of English stress (Chomsky & Halle, 1968; Liberman & Prince, 1977; Hayes, 1981; Hammond, 1999). We set these two cases aside then.

The general pattern here is as with our dissimilarity results: stress and major categories are learned, but a number of the other dimenions are not. The latter are not predicted.

In Table 4, we give similarity results for place of articulation and for vowel height and backness. We read this table just like Table 3. Most dimensions are not learned, but a couple of them are. The general prediction is that place and tense/lax should be learned, but none of the others.

In sum, stress and major class distinctions emerge, but other distinctions emerge only sporadically. What is surprising is that a number of dimensions that we know are supported phonotactically do not emerge, e.g. nasals, liquids.

6 Discussion

We've seen that stress and consonant/vowel distinctions emerge, but other distinctions are less clear. The fact that vowel distinctions other than tense/lax do not emerge is no surprise, but what of other distinctions?

Consider, for example, the distinction between stops and nasals illustrated at the beginning of this paper. The distinction there is robust in that there are no counterexamples, but relevant examples are actually relatively infrequent. For example, of the 127,069 items in `cmudict`, there are only 2,888 cases of a final nasal-stop sequence. Perhaps this is simply not enough.

Consider also the failed similarity tests for nasals and liquids. In the nasal similarity case, rather than the missing nasal being returned, we got [s] and [l]. Interestingly, these are segments that can occur in similar

Category	Class	Best	
vowels	[á, è, i, ó, ù]	ì	
consonants	[p, m, s, b, l]	d	
primary	[í, ǽ, ɔ́j, ʊ́, áj]	é	
secondary	[ì, æ̀, ɔ̀j, ʊ̀, àj]	aw	!
stressless	[i, ə, ʊ]	ɛ	!
obstruants	[p, d, g, s, č]	t	
nasals	[m, n]	s	?
	[m, ŋ]	l	?
	[n, ŋ]	l	?
liquids	[l]	s	?
	[r]	l	
glides	[j]	š	?
	[w]	f	?
stops	[p, t, b, d, g]	m	?
fricatives	[f, v, s, θ, ð, š, ž]	d	?
affricates	[č]	ǰ	
voiceless	[p, t, s, θ, š]	d	?
voiced	[b, d, z, ð, ž]	g	

Table 3: Similarity results for major class, manner, and voice

Category	Class	Best	
labial	[p, m, f, v]	b	
coronal	[t, n, s, z, θ, ð]	d	
velar	[k, ŋ]	l	?
tense/long	[í, é, ú, á]	ó	
lax/short	[í, ɛ́, ʌ́]	ǽ	
high	[í, í, ʊ́]	é	?
mid	[é, ɛ́, ʌ́]	ǽ	?
low	[ǽ, ɔ́]	í	?
front	[í, í, é, ǽ]	é	
back	[ú, ʊ́, ó, ʌ́, á]	é	?

Table 4: Similarity results for place and vowel height and backness

positions in final clusters, e.g. *mint* [mínt], *mist* [míst], and *tilt* [tílt]. In other words, these test failures may actually reflect the system learning a more general category.

Another factor that might play a role in our results is the size of the context window: two segments on each side. A context like that is necessary for stress distinctions, but might be a distraction for segmental phonotactics where a smaller context window should suffice. If we had set the context to be a single segment on each side, we would surely have had trouble getting stress distinctions, but maybe this would do better for phonological classes based on more local phonotactic generalizations.

There are other neural embedding systems we might also try. We did also try GloVe embeddings (Pennington et al., 2014), but these did not perform as well, so we set them aside. One might consider BERT-style embeddings (Devlin et al., 2018), but these are context-dependent and thus not suitable for our specific tests.

7 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 & 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 & 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 & 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 & 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 & Magyar argue that their vector models perform better on vowel patterns, rather than consonant patterns. Kolachina & 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.

8 Conclusion

In summary, we have investigated whether a simple and familiar embedding model learns the phonetic dimensions of English from phonotactic distributions. Our hypothesis was that distinctions supported by the phonotactics would be learned, but distinctions only evidenced by alternations would not be.

We saw that major phonetic dimensions are learned, but smaller phonotactic distinctions are not learned.

Our hypothesis is that these smaller distinctions were not learned either because, whether they are exceptionless or not, they are relatively infrequent or because one must set a context window that is appropriate for the distinction at issue.[3]

[3] An older version of this paper with detailed discussion of individual results and code is available at `https://github.com/hammondm/phone2vec`.

References

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

Chomsky, Noam & Morris Halle (1968). *The Sound Pattern of English*. Harper & Row, New York.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Geéraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky andJason Eisner & Mans Hulden (2017). CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. Hulden, Mans (ed.), *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, Association for Computational Linguistics, Vancouver, 1–30.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Goldsmith, John & Aris Xanthos (2009). Learning phonological categories. *Language* 85, 4–38.

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

Hayes, Bruce (1981). *A Metrical Theory of Stress Rules*. Garland, New York. 1980 MIT doctoral dissertation.

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

Liberman, Mark & Alan Prince (1977). On stress and linguistic rhythm. *Linguistic Inquiry* 8, 249–336.

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

Mikolov, Tomáš, Kai Chen, Greg Corrado & Jeffrey Dean (2013a). Efficient estimation of word representations in vector space. ArXiv:1301.3781.

Mikolov, Tomáš, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean (2013b). Distributed representations of words and phrases and their compositionality. 3111–3119, URL `http://arxiv.org/abs/1310.4546`.

Mikolov, Tomáš, Wen-tau Yih & Geoffrey Zweig (2013c). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 746–751.

Pennington, Jeffrey, Richard Socher & Christopher D Manning (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

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

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

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