Interaction of lexical strata in hybrid compound words through gradient phonotactics

Eric Rosen[1] Matthew Goldrick[2]
[1]*University of Leipzig and* [2]*Northwestern University*

1 Introduction

We analyse hybrid compound words in Japanese, where a hybrid compound is one formed from stems that belong to more than one lexical stratum. In native-foreign compounds, where one stem belongs to one of the three native strata: Yamato, Sino Japanese and mimetic, as identified by Itô & Mester (1995) and the other to the foreign stratum, we observe that violations of phonological wellformedness constraints in the foreign stem are significantly less probable than in pure foreign words. These observations are explainable through gradient phonotactic probability, where the probability of a phoneme is determined by the whole sequence of phonemes that precedes it. We shall argue that this observed phonotactic behaviour of hybrid compounds is best explained by the hypothesis that both lexical strata distinctions and phonotactics are graded rather than categorical.

In §2, we introduce the lexical strata of Japanese and some phonotactic constraints that apply differently in different strata. In §3 we introduce our corpus data and show some statistical properties that point to an effect of differing phonotactics between strata gradiently affecting stem choice in hybrid compounds. In §4, we introduce the concept of graded phonotactic probability. In §5, we introduce the RNN phonotactic model we use and how it operates. In §6 we discuss probabilistic models of phonotactics with an example of a word passed through an RNN model, and show how the model's probabilities of a stem in a compound vary depending on the nature of the stem that precedes it. In §7 and §8 we present results of tests with random pairings of stems from different strata that are fed to the model and how the results support our hypothesis of graded inter-stratal phonotactic effects. In §9 we show that the divide between strata when measure on phonotactic probabilities is graded rather than categorical. §10 concludes.

2 Lexical strata in Japanese

There is good evidence (Smith & Tashiro, 2019) that lexical strata (Itô & Mester, 1995) are a productive part of the synchronic knowledge of Japanese speakers. In table 1 are examples from each of the four lexical strata identified by Itô & Mester (1995)

Stratum:	Yamato	Sino-Japanese	Mimetic	Foreign
Example:	*inu*	*ryokoo*	*pikpika*	*firumu*
	'dog'	'travel, trip'	'glittering'	'film'

Table 1: Japanese lexical strata

Here, we focus on differences between the foreign stratum and the other three and collapse the Yamato, Sino-Japanese and mimetic strata into a single 'native' group that contrasts with the foreign stratum.

The foreign stratum allows certain phonotactic patterns such as the three shown with examples below, that are disallowed in the three native strata.

* We wish to thank three anonymous reviewers and audiences at AMP 2021 for helpful comments and suggestions. All errors are our own.

Proceedings of *AMP 2021*

Native Yamato, Sino-Japanese, Mimetic	Foreign

Table 2: Collapsing the three 'native' strata

Pattern allowed only in foreign stratum	Example
di/ti	di *iraa* 'dealer'
f{a,i,e,o}	fi *rumu* 'film'
tautopmorphemic [+voi, -son] ... [+voi, -son]	b *a* d *ominton* 'badminton'

Table 3: Constraints violated in foreign stratum only

3 Corpus data and observations

To investigate the phonotactic behaviour of words in different strata we use a corpus of 70,000+ Japanese words from the NHK pronunciation dictionary, in phonemic representations.

Table 4 shows our calculations of violations per character (VPC) of the above three phonotactic constraints in pure foreign words vs. the foreign part of hybrid compounds within the corpus. We find that for each constraint, the foreign part of hybrid compounds avoids these patterns to a statistically significant degree as compared to their per-character violation rate in pure foreign words.

VPC = violations per character/bigram in foreign words vs. *foreign part* of hybrids					
Constraint	VPC pure foreign	VPC foreign part of hybrid	Ratio	$\chi^2_{df=1}$ N=41,618	p-value
*[+voi, -son] ... [+voi, -son]	0.017	0.010	0.569	11.71	0.001
di/ti	0.003	0.001	0.241	5.84	0.016
f{a,e,i,o}	0.004	0.002	0.455	6.06	0.014

Table 4: Statistics on markedness violations

4 Graded phonotactic probability

Rather than a discrete delineation between lexical strata, we find that through the phonotactic model we shall present, the resulting phonotactic probabilities define a gradient continuum of allowed markedness among words: what Smith & Tashiro (2019:2), citing Kiparsky (1973) refer to as a "hierarchy of foreignness", similar to Itô & Mester (1995)'s 'core-periphery structure'. We account for these tendencies by positing graded degrees of well-formedness arising from graded phonotactic probability.[1]

As an example of graded phonotactic properties within a stratum, our model assigns a per-phoneme exponentiated average probability, calculated as $\exp(\frac{1}{n}\sum_i^n \text{logprob}(x_i))$, x_i= the ith phoneme, to foreign word *byuffe* 'buffet' of only 0.008, where /fe/ and /ff/ are marked, but 0.215 to less foreign-sounding but foreign word *koodo* 'cord; code' which is homophonous with several non-foreign words.

We adopt a model of gradient phonotactics (Mayer & Nelson, 2020) in which the occurrence of a phoneme in a word is assigned a probability based on all the phonemes that precede it. This leads to a bias in compounds, where the phonotactics of the first morpheme influence the probability of the second. Two results of these observations are (a) that strata membership is graded and (b) that the phonotactics of the first morpheme in a hybrid compound can influence the choice of the second morpheme so that it avoids phonotactics that are too improbable given the phonotactics of the preceding first morpheme. This means that even if morphemes from different strata observe different phonotactic constraints, a morpheme from

[1] See also Hayes (2016), who uses a MaxEnt model to measure gradient membership in the Latinate vs. Native strata of English, which he measures as scores on a scale based on weighted constraints that favour or disfavour membership in one of the strata.

one stratum can affect the choice of the other from a different stratum, even if the two strata have different rankings of phonotactic markedness constraints such as those shown above.[2,3]

5 The model: "A stratified RNN"

We use a recurrent neural network language model (RNNLM), designed to learn gradient phonotactics by optimizing on the probability of a phoneme occurring in a word based on the phonemes that precede it. This kind of phonotactic model was first proposed by Mayer & Nelson (2020) (henceforth M&N). We use Max Andrew Nelson's github code (Nelson & Mayer, 2020), to which we made minor changes so that it processes the Japanese corpus instead of the language data analysed by M&N.

We adopt a maximally simple version of an RNN that is unidirectional and has just a single hidden layer because, as discussed by M&N, a 1-layer RNN of finite precision is unable to learn unattested phonological patterns such as $a^n b^n$ (Merrill et al., 2020; Weiss et al., 2018). Each cell is fed (a) a vector-encoding of the input segment (explained below) and (b) the vector output of the previous hidden state. It applies a separate linear transformation to each, sums them, applies a non-linear $tanh$ function, and outputs a vector which is softmaxed to give a probability distribution over candidate phonemes. M&N have two possibilities for the vector-encoding of each phoneme: (a) an encoding based on phonological features and (b) an encoding that learns abstract representations of features. We found that the latter encoding produced better results. To avoid overfitting with the embedding model, we follow M&N's practice of 'tying' the embeddings of phonemes with the output weight matrix, each with dimension 24. (See M&N p. 151 for detailed discussion.) The objective of the model is to maximize the overall probability of each phoneme, averaged across positions in words and words in the database. We trained the model on the 75,000 words in the NHK database excluding a held out a test set of 309 native-foreign hybrids, 302 foreign-native hybrids and also excluding all the foreign and native stems that occurred in the hybrids. That way, the model will not have learned anything about the specific hybrids we are examining in testing.

6 Probabilistic models of phonotactics through an RNN model

Probabilistic models of phonology (Hayes & Wilson, 2008; Pierrehumbert, in press) can capture fine-grained phonotactic properties that go beyond what categorical constraints can capture. For example, foreign words *nikkaabokkaa* 'knickerbocker', *baatendaa* 'bartender' and *maakettorisaati* 'market research' violate none of Itô & Mester (1995)'s strata-distinguishing constraints but have phonotactic patterns uncharacteristic of native words with their repeated and/or word-final occurrences of long /a/. Our model feeds each word, one phoneme at a time, into a recurrent neural network that is trained to predict the next phoneme based on the phonemes that precede it. As shown in figure 1, at each timestep, it outputs a probability distribution over possible phonemes in the language. The probability of the word *byuffe* assigned by the model can be found by taking the exponential of the average log probabilities of each of the target phonemes in the top row. The probabilities shown along the top of the diagram are actual rather than log probabilities.

As an example of the lower probabilities the model assigned to marked phonotactic patterns in the hybrids as opposed to in pure foreign words, we found that in held-out testing, the probabilities given by the model to the /fi/ sequence (disallowed in non-foreign words) in native+foreign hybrids *funenesee-firumu* 'non-flammable film' and *sekigaisen-firumu* 'infrared film' are considerably less than those assigned to the same sequences in pure foreign words *karaaa-firumu* 'colour film', *maikuro-firumu* 'microfilm'. We show in table 5 below the model probabilities for initial /fi/ in these examples.

[2] An anonymous reviewer asks "if we need abstract phonological constraints like Lyman's Law and *di/ti when the observed patterns are so well-explained in terms of transitional probabilities." Something that a more abstract constraint like *[+voi,-son]…[+voi, -son] still achieves, whether it is weighted as in HG or categorical as in OT, is to make the kinds of generalizations that we predict a speaker would make over the more specific segmental probabilities that form the input to our RNN model.

[3] An anonymous reviewer notes that our model could also be used to test the kinds of experimental results produced by Moreton & Amano (1999), who find that Cy sequences that occur early in a word and are indicative of the Sino-Japanese stratum affect the perception of vowel length later in the word.

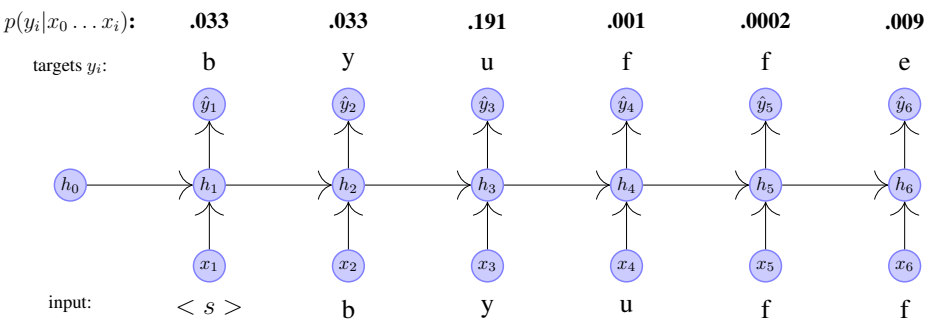

Figure 1: Model being fed word *byuffe* 'buffet'

hybrid		foreign		foreign to hybrid
funensee-**fi**rumu		karaa-**fi**rumu		probability ratios
f	i	f	i	
0.0005	0.0237	0.0043	0.0776	f: 8.6
sekigaisen-**fi**rumu		maikuro-**fi**rumu		
f	i	f	i	
0.0002	0.0363	0.0017	0.0971	i: 2.9

Table 5: Phoneme probabilities assigned by the model to held-out items

7 Selection of hybrid stem2s is biased by native phonotactics

How might our model give an explanation of our statistical observations regarding the avoidance of marked patterns in hybrids? As a first test, we used a Monte Carlo method to see if the differences in the hybrid vs. pure foreign second stem's phonotactics were different than expected by chance. We estimated the chance distribution by creating 1000 sets of random pairings of native stems with foreign stems. We then compared the model's mean exponentiated average phoneme probability of these 1000 random sets with those of the real foreign-native hybrids. In the random pairings, the first (foreign) conjunct was randomly chosen from the stems that occur in real foreign+native hybrids. The native stem was chosen from all listed native words.

The mean per-word phoneme probabilities for the random pairings ranged from 0.0859 to 0.0891. The mean per-phoneme probability for the real hybrids was 0.0901; as this exceeds the chance distribution, there is a less than 1/1000 probability that the change in phonotactic probabilities occurred by chance (i.e., $p < .001$).

The kernel density plot in figure 2 shows the distribution of 1000 mean word probabilities in the random pairings compared with the mean word probability for the real hybrids (vertical line.) As we can see, the mean of the real hybrids is well to the right of the distribution of 1000 mean probabilities on the graph.

We also tested 1000 sets of random pairings of native+foreign compounds, where the native stem was chosen from stems that occur in native+foreign real hybrids and the foreign stem chosen from all listed foreign words. In this test, the mean per-phoneme probability for the real native hybrids was 0.102. In contrast, the mean per-phoneme probabilities for 1,000 sets of random pairings had a minimum value of 0.075 and a maximum value of 0.080: distinctly below those of the real native hybrids. Figure 3 shows a kernel density plot comparing the distribution of these random pairings with the mean word probability of the native-foreign real hybrids. Here, we see that the mean probability of the real hybrids is even further to the right of the distribution of random pairings than in the previous graph.

The differences seen in these pairs of distributions indicate that in hybrid compounds there is a bias towards selecting morphemes from the two strata such that the phonotactics of the first do not make the

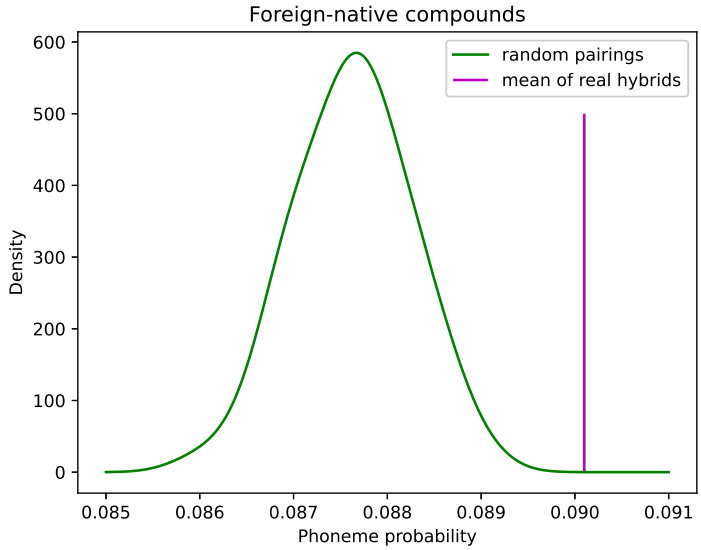

Figure 2: Kernel density plot comparing the random pairings to real hybrids

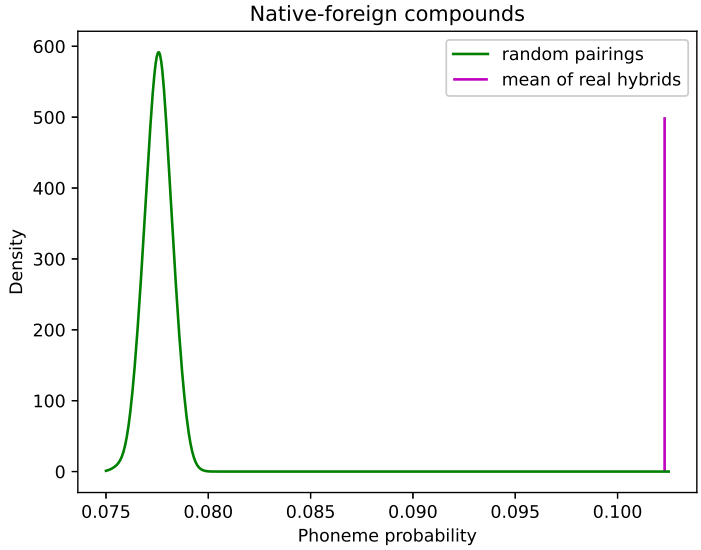

Figure 3: Kernel density plot comparing the random pairings to real hybrids

phonotactics of the second too improbable.[4]

The foreign stem that follows a native one in a hybrid compound does not exhibit, on average, the same gradient phonotactics as just any random foreign stem. The same appears to be the case for a native stem that follows a foreign one in a hybrid compound. The real hybrids have greater mean inverse perplexity compared to the random pairings – more so in the case of the native+foreign hybrids. We see this in the two test results,

[4] See also Breiss & Hayes (2020) for a study of how phonological markedness affects speakers' choices of syntactic constructions and of synonymous words.

where certain marked configurations are less common in hybrids than in pure foreign words.

The avoidance of certain phonotactic patterns in hybrid compounds is not just limited to foreign stems in hybrids avoiding patterns that do not occur in native words. We also find that certain patterns that occur frequently in native words are avoided to a significant degree in the native part of native-foreign hybrids. We test this for consonant-glide sequences /ky/, /sy/ and /zy/, which occur frequently in words of Sino-Japanese origin (which we have classified with the 'native' group.) The occurrences of these sequencdes have a significant reduction in per-bigram frequency in native-initial hybrids in comparison to all listed words. Table 6 shows the frequencies of these sequences relative to all bigrams in the set of native-initial hybrids as compared with their frequencies in the whole corpus.

Bigrams in whole corpus	Bigrams in native-foreign	ky	sy	zy
520,996	4,104			
Pattern counts in whole corpus		2,270	6,086	2,966
Frequency		0.0044	0.0117	0.0577
Pattern counts in native-foreign		9	24	10
Frequency		0.0022	0.0058	0.0024
χ^2		3.93	5.23	7.09
p-value		0.0475	0.022	0.0077

Table 6: Frequency comparisons for 3 consonant-glide patterns

If the native part of a native-initial hybrid contains one of these sequences, which are very frequent in Sino-Japanese words, a gradient phonotactic model will anticipate following sequences to contain patterns that are typical of Sino-Japanese words and which are not expected in foreign morphemes, thus making the whole compound marked with respect to gradient phonotactics if a foreign stem were chosen as the second conjunct. This would result in the avoidance of the occurrence of native morphemes with these sequences in hybrid compounds.

Moreover, we find that the foreign stems listed in table 7 that *do* occur with in hybrids with an initial native stem containing, for example, /sy/, generally have phonotactic patterns that robustly occur in native words, the only salient exceptions being the word-final /aa/ sequence in *reedaa* and the single /p/ in *aisotoopu*. These specific observations are consistent with the above findings that the mean per-phoneme probability of real native-foreign hybrids produced by the model was significantly higher than for random pairings of native and foreign morphemes. None of these words contain any other phonemic patterns that are exclusive to foreign words such as voiced geminates, /fi/, /dyi/ or multiple voiced obstruents.

pan	'bread'	saabisu	'service'
daiya	'diamond; diagram'	horumon	'hormone'
uran	'uranium'	dansu	'dance'
reguhon	'leghorn'	nyuusu	'news'
tero	'terrorism'	basu	'bus'
anmoniumu	'ammonium'	kurabu	'club'
aisotoopu	'isotope'	reedaa	'radar'
kariumu	'kalium (potassium)'	ene	'energy'
tesuto	'test'		

Table 7: Foreign stems2 occurring after native words with /sy/

8 In foreign-native hybrids, the native stem2s are biased to be gradiently more foreign-like

So far, we have looked at the effects of native stems on foreign stems in native-foreign hybrids, where the presence of the native stem makes phonotactically marked patterns avoided in the choice of a foreign stem, even though the occurrence of these patterns would not have violated any constraints that apply to foreign words. We now investigate to what extent the converse effect holds: that is, does the presence of a foreign

stem1 affect the choice of a native stem 2 in native foreign hybrids with respect to its phonotactics. To do this, we trained the model 10X each on random partitions of stem2 of (a) native-native compounds and (b) foreign-native compounds. There were 1993 native stem2s and 256 hybrid stem2s. We eliminated longer words from the set of native-native stem2s so that the mean word lengths were around 4.5 in both sets. We ran 10 randomly partitioned train and test sets, with 256 items in test sets for both native and hybrids.

We tested on each pair of held-out sets of stem2s of native-native and foreign-native compounds, with mean word lengths equalized. The plots in figure 4 show the mean likelihoods of hybrid stem2s vs. native stem2s in the held-out test sets across 10 random data partitions. The boxplot on the left shows results for training on foreign stem2s; the rightmost for training of native stem2s.

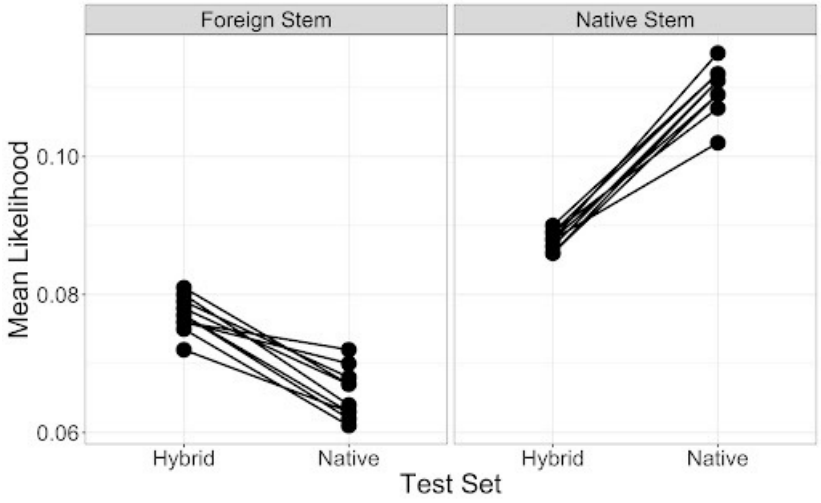

Figure 4: Boxplots of pairs of test items

We see that when the model is trained on foreign stems (left boxplot), the mean probabilities are consistently higher for foreign-native hybrids than for pure native words. When the model is trained on native stems, the opposite holds (right boxplot): pure native words are more probable than foreign-native hybrids. These results suggest that in foreign-native hybrids, the native stem2s are biased to be gradiently more foreign-like, even if they do not violate any categorical constraints that are violable only for foreign morphemes.

9 Divide between stem2 types is not categorical

The scatter plot in figure 5 shows per-word results from partition 10 of the above random pairings. Despite a bias for hybrid stems to be more foreign-like (below x=y line) and native stems to be more native-like (above x=y line), there is no categorical separation of the two. (Red and blue triangles on axes are means of each set of stem2s.)

10 Conclusions

We have examined patterns of phonotactic markedness in Japanese hybrid compounds, in which one stem is from the foreign stratum and the other from either the Yamato or Sino-Japanese stratum. We found that in native-foreign hybrids, there is a bias for the foreign part to be less marked phonotactically than pure foreign words. These results are apparent statistically, where marked patterns that are permitted only in foreign words occur significantly less frequently in the foreign part of hybrids than in pure foreign words. Moreover, we find that in applying a gradient phonotactic model to native-foreign hybrids and pure foreign words, the predicted probabilities of the foreign part of hybrids are on average lower than in pure foreign words. We find parallel tendencies in in foreign-native hybrids, where there is a bias for the native part to be phonotactically more foreign-like, but without violating any constraints that apply to native words. Both

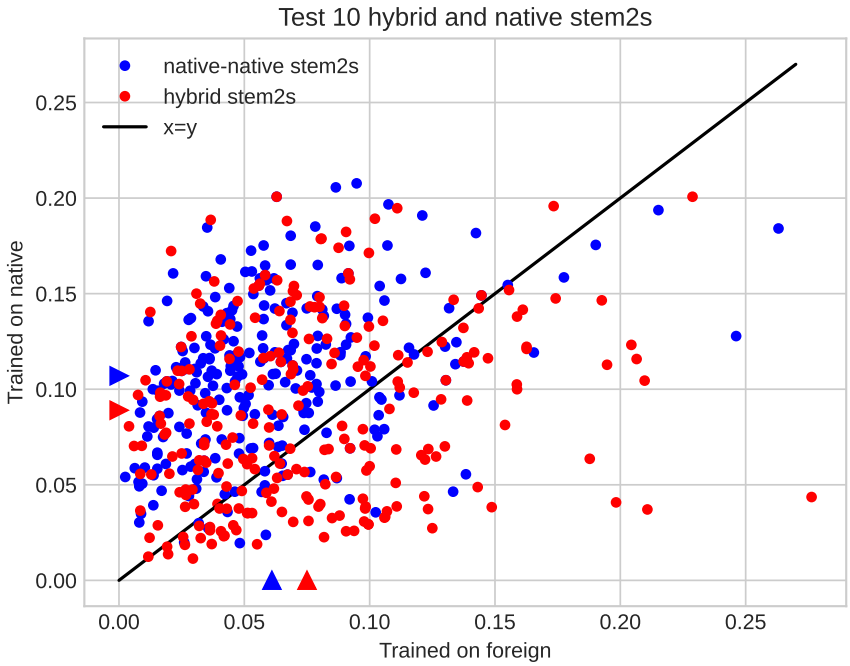

Figure 5: Scatter plot of per-word results from partition 10

tendencies are explainable by a gradient model of phonotactics, which we modeled through an RNN, where stem1's phonotactics bias stem2's.

References

Breiss, Canaan & Bruce Hayes (2020). Phonological markedness effects in sentence formation. *Language* 96:2.

Hayes, Bruce (2016). Comparative phonotactics. *Proceedings of the 50th meeting of the Chicago Linguistic Society*, 265–285.

Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:3, 379 – 440.

Itô, Junko & Armin Mester (1995). The core-periphery structure of the lexicon and constraints on reranking. *University of Massachusetts Occasional Papers in Linguistics*, vol. 18.

Kiparsky, Paul (1973). *Three dimensions of linguistic theory*, Tokyo Institute for Advanced Studies of Language, chap. How abstract is phonology?

Mayer, Connor & Max Nelson (2020). Phonotactic learning with neural language models. *Proceedings of the Annual Meeting of the Association for Computation in Linguistics*, vol. 3.

Merrill, William, Gail Weiss, Yoav Goldberg, Roy Schwartz, Noah A. Smith & Eran Yahav (2020). A formal hierarchy of RNN architectures. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 443 – 459.

Moreton, Elliott & Shigeaki Amano (1999). Phonotactics in the perception of Japanese vowel length: evidence for long-distance dependencies. *Proceedings of the IEEE* .

Nelson, Max & Connor Mayer (2020). Phonotactic language model. https://github.com/MaxAndrewNelson/PhonotacticLM.

Pierrehumbert, Janet (in press). *Oxford Handbook on the the History of Phonology*, Oxford University Press, chap. 70+ years of probabiliistic phonology.

Smith, Jennifer & Yuka Tashiro (2019). Nonce-loan judgments and impossible-nativization effects in Japanese. *Proceedings of the LSA* 4:26, 1–14.

Weiss, Gail, Yoav Goldberg & Eran Yahav (2018). On the practical computational power of finite precision rnns for language recognition. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, 740 – 745.