

# Lexical Strata and Phonotactic Perplexity Minimization

Eric Rosen

*Johns Hopkins University*

## 1 Introduction

We test the hypothesis that in some languages the lexicon is stratified (Itô & Mester, 1995a) and that multiple phonotactic subgrammars based on gradiently measured phonotactics not only reduce average phoneme uncertainty, but align well with proposed lexical strata that are based on categorical constraint ranking differences.

Whereas some recent studies (Smith, 2018; Hsu & Jesney, 2017; Hearn, 2016) address the question of lexical stratification directly through interactions of categorical or gradient phonotactic and faithfulness constraints, here we adopt a neural network approach, originating with Elman (1990) and most recently implemented by Mayer & Nelson (2020) (henceforth M&N) which captures phonotactic knowledge through relatively simple recurrent neural language models (RNNLMs) that predict the next phoneme given the previous phonemes in the word. M&N propose that the networks they test on Finnish, Cochabamba Quechua and English can “learn and generalize phonotactic patterns as well as or better than” constraint-based Max-Ent models.

Hayes & Wilson (2008)’s model of phonotactics introduced into mainstream phonological theory the conception of phonotactic knowledge as probabilistic gradience. For example, in English [pr] is a more probable onset cluster than [θw], but both are possible. Here, we ask: if a grammar can account for phonotactic patterns probabilistically, and having multiple subgrammars achieves a greater overall probability of the data of a language, how might such probabilistically optimal subgrammars place words into phonotactically differing lexical strata?

We test this idea on the well-known hypothesis of lexical stratification in Japanese (Itô & Mester, 1995a), in which the proposed strata – Yamato (native), Sino-Japanese, mimetic and foreign – exhibit different phonotactic properties. We apply a modification of M&N’s code (Nelson & Mayer, 2020), to a corpus of 25,000+ words from NHK (1999), converted to phonemic representations. The model learns a RNNLM whose objective function is to minimize the overall phoneme perplexity<sup>1</sup>, averaged across positions in each word and across words in the database. We then bifurcate the model into two separate RNNLMs, with no prior bias given to each, and the model calculates the perplexity of each word as the minimum value of what each of the models comes up with, in effect assigning each word to one of two grammars/models, with no supervision about a word’s lexical stratum.

## 2 The experiment

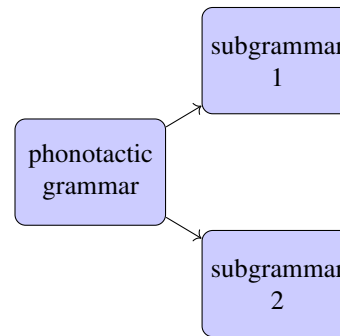
Our hypothesis, then, is that a learner, faced with sets of words that exhibit very divergent phonotactic properties, would allow their phonotactic grammar to diverge into sub-modules that align with each divergent set of words, as shown in figure 1.

We ask, to what extent would these sub-modules align with the lexical strata such as those proposed by Itô & Mester (1995a) for Japanese, which subdivides the lexicon into strata (Yamato, Sino-Japanese, etc.),

---

\* After being presented at AMP 2020, an extended abstract of this paper was submitted to and will be published in the Proceedings of the Society for Computation in Linguistics 2021, which allowed short submissions that were previously presented at a conference with a different audience. Three anonymous reviewers each from AMP and SCiL provided valuable suggestions. Thanks also to Paul Smolensky, Caitlin Smith, Matt Goldrick, Connor Mayer, Max Nelson, Jennifer Smith, Jason Shaw, John Alderete and Shigeto Kawahara for helpful feedback and suggestions. Research was generously sponsored by Microsoft Research and NSF INSPIRE grant BCS-1344269. All errors are my own.

<sup>1</sup> M&N calculate the perplexity as “the exponentiated entropy, or inverse of the mean log likelihood, of all phonemes in the test word.”



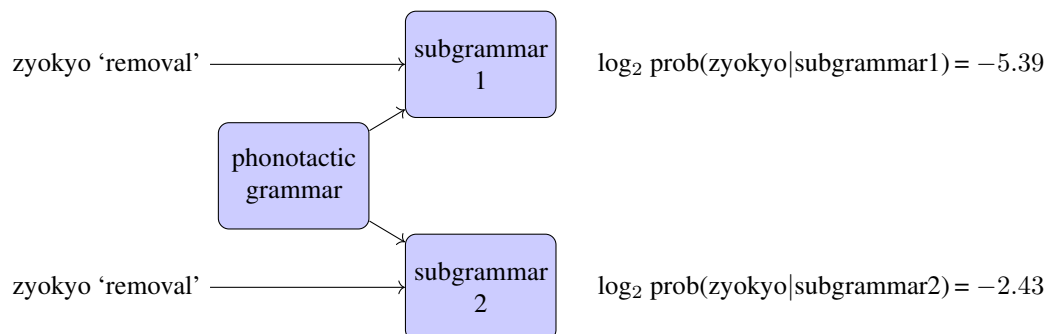
**Figure 1:** Bifurcation of grammar into sub-grammars

each of which has a different ranking of some constraints in Optimality Theory (Prince & Smolensky, 1993)? Figure 2 recreates a similar table to Itô & Mester (1995a:184) which shows their proposed strata and how they differ with respect to how words in each stratum obey certain categorical phonotactic constraints. (d.n.a. == ‘does not apply’.)

	SYLLSTRUC	NO-DD	NO-P	NO-NT
Yamato	✓	✓	✓	✓
Sino-Japanese	✓	✓	✓	d.n.a
Foreign	✓	✓	d.n.a	d.n.a
Unassimilated Foreign	✓	d.n.a	d.n.a	d.n.a

**Figure 2:** Itô and Mester’s constraint violations in lexical strata

Here we adopt a probabilistic model of phonology (Pierrehumbert, 2015) because it can capture fine-grained phonotactic properties that go beyond what categorical constraints can capture. For example, Sino-Japanese word *zyokyo* 除去 ‘removal’ violates none of the constraints in Itô and Mester’s tableau but has a phonotactic pattern (offglide after onset consonant) seldom seen in Yamato words. In Japanese, offglides **can** occur as part of onsets in Yamato (native) words (e.g., *kyuuri* 胡瓜 ‘cucumber’ (Martin, 1987:469)), but such sequences are rare in Yamato words and extremely common in Sino-Japanese words. In our experiment, as illustrated in figure 3, we simulate a putative divergence of a phonotactic grammar into sub-modules by feeding a corpus of Japanese words into a dynamic probabilistic model that is allowed to fork into two sub-models.



**Figure 3:** Sample word *zyokyo* ‘removal’ fed into two sub-grammars

**2.1 Outline of the experiment** We use a corpus of 25,000+ Japanese words from NHK (1999), converted to phonemic representations:

... 除去 → ジョキョ → zyokyo ...

We feed them into a maximally simple recurrent neural network, modeled after Mayer & Nelson (2020); Nelson & Mayer (2020), whose one-layer RNN of finite precision has been shown to be unable to learn unattested patterns such as  $a^n b^n$  (Weiss et al., 2018; Merrill et al., 2020). Each cell  $h_i$  of the RNN is fed (a) a vector-encoding of the input segment  $x_i$  and (b) the vector output of the previous hidden state  $h_{i-1}$ . It applies a separate linear transformation to each, sums them, applies a non-linear function such as tanh, and outputs a vector which is softmaxed to give a probability distribution over candidate phonemes  $y_i$ . Its objective is to minimize the overall negative log probability of each phoneme, averaged across positions in words and words in the database. The model is initialized as two subnetworks, identical except that each has a different random initialization. Each word is fed into both submodels, each of which tries to predict each segment based on the string that precedes it.

Figure 4, copied from M&N, illustrates the architecture of one timestep of a simple RNN.  $x_t$  is a phoneme input at timestep  $t$ ,  $h_{t-1}$  is the output of the network’s hidden layer at time  $t - 1$ , recycled back on the next timestep,  $W_h$  and  $W_x$  are linear transformations with an added nonlinearity, and  $W_y$  is a linear transformation to produce output  $y_t$  for each timestep.

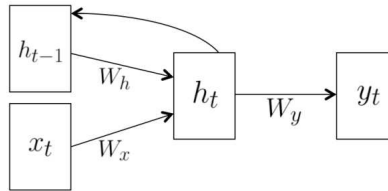
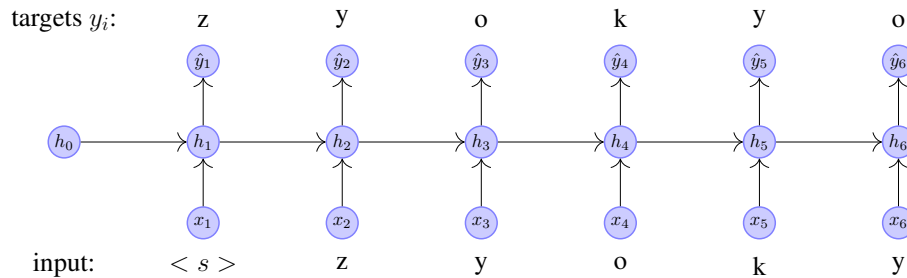


Figure 1: Schematic sRNN architecture

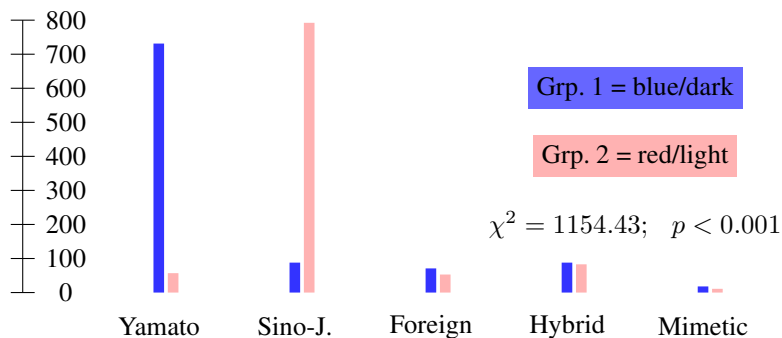
**Figure 4:** Mayer and Nelson’s diagram of an RNN cell

As phoneme vectors are input to the model over time, an unrolled model that is fed example word *zyokyo* 除去 ‘removal’ looks as shown in figure 5:



**Figure 5:** Unrolled Model over time

Each word in the dataset is fed to each of two randomly initialized submodels. The submodel that a given word performs best on is updated with backpropagation to improve that word’s predicted probability. But the other submodel is not updated. If the words diverge enough in their phonotactics, the submodels will also diverge, with some words being more predictable with one submodel and other words with the other. The learning is unsupervised, in that the words are not tagged with any strata labels such as ‘Yamato’ or ‘Sino-Japanese’. The model quickly plateaus after running through all the data for only 3 epochs. The words end up in two groups, with membership of each word determined by the model that gave it the highest probability at the end of learning. In a random sample of 1,000 words from each of the resulting groups, group 1 has a strong presence (73.2%) of Yamato words (words that Itô and Mester’s stratification would place in a Yamato stratum) but few Sino-Japanese words, which dominate group 2 (79.3%) which, in turn, has few Yamato words, as shown in figure 6:



**Figure 6:** Membership in strata of 1000 words assigned by each sub-model

Many of the misclassified words could phonotactically occur in either stratum: misclassified Sino-Japanese words *yaku-ri* ‘pharmacology’ 薬理 and *sui-ro* 水路 ‘watercourse’, are homophonous with fictitious Yamato compounds *ya-kuri* 家栗 ‘house-chestnut’ and *su-iro* 巢色 ‘nest-colour’.

The outputs of each RNN at each timestep reveal differences in predictions that mirror gradient phonotactic differences between Yamato and Sino-Japanese words. Among  $\sim 4000$  nouns and  $\sim 2000$  verbs Martin (1987)’s diachronic study of Yamato Japanese, only 15 lexemes have a word-initial consonant-offglide sequence such as [#ky–]. Such [Cy] sequences are extremely common among Sino-Japanese words (e.g. city name 京都 *kyoto* ‘Kyoto’.) Conversely, diphthong [ae] which occurs frequently in the Yamato lexicon (e.g., *mae* 前 ‘before’) occurs rarely if at all tautomorphemically in Sino-Japanese words.<sup>2</sup>

For comparison, we ran a bigram model that predicts only from the previous segment. It misclassifies Sino-Japanese words at a 68% higher rate than the n-gram model, suggesting that n-gram segmental patterns contribute to the gradient phonotactics of the language. (E.g., bigrams will not detect the fact that few Yamato words have /e/ in the first syllable.) (Martin, 1987:48)

Table 1 shows the ratio of probabilities assigned by  $RNN_1$  relative to  $RNN_2$  for offglide [y] to occur after selected word-initial consonants (column 2) and for [e] to follow a word-initial [Ca] sequence (column 3). We see that  $RNN_1$  favours the occurrence of offglides much more than  $RNN_2$  and  $RNN_2$  favours diphthong [ae] much more than  $RNN_1$ .

	#Cy sequences	#Cae sequences
#C	$p(y \#C; RNN_1)/p(y \#C; RNN_2)$	$p(e \#Ca; RNN_1)/p(e \#Ca; RNN_2)$
k	17.26	.061
s	7.62	.111
t	4.20	.281
n	47.61	.212
h	20.51	.095
b	3.32	.051
m	10.25	.169

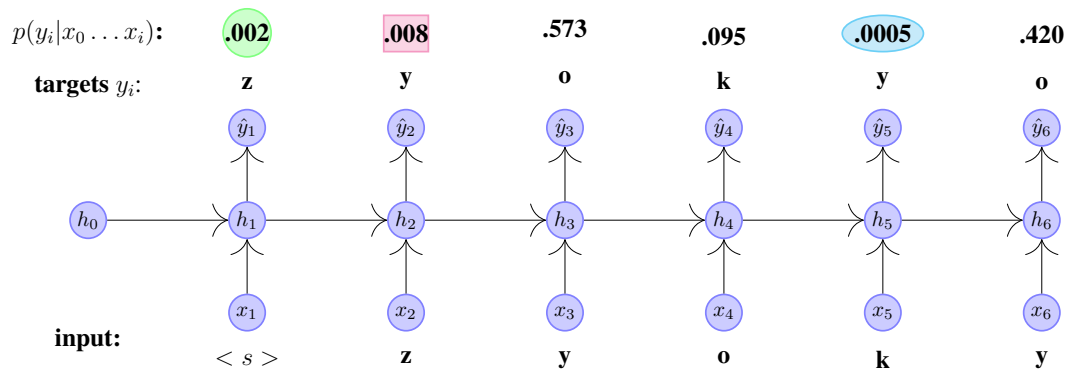
**Table 1:** Ratios of probabilities assigned by each of the two models to #Cy and #Cae sequences

These results suggest that the two-RNN model has encoded gradient phonotactic differences between Yamato and Sino-Japanese words.<sup>3</sup>

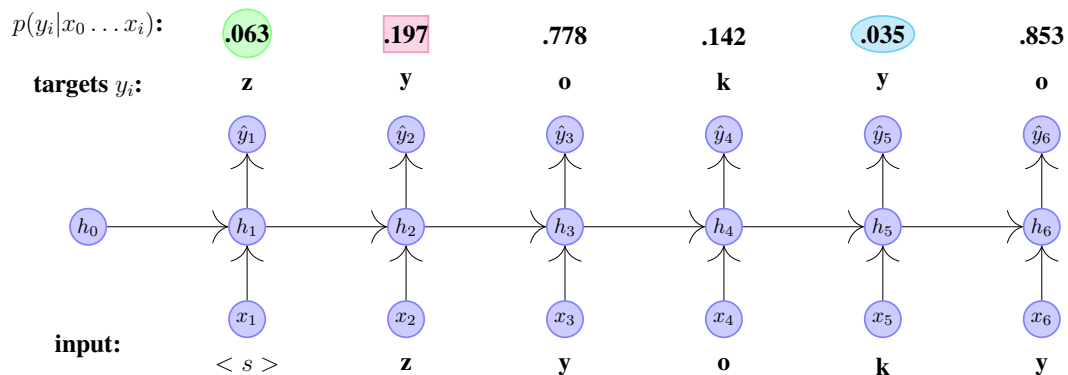
<sup>2</sup> See also Moreton & Amano (1999) whose psycholinguistic experiments use initial Cy sequences to trigger perception of a Sino-Japanese stratum, which in turn affects perception of vowel length later in the word.

<sup>3</sup> Not all languages that experience borrowing will necessarily exhibit strata: arguably, only if the phonotactics of the source and languages differ enough.

**2.2 Schematic of the RNN model** Sample word, *zyokyo* 除去 ‘removal’ is shown in figures 7 and 8 processed by each of the two submodels. Its overall probability, calculated as the mean log probability of each segment, is 7.78 times higher for submodel 2 than with submodel 1. ( $2^{-2.43}/2^{-5.39}$ )



**Figure 7:** Model 1 Mean per-phoneme  $\log_2$  probability = -5.39



**Figure 8:** Model 2 Mean per-phoneme  $\log_2$  probability = -2.43

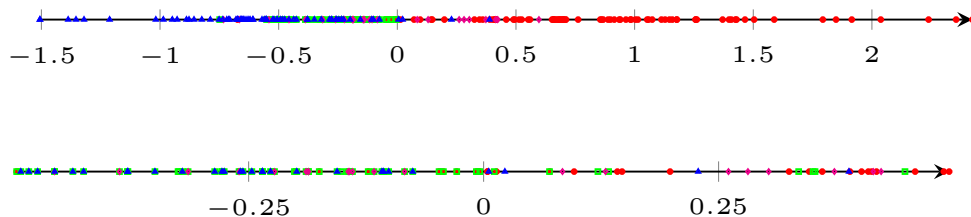
Pairs of segments with corresponding colours/box-shapes across the two models show a greater likelihood for group 2 than for group 1 by factors of 31, 24 and 70.

One source of this difference is that the word-initial /z/ is uncommon in Yamato words, which clustered with submodel 1, but is not uncommon in Sino-Japanese words. And the offglides that follow both the z and also the k, as mentioned on page 2, occur frequently in Sino-Japanese words but rarely in Yamato words. What we see, then, is that unsupervised clustering with phonotactic submodels that are allowed to fork aligns strongly with strata that have been proposed on the basis of categorical constraints. Not all languages that have experienced borrowing will necessarily do this, if the phonotactics of the source language are not as different from those of the native language as is the case for Japanese borrowing from Chinese.

### 3 Gradient membership in strata

Hayes (2016) and Jennifer Smith (p.c.) both cite Itô & Mester (1995b:821) suggesting that membership in lexical strata may be gradient. Hayes (2016) explores, using a MaxEnt model, gradient membership of English words in Native vs. Latinate vocabularies as scores on a scale based on weighted constraints that favour or disfavour membership in one of the strata. Whereas Hayes’ model uses heuristics to pre-classify a word’s stratum membership and pre-defines phonotactic constraints, our model allows strata to emerge on their own without pre-assignment and constraints to emerge latently by the probabilities the model assigns to segment in a particular environment.

To examine how our model might assign words gradually into strata<sup>4</sup>, we took random samples of 100 words each assigned to groups 1 (mostly Yamato) and 2 (mostly Sino-Japanese), with differences of  $\text{perplexity}_2 - \text{perplexity}_1$  shown in the first plot, and the most marginal words ( $|\text{diff}| < 0.5$ ) in the second plot. ▲ = Yamato, ● = Sino-Japanese, ■ = Foreign, ◆ = hybrid or ambiguous.



**Figure 9:** Perplexity margins for sample of words in groups 1 and 2

The four most marginal, misclassified Sino-Japanese words in group 1 (red dots left of 0), are *hi-dai* 肥大 ‘corpulence’ (lit. ‘fatten-big’), *ei-yo* 栄誉 ‘honour’ (lit. ‘honour-honour’), *ku-iki* 区域 ‘district’ (lit. ‘ward-level’) and *ki-matu* 期末 ‘end-of-term’ (lit. ‘term-end’) with margins of -0.004, -0.008, -0.047 and -0.043 respectively, which are homophonous with fictitious Yamato compounds *hida-i* 襞胃 ‘pleat-stomach’, *ei-yo* 鱒夜 ‘ray(fish)-night’, *kui-ki* 杭木 ‘stake-tree’ and *ki-matu* 木松 ‘tree-pine’.<sup>5</sup> On one hand, the abundance of morphemes that have different Sino-Japanese and Yamato readings of the same kanji (e.g., *moku* and *ki* for 木 ‘tree’), means that the stratum membership of a given reading is discretely determined by which side of the pronunciation contrast it is on: *moku* is Sino-Japanese because it contrasts with *ki* for the same kanji/morpheme. On the other hand, many readings of either type, Sino-Japanese or Yamato, not only satisfy all the constraints that Itô and Mester proposed for distinguishing the two strata, but show marginal differences in the phoneme perplexity assigned by each model. This makes these phoneme sequences ambiguous as to their stratum. Because of the abundance of homophony in Japanese, it is easy to find homophonous pairs such as *atu* with a Sino-Japanese reading for a morpheme meaning ‘pressure’ (as in *si-atu* ‘finger-pressure, shiatsu’) and a Yamato reading for an unrelated adjectival morpheme meaning ‘hot’. The fact that such phoneme sequences do not have a shape that is characteristically Yamato or Sino-Japanese makes them good candidates for having gradient strata membership in a way analogous to English words that Hayes judges to be ‘intermediate in Latinity.’

To put the inconclusiveness another way, on one hand, the dichotomies between *On* (Sino-Japanese) and *Kun* (Yamato) readings of kanji characters are discrete rather than continuous, which makes it difficult to classify a Sino-Japanese word like *sin-setu* 親切 ‘kindness’ as only partly Sino-Japanese, when the *Kun* readings of the two characters: 親 *oya* ‘parent’ and 切 *ki(ri)* ‘cut’ are clearly different both phonologically and semantically. On the other hand, words whose form is ambiguous between Sino-Japanese and Yamato such as the above examples *ku-iki* and *ki-matu* could be considered intermediately Sino-Japanese, given that their form, with syllables being CV, is actually more Yamato-like.

Also in the marginal group with a margin of 0.084 is *tooya* which is ambiguous between Sino-Japanese 陶冶 ‘cultivate’ (lit. ‘porcelain-melting’) and Yamato 遠矢, a family name and placename (lit. ‘distant-arrow’).

If we look at misclassified Yamato words in group 2 (blue dots right of 0) we find fewer marginal words. We do find *tooku* ‘far’ (adv.), (which is also homophonous with foreign borrowing ‘talk’), and *atude* ‘thick’ (lit. thick-hand) with margins 0.023 and 0.228 respectively. *tooku* has many candidates for homophonous fictitious compounds, including what appears to be a recently coined compound 投句 ‘posting a haiku poem in the internet’ (lit. ‘throw-stanza’). In the marginal group are also two hybrid compounds, *modosi-zee* ‘tax refund’ (lit. ‘return(trans.)-tax’, Yamato+Sino-Japanese) and *zyo-no-kuti* ‘beginning’ (lit. ‘beginning-entrance’, Sino-Japanese+Yamato) with margins of 0.084 and 0.130.

<sup>4</sup> There will be some oversimplification in that so far, we have only used two RNN models in spite of evidence of more than two strata in Japanese.

<sup>5</sup> The last one is not quite fictitious, having been coined as the actual name of a hotel in Hiroshima.

## 4 Discussion and directions for future research

One of the limitations of the present study is that the dataset being tested was taken from the full set of 116,000+ entries in NHK (1999) but then reduced to a smaller set of words that exceeded a frequency threshold. Words in the dataset were fed into the model one at a time, each word once per epoch. This resulted in foreign words making up a relatively small proportion of the corpus – so small that when the experiment was performed with three RNNs, there were not enough foreign words to command one of the three groups into which the three sub-models sorted words. The same goes for mimetic words, which made up an even smaller proportion of the dataset as can be seen in figure 6 above. Further research with three or four RNNs and a different dataset in which foreign and mimetic words are more strongly represented might test whether a third foreign-dominated group or fourth mimetic group might emerge.

In Itô & Mester (1995b)'s core-periphery model of the Japanese lexicon, we can see a kind of gradient phonotactics based on categorical constraints, where words become increasingly foreign in nature as their location moves further to the periphery. (See Itô & Mester (1995b:834) example (34).)

The values given above in figure 9 show *relative differences* in average phoneme perplexity for a word between models, not the absolute average phoneme perplexity. If we look at absolute numbers, we do find that the words with higher absolute average phoneme perplexities would be located on the periphery of Itô and Mester's model. For example, foreign word *figyua* 'figure' has one of the highest average perplexities in both models of 4.20 and 4.72. It violates Itô and Mester's \*F constraint, whose boundary is near the periphery of the figure. Other words with high average perplexities are *uddo* 'wood' (3.98 and 5.60) *eabaggu* 'airbag' (3.94 and 4.05) which violate \*DD, and *refarensu* 'reference' (2.60 and 3.93), which violates \*F. Also with high perplexity is *doa* 'door' (3.54 and 3.99) whose *loa/* sequence does not occur tautomorphemically on the surface in Sino-Japanese or Yamato words. In the latter, the sequence is broken by glide /w/, as can be seen in a {e ~ a} alternation between transitive and inchoative verb pairs such as *oeru* ~ *owaru* 'finish'.

Because the model is calculating *average* phoneme perplexity of a word, the effects of a single phoneme with a low contextual probability could be diluted by having other phonemes with a higher contextual probability in the word. So the present model will not assign words a position on a core-periphery continuum in the same way that in Itô and Mester's model, a violation of a single constraint such as \*DD will assign it to a position near the periphery. As it turns out, though, we find that foreign words that violate this constraint will tend to have high average perplexities anyway, because other low-probability gradient phonotactic patterns will occur. The first 10 words in the dataset with geminate /dd/ consonants all had high perplexities in at least one of the two models, (and in fact, the low frequency of foreign words in the dataset would tend to give their phonotactic patterns a lower probability): *sarabureddo* 'thoroughbred' (2.36 and 3.51), *beddo* 'bed' (3.57 and 4.15), *suiaddoresu* 'studless' (2.57 and 3.40) *sofaabeddo* 'sofa bed' (3.39 and 4.38), *guddobai* 'good bye' (2.88 and 3.61), *heddohon* 'headphone(s)' (3.10 and 3.39), *deddobooru* 'dead ball' (2.64 and 3.25), *heddorampu* 'headlamp' (3.17 and 3.58) and *daburuheddaa* 'double header' (3.23 and 3.61).

Because the present model operates on the basis of contextual probabilities rather than discrete and categorical constraints, it will not create the same kind of lexical organization that is depicted in Itô and Mester's graphical depiction. If we were to graph the model's calculations for each word in the dataset in two dimensions, with the two x and y axes being the perplexities assigned by model 1 and model 2 respectively, we should find not only the group 1 words on the southeast side of an x=y diagonal and the group 2 words on the northwest side, but also the most peripheral words in the northeast corner on either side of the diagonal and the most core words of each group close to its axis. Because the model was bifurcated into two separate models, with words encouraged to cluster with one or the other, we tend not to find words that have low perplexity in both models. This is also because low perplexity means high contextual probability, not lack of constraint violations. For example, words that have perplexity of less than 1.0 for group 2, do not necessarily have low perplexity with group 1: e.g., *zyoozyoo* 上上 'excellent; other kanji and meanings' (grp. 1: 3.16), *tokusyoku* 特色 'characteristic(s)' (grp. 1: 2.40), *kansyoo* 干涉 'interference; other kanji and meanings' (grp. 1: 2.66) – all Sino-Japanese. There is a handful of words with perplexities of below 2.0 in both groups: Sino-Japanese *sikaku* 四角 'square; other kanji and meanings' (1.37 and 1.30), Sino-Japanese *hakuraku* 伯樂 'horse trader' (1.44 and 1.42), hybrid *yakusuru* 訳する 'translate' (1.46 and 1.43), Yamato *kasikari* 貸し借り 'lending and borrowing' (1.18 and 1.67) and Sino-Japanese *hankai* 半開 'half-open; other kanji and meanings' (1.90 and 1.12).

The experiments described in this paper are a first pass at testing how simple neural networks can learn

gradient phonotactic properties of words such as the probability of a given phoneme to occur after a given string, and in what ways they might be useful tools for capturing the ways in which gradient phonotactics separate words in a language into strata in both discretely and continuously. Hayes (2016:3) suggests that speakers of a stratified language internalize stratal divisions for stylistic reasons. Further research might examine whether this applies to Japanese, where there can be a choice among a Yamato, Sino-Japanese and foreign word for expressing the same meaning (e.g., *kuruma* 車, *zidoosya* 自動車, *kaa* カア for ‘car, automobile’).

## References

- Elman, Jeffrey L. (1990). Finding structure in time. *Cognitive Science* 14:2, 179 – 211.
- 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.
- Hearn, Ryan (2016). Rethinking the Core-Periphery Model: Evidence from Japanese and English. *Proceedings of the 24th Manchester Phonology Meeting*.
- Hsu, Brian & Karen Jesney (2017). Loanword adaptation in Québec French: Evidence for weighted scalar constraints. *Proceedings of the West Coast Conference on Formal Linguistics*, vol. 34, 249 –258.
- Itô, Junko & Armin Mester (1995a). The core-periphery structure of the lexicon and constraints on reranking. *University of Massachusetts Occasional Papers in Linguistics*, vol. 18.
- Itô, Junko & Armin Mester (1995b). *The Handbook of Phonological Theory*, Blackwell, chap. Japanese Phonology, 817 – 838.
- Martin, Samuel E. (1987). *The Japanese Language Through Time*. Yale University Press.
- 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 (Github repository). [https://github.com/MaxAndrewNelson/Phonotactic\\_LM](https://github.com/MaxAndrewNelson/Phonotactic_LM).
- NHK (1999). *NHK Hatsuon Akusento Jiten (NHK Pronunciation and Accent Dictionary)*. NHK (Japanese Broadcasting Corporation).
- Pierrehumbert, Janet (2015). *Oxford Handbook on the the History of Phonology*, Oxford University Press, chap. 70+ years of probabilistic phonology.
- Prince, Alan & Paul Smolensky (1993). Optimality Theory: Constraint Interaction in Generative Grammar. Tech. rep., Rutgers University Center for Cognitive Science and Computer Science Department, University of Colorado at Boulder.
- Smith, Jennifer (2018). Stratified faithfulness in Harmonic Grammar and emergent core-periphery structure. *Hana-bana: A festschrift for Junko Itô and Armin Mester* 13.
- 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.