Lexical strata and phonotactic perplexity minimization

Eric Rosen
Johns Hopkins University
erosen27@jh.edu

We test the hypothesis that in some languages the lexicon is stratified (Itô and 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 and Jesney, 2017; Hearn, 2016; Hayes, 2016) address the question of lexical stratification directly through interactions of categorical or gradient phonotactic and/or faithfulness constraints, here we adopt a neural network approach, originating with Elman (1990) and most recently implemented by Mayer and 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.

Hayes and Wilson (2008)'s model of phonotactics introduced into mainstream phonological theory the conception of phonotactic knowledge as probabilistic gradience.[1] 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 and 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 and Mayer, 2020), to a corpus of 75,000+ words from NHK (1999), converted to phonemic representations. The model learns a RNNLM whose objective function is to minimize the overall phoneme perplexity[2], 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 result between the two models, in effect assigning each word to one of two grammars/models, with no supervision about a word's lexical stratum.

The experiment We propose that a learner, faced with sets of words that exhibit divergent phonotactic properties, would allow their phonotactic grammar to diverge into sub-modules that align with each divergent set.

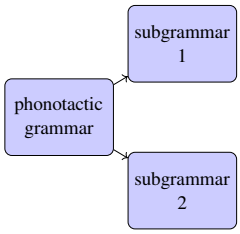


Figure 1: Bifurcation of grammar into sub-grammars

We ask, to what extent would these sub-modules align with the lexical strata proposed by Itô and Mester (1995, 1999) for Japanese, which subdivides the lexicon as shown in figure 2, where each stratum has a different ranking of some constraints in OT?

Here we adopt a probabilistic model of phonology (Pierrehumbert, 2015) which can capture fine-grained phonotactic properties that go beyond what categorical constraints can capture. For

[1] e.g., in English [pr] is a more probable onset cluster than [θw], but both are possible.

[2] M&N calculate the perplexity as "the exponentiated entropy, or inverse of the mean log likelihood, of all phonemes in the test word."

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 415-419.
Held on-line February 14-19, 2021

(16)

	SyllStruc	No-dd	No-p	No-nt
Yamato	✓	✓	✓	✓
Sino-Japanese	✓	✓	✓	violated
Assimilated foreign	✓	✓	violated	violated
Unassimilated foreign	✓	violated	violated	violated

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

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. Offglides occur robustly in Sino-Japanese words but rarely in Yamato (native) words such as *kyuuri* 胡瓜 'cucumber' (Martin, 1987, 469)). 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 two diverging RNNs.

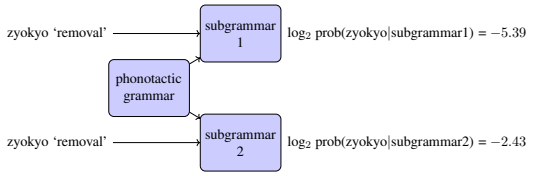


Figure 3: Sample word *zyokyo* 'removal' fed into two sub-grammars

Outline of the experiment We use a corpus of 24,000+ 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 and Nelson (2020); Nelson and 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, each with 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.

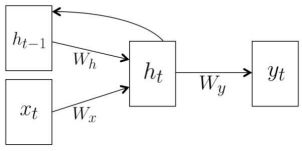


Figure 1: Schematic sRNN architecture

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:

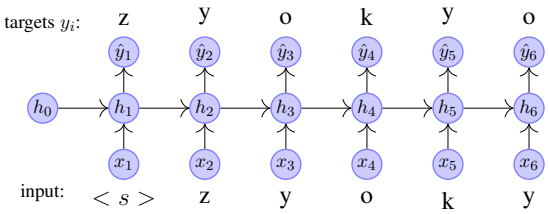


Figure 5: Unrolled Model over time

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 but few Sino-Japanese words, which

dominate group 2 (79.3%), which has few Yamato words, as shown in figure 6:

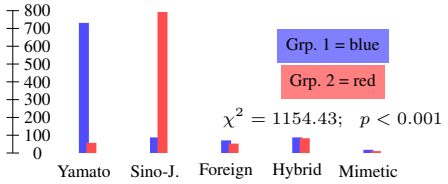


Figure 6: Membership in strata of 1000 words assigned by each sub-model

Many of the misclassified words could phonotactically occur in either stratum: misclassified SJ 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* 巣色 '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 ∼4000 nouns and ∼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 *kyooto* '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.[3]

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 with $n > 2$ contribute to the gradient phonotactics of the language.[4]

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). RNN_1 favours the occurrence of offglides much more than RNN_2 and RNN_2 favours diphthong [ae] much more than RNN_1.

These results suggest that the two-RNN model has encoded gradient phonotactic differences between Yamato and Sino-Japanese words.[5]

	#Cy sequences	#Cae sequences
#C	$\frac{p(y\|\#C;\mathrm{RNN}_1)}{p(y\|\#C;\mathrm{RNN}_2)}$	$\frac{p(e\|\#Ca;\mathrm{RNN}_1)}{p(e\|\#Ca;\mathrm{RNN}_2)}$
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

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. ($2^{-2.43}/2^{-5.39}$)

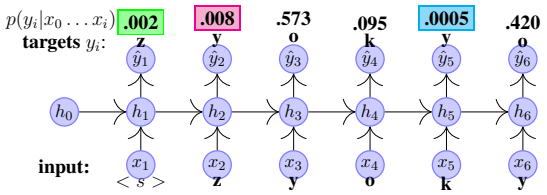


Figure 7: Model 1 Mean per-phoneme $\log_2$ probability = -5.39

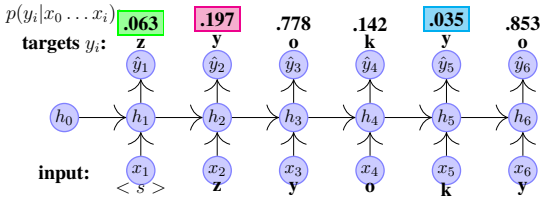


Figure 8: Model 2 Mean per-phoneme $\log_2$ probability = -2.43

Corresponding coloured pairs of segments across the models show a greater likelihood for group 2 than 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 not in Sino-Japanese words. And the offglides that follow both the z and the k are much more common in Sino-

[3] See also Moreton and 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.

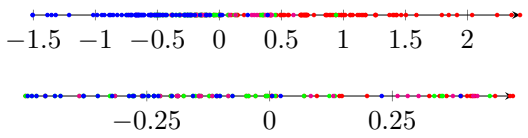
[4] E.g., bigrams will not detect the fact that few Yamato words have /e/ in the first syllable. (Martin, 1987, 48)

[5] Not all languages that experience borrowing will necessarily exhibit strata: arguably, only if the phonotactics of adapted forms of borrowings differ enough from those of native words.

Japanese words than Yamato words. In sum, unsupervised clustering with diverging phonotactic submodels aligns strongly with strata based on categorical constraint rankings.

Gradient membership in strata Hayes (2016) and Jennifer Smith (p.c.) both cite Itô and 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 gradiently into strata[6], we took random samples of 100 words each assigned to groups 1 (mostly Yamato) and 2 (mostly Sino-Japanese), with differences of perplexity$_2$ − perplexity$_1$ shown in the first plot, and the most marginal words (|diff| < 0.5) in the second plot. (● = Yamato, ● = Sino-Japanese, ● = foreign, ● = hybrid or ambiguous.



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* 鱏夜 'ray(fish)-night', *kui-ki* 杭木 'stake-tree' and *ki-matu* 木松 'tree-pine'.[7] On one hand, the abundance of morphemes with different Sino-Japanese and Yamato readings of the same kanji (e.g., *moku* and *ki* for 木 'tree'), discretely determines the stratum membership of a given reading by the pronunciation contrast: Sino-Japanese *moku* contrasts with Yamato *ki*. On the other hand, many readings of either type, Sino-Japanese or Yamato, not only satisfy all of Itô and Mester's strata-distinguishing constraints, but show only marginal differences in the phoneme perplexity assigned by each model, making their phoneme sequences ambiguous as to their stratum. In Japanese, one easily finds strata-straddling homophones like Sino-Japanese *atu* 圧, 'pressure' (as in *si-atu* 'finger-pressure, shiatsu') and Yamato *atu-i* 熱い 'hot'. The lack of a characteristically Yamato or Sino-Japanese shape makes them good candidates for gradient strata membership in a way analogous to English words that Hayes judges to be 'intermediate in Latinity.'

If we look at misclassified Yamato words in group 2 (blue dots right of 0) we find fewer marginal words. We do find *tooku* 遠く 'far' (adv.), (which is also homophonous with foreign borrowing 'talk'), and *atude* 厚手 '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* 戻し税 'tax refund' (lit. 'return(trans.)-tax', Yamato+Sino-Japanese) and *zyo-no-kuti* 序の口 'beginning' (lit. 'beginning-entrance', Sino-Japanese+Yamato) with margins of 0.084 and 0.130.

Summary Simple neural networks which can learn gradient phonotactic properties of words such as the probability of a given phoneme to occur after a given string are shown to be useful tools in capturing the ways in which gradient phonotactics separate words in a language into strata in both discrete and continuous ways. 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 is 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').

[6]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.

[7]The last one is not quite fictitious, having been coined as the actual name of a hotel in Hiroshima.

References

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179 – 211.

Bruce Hayes. 2016. Comparative phonotactics. In

Proceedings of the 50th meeting of the Chicago Linguistic Society, pages 265–285.

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

Ryan Hearn. 2016. Rethinking the Core-Periphery Model: Evidence from Japanese and English. In *Proceedings of the 24th Manchester Phonology Meeting*.

Brian Hsu and Karen Jesney. 2017. Loanword adaptation in Québec French: Evidence for weighted scalar constraints. In *Proceedings of the West Coast Conference on Formal Linguistics*, volume 34, pages 249 –258.

Junko Itô and Armin Mester. 1995a. The core-periphery structure of the lexicon and constraints on reranking. In *University of Massachusetts Occasional Papers in Linguistics*, volume 18.

Junko Itô and Armin Mester. 1995b. *The Handbook of Phonological Theory*, chapter Japanese Phonology. Blackwell.

Samuel E. Martin. 1987. *The Japanese Language Through Time*. Yale University Press.

Connor Mayer and Max Nelson. 2020. Phonotactic learning with neural language models. In *Proceedings of the Annual Meeting of the Association for Computation in Linguistics*, volume 3.

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

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

Max Nelson and Connor Mayer. 2020. Phonotactic language model.

NHK. 1999. *NHK Hatsuon Akusento Jiten (NHK Pronunciation and Accent Dictionary)*. NHK (Japanese Broadcasting Corporation).

Janet Pierrehumbert. 2015. *Oxford Handbook on the the History of Phonology*, chapter 70+ years of probabiliistic phonology. Oxford University Press.

Jennifer Smith. 2018. Stratified faithfulness in Harmonic Grammar and emergent core-periphery structure. *Hana-bana: A festschrift for Junko Itô and Armin Mester*, 13.

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