2024. *Proc Ling Soc Amer* 9(1). 5725. https://doi.org/10.3765/plsa.v9i1.5725.

A phonotactic-tonotactic grammar for Tokyo Japanese that clusters by lexical strata offers a good trade-off between model size and likelihood

Satoru Ozaki*

Abstract. The Japanese lexicon is typically classified into at least three etymological strata: native, Sino-Japanese and foreign words. In Tokyo Japanese, nouns from different strata are known to have different phonotactic as well as tonotactic properties. Should one analyze Tokyo Japanese nouns using a non-clustering grammar that generates all nouns using the same phonological grammar, or should one analyze them using a clustering grammar that generates nouns from different strata using different grammars? In this study, I address this question from a probabilistic and a model selection perspective: the better probabilistic grammar is one that better balances fit to data and the number of parameters in the grammar. Using the UCLA Phonotactic Learner, I train two kinds of MaxEnt grammars that correspond to non-clustering and clustering grammars. I compare the two kinds of grammar using the Bayesian Information Crierion (BIC), and show that the non-clustering grammars make a better trade-off between fit to data and model size than non-clustering grammars. Consequently, different etymological strata of the Tokyo Japanese nominal lexicon are better analyzed as being generated from different MaxEnt grammars than from the same MaxEnt grammar.

Keywords. phonology; phonotactics; tonotactics; lexical strata; model selection; MaxEnt; Japanese

1. Introduction. Japanese words are typically classified into at least three categories by their etymological sources: native Japanese words, loanwords from Old Chinese and loanwords from languages other than (Old) Chinese. That is, the Japanese lexicon consists of at least three etymological strata. The Tokyo Japanese pronunciations of nouns from different strata are known to have different phonotactic as well as tonotactic properties, i.e., accent distributions. It is possible to analyze all Tokyo Japanese nouns as generated by the same phonological grammar, which I call a *non-clustering grammar*, or analyze nouns from different strata as generated by different phonological grammars, which I call a *clustering grammar*. Should we choose a non-clustering grammar or a clustering grammar for our analysis of Tokyo Japanese nouns?

In this study, I address this question from a probabilistic and a model selection perspective: the better probabilistic grammar is one that better balances fit to data and the number of parameters in the grammar. Using the UCLA Phonotactic Learner, I train two kinds of MaxEnt grammars that correspond to non-clustering and clustering grammars. I compare the two kinds of grammar using the Bayesian Information Crierion (BIC), and show that the non-clustering grammars make a better trade-off between fit to data and model size than non-clustering grammars. Consequently, different etymological strata of the Tokyo Japanese nominal lexicon are better analyzed as being generated from different MaxEnt grammars than from the same MaxEnt grammar.

This paper is structured as follows. In Section 2, I provide background information for this study. In Section 3, I describe my study. In Section 4, I discuss the conclusions warranted by my results and address some limitations in my study. I conclude in Section 5.

* I would like to thank Gaja Jarosz, Michael Becker, my classmates for LING 603 in Fall 2022 and LING 606 in Spring 2023 at UMass Amherst, Shigeto Kawahara and the audience at the 2024 LSA Annual Meeting. Author: Satoru Ozaki, University of Massachusetts Amherst (sozaki@umass.edu).

2. Background. In this section, I discuss the phonotactic and tonotactic differences between Tokyo Japanese etymological strata. I also put forth the central research question in this paper. As the reader will see later, most work in the literature answers this question from a learnability perspective, but this study will address this question from a model selection perspective.

2.1. ETYMOLOGICAL STRATA. Typically, the Japanese lexicon is assumed to consist of at least three etymological strata: (a) native Japanese words, (b) Sino-Japanese words, i.e., loanwords from Old Chinese, and (c) *gairaigo* 'foreign words', i.e., loanwords from languages other than (Old) Chinese. In this paper, I often abbreviate Sino-Japanese words as Sino-Japanese words, and refer to the *gairaigo* stratum as foreign words. I provide some examples of Japanese nouns from each stratum in (1).

(1) a. Examples of native nouns
kami 'paper', *tobira* 'door', *madoromi* 'drowse'

b. Examples of Sino-Japanese nouns
sen 'thousand', *dempa* 'phone signal', *gengogaku* 'linguistics'

c. Examples of foreign nouns (*gairaigo*)
mainootaa 'Minotaur', *korubeeru* 'Colbert', *syaaman* 'Sherman'

In the Tokyo Japanese variety, nouns from different strata are known to have different phonotactic as well as tonotactic properties (Ito & Mester 1995a,b; Fukuzawa 1998; Ito & Mester 1999; Moreton & Amano 1999; Gelbart 2005; Gelbart & Kawahara 2007; Frellesvig 2010; Morita & O'Donnell 2022). In this study, I focus on the cross-stratal differences in syllable weight and accent. Heavy syllables are more common in Sino-Japanese and foreign words than native words (todo: cite). Native words are the least likely to be accented (29%), Sino-Japanese words are more likely (49%), while foreign words are the most likely to be accented (93%) (Kubozono 2006, 2011). There are also segmental differences across strata, which I will not focus on the study. For example, a voiceless obstruent does not immediately follow a nasal in native words (cf. Sino-Japanese *sintai* 'body' and foreign *ranku* 'rank'). Sounds such as [ɸa], [ɸi], [ɸe], [ɸo] and nongeminate [p] only occur in foreign words (e.g., *fairu* 'file', *finrando* 'Finland', *feruto* 'felt', *forumu* 'form', *pai* 'pie').

2.2. CLUSTERING VS. NON-CLUSTERING GRAMMARS. Given that the etymological strata are associated with different properties, a grammar for the Tokyo Japanese nominal lexicon can either (a) have just one grammar that models the distribution over all Tokyo Japanese nouns, or (b) consist of three separate grammars that each model the distribution over Tokyo Japanese nouns from one of the three strata. I refer to grammars of the first kind as *non-clustering grammars* and to those of the second kind as *clustering grammars*. The grammars that make up a clustering grammar are called *clusters*. The natural question to ask is:

(2) Which kind of grammars should be preferred, and why?

Most approaches to (2) in the literature come from a *learnability* perspective. One could view the process of acquiring the Tokyo Japanese nominal lexicon as an algorithm that learns a grammar for that lexicon. If grammars of one kind are more difficult to learn than grammars of the other kind, then that is a good reason to believe we are learning the easier kind of grammars.

While a learner of a non-clustering grammar just has to figure out the grammar for the entire lexicon, a learner of a clustering grammar has more information to figure out: (a) how many clusters there are in the first place, and (b) cluster assignment, i.e., which noun belongs to which cluster. There is much debate on whether these additional pieces of information are difficult to learn, which would make clustering grammars more difficult to learn than their non-clustering counterparts. For example, Rice (1997) and Ota (2004) consider a particular kind of learner, and argue that once this kind of learner hypothesizes that there is only one cluster to which all nouns belong, it will never be able to reject that hypothesis (Becker 2009). However, many methods have been proposed that would make learning a clustering grammar possible, including providing data in a specific order (Ito & Mester 1999) and semi-supervised learning (Shaw 2006), where the learner is given the correct cluster assignment for certain words. Morita & O'Donnell (2022) develop a fully unsupervised probabilistic learning approach, where they use n-gram models to model the grammar for each cluster.

The question in (2) can also be addressed from a *model selection* perspective. A clustering grammar gives a fit to the data that is at least as good as the fit given by a non-clustering grammar to the same data. This is because any non-clustering grammar can be converted into a clustering grammar with one cluster. The strength of clustering grammars comes with a cost: they use more statistical parameters. An ideal grammar should balance fit to data with the number of parameters. Quantitative criteria such as the Bayesian Information Criterion (BIC; Schwarz 1978) measures such trade-offs between likelihood and model size. In this study, I attempt to answer (2) from a model selection perspective. I will train non-clustering and clustering grammars in certain ways, compare them using the BIC, and show that the clustering grammars offer a better trade-off between likelihood and model size.

3. Study. In this section, I describe my study. I start by describing the corpora I source my data from and the preprocessing and transformation I apply to the data. Next, I provide some formal details on MaxEnt grammars, and their non-clustering and clustering extensions that I use in my study. Then, I describe how I use the UCLA Phonotactic Learner to train the MaxEnt grammars. Finally, I present my learning results and my model selection results.

3.1. DATA. My data comes from two corpora: (a) the Balanced Corpus of Contemporary Written Japanese (BCCWJ; Maekawa et al. 2013) and (b) the New Dictionary of Japanese Pronunciation and Accentuation published by the Japan Broadcasting Corporation (NHK), also known as the NHK Accent Dictionary. BCCWJ is a balanced corpus with 100 million words, covering texts from a wide range of registers including books, newspapers, governmental white papers, internet texts, legal texts as well as poetry. The text is segmented into words, and each word is annotated with information such as part-of-speech and etymological stratum. The NHK Accent Dictionary is a dictionary with 75 thousand words. For each word, it indicates whether an accent is present on that word and its position if present.

I also provide each word with a phonological representation. These representations abstract away the segmental content of words and instead represent each word as a sequence of mora types. I classify moras into five types: V, Q, N, R and J; this classification is explained in detail in (3) with examples. The representations also specify whether each mora is accented; an accented mora is coded with an uppercase letter, and an unaccented mora with a lowercase letter.[1] It is

[1] Note on notation: each mora is prefixed with $_{\mu}$. Uppercase denotes an accented mora, and lowercase denotes an

popular practice in the Japanese phonology literature to classify the second mora in bimoraic syllables into the four types Q, N, R and J (Vance 2008, among others); my classification system is based on this four-way classification system.

(3) a. V: a (C)V structure
E.g./$_\mu$**A**$_\mu$ki/, /$_\mu$**ta**$_\mu$KI/ → **V**v, **v**V

b. R: second half of a long vowel
E.g. /$_\mu$SE$_\mu$n$_\mu$ta$_\mu$**a**/, /$_\mu$to$_\mu$**o**$_\mu$kyo$_\mu$**o**/ → Vnv**r**, v**r**v**r**

c. J: second half of a diphthong
E.g. /$_\mu$ga$_\mu$**i**$_\mu$ko$_\mu$ku/, /$_\mu$KO$_\mu$**i**/ → v**j**vv, V**j**

d. N: moraic nasal
E.g. /$_\mu$a$_\mu$**m**$_\mu$pa$_\mu$**n**/ → v**n**v**n**

e. Q: first half of a geminate consonant
E.g. /$_\mu$na$_\mu$**t**$_\mu$TO$_\mu$o/, /$_\mu$ma$_\mu$**p**$_\mu$pu/ → v**q**Vr, v**q**v

For each word, I extract its stratum information from BCCWJ, and its accent information from the NHK Accent Dictionary. This gives me a list of 35,832 word entries, where each entry contains three pieces of information: (a) the word, (b) its phonological representation, and (c) its etymological stratum (native, Sino-Japanese or foreign). Some sample entries from the list are provided in Table 1.

Word	Representation	Stratum
koto	vV	Native
zyuu	Vr	Sino-Japanese
sentaa	Vnvr	foreign

Table 1. Sample entries from the list of word entries

I then calculate the type frequency for each attested phonological representation. The result is a list of 716 unique phonological representations, annotated with their type frequencies; this represents the Tokyo Japanese nominal lexicon. I call this data D. I also create three sublexicons D_N, D_S, D_G which represent the native, Sino-Japanese and foreign nouns respectively. The sizes of the native, Sino-Japanese and foreign sublexicons are 241, 191 and 541 respectively.

The UCLA Phonotactic Learner requires each phonological representation to be a sequence of feature-value matrices, where all features have binary values. Since I use the UCLA Phonotactic Learner, my representations must meet this requirement. I represent each mora with a matrix. Each mora type m is represented with a binary feature $[m]$; each mora is valued $[+m]$ iff a mora is of type m, and $[-m]$ otherwise. I also represent accentedness with a binary feature [acc]. In (4), I provide an example of a feature-value matrix representation of a hypothetical phonological representation /Vqvnvrvj/.

unaccented mora. For example, /$_\mu$A$_\mu$ki/ denotes the two-mora word *aki*; the first mora is an accented /a/, and the second mora is an unaccented /ki/. The nouns on the left hand side of each arrow are coded as the phonological representations on the right hand side of that arrow.

(4) Feature-value matrix representation of the phonological representation /Vqvnvrvj/

$$\begin{bmatrix} +\text{v} \\ -\text{q} \\ -\text{n} \\ -\text{r} \\ -\text{j} \\ +\text{acc} \end{bmatrix} \begin{bmatrix} -\text{v} \\ +\text{q} \\ -\text{n} \\ -\text{r} \\ -\text{j} \\ -\text{acc} \end{bmatrix} \begin{bmatrix} +\text{v} \\ -\text{q} \\ -\text{n} \\ -\text{r} \\ -\text{j} \\ -\text{acc} \end{bmatrix} \begin{bmatrix} -\text{v} \\ -\text{q} \\ +\text{n} \\ -\text{r} \\ -\text{j} \\ -\text{acc} \end{bmatrix} \begin{bmatrix} +\text{v} \\ -\text{q} \\ -\text{n} \\ -\text{r} \\ -\text{j} \\ -\text{acc} \end{bmatrix} \begin{bmatrix} -\text{v} \\ -\text{q} \\ -\text{n} \\ +\text{r} \\ -\text{j} \\ -\text{acc} \end{bmatrix} \begin{bmatrix} +\text{v} \\ -\text{q} \\ -\text{n} \\ -\text{r} \\ -\text{j} \\ -\text{acc} \end{bmatrix} \begin{bmatrix} -\text{v} \\ -\text{q} \\ -\text{n} \\ -\text{r} \\ +\text{j} \\ -\text{acc} \end{bmatrix}$$

3.2. GRAMMAR. I use MaxEnt grammars in this study. Formally, a *MaxEnt grammar* G is a 3-tuple $\langle \Omega, C, \mathbf{w} \rangle$. Ω is the space of phonological representations. $C = \langle c_1, \cdots, c_n \rangle$ is a vector of *constraints*, where each constraint $c_k : \Omega \to \mathbb{N}$ is a function that assigns a non-negative integer known as *violation count* to each phonological representation in Ω. $\mathbf{w}$ is a n-dimensional real vector, called the *weights*. The kth weight w_k describes how undesirable each violation of the kth constraint c_k is; the higher the value of w_k, the lower the probability of a phonological representation that violates c_k.

G defines a probability distribution over Ω. The probability $p_G(x)$ of a phonological representation $x \in \Omega$ is given by:

$$p_G(x) = \frac{\exp(-\mathbf{w}^\top C(x))}{\sum_{x' \in \Omega} \exp(-\mathbf{w}^\top C(x'))} \tag{5}$$

where $C(x) = \langle c_1(x), \cdots, c_n(x) \rangle$. In this study, Ω is the set of theoretically possible Tokyo Japanese nouns, which can be indefinitely long. This makes Ω an infinite set. However, I restrict Ω to contain sequences no longer than the longest sequence in the training data D, following (Hayes & Wilson 2008).

My lexicon D (or any sublexicon) is a list of pairs $\langle \langle x_1, f_1 \rangle, \cdots, \langle x_{|D|}, f_{|D|} \rangle \rangle$, where each pair $\langle x_i, f_i \rangle$ consists of the phonological representation x_i and the type frequency f_i. Given D, the likelihood $p_G(D)$ of D given by the grammar G is simply:

$$p_G(D) = \prod_{i=1}^{|D|} p_G(x_i)^{f_i} \tag{6}$$

A non-clustering grammar uses the same MaxEnt grammar to assign likelihood to all nouns. Formally, a non-clustering grammar G is just a MaxEnt grammar. The likelihood of data under G is simply $p_G(D)$ as defined in (6).

A clustering grammar with K clusters consists of K MaxEnt grammars, and it assigns the likelihood to each noun using the MaxEnt grammar for the cluster this nouns belongs to. Formally, a clustering grammar G with K clusters is a K-tuple $G_1, \cdots, G_K$, where each G_k is a MaxEnt grammar, and all G_k's share a common Ω. Given a lexicon D, a *cluster assignment* is a function $g : [|D|] \to [K]$ such that for each $i \in [|D|]$, $g(i)$ indicates the cluster that the ith example in D is assigned to. The likelihood $p_G(D \mid g)$ of D given by the grammar G with the cluster assignment g is:

(7) $$p_G(D \mid g) = \prod_{i=1}^{|D|} p_{G_{g(i)}}(x_i)^{f_i}$$

As discussed in Section 2.2, learning a clustering grammar is more work than learning a non-clustering grammar, since the learner needs to decide on the number of clusters K as well the cluster assignment g. It is very challenging to design an algorithm that learns K and g under a MaxEnt framework without any supervision. Furthermore, a learner of a clustering grammar can end up with a grammar that doesn't give a good fit to the data if it learns a suboptimal combination of K and g. Thus, it is desirable to prevent a suboptimal choice of K and/or g, since this might affect my comparison between the non-clustering and clustering grammars I obtain from my learning trials. I do so by simply providing the clustering grammars with the correct values of K and g. Specifically, in this study, I only consider clustering grammars with three clusters, where the first, second and third grammars are respectively used to assign likelihoods to native, Sino-Japanese and foreign words. The likelihood of the data given by this kind of grammar G would be:

(8) $$p_G(D) = p_{G_1}(D_N)p_{G_2}(D_S)p_{G_3}(D_F)$$

3.3. LEARNING AND EVALUATION. I use the UCLA Phonotactic Learner (Hayes & Wilson 2008). For each learning trial, I learn d constraints, where d is a hyperparameter. I experiment with three values of d: 50, 75 and 100.

To learn a non-clustering grammar, I learn one set of d constraints over D, the entire Tokyo Japanese nominal lexicon. To learn a clustering grammar, I learn one set of d constraints over each sublexicon (D_N, D_S, D_F). As a result, for each value of d, the clustering grammars have three times the constraints of the non-clustering grammars. I perform five trials for each setup.

The Bayesian Information Criterion (BIC; Schwarz (1978)) is a number that scores the trade-off a model makes between fit to data and model size. For a MaxEnt grammar G with n constraints, the BIC value BIC(G) is given as in (9):

(9) $$\text{BIC}(G) = n \log |D| - 2 \log p_G(D)$$

One can see from the formula in (9) that a high likelihood and a low number of constraints will cause a grammar to have a low BIC value. A grammar with a lower BIC value makes a better trade-off between fit to data and grammar size.

3.4. RESULTS. I report my results in Table 2.

As d, i.e., the number of constraints increases, the log-likelihood increases (which is an improvement) and the BIC value decreases (which is also an improvement). This is expected. For each value of d, the clustering grammar has a higher log-likelihood than the non-clustering grammar. This is also expected. The crucial result is that the clustering grammar has a lower BIC value than the non-clustering grammar for each d. This suggests that, given my setup, my clustering grammars always offer a better trade-off between likelihood and model size. Consequently, it is advantageous to analyze Tokyo Japanese nouns from different etymological strata as generated by distinct MaxEnt grammars, rather than analyzing them as all generated by the same MaxEnt grammar.

Grammar			$d = 50$	$d = 75$	$d = 100$
Non-clustering	$\log p$	Avg.	-401,156	-364,282	-309,266
		Std.	2,454	7,398	18,491
	BIC	Avg.	802,841	729,356	619,589
		Std.	4,908	14,795	36,982
Clustering	$\log p$	Avg.	-327,047	-309,081	-288,158
		Std.	2,087	7,237	14,354
	BIC	Avg.	655,679	620,540	579,486
		Std.	4,174	14,473	28,707

Table 2. Results. $\log p$ stands for log-likelihood. Averages and standard deviations are obtained over five runs.

4. Discussion. In this section, I discuss my results, point out limitations of my current study and suggest how to address them in future work.

4.1. ON GRAMMARS. The clustering grammars in this study were given the correct number of clusters K and cluster assignments g, and they have a lower BIC value than the non-clustering grammars. However, it is unknown whether clustering grammars learned in a fully unsupervised setting, where K and g are not given, also have a lower BIC value than the non-clustering grammars. I can examine existing learners of clustering grammars to see if this is indeed the case.

It is also debatable to what extent the given K and g are "correct". There is independent reason, namely etymology, to assume that $K = 3$ and g maps nouns to their etymological strata, but it is in principle possible for a different combination of K and g to give a better fit to the data than our clustering grammars (Shih & Inkelas 2016; Shih 2018).[2] To investigate this possibility empirically, I need a learning algorithm that learns K and g for clustering MaxEnt grammars.

The clusters in the clustering grammars are allowed to vary in both their constraints and weights. One can imagine a different kind of clustering grammars where the clusters share some but not all parts of the grammar; such sharing reduces the size of the grammar. For example, the clusters in a clustering grammar could share the same set of constraints, but be allowed to have different weights.[3] Or, one could constrain the clusters to share the same constraints that target phonotactics (e.g. *[+V][+N]), but allow the clusters to have different constraints that either only target accent (e.g. *#[+acc]) or link accent and phonotactics (e.g. *[+Q,+acc]). Perhaps these kinds of grammars would offer an even better trade-off between likelihood and model size.

4.2. ON REPRESENTATIONS. The phonological representations I use in this study contain little to no segmental information. However, nouns from different etymological strata are known to have different segmental distributions, as detailed in Section 2. It is possible for the BIC value to favor the non-clustering grammar once I enrich the representations with segmental information.

The BIC formula I use in this study penalizes a model with many constraints, but it doesn't penalize a model that uses very rich representations. A more holistic approach to model selection might take into account richness of the representations as well as the number of constraints.[4]

[2] Thanks to Mits Ota for this comment.

[3] Thanks to Jennifer Kuo for this comment.

[4] Thanks to Scott Nelson for this comment.

5. Conclusion. In this study, I train two kinds of MaxEnt grammars on the Tokyo Japanese nominal lexicon: non-clustering grammars that generate each noun in the lexicon using the same MaxEnt grammar, and clustering grammars that generate nouns from different etymological strata using different MaxEnt grammars. The clustering grammars are given a number of clusters that corresponds to the number of etymological strata, as well as a cluster assignment that maps each noun to the cluster corresponding to its etymological stratum. The clustering grammars have a lower BIC value than the non-clustering grammars, suggesting that clustering grammars make a better trade-off between likelihood and model size than the non-clustering grammars. The consequence of this is that different etymological strata of the Tokyo Japanese nominal lexicon are better analyzed as being generated from different MaxEnt grammars than from the same MaxEnt grammar.

References

Becker, Michael. 2009. *Phonological trends in the lexicon: The role of constraints*. Amherst: UMass dissertation.

Frellesvig, Bjarke. 2010. *A history of the Japanese language*. Cambridge: Cambridge University Press.

Fukuzawa, Haruka. 1998. Multiple input-output faithfulness relations in Japanese. Rutgers Optimality Archive ROA-260-0698.

Gelbart, Ben. 2005. *Perception of foreignness*. Amherst: UMass dissertation.

Gelbart, Ben & Shigeto Kawahara. 2007. Lexical cues to foreignness in Japanese. In Yoichi Miyamoto & Masao Ochi (eds.), *MIT working papers in linguistics 55*, 49–60.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39. 379–440. https://doi.org/10.1162/ling.2008.39.3.379.

Ito, Junko & Armin Mester. 1995a. The core-periphery structure of the lexicon and constraints on reranking. In Jill Beckman, Suzanne Urbanczyk & Laura Walsh Dickey (eds.), *UMass occasional papers in linguistics 32*, 180–210. Amherst: UMass Graduate Linguistics Student Association.

Ito, Junko & Armin Mester. 1995b. Japanese phonology. In John Goldsmith (ed.), *A handbook of phonological theory*, 817–838. Oxford: Blackwell.

Ito, Junko & Armin Mester. 1999. The phonological lexicon. In Natsuko Tsujimura (ed.), *The handbook of Japanese linguistics*, 62–100. Oxford: Blackwell.

Kubozono, Haruo. 2006. Where does loanword prosody come from? A case study of Japanese loanword accent. *Lingua* 116(7). 1140–1170. https://doi.org/10.1016/j.lingua.2005.06.010.

Kubozono, Haruo. 2011. Japanese pitch accent. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.), *The Blackwell companion to phonology* (5th edn.), 2879–2907. Oxford: Blackwell.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka & Yasuharu Den. 2013. Balanced corpus of contemporary written Japanese. In *Proceedings of LREC 48*, 345–371.

Moreton, Elliott & Shigeaki Amano. 1999. Phonotactics in the perception of Japanese vowel length: Evidence for long-distance dependencies. *Proceedings of the European Conference on Speech Communication and Technology* 5. 2679–2682.

Morita, Takashi & Timothy J. O'Donnell. 2022. Statistical evidence for learnable lexical subclasses in Japanese. *Linguistic Inquiry* 53(1). 87–120. https://doi.org/mtx2.

Ota, Mitsuhiko. 2004. The learnability of the stratified phonological lexicon. *Journal of Japanese Linguistics* 20(1). 19–40. https://doi.org/10.1515/jjl-2004-0104.

Rice, Keren. 1997. Japanese NC clusters and the redundancy of postnasal voicing. *Linguistic Inquiry* 28(3). 541–551. https://www.jstor.org/stable/4178991.

Schwarz, Gideon. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6. 461–464.

Shaw, Jason. 2006. Learning stratified lexicon. In Christopher Davis, Amy Rose Deal & Youri Zabbal (eds.), *Proceedings of NELS 36*, vol. 2, 519–530.

Shih, Stephanie S. 2018. Learning lexical classes from variable phonology. *Proceedings of the Asian Junior Linguists Conference* 2. 1–15.

Shih, Stephanie S. & Sharon Inkelas. 2016. Morphologically-conditioned tonotactics in multi-level Maximum Entropy grammar. In Gunnar Hansson, Ashley Farris-Trimble, Kevin McMullin & Douglas Pulleyblank (eds.), *Proceedings of the 2015 Annual Meeting on Phonology*. Washington, DC: Linguistic Society of America. https://doi.org/mtx3.

Vance, Timothy J. 2008. *The sounds of Japanese*. Cambridge: Cambridge University Press.