Statistical Evidence for Learnable Lexical Subclasses in Japanese

Takashi Morita
Timothy J. O’Donnell

It has been proposed that the Japanese lexicon can be divided into etymologically defined sublexica on phonotactic and other grounds. However, the psychological reality of this sublexical analysis has been challenged by some authors, who have appealed to putative problems with the learnability of the system. In this study, we apply a commonly used clustering method to Japanese words and show that there is robust statistical evidence for the sublexica and, thereby, that such sublexica are learnable. The model is able to recover phonotactic properties of sublexica previously discussed in the literature, and also reveals hitherto unnoticed phonotactic properties that are characteristic of sublexical membership and can serve as a basis for future experimental investigations. The proposed approach is general and based purely on phonotactic information and thus can be applied to other languages.

Keywords: Japanese, phonotactics, lexical strata, Bayesian learning, naturalistic learnability, variational inference

1 Introduction

Languages frequently borrow words from one another. Since the donor and recipient languages typically differ in their phonological properties, native words and loanwords often exhibit systematically differing phonological patterns. A classical example of such a mixed system is the lexicon of Japanese, which contains large numbers of Sino-Japanese (SJ) loanwords from old Chinese, as well as many other Foreign loanwords, primarily from English. Researchers have observed that Native, SJ, and Foreign words exhibit markedly different phonotactics (Ito and Mester 1995a,b, 1999, 2003b, 2008, Fukazawa 1998, Fukazawa, Kitahara, and Ota 1998). Table 1 lists some phonotactic patterns discussed in previous literature whose presence or absence distinguishes these three sublexica. For example, voiceless obstruents after nasals are prohibited (i.e., *NC) in the Native sublexicon whereas they are allowed (i.e., NČ) in loanwords (SJ and Foreign; we

We gratefully acknowledge helpful feedback and discussion from Roger Levy, Adam Albright, Michael Kenstowicz, and Edward Flemming. We also wish to express our appreciation to the participants in MIT CompLang and Phonology in the NorthEast 2018, and also to Kate McCurdy for her help improving the presentation of this article. We gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Fonds de Recherche du Québec, Société et Culture (FRQSC), the Canada CIFAR AI Chairs program, MEXT Grant-in-aid for Scientific Research on Innovative Areas #4903 (Evolinguistic; JP17H06380), and the JST Core Research for Evolutional Science and Technology 17941861 (#JPMJCR17A4).
review phonotactic properties in more detail in section 2.1). Such observations have led researchers to incorporate references to the sublexicon membership of words in their analyses of Japanese phonology.

However, a sublexical analysis of Japanese phonotactics raises a potential learnability problem: the sublexicon membership of words is not directly observable, and some researchers have argued that the proposed phonotactic constraints do not provide sufficient cues for learning to take place (Rice 1997, Ota 2004). In this study, we apply a commonly used clustering method to Japanese words extracted from a corpus and show that (i) the data include abundant and reliable statistical evidence indicating the sublexicon membership of individual words; thus, (ii) the sublexica are learnable, (iii) the learned sublexica recover phonotactic properties previously observed in the literature, and (iv) the learned sublexica also reveal hitherto unnoticed phonotactic properties that can be studied in future empirical work. While the current study focuses on the Japanese data, our learning method is general and can be applied crosslinguistically to similar sublexicon-learning problems (e.g., the Germanic vs. Latinate differences in English; Grimshaw 1985, Anshen et al. 1986, Grimshaw and Prince 1986, Fabb 1988, Gropen et al. 1989, Levin 1993, O’Donnell 2015).

The rest of this article is organized as follows. In section 2, we review previously reported properties of the Japanese sublexica in more detail. We also review previous discussions about learnability of the sublexica. In section 3, we introduce our clustering method, and in section 4, we apply it to naturalistic Japanese data. Section 5 concludes.

2 Background

In this section, we first describe the properties of Japanese sublexica previously discussed in the literature (section 2.1). We then review an earlier experimental study that suggests that sublexica are psychologically real for speakers of Japanese (section 2.2). After this, we review earlier arguments about how sublexica might (or might not) be learnable on the basis of phonotactic patterns (section 2.3). Finally, we give an overview of the approach we adopt to the computational modeling of language learning (sections 2.4–2.5) and phonotactics (section 2.6).
2.1 Linguistic Differences among Japanese Sublexica

As introduced in section 1, the Japanese sublexica exhibit different phonotactic constraints: for example, the Native sublexicon obeys the \(^{*}NC\) constraint whereas the SJ and Foreign sublexica violate it (i.e., \(<^{*}NC\)). Table 1 also shows that the following three types of substrings occur only in the Foreign sublexicon:

- [fi] before [a], [i], [e], [o] (non-[u] vowels): fi (Ito and Mester 1995b, Moreton and Amano 1999)

In addition to these categorical contrasts, some statistical differences among sublexica have been observed. Moreton, Amano, and Kondo (1998) performed a more comprehensive study on inter-sublexical differences in the probability of a phonetic segment appearing in a word as well as the consonant-to-vowel transitional probabilities. Moreton and Amano (1999) point out that [rj] and [çj] are significantly more frequent in SJ than in the other sublexica. They also observe that the long low vowel [ar] appears more frequently in the Foreign sublexicon “irrespective of position” in words. Gelbart (2005) and Gelbart and Kawahara (2007) make a slightly different generalization about [ar], arguing that the elevated rate of the segment is limited to word-final positions. Our analysis will provide support for Moreton and Amano’s (1999) generalization.

The phonotactic constraints listed in table 1 refer only to substrings of segments or classes of segments. There are also sublexicon-specific constraints that are sensitive to morphological structure. In the Native and SJ sublexica, a single morpheme can contain no more than one voiced obstruent (Lyman’s Law; Suzuki 1998, Ito and Mester 2003a, 2008). SJ morphemes have the further restriction that they are at most one syllable long (Fukazawa 1998, Ota 2004). SJ morphemes have the further restriction that they are at most one syllable long (Fukazawa 1998, Ota 2004).\(^1\) There is also a morphophonological alternation called rendaku or sequential voicing that is sensitive to sublexical affiliation. Rendaku turns word-initial voiceless obstruents into voiced obstruents when a word is used as the second half of a compound (e.g., Vance 2016): for example, [jama] ‘mountain’ + [tera] ‘temple’ \(\rightarrow\) [jamadera] ‘temple on a hillside’. This alternation applies primarily to Native words (see Kawahara and Zamma 2016), although there are exceptional SJ and Foreign words that exhibit rendaku: for example, [jii] ‘hot water’ [tô:phiû] ‘tofu (SJ)’ \(\rightarrow\) [judo:phiû] ‘boiled tofu’ and [ame] ‘rain’ + [kap:]] ‘raincoat (Foreign)’ \(\rightarrow\) [amagap:] ‘raincoat’. An experimental study by Suzuki, Maye, and Ohno (2000), reviewed in Kawahara 2016, also calls into question this sublexicon-based generalization.

\(^1\) The constraint on the length of SJ morphemes applies to underlying representations rather than surface forms, which can be disyllabic owing to epenthesis. There is also a slightly different generalization that SJ morphemes are maximally bimoraic (Kurisu 2000).
There are also prosodic differences among the sublexica, including differences in pitch-accent patterns. Japanese words have zero or one pitch-accented syllables. Following a pitch-accented syllable, there is a steep fall in F0 (the lack of a pitch-accented syllable means there will be no fall in F0 within a word, except for gradual falls owing to other factors such as phonetic interpolation and declination; Haraguchi 1975, 1999, Pierrehumbert and Beckman 1988). Focusing on trimoraic words, Kubozono (2006, 2011) reports that Native words tend to lack accented syllables (29% are accented) whereas Foreign words are likely to have one (93% are accented). SJ words seem unbiased: about half (49%) are accented and the others are unaccented.

Finally, the majority of Native, SJ, and Foreign words are written with different sets of characters in Japanese. Some researchers have proposed that these orthographic differences are essential for sublexicon learning (Gelbart and Kawahara 2007).

The properties of the sublexica mentioned above (except orthography) are summarized in table 2. The diagnoses headed by the white bullet point ○ refer to morphological or prosodic structure, which will not be analyzed here.

2.2 Psychological Reality of Japanese Sublexica

There is also empirical evidence that native speakers of Japanese are aware of sublexical generalizations. Moreton and Amano (1999) showed that segments that are representative of particular sublexica cause systematic biases in the perception of other segments.

Recall that the long vowel [aː] is representative of the Foreign sublexicon in Japanese (ْا). Moreton and Amano (1999) studied the perceptual threshold between this vowel and the corresponding short vowel [a], which occurs in all three sublexica. In their experiment, they considered

### Table 2

Summary of phonotactic/(morpho)phonological properties of Japanese sublexica. The properties marked by the black bullet points are purely segmental phonotactics; those marked by the white bullet points require either morphological or prosodic analysis.

<table>
<thead>
<tr>
<th>Sublexicon</th>
<th>Diagnoses</th>
</tr>
</thead>
</table>
| Native     | • No NC clusters (*NC)  
○ Rendaku  
○ High frequency of unaccented words |
| SJ         | • High frequency of [ɾj] and [ɕj] (´ɾj, ɕj)  
○ No more than one syllable per morpheme  
○ Even frequencies of accented and unaccented words |
| Foreign    | • Exclusive occurrences of nongeminate [p] (´p)  
• Exclusive occurrences of voiced geminates (´C[+long, voi])  
• Exclusive occurrences of [ɸ] before [a], [ɪ], [ɛ], and [o] (´ɸ + NON-ε)  
• High frequency of [aː] (´aː)  
○ Exclusive occurrences of more than one voiced obstruent per morpheme  
(i.e., no Lyman’s Law)  
○ High frequency of accented words |
nonce words of the form $C_1oC_2V_2$, where $V_2$ could be either [a] or [aː], and systematically varied the values of $C_1$ and $C_2$ to provide cues for and against the identification of the forms as Foreign sublexicon words:

- $C_1=[p]$ and/or $C_2=[\phi]$: representative of Foreign words
- $C_1=[rj]$ and/or $C_2=[\epsilon j]$: representative of SJ words
- $C_1=[r]$ and $C_2=[t]$: possible in both Foreign and SJ words

Moreton and Amano found that in Foreign-cued contexts (i.e., after [p] and [ϕ]), vowels were more readily perceived as the long vowel [aː] even when their duration was relatively short, while in SJ-cued contexts (after [rj] and [εj]) longer durations were required to perceive the vowel as [aː]. Finally, in neutral contexts (after [r] and [t]), the duration threshold required to perceive a vowel as long [aː] was intermediate between the two other cued contexts. One explanation for this pattern of results is that the presence of [p] and [ϕ] makes a word more likely to be perceived as Foreign, which in turn makes [aː] more likely, thus reducing the duration required to perceive a vowel with this quality as long. Similarly, [rj] and [εj] make the nonce form less likely to be Foreign (and more likely to be SJ), which makes [aː] less likely, thus increasing the duration required to perceive this long vowel. Finally, in the unbiased condition an intermediate duration is required to perceive vowel length. This pattern of results, to which we will return in section 4.3.3, is illustrated in figure 1.

**Figure 1**
The duration threshold between [a] and [aː] (Moreton and Amano 1999). Stimulus carriers were in the form of $C_1oC_2a(t)$. 
2.3 Previous Studies on Sublexicon Learning

Since the sublexical affiliations of individual words are not directly observable, they must be inferred from other, observable information in the data. Accordingly, an important question is what allows such learning to take place, in terms of both the data and the biases that learners bring to the table. Linguists have primarily examined the phonological evidence that learners might use in acquiring the Japanese system: the presence or absence of certain segments and segment combinations, as well as patterns of prosody.\(^2\) In this section, we review previous studies of sublexicon learning from phonological evidence and point out some of their limitations.

Rice (1997) argues that Japanese sublexica are not learnable because they stand in subset-superset relations. For example—at least according to the patterns shown in table 1—all possible Native words are also possible Foreign words, so Native $\subset$ Foreign; similarly, SJ $\subset$ Foreign. Thus, the existence of the Native and SJ sublexica is obscured when words from all the sublexica are presented together because Native $\cup$ SJ $\cup$ Foreign = Foreign. In other words, a learner would face no contradiction in the data even if they hypothesized a single lexicon equivalent to Foreign. Ota (2004) draws a similar conclusion in the framework of Optimality Theory (Prince and Smolensky 1993).

Rice’s (1997) and Ota’s (2004) observations center on learning algorithms that require a hypothesis to be logically inconsistent with the data in order to be rejected—\textit{strict-inconsistency learners} (SILs). As these authors note, such algorithms face a critical problem with nested patterns such as those of the Japanese systems. If the patterns in table 1 exhaust the phonological differences among Japanese sublexica, then once an SIL selects the single-lexicon hypothesis, it can never be rejected again (e.g., Becker 2009). Moreover, if an SIL has a general bias in favor of having a single sublexicon, such a single-lexicon solution will be globally optimal as well. However, the strict-inconsistency requirement is a very strong one that is not representative of most current work in artificial intelligence, machine learning, and other areas that study mathematical models of learning. If, instead, a learning model accepts or rejects hypotheses on the basis of some less strict goodness-of-fit criterion, the nested-lexicon problem can be avoided. It can also be avoided if there are nonnested patterns, beyond those in table 1, that distinguish the sublexica. In sections 3 and 4, we introduce a model that goes beyond strict inconsistency and, using this model, show that there are in fact many hitherto unnoticed patterns representative of each of the Japanese sublexica.

Several other researchers have also attempted to provide models of learning that can overcome the nested-sublexicon problem. Ito and Mester (1999) propose that an online learner can identify sublexicon-specific phonotactic constraints if the data are presented in a specific order. For example, the learner might identify the constraint against NC_{\text{underring}} clusters if no initial data included them. Later, when data that violated this constraint were presented, the learner would incorporate these data into a new sublexicon. However, it seems unlikely that Japanese-learning children receive

\(^2\) As mentioned in section 2.1, another possible source of information for adult speakers is orthography (Gelbart and Kawahara 2007).
data in the order the algorithm requires. For example, Ota (1999, 2003, 2004) reports that actual learners of Japanese are exposed to a fair number of NC/underring clusters in the early stages of acquisition.

Another approach is presented by Pater (2005), who proposes an algorithm that initially treats violations of phonotactic constraints as exceptions rather than as evidence against the existence of the constraints. Each word violating some phonotactic constraint constitutes a singleton exception cluster. Over time, these clusters are merged together to produce sublexica. However, this study used only nine hypothetical words in an idealized demonstration of the algorithm; it remains unclear whether it will scale to accommodate data more similar to those observed by children, which are more plentiful and more varied in phonological structure. For example, the actual Japanese lexicon includes many SJ and Foreign words that do not violate any of the Native-specific constraints proposed in the previous literature (e.g., *NC/underring). Such words would be indistinguishable from Native words under Pater’s algorithm unless additional constraints—targeting only the misclassified SJ/Foreign words—were included in the analysis. The algorithm may also overclassify words when more constraints are taken into account, failing to merge the constraint-based clusters into the appropriate classes.

Ito, Mester, and Padgett (2001) suggest that learners do not need to rely on phonotactic information alone, but also make use of morphophonological alternations such as rendaku (section 2.1) to identify sublexica. For instance, the *NC constraint causes a voicing alternation when concatenation of two morphemes would otherwise create a NC substring: for example, /sin + -ta/ $\mapsto$ [cin\textipa{\textnta}] ‘die (Native) + PAST’ (where [n] represents the coda nasal, following Vance (2008); see section 4.1 for more details on our transcriptions); /siN + tai/ $\mapsto$ [cin\textipa{\textntai}] ‘body (SJ) + body (SJ) $\mapsto$ body’. Observing that this alternation applies only to a limited class of words, learners would identify the existence of the word class. Such alternations can certainly help learners identify sublexica. However, since most alternations are not applicable to most words, we propose that alternations are likely to play a less important role than overall phonotactics. To again take the *NC voicing constraint as an example: most Native morphemes neither end with a nasal nor start with a voiceless obstruent, and so the concatenation of most pairs of Native morphemes will not trigger this alternation. The rare times that alternations are triggered may play an important but supportive role to phonotactic-based learning (see section 4.4 for discussion about the general effectiveness of synergistic learning from multiple correlated sources of evidence).

The learnability studies introduced above investigated fully unsupervised learning of Japanese sublexica, and the current study shares this objective. In unsupervised learning, the sublexical affiliations of individual lexical items are not provided in any way to the learning model. By contrast, supervised learning algorithms are provided with the target affiliations for each word. Shaw (2006) proposes one such algorithm in the framework of analogical modeling (Skousen 1989, 1992, 2002). Suppose that words are classified according to one of the unsupervised algorithms above. As noted, unmarked SJ and Foreign words without constraint violations would be misclassified into the Native sublexicon. Shaw’s proposal concerns how such misclassified words could be salvaged on the basis of their similarity to correctly classified SJ/Foreign words. Shaw demonstrated this learning procedure by training the model with 38 existing Japanese words (18
Native and 20 SJ) and then correctly predicting the class of 5 test words (2 Native and 3 SJ). Again, this study was limited in terms of its scale, and the choice of the training and test data was idealized: items were chosen that had several segments in common in the same positions. Perhaps most importantly, the approach is supervised in that the algorithm was provided training items with their correct classifications.

In summary, earlier work has argued that because the phonotactic patterns that characterize Japanese sublexica learning are nested, the learning problem is ill-posed. There have been various attempts to deal with the nested-sublexicon problem that have generally not scaled to the full complexity of the Japanese lexicon. As we have noted, however, the nested-sublexicon problem only arises for models that make use of strict logical consistency with the data to accept/reject hypotheses. It also only arises if the phonological cues distinguishing sublexica lead to nesting as in table 1. Below, we introduce a model that does not make the strict-inconsistency assumption and, using it, show that in fact a much greater variety of phonotactic patterns characterize Japanese sublexica than have been identified in previous discussions.

2.4 Sublexicon Learning as a Trade-Off

At a general level, models of learning often involve the specification of a principle of inductive inference that measures the goodness of particular hypotheses relative to some input data. In linguistics, these principles are often encoded by an evaluation metric measuring the quality of an inferred grammar. Evaluation metrics naturally consider two different kinds of information. One kind of information, which we will call prior information, measures the quality of a grammatical theory independent of the data itself, encoding preferences for simple, elegant, or parsimonious theories. Another kind of information, which we will call the likelihood, measures the fit to the data of individual hypotheses. The likelihood measures how well a hypothesis explains the particular observed data versus other possible, but unobserved, datasets. Although we have borrowed terminology from Bayesian statistics here, most modern theories of inductive inference make some trade-off between fit to the data and a priori hypothesis quality (Bernardo and Smith 1994, Rissanen and Ristad 1994, Vapnik 1998, Jain et al. 1999, Grünwald 2007, Li and Vitányi 2008, Clark and Lappin 2011, Shalev-Shwartz and Ben-David 2014).

As Rasin and Katzir (2016) point out, early generative grammar often emphasized the prior—favoring simplicity concerns in grammar at the expense of fit to the data. This led to a number of well-known learnability problems. One such problem is that overly general grammars will sometimes be selected on the basis of simplicity. This is the learnability problem identified by Rice (1997) and discussed in the preceding section. Since the simplest grammar is one that

---

3 A complete model of learning requires many other ingredients, such as the specification of a hypothesis space, a data presentation method, and search algorithms. Many learning frameworks also do not specify an explicit evaluation metric; rather, they leave comparison between hypotheses implicit in search algorithms or representational assumptions. Nevertheless, even these frameworks induce some implicit metric on the hypothesis space.

4 An exception to this was Chomsky 1979, which includes a secondary criterion of length of derivations that provides an indirect incorporation of fit to data. Since this criterion is nested inside a criterion of grammatical simplicity, it does not lead to the same sort of trade-offs that are present in modern theories of inductive inference.
groups all words together and only makes use of the Foreign sublexicon constraints, why not simply choose that as the phonotactic grammar for the whole language? This problem of overgeneralization was also famously identified by Gold (1967) in the mathematical setting and discussed by Baker (1979) in a linguistics setting (see also Pinker 1979, Dell 1981, Berwick 1985). The solution to this problem is straightforward. In addition to a bias for simple grammars, we must include an opposing bias for fit to the data—and then balance the two. In the current study, we develop such a model. By including both prior and likelihood terms in our evaluation metric, our model is able to learn Japanese sublexica from naturalistic data extracted from a corpus.

2.5 Naturalistic Empirical Learnability

Note that unlike the previous studies on the Japanese sublexica, or much of the work on learnability discussed in the last section, our goal is not to design an algorithm that induces certain aspects of phonotactics from purely hypothetical data, nor to mathematically prove the absence of such an algorithm under hypothetical conditions. Instead, we examine whether sufficient evidence for the existence of multiple sublexica is present in existing Japanese words. We call our simulation-based approach a naturalistic empirical learnability study (see, e.g., Perfors et al. 2011, Pearl, Ho, and Detrano 2017, Bergen, Gibson, and O’Donnell 2015), to contrast it with learnability studies that attempt mathematical proof of learnability under general conditions or simulations with toy data (e.g., Gold 1967, Angluin 1988).

2.6 Generative Phonotactics

To build our model, we also adopt another assumption that differs markedly from most earlier studies. Most work in phonotactics, including discussions of Japanese sublexica, have focused on relatively small numbers of phonotactic constraints, such as those outlined in section 2.1. Since each constraint only makes reference to a local configuration of natural classes, constraint-based theories typically do not provide a complete model of the structure of all words. In order to do this, they must be augmented with some account of the structure of the underlying forms to which constraints are applied. One example is the richness of the base assumption employed in Optimality Theory. Richness of the base proposes that all strings of phones are equiprobable prior to constraint application (see Rasin and Katzir 2015 for discussion of problems with this assumption). This approach is also implicit in maximum entropy models such as that of Hayes and Wilson (2008). The principle of maximum entropy seeks the distribution over forms that is

---

5 More accurately, one of Gold’s (1967) results was that under his model of learning, no class of formal languages that contains an infinite chain of finite languages and the infinite union of those languages can be learnable. This is essentially a more mathematically precise version of the problem discussed by Rice (1997).

6 In the history of linguistics, recognition of overreliance on grammatical simplicity led to the introduction of a generation of learning models that instead emphasized fit to the data (likelihood). These included models that made use of one variety or another of the Subset Principle, which suggested always choosing the grammar that generalized least (Dell 1981, Berwick 1985). Unlike earlier proposals that systematically overgeneralized, these approaches systematically overfit.
as close to uniform as possible while still respecting the constraints. Assumptions such as these imply that all words without constraint violations are equiprobable and thus risk missing important statistical generalizations about sublexical affiliation.\footnote{It is of course possible to imagine a constraint-based theory that is exhaustive in the sense that it specifies a constraint for every possible combination of natural classes and thus addresses the problem discussed in the text. However, we know of no attempt to do this, and in practice it seems that generative models are typically used when the modeler wishes to give an explicit formula for generating every word form.}

We take an alternative approach, building a generative model of phonotactic structure (Futrell et al. 2017). This approach, which revives an earlier tradition of morpheme structure rules or conditions (Halle 1959, Stanley 1967, Chomsky and Halle 1968, Booij 2011, Rasin and Katzir 2015, 2020), sees phonotactics as a module of grammar that generates phonological forms of lexical items and, as such, provides a complete account of statistical redundancies in the set of allowable lexical items. In contrast to constraint-based models, generative models typically make use of a more expressive likelihood that takes into account a wider variety of cooccurrence patterns. Thus, rather than describing Japanese sublexica using a relatively small number of phonotactic constraints that rule out certain patterns, our model provides a complete generative process for word forms in each sublexicon, assigning a probability to all possible patterns. The model represents a lexicon-specific conditional probability for every segment in each word and thus characterizes the probability of all segments exhaustively. This representation is less sparse than those typically used in constraint-based models and thus allows our model to capture a greater number of patterns characteristic of each sublexicon. In the next section, we introduce our model.

3 Methods
In this section, we describe our phonotactic model and sublexicon-learning method, which we will use in section 4 to show the naturalistic learnability of the Japanese sublexica. First, we outline our approach (section 3.1). Then, after defining our mathematical notation (section 3.2), we provide a high-level description of our phonotactic model and clustering method (section 3.3). The full mathematical description is provided in section S1 of the supporting information at https://doi.org/10.1162/ling_a_00401.

3.1 Overview
As mentioned in section 2.3, we model learning of Japanese sublexica by balancing a trade-off between prior and likelihood. The prior encourages model simplicity by preferring systems with small numbers of sublexica. It is implemented by a distribution over grammars that is maximized when the number of sublexica is small. The likelihood measures the probability of the data given assignments of words to sublexica. It is high when words within a sublexicon share similar phonotactics and are thus assigned high probability by the sublexicon. These are opposing biases and thus lead to a trade-off. To understand the trade-off, it is useful to consider two extreme cases.
At one extreme, imagine that every word is assigned to its own sublexicon. Under this condition, the likelihood will be maximized since each cluster will have a very specific phonotactic distribution that places very high probability on the single word in that sublexicon. At the same time, the prior will disfavor such a solution since it will involve an enormous number of sublexica. At the other extreme, imagine that all words are assigned to a single sublexicon. The prior will now be maximized, since this is the minimum number of possible sublexica. However, a single, very general phonotactic distribution will be needed to account for a large diversity of different—and potentially dissimilar—word forms. Thus, the likelihood will penalize such a solution. The goal of inference is to find an assignment of words to sublexica that balances this trade-off—that is, where each sublexicon tightly predicts word forms that it contains, while at the same time avoiding too many clusters.

3.2 Mathematical Notation

Unless otherwise specified, we follow standard conventions. Random variables are denoted by capital letters (e.g., $X$), while particular (i.e., sampled) values are denoted by corresponding lowercase letters (e.g., $x$). We overload notation and allow Greek letters (e.g., $\alpha$) to mean either random variables or their values depending on context. Boldface symbols denote a finite/infinite sequence of random variables or values: for example, $X := (X_1, \ldots, X_l)$ and $x := (x_1, \ldots, x_l, \ldots)$. We use $p(x)$ to denote either a probability mass or a probability density function of $x$, depending on whether $X$ is a discrete or continuous random variable.

3.3 High-Level Description

Our generative process samples a word in two steps: (i) choose a sublexicon from the prior distribution over sublexica, and then (ii) sample a word as a string of phonetic segments using the phonotactic distribution associated with this sublexicon—that is, the likelihood conditioned on the sublexicon. This section introduces these prior and likelihood probability distributions. We will keep the introduction intuitive, leaving mathematical details to S1 of the online supporting materials. We first describe the prior (section 3.3.1) and then the likelihood (section 3.3.2). Finally, we illustrate inference in this model with a toy example (section 3.3.3).

3.3.1 Prior on Sublexica

The first component of our model is a prior distribution that prefers fewer, more reusable sublexica. We assume that the number of sublexica is finite but unbounded and that each sublexicon is indexed by a positive integer $t$. With this assumption, we can define a distribution over sublexica by defining a distribution on the positive integers. Let $\pi_t$ be the probability of the $t$-th sublexicon. We thus seek a distribution over the infinite sequence of probabilities $\pi := (\pi_1, \pi_2, \ldots, \pi_t, \ldots)$, such that $\sum_{t=1}^{\infty} \pi_t = 1$. A popular way to construct a prior on $\pi$ is by the Dirichlet process (DP; Ferguson 1973, Antoniak 1974, Sethuraman 1994). We leave the mathematical definition of the DP to section S1.1 in the online supporting information. However, we note an important property of this prior distribution. Under the DP, $\pi_t$ shrinks exponentially quickly (in probability) as the index $t$ increases. Accordingly, the probability of sampling sublexica indexed by large $t$ becomes vanishingly small very quickly under this prior.
This results in a bias toward reusing a small number of sublexica (i.e., indexed by small integers). The DP has been widely used as a prior over linguistic inventories such as sets of words, phones, and lexica in linguistics and other domains (e.g., Anderson 1990, Goldwater, Griffiths, and Johnson 2006, Teh 2006, Teh et al. 2006, Kemp, Perfors, and Tenenbaum 2007, Goldwater, Griffiths, and Johnson 2009, Feldman et al. 2013, O’Donnell 2015).

3.3.2 Likelihood In our model, the likelihood defines a probability distribution over words given a sublexicon. Following previous studies such as those of Hayes and Wilson (2008) and Futrell et al. (2017), we formalize the likelihood in terms of a model of strings of phonetic segments in order to capture phonotactic well-formedness: words with greater probability are phonotactically more well-formed (for that sublexicon) than those with smaller probability. The main difference from the earlier models is that we allow multiple sublexica and thus multiple likelihood distributions for different subsets of words in the lexicon.

We adopt a simple and widely used model of strings, n-gram models, which exhaustively express local, adjacent phonotactic dependencies among segments. See section 4.4 for a discussion of the linguistic implications of this choice. We provide a basic description of n-gram models below. The full mathematical definition of the particular version of n-gram models we adopt is given in section S1.2 of the online supporting information.

Our n-gram model generates words segment by segment from left to right. The generation of each segment is conditioned on the preceding \(n-1\) segments (which we call the context). Let us consider a toy example with an inventory consisting of two symbols, \(a\) and \(b\), and \(n = 3\) (a trigram model). To generate a word, we initialize the context as \((\text{PREWD}, \text{PREWD})\), where \(\text{PREWD}\) is a distinguished symbol that pads out a context when it would otherwise be shorter than \(n-1\) symbols. We also assume another special symbol, END, that indicates the end of the word. The probability of getting \(a\) as the first segment of the word is \(p(a | \text{PREWD}, \text{PREWD})\), the probability of \(b\) is \(p(b | \text{PREWD}, \text{PREWD})\), and the probability of \(\text{END}\) is \(p(\text{END} | \text{PREWD}, \text{PREWD})\). We first sample this initial symbol. If \(\text{END}\) is chosen, we terminate this segment-sampling process, and the resulting word is the empty string, \(\varepsilon\).\(^8\) Otherwise, we update the context to \((\text{PREWD}, a)\) or \((\text{PREWD}, b)\) depending on the value of the first sample, and we sample the second segment according to the probability distribution conditioned on this updated context. We continue this segment-sampling process until we eventually sample the \(\text{END}\) symbol. Accordingly, the joint probability of the entire word is the product of the trigram probability of the component segments (including \(\text{END}\)). For instance, the probability of \(aabaab\) is factorized as follows:

\[
p(aabaab) = p(\text{END} | a, b)p(b | a, a)p(a | b, a)p(a | a, b) \\
p(b | a, a)p(a | \text{PREWD}, a)p(a | \text{PREWD}, \text{PREWD})
\]

Note that we have two occurrences of \(p(b | a, a)\) on the right-hand side of the equation above. The trigram model must assign the same value to them; thus, it cannot distinguish occurrences

\(^8\) Note that real data would never include the empty string. In this article, however, we leave the empty string as a possible sequence generable by the model since removing it would introduce additional mathematical complexity. This choice does not affect any results.
of b after aa (or any other bigram) in various positions in words. In formal language-theoretic terms, n-gram models correspond to strictly local grammars, which are said not to be powerful enough to capture all phonotactic patterns in the world’s languages (Hansson 2001, Rose and Walker 2004, Heinz 2010). Nevertheless, such grammars can capture many patterns and have been employed in a variety of earlier phonotactic models (Gafos 1999, Ní Chiosáin and Padgett 2001, Hayes and Wilson 2008), and they can be understood as providing a way to lower-bound the phonotactic information that is available to learners. As we will show, such patterns are rich enough to allow our learner to reliably acquire Japanese sublexica. Of course, it will be interesting to pursue more linguistically motivated (e.g., Gorman 2013, Futrell et al. 2017) or more statistically powerful (e.g., Pimentel, Roark, and Cotterell 2020) models in future work.

3.3.3 Posterior Inference Once we define the prior and likelihood distributions, sublexicon learning can be formulated by finding an assignment of words to sublexica that optimally balances the trade-off between the two. Here, we provide a toy example of such an assignment using the DP prior and trigram likelihood (i.e., \( n/H11505 \)) introduced above.

We consider a case where all words take one of the following two forms: \((ab)^m\) or \((aab)^m\), where \(m\) denotes an arbitrary number of repetitions of the substrings in the parentheses. We compare two particular cases of sublexical assignment: (i) all the words are generated from a single sublexicon, and (ii) the \((ab)^m\) words and \((aab)^m\) words are generated from different sublexica, SUBLEX\(_{ab}\) and SUBLEX\(_{aab}\), respectively. These two possibilities are illustrated in table 3. Note that in the actual simulation of Japanese sublexicon learning, we do not pick specific configurations and compare them; rather, we approximate the posterior distribution over all possible assignments (see section S1.3 of the online supporting information for details). The current example is merely intended to provide intuitions.

When there is only one sublexicon (configuration (i) in table 3), the model must allow a transition to both a and b from the context a. Since probability distributions sum to 1 (i.e., \( p(b \mid ba) + p(a \mid ba) + p(END \mid ba) = 1 \)), this means that the transition probability from the context a to both a and b must be less than 1 (i.e., \( p(a \mid ba) < 1 \) and \( p(b \mid ba) < 1 \)). By separating the two languages into distinct sublexica (configuration (ii) in table 3), the model can assign a higher probability to strings by noting that in the \((ab)^m\) subset of words only b is possible after ba and in the \((aab)^m\) subset of words only a is possible after ba. That is, \( p(b \mid ba, \text{SUBLEX}_{ab}) = 1 \) and \( p(a \mid ba, \text{SUBLEX}_{aab}) = 1 \). Therefore, dividing these strings into two subsets allows the assignment of higher likelihoods to forms. If this advantage outweighs the prior cost of multiple sublexica, then (ii) will be preferred to (i).

In a Bayesian setting, the balance between likelihood and prior is captured by the definition of the posterior distribution. The posterior is the conditional probability of a model configuration (hypothesis) given data—in our case, the probability of assigning each word to some particular sublexicon, conditioned on the segmental information in the set of words. Let \( z = (z_1, \ldots, z_N) \)

\(9\) This is a statement of a property of strictly local grammars known as \(n\)-local suffix substitution closure.
Table 3
The extremely simplified example that motivates sublexical classification of words

<table>
<thead>
<tr>
<th></th>
<th>(i) 1 lexicon</th>
<th>(ii) 2 sublexica</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SUBLEX&lt;sub&gt;ab&lt;/sub&gt;</td>
<td>SUBLEX&lt;sub&gt;aab&lt;/sub&gt;</td>
</tr>
<tr>
<td>a bab</td>
<td>a bab</td>
<td>a bab</td>
</tr>
<tr>
<td>p(b</td>
<td>ba)&lt;1</td>
<td>p(b</td>
</tr>
<tr>
<td>aa baa b</td>
<td>aa baa b</td>
<td>aa baa b</td>
</tr>
<tr>
<td>p(a</td>
<td>ba)&lt;1</td>
<td>p(a</td>
</tr>
<tr>
<td>aa baa baa b</td>
<td>aa baa baa b</td>
<td>aa baa baa b</td>
</tr>
<tr>
<td>p(a</td>
<td>ba)&lt;1</td>
<td>p(a</td>
</tr>
<tr>
<td>aba bab</td>
<td>aba bab</td>
<td>aba bab</td>
</tr>
<tr>
<td>p(b</td>
<td>ba)&lt;1</td>
<td>p(b</td>
</tr>
</tbody>
</table>

be a random variable indicating the sublexical assignment of the i-th word in the data (i \in \{1, \ldots, N\}, where N is the total number of word tokens); that is, its value z_i denotes the sublexicon for word i. And let x_i be the segmental representation of the word i and x := (x_1, \ldots, x_N) the segmental information for all words in the corpus. By Bayes’s Rule, the posterior probability p(z | x) is proportional to the product of prior p(z) and likelihood p(x | z):

\[
p(z \mid x) \propto p(z)p(x \mid z) = \frac{p(z)p(x \mid z)}{\sum z' : p(z')p(x \mid z')} \propto p(z)p(x \mid z)
\]

In order to achieve a high posterior probability, we need to choose an assignment of words to sublexica—that is, z—that makes both the prior and likelihood terms in the expression above as high as possible. However, since increasing one term tends to decrease the other, such maximization will occur by optimizing this trade-off. See section S1.3 of the online supporting information for an algorithm to approximate this posterior probability distribution for the specific models used in this article.

4 Simulations

To study the learnability of Japanese sublexica, we apply our clustering method to a corpus of Japanese words. Here, we introduce the data we use (section 4.1), define the values of the free parameters of the model (section 4.2), report the results (section 4.3), and discuss our findings (section 4.4).
4.1 Data

Our data consisted of Japanese words appearing in the Balanced Corpus of Contemporary Written Japanese (BCCWJ, ver 1.0). The word frequency list of the BCCWJ contains information on the sublexical affiliation (in the wType column) and syllabary representation (katakana, in the 1Form column) of the words.\textsuperscript{10} We translated the syllabary representation into IPA notation following Vance (2008), and our system was provided with only the phonetic representation of the words (i.e., no orthographic information).\textsuperscript{11} Note that Vance adopts some nonstandard transcriptions of consonants: onset consonants are represented as short (e.g., [n], as in [naka] ‘inside (Native)'), codas are marked by a single extra length mark (e.g., [n:], as in [hoton:do] ‘most (Native)'), and geminates (coda followed by identical onset) are marked by two extra length marks (e.g., [n::], as in [on::a] ‘female (Native)'). Long segments (i.e., long vowels, codas, geminates) are treated as single symbols rather than sequences of a short consonant and the length mark ([:]). Compounds are listed as individual words and are not morphologically segmented into their root morphemes.

We focused on the \textit{core} words (with nonzero core_frequency in the frequency list), which were checked by hand and are more accurate than the other portions of the data according to the database manual. We further limited our data to nouns that are noninflected in Japanese and thus are suitable for morphologyless models like ours. We do not believe that our focus on nouns makes sublexicon learning significantly easier, because the vast majority of core SJ and Foreign words are nouns: 92.9\% of the SJ core words and 97.6\% of the Foreign core words are nouns. Finally, we excluded words containing personal information (e.g., proper names of persons) from the analysis because their syllabary representation was masked (replaced with a sequence of “_”s) in the database and we were thus unable to extract their phonetic representations. After filtering, the distribution of sublexica was as in Table 4.

\begin{table}[h]
\centering
\caption{Type frequency of the core nouns in the BCCWJ word list, categorized by their sublexical affiliation}
\begin{tabular}{l|c}
\hline
Sublexicon & Type frequency of words \\
\hline
Native & 5,107 \\
SJ & 12,427 \\
Foreign & 4,316 \\
Mixed & 820 \\
Proper Names & 7,322 \\
Symbols & 562 \\
\hline
Total & 30,554 \\
\hline
\end{tabular}
\end{table}

\textsuperscript{10} The word frequency list (BCCWJ\_frequencylist\_suw\_ver1\_0.zip) and its manual (BCCWJ\_frequencylist\_manual\_ver1\_0.pdf) are available at http://pj.ninjal.ac.jp/corpus\_center/bccwj/en/freq-list.html. The BCCWJ was the only database that provided etymological information for its words when the current research project started. More recently, a similar word list became available, called the Corpus of Spontaneous Japanese.

\textsuperscript{11} Our Python code for the translation is available at https://github.com/tkc-morita/variational_inference_DP_mix_HDP_topic_ngram.
In addition to the three etymological sublexica (Native “和,” SJ “漢,” and Foreign “外”), the word list annotates three other classes of nouns: Mixed “混,” Proper Names “固,” and Symbols “記号.” Mixed words are morphologically complex words consisting of morphemes from different sublexica: for example, [mai] ‘every (SJ)’ + [toci] ‘year (Native)’ \( \mapsto \) [mai-toci] ‘every year’. Proper Names are put into a single class though they may come from different languages: for example, [rjut:kjut:] ‘Okinawa’ sounds SJ while [fer:ari] ‘Ferrari’ sounds Foreign. Finally, words written in non-Japanese orthography in the BCCWJ are classified as Symbols. They consist primarily of initialisms (e.g., [di:bu:idi:] ‘DVD’) and acronyms (e.g., [ra:n:] ‘LAN’), which are written in Latin script.

### 4.2 Parameters and Simulation Runs

We set our \( n \)-gram order to \( n = 3 \), adopting a trigram model for the phonotactic likelihood distribution. See S4 of the online supplementary materials for the other free parameters of this simulation (S1 provides background on the mathematical details of the model).

### 4.3 Results

In this section, we report the results of our simulations. We first examine whether our sublexicon-learning system can recover the correct classification of Japanese words in our database, showing that our method can indeed recover the major divisions of Japanese words (section 4.3.1). We then turn to the question of which phonological patterns the model is using to sort the words into sublexica (section 4.3.2). We show that the model makes use of several of the patterns mentioned in the earlier work discussed in section 2.1. We also perform a more systematic investigation of the phonotactic patterns that are characteristic of different Japanese sublexica, identifying previously unnoticed patterns that can be tested in future experimental studies. Finally, in section 4.3.3 we analyze whether our model predicts the experimental results reported by Moreton and Amano (1999), discussed in section 2.2.

#### 4.3.1 Classification of Words

We first evaluate the classification of the words into sublexica predicted by our method. Our analyses make use of the maximum a posteriori (MAP) classification of words under our model. The MAP classification refers to the assignment of each word to a sublexicon that is most probable under the posterior distribution of the learned model. We compare these MAP sublexicon assignments with the gold-standard sublexicon annotations in the BCCWJ database.\(^{12}\)

The heatmap in figure 2 shows the MAP classification of the data.\(^ {13}\) The rows correspond to the gold-standard sublexica from the BCCWJ database, the columns correspond to the categories

\(^{12}\) The posterior distribution over cluster assignments (i.e., \( p(\mathbf{z} \mid \mathbf{x}) \)) for each word \( i \) is approximated in practice using the variational distribution \( q(\mathbf{z}) \) (see section S1.3 of the online supporting information). We compare the MAP classification of the words, argmax \( q(\mathbf{Z}_i = t) \), against the etymological classification given in the database.

\(^{13}\) The integer indices assigned to the predicted word categories by the DP are as follows: \( \text{SUBLEX}_{\text{Native}} = 4 \), \( \text{SUBLEX}_{\text{SJ}} = 3 \), \( \text{SUBLEX}_{\text{Foreign}} = 0 \), and \( \text{SUBLEX}_{\text{Symbols}} = 5 \).
predicted by the model (which we call SUBLEX\(_{\text{Native}}\), SUBLEX\(_{\text{SJ}}\), SUBLEX\(_{\text{Foreign}}\), and SUBLEX\(_{\text{Symbols}}\)), and the darkness of each cell in the heatmap represents the frequency of the corresponding combination, with a maximum value of 3,000. As the figure indicates, the model discovered three major categories of word, labeled SUBLEX\(_{\text{Native}}\), SUBLEX\(_{\text{SJ}}\), and SUBLEX\(_{\text{Foreign}}\). These three predicted word categories largely capture the three major gold-standard etymological sublexica (see the top-left corner boxed with a dashed line): Native words correspond largely to SUBLEX\(_{\text{Native}}\), SJ words to SUBLEX\(_{\text{SJ}}\), and Foreign words to SUBLEX\(_{\text{Foreign}}\). Almost all of the Mixed words and the Proper Names were also assigned to these predicted categories, but without any noticeable bias toward one of the three. This is expected because Mixed words by definition involve morphemes with etymological origins across sublexica and Proper Names can also refer to persons, places, and the like, either in Japan (Native or SJ) or in other countries (Foreign). The MAP classification predicted an additional minor category, SUBLEX\(_{\text{Symbols}}\), whose members are mostly Symbols (specifically, acronyms such as [di:enue:] ‘DNA’).

Figure 2

*Maximum a posteriori* classification of words into categories (columns) against the etymological sublexica (row). The darkness of each cell in the heatmap represents the frequency of the corresponding combination.
In order to quantitatively evaluate the alignment between the word categories predicted by our clustering and the etymological sublexica according to the database, we compared the performance of our system with random baselines using the V-measure score (Roseberg and Hirschberg 2007). The V-measure score evaluates the similarity of a predicted clustering to a ground truth classification based on two desiderata: (i) each of the predicted clusters should only contain members of a single ground truth class (homogeneity); (ii) the members of each ground truth class should be clustered into the same category (completeness). Homogeneity and completeness are both defined to fall on a scale of 0 (worst) to 1 (best) based on the conditional entropy. The V-measure is their harmonic mean (by analogy with the F-score).

We computed V-measures with respect to two datasets, one consisting of the whole dataset used by our model and the other containing only the Native, SJ, and Foreign words (the top three rows of figure 2), which have been the target of most earlier studies (section 2.1). We report three types of random baselines. The first is uniform, i.i.d. (independent and identically distributed) random assignment of each word to one of the six/three sublexica. The second baseline also classified words i.i.d., but the assignment is not uniform; rather, it is weighted by the frequency of the gold-standard classifications. The last baseline randomly permutes the gold-standard classification of the words. The random baselines are all based on 100,000 independent samples. The results are shown in table 5. As the table indicates, the V-score of the model prediction is greater than the random baselines by three to four orders of magnitude. This result indicates that the predicted categories matched the gold-standard etymological sublexica far beyond the level expected by any form of chance classification.

### Table 5

The V-measure scores of the classifications by the model and random baselines. (a) The data consisted of all the words used in the learning. (b) The data were limited to the Native, SJ, and Foreign words. (SD = standard deviation, i.i.d. = independent and identically distributed)

#### (a) All words

<table>
<thead>
<tr>
<th>Baselines</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform i.i.d.</td>
<td>0.000701</td>
<td>0.000252</td>
<td>0.000071</td>
</tr>
<tr>
<td>Proportional i.i.d.</td>
<td>0.000731</td>
<td>0.000282</td>
<td>0.000080</td>
</tr>
<tr>
<td>Shuffle</td>
<td>0.000819</td>
<td>0.000282</td>
<td>0.000080</td>
</tr>
<tr>
<td><strong>Model prediction</strong></td>
<td><strong>0.375130</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (b) Only Native, SJ, and Foreign words

<table>
<thead>
<tr>
<th>Baselines</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform i.i.d.</td>
<td>0.000671</td>
<td>0.000088</td>
<td>0.000063</td>
</tr>
<tr>
<td>Proportional i.i.d.</td>
<td>0.000659</td>
<td>0.000093</td>
<td>0.000066</td>
</tr>
<tr>
<td>Shuffle</td>
<td>0.000749</td>
<td>0.000093</td>
<td>0.000066</td>
</tr>
<tr>
<td><strong>Model prediction</strong></td>
<td><strong>0.526670</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Some of the putatively misclassified words also exhibited interesting patterns. First, the dataset included 23 exceptional Native words containing a NC/underring substring: for example, [tam:popo] ‘dandelion’, [buran:ko] ‘swing’, and [in:teiki] ‘cheating’. Interestingly, none of them were classified into \textsc{Sublex\_Native}: 13 were classified into \textsc{Sublex\_SJ} and 8 into \textsc{Sublex\_Foreign}. In section 4.3.2, we will show that *NC is in fact highly indicative of \textsc{Sublex\_Native}; thus, this classification of the exceptional Native words is not accidental. Second, the model discovered some mistakes in the ground truth labels of the database: [mat:e:wi] ‘subdivision of cities’ and [umibei] ‘coast’ are labeled as SJ, but they are in fact Native. Our model “misclassified” these words by correctly assigning them to the Native sublexicon \textsc{Sublex\_Native}.

To summarize, we have demonstrated that the gold-standard etymological classification of Japanese words is recoverable from their phonotactic properties. Importantly, our model was completely unsupervised: the system never had any access to the ground truth classification provided in the database, and the predicted clusters were built purely on the statistical information present in the pattern of segments in the Japanese lexicon.

4.3.2 Phonotactic Properties of the Learned Sublexica

We next investigate the phonological properties of the sublexical categories inferred by the model. We will focus on the three sublexica \textsc{Sublex\_Native}, \textsc{Sublex\_SJ}, and \textsc{Sublex\_Foreign}. To study which phonological patterns are indicative of each of these sublexica, we compute the \textit{representativeness} of particular phonological patterns with respect to each cluster of word forms (Good 1950, Tenenbaum and Griffiths 2001).

(1) defines the representativeness $R(a, t)$ of the sequence of segments $a := (a_1, \ldots, a_m)$ with respect to the predicted sublexicon $t$.\[
R(a, t) : = \log \frac{p(\ldots a \ldots | t, x)}{\sum_{t' \neq t} p(\ldots a \ldots | t', x)p(t' | x, t' \neq t)},
\]

where $x$ is the training data. $R(a, t)$ is the log ratio between the posterior predictive probability of $a$ given that the string is a part of a word belonging to the sublexicon $t$ vs. the average posterior predictive probability of $a$ over the other sublexica, given that $t$ was not the sublexicon. $R(a, t)$ is positive if $a$ is more probable in sublexicon $t$ than in other sublexica. Note that representativeness is highest for those patterns that are highly probable for the target sublexicon while being improbable under the other sublexica. Thus, it tells us what strings of segments would be especially informative cues for classification. See section S2 of the online supporting information for details on computation of $R(a, t)$.

In table 6, we first report the representativeness of segment sequences that have been discussed in previous literature (see table 1). Our computed scores correctly recover the (non)representativeness of most of the 25 patterns that we identify in the table as having been discussed in prior studies (see section 2.1).

We organize our discussion by the blocks identified in the leftmost column of table 6. As Block 1 indicates, NC sequences exhibit systematically negative representativeness with respect to \textsc{Sublex\_Native} just as discussed in the literature. Block 2 shows that the representativeness of [rj] and [cj] is positive only for \textsc{Sublex\_SJ}.
Table 6
The representativeness of substrings of particular interest (table 1). Values in the shaded cells are predicted to be small according to the previous literature.

<table>
<thead>
<tr>
<th>Block</th>
<th>Substring</th>
<th>Freq.</th>
<th>( \text{SUBLEX}_{\text{Native}} )</th>
<th>( \text{SUBLEX}_{\text{SJ}} )</th>
<th>( \text{SUBLEX}_{\text{Foreign}} )</th>
<th>Generalization in the literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mp</td>
<td>273</td>
<td>-7.5980</td>
<td>3.6860</td>
<td>-2.4883</td>
<td>* only in Native</td>
</tr>
<tr>
<td></td>
<td>nt</td>
<td>449</td>
<td>-2.9391</td>
<td>-0.2216</td>
<td>1.3501</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ns</td>
<td>18</td>
<td>-1.6533</td>
<td>-1.4371</td>
<td>2.1788</td>
<td></td>
</tr>
<tr>
<td></td>
<td>nk</td>
<td>611</td>
<td>-4.9233</td>
<td>1.2612</td>
<td>-0.0316</td>
<td></td>
</tr>
<tr>
<td></td>
<td>nc</td>
<td>153</td>
<td>-6.5132</td>
<td>5.4464</td>
<td>-4.3796</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uq:s</td>
<td>452</td>
<td>-8.8335</td>
<td>1.6871</td>
<td>-0.4558</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uq:th</td>
<td>11</td>
<td>-7.9877</td>
<td>1.8977</td>
<td>-0.6605</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uq:q</td>
<td>3</td>
<td>-6.6560</td>
<td>2.5832</td>
<td>-1.3859</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uq:ph</td>
<td>35</td>
<td>-7.2159</td>
<td>0.6721</td>
<td>0.5489</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uq:q</td>
<td>318</td>
<td>-8.6983</td>
<td>2.5207</td>
<td>-1.3161</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>rij</td>
<td>512</td>
<td>-6.1160</td>
<td>1.5819</td>
<td>-0.5465</td>
<td>✓ only in SJ</td>
</tr>
<tr>
<td></td>
<td>çj</td>
<td>116</td>
<td>-5.0908</td>
<td>4.0184</td>
<td>-3.0207</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>p</td>
<td>1,326</td>
<td>-3.3707</td>
<td>-1.2437</td>
<td>1.8241</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a:</td>
<td>1,123</td>
<td>-2.3588</td>
<td>-2.5074</td>
<td>2.5131</td>
<td>✓ only in Foreign</td>
</tr>
<tr>
<td></td>
<td>a:END</td>
<td>510</td>
<td>-2.3734</td>
<td>-2.0740</td>
<td>2.9153</td>
<td></td>
</tr>
<tr>
<td></td>
<td>φa</td>
<td>78</td>
<td>-0.1194</td>
<td>-2.2192</td>
<td>1.6128</td>
<td></td>
</tr>
<tr>
<td></td>
<td>φe</td>
<td>59</td>
<td>-2.2314</td>
<td>-2.7172</td>
<td>3.2281</td>
<td></td>
</tr>
<tr>
<td></td>
<td>φi</td>
<td>65</td>
<td>-4.7132</td>
<td>-1.7469</td>
<td>2.9279</td>
<td></td>
</tr>
<tr>
<td></td>
<td>φo</td>
<td>34</td>
<td>-0.8623</td>
<td>-2.3328</td>
<td>2.2397</td>
<td></td>
</tr>
<tr>
<td></td>
<td>φu</td>
<td>908</td>
<td>-0.4153</td>
<td>-0.3846</td>
<td>0.5256</td>
<td>✓ in all</td>
</tr>
<tr>
<td>4</td>
<td>b::</td>
<td>4</td>
<td>-0.6465</td>
<td>-0.3553</td>
<td>0.1576</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d::</td>
<td>51</td>
<td>-1.3744</td>
<td>-1.2067</td>
<td>1.4787</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d::</td>
<td>5</td>
<td>-0.5507</td>
<td>-0.0552</td>
<td>-0.2357</td>
<td>✓ only in Foreign</td>
</tr>
<tr>
<td></td>
<td>g::</td>
<td>18</td>
<td>-0.9323</td>
<td>-0.6944</td>
<td>0.7624</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d::</td>
<td>18</td>
<td>-0.8060</td>
<td>-0.6934</td>
<td>0.7012</td>
<td></td>
</tr>
</tbody>
</table>

Block 3 shows several patterns that are characteristic of the Foreign sublexicon. The representativeness of [p] (first sub-block) and [a:] (second sub-block) is positive only for \( \text{SUBLEX}_{\text{Foreign}} \). The representativeness of [φ] is also only positive for \( \text{SUBLEX}_{\text{Foreign}} \) (third sub-block). However, this effect is weaker when [φ] is followed by [tu], which is said to be allowed exceptionally in the Native and SJ sublexica as well (last sub-block).

\(^{14}\) Table 6 shows that the representativeness of [m:p] is the greatest for \( \text{SUBLEX}_{\text{SJ}} \) (Block 1), while [p] in general has the greatest bias toward \( \text{SUBLEX}_{\text{Foreign}} \) (first sub-block of Block 3). This exceptional behavior of [m:p] has not attracted attention from theoretical linguists, but it has been well-recognized by historical linguists: [p] used to be included in SJ words but was changed into [h] except after [m:] (Frellesvig 2010).
However, Block 4 of table 6 indicates that our model predictions do not strongly support the representativeness of the voiced geminates for the Foreign sublexicon. Those segments have systematically negative representativeness with respect to \textsc{Sublex}=\textsc{Native} and \textsc{Sublex}=\textsc{SJ}, and positive representativeness with respect to \textsc{Sublex}=\textsc{Foreign} (with the exception of [dːː]).\footnote{[dːː] had high representativeness with respect to \textsc{Sublex}=\textsc{Symbols} (2.1856), which makes it possible that its representativeness is always negative with respect to the other three categories.} However, these representativeness scores are close to 0 (with the exception of [dːː], which is the fourth most representative segment of \textsc{Sublex}=\textsc{Foreign}). Also, note that the Foreign-representativeness of the (non-[dːː]) voiced geminates is close to or smaller than that of [ʃu], which is known to be allowed in all sublexica. Hence, the results suggest that the voiced geminates are not strongly indicative of the Foreign sublexicon.

The low degree of representativeness of voiced geminates for the Foreign sublexicon does not imply that words containing them are necessarily misclassified.\footnote{In a similar vein, the low degree of Foreign-representativeness of voiced geminates also does not imply that they are not marked in the Native and SJ sublexica (Kuroda 1965, Kurisu 2000), since the probability of these segments is extremely low in these sublexica as well.} As noted above, their representativeness for \textsc{Sublex}=\textsc{Foreign} is nonnegative except for [dːː]; thus, they are unbiased with respect to classification into \textsc{Sublex}=\textsc{Foreign}. In fact, the words containing voiced geminates were correctly classified into \textsc{Sublex}=\textsc{Foreign} overall: The expected number of voiced geminates that were classified into the \textsc{Sublex}=\textsc{Foreign} sublexicon was \(17.05\) for [bːː], \(50.88\) for [dːː], \(17.99\) for [gːː], and \(17.05\) for [dːː]. Only [dːː] occurred in words typically classified into other sublexica: \(3.99\) in \textsc{Sublex}=\textsc{SJ} and \(0.98\) in \textsc{Sublex}=\textsc{Native}. Such correct classifications are possible because our model does not rely on a single phonotactic pattern, but rather integrates information across all segments in words. For example, words including voiced geminates may also contain other Foreign-representative segments (as exemplified by [aːː] in [kaːtoridːːi] ‘cartridge’). If a word contains several weakly biased patterns (including geminates), these can accumulate and lead to a strong Foreign classification (e.g., the trigram [aqːːu], as in [doɾaːqːːu] ‘drag’, has a representativeness of 3.8097 with respect to \textsc{Sublex}=\textsc{Foreign}).

We hypothesize that the low degree of representativeness of the voiced geminates for the Foreign sublexicon is simply due to the small number of forms in our training data that contain these patterns:\footnote{An anonymous reviewer points out that voiced geminates are not limited to Foreign words, given that Native words have an emphatic form with gemination (e.g., [kudarana] ‘trivial, useless’ \[\rightarrow \] [kudarana]; Gelbart and Kawahara 2007). Such emphatic gemination, however, would have had little effect on our model predictions. We showed that the words containing the voiced geminates were correctly classified into \textsc{Sublex}=\textsc{Foreign} overall, so the low representativeness for the sublexicon is unlikely to be due to the ambiguity between \textsc{Sublex}=\textsc{Native} and \textsc{Sublex}=\textsc{Foreign}. In addition, the emphatic gemination is seldom observed in nouns (Gelbart and Kawahara 2007).} with the exception of [dːː], which is in fact representative of \textsc{Sublex}=\textsc{Foreign}, these geminate patterns each occurred fewer than 20 times in our dataset. Since our model makes use of no phonological or phonetic information to group segments, it must independently learn that each of these gemination patterns is representative of \textsc{Sublex}=\textsc{Foreign}. Real learners may exploit phonetic similarities among these patterns and thus be able to group together the total of
96 geminate occurrences. Since \([d\text{-}}\), which accounts for 51 occurrences, exhibits high representativeness for the Foreign sublexicon and the vast majority of words containing other voiced geminates were correctly classified into SUBLEX\textsubscript{Foreign}, real learners may be able to infer that the entire class of voiced geminates is representative of the Foreign sublexicon.

There is another claim in the literature that is more clearly inconsistent with the predictions of our model: Gelbart and Kawahara (2007) have claimed that the representativeness of \([a\text{-}}\) is in fact representative of SUBLEX\textsubscript{Foreign} in any position in a word. The representativeness score of \([a\text{-}}\) is only slightly higher than that of \([a\text{-}}\] in other positions (see the second sub-block of Block 3). These results support Moreton and Amano’s (1999) claim that \([a\text{-}}\] is characteristic of Foreign words “irrespective of position” in words. 18

So far, we have discussed phonotactic patterns that have appeared in previous literature. We can also use representativeness as a way to discover previously unknown patterns that are characteristic of each lexicon. When choosing representative patterns of interest, it is also important to consider the absolute probability of the pattern, not merely its representativeness; these two quantities may not be well-aligned. To see this point, suppose that the probability of one substring, \(a\), is \(p(...a...|t,x) = 0.1\) in the sublexicon \(t\) whereas its average probability over the other sublexica is \(\sum_{t'} p(...a...|t',x)p(t'|x, t' \neq t) = 0.001\). Accordingly, the representativeness of \(a\) is \(R(a, t) = \log \frac{0.1}{0.001} = 2 \log 10\). Suppose, on the other hand, that the probability of another substring, \(b\), is much smaller, \(p(...b...|t,x) = 10^{-101}\). However, if \(b\) is also extremely improbable in the other sublexica and \(\sum_{t'} p(...b...|t',x)p(t'|x, t' \neq t) = 10^{-103}\), the representativeness of \(b\) is the same as that of \(a\): \(R(b, t) = \log \frac{10^{-101}}{10^{-103}} = 2 \log 10\), even though substring \(b\) is improbable for the language as a whole (i.e., words containing \(b\) are not particularly probable).

This scenario is not unusual for our model, which assigns nonzero probability to all logically possible substrings (due to our backoff scheme; see section S1.2 of the online supporting information; Kneser and Ney 1995, Goldwater, Griffiths, and Johnson 2006, Teh 2006). For example, according to our model \([je]\) is the most representative bigram for SUBLEX\textsubscript{Native}. However, it did not appear in our dataset at all. Its predicted representativeness is purely based on how our model handles sparsity (in fact, \([je]\) is more likely to be characteristic of the Foreign lexicon).

Tables 7a–c list the most representative patterns for each lexicon that appear in the input at least once (see section S5 of the online supporting information for the ranking without the frequency filter). The tables reflect some of the existing generalizations we remarked upon earlier: \([a\text{-}}\), \([p]\), and their combinations are representative of SUBLEX\textsubscript{Foreign}; \([ej]\) is the fourth most representative bigram with nonzero frequency (and the tenth among all) in SUBLEX\textsubscript{SJ}. More importantly, these tables also show that we have discovered several new characteristic patterns.

18 Note, however, that Moreton and Amano (1999) also placed \([a\text{-}}\] at the end of their stimuli in their experiment (see sections 2.2 and 4.3.3); thus, there has been no experimental adjudication done between the two slightly different generalizations.
TABLE 7  
1- to 3-gram substrings with nonzero frequency yielding the greatest representativeness (rep.)  

(a) SUBLEX—Native

<table>
<thead>
<tr>
<th>1-gram</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>warai</td>
<td>maka</td>
<td>jama</td>
<td>çito</td>
<td>mizu</td>
</tr>
<tr>
<td>Gloss</td>
<td>‘laugh’</td>
<td>‘inside’</td>
<td>‘mountain’</td>
<td>‘human’</td>
<td>‘water’</td>
</tr>
<tr>
<td>Rep.</td>
<td>1.5753</td>
<td>1.3608</td>
<td>1.2742</td>
<td>1.2315</td>
<td>0.8678</td>
</tr>
<tr>
<td>Freq.</td>
<td>1,079</td>
<td>3,035</td>
<td>5,992</td>
<td>919</td>
<td>1,183</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2-gram</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>kawa</td>
<td>najami</td>
<td>ao</td>
<td>jane</td>
<td>iwa</td>
</tr>
<tr>
<td>Gloss</td>
<td>‘river’</td>
<td>‘worry’</td>
<td>‘blue’</td>
<td>‘roof’</td>
<td>‘rock’</td>
</tr>
<tr>
<td>Rep.</td>
<td>2.7892</td>
<td>2.7484</td>
<td>2.6555</td>
<td>2.5250</td>
<td>2.3772</td>
</tr>
<tr>
<td>Freq.</td>
<td>419</td>
<td>227</td>
<td>188</td>
<td>546</td>
<td>157</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3-gram</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>nezumi</td>
<td>tomoenage</td>
<td>omowaku</td>
<td>kaotsuki</td>
<td>ciwajose</td>
</tr>
<tr>
<td>Gloss</td>
<td>‘mouse’</td>
<td>‘tome nage’</td>
<td>‘expectation’</td>
<td>‘expression’</td>
<td>‘bad influence’</td>
</tr>
<tr>
<td>Freq.</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

(b) SUBLEX—SJ

<table>
<thead>
<tr>
<th>1-gram</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>kānpuka</td>
<td>dairjok</td>
<td>kibo</td>
<td>cimbuin</td>
<td>deqqua</td>
</tr>
<tr>
<td>Gloss</td>
<td>‘patient’</td>
<td>‘ingredient’</td>
<td>‘hope’</td>
<td>‘newspaper’</td>
<td>‘telephone’</td>
</tr>
<tr>
<td>Rep.</td>
<td>2.3543</td>
<td>1.5910</td>
<td>1.5169</td>
<td>1.4594</td>
<td>1.3793</td>
</tr>
<tr>
<td>Freq.</td>
<td>404</td>
<td>566</td>
<td>6,461</td>
<td>533</td>
<td>1,231</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2-gram</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>depiti</td>
<td>dōk</td>
<td>φumcjo</td>
<td>čjaku</td>
<td>seqcinn</td>
</tr>
<tr>
<td>Gloss</td>
<td>‘battery’</td>
<td>‘support’</td>
<td>‘ rumor’</td>
<td>‘hundred’</td>
<td>‘product’</td>
</tr>
<tr>
<td>Rep.</td>
<td>5.4464</td>
<td>5.3251</td>
<td>4.5851</td>
<td>4.0184</td>
<td>3.9432</td>
</tr>
<tr>
<td>Freq.</td>
<td>153</td>
<td>251</td>
<td>8</td>
<td>116</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3-gram</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>dempjo</td>
<td>tepδo</td>
<td>očjkin</td>
<td>pteo</td>
<td>cuiŋohi</td>
</tr>
<tr>
<td>Gloss</td>
<td>‘check’</td>
<td>‘ceiling’</td>
<td>‘queen’</td>
<td>‘tension’</td>
<td>‘instant’</td>
</tr>
<tr>
<td>Rep.</td>
<td>7.6487</td>
<td>7.4778</td>
<td>7.4300</td>
<td>7.4201</td>
<td>7.3457</td>
</tr>
<tr>
<td>Freq.</td>
<td>8</td>
<td>46</td>
<td>8</td>
<td>48</td>
<td>4</td>
</tr>
</tbody>
</table>
Most interestingly, table 7a reveals several new patterns representative of SubLEX−Native; for example, [aw] as in [kawa] ‘river’ and [ezu] as in [nezumi]. Some of these patterns may be the result of historical patterns of change in the Native sublexicon. An anonymous reviewer points out that the high representativeness of the hiatuses, [ao] and [oe] (in [meo]), could be a result of the series of historical changes, p\[^w\]w\[^null\], that happened at intervocalic positions (except before [a]) in Early Middle Japanese (Martin 1987, Miyake 2003, Frellesvig 2010). Importantly, the change was not applied to SJ words (Martin 1987, Miyake 2003) and thus became characteristic of the Native sublexicon.

The discovery of the patterns that are representative of the Native sublexicon is interesting because the Native sublexicon has been described previously only in terms of constraints. Specifically, it has been described solely in terms of disallowed phonotactic patterns (i.e., as being nested in the other sublexica; Ito and Mester 1995a,b, 1999, 2008, Rice 1997, Ota 2004) or low-frequency substrings (Moreton, Amano, and Kondo 1998); no positive patterns characteristic of this sublexicon have been identified in previous work. Our results suggest that there are substrings that provide an explicit cue to Nativeness and deserve future experimental investigation (cf. Moreton and Amano 1999).

4.3.3 Moreton and Amano 1999 In this section, we evaluate our model’s ability to predict the experimental results reported by Moreton and Amano (1999), discussed in section 2.2. Moreton and Amano examined the perceptual thresholds in duration needed to discriminate between [a] and [at] in the context C\(_1\)oC\(_2\) (figure 1). Recall that [at] is representative of the Foreign sublexicon while [a] is not. Moreton and Amano found that these thresholds vary depending on the identity of the consonants, C\(_1\) and C\(_2\)—and thus, by hypothesis, the sublexicon. The results of their experiment indicated that the thresholds are
We tie these results to our model by linking thresholds to a probabilistic quantity: the log ratio of the posterior predictive probability of \([a]\) to that of \([\text{a}]/\text{lengthmark}\) given the context.

\[
\log \frac{p(a | C_1oC_2\_\text{END}, x)}{p(a: | C_1oC_2\_\text{END}, x)}
\]

This quantity has several interpretations and uses in the literature. Log probability ratios are frequently used in models of perceptual judgment (e.g., Feldman, Griffiths, and Morgan 2009). This quantity can be viewed as the difference between the surprisals of the long and short vowels in context. Surprisal has been proposed and successfully studied as a linking hypothesis between probabilistic models and various kinds of linguistic and psychological data—especially response time data (Attneave 1959, Hale 2001, Levy 2008). In our case, the intuition behind this linking hypothesis is that a longer duration is necessary to perceive \([\text{a}]/\text{lengthmark}\) instead of \([a]\) in contexts where \([\text{a}]/\text{lengthmark}\) is improbable (i.e., has higher surprisal) relative to \([a]\). See section S3 of the online supporting information for details on the computation of the log probability ratio.

We evaluate the correlation between the predictions of our model and the mean thresholds reported by Moreton and Amano. Figure 3 shows the scatter plot of the duration threshold against

![Figure 3](image-url)

Figure 3
Scatter plot of the duration threshold between \([a]\) and \([\text{a}]/\text{lengthmark}\) (Moreton and Amano 1999) against log posterior predictive probability ratio of \([a]\) to \([\text{a}]/\text{lengthmark}\).

- **low** when \(C_1/C_2\) are \([p]/[\phi]\) (✓ only in Foreign),
- **intermediate** when \(C_1/C_2\) are \([r]/[t]\) (✓ in both Foreign and SJ), and
- **high** when \(C_1/C_2\) are \([\text{r}]/[\text{qj}]\) (✓ only in SJ).
the predictor, the log probability ratio obtained from our simulations. The result shows that our model is able to capture the core pattern in the threshold data: perception of \( [a:] \) requires more time when \( [a:] \) is less probable in context than \( [a] \). The probability of \( [a:] \), which is representative of the Foreign sublexicon, is small (and the log ratio is high) when it cooccurs with the SJ sounds \([r]\) and/or \([\tilde{c}]\). In contrast, the probability of \( [a:] \) is high (and the log ratio is low) when other Foreign sounds, \([p]\) and/or \([\tilde{g}]\), precede it. And the probability of \( [a:] \) (and the log ratio) takes on an intermediate value when the carrier word contains \([r]\) and \([t]\), which are possible in both the Foreign and SJ sublexica. The Spearman correlation coefficient between the experimental data and the log probability ratio was \( r_s = 0.912 \) (\( p = 0.001 \)), which indicates that the log ratio is a significant predictor of the experimental data.\(^{19}\)

4.4 General Discussion

The results of our classification experiments in section 4.3.1 demonstrate that there is statistical evidence for the division of most of the Japanese lexicon into three sublexica, which largely correspond with the etymological sublexica Native, SJ, and Foreign. At the same time, the results suggest that the Japanese sublexica are learnable from naturalistic data: a learner that had the ability to use information in a way similar to our model could exploit the statistical evidence in the input to discover the sublexica. This result addresses earlier skepticism about the learnability of such a system by Rice (1997) and Ota (2004), and provides evidence for learnability in a more naturalistic setting than in Ito and Mester 1999 and Pater 2005.

The key to this success is the way that the model implemented the trade-off between model simplicity (prior) and fit to the data (likelihood)—an idea that is inherent in every modern theory of inductive inference (Rissanen and Ristad 1994, Stolcke and Omohundro 1994, Osborne and Briscoe 1997, Goldwater, Griffiths, and Johnson 2007, 2009, Johnson, Griffiths, and Goldwater 2007, Li and Vitányi 2008, Hsu and Chater 2010, Hsu, Chater, and Vitányi 2011, O’Donnell 2015, Rasin and Katzir 2016). Rice’s and Ota’s arguments against learnability, on the other hand, were built on a learning model that (i) only considered whether or not data were possible given a hypothetical assignment of words to sublexica and (ii) always chose the simpler hypothesis with a smaller number of sublexica when multiple hypotheses were consistent with the data. Accordingly, a single sublexicon was the best solution under their approach because it was the simplest hypothesis that was consistent with the data. Our result suggests that sublexicon learning becomes possible once a more general notion of fit to the data is taken into account and when fit to the data is balanced against simplicity in a trade-off.

The word categories predicted by our model recovered most of the phonotactic properties of the corresponding etymological sublexica that have been discussed in the previous literature (section 4.3.2). Furthermore, with additional linking hypotheses our model also provides an explanation for the experimental data reported by Moreton and Amano (1999) (section 4.3.3). In

\(^{19}\) The exact \( p \)-value for the Spearman correlation was calculated with R’s cor.test (R Core Team 2018). The Pearson correlation was \( r = 0.868 \).
particular, we showed that variation in the perceptual threshold between [a] and [a lengthmark] reported in that paper was captured by the relative posterior predictive probabilities of the two vowels.

One potential worry about this latter result is that it may not require a model that makes use of sublexica, but instead might be derivable using a single-lexicon phonotactic model, such as a trigram model. To see this, note that the quantity we used to predict the experimental results was the log ratio between the posterior predictive probabilities of [a] and [a lengthmark]:

$$\log \frac{p_{single}(a | C_1oC_2__END, x)}{p_{single}(a lengthmark | C_1oC_2__END, x)}$$

This quantity does not directly refer to any sublexicon and is in fact compatible with any model of phonotactics (see also Feldman, Griffiths, and Morgan 2009). The question becomes whether a model that does not make explicit use of sublexica can or cannot represent the relative difference in probability between the two vowels in the context $C_1oC_2__END$.

To capture this difference, a model will need to be able to take into account this entire context. That is, it will need to be able to capture the dependency between the word-initial conditioning consonant $C_1$ and the target vowels [a]/[a lengthmark] in the stimuli in the form of $C_1oC_2a()$. Section S3 of the online supporting information shows in detail how our sublexical model made this possible. Many models with limited contextual information will not be able to capture this difference. For example, a single-lexicon trigram model will fail to predict the difference because $C_1$ is out of the window of trigram context of the target vowel, as shown by the following mathematical expression:

$$\log \frac{p_{single}(a | C_1oC_2__END, x)}{p_{single}(a lengthmark | C_1oC_2__END, x)} = \log \frac{p_{single}(a END | C_1oC_2__, x)}{p_{single}(a END | C_1oC_2__, x)}$$

$$= \log \frac{p_{single}(a | C_1oC_2__, x)p_{single}(a | C_1oC_2, x)}{p_{single}(a lengthmark | C_1oC_2__, x)p_{single}(a lengthmark | C_1oC_2, x)}$$

$$= \log \frac{p_{single}(END | C_1oC_2a, x)p_{single}(a | C_1oC_2, x)}{p_{single}(END | C_2, x)p_{single}(a | oC_2, x)}$$

Here, $p_{single}$ denotes the probability under the single-lexicon trigram model (Moreton (2002) argues against single-lexicon bigram models for the same reason). For a single-lexicon $n$-gram model to be able to capture the dependency between $C_1$ and [a]/[a lengthmark], a larger window length $n \geq 5$ is necessary (such that $C_1 = [\text{rr}]$ is included in the context). Such large-order $n$-gram models, however, require many more latent variables than our sublexical trigram model: the number of possible contexts that must be represented in the 5-gram backoff model is $\sum_{d=0}^{10} (# of contexts of length d) = 10,936,367$ when the size of the phonetic inventory is 57 (our data), whereas the number is just 10,098 for the trigram model with three sublexica.

Likewise, many phonotactic models allowing unboundedly long interactions would not be able to explain Moreton and Amano’s (1999) results without assuming sublexica. This is because the models typically impose restrictions on the pairs of phonetic segments that can interact at a distance. In particular, many phonological theories, including tier-based interaction (Goldsmith 1976, Heinz, Rawal, and Tanner 2011, Futrell et al. 2017) and Agreement by Correspondence (Hansson 2001, Rose and Walker 2004), respect the crosslinguistic generalization that long-distance dependencies are limited to phonetically related segments (e.g., anteriority agreement between sibilants in Navajo; Sapir and Hoijer 1967). What Moreton and Amano observed, on
the other hand, was an interaction between phonetically unrelated segments, \([p]/\overline{f}/[r]/\overline{r}\j] and \([a]/\overline{a}\j] \) (see Hayes et al. 2009, Becker, Ketrez, and Nevins 2011 for more general discussion on the possibility of interactions among phonetically unrelated segments, including local interactions). Therefore, a single-lexicon model would not be able to explain the experimental results as long as it respects this crosslinguistic generalization about long-distance dependencies.

Our model predictions also adjudicate between two slightly different phonotactic generalizations made in previous literature. Gelbart and Kawahara (2007) claim that \([a]:\) is representative of the Foreign sublexicon only in word-final positions. Moreton and Amano (1999), on the other hand, state that the representativeness of the long vowel is independent of its position in words. Our results suggest that Moreton and Amano are correct. This can be investigated in more detail in future experimental studies.

In addition to recovering previous observations, we found hitherto unrecognized substrings that are representative of particular sublexica (section 4.3.2). Most importantly, we have produced the first list of phonotactic patterns that are characteristic of the Native sublexicon. Prior literature has not investigated such patterns because it has been assumed that since the set of allowable strings for the Native sublexicon is merely a subset of the sets for the other sublexica, it can possess no characteristic patterns. Our results show this assumption to be unjustified and provide predictions that can be studied in future experimental work. Full lists of 1- to 3-grams together with their representativeness with respect to each sublexicon are available in supporting information table S5.2.

We emphasize that we have demonstrated this sublexicon learnability from phonotactic information alone. In particular, the proposed learning method does not rely on orthographic information—which has sometimes been believed to be indispensable for Japanese sublexicon learning (Gelbart and Kawahara 2007)—or, indeed, any representation particular to Japanese. This implies that our sublexicon-learning method is applicable to any other language. Just as we did with the Japanese data here, researchers can test for the existence of proposed lexical classes in other languages. Morita (2018), for instance, applied exactly the same model to English data and discovered word clusters aligning with the Germanic and Latinate sublexica that have been proposed to underlie the English lexicon (Grimshaw 1985, Anshen et al. 1986, Grimshaw and Prince 1986, Fabb 1988, Gropen et al. 1989, Levin 1993, O’Donnell 2015).

As we discussed in section 1, Japanese sublexica are said to exhibit other (morpho)phonological properties beyond those phonotactic properties captured by our model (see those headed by the white bullet points in table 2). For example, actual learners would also be able to exploit the orthographic differences, which have been considered an essential factor for sublexicon learning (Gelbart and Kawahara 2007, Kawahara 2016). While it remains an open question how our results might change were we to incorporate these nonsegmental properties of the sublexica, previous studies have shown that joint learning of multiple aspects of linguistic structure often leads to more successful learning outcomes, as it can exploit various correlations present in multimodal data (Maurits, Perfors, and Navarro 2009, Johnson et al. 2010, Feldman et al. 2013). These are what researchers call synergies of learning (Johnson 2008, Johnson et al. 2010, O’Donnell 2015). If prosodic, morphological, and orthographic characterizations of the Japanese sublexica are also
exhibited by naturalistic data, they will correlate with the segmental properties covered in this study. Hence, jointly modeling segmental, prosodic, and morphological structures may also help identify the sublexical class of words.

We may also improve the model by pairing the DP prior with a more sophisticated phonotactic model (i.e., likelihood) with feature representation of segments. The current likelihood distribution (called the HDP n-gram backoff model; Goldwater, Griffiths, and Johnson 2006, Teh 2006) does not exploit the inherent phonetic similarities/differences among segments (e.g., similarities observed in acoustics or articulation, or their discrete idealization by phonetic features). Futrell et al. (2017) propose an extension of the HDP n-gram backoff model, which can make feature-based generalizations and was shown to outperform the segmental model with data from diverse languages. Such feature-based generalizations may also improve the learning of sublexica. For example, our version of the sublexical model failed to predict the high degree of Foreign-representativeness of voiced geminates (section 4.3.2), likely due to their low frequency. Feature-based models, on the other hand, are able to represent the natural class of voiced geminates as \{voiced: +, LENGTH:(super)long\}, and by representing this natural class, these problems of data sparsity could be avoided. In our study, (i) the most frequent voiced geminate, [dː], already exhibited high representativeness for the Foreign sublexicon and (ii) words containing the other voiced geminates (besides [Nː]) were also correctly classified into the Foreign sublexicon (on the basis of this cue and other cues present in the words). Thus, the representation of a natural class would allow this statistical information to be pooled across the system.

5 Conclusion

Our study applied a commonly used clustering method, based on a prior-likelihood trade-off, to naturalistic data of Japanese words, and found that the predicted clusters largely corresponded to the etymological sublexica of the language. This result demonstrates that (i) statistical evidence for the sublexica exists in naturalistic Japanese data and thus (ii) the sublexica are learnable by learners that balance a trade-off between a prior bias for a small number of sublexica (simplicity) against a bias for good fit to the data (likelihood). In addition to our classification results, we (iii) recovered most of the previously observed phonotactic properties of the sublexica, (iv) predicted the experimental results of Moreton and Amano (1999), and (v) generated new predictions that can be tested in future experimental investigations.

References

Anshen, Frank, Mark Aronoff, Roy Byrd, and Judith Klavans. 1986. The role of etymology and word-length in English word formation. Ms., IBM Thomas J. Watson Research Center.


Moreton, Elliott, Shigeaki Amano, and Tadahisa Kondo. 1998. Statistical phonotactics of Japanese: Transitional probabilities within the word. Transactions of the Technical Committee on Psychological and


Shalev-Shwartz, Shai, and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge: Cambridge University Press.


**Takashi Morita**  
The Institute of Scientific and Industrial Research  
Osaka University  
tmorita@alum.mit.edu

**Timothy J. O’Donnell**  
Department of Linguistics, McGill University  
Québec Artificial Intelligence Institute (Mila)  
Canada CIFAR AI Chair, Mila  
timothy.odonnell@mcgill.ca