

# A phonotactic/tonotactic grammar for Tokyo Japanese that clusters by lexical strata does not overfit

LSA 2024 presentation

Satoru Ozaki

University of Massachusetts Amherst

`sozaki@umass.edu`

`ikazos.gitlab.io`

January 5, 2024

At least three etymological strata in Tokyo Japanese (TJ):

- (1)
  - a. Native Japanese words
  - b. Sino-Japanese words
  - c. Foreign loanwords

Different strata, different phonotactic and tonotactic properties.

Should we analyze TJ with...

- (2)
  - a. A **non-clustering** grammar that treats all strata equally? Or...
  - b. A **clustering** grammar that can treat the strata differently?

**Result:** Clustering MaxEnt grammars don't overfit.

- (3) a. **Background**
  - i. TJ strata and their properties
  - ii. Two kinds of grammars: non-clustering vs. clustering
  - iii. The model comparison problem
- b. **Study**
  - i. Data
  - ii. Learning MaxEnt grammars
  - iii. Comparing the learned grammars
- c. **Future work & conclusion**

(4) a. **Native Japanese words**

Examples: *kami* 'hair', *tobira* 'door', *madoromi* 'drowse'

b. **Sino-Japanese (SJ) words**

Examples: *sen* 'thousand', *dempa* 'phone signal', *gengogaku* 'linguistics'

c. **Foreign loanwords (loanwords from languages other than Chinese)**

Examples: *pen* 'pen', *piiman* 'bell peppers', *budda* 'Buddha'

Many differences (Frellesvig 2010; Fukuzawa 1998; Gelbart 2005; Gelbart and Kawahara 2007; Ito and Mester 1995a,b, 1999; Moreton and Amano 1999; Morita and O'Donnell 2022):

- (5) a. **No voiceless obstruent after nasals (e.g. \*[nt]) in native words.**  
Examples: SJ *sintai* 'body', foreign *ranku* 'rank'
- b. **Nongeminate [p] only occurs in foreign words.**  
Examples: *pai* 'pie', *apo* 'appointment'
- c. **[ɸa], [ɸi], [ɸe], [ɸo] only occur in foreign words.**  
Examples: *ɸairu* 'file', *ɸinrando* 'Finland', *ɸeruto* 'felt', *ɸorumu* 'form'
- d. **Likelihood of accent (Kubozono 2006, 2011):**  
Native: 29%, SJ: 49%, foreign: 93%.

(6) a. **A non-clustering grammar**

Use a single grammar to predict the distribution of all TJ nouns.

b. **A clustering grammar**

Use one grammar to predict the distribution of TJ nouns in each stratum.

Two questions:

(7) a. **Learnability** ( $\leftarrow$  lots of previous work)

How do you build a clustering learner?

b. **Model comparison** ( $\leftarrow$  this work!)

Which kind of grammar makes a better trade-off between model size and likelihood?

A clustering learner must decide on:

- (8) a. Number of clusters.
- b. Which word belongs to which cluster (i.e., assignment).
- c. The grammar for each cluster.

**Unsupervised learner:** figures everything out by itself (Ito and Mester 1999; Morita and O'Donnell 2022).

**(Semi-)supervised learner:** some information is given (Shaw 2006).

Learners have morphological and orthographic cues to figure out assignment (Gelbart and Kawahara 2007; Ito, Mester, and Padgett 2001).

Each grammar makes a trade-off between:

- (9) a. Maximizing the predicted **likelihood** of the observed data
- b. Minimizing the **number of parameters**

There are quantitative criteria that measures such trade-off, e.g. the **Bayesian Information Criterion** (BIC) (Schwarz 1978).



Does the clustering grammar (with the correct number of clusters and assignment) make a better trade-off than the non-clustering grammar, w.r.t. such criteria?

Specifically:

- |      |  |                |
|------|--|----------------|
| (10) | a. Number of clusters.                                     | <b>Given</b>   |
|      | b. Which word belongs to which cluster (i.e., assignment). | <b>Given</b>   |
|      | c. The grammar for each cluster.                           | <b>Learned</b> |

By giving away (10a) and (10b), I show what performance a learner can achieve in principle.

(11) a. **Data**

Use corpora to build:

- i. The TJ nominal lexicon.
- ii. The native, SJ and foreign sublexicons.

b. **Learning grammars**

Use the UCLA Phonotactic Learner (Hayes and Wilson 2008) to learn:

- i. **A non-clustering grammar.**  
One set of constraints over the entire TJ lexicon.
- ii. **A clustering grammar.**  
One set of constraints over each sublexicon.

c. **Compare the grammars**

Use the BIC to compare the two grammars.

I combine two corpora:

- (12) a. **Balanced Corpus of Contemporary Writtern Japanese** (Maekawa et al. 2013)  
100m words. Provides **etymological stratum** for each word.
- b. **NHK's New Dictionary of Japanese Pronunciation and Accentuation**  
75k words. Provides **accent position** for each word.

This allows me to build (i) a TJ lexicon and (ii) the native, SJ and foreign sublexicons separately.

# Data: phonological representations

I represent each word as (i) a sequence of mora types and (ii) the presence/position of the accent. Five mora types (Vance 2008):

- (13)
- |   |   |            |
|---|---|------------|
| a. <b>V</b> – Optional consonant + vowel  |   |            |
| E.g. / <sub>μ</sub> <b>A</b> <sub>μ</sub> ki/, / <sub>μ</sub> <b>ta</b> <sub>μ</sub> Kl/  | → | Vv, vV     |
| b. <b>Q</b> – First half of a geminate consonant  |   |            |
| E.g. / <sub>μ</sub> na <sub>μ</sub> <b>t</b> <sub>μ</sub> TO <sub>μ</sub> o/, / <sub>μ</sub> ma <sub>μ</sub> <b>p</b> <sub>μ</sub> pu/                          | → | vqVr, vqv  |
| c. <b>N</b> – Moraic nasal  |   |            |
| E.g. / <sub>μ</sub> a <sub>μ</sub> <b>m</b> <sub>μ</sub> pa <sub>μ</sub> <b>n</b> /   | → | vnvn       |
| d. <b>R</b> – Second half of a long vowel   |   |            |
| E.g. / <sub>μ</sub> SE <sub>μ</sub> n <sub>μ</sub> ta <sub>μ</sub> <b>a</b> /, / <sub>μ</sub> to <sub>μ</sub> <b>o</b> <sub>μ</sub> kyo <sub>μ</sub> <b>o</b> / | → | Vnvr, vrvr |
| e. <b>J</b> – Second half of a diphthong  |   |            |
| E.g. / <sub>μ</sub> ga <sub>μ</sub> <b>i</b> <sub>μ</sub> ko <sub>μ</sub> ku/, / <sub>μ</sub> KO <sub>μ</sub> <b>i</b> /  | → | vjvv, Vj   |

**Notation:** lowercase = unaccented, UPPERCASE = accented.

Sequence of feature-value matrices (as required by the UCLA Phonotactic Learner).

Five features for five mora types, one feature for accentedness.

Example: vqNrj

$$(14) \quad \begin{bmatrix} [+v] \\ [-q] \\ [-n] \\ [-r] \\ [-j] \\ [-acc] \end{bmatrix} \begin{bmatrix} [-v] \\ [+q] \\ [-n] \\ [-r] \\ [-j] \\ [-acc] \end{bmatrix} \begin{bmatrix} [-v] \\ [-q] \\ [+n] \\ [-r] \\ [-j] \\ [+acc] \end{bmatrix} \begin{bmatrix} [-v] \\ [-q] \\ [-n] \\ [+r] \\ [-j] \\ [-acc] \end{bmatrix} \begin{bmatrix} [-v] \\ [-q] \\ [-n] \\ [-r] \\ [+j] \\ [-acc] \end{bmatrix}$$

I use the UCLA Phonotactic Learner (Hayes and Wilson 2008).

$d$  = maximum number of learned constraints: 50, 75 and 100.

Two setups, five runs per setup:

(15) a. **Non-clustering grammar**

One set of  $d$  constraints over the entire TJ lexicon.

Likelihood is the likelihood of the entire lexicon.

b. **Clustering grammar**

One set of  $d$  constraints over each sublexicon.

Likelihood is the product of the sublexicon likelihoods.

**BIC:**  $k \log N - 2 \log \mathcal{L}$ , where:

- (16)
- a.  $k$  is the number of parameters, i.e. the number of constraints;
  - b.  $N$  is the number of observations;
  - c.  $\mathcal{L}$  is the likelihood.

$k$  for clustering grammars is 3 times that for non-clustering grammars.

Lower BIC = better trade-off between grammar fit and grammar size.

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

Higher  $d \Rightarrow$  lower BIC. Clustering BIC  $<$  non-clustering BIC.



Given the correct number of clusters and cluster assignments, a clustering MaxEnt grammar for TJ nominal phonotactics/tonotactics makes a better trade-off between likelihood and grammar size than a non-clustering grammar.

**Consequence:** It is theoretically advantageous to analyze the etymological strata as generated by distinct MaxEnt grammars.

(17) **What about empirical results?**

Does a clustering learner actually get to a good grammar?

(18) **A “split” grammar**









Clustering on the phonotactics, no clustering on the tonotactics.






(19) **Try other kinds of representations and grammars**




Add segmental information? Try a n-gram grammar?

Thanks to: Gaja Jarosz, Michael Becker, Shigeto Kawahara!

Thanks to: my reviewers, the LSA 2024 organizers and the audience!

-  Frellesvig, Bjarke (2010). *A history of the Japanese language*. CUP.
-  Fukuzawa, Haruka (1998). “Multiple input-output faithfulness relations in Japanese”. *Rutgers Optimality Archive ROA-260-0698*.
-  Gelbart, Ben (2005). “Perception of foreignness”. *PhD thesis*. UMass Amherst.
-  Gelbart, Ben and Shigeto Kawahara (2007). “Lexical cues to foreignness in Japanese”. In: *MITWPL 55*. Ed. by Yoichi Miyamoto and Masao Ochi, pp. 49–60.
-  Hayes, Bruce and Colin Wilson (2008). “A maximum entropy model of phonotactics and phonotactic learning”. In: *LI 39*, pp. 379–440.
-  Ito, Junko and Armin Mester (1995a). *Japanese phonology*. Ed. by John Goldsmith.
-  — (1995b). *The core-periphery structure of the lexicon and constraints on reranking*. Ed. by Jill Beckman, Suzanne Urbanczyk, and Laura Walsh Dickey.
-  — (1999). *The phonological lexicon*. Ed. by Natsuko Tsujimura.

-  Ito, Junko, Armin Mester, and Jaye Padgett (2001). “Alternations and distributional patterns in Japanese phonology”. In: *Journal of the Phonetic Society of Japan* 5, pp. 54–60.
-  Kubozono, Haruo (2006). “Where does loanword prosody come from? A case study of Japanese loanword accent”. In: *Lingua* 116.7, pp. 1140–1170.
-  — (2011). *Japanese pitch accent*. Ed. by Marc van Oostendorp et al.
-  Maekawa, Kikuo et al. (2013). “Balanced corpus of contemporary written Japanese”. In: *Proceedings of LREC 48*, pp. 345–371.
-  Moreton, Elliott and Shigeaki Amano (1999). “Phonotactics in the perception of Japanese vowel length: Evidence for long-distance dependencies”. In: *Proceedings of the 6th European Conference on Speech Communication and Technology*, pp. 2679–2682.
-  Morita, Takashi and Timothy J. O’Donnell (2022). “Statistical Evidence for Learnable Lexical Subclasses in Japanese”. In: *Linguistic Inquiry* 53.1, pp. 87–120.

-  Schwarz, Gideon (1978). “Estimating the dimension of a model”. In: *The Annals of Statistics* 6, pp. 461–464.
-  Shaw, Jason (2006). “Learning stratified lexicon”. In: ed. by Christopher Davis, Amy Rose Deal, and Youri Zabbal. Vol. 2, pp. 519–530.
-  Vance, Timothy J. (2008). *The sounds of Japanese*. CUP.