Rule-ambiguity and distributional play in asynchronous L2 shiritori

Joseph L. Austerweil[†], Jumpei Nishikawa[‡], Junya Morita[§]

[†] Chiba Institute of Technology, [‡] Okayama Prefectural University, [§] Shizuoka University
joseph.austerweil@gmail.com

Abstract

Shiritori chains words so that each begins with the final mora of the previous (e.g., *sakura* → *rakuda* → *daikon*); a word ending in *n* (ん) loses. The game imposes a dynamic phonological constraint on lexical retrieval. We analyze 9,891 moves from 568 second-language (L2) Japanese learners playing asynchronously on a public web forum, and argue that the phonological constraint is processed at three dissociable levels. (H1) At the level of *content*, higher-proficiency players produce rarer words (Bayesian posterior $P(\beta<0)=1.00$). (H2) At the level of *strategy*, avoidance of trap-ending morae is individually learned from experience but unrelated to proficiency; in contrast, vocabulary sophistication tracks the community, not personal history—a community/individual dissociation. (H3) At the level of *rule interpretation*, when the prior word ends in a yōon mora the rule is formally ambiguous, and players produce a distribution over rule-variants (mora-match 48%, small-kana 7%, base-kana 2%, voicing-confusion 2%) rather than converging on a single match; 90.5% of players contributing ≥ 3 such transitions use multiple variants, arguing against a fixed-but-wrong internal rule. (H4) A multiplex random-walk model jointly weighting semantic and transition-graph layers decisively wins model comparison (ω=0.67; AIC weight .9995), a single computational account for H1–H3.

Keywords: shiritori, lexical retrieval, L2 Japanese, phonological constraint, multiplex network

1. Introduction

Shiritori is a turn-based Japanese word game in which each new word must begin with the final mora of the previous word (e.g., さくら/*sakura* → らくだ/*rakuda* → だいこん/*daikon*), and any word ending in the mora ***n*** (ん) loses. Japanese *mora* is the language's timing unit: roughly a CV pair, with ん (*n*) counting as a mora of its own and losing because no Japanese word begins with it. *Yōon*—the palatalized CV+*ya*/*yu*/*yo* sequences written as a full-size kana plus a half-size ゃ/ゅ/ょ (e.g., きょ = *kyo*)—form a *single* mora spanning two kana. Yōon boundaries are thus rule-ambiguous: 金魚 (*kingyo*, "goldfish") ends in *gyo*, and the next word can legitimately start with *gyo* (whole mora), *yo* (small-kana alone), *ki* (base kana), or—modulo the *dakuten* voicing diacritic (゛, distinguishing e.g. し/*shi* from じ/*ji*)—*kyo*. The game is thus a minimal naturalistic paradigm for *constrained lexical retrieval*: players search semantic memory under a time-varying phonological cue, avoid a dynamically-shifting trap, and satisfy a rule whose formal edges admit several equally "legal" continuations.

Prior semantic-search work treats the retrieval cue as either free (semantic fluency) or categorical (category-cued fluency); shiritori adds a phonological self-conditioning rule linking search-on-a-semantic-network accounts (Abbott et al., 2015) with multiplex lexical networks (Vitevitch, 2008).

Shiritori has been modeled with ACT-R cognitive architectures for phonological-awareness formation and error (Nishikawa & Morita, 2022) and for activation- and association-based factors shaping sequence length and coherence (Nishikawa et al., 2025); ours is a complementary distributional/network-level account on naturalistic data, with parameters qualitatively tracking those mechanisms (§ 4, § 5).

We analyze a public corpus of asynchronous play by L2 Japanese learners on the Bunpro community forum—a useful lens: turns span minutes to days, removing time pressure; players self-report a proficiency index; and the forum's posted rules explicitly acknowledge multiple legal yōon continuations, exposing rule-variant behavior that synchronous play would drown out.

We argue the phonological constraint is processed at three dissociable levels—content (H1), strategy (H2), and rule interpretation (H3)—and show a single multiplex-random-walk model (H4) accommodates all three.

2. Data and setting

Bunpro is a paid spaced-repetition app for Japanese grammar; its proficiency level is an internal 1–101 scale tied to grammar-point mastery, loosely aligned with JLPT N5–N1. The Bunpro community forum (community.bunpro.jp) hosts three long-running shiritori threads with data licensed CC-BY-NC-SA 3.0. We scraped the public Discourse JSON API and parsed 9,891 valid game moves from 568 unique users between October 2019 and February 2026. Each user has a self-reported Bunpro proficiency level scraped at analysis time; the level is a snapshot, not a post-time value, which we treat as a known-bias caveat rather than a correctable covariate. Turn cadence is asynchronous (median gap of hours), and the forum explicitly posts three permissible yōon continuations: match the full mora (きょ → きょ . . . , *kyo*→*kyo*-. . .), match the small-kana alone (きょ

→ よ..., *kyo*→*yo*-...), or match the base kana (きょ → き..., *kyo*→*ki*-...). This permissiveness is atypical for native play but exposes rule-variant navigation to measurement.

3. Three levels of processing

3.1 Content: proficiency → word rarity (H1)

Using a Bayesian mixed-effects re-estimation (bambi/PyMC; random intercepts and slopes by user, thread as a fixed covariate, position-in-game as a continuous covariate), proficiency negatively predicts Zipf word frequency: posterior mean $\beta_{\text{level}} = -0.004$, 95% HDI $[-0.006, -0.002]$, $P(\beta<0) = 1.00$, $\widehat{R}\leq 1.01$. Higher-proficiency players produce measurably rarer words (Figure 1). This naturalistically recapitulates Laufer & Nation (1995)'s classical lexical-breadth result.

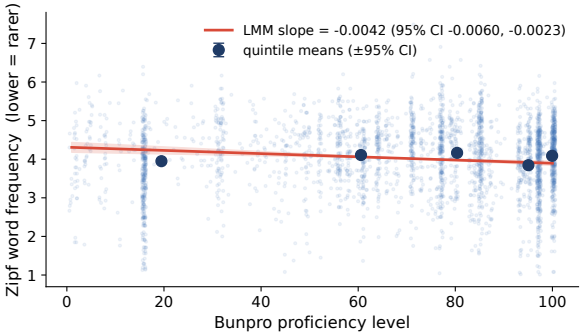

Figure 1 Word-rarity vs. proficiency. Blue dots: per-move Zipf frequency (jittered, subsample). Dark points: quintile means with 95% CIs. Red line: the LMM fixed-effect slope ($\beta = -0.004$, 95% CI $[-0.006, -0.002]$). Lower Zipf = rarer word.

3.2 Strategy: community-driven vocabulary vs. individually-learned trap avoidance (H2)

We fit two parallel mixed models (following Barr et al. (2013)) that decompose a player's move-level outcome into an *individual-experience* slope and a *community* exposure term. Individual experience is the player's own cumulative move index across the corpus (pooled across the three forum threads of § 2, min–max scaled within player so the slope is identified within-player; between-player proficiency differences are absorbed by user random intercepts and slopes). Community exposure is the monthly mean of the outcome across all players' moves that month—the register a new move is entering. The two outcomes show a clean dissociation (Figure 2).

For *vocabulary sophistication* (Zipf frequency of the produced word, lower = rarer), the community slope is strongly negative and highly reliable ($\beta_{\text{community}} = 0.683$, $p = 2.2\times10^{-10}$) while the individual slope is indistinguishable from zero ($\beta_{\text{experience}} = -0.019$, $p = .84$). Players do not become more sophisticated over their own participation; they match the current community register.

For *trap-mora avoidance*—a binary outcome: did the player's word avoid morae whose continuations disproportionately chain toward "ん"-ending follow-ups? Trap status is the (as-initial)/(as-final) ratio for a mora in our L2-player corpus; the L2 trap set differs from what Japanese phonotactics predicts for L1 play. Classically rare-initial morae in Japanese—notably り (Murata & Isahara, 2002)—do *not* behave as traps here: り appears *more* often as initial than as final (ratio ≈ 1.24), likely because L2-accessible vocabulary (鳥 *tori*, 利用 *riyō*, 料理 *ryōri*) reseeds り as initial, whereas the rare-initial status in Murata & Isahara (2002) rests on a much broader native lexicon. The highest-trap morae for *L2 players* are instead ん (*n*, ratio .016), よ (*yo*, .295), and ー (the long-vowel mark *chōonpu*, .024); in L2 play these appear often word-finally but rarely word-initially, structurally cornering the next player. The L1/L2 divergence is itself a prediction: trap structure depends on the population's accessible vocabulary, not on Japanese per se. A GEE logistic regression finds the opposite pattern to vocabulary: the individual-experience slope is reliably negative (odds ratio =0.272, $p = .010$)—a ~73% within-player reduction over participation. Players individually learn to avoid traps; vocabulary sophistication comes from the room.

Proficiency itself does not predict trap avoidance: the posterior on the level slope sits at conventional reliability's edge ($P(\beta<0) \approx .95$); we decline to interpret it as a proficiency effect.

3.3 Rule interpretation: yōon transitions are distributional (H3)

For every same-game consecutive pair whose prior word ends in a yōon mora ($n = 252$), we classify the continuation into one of five variants (Table 1). Mora-match is the plurality but not the majority: 48.0% of yōon-context transitions invoke an alternative legal continuation or break the chain.

Table 1 Yōon-transition variant categories (n=252). Prior-word final mora and next-word initial mora shown after the dot.

Variant	Rule	Example (prior → next)	*n*
mora-match	match full CV+small kana	領収書・しょ → 証明・しょ (*ryōshūsho* "receipt" → *shōmei* "proof")	121
small-kana	match small kana alone	金魚・ぎょ → 夜・よる (*kingyo* "goldfish" → *yoru* "night")	18
base-kana	match base consonant kana	労働者・しゃ → 師匠・し (*rōdōsha* "worker" → *shishō* "master")	6
voicing	match modulo dakuten	神社・じゃ → 社宅・しゃ (*jinja* "shrine" → *shataku* "co. housing")	6
other	chain break, sound-not-script, etc.	—	101

Against an error account. Of players contributing ≥ 3 yōon transitions, 19 of 21 (90.5%) used at least two

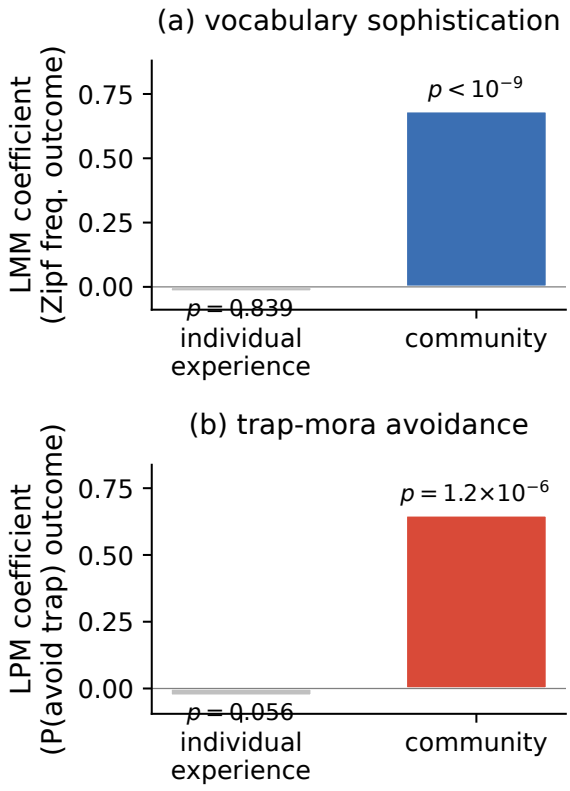

Figure 2 Community-vs-individual dissociation. "Community" is the monthly mean of the outcome across all players; "individual experience" is the player's own cumulative move index, min–max scaled within player. **(a)** For vocabulary sophistication (Zipf frequency), only the community exposure term is reliably non-zero; individual experience is not. **(b)** For trap-mora avoidance (LPM shown for scale-matched display; the GEE logistic OR is 0.272, p=.010), the pattern is reversed: individual experience is the load-bearing term for strategy, as it is not for vocabulary.

distinct classification variants across their own moves. That is, the same player is not consistently picking (say) small-kana continuations; players sample from the variant distribution within themselves. This is inconsistent with an account on which non-mora-match outcomes reflect a fixed-but-wrong internal rule for yōon boundaries; it is consistent with genuine ambiguity navigation, where the posted forum rules legitimize multiple variants and players move through them move by move, and echoes qualitatively the partial-match mechanism by which mora-level phonological variants arise in an ACT-R account of shiritori (Nishikawa & Morita, 2022).

The "other" 40% is heterogeneous: chain breaks, semantic leaps, and sound-not-script matches (e.g., 投手・しゅ → 日本語・に, *tōshu* → *nihongo*; 金魚・ぎょ → 鰻・う, *kingyo* → *unagi*). A principled decomposition awaits finer annotation.

3.4 Joint summary

The three levels are dissociable: proficiency indexes content; individual experience indexes strategy; within-player heterogeneity indexes rule interpretation. None of the three is reducible to the others in these data: proficiency does not predict trap avoidance; individual experience does not predict vocabulary rarity; neither proficiency nor cumulative experience predicts yōon-variant choice (not shown). The next section asks whether a single computational account can capture all three.

4. A multiplex-network synthesis (H4)

Each move is a step on a censored random walk over a multiplex lexical network with three layers: an *orthographic-neighborhood* layer (Vitevitch-style kana edit-distance neighborhoods), a *semantic* layer (multilingual-e5 embedding (Wang et al., 2024) k-nearest neighbors, validated against JWSAN similarity norms), and a *transition* layer (directed shiritori legal-next-move graph, induced by mora rules). The orthographic layer is a direct port of the English Vitevitch construct to kana, and coarser than the mora-level distinctive-feature similarity used in Nishikawa & Morita (2022)'s ACT-R shiritori model; we return to this limitation below. At each step the walker chooses a next word by mixing layer neighbors; moves that violate the shiritori rule are down-weighted by a small residual probability ϵ, which absorbs genuine rule-variant play as noise. Formally, the walker's next-move distribution is a mixture over layers weighted by a parameter ω, censored by the shiritori rule (illegal next-moves receive probability $\epsilon/|V|$, constrained $0 \le \epsilon \le 0.3$), with a Zipf-frequency bias term α and an exponential memory decay λ. We fit five variants by MLE (L-BFGS-B) on 7,298 moves and compare by AIC: Model A (transition only), B (transition + semantic), C (transition + semantic + orthographic), D (orthographic only), and Uniform (no structure).

Model B (transition + semantic) wins decisively: AIC weight .9995; $\omega_{\text{transition}} = 0.674$, implying $\omega_{\text{semantic}} = 0.326$; $\lambda = 0.10$ (slow decay); $\alpha = -0.055$ (a *negative* frequency bias, consistent with the community-driven rare-word norm documented in § 3.2); and $\epsilon = 0.30$ saturated at its ceiling. Adding the orthographic layer (Model C) *hurts* (ΔAIC = +15); removing both layers (Uniform) is much worse (ΔAIC = +45); an orthographic-only walker is the worst fit (ΔAIC = +254). The transition layer (which is itself mora-rule-conditioned) supersedes kana-edit-distance neighborhoods (Figure 3); we discuss below why this is not a general phonology-null.

The ϵ parameter saturates at 0.30, the upper bound set by our prior, and functions as a uniform rule-error residual. This absorbs the yōon-variant distribution reported in § 3.3 without structurally representing it: the model

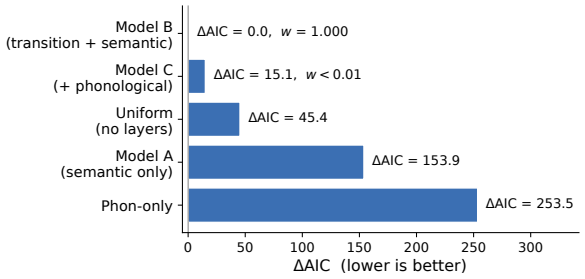

Figure 3 Multiplex-network model comparison (ΔAIC, lower is better). Model B (transition + semantic) wins decisively; adding kana-edit-distance orthographic neighborhoods (Model C) hurts.

accommodates but does not explain H3's sub-structure. A layered rule-variant representation (small-kana and base-kana as separate sub-edges in the transition graph) would make ϵ predictive rather than a catch-all.

The loss of Model C to Model B is narrower than it appears, not a general phonology null. Our orthographic layer is a kana-string edit-distance measure: a direct import of the Vitevitch-style construct from English orthographic-neighborhood work. For Japanese, a principled phonological-similarity construct operates at the *mora* level, with each mora represented as a vector of distinctive features compared by cosine similarity—the approach used in the ACT-R shiritori model of Nishikawa & Morita (2022), where it is designed to capture within-mora confusions that produce rule-violating *production* errors (e.g., voicing confusions し ↔ じ). In a text-based asynchronous forum, players type and re-read before posting, so production errors of that kind are rare; neither the mora-level construct nor ours should be expected to carry much of the predictive load here. Integrating Nishikawa's mora-level distinctive-feature similarity as the phonological layer is an important direction for future work: it is the construct that actually matches the phenomenon in synchronous or spoken shiritori, and the natural path to a principled combined account whose layer weights could be read against the ACT-R activation dynamics of Nishikawa & Morita (2022), Nishikawa et al. (2025).

A further limitation of H4 is exposed generatively. When we simulate games from the fitted Model B, simulated chains terminate in ん-endings at ~80% vs. ~4.5% in the observed corpus, and average simulated game length is 16.3 moves vs. 26.2 observed. Real players appear to apply a two-stage procedure—generate candidates via network navigation, then screen against ん-endings before committing—that the single-stage random walk cannot replicate. The dissociation between *individually-learned* trap avoidance (§ 3.2) and *community-driven* vocabulary (§ 3.2) suggests individual experience drives the screening stage.

5. Discussion

First, the asynchronous L2-forum setting is a feature, not a limitation: it removes time pressure, thereby separating lexical-retrieval *content* from the real-time cadence effects that ordinarily bundle proficiency, strategy, and rule-interpretation together. Second, the three-level dissociation (H1/H2/H3) is post-hoc: we report these as findings from exploratory analyses, not as pre-registered predictions. Third, we cannot adjudicate between the distributional and error-based accounts of yōon variation from L2 data alone; the within-player heterogeneity result (§ 3.3) is our strongest evidence that it is not pure error, but the cleanest test is a matched-rule comparison with a different player population. That comparison remains future work.

Two caveats qualify these claims without undermining the qualitative dissociations: the Bunpro corpus selects for motivated self-study L2 learners (limiting generalization), and the proficiency snapshot noted in § 2 means a player who leveled up mid-window appears artificially stable within themselves—inflating apparent within-player proficiency consistency, but not between-player comparisons. Finally, the network-level account offered here and the ACT-R accounts of Nishikawa & Morita (2022), Nishikawa et al. (2025) make different commitments—population-level fit vs. process-level simulation—and are most informative read together.

References

Abbott, J.T., Austerweil, J.L., & Griffiths, T.L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122* (3), 558–569.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68* (3), 255–278.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16* (3), 307–322.

Murata, M., & Isahara, H. (2002). A quantitative investigation of shiritori (*shiritori ni kansuru sūryoteki chōsa*). *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing (NLP2002)*, pp. 49–52. Japan: In Japanese.

Nishikawa, J., & Morita, J. (2022). Cognitive model of phonological awareness focusing on errors and formation process through Shiritori. *Advanced Robotics*, *36* (5–6), 318–331.

Nishikawa, J., Sasaki, K., & Morita, J. (2025). A cognitive model of the factors controlling the characteristics of Shiritori word sequences. *Proceedings of the 47th Annual Conference of the Cognitive Science Society*, pp. 2945–2951.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval?. *Journal of Speech, Language, and Hearing Research*, *51* (2), 408–422.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., & Wei, F. (2024). Multilingual E5 text embeddings: A technical report. arXiv:2402.05672.