Interaction of lexical strata in hybrid compound words through gradient phonotactics

Eric Rosen (University of British Columbia) and Matthew Goldrick (Northwestern University)

Japanese morphemes fall into several subclasses, known as lexical strata (Itô and Mester 1995). The phonological and morphological distinctiveness of these strata are a productive part of the synchronic knowledge of Japanese speakers (Smith and Tashiro 2019). In this study, we show how words composed of morphemes from two different lexical strata ('hybrid compounds') gradiently resist marked phonotactic patterns through the influence of the phonotactics of one stratum on another. We show that certain marked phonotactic patterns that are banned in native[1] words are avoided by hybrid words to a greater degree than by purely foreign words, suggesting an influence of the native on the foreign morpheme.

NHK (1999) lists 825 words that can be analysed as foreign/non-foreign hybrid compounds, where foreign morphemes are identifiable by their katakana orthography. The table below shows the frequency per character/bigram **in foreign parts only** of three Markedness violations for the two sets: Lyman's Law (no multiple voiced obstruents in same morpheme), *di/ti (no unaffricated coronal stops before /i/) and *f (no labial/labiodental fricative other than before /u/). The highlighted numbers show that these violations are significantly ($p < 0.016$) less frequent in hybrid compounds than in pure foreign words.

VPC = violations per character/bigram					
Constraint	VPC foreign	VPC hybrid	$\frac{\text{hybrid}}{\text{foreign}}$	$\chi^2_{df=1}$	p-value
Lyman	0.017	0.010	0.569	11.71	0.001
*di/ti	0.003	0.001	0.241	5.84	0.016
*f{a,e,i,o}	0.004	0.002	0.455	6.06	0.014

To account for these patterns, one could adopt an analysis where constraints are indexed to strata. However, this would require that constraints for one stratum can apply to an adjacent heterostratal morpheme, which would be problematic in light of arguments by Pater (2007) and Round (2017) that allowing constraint indexing to be non-local makes incorrect predictions (e.g., Pater 2007:17). Our analysis instead relies on graded degrees of well-formedness arising from graded phonotactic probability. Following Mayer and Nelson (2019) we adopt a model of gradient phonotactics in which the occurrence of a phoneme in a word is assigned a probability based on all the phonemes that precede it. This leads to a bias in compounds, where the phonotactics of the first morpheme influence the probability of the second.

We estimate phonotactic probabilities using a recurrent neural network (Elman 1990, Mayer and Nelson 2019). The inputs at each timestep are (a) a feature-based vector representation of the previous phoneme and (b) the output of a hidden layer at the previous timestep, each passed through a linear transformation with an added nonlinearity to a hidden layer. The hidden layer is passed through another linear transformation to yield an output vector which is softmaxed to give a probability distribution over possible phonemes at that timestep.

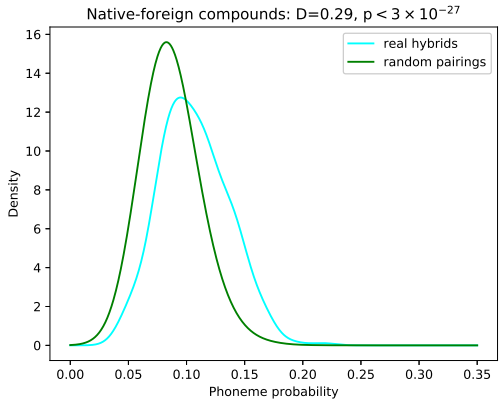
Rather than a discrete delineation between lexical strata, the resulting phonotactic probabilities define a gradient continuum of allowed markedness among words: what Smith and Tashiro (2019:2), citing Kiparsky (1973) refer to as a "hierarchy of foreignness", similar to Itô and Mester (1995)'s 'core-periphery structure'. For example, our model assigns a per-phoneme exponentiated average probability, calculated as $\exp(\frac{1}{n}\sum_i^n \text{logprob}(x_i))$, x_i= the ith phoneme, to foreign word *diiraa* 'dealer' of only 0.028, where /di/ is marked and /aa/ is infrequent, but 0.205 to *koodo* 'cord; code' which is homophonous with several non-foreign words.

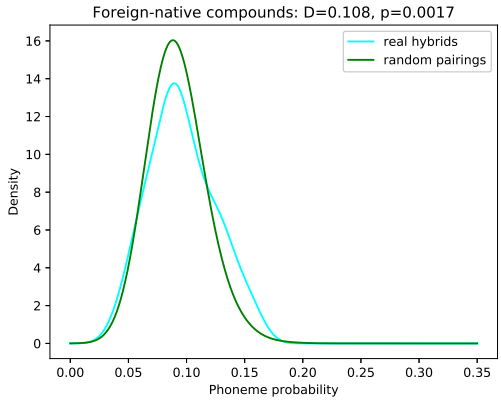
Our model predicts the observed lower per bigram frequency of /fi/ in hybrid compounds than in pure foreign words because, for example, in a native+foreign hybrid, a sequence of phonemes typical of a native word preceding the /fi/ sequence will make /fi/ less likely than if it were preceded by a sequence typical of a foreign morpheme. In hybrid compound *funensee-firumu* 'noncombustible film, safety film', our model gives probabilities of 0.0019 and 0.0027 to the first

[1]Here, we collapse Yamato and Sino-Japanese words into one 'native' stratum.

two phonemes, /f/ and /i/ in *firumu* 'film'. But in purely foreign word *karaa-firumu* 'colour film', assigned probabilities of the /f/ and /i/ are 0.0030 and 0.0686, where the /i/ following the /f/ is 36 times more probable than in the hybrid.

To examine if these patterns held more broadly across the set of compounds, we utilized a Monte Carlo procedure to assess if hybrid compounds have graded phonotactic probability that is higher than predicted by chance. For each of the foreign+native and native+foreign compound sets, we created 1000 sets of random pairings, where the first conjunct is taken from each of the real hybrids and the second is chosen randomly from the set of words in the opposite stratum. The model, pre-trained on 4 epochs of the entire 78,000-word NHK lexicon, calculates the probability of each phoneme in each set. We then did a Kolmogorov-Smirnov test to compare the probability distributions of the real and randomly-paired hybrids and found a significant difference within each pair of distributions, as shown in the graphed kernel density estimates below.





The difference within each pair of distributions indicates that in hybrid compounds there is a bias towards selecting morphemes from the two strata such that the phonotactics of the first do not make the phonotactics of the second too improbable. For example, *funenseefirumu* 'safety film' mentioned above has the second lowest exponentiated average phoneme probability among native+foreign hybrids of 0.0452, but in the randomly paired nonexistent compound *kao-ueitaa* 'face-waiter' it's only 0.0148: well below anything seen in real hybrids.

In summary, our model measures phonotactic well-formedness by calculating phoneme probabilities, and in so doing, makes lexical strata differentiation continuous rather than discrete. Moreover, without requiring constraint indexation, it provides an account of our observed intra-word interactions between morphemes of different strata.

References Jeffrey L. Elman. 1990. Finding structure in time. Cognitive Science, 14(2):179–211. ⊙ Junko Ito and Armin Mester. 1995. The core-periphery structure of the lexicon and constraints on reranking. In J. Beckman et al (eds.), Papers in Optimality Theory, 181–209. Amherst: GLSA. ⊙ Paul Kiparsky. 1973. How abstract is phonology? In O. Fujimura, ed., *Three dimensions of linguistic theory*. 5–56. Tokyo: Tokyo Institute for Advanced Studies of Language. ⊙ Connor Mayer and Max Nelson. 2020. Phonotactic learning with neural language models. Proceedings of SCiL. Vol.3. Article 16. ⊙ NHK (1999). *NHK Hatsuon Akusento Jiten* (NHK Pronunciation and Accent Dictionary). Japanese Broadcasting Corporation. ⊙ Joe Pater. 2007. The Locus of Exceptionality: Morpheme-Specific Phonology as Constraint Indexation. In U. Mass. Occasional Papers 32: Papers in Optimality Theory III. Leah Bateman et al, eds. pp. 259-296. ⊙ Erich R. Round. 2017. Phonological exceptionality is localized to phonological elements: the argument from learnability and Yidiny word-final deletion. In C. Bowerns et al, eds. *On looking into words (and beyond)* 59-98. Berlin: Language Science Press. ⊙ Jennifer Smith and Yuka Tashiro. 2019. Nonce-loan judgments and impossible-nativization effects in Japanese. Proceedings of the LSA 4. 26:1-14.