

Strata Classification by variegated similarity: the case of Sino-Japanese compounds

Since McCawley's (1968) phonological grammar of Japanese, researchers have recognized that the Japanese lexicon is divided into at least three major lexical subclasses, or strata, including Sino-Japanese (SJ), Yamato Japanese (YJ) and Non-Chinese Foreign (F), each with its own phonological properties. Ito and Mester (1995, 1999) have developed a model that captures the differences between these sub-classes in terms of overlapping markedness constraints. The model, dubbed the "core-periphery" model, offers a view of these sub-classes as stages in nativization, while providing a principled mechanism for identifying strata membership within the synchronic grammar. Formalized into an OT framework, the grammar allows faithfulness constraints to be indexed to particular lexical items or classes of items, extending the notion of correspondence laid out in McCarthy and Prince (1995), but effectively treating exceptions to non-indexed faithfulness constraints.

Opponents of the core-periphery model argue that allowing faithfulness constraints to refer to a subset of the lexicon unnecessarily complicates the grammar while over-generating lexical subclasses (Rice 1997, Ota 2004). In addition, several scholars have noted exceptions to the sub-classes produced by the grammar (Tateishi 2003, Fukazawa and Kitahara 2002) and Ota (2004) identifies learnability problems for these exceptions. Tateishi (2005) argues against the synchronic formalization of lexical strata, claiming that phenomena described by class-specific faithfulness constraints within the core-periphery model can be expressed by more general constraints on the phonological and morphological structure of Japanese. None of these arguments, however, consider the data presented here, a direct comparison of SJ compounds with YJ compounds.

The class of compounds referred to as Sino-Japanese, are composed of roots that historically derive from Chinese (Tateishi 1990), and, as a group, violate phonotactic constraints, such as *NT, that are typically obeyed by Yamato Japanese vocabulary. Since SJ roots undergo different phonological processes during compounding (Ito and Mester 1996, Kuriso 2000), this lexical stratum appears to have reality beyond the realm of phonotactics. Ito and Mester (1996) identify a generalization that captures the differences between the respective compounding paradigms in the SJ and YJ compounds in (1) and (2), dubbing the SJ operation "root fusion" and the YJ operation "root spreading."

The puzzling fact about SJ compounds is that even those that do not provide any overt phonotactic evidence for Sino-Japanese-ness, are formed using root fusion. These roots behave like all other SJ roots (and never like YJ roots) with respect to the phonological operation involved in compounding, while obeying the phonotactic constraints of the YJ stratum. Thus, the data require a mechanism to determine the phonological output of compounds that is independent of phonological structure. Although the correlation between Sino-Japanese-ness and root fusion is perfect, it is a mystery as to how the grammar knows what is Sino-Japanese.

My solution to this problem derives from a hypothesis that knowledge of Sino-Japanese-ness and, thus, when to use root fusion to form compounds, comes from the phonology even though it is not based on a uniform structured similarity. Hayes and Albright (2003) use the term "variegated similarity" to describe when a form is similar in different ways to various comparison forms. In order to test the variegated similarity hypothesis, I implemented a computational model of YJ and SJ compounds based on Skousen's Analogical Model (1989). In the model, compounds are represented by variables for distinctive features, word length (in mora), and (mora-sized) lexical chunks and associated with a lexical sub-class ($N_{SJ} = 20$; $N_{YJ} = 18$). The assumption is that these exemplars are classified (or indexed to the same faithfulness constraints) as SJ or YJ by the phonology and form the anchor for their respective strata. As is illustrated by the data in (1), however, not all compounds that behave like SJ vocabulary can be unambiguously identified by the phonology. These vocabulary enter the lexicon underdetermined for lexical class, but, in this model are immediately classified based upon variegated similarity to stored exemplars.

The results of a simulation for five novel compounds that are phonotactically ambiguous between SJ and YJ are given in (3). Despite relying on an impoverished exemplar set, the model identifies the variegated similarities between the novel and exemplar compounds, successfully predicting the correct lexical subclass and allowing for productive use of SJ roots. In modular terms, the work performed by analogy in this model cannot be simply characterized as morphological or phonological, but, rather, exists in the periphery of these systems, an area left untouched by the debate between instance-based/neural network vs. rule-based learning models. Analogy here functions to organize the lexicon, matching phonological forms to lexical sub-classes that are independently required and determined by the phonological grammar.

(1) Sino-Japanese Compounds

- a) hatu + tatu = hattatu *hatutatu ‘development’
- b) hatu + hyoo = happyoo *hatuhyoo ‘presentation’
- c) hatu + sya = hassya *hatusya ‘departure’
- d) hatu + den = hatuden *hadden ‘generation’
- e) hatu + ba = hatuba *habba ‘start’
- f) hatu + mei = hatumei *hammei ‘invention’
- g) gaku + koo = gakkoo *gakukoo ‘school’
- h) gaku + to = gakuto *gatto ‘student’

(2) Yamato Japanese Compounds

- a) butu + kakeru = bukkakeru *butukakeru ‘cover violently’
- b) butu + naguru = bunnaguru *butunaguru ‘beat forcefully’
- c) tuku + kakaru = tukkakaru *tukukakaru ‘turn on’
- d) hiku + tuku = hittuku *hikutuku ‘stick to’
- e) hiku + haru = hipparu *hikuharu ‘jerk/pull’
- f) tuku + nomeru = tunnomeru *tukunomeru ‘fall forward’

(3) Results of the categorization of 5 novel compounds

The column on the left lists the compounds that were compared to the 38 stored exemplars. The # column lists the number of variables in context (supracontexts in the language of Skousen 1989) shared by members of the lexical class; % lists the present shared by the lexical class. *Root fusion*, by definition (Ito and Mester 1996), is blocked by voicing. Thus SJ are divided amongst null and root fusion processes.

Novel Compound	# and % of shared variables in context by outcome					
	Sino-Japanese				Yamato Japanese	
	null		root fusion		root spreading	
	#	%	#	%	#	%
gakuto (SJ)	42428	94.98%	2208	4.94%	36	0.08%
hakka (SJ)	12	0.08%	15108	99.92%	0	0.00%
gakuho (SJ)	205224	94.41%	12160	5.59%	0	0.00%
tukkaesu (YJ)	0	0.00%	704	0.03%	2502632	99.97%
tupparu (YJ)	112	0.04%	24	0.01%	256718	99.95%

References

Anttila, Arto. (2002) Variation and Phonological Theory. *The Handbook of Language Variation and Change* Chambers, Trudgill and Schilling-Estes (eds.) Blackwell Publishers, Cambridge, MA. 206-243

Albright, A. and B. Hayes. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90, 119-161.

Ito, Junko and Armin Mester (1995) “Japanese Phonology.” in J. Goldsmith (ed.). *The Handbook of Phonological Theory*, Blackwell Publishers, Cambridge, MA, pp. 817-838.

Ito, Junko and Armin Mester (1996) “Stem and Word in Sino-Japanese,” in T. Otake and A. Cutler (eds.), *Phonological Structure and Language Processing: Cross-Linguistic Studies*, Mouton de Gruyter, Berlin and New York, pp. 13-44

Ito, Junko and Armin Mester. (1999) The phonological lexicon. In Natsuko Tsujimura (ed.), *The handbook of Japanese linguistics*. 62-100. Cambridge, MA: Blackwell.

Fukazawa, Haruka and Kitahara, Mafuyu (2002). “Ranking Paradox in Consonant Voicing in Japanese,” ms., Kyushu Institute of Technology and Yamaguchi University.

Kuriso, Kazutaka. (2000) Richness of the Base and Root fusion in Sino-Japanese. *Journal of East Asian Linguistics* 9, 147-185.

McCarthy, John and Alan Prince (1995) “Faithfulness and reduplicative Identity,” in J. Beckman, L. Walsh-Dickey, and S. Urbanczyk (eds.), *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*, pp. 249-384.

McCrawly, James (1968) *The Phonological Component of a Grammar of Japanese*, Mouton, The Hague.

Ota, Mitsuhiko. (2004) The learnability of the stratified phonological lexicon. *Journal of Japanese Linguistics*, 20

Rice, Keren. (1997) Japanese NC clusters and the redundancy of postnasal voicing. *Linguistic Inquiry* 28. 541-551.

Skousen, Royal (1989). *Analogical modeling of Language*. Dordrecht: Kluwer Academic Publishers.

Tateishi, Koichi (2005). “Class-Specific” Phonology without reference to Lexical Subclasses. Handout: *Linguistic Society of America 2005 Annual Meeting*. Oakland, CA.

Tateishi, Koichi (1990). Phonology of Sino-Japanese Morphemes. *University of Massachusetts occasional papers in linguistics*. 13, pg 209-235.