Predicting semi-regular patterns in morphologically complex words

Eric Rosen

Johns Hopkins University
errosen@mail.ubc.ca

Abstract

We expect generative models of language to correctly predict surface forms from underlying forms, but morphologically complex words, especially compounds, can exhibit idiosyncratic outputs, which require an extra lexical listing. This results in (a) a poorer Minimum Description Length of our model (Goldsmith 2011) and (b) failure of a grammar to capture patterning among exceptions. To solve an instance of this problem, we examine pitch-accent patterns of 2-mora-2-mora Japanese Yamato (native) noun-noun compounds, hitherto considered semi-predictable but which show gradient tendencies among constituents to trigger a particular accent pattern. In the framework of Gradient Symbolic Computation (Smolensky and Goldrick 2015), a type of harmonic grammar which allows partially activated feature values and weighted constraints, such gradient patterns can be captured through the additive combination of coalescing features on each conjunct, which results in a pitch accent when the summed activations surpass a threshold determined by the grammar. The ability of this framework to completely predict these semi-regular patterns holds promise that it can also explain similar kinds of patterns in other languages.

Keywords: Gradient Symbolic Computation, pitch-accent, lexicalization, Minimum Description Length, predictability

1 Introduction

Generative models of language are designed to predict surface forms through the action of a grammar on underlying forms. Ideally, outputs should be completely predictable from their inputs. When outputs consist of several morphemes, each with its own lexical entry, the grammar does not always correctly predict the exact form of those morphologically complex outputs. This is particularly true for compound words, which often exhibit idiosyncrasy not just in meaning but in phonological shape relative to the shapes of the constituents. When the grammar is unable to predict output forms of complex items, they need to be lexically listed beyond the individual specifications of the constituent morphemes. Such a situation weakens a model of language in two ways: (a) it achieves a poorer Minimum

Description Length (Goldsmith 2011) than if those outputs were predictable without lexical specification; (b) exceptions to predictable phonological processes often occur in a gradient fashion (Zuraw 2000, Coetzee and Pater 2008), where the patterning of exceptional outputs is not completely random but follows identifiable tendencies. Lexical listing of exceptions will arguably account for only categorical but not gradient patterning, and therefore misses important generalizations about the patterns of that language.

To illustrate, we examine pitch-accent patterns of two-member, 2μ-2μ Japanese Yamato (native) noun-noun compounds, hitherto considered at best semi-predictable (Kubozono and Fujiura 2004) Although some second conjuncts (henceforth 'N2s') predictably determine compound accent through their status as preaccenting (accent precedes morpheme boundary) (e.g. *tori* 'bird': *yamá-dori* 'mountain bird'), accent-keeping (e.g. *túbu* 'granule': *kome-túbu* 'grain of rice') or deaccenting (e.g. *tabí* 'trip, *hito* 'person': *tabi-bito* (unacc.) 'traveller'), many other N2s show accenting tendencies that are predisposed in a certain direction but which do not trigger the same accent pattern in all compounds. For example, *otó* 'sound' deaccents in 7 out of 10 compounds and postaccents in the other three. Noun *monó* 'thing' deaccents in 14 N-N compounds, but preaccents in another 12: e.g. *nabé-mono* (preacc.) 'hot-pot' vs. *sina-mono* (unacc.) 'merchandise'. And noun *miti* 'path' preaccents in 8 of 12 compounds and deaccents in the other four. (Data from *Nippon Hoosoo Kyokai* 1998 [Japanese Broadcasting Corporation][1]) We thus find a continuum of accenting tendencies among nouns that form compounds, from nouns that always preaccent or postaccent to nouns that always deaccent, with other nouns at various intermediate positions in between. A model that either lexically specifies compound accent or lexically marks a noun as deaccenting or preaccenting does not account for gradient tendencies of nouns like *oto*, *mono* or *miti* to trigger compound accent in variable ways.

As suggested by an anonymous reviewer, we show graphically, in the two ternary plots below, the gradient accenting behaviour in compounds of the N1s and N2s in the database. In the first plot, each colour-coded dot represents the accenting behaviour of a particular N1. Each corner of the triangle represents an absolute tendency of an N1 to trigger unaccenting, preaccenting or postaccenting in a compound. N1s that occurred only once or twice in the database and with the same accent pattern were given an perturbed position along with some random noise in the direction of the probability that they could occur with a different accent pattern if they occurred in more compounds. For example, an N1 that only occurred once and in an unaccented pattern was moved in the direction of the preaccented corner according to the frequency of preaccenting by N1s that occur in at least one unaccented compound and in the direction of the postaccented corner according to the frequency of postaccenting by N1s that occur in at least one unaccented compound. Therefore, infrequently occurring N1s that show absolute, non-gradient tendencies towards one accent pattern are separated and distinguishable in the plot. Non-gradient N1s are represented by open circles and gradiently-behaving N1s by solid circles. N1s that occur in two out of three possible patterns occur along the sides of the triangle. The plot shows that the majority of the N1s behave gradiently. The second plot shows the behaviour of N2s in exactly the same way.

[1] The accent patterns of words in the database were based solely on listings in this dictionary. We acknowledge that the NHK dictionary gives some of the words more than one possible accent pattern, as discussed below in section 9.2.

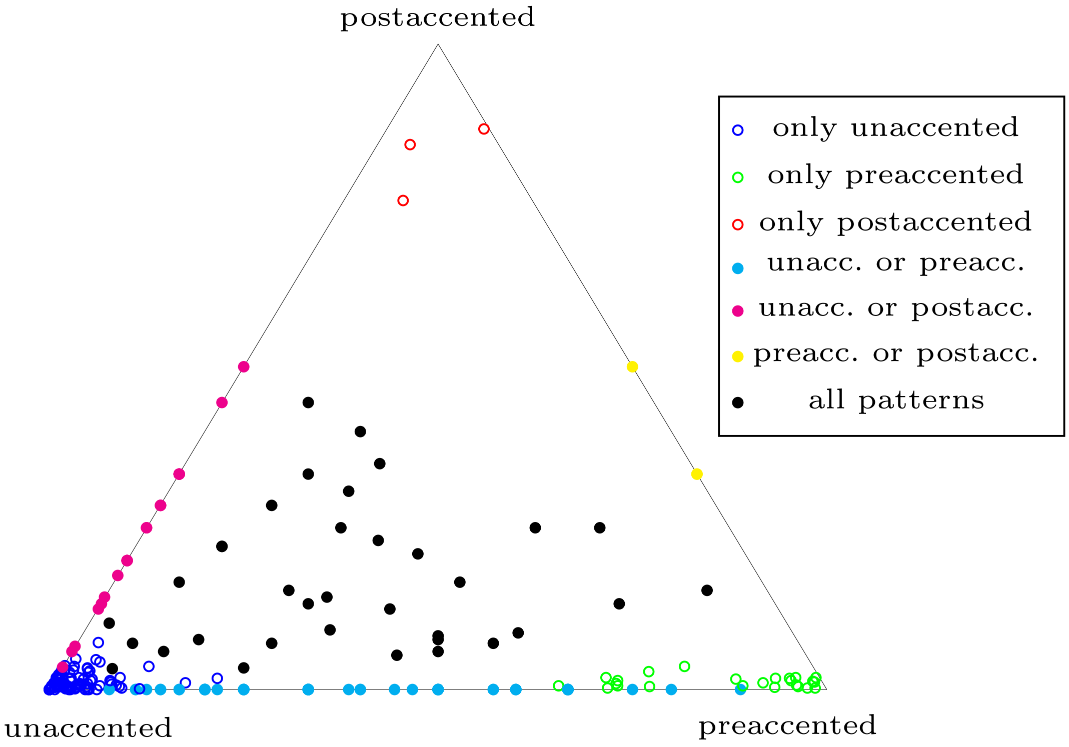


Figure 1: Gradient behaviour of N1s

These gradient patterns can be captured in the framework of Gradient Symbolic Computation (Smolensky and Goldrick 2015, henceforth GSC), a type of harmonic grammar which allows partially activated feature values and weighted constraints. GSC's ability to predict these patterns holds promise that it can similarly predict other observed gradient patterns in this and other languages. A GSC analysis in Rosen (2016) also explains gradient patterns of rendaku voicing in Japanese that are not otherwise predictable.

2 Patterning of pitch accent in Japanese 2μ-2μ N-N Yamato compounds

This dataset of 2μ-2μ compounds was chosen to control for the effects of prosody, syntactic category and lexical stratum on surface pitch accent. If we factor out dvandva compounds, the pitch-accent falls overwhelmingly into three possible patterns: unaccented, prejunctural accent (accent on the second mora of N1) or postjunctural accent (accent on the first mora of N2.) Following Itô and Mester (2016) and related work, we take the absence of initial and final accent in these compounds to result from high ranking of constraint NONFINAL(SYLL), which rules out final accent and RIGHTMOST, violated by any Foot following the head Foot, combined with INITIAL FOOT (the Pwd begins with a Foot) which rule out initial accent. Whereas the pitch accent of compounds in which at least one conjunct exceeds two moras in length is predictable, the accent of these shorter compounds is generally considered at best semi-predictable.

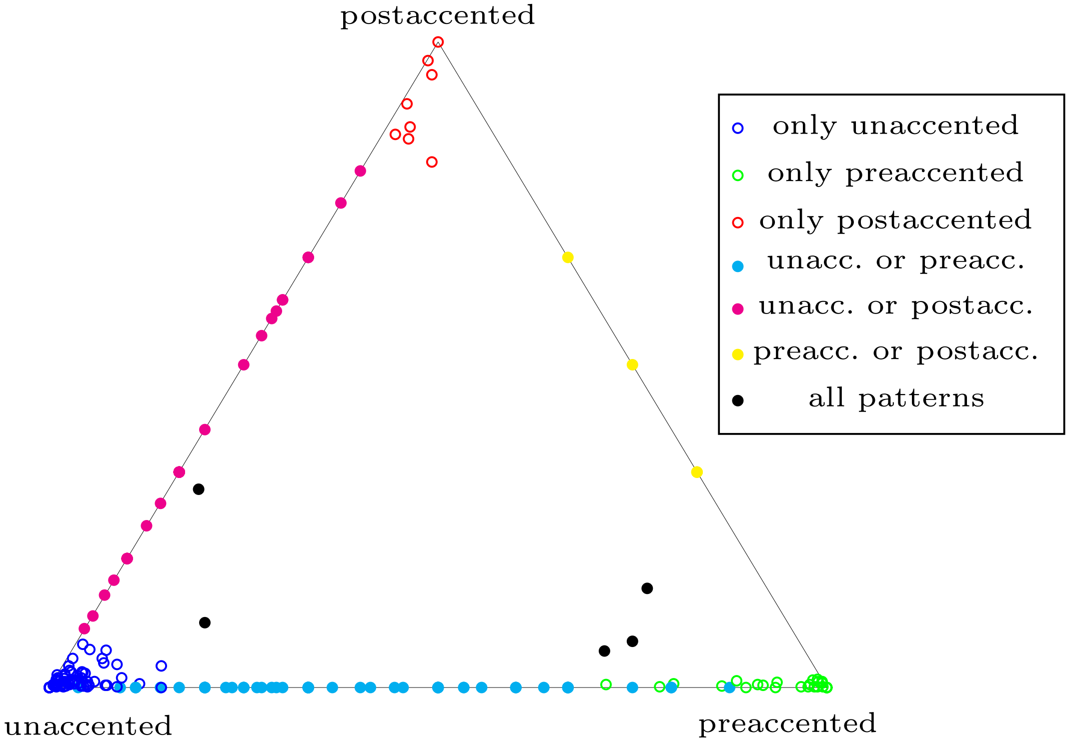


Figure 2: Gradient behaviour of N2s

In this dataset, the gradient tendencies of pitch accent are a function of not only N2, but also N1, regardless of which conjunct the accent surfaces on. For example, nouns *mizu* 'water', *isi* 'stone' and *húne* 'boat', when they occur as N1s, deaccent in 22/23, 9/10 and 12/15 compounds in the dataset respectively.

To explain how one conjunct can affect accent placement on the other, we posit floating accent features protruding over the left and right edges of each stem. A feature on the right edge of N1 can coalesce with a feature on the first mora of N2 and vice versa. The resulting feature activation for accent will be the sum of the activations of the two coalescing features.[2] Because of the proposed high ranking of constraints that rule out initial or final accent, we only look at three possibilities: no accent, accent on the second mora and accent on the third mora. Details of how this works are shown below in (4) and (5).

3 The GSC framework

This framework has both a symbolic and a sub-symbolic level. At the symbolic level, the activation level at which a feature surfaces is determined by Max and Dep constraints that are familiar from Optimality Theory (Prince and Smolensky 1993), but which have weighted

[2] Deriving a surface pattern through the coalescence of two features on either side of a morpheme boundary is also proposed in Smolensky and Goldrick (2016) who account for the occurrence of liaison consonants in French in the GSC framework.

values and are evaluated somewhat differently. For a MAX constraint of weight M_i for an input feature a_j with output a_k, the system gains positive harmony of $M_i \min(a_k, a_j)$: the amount of the feature that surfaces, up to a maximum of the input, weighted by the weight of the constraint. (For the feature to surface with a value greater than the input does not contribute to Harmony.) For a DEP constraint of weight D_i, for an input feature a_j with output a_k, the system loses harmony to the amount $D_i \max(0, a_k - a_j)$: the amount by which the output activation exceeds the input activation but with a minimum of 0.

GSC also employs a process called quantization, discussed further on page 7, which drives partially activated features to settle on values at or close to 1 or 0, depending on which of the two results in greater Harmony. Because Japanese never has more than one pitch accent per accentual phrase, we shall view quantization as occurring globally across the whole phrase, where all units compete for a value of 1 and at most one wins. (See Cho, Goldrick and Smolensky, to appear, for details of the mechanics of quantization in the GSC framework and the supplementary materials for details on how the quantization constraints work.)

4 Derivations of pitch accent patterns

We now show how pitch accent patterns in compounds in the database can be derived from input activation values of the two conjuncts. Consider the following three compounds:

(1) *yuki-dama* 'snow+ball=snowball' (unaccented)

(2) *yukí-gutu* 'snow+boots/shoes=snow boots' (preaccented)

(3) *kawa-gutu* 'hide+boots=leather boots' (unaccented)

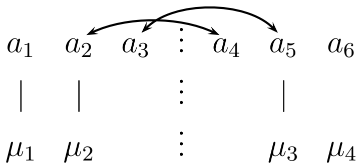
- Noun *yukí* 'snow' deaccents in the first compound but preaccents in the second.
- Noun *kutú* 'boots; shoes' preaccents in the second and deaccents in the third.

Neither *yukí* nor *kutú* predictably determines the accent of a compound on its own. But if compound accent is determined by partially activated input features coalescing on a particular mora, the occurrence or non-occurrence of accent will be determined by whether the sum of the two input features surpasses a threshold.[3] The following diagrams show how coalescence will occur in compounds when there are floating accent features protruding over the edges of the constituents. In the first diagram, vertical dashed lines indicate the morpheme boundary between the two constituents. a_i represents the activation of an accent feature. Arrows connecting two features indicate that they coalesce in the output.

[3] The nature of the threshold, which is an epiphenomenon in GSC that occurs as a result of the effects of MAX and DEP constraints, is explained further on page 8.

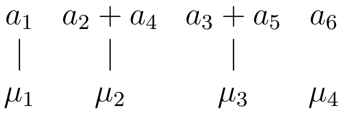
Input:

(4)


a_1 a_2 a_3 a_4 a_5 a_6
μ_1 μ_2 μ_3 μ_4

Output:

(5)

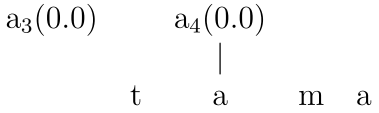

a_1 $a_2 + a_4$ $a_3 + a_5$ a_6
μ_1 μ_2 μ_3 μ_4

If accent occurs on μ_2 or μ_3, it will depend on whether the sum of the activations of the two contributing input activations exceeds the threshold, which in this case is 0.5, if Max and Dep constraints have unit values.

We rule out coalescence of a_2 with a_3 or a_4 with a_5 through a highly-ranked, strict Linearity constraint, as proposed in Pater (1999) and Buchwald et al. (2002) in which precedence $x \prec y$ in the input requires absolute precedence (i.e. $x \prec y$) in the output rather than lack of reversed precedence (i.e. $\neg[y \prec x]$). This rules out coalescence between any two features that have some precedence relation in the input. Following De Lacy (1999) and Zukoff (2016), we consider two compound-forming stems not to have any relative ordering in the input; therefore, there is no precedence relation in the input between a_3 and a_5 or a_2 and a_4.

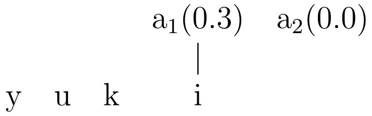
In (6) - (9) we posit hypothetical feature activations on the nouns in the compounds listed above, based on their behaviour in compounds across the dataset. *tamá* 'ball', which only occurs in unaccented compounds as an N2, has low activation on both its first mora and on a floating feature at the left edge.

(6)



yukí 'snow' has a higher activation on its second mora but not high enough to surpass the threshold of 0.5 when it coalesces with a_3 on *tamá*:

(7)



kutú 'boots, shoes' has an activation on its left-leaning floating accent feature high enough to surpass the threshold when it coalesces with the feature on the second mora of *yukí* but not with the lower-activated second mora on *kawá* 'hide'.[4]

(8)
$$\begin{array}{cccccc} a_5(0.35) & & a_6(0.0) & & a_7(0.1) & a_8(0.0) \\ & & | & & | & \\ & \mathrm{k} & \mathrm{u} & \mathrm{t} & \mathrm{u} & \end{array}$$

(9)
$$\begin{array}{cccccc} a_9(0.0) & & a_{10}(0.0) & & a_{11}(0.1) & a_{12}(0.0) \\ & & | & & | & \\ & \mathrm{k} & \mathrm{a} & \mathrm{w} & \mathrm{a} & \end{array}$$

Importantly, the input activation values of a constituent need to be the same for both simplex and compound derivations. In section 8 we show how these activation values also determine simplex accent. Because the unaccented pattern is much less common and initial accent more common than final accent in bimoraic simplex nouns, we assume constraints based on the prosody of bimoraic nouns to derive their surface accent differently than by simply choosing the mora with the highest activation above the threshold for simplex accent. We posit a threshold of somewhere between 0 and 0.3 for accent to surface on the second mora of simplex words, with some words having negative input activation on that mora, thus accounting for the final accent that surfaces on *yukí*, *kutú* and *kawá* as simplex words. This provides a unified account of compound accent and the simplex words that constitute them, with the input activations of both being learned simultaneously.

The following harmonic tableaux show derivations for the three compounds in (1), (2) and (3) above. Input activations of accent features are represented by decimal numbers. We assume initial and final accent to be ruled out by highly ranked prosodic constraints proposed by Itô and Mester (2016), given above in section 2, which we omit from the tableaux. Given that quantization will drive output activations to either zero or 1, only candidates with those activations are considered in the tableaux. In (10) only we show how a highly-ranked Linearity constraint rules out coalescence of tautomorphemic features.

(10)

	-2	1	1	1	-1	
$y\overset{0.0}{u}k\overset{0.3}{\imath}\overset{0.0}{}+\overset{0.1}{}t\overset{0.0}{a}m\overset{0.4}{a}$	Linearity	Max-Acc$_2$	Max-Acc$_3$	Max-Acc$_4$	Dep-Acc	H
☞ **yuki-dama**						**0**
yukí-dama		0.3		0.1	−0.6	−0.2
yuki-dáma					−1.0	−1.0
$yuk\overset{a_2+a_3}{\acute{\imath}}-dama$	-2	0.3	0.0		−0.7	−2.4

(11)

	1	1	-1	
$y\overset{0.0}{u}k\overset{0.3}{\imath}\overset{0.0}{}+\overset{0.35}{}k\overset{0.0}{u}t\overset{0.1}{u}$	Max-Acc$_2$	Max-Acc$_4$	Dep-Acc	H
yuki-gutu				0
☞**yukí-gutu**	0.3	0.35	−0.35	**0.3**
yuki-gútu			−1.0	−1.0

[4] The rendaku voicing of the initial obstruent on *kutú* is orthogonal to this analysis.

(12)

	1	1	-1	
$k\overset{0.0}{a}w\overset{0.1}{a}\overset{0.0}{\ }+\overset{0.35}{\ }k\overset{0.0}{u}t\overset{0.1}{u}$	Max-Acc$_2$	Max-Acc$_4$	Dep-Acc	H
☞**kawa-gutu**				**0**
kawá-gutu	0.1	0.35	−0.55	−0.1
kawa-gútu			−1.0	−1.0

If the input activation on the second mora of N1 a_2 plus the input activation on the left edge of N2 a_4 exceeds 0.5, as with *yukí-gutu*, for the candidate with prejunctural accent, the harmony due to Max constraints for those two positions will exceed 0.5 and the penalty on Harmony from the Dep constraint will be $1 - a_2 - a_4$ which is less than 0.5. Net Harmony will be positive, making that candidate more optimal than an unaccented candidate, which has zero net Harmony. Similarly, postjunctural accent will surface if the input activation on the right edge of N1 added to the activation on the first mora of N2 exceeds both 0.5 and the sum of the two input activations that trigger prejunctural accent. It can be shown algebraically that the effective threshold will be $\frac{D}{M+D}$ where M and D are the absolute values of the weights of the Max and Dep constraints. The threshold itself is an epiphenomenon in the GSC framework but is a convenient way of viewing the combined effects of Max and Dep constraints.

5 Possible alternative analyses

If the observed accent patterns were derived without coalescing floating accent features, the appearance of pitch accent on the prejunctural or postjunctural mora would depend solely on the input activation on that mora, modulated by constraints that would remain the same across the dataset. Such an account is ruled out by the occurrence of accent in simplex words that does not always match their behaviour in compounds. For example, *kitá-guni* ‘north country’ is preaccented but the first conjunct, *kita* ‘north’, is unaccented alone. Conversely, well over 100 compounds such as *hana-zono* ‘flower-garden’ are unaccented with a first conjunct that has final accent in simplex form. If the occurrence of an accent on the second mora of *kitá-guni* is based solely on the degree of input activation on the second mora of *kita* then that mora must have a higher activation than the second mora of *haná* ‘flower’, which contradicts that fact that *haná* is accented but *kita* is not.

6 Input activations across the lexicon

In order for the pitch accent patterns in our data to be completely predictable, two things need to occur:

1. Input activations for individual nouns stay constant across the lexicon. In the above tableaux, for example, the relevant input activations on *yukí* ‘snow’ are 0.1, 0.3 and 0.1 in both compounds.

2. There are no domination paradoxes among the nouns in the dataset. In the compounds below, *mizu* 'water' must have a higher leftmost input activation than *así* 'foot, and *así* higher than *kuti* 'mouth' since the former of each pair preaccents with an N1 with which the latter doesn't accent. Similarly, *wáni* 'crocodile' must have a higher second-mora input activation than *áto* 'after' and *áto* than *áme* 'rain' because the former of each pair preaccents with an N2 with which the latter doesn't accent.[5]

(13) *amá-mizu* 'rain-water' (preaccented)

(14) *ama-asi* (lit. rain-foot) 'beating of the rain' (unaccented)
mizu 'water' ≫ *así* 'foot' as an N2.

(15) *ató-asi* 'hind leg' (preaccented)
áto 'after' ≫ *áme* 'rain' as an N1.

(16) *ato-kuti* 'aftertaste' (lit. 'after-mouth') (unaccented)
ási 'foot, leg' ≫ *kuti* 'mouth' as an N2.

(17) *waní-guti* (lit. crocodile mouth) 'a Shinto folklore creature' (preaccented)
wáni 'crocodile' ≫ *áto* 'after' as an N1.

(18) *wani-gawa* 'crocodile hide' (unaccented)
kuti 'mouth' ≫ *kawá* 'after' as an N2.

(19) *mizu* 'water' ≫ *así* 'foot' ≫ *kuti* 'mouth' ≫ *kawá* 'hide' as N2s

(20) *wáni* 'crocodile' ≫ *áto* 'after' ≫ *áme* 'rain' as N1s.

But we never find pairs of compounds that contradict such a hierarchy. For example, no noun preaccents as an N1 with *kawá* 'hide' but not with *mizu* 'water'. This fact is significant inasmuch as the explanatory success of this model absolutely depends on a strict hierarchy of accenting tendencies which does not seem to have been observed before in the literature.

7 The relationship between compound accent and accent of simplex words

Consider now how our proposed input activations affect the accentuation of the conjuncts occurring as simplex words. Among the 856 noun-noun compounds we examined, 72% are unaccented, 22.5% are preaccented and 5.4% are postaccented. The predominance of the unaccented pattern in these four-mora compounds is consistent with the predominance of the unaccented pattern in four-mora words analysed by Itô and Mester (2016) and can be arguably accounted for by the same kinds of constraints they propose. Among two-mora native nouns, 27% are unaccented, 43% have initial accent and 30% have final accent. The difference in frequency of the unaccented pattern between four-mora compounds and two-mora simplex words suggests that we should not expect that a low activation on an accent

[5] The vowel alternation in *áme* rain is orthogonal to this analysis.

feature for a constituent of a compound will necessarily prevent that constituent from having an accented pattern when it occurs alone. We expect that a noun needs a higher activation in order to contribute to an accented pattern in a compound than it does as a simplex noun if forces in the grammar act more strongly against the surfacing of accent in four-mora words than they do in two-mora words. (See Itô and Mester (2016) for an analysis of why four-mora simplex words in Japanese tend to be unaccented.) It is therefore not inconsistent with our analysis that among the unaccented compounds in the database, only about 10% of them are composed of conjuncts both of which are unaccented alone.

We do need to examine cases where a noun that is unaccented alone contributes to an accented pattern in a compound, if that requires greater input activation than when it accents alone. Recall that the floating accent features we proposed will not figure in the accentuation of the simplex word, if a Linearity constraint prevents them from coalescing with a feature that has a path to a mora in the input. For an N1 that is unaccented alone, the crucial feature is the accent activation on its second mora, assumed to have a low activation since there is no accentuation in the simplex word. Among the compounds in the database with an N1 that is unaccented alone, the overwhelming majority are unaccented as well. Only in the following handful of cases is the compound preaccented. Words marked with a dagger below were considered unknown or obscure to one middle-aged Tokyo speaker who was consulted.

(21) *suzú-musi* ‘bell cricket’ (lit. ‘bell + ‘insect’) *musi* ‘insect’ preaccents in 7 out of 8 compounds.

(22) *tuzí-huda* ‘crossroads sign’ *huda* ‘tag; sign’ preaccents in all database compounds.

(23) *sodé-take* ‘sleeve length’ *take* ‘height; length’ preaccents in all database compounds.

(24) *mizú-gai* ‘water seashell: a freshwater shellfish’ *kai* ‘shellfish’ preaccents in all database compounds.

(25) *sibá-gaki* ‘brushwood fence’ *kaki* ‘fence’ preaccents in all compounds except with an N1 that always deaccents.

(26) *sibá-guri*† ‘brushwood chestnut’ *kuri* ‘chestnut’ preaccents in all database compounds.

(27) *sodé-haba*† ‘sleeve width’ is the only problematic case. *haba* ‘width’ ranks at the third level for preaccenting.

With the exception of *sodé-haba* ‘sleeve width’, preaccenting results from high input activation on the leftmost accent feature of the N2 in the compound, in spite of the putative low activation on the second mora of the N1.

When an N2 is unaccented alone, the accent activation on its first mora should be low. The database contains only two exceptions to this: *musi-búe* ‘insect-whistle’ and *asi-búe*† ‘reed flute’, both postaccented, with *hue* ‘flute’ unaccented alone. But the NHK dictionary lists alternative possible accents for each compound: *musi-bue* (unaccented) and *asi-bue* with all three patterns possible.

We now discuss a learning algorithm for input activations of the nouns that can find levels that are consistent with their accenting behaviour both as simplex words and as constituents of compounds.

8 A learning algorithm for activation levels

An error-driven learning algorithm was able to learn, with 100% accuracy, activation levels for the N1s and N2s in the database that correctly derived not only the accent pattern in all the compounds but also the accenting behaviour of each individual noun when it occurs alone.[6] For compound accent, the two relevant positions on N1s are the second mora and the proposed floating feature at the right edge; on N2s, the floating feature at the left edge and the first mora.

8.1 Initialization

Activation levels for accent were initialized at zero, based on initial lack of evidence for accentuation. MAX and DEP constraints were initialized with unit values. Threshold levels for the two moras for simplex accent were set at 0.1 for μ_1 and 0.2 for μ_2 after testing various values. This pair had the lowest average number of iterations (about 19) required to correctly derive all the compounds and their simplex constituents.

8.2 Steps on each iteration

On each iteration, the compounds are examined one at a time to check if each constituent's activations correctly derive (a) its simplex accent pattern and (b) the compound accent pattern. In each case, if the correct pattern is not derived, (i) activations on each relevant mora are incremented or decremented by a stepsize of 0.05 in the direction of a correct result (ii) MAX and DEP have their weights slightly adjusted through a simulated annealing process with a decaying temperature T, stepsize η and random Gaussian noise N with mean 0 and s.d. 0.5. Where two coalescing activations require adjusting, we randomly adjust one and adjust the second only if the value still requires adjustment. Since activations values can be either decremented or incremented, we end up with some negative activation values. The algorithm halts when every compound and simplex accent is correctly derived.

Because an unaccented pattern is much less common in simplex nouns than in the compounds, we posit lower thresholds for accentuation for two-mora simplex nouns than for 2μ-2μ compounds. Space limitations preclude a detailed investigation of what constraints would result in differing compound and simplex accenting tendencies. As an anonymous reviewer has pointed out, the constraint MINWORDACCENT proposed by Itô and Mester (2016) requires that a minimal prosodic word be accented. They show that ranking that constraint above NONFINALFT allows final accent in bimoraic words. High ranking of MINWORDACCENT would lower the threshold for accent in simplex bimoraic words.

Results of one run of simulating the learning algorithm:

[6] The idea of finding a learning algorithm that simultaneously learns activations for both simplex and compound accent patterns, to ensure that values are consistent between the two, was suggested by Paul Smolensky (personal communication).

(28)

Parameters	
Interval between activation levels on N1 and N2	0.05
Threshold on μ_1 for simplex words	0.1
Threshold on μ_2 for simplex words	0.2
Results	
Average number of iterations in 10 runs	19
Final value of Dep on one run	1.016
Final value of Max on one run	0.984
Range of number of levels for each of four positions	8 to 12

9 Residual issues and directions for further work

9.1 Compounds with 1μ constituents

Limiting the data to 2μ-2μ compounds abstracts away from the effects on accent of prosodic differences between compounds. In addition to 2μ-2μ compounds, it is well known that compounds whose prosodic structure is 1μ-2μ, 2μ-1μ or 1μ-1μ also show semi-regular accent patterns.

As suggested by an anonymous reviewer, we applied the same analysis to a set of 366 Yamato noun-noun compounds of this type, culling out words that a native speaker judged as obsolete or obscure. There were 151 1μ-2μ compounds, 198 2μ-1μ compounds and 17 1μ-1μ compounds, which were given previously learned activation values for 2μ constituents occurring in 2μ-2μ compounds. The algorithm learned values for their single-mora constituents that could derive both the accent locus of the compound and the simplex accent of the single-mora constituents. A small handful (4 or 5 compounds in each set) resisted a correct analysis but the rest (roughly 98%) analysed correctly. The resistant compounds have possibly adopted a lexicalized accent pattern because of their high frequency. Examples are *me-ue* (unaccented) ‘superior’ (i.e. social rank, lit. ‘eye-above’), *se-naka* ‘back’ (unaccented) (i.e. the back in human anatomy, lit. ‘back-middle’), *haka-bá* ‘graveyard’ (lit. grave-place) and *nama-tí* ‘fresh blood’.

9.2 Variation within items

The NHK Accent Dictionary lists multiple accent patterns for 186 of the 2μ-2μ compounds in the database. As pointed out by an anonymous reviewer, we should expect that in these cases, the sum of relevant activations should yield output values close to the borderline between the expected values for each possible pattern. To test this, the learning algorithm was run so that these compounds would seek values close to the threshold.

All but two compounds, *siba-bue* ‘brushwood-flute’ (unaccented or preaccented) and *aki-same* ‘autumn rain’ (unaccented), were correctly derived but only if the margin between the activation sum and the threshold was at least 0.15. The stepsize of adjustments to activation values also was halved from 0.05 to 0.025. We might ask, what kind of optionality is represented by the dictionary’s listing of multiple accent choices? Does a variable pattern mean that all speakers freely choose one or the other or do some speakers choose one pattern and others the other pattern? If optionality is mainly situated in variation from speaker to speaker, then the wide margin from the threshold would be expected.

Some of the variability could also be due to some speakers lexicalizing an accent pattern for compounds of high frequency such as *ama-gasa* 'rain-umbrella' (unaccented or postaccented) or *ama-gumo* 'rain-cloud' (postaccented or unaccented). And compounds with an infrequently-occurring constituent might have an indeterminate activation value for the floating features if speakers had little chance to learn a standard activation value based on its occurrence in other compounds. An example is *hana-gasa* 'a conical hat adorned with flowers' (unaccented or postaccented) for which the second constituent *kǻsa* 'bamboo hat' does not appear frequently in compounds.

9.3 Testing with cross-validation

As suggested by an anonymous reviewer, we tested the model through cross-validation, first training on a subset of the data in which every N1 and N2 was represented in at least one compound for each accent pattern it occurs in; otherwise, some N1 or N2 that occurred only in the test set could have no learned values. Training and then testing the model on the holdout data was repeated ten times, with the data randomly reshuffled between each test to ensure a different training set each time. The test sets varied from 298 to 316 items (approximately 37% to 39% of the total.) The accuracy on the test set ranged from 90% to 95% and averaged 93.2%. Most of the errors had an activation level only slightly too high or too low – for example, the threshold for compound accent was 0.495 and some activations at 0.5 were only slightly too high. Some errors also involved variably-accented compounds: for example, *ato-kuti* 'aftertaste' is listed by NHK first as unaccented and alternatively as preaccented. Every time it occurred in the test set among the ten cross-validation tests, it produced an error, with its activation for preaccenting slightly above the threshold.

Viewed in terms of precision and recall, the following table shows the results of a typical run.

Table 1: Precision and recall

	Precision	Recall
Unaccented	96.1%	96.1%
Preaccented	79.1%	77.3%
Postaccented	75.0%	100.0%

Precision and recall were not as good as overall accuracy for the preaccented pattern and precision was not as good for the postaccented pattern. This is due to the relatively low numbers of preaccented and postaccented compounds that end up in the test set, after making sure that each accent pattern that occurs for each N1 and N2 is represented in the training set.

We subsequently ran a cross-validation simulation that allowed the algorithm to go back over all the compounds and adjust activation levels to derive all the compounds in both the training set and the test set. On a run with 20 errors in cross validation, it took only one iteration back over all the data to get a correct derivation for the total set of compounds.

These results suggest that speakers may routinely vary the input forms of constituents by small degrees in order to derive new compounds as they are coined or encountered.

10 Discussion

Unlike Itô and Mester's (2016:43) examination of accent patterns of four-mora loanwords in Japanese, in which different patterns are seen as subgrammars of Japanese with different constraint rankings, here, assigning different compounds to different subgrammars is not possible, since the same conjunct can occur in compounds with different accent patterns, so there would be no way to correlate a lexical listing for a simplex noun with a given subgrammar.

The computational simulations described above show that these compound accent patterns can be derived from an additive effect of proposed activation values on both N1 and N2 that remain constant for each lexeme. Moreover, these activation values occur in a hierarchy that reflects each noun's tendency to trigger a certain type of accent. The capability of the GSC framework to represent pitch accent with continuous activation levels makes it ideally suited to capturing this kind of hierarchy.

Acknowledgements: Helpful discussion and suggestions were provided by Paul Smolensky, Matthew Goldrick, Armin Mester and participants at the Strength in Grammar workshop at the University of Leipzig where my attendance was generously supported by NSF grant BCS1344269. I also thank Shigeto Kawahara and two anonymous reviewers at Linguistics Vanguard for helpful comments and suggestions. All errors are my own.

11 References

Cho, Pyeong Whan, Matthew Goldrick and Paul Smolensky. Incremental parsing in a continuous dynamical system: Sentence processing in Gradient Symbolic Computation. To appear in *Linguistics Vanguard.*

Buchwald, Adam, Oren Schwartz, Amanda Seidl, and Paul Smolensky. 2002. Recoverability Optimality Theory: Discourse Anaphora in a Bidirectional framework. In Johan Jos, Mark Ellen Foster and Colin Matheson (eds.), Proceedings of the sixth workshop on the semantics and pragmatics of dialogue (EDILOG 2002), 37-44. 4-6 September. Edinburgh, UK.

Coetzee, Andries and Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. NLLT 26(2). 289-337.

de Lacy, Paul. 1999. A correspondence theory of morpheme order. In Peter Norquest, Jason D. Haugen, and Sonya Bird (eds.), WCCFL (West Coast Conference in Formal Linguistics) XVIII. Arizona: Coyote Working Papers in Linguistics, pp. 27-45. Rutgers Optimality Archive 338.

Goldsmith, John. 2011. The evaluation metric in generative grammar. Paper presented at the 50th anniversary celebration for the MIT Department of Linguistics, December 2011.

Itô, Junko and Armin Mester. 2016. Unaccentedness in Japanese. *Linguistic Inquiry* 47. 471-526.

Kubozono, Haruo and Yayoi Fujiura. 2004. Morpheme-dependent nature of compound accent in Japanese: An analysis of “short” compounds. *On-in Kenkyuu* [phonological studies] 7. 916.

Pater, Joe. 1999. Austronesian Nasal Substitution and other NC Effects. In Harry van der Hulst, René Kager and Wim Zonneveld (eds.). *The Prosody Morphology Interface.* Cambridge University Press. 310-343.

Prince, Alan and Paul Smolensky. 1993. Optimality Theory. Ms. Rutgers University.

Rosen, Eric. 2016. Predicting the unpredictable: Capturing the apparent semi-regularity of rendaku voicing in Japanese through harmonic grammar, in E. Clem, V. Dawson, A. Shen, A. H. Skilton, G. Bacon, A. Cheng and E. H. Maier, (eds.). Proceedings of BLS 42, Berkeley Linguistic Society, pp. 235-249.

Smolensky, Paul and Matthew Goldrick. 2015. Gradient Symbolic Computation. LSA Summer Institute Workshop. Chicago.

Smolensky, Paul and Matthew Goldrick. 2016. Gradient Symbolic Representations in Grammar: The case of French Liaison. Rutgers Optimality Archive 1552.

Zukoff, Sam. 2016. The Mirror Alignment Principle: Morpheme ordering at the morphosyntax-phonology interface. MIT Generals Paper. January 26.