# Predicting semi-regular Japanese accent patterns through gradient strengths of inputs

Eric Rosen

This talk was supported by NSF grant BCS1344269, thanks to Paul Smolensky and Johns Hopkins Cognitive Science Dept..

Research-in-progress to appear in Linguistics Vanguard, subject to revisions

Strength in Grammar workshop
University of Leipzig
November 11th, 2017

# Motivation for feature strength

- Japanese noun-noun compounds
- Each constituent is 'short' ($\leq 2\mu$)
- Here, we look at $2\mu + 2\mu$ compounds of Yamato (native) origin.
- e.g. $\overset{\mu}{\text{y a}}\ \overset{\mu}{\text{m a}}$ + $\overset{\mu}{\text{m i}}\ \overset{\mu}{\text{z u}}$ 'mountain-wind'
- Accent of these compounds falls overwhelmingly into three possible patterns:
  - Unaccented: *ato-kuti* 'after-mouth' ( lit. aftertaste)
  - Preaccented: *amá-mizu* 'rain-water'
  - Postaccented: *asi-áto* 'footprint'
- considered at best semi-predictable
- Haruo Kubozono (personal communication): compounds with no conjunct exceeding 2 moras need to be treated like single lexical items with respect to accent.
- Initial accent occurs on dvandva compounds: e.g. *nábe-kama* 'pots and pans', not analysed here.

# Motivation for feature strength

- Proposal: surface accent of these compounds can be predicted from input forms of constituents if accent input features can have varying strengths.

e.g.

| | | | | |
|---|---|---|---|---|
| | acc·(0.05) | | acc·(0.3) | |
| | $\mu$ | | $\mu$ | |
| y | a | m | a | 'mountain' |

- Feature strengths can capture observed tendencies for a constituent of such a compound to trigger or not trigger an accent pattern.

- Here we adopt Gradient Symbolic Computation (Smolensky and Goldrick 2015) – a framework in which phonological features can have **partial levels of activation.**

# Japanese pitch accent

- A pitch fall from H to L tone. $\ldots \overset{\mu}{\underset{|}{\mathrm{H}}}\ \overset{\mu}{\underset{|}{\mathrm{L}}} \ldots$
- We abstract away from tonal representations and consider 'accent' as a phonological feature, following Itô and Mester 2016 inter alia.
- Won't try to determine typological nature of pitch accent. (See Hyman 2009)
- Words/phrases can be unaccented: e.g. *kitune* 'fox'
- At most one accent per accentual phrase. e.g. *hatu-kao-áwase* 'first face-to-face meeting' (Itô and Mester 2006 (12))

# Limiting the data to Yamato (native Japanese) 'short-short' noun-noun compounds: reduce number of variables affecting accent

- Sino-Japanese compounds behave differently:
  - *"Sino-Japanese compounding obeys very rigid restrictions on size and segmental combinatorics unknown to the rest of the lexicon" (Itô and Mester 2015:22)*
  - SJ words typically consist of two bound morphemes: e.g. *hoo-soo* 'broadcasting';
  - Compounding these results in prosodically longer words: e.g. *hoo-soo-ki-sya* 'radio/TV reporter'
  - Sino-Japanese compounds show very different accenting tendencies (Kawahara 2015:17)
  - Foreign borrowings that look like compounds:
    - tend to be of greater prosodic length because of epenthesis (shown in angled brackets) to respect Japanese phonotactics

      e.g. *h⟨o⟩wait⟨o⟩ k⟨u⟩ris⟨ú⟩mas⟨u⟩* 'white Christmas' (Kubozono and Mester 1995 (13c))
  - Verb compound accent is shown to be affected by adjunct vs. argument status of the first conjunct (Yamaguchi 2011)

# Accenting tendencies of N1s

Some examples:

| N1s that usually un accent | | | |
|---|---|---|---|
| Noun | Unacc. freq. | Unaccented example | Accented exception |
| *húne* 'boat' | 12/15 | *huna-zoko* 'ship's bottom' | *hunā-nusi* 'ship's master' |
| *isi* 'stone' | 9/10 | *isi-basi* 'stone bridge' | *isí-miti* 'stone road' |
| *mizu* 'water' | 22/23 | *mizu-tama* 'water droplet' | *mizú-gai* 'seashell in cold water' |

| N1s that usually accent | | | |
|---|---|---|---|
| Noun | Acc. freq. | Accented example | Unaccented exception |
| *sio* 'salt' | 9/10 | *sió-azi* 'salty taste' | *sio-zawa* 'salt marsh' |

# Accenting tendencies of N2s

| N2s that usually unaccent | | | |
|---|---|---|---|
| Noun | Unacc. freq. | Unaccented example | Accented exception |
| *ási* 'foot' | 9/12 | *ama-asi*<br>'beating of the rain'<br>(lit. rain-foot) | *ató-asi*<br>'hind legs' |

| N2s that usually accent | | | |
|---|---|---|---|
| Noun | Acc. freq. | Accented example | Unaccented exception |
| *kazé* 'wind' | 14/20 | *yamá-kaze*<br>'mountain wind' | *kita-kaze*<br>'north wind' |

# Strict hierarchies of preaccenting tendencies: seen in morphologically minimal pairs

Same N1s looking horizontally. Same N2s looking right and up.

| | Preaccented | | | Unaccented | |
|---|---|---|---|---|---|
| a. | *amá-mizu* | 'rain-water' | b. | *ama-asi* | 'rain-foot' (beating of the rain) |
| c. | *ató-asi* | 'hind foot' | d. | *ato-kuti* | 'after-mouth' (aftertaste) |
| e. | *waní-guti* | 'crocodile-mouth' (folklore creature) | f. | *wani-gawa* | 'crocodile hide' |

*wáni* 'crocodile' $\gg$ *áto* 'after' $\gg$ *áme* 'rain' (as N1s)
(by comparing e. to d. and c. to b.)

*mizu* 'water' $\gg$ *así* 'foot' $\gg$ *kuti* 'mouth' $\gg$ *kawá* 'hide' (as N2s)
(by comparing a. to b., c. to d. and e. to f.)

We never find compounds that contradict this hierarchy: e.g. no $N_1$ preaccents with *kawá* 'hide' but not with *mizu* 'water'.

# Strict hierarchies of postaccenting tendencies

| Postaccented | | | Unaccented | | |
|---|---|---|---|---|---|
| a. | *haru-góro* | 'spring-time' | b. | *haru-same* | 'spring-rain' |
| c. | *nuka-áme* | 'bran-rain' (drizzle) | d. | *nuka-miso* | 'bran-miso' |

*nuka* 'bran' ≫ *haru* 'spring' (as N1s)

*koro* 'time' ≫ *ame* 'rain' ≫ *miso* 'bean paste' (as N2s)

There is no N1 that postaccents with *miso* but not with *koro*.

| Postaccented | | | Unaccented | | |
|---|---|---|---|---|---|
| | | | b. | *huyu-zora* | 'winter-sky' |
| c. | *aki-zóra* | 'autumn-sky' | d. | *aki-same* | 'autumn rain' |
| e. | *nuka-áme* | 'bran-rain' | | | |

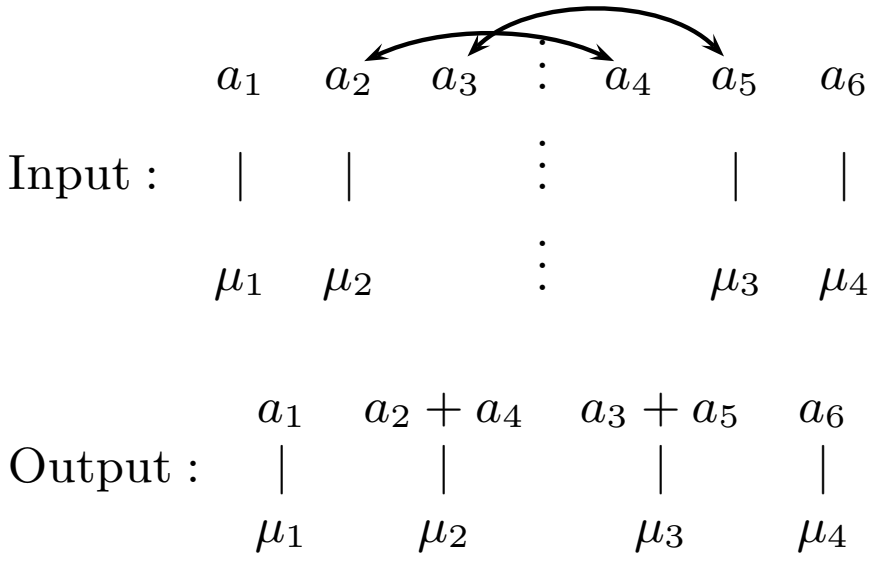
*sora* 'sky' ≫ *ame* 'rain'

*nuka* 'bran' ≫ *aki* 'autumn' ≫ *huyu* 'winter'

N1 can affect accent on N2: *huyu-zora* vs. *aki-zóra*

N2 can affect accent on N1: *amá-mizu* vs. *ama-asi*

# Proposed floating accent features at edges




Feature on mora at juncture edge of one conjunct coalesces with anchored feature at juncture edge of other conjunct.

The idea of features coalescing in the output at a morpheme juncture is also proposed in two other GSC accounts: Smolensky and Goldrick (2015) and Rosen (2016).

## How can these accent features determine surface accent?

- Features have gradient strengths e.g. an accent feature of 0.3
- Two coalescing features have the additive strength of the two features: $a_i \cdot (0.3) + a_j \cdot (0.4) \rightarrow a_{ij} \cdot (0.7)$
- Accent surfaces when the additive strength exceeds some threshold.

# Gradient Symbolic Computation (Smolensky and Goldrick 2015)

A framework in which phonological features can have **partial levels of activation.**

Features of GSC:

- A type of Harmonic Grammar
- Two levels:
  - Sub-symbolic level: a connectionist network
  - Symbolic level of (usually) discrete structures
- Here, we focus mainly on the symbolic level.
- Effect of the sub-symbolic level at the symbolic level is:
  - Weighted constraints: e.g. MAX-ACCENT=1.0
  - Varying degrees of activation for symbolic elements: e.g. $a \cdot 0.4$
- Process of *quantization* forces output activations to be at or near 0 or 1.

# Max and Dep constraints in GSC

- Max constraints contribute positive Harmony to the amount of an input feature that surfaces.
- Dep constraints penalize with negative Harmony to the amount of the deficit between an output activation and its corresponding input activation.
- If two features coalesce with input strengths $a_i$ and $a_j$:
  - If the feature surfaces with value 1 in an output candidate, Harmony due to Max is $M(a_i + a_j)$, Harmony due to Dep is $-D[1 - (a_i + a_j)]$. Net Harmony is $(a_i + a_j)(M + D) - D$.
  - If the feature does not surface in a candidate (value 0), Harmony is 0.
  - Candidate with value 1 will win iff $(a_i + a_j)(M + D) - D > 0$ or $a_i + a_j > \frac{D}{M+D}$
  - If $M$ and $D$ have unit weights, then the threshold is $\frac{1}{2}$

# Harmonic tableaux for preaccenting

| $(0.35)_1$( 0.3 )$_2$ $(0)_3$ ( 0.25 )$_4$ $(0)_5$ $(0.2)_6$<br>a t o + a s i | $\text{Max}_1$ | $\text{Max}_2$ | $\text{Max}_3$ | $\text{Max}_4$ | $\text{Max}_5$ | $\text{Max}_6$ | Dep | H |
|---|---|---|---|---|---|---|---|---|
| (a) ☞ **ató-asi** | | 0.3 | | 0.25 | | | -0.45 | **0.1** |
| (b) ato-ási | | | | | | | $-1.0$ | $-1.0$ |
| (c) ato-asi | | | | | | | | 0 |

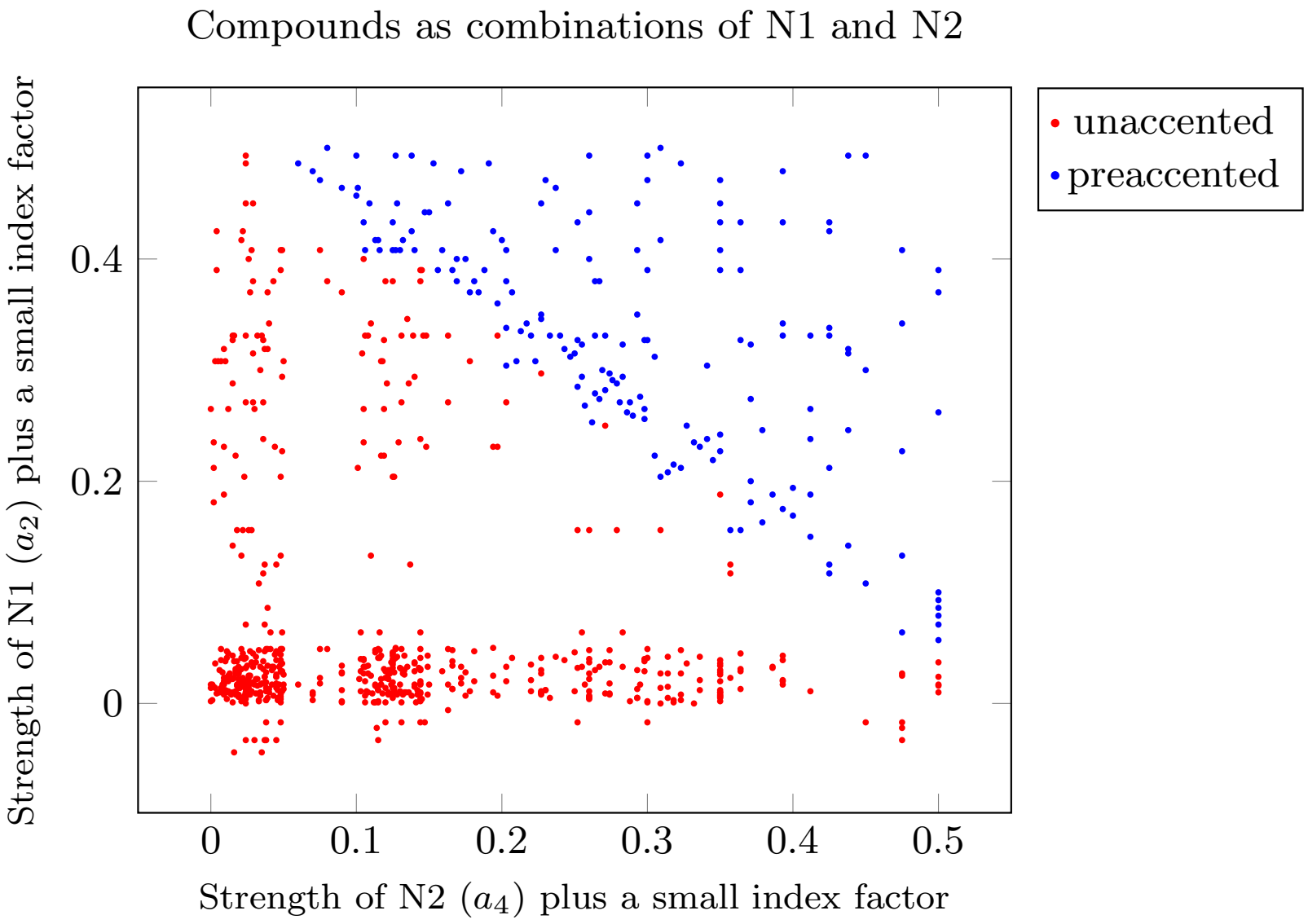
- A separate MAX constraint calculation for each indexed feature.
- Because of quantization, we only consider candidates with accent activations of 0 or 1.
- For cand. (a), activation sum of coalescing $a_2$ and $a_4$, $(0.3 + 0.25 = 0.55)$ due to MAX exceeds deficit $(1.0 - 0.55 = 0.45)$ due to DEP, giving net Harmony 0.1.
- For cand. (b), zero input activation for accenting initial mora of *asi* results in zero Harmony due to MAX.

# Harmonic tableaux for unaccented compound with same N1

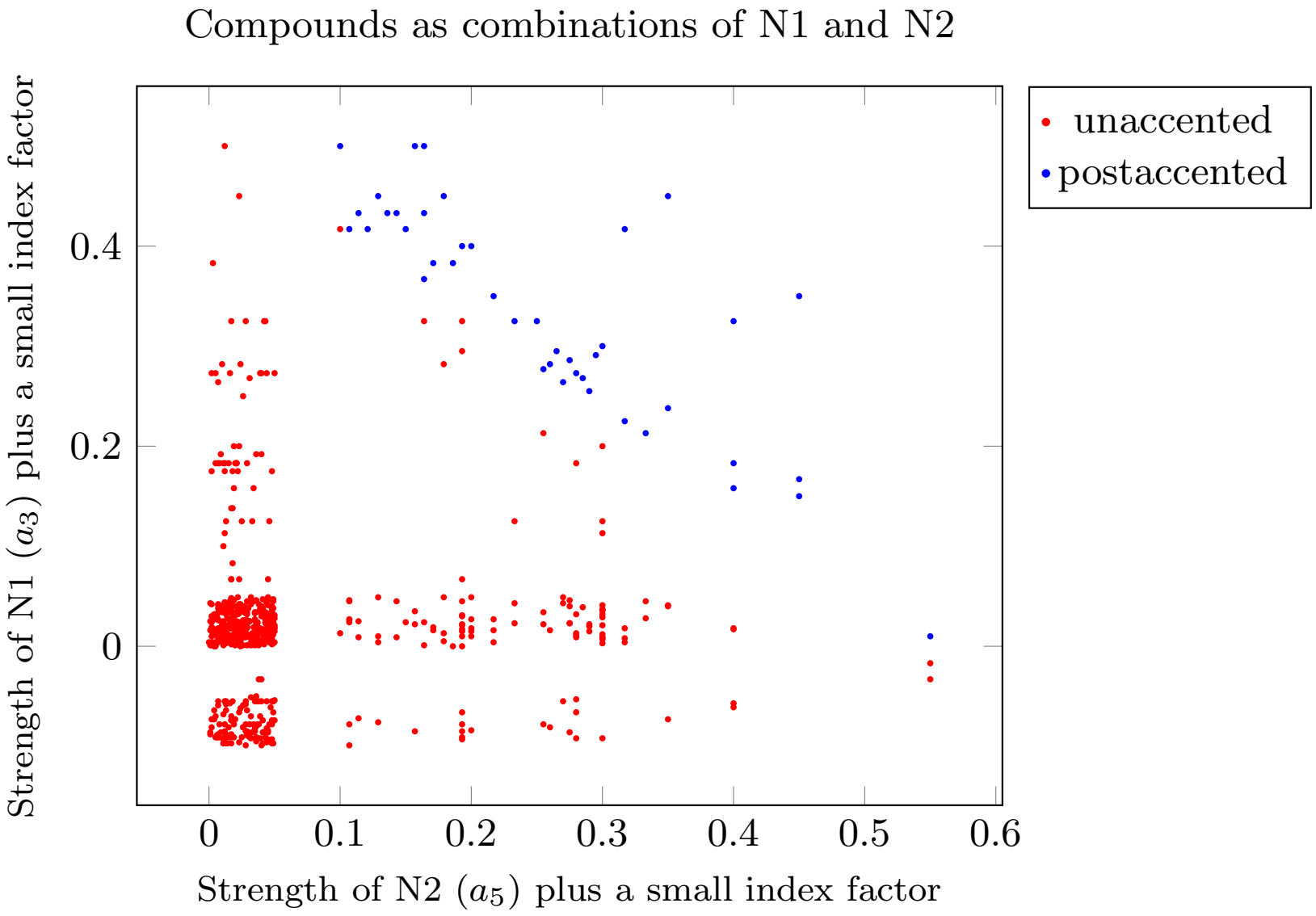
| $(0.35)_1$ ( 0.3 $)_2$ $(0)_3$ ( 0.15 $)_4$ $(0)_5$ $(0)_6$<br>a t o + k u t i | $\text{Max}_1$ | $\text{Max}_2$ | $\text{Max}_3$ | $\text{Max}_4$ | $\text{Max}_5$ | $\text{Max}_6$ | Dep | H |
|---|---|---|---|---|---|---|---|---|
| (a) ató-kuti | | 0.3 | | 0.15 | | | -0.55 | -0.1 |
| (b) ato-kúti | | | | | | | $-1.0$ | $-1.0$ |
| (c) ☞**ato-kuti** | | | | | | | | **0** |

Lower activation on floating $a_4$ of *kuti* (0.15) than on *asi* (0.25) results in negative Harmony for preaccented candidate (a), making unaccented candidate (c) optimal.

# How unaccented vs. preaccented compounds look w.r.t. the features that could make them preaccented ($a_2 + a_4$)



# How unaccented vs. postaccented compounds look w.r.t. the features that could make them postaccented ($a_3 + a_5$)



# Why can't features from the same morpheme coalesce?

- Propose a highly-ranked, **strict** LINEARITY constraint as in Pater (1999) or Buchwald et al. (2002).
- Precedence $x \prec y$ in the input requires absolute precedence (i.e. $x \prec y$) in the output, NOT lack of reversed precedence (i.e. $\neg[y \prec x]$).
- 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 stems that combine to form a compound 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 $a_4$.
- LINEARITY also rules out triple coalescence such as $a_2$ with $a_4$ and $a_5$

# Why can't multiple accents surface?

We consider quantization to operate globally, where a maximum of one mora can surface with accent activation of 1.

Global quantization constraint (Paul Smolensky, personal communication)

$H_{Q_1} = -(\sum_k a_k^2)[(\sum_k a_k^2) - 1]^2$

- First term is sum of squares of all accent activations.
- Second term is square of sum of squares of activations less 1.
- If all activations are zero, first term is 0 so Harmony is 0.
- If exactly one activation is 1, second term is 0 so Harmony is 0.
- If more than one activation is 1, second term is positive; so is first term, so Harmony is negative.
- So maximum Harmony (zero) occurs iff there is either 0 or 1 activation of 1.

# A learning algorithm for feature strengths

- A database of $2\mu - 2\mu$ noun-noun compounds.

| Compound | Acc. $\mu$ | Acc. N1 alone | Acc. N2 alone |
|---|---|---|---|
| ama-mizu ‘rain-water’ | 2 | 1 | 0 |
| ama-asi ’rain-foot’ | 0 | 1 | 2 |
| ato-asi ‘after-foot’ | 2 | 1 | 2 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

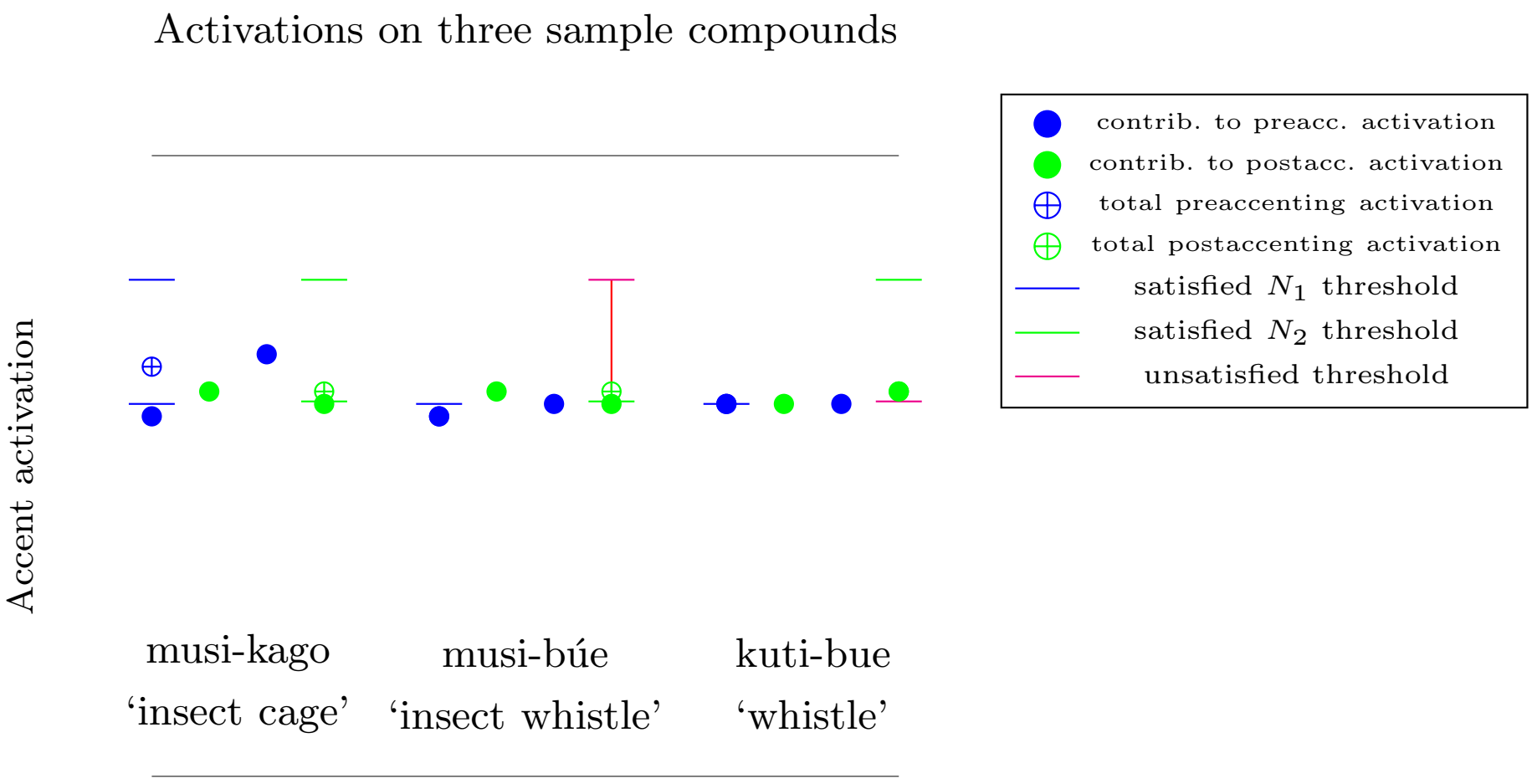
- Relevant accent features on N1 are $\mu_2$ and floating accent at right edge.
- Relevant accent features on N2 are floating accent at left edge and $\mu_1$.

$$\boxed{\begin{array}{ccc} & a_2 & a_3 \\ & | & \\ \mu & \mu & \\ \multicolumn{3}{c}{\leftarrow N_1 \rightarrow} \end{array}} \qquad \boxed{\begin{array}{ccc} a_4 & a_5 & \\ & | & \\ & \mu & \mu \\ \multicolumn{3}{c}{\leftarrow N_2 \rightarrow} \end{array}}$$

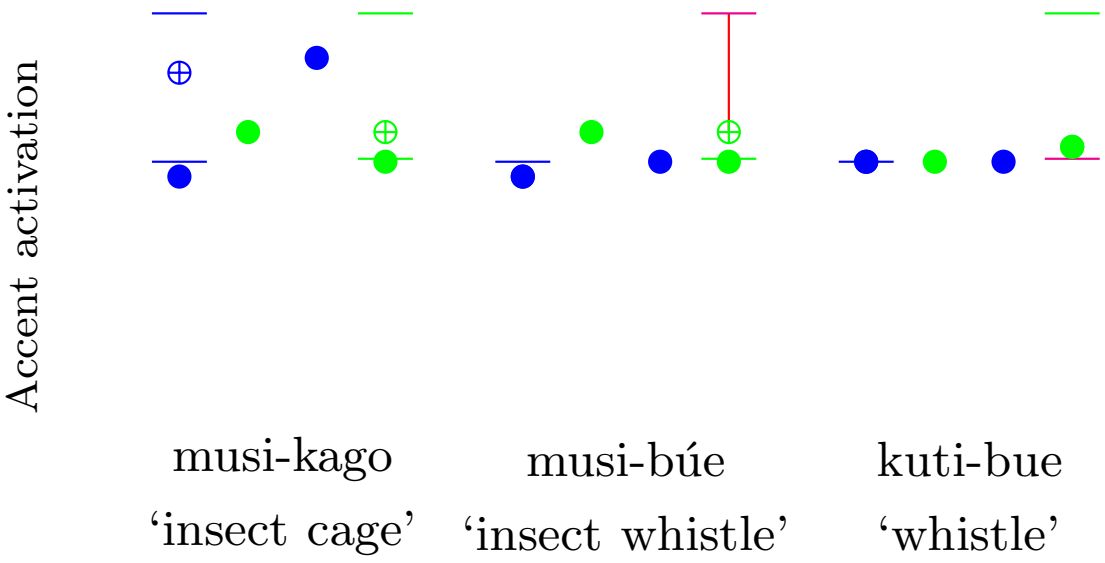
- Initialize accent features at zero
- Initialize MAX at 1 and DEP at $-1$.

# A learning algorithm for feature strengths

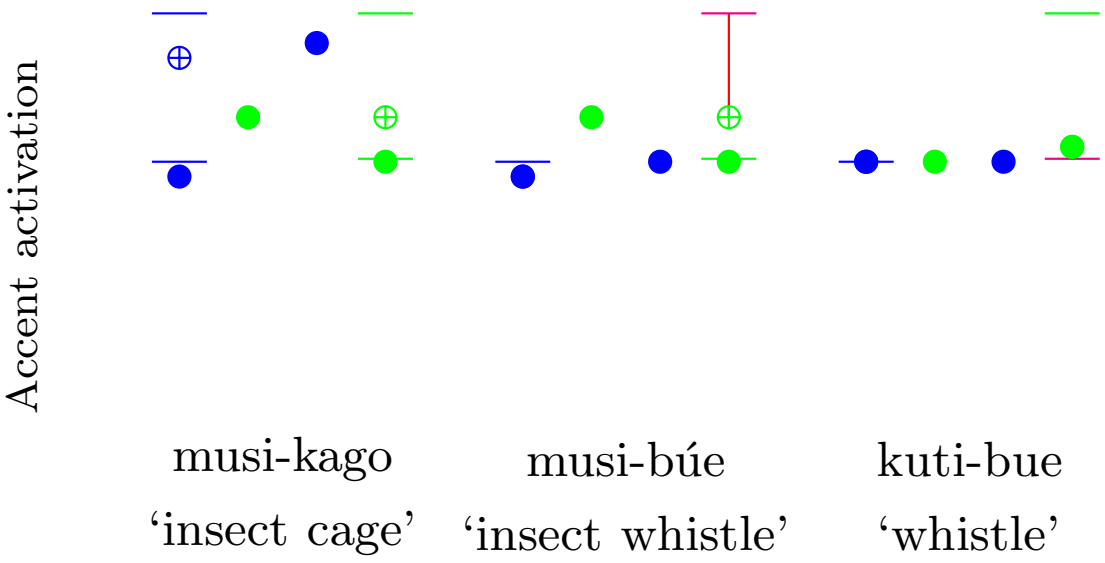
- Algorithm is error-driven.
- Run repeated iterations until accent strengths derive all cases of compound accent as well as accent of simplex words.
- For each compound, if incorrect result is derived, modify strengths of relevant features by a small stepsize in the direction of a correct result.
- Also modify weights if MAX and DEP constraints in the direction of correct results.


Activations on three sample compounds
Accent activation
musi-kago
'insect cage'
musi-búe
'insect whistle'
kuti-bue
'whistle'
contrib. to preacc. activation
contrib. to postacc. activation
total preaccenting activation
total postaccenting activation
satisfied $N_1$ threshold
satisfied $N_2$ threshold
unsatisfied threshold

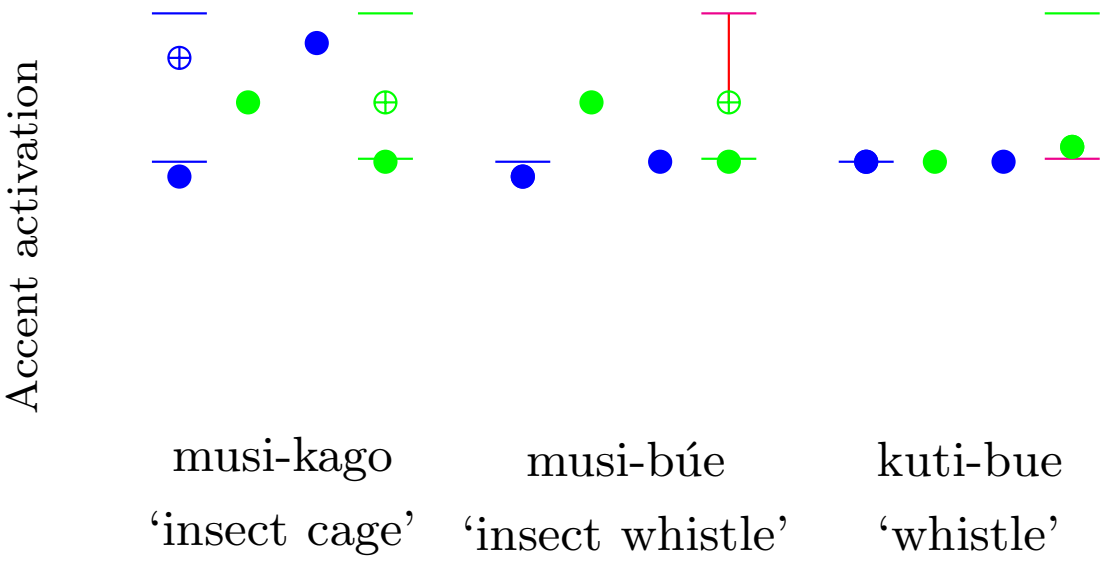
Activations on three sample compounds



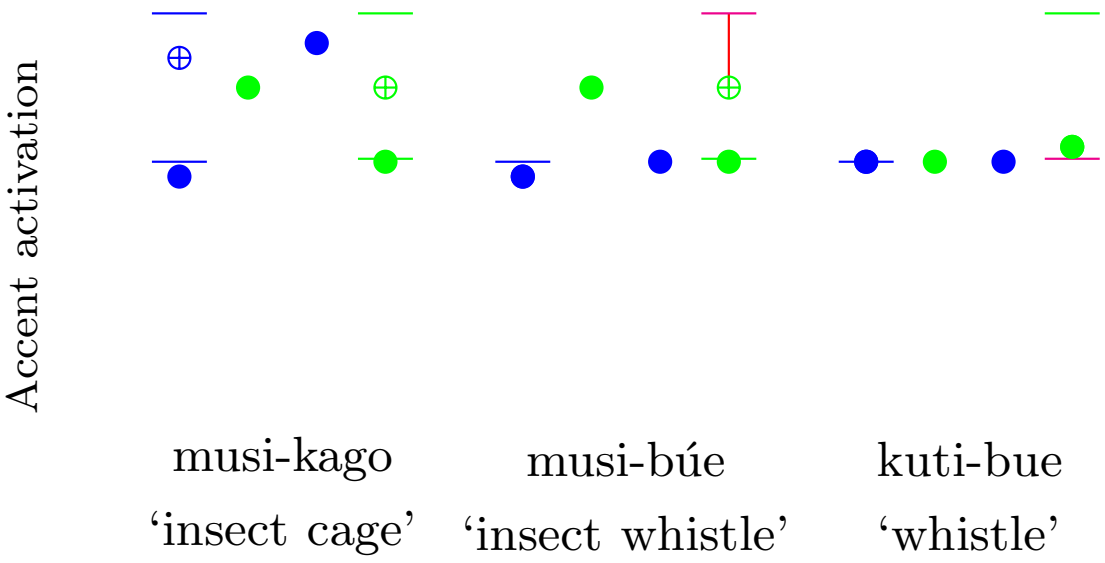
Activations on three sample compounds



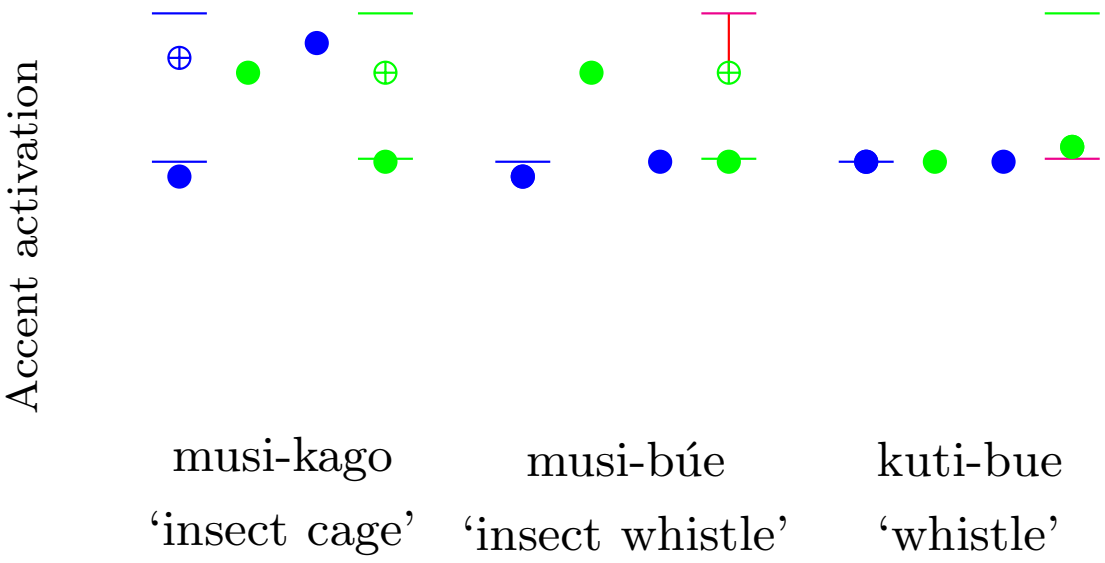
Activations on three sample compounds



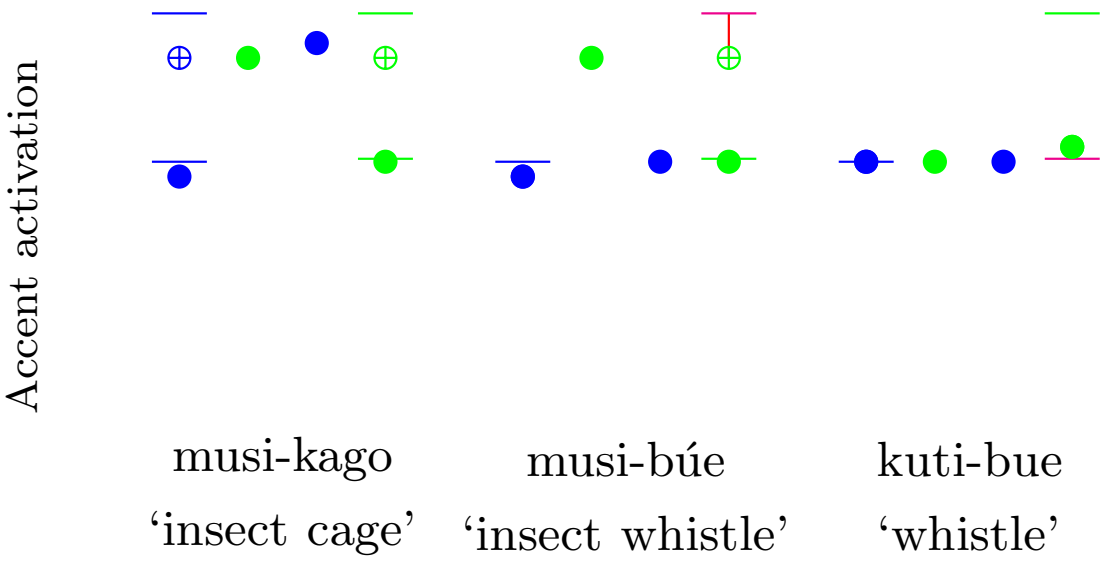
Activations on three sample compounds



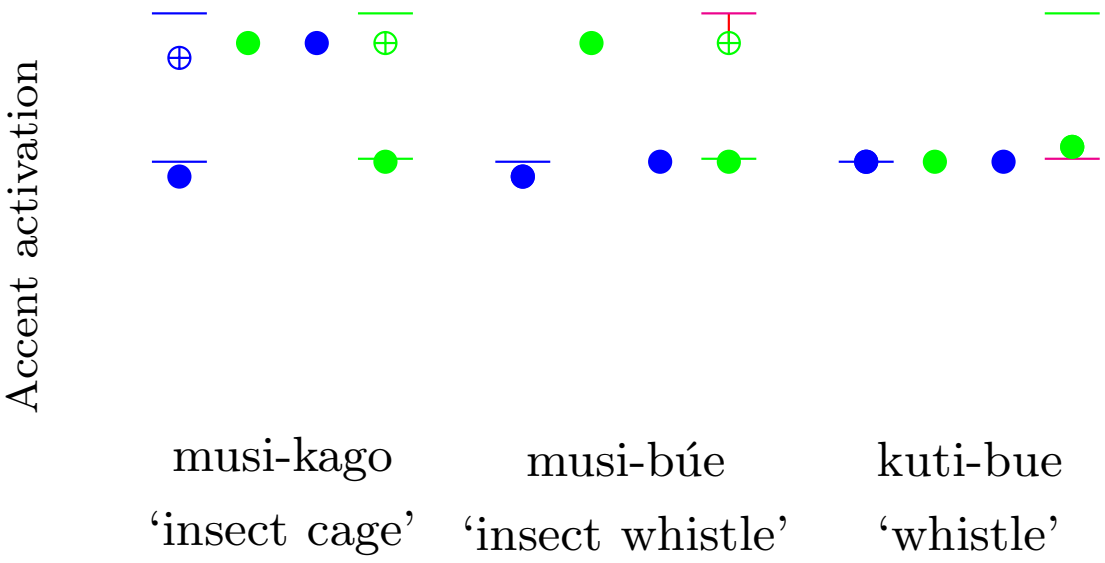
Activations on three sample compounds



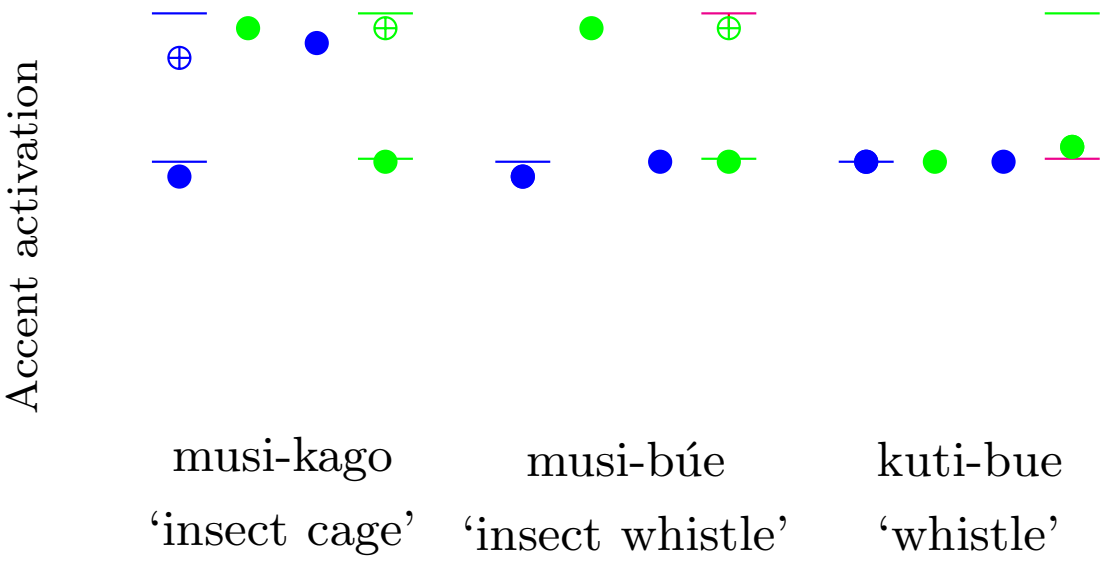
Activations on three sample compounds



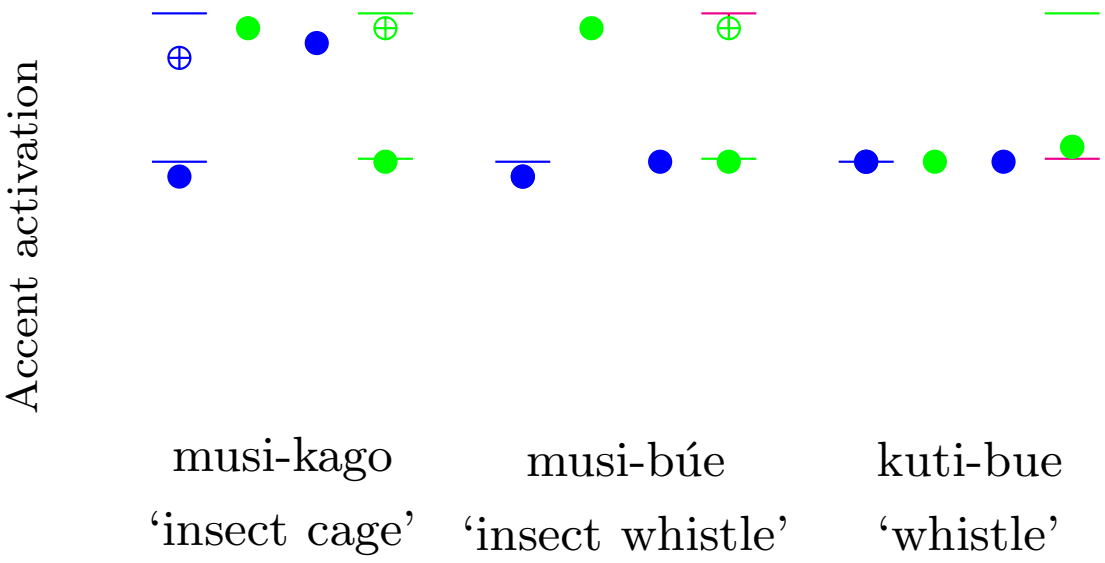
Activations on three sample compounds



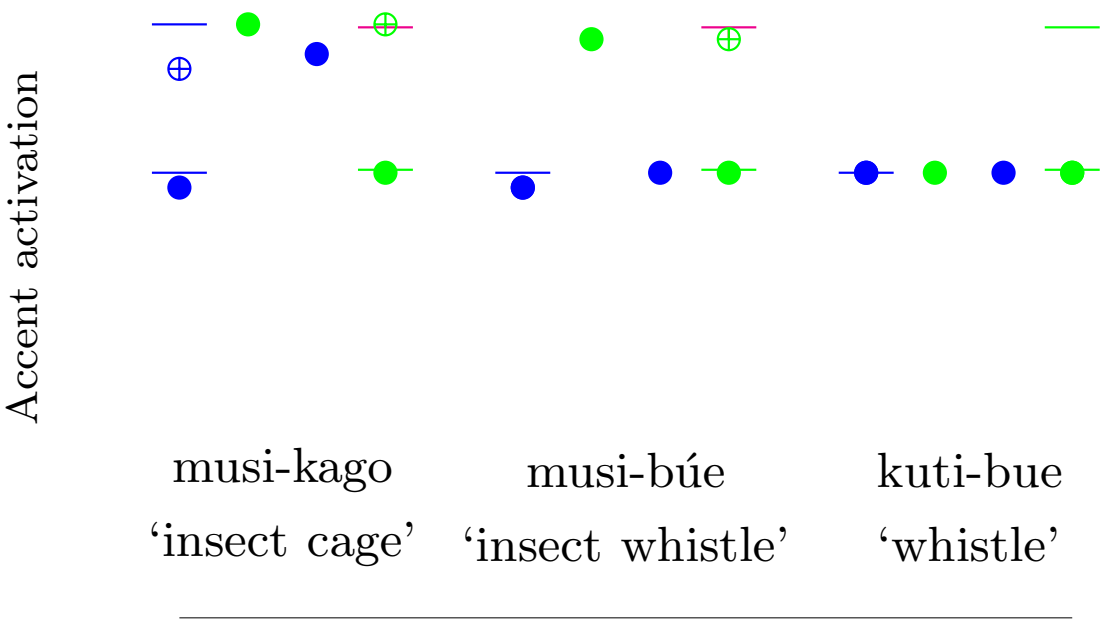
Activations on three sample compounds



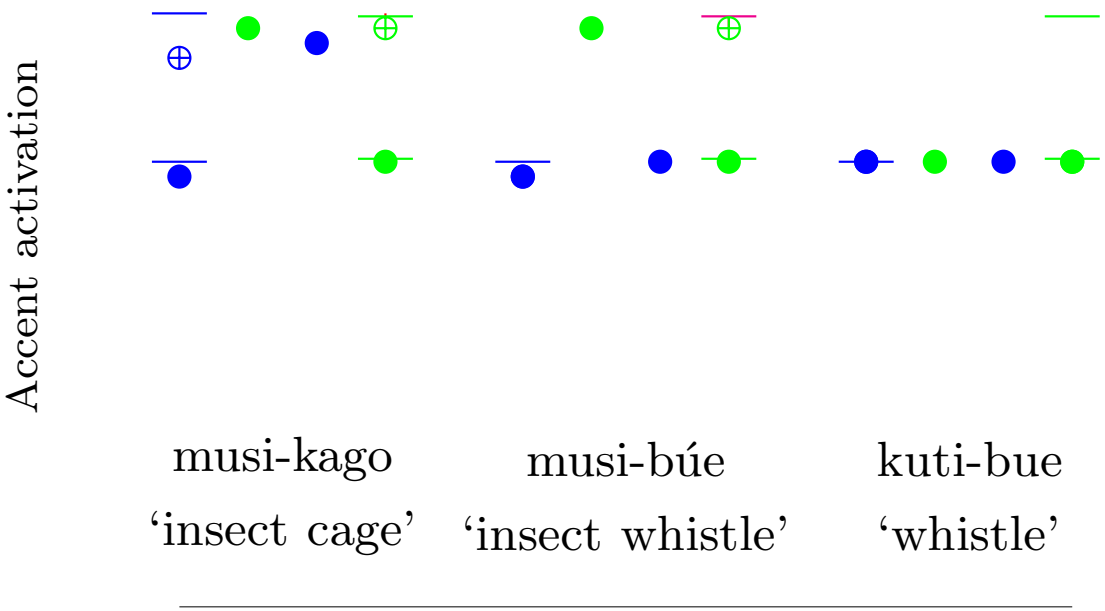
Activations on three sample compounds



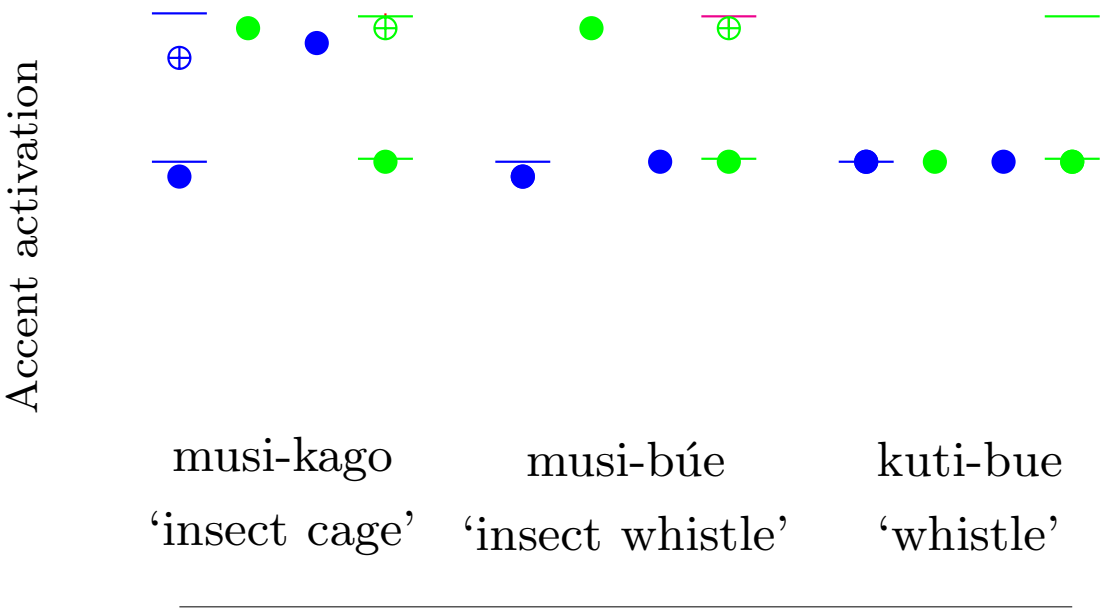
Activations on three sample compounds



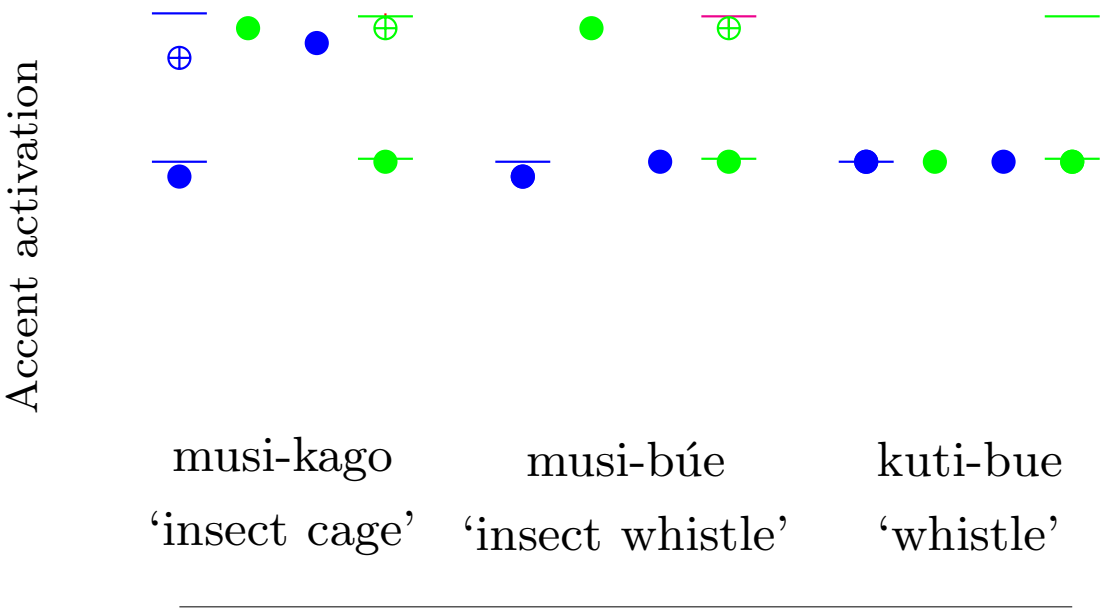
Activations on three sample compounds



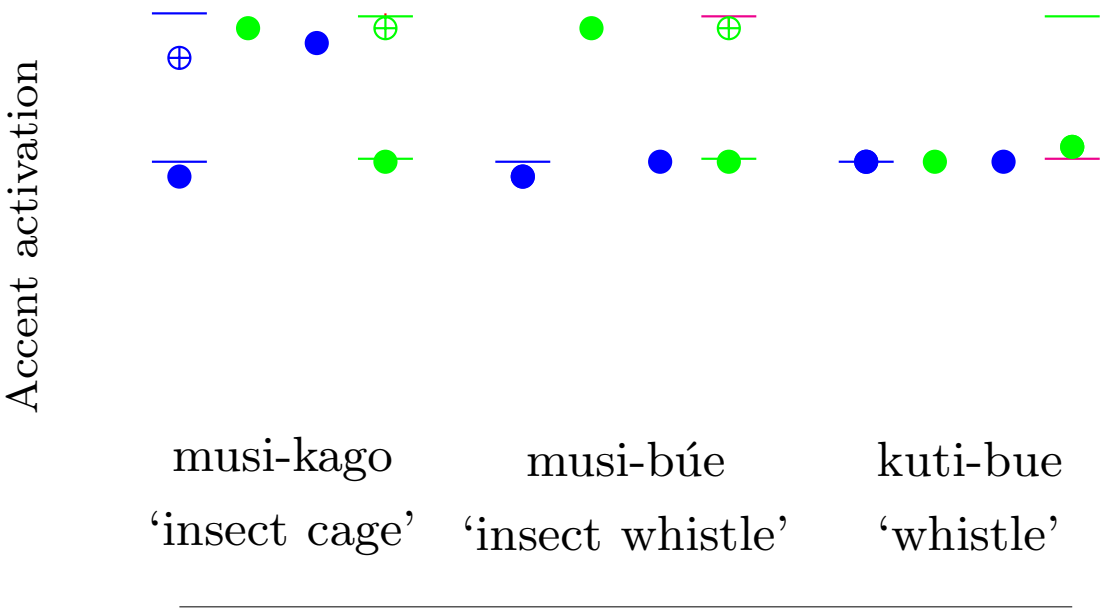
Activations on three sample compounds



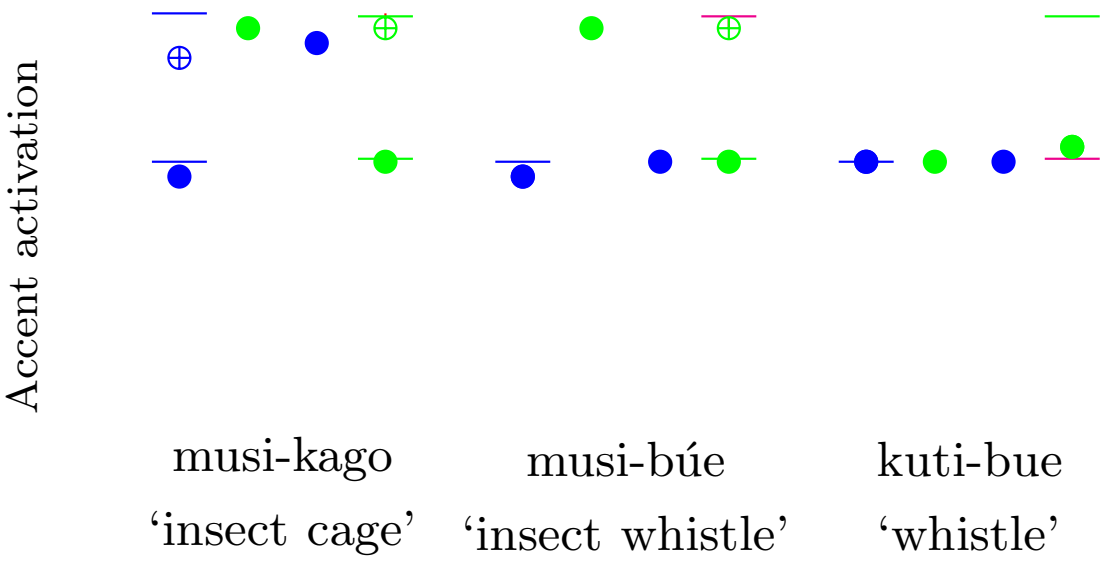
Activations on three sample compounds



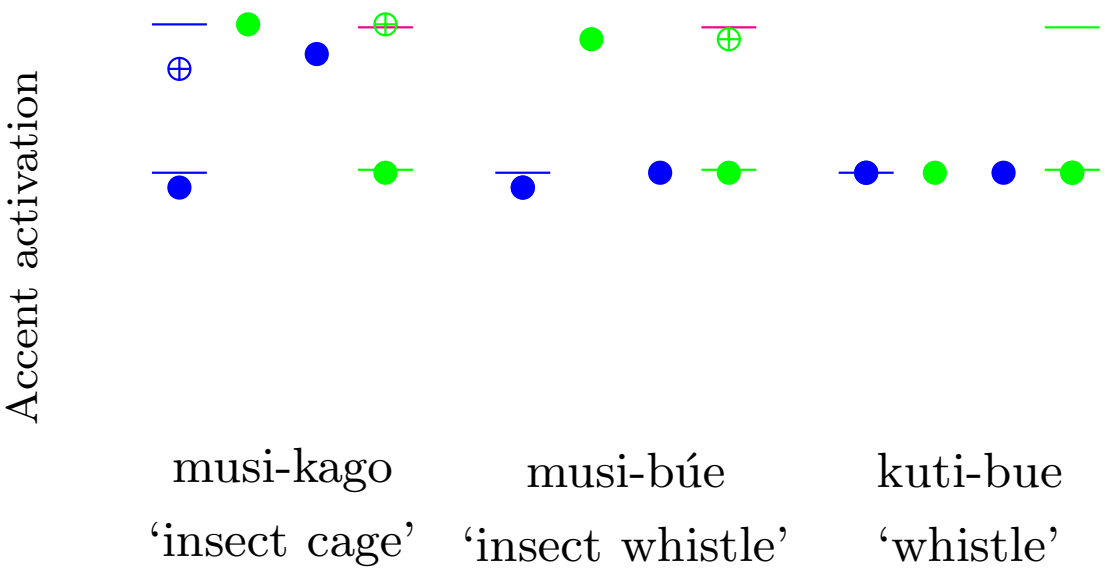
Activations on three sample compounds



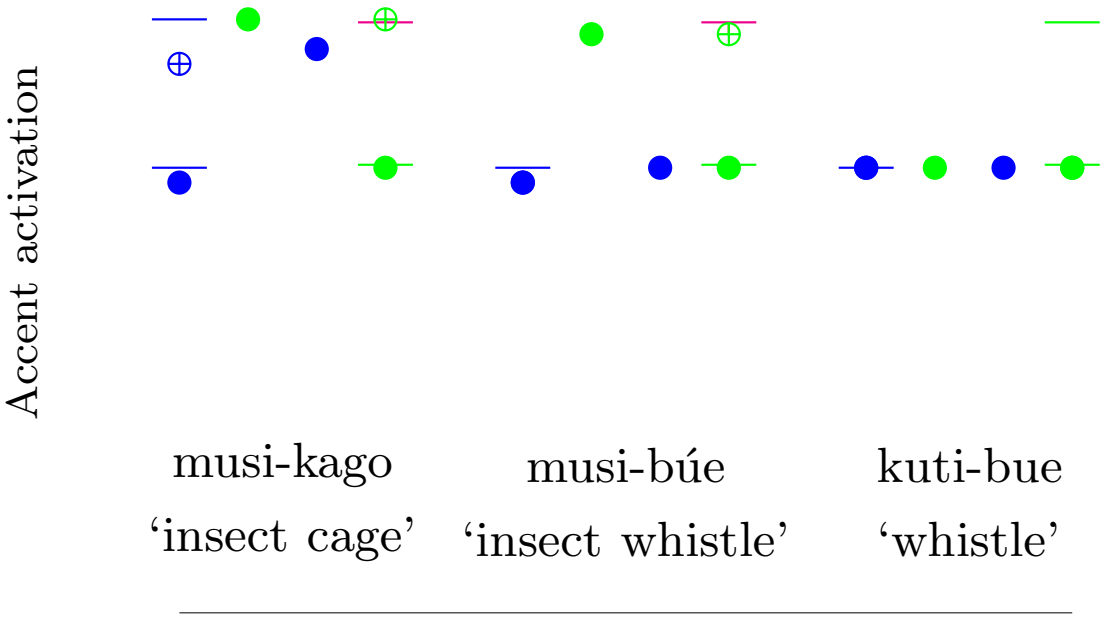
Activations on three sample compounds



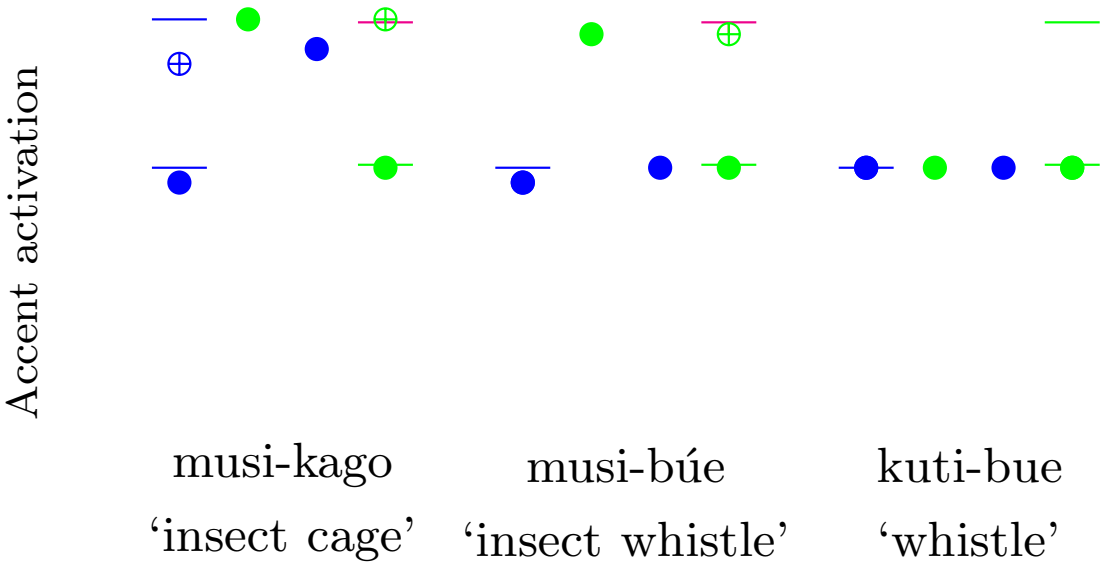
Activations on three sample compounds



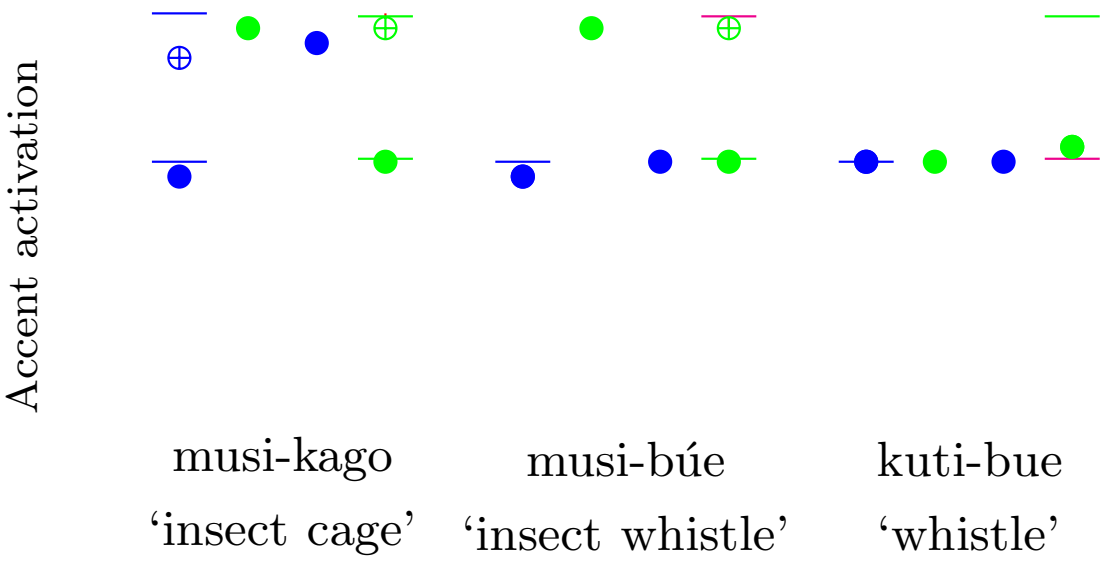
Activations on three sample compounds



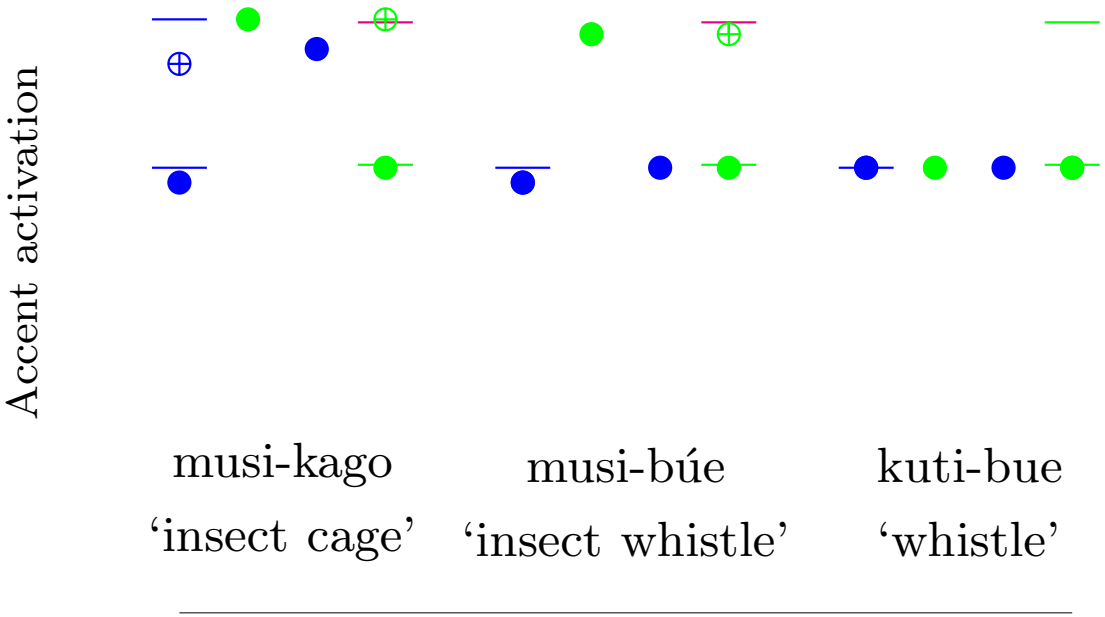
Activations on three sample compounds



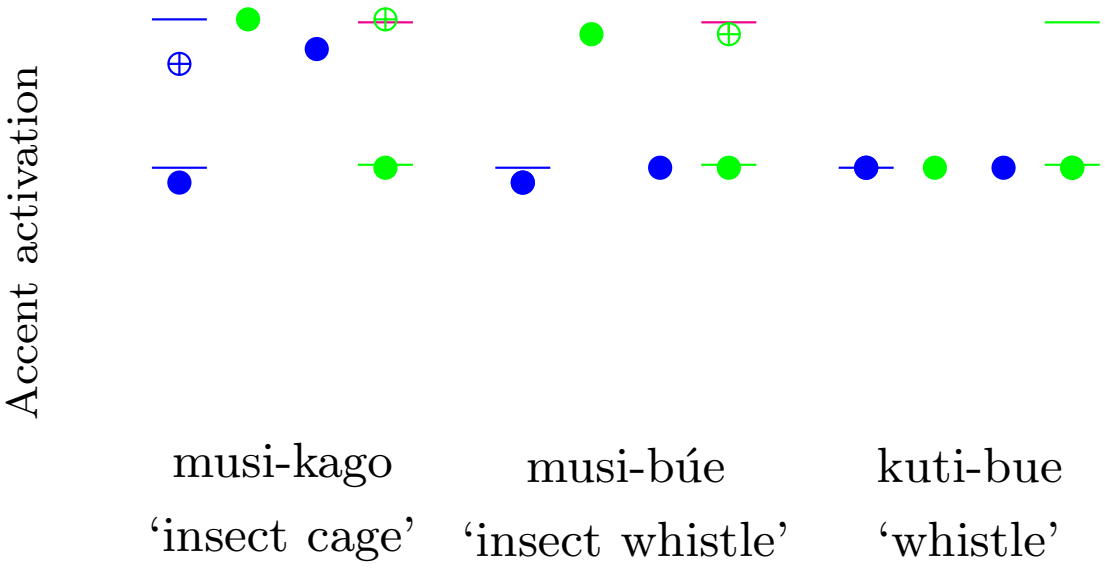
Activations on three sample compounds



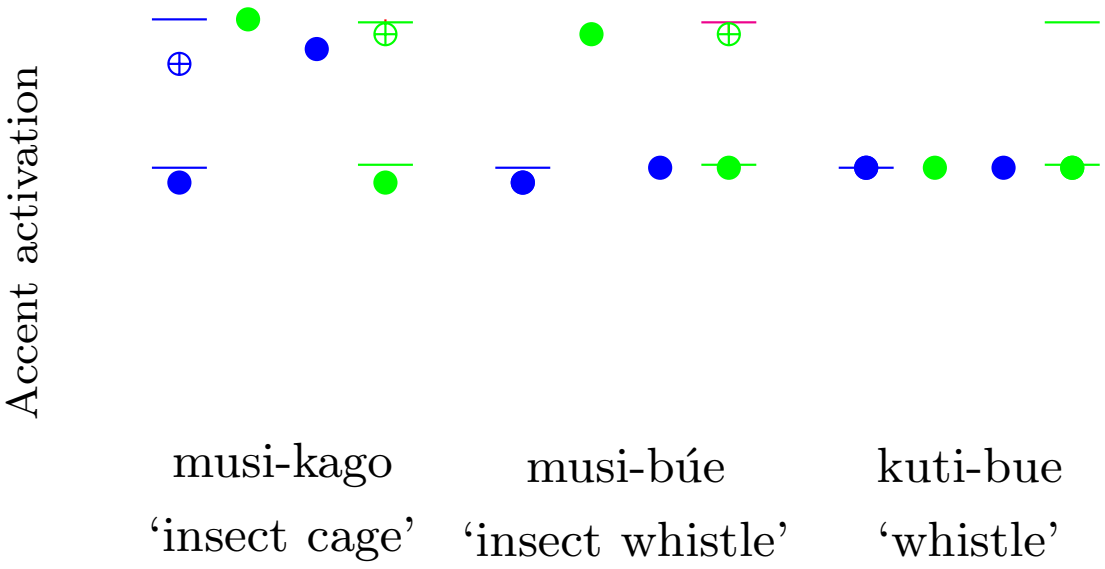
Activations on three sample compounds



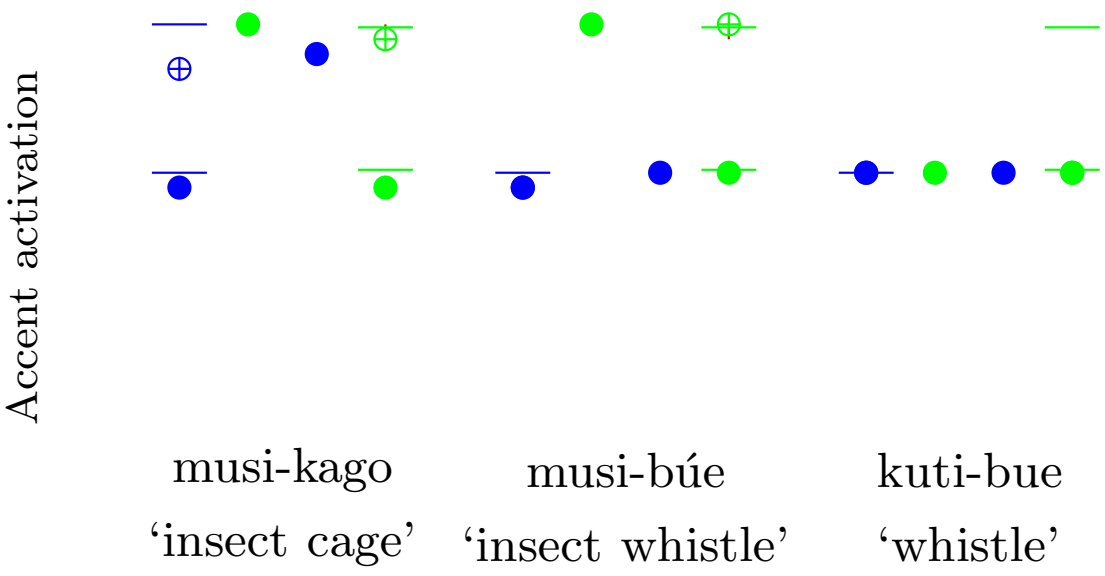
Activations on three sample compounds



Activations on three sample compounds



Activations on three sample compounds



# Further work and lingering issues

- Test this hypothesis with native speaker judgements of fictitious compounds to see if their choice of accent is consistent with observed strength tendencies in real compounds.
- This account works with 100% accuracy for the $2\mu + 2\mu$ words in the database.
- It works not quite as well for $2\mu + 1\mu$, $1\mu + 2\mu$ and $1\mu + 1\mu$ words. (Roughly 90%+ accuracy)
- A native speaker who was consulted considers these even shorter compounds to be just as polymorphemic as the $2\mu + 2\mu$ compounds.

# References

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.

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. ROA 338.

Hyman, Larry. 2009. How (not) to do phonological typology: the case of pitch-accent. Language Sciences 31.213-238.

Itô, Junko and Armin Mester. 2006. Prosodic Adjunction in Japanese Compounds. Proceedings of FAJL (Formal Approaches to Japanese Linguistics) 4, Osaka 2006.

# References

Itô, Junko and Armin Mester. 2015. Word formation and phonological processes. Chapter 9 of Handbook of Japanese Phonetics and Phonology (2015) ed. by H. Kubozono, Mouton de Gruyter Series Handbooks of Japanese Language and Linguistics (Series editors, M. Shibatani, T. Kageyama)

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

Kawahara, Shigeto. 2015. The Phonology of Japanese Accent. Chapter 12 of Handbook of Japanese Phonetics and Phonology. Kubozono, Haruo, ed. De Gruyter Mouton.

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.

Kubozono, Haruo and Armin Mester. 1995. Foot and Accent: New Evidence from Japanese Compound Accentuation. Ms. University of California, Santa Cruz.

# References


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.

Rosen, Eric. 2016. Predicting the unpredictable: capturing the apparent semi-regularity of rendaku voicing in Japanese through harmonic grammar. Proceedings of the 42nd annual meeting of the Berkeley Linguistic Society, February 5-7, 2016.

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.

Yamaguchi, Kyoko. 2011. Accentedness and Rendaku in Japanese Deverbal Compounds. Gengo Kenkyuu 140:117-133.

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

# Acknowledgements

Thanks to Paul Smolensky, Matt Goldrick and two anonymous reviewers from Linguistics Vanguard for helpful comments and suggestions. All errors are my own.