A new model of loanword accentuation in Japanese, based on faithfulness, frequency, and pseudo-compound structure

Hironori Katsuda (University of California, Los Angeles)

Synopsis: It has been widely assumed that loanword accentuation in Tokyo Japanese exhibits a highly regularized pattern, which was first formalized as the antepenultimate accent rule (McCawley 1968), and more recently analyzed in a more elaborate Optimality Theoretic model by Ito and Mester (2016). The assumption shared in this literature is that non-native contrasts in source words, in particular the contrast between stressed and unstressed syllables in English, plays at best a minor role in loanword accentuation. However, this assumption has never been checked against a corpus. My corpus analysis suggests that the accentuation of loanwords borrowed from English is in fact heavily influenced by the non-native contrasts in the source words. Furthermore, my analysis suggests that these faithfulness effects are modulated by the frequency of the source English words (high exposure → high faithfulness) as well as that of the loanwords as Japanese words (frequent use → more nativization). I show that a probabilistic model that integrates these faithfulness and frequency effects covers the data more accurately than existing published models. Finally, my corpus data suggest that Ito and Mester's model wrongly predicts unaccentedness for sequences like LHLL or HHLL. I suggest that Japanese speakers treat loanwords longer than four moras as pseudo-compounds (e.g., [LH-LL], [H-HLL]) and assign an accent based on the compound accent rule.

Data: I extracted all loanwords listed in the *NHK Pronunciation and Accent Dictionary*, and recorded their phonological shapes and accent patterns. In cases where more than one accent pattern were listed, I recorded all accent patterns and treated each variant as an independent word. I then excluded loanwords longer than four syllables and ones that were borrowed from languages other than English, leaving 3265 loanwords. I used the Carnegie Mellon University (CMU) dictionary to annotate the status of each loanword syllable in terms of what it corresponds to in the English source words: primary stressed, secondary stressed, unstressed, or epenthetic. Finally, I used the SUBTLEX corpus and the Balanced Corpus of Contemporary Written Japanese (BCCWJ) to annotate the frequency of the English source words and that of the loanwords as Japanese words, respectively.

Modeling: My models build on Ito and Mester's (2016) classical OT model, which nicely captures what they call the "default" accent patterns of loanwords by the interaction of the foot-based markedness constraints. As a reference model, I first rendered their model probabilistic by employing the Maximum Entropy (maxent) grammar model (Smolensky 1986, Goldwater & Johnson 2003). I fed my corpus data to the model and ran the model to optimize the constraint weights to match the predicted probabilities to the observed probabilities. Figure 1 shows the distribution of the predicted probabilities over that of the observed probabilities.

In integrating the faithfulness effects, I made two assumptions regarding the underlying representation: (i) every loanword has an underlying accent on the syllable stressed in English, (ii) word-initial epenthetic syllables (derived from word-initial consonant clusters in English, such as ***sp****in* and ***sk****y*) are specified with a "weak" vowel in their underlying representation (e.g., /su$_{weak}$pin/). Based on these assumptions, I introduced three faithfulness constraints, shown in (1).

(1) a. IDENT[ACCENT]: Accent should be faithful to underlying representation (i.e., violated by any outputs that have an accent on syllables that are not underlyingly accented)
b. MAX[ACCENT]: Underlying accent should be present (i.e., violated by any unaccented

outputs)

c. DEP[ACCENTWEAKV]: Accent should not be put on a weak vowel (i.e., violated by any outputs that have an accent on an initial epenthetic syllable)

The model with these faithfulness constraints significantly outperforms the reference model, with all the faithfulness constraints gaining a reasonable weight. Contrary to earlier literature, this suggests that the faithfulness effects play a significant role in loanword accentuation.

I also tested two hypotheses regarding the effects of frequency: (i) faithfulness to the stress position, formalized as IDENT[ACCENT], becomes *stronger* as the frequency of the source word goes up, (ii) faithfulness to accentedness, formalized as MAX[ACCENT], becomes *weaker* as the loanword frequency goes up. The second hypothesis is based on the statistical tendency that there are significantly more unaccented words in the native vocabulary than in the loan vocabulary (Kubozono 2006). To test these hypotheses, I divided the loanwords into three bins once by the English frequency and once by the loanword frequency, and took the data per bin as an input. To this end, I split each of IDENT[ACCENT] and MAX[ACCENT] into three sub-constraints corresponding to the three frequency bins. I then fit a maxent grammar with these six sub-constraints in addition to the other constraints for the previous model. This model significantly outperformed the previous model, and the weights of the sub-constraints learned in the model were consistent with the hypotheses: high English frequency induces a greater weight of IDENT[ACCENT], while high frequency as Japanese words means a lower weight of MAX[ACCENT], as shown in Figure 3.

Finally, I address the fact that Ito and Mester's categorical OT model overpredicts unaccentedness. Specifically, their model predicts any loanwords that end with a HLL sequence (e.g., HLL, LHLL, HHLL) as well as the LLLL shape to be unaccented, but the corpus data show that only *four-mora* words that end with a sequence of two light syllables (i.e., LLLL, HLL) are unaccented. I argue that Japanese speakers treat loanwords longer than four moras as pseudo-compounds (Martin 2005) (e.g., [LH-LL], [H-HLL]) as such words are almost always compounds in the native vocabulary, and assign an accent based on the compound accent rule (Kubozono 2006). I formalize this by introducing pseudo-compound structure for longer words into the model and show that this repairs the underprediction of the unaccentedness. The scattergram for the final model is shown in Figure 2.

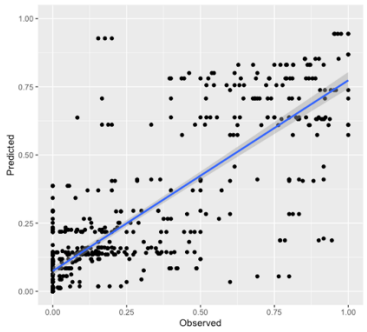

Figure 1

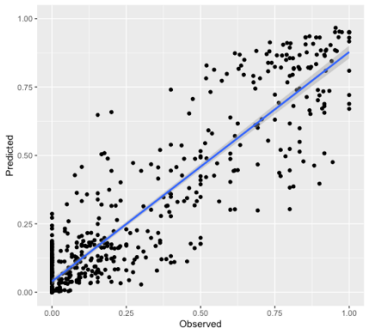

Figure 2

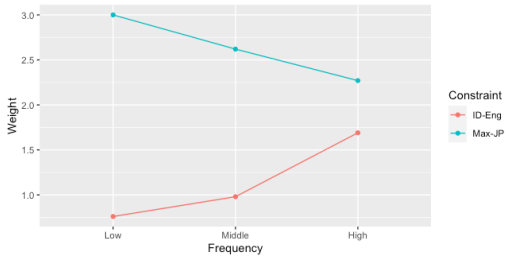

Figure 3

Conclusion: This paper shows that Japanese speakers are faithful to the non-native contrasts in loanword accentuation. Furthermore, it showed that frequency effects exist and have a natural explanation in terms of the borrowing process. Finally, it also showed that positing pseudo-compound status for loanwords longer than four moras better explains the corpus data than the earlier, purely metrical alternative.