Assessing the prosodic licensing of wh-in-situ in Japanese: A computational-experimental approach*

Shigeto Kawahara, Jason Shaw, Shinichiro Ishihara
Keio University, Yale University, Lund University

Abstract

The relationship between syntactic structure and prosodic structure has received increased theoretical attention in recent years. Richards (2010) proposes that Japanese allows wh-elements to stay in-situ because of a certain aspect of its prosodic system. Specifically, in contrast to some other languages like English, Japanese can prosodically group wh-elements together with their licensers. This prosodic grouping is phonetically signaled by eradication or reduction of the lexical pitch accents of intervening words. In this theory, a question still remains as to whether each syntactic derivation is checked against its phonetic realization, or what allows Japanese wh-elements to stay in-situ is more abstract phonological prosodic structure, whose phonetic manifestations can potentially be variable. This paper reports an experiment which addressed this question, by testing whether there is eradication or reduction of lexical pitch accents based on the detailed analyses of F0 contours. Our analysis makes use of a computational toolkit that allows us to assess the presence of tonal targets on a token-by-token basis. The results demonstrate that almost all speakers produce some wh-sentences which show reduction or eradication of the lexical pitch accents, as well as some that do not. Those tokens that show reduction or eradication directly support the prediction of Richard's (2010) theory. The variability observed in the results suggest that the property of Japanese that allows their wh-elements to stay in-situ must be abstract, phonological prosodic structure, whose phonetic realizations can vary within and across speakers. We discuss several possible mechanisms through which such phonetic variation can arise.

*We received helpful comments from the participants at the following conferences, AMP 2019, HisPhonCog 2019, and ICPP 2019, especially Mary Beckman, Ryan Bennett, Edward Flemming, Haruo Kubozono and Mariko Sugahara. Three NLLT anonymous reviewers as well as the Associate Editor Arto Anttila provided very helpful comments which helped us improve the paper. They bear no responsibilities for the remaining errors. This research is supported by NINJAL collaborative research project 'Cross-linguistic Studies of Japanese Prosody and Grammar.'

1 Introduction

The relation between syntactic structure and prosodic structure has received increased theoretical attention in recent years. The standard feedforward model, in which the syntactic derivation is computed first and prosody follows, has been challenged by apparent cases in which prosody and other phonological factors influence the syntactic derivation (e.g. Anttila, Adams, and Speriosu 2010; Bennett, Elfner, and McCloskey 2016; Breiss and Hayes 2020; Shih and Gribanova 2016; Shih and Zuraw 2017 among others).[1] We take up an example from the widely discussed syntactic phenomenon of wh-movement.

Languages differ with respect to whether wh-phrases move overtly or not: Tagalog wh-phrases move overtly, whereas Japanese wh-phrases can stay in-situ (or move covertly at LF). In Minimalist Syntax (Chomsky 1995), this difference was stipulated to derive from a difference in feature strength. Strong (or uninterpretable) features, which Tagalog wh-elements have, need to be checked in the syntax, which requires overt movement, whereas weak features associated with Japanese wh-elements can be checked at LF. This feature-based account restates the difference between overt and covert movement in terms of feature strength. Richards (2010) pursues a more explanatory theory of this cross-linguistic variation in wh-movement and attempts to derive differences between overt movement and covert movement (or lack of overt movement) from independent properties of each language (see also Richards 2016 for an extension of this proposal to a wider range of syntactic phenomena).

More specifically, Richards (2010: 145) argues that there is a universal principle—all languages attempt "to create a prosodic structure for wh-questions in which the wh-phrase and corresponding complementizer are separated by as few prosodic boundaries as possible." This prosodic grouping is accomplished in Tagalog via wh-movement. Inspired by a body of work on Japanese intonational patterns (Deguchi and Kitagawa 2002; Ishihara 2003; Hirotani 2005; Sugahara 2003; Smith 2005), Richards (2010) proposes that Japanese has a prosodic means to group the wh-phrase and its complementizer, and hence does not need to resort to overt wh-movement.[2]

This theory can be interpreted in two ways. One interpretation is that whether a wh-element moves or not is determined for each syntactic derivation against its phonetic realization. The other interpretation is that there is some abstract prosodic feature of Japanese, specifically phonological phrasing, that allows wh-elements to stay in-situ, and it is not necessarily the case that each syntactic derivation is checked against its phonetic realization. To address this question, one of the

[1] See footnote 1 in Bennett, Elfner, and McCloskey (2016) for an extensive list of relevant proposals in which phonological factors seem to influence word order.

[2] Here and throughout, we use the shorthand term "Japanese" to refer to "Tokyo Japanese." Smith (2011) argues based on data from Fukuoka Japanese that it is the complementizer, not the wh-elements, that derives this phrasing pattern. We are not concerned in this paper with what triggers this prosodic grouping. Our concern is instead on whether this prosodic grouping indeed occurs or not in Tokyo Japanese, and if so, how this prosodic grouping manifests itself phonetically.

aims of this paper is to elicit prosodic patterns from naive participants in a controlled experiment in order to examine if—and how—wh-sentences in Japanese show phonetic evidence of prosodic grouping between wh-elements and their licensers. Our results show that there is phonetic variation both within and across speakers. Those tokens that show eradication or reduction of lexical pitch accents can be taken to directly support Richards' (2010) proposal. To accommodate the observation that there are tokens which do not show either eradication or reduction, we maintain that Richard's (2010) theory has to be based on abstract phonological prosodic structure, which can be variably realized phonetically. We discuss some concrete mechanisms through which such phonetic variations can arise.

Some background information regarding the Japanese prosodic system is in order. In Japanese, a Minor Phrase (MiP) contains at most one accented lexical item, and is signaled by a phrase-initial %LH rise and a H*+L accentual fall (e.g. Igarashi 2015; Ishihara 2015; Pierrehumbert and Beckman 1988; Venditti 2005; Venditti, Maekawa, and Beckman 2008).[3] Deguchi and Kitagawa (2002), one source of inspiration for Richards (2010), argue that these tonal events associated with Minor Phrases are *eradicated* after wh-elements up to the complementizer that licenses the wh-elements, effectively grouping the wh-phrase and complementizer within a single Minor Phrase.[4] This eradication is accompanied by a boost of accentual rise on the wh-element itself, which is arguably an instance of a more general focus-induced prominence boost (Igarashi 2015, Ishihara 2015 and Venditti, Maekawa, and Beckman 2008 and many references cited therein). Deguchi and Kitagawa (2002: 74) state:

> Another important prosodic effect of focus pointed out by Ishihara (2000) (extending the original observation by Ladd (1996)) is that an emphatic accent is accompanied by what we label as "**eradication**" of lexical accents. That is, when one or more of [the] lexical accents follow an emphatic accent, their H tones (H*) are all suppressed. As a result, the lowest F0 induced by the emphatic accent is inherited and prolonged with further gradual declination up to the right boundary of some clausal structure (emphasis in the original).

[3] A Minor Phrase is also known as an Accentual Phrase. A Major Phrase, the level above a Minor Phrase, is also known as an Intermediate Phrase. Terminological differences do not concern us much here (see Igarashi 2015 for a recent systematic review). We use the term Minor Phrase, because this is what Richards (2010) uses. See Richards (2016) for a proposal which deploys a recursive prosodic structure (ϕ) without a Minor Phrase/Major Phrase distinction (e.g. Ito and Mester 2012). In this paper we follow Richards' (2010) conventions, as the predictions regarding phonetic realizations are straightforward to illustrate. Nothing in this paper hinges upon this choice of this particular set of terminologies, however.

[4] See also Igarashi (2015), Pierrehumbert and Beckman (1988), Venditti et al. (2008) and Ishihara (2015) and works cited therein for (de)phrasing that may occur in post-focal positions in general. Most of these studies, however, posit that dephrasing occurs at the level of the Major Phrase rather than the Minor Phrase. Here we focus on the proposal by Deguchi and Kitagawa (2002), which Richards (2010) builds upon. This paper specifically analyzes those contexts that are relevant to wh-constructions in Japanese.

An oft-cited pair of pitch tracks to illustrate this observation is provided in Figure 1, which is reproduced from Ishihara (2003: 53). The top panel shows a declarative sentence consisting of four words which are all lexically accented. The bottom panel shows a corresponding wh-sentence, in which the second word is a wh-element, *nani-o* "what-ACC," shown by the thick arrow. The domain of eradication (or reduction: see below) in the wh-sentence is shown in grey.

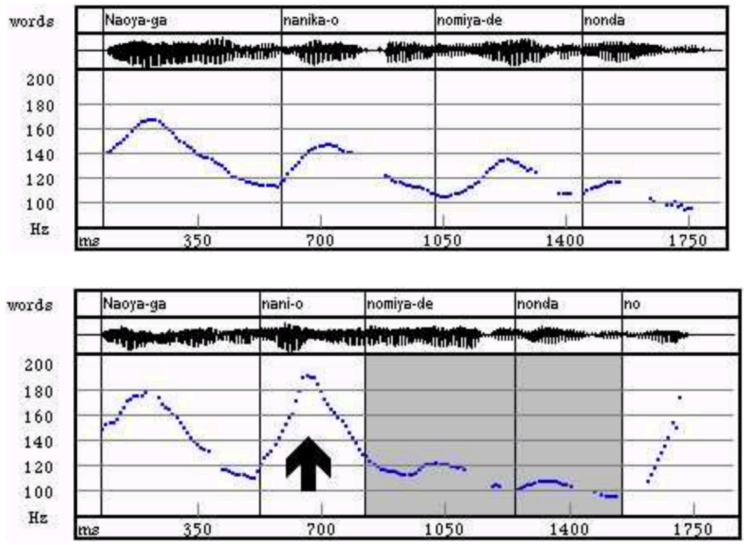

Figure 1: Illustrative pitch tracks of a declarative sentence and a wh-sentence in Japanese. Reproduced from Ishihara (2003: 53). These tokens are based on the production of Ishihara himself. Top: declarative, "Naoya drank something at a bar." Bottom: wh-sentence, "What did Naoya drink at a bar?" The wh-element (=*nani-o*) is shown by the thick arrow. The domain of eradication is shown in grey.

If Deguchi and Kitagawa's (2002) observation is correct, then we have a very simple story in accordance with Richards' (2010) theory—Japanese groups wh-elements and the complementizer within a single Minor Phrase without any intervening Minor Phrase boundaries. The prosodic structure that would reflect Deguchi and Kitagawa's observation is schematically depicted in (1). This structure is also entertained by Richards (2010: 145), although he ultimately adopts a different structure for Japanese, to be discussed below, in (3).

(1) $_{\text{MiP}}$[wh DP DP V Comp]

Compare (1) with the prosodic structure of the declarative sentence in (2). In the declarative sentence, each lexical item is parsed into a separate MiP, and the whole predicate is parsed into a higher phrase (MaP or a higher recursive ϕ in Ito and Mester's 2012 model; see footnote 3). Given this structure, the accent of each lexical item is predicted to be realized.

(2) $_{\text{MaP}}[_{\text{MiP}}[\text{DP}]\ _{\text{MiP}}[\text{DP}]\ _{\text{MiP}}[\text{V}]]$

If (1) is correct, any tonal events in the intervening DPs should be lost, *à la* Deguchi and Kitagawa (2002); i.e. the intonational contour should be "flat" throughout the whole Minor Phrase. We note at this point that the claim by Deguchi and Kitagawa (2002) is based on the intonational contours produced by the authors (p. 75). They themselves note the preliminary nature of the data, and it seems necessary that we test the claim about tone eradication with more objective methodology. We provide this test in the current study.

Some later studies cast doubt on Deguchi and Kitagawa's (2002) claim that accent following wh-elements is completely eradicated; instead, the post-wh accents may simply be reduced (Hirotani 2005; Ishihara 2003; Ishihara 2011a; Sugahara 2003—see also Maekawa 1994). Ishihara (2003: 32) for example notes that if all tonal events after wh-elements are eradicated, the lexical distinction between accented and unaccented words should be neutralized after wh-elements. Though he does not report any experimental results, Ishihara suggests that this neutralization does not occur. Hirotani's (2005) production study showed substantial variation across speakers in the degree of compression after wh-elements. These studies have thus raised the possibility that total eradication of lexical pitch accents in relevant wh-sentences may not be observed, contrary to the claim by Deguchi and Kitagawa (2002).

Citing Kubozono (2007), Richards (2010: 149-151) posits the prosodic structure schematized in (3) for Japanese wh-sentences, in which wh-elements and the complementizers are grouped in a higher recursive Minor Phrase, making use of the recursive Minor Phrase first proposed for Japanese by Kubozono (1988).[5] Unlike (1), the prosodic structure in (3) predicts that accentual rises in the intervening DPs should be reduced, compared to when they occur in non-wh-contexts.

(3) $_{\text{MiP}}[_{\text{MiP}}[\text{wh}]\ _{\text{MiP}}[\text{DP}]\ _{\text{MiP}}[\text{DP V Comp}]]$

Ishihara (2011a) reports an instrumental study that suggests that reduction (instead of eradication) takes place after wh-elements, as predicted by the structure in (3). However, as we will discuss more fully below, Ishihara (2011a) only examines averaged contours, and a token-by-token

[5]Recent work has identified potential problems with recursive Minor Phrase, raising an alternative possibility that the higher prosodic level may be a Major Phrase. The issues with the recursive Minor Phrase is that since Minor Phrases are usually defined in terms of accent culminativity (i.e. at most one accent), a recursive structure should not be possible (Ito and Mester 2012), except for very special cases in which all the terminal Minor Phrases contain an unaccented item. For our purposes, we follow Richards (2010) in positing MiP as the higher prosodic level. If the higher level prosodic level is indeed a Major Phrase, then we would need to posit that the tonal events of the intervening DPs should be reduced due to post-focal reduction (e.g. Ishihara 2011b; Pierrehumbert and Beckman 1988; Sugahara 2003) in addition to independently observed downstep, whose domain is a Major Phrase (McCawley 1968; Pierrehumbert and Beckman 1988; Poser 1984) (cf. Ishihara 2016).

If we adapt the model proposed by Ito and Mester (2012), which does not distinguish MiP and MaP and instead posits recursive ϕ, as Richards (2016: 81-83) does, we would still have to posit that both downstep, whose domain is ϕ, as well as post focal reduction to distinguish between declarative sentences and wh-sentences.

analysis is not reported. As we show below, however, looking at averaged contours can be misleading. From averages, it is not possible to differentiate between reduction and variability, phonetic outcomes which often map onto distinct theoretical proposals both in this case and more generally (Cohn 2006; Shaw and Kawahara 2018a). In other words, if the phonetic realizations are variably either "fully realized" or "completely eradicated," averaging over these two categorical generalizations would not be distinguishable from "reduction." To tease apart these two distinct possibilities, we need to analyze intonational contours on a token-by-token basis.

To summarize, Richards' (2010) theory of wh-movement assumes that Japanese groups wh-elements and their licensers in one way or another. Since past work has shown variation in F0 patterns following wh-elements (e.g. Hirotani 2005), we feel that it is important to test for tone eradication/reduction on a token-by-token basis. The time is particularly appropriate given the recent development of new computational methods for assessing on a token-by-token basis whether a phonetic target is reduced, deleted or fully realized (Shaw and Kawahara 2018a). This new method allows us to tease apart two possible interpretations of Richards' (2010) theory as well: whether each syntactic derivation is checked against its phonetic realization, or whether what allows Japanese wh-elements to stay in-situ is more abstract phonological structure, whose phonetic realizations can potentially be variable.

2 The current study

Against this theoretical background, this paper reexamines wh-phrase conditioned tone eradication/reduction in Japanese. We adopt a token-by-token analysis of intonational contours, using the computational toolkit developed by Shaw and Kawahara (2018a). The basic approach is to classify intonational contours based upon competing phonological hypotheses, in this case the presence/absence of an H tone. Each intonational contour is assigned a probability of being generated from a LHL pitch accent or a LΦL pitch accent, where Φ represents syllables unspecified for F0. A key step in mapping the continuous trajectories to discrete phonological hypotheses is a low-dimensional representation of the F0 signal, which is accomplished using Discrete Cosine Transform (e.g. Jain 1989). More broadly, the approach fits within the broader analytical strategy of functional data analysis (e.g. Beddor et al. 2018; Krivokapić, Styler, and Parrell 2020; Lee, Byrd, and Krivokapić 2006), by which the low dimensional representation is achieved by fitting non-linear functions to the data.

A key innovation of the approach is that phonetic interpolation, characteristic of LΦL, is formalized stochastically. First developed to assess the presence/absence of a lingual vowel target of devoiced vowels in articulatory trajectories (Shaw and Kawahara 2018b), the approach is general and has been extended to other types of continuous phonetic data, including nasal reduction in

Ende (Brickhouse and Lindsey 2020) and tone reduction in Mandarin Chinese (Zhang, Geissler, and Shaw 2019). In addition to being able to analyze each tonal contour on a token-by-token basis, this method has an advantage of being able to analyze the whole intonational contour without relying on "magic moments" (Mücke, Grice, and Cho 2014), i.e. particular aspects of the phonetic signal, such as F0 minima or maxima, are not given any privileged status in the analyses, eschewing the potential danger of missing important aspects of dynamic speech (Cho 2016; Mücke, Grice, and Cho 2014; Vatikiotis-Bateson, Barbosa, and Best 2014).

To preview the results, our analysis shows that there is prosodic variation. Almost all speakers produce some sentences that behave precisely as predicted by Richards' (2010) theory, as well as some that do not. Those tokens that show eradication or reduction of the intervening lexical pitch accents directly support Richards' (2010) theory. Those tokens that show neither reduction nor eradication may be taken to present a challenge to Richard's (2010) theory. To accommodate this observation, we maintain that what licenses Japanese wh-elements in-situ should be abstract phonological prosodic structure, but that that prosodic structure can be variably realized phonetically. To the extent that Richards' (2010) theory is on the right track, it is not the case that each syntactic derivation is checked against its phonetic realizations—evidence for prosodic grouping is not always "visible" from surface phonetics alone for each and every sentence.[6] The overall conclusion with respect to the case under study is that, while language-specific syntax may operate in the presence of language-specific prosodic patterns, what matters is abstract prosodic structures rather than their token-by-token phonetic realizations.

2.1 Methods

2.1.1 Materials

The current study reanalyzes a subset of the data recorded by Ishihara (2011a). The corpus features carefully controlled pairs consisting of a declarative sentence and wh-question counterpart. All sentences consisted of five words. Schemata of the item pairs are given in (4) and (5), together with one example sentence for each condition. Accent is shown by an apostrophe (') following accented vowels. Word2, Word3, and Word4 were all lexically accented. For wh-questions, the second word was the wh-phrase. There were six types of sentences with different lexical items for (4) and (5).

[6] As we discuss below in §3.2 in some detail, positing that the same prosodic structure can receive various phonetic realizations is not necessarily an ad hoc stipulation, because mechanisms which can derive this sort of variation are independently motivated. It suffices to point out at this point that generally speaking, it is not uncommon to observe variable phonetic realizations of one phonological structure.

(4) Declarative sentence (control): Word1 Word 2[-wh] Word3 Word4 Verb

Maruyama-wa e'rumesu-no eri'maki-ni nomi'mono-o kobo'shi ma'shi-ta
NAME-TOP Hermes-GEN scarf-DAT drink-ACC spill POL-PAST

"Maruyama spilled drink over Hermes scarf."

(5) Wh-question sentence (test): Word 1 Word 2[+wh] Word3 Word4 Verb

Maruyama-wa do'nohito-no eri'maki-ni nomi'mono-o koboshi ma'shi-ta ka?
NAME-TOP whose scarf-DAT drink-ACC spill POL-PAST COMP

"Whose scarf did Maruyama spill drink over?"

The declarative sentences, as in (4), serve as the control sentences. Both Word3 and Word4 are Minor Phrases; as such, the phrasal tones (%LH) and lexical accent (H*+L) of both Word3 and Word4 are realized (i.e. full target), as shown in a sample pitch track given in Figure 2 (top).

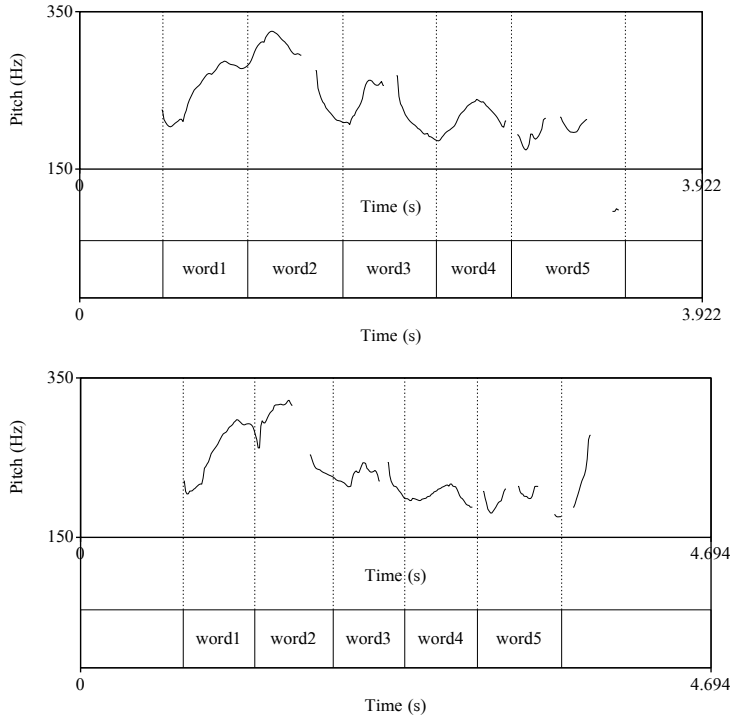

Figure 2: Sample pitch tracks from the current recordings. Top: declarative sentence (=sentence (4)). Bottom: wh-sentence (=sentence (5)).

We are interested in whether Word3 and Word4 in (5) (=Figure 2, bottom) also fully retain these %LH and H*+L tones, or whether these tones are completely eradicated or reduced.

2.1.2 Recording procedure

The stimuli contained six items per each of the two conditions shown in (4) and (5) (i.e. $2 \times 6 = 12$ target sentences). They were recorded together with another 142 filler sentences (which were recorded for other studies).

Nine native speakers of Tokyo Japanese (4 females and 5 males) read all the stimulus sentences, twice each. One stimulus sentence was presented to a speaker per trial on a computer screen. Speakers were allowed to repeat the sentence when they made a mistake or they felt that their utterance was unnatural, in which case only the last rendition was used for the following analysis. The order of the stimuli was pseudo-randomized. Two recordings per speaker were made in two different randomized orders. A total of 432 tokens (12 target sentences read by 9 speakers twice each for Word3 and Word4) entered into the subsequent analysis.

2.1.3 The computational analysis

The intonational contours of Word3 and Word4, delimited by %L and +L, were extracted using YAAPT, a robust pitch tracking algorithm (Yet Another Algorithm for Pitch Tracking: Zahorian and Hu 2008). We then applied the computational toolkit developed by Shaw and Kawahara (2018a), whose details are illustrated below.

The starting point of the approach is to recognize that, given a phonetic trajectory, it is often hard to decide, especially if we rely on visual inspection of pitch tracks, whether that trajectory should be characterized as linear interpolation between two targets (with declination), or whether it has a distinct phonetic target (i.e. in the case at hand, a H tone), as schematically illustrated in Figure 3. This is particularly so because actual intonational contours always involve natural variability, due to various factors such as consonantal perturbations (Hombert, Ohala, and Ewan 1979), the influences of vowel height (Whalen and Levitt 1995), and others. In other words, intonational contours are usually "bumpy", and it is often hard to tell whether a "bump" comes from an actual phonological specification or merely due to random variation (noise).

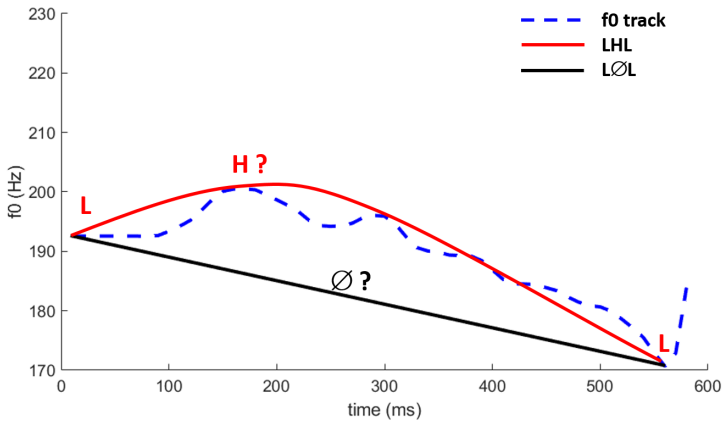

Figure 3: Deciphering whether a contour has a H(igh) tonal target. The dotted blue line, representing an actual phonetic contour, is shown together with phonetic schemata for competing phonological hypotheses, LΦL (solid black line) and LHL (solid red line).

To address this question in an objective fashion, we first transform the phonetic trajectories (i.e. intonational contours) into a low-dimensional space, so that we have a mathematical handle on them. For this purpose, we use Discrete Cosine Transform (DCT) to transform phonetic signals from the time domain (changes in F0 over time) to the frequency domain (sums of cosines of different frequencies and amplitudes). The numerical expression of DCT is given in formula (1)-(2). Like Fourier Transform, this analysis decomposes trajectories into a set of DCT components with different frequencies and different amplitudes (Jain 1989).

$$y(k) = w(k) \sum_{n-1}^{L} \cos \frac{\pi(2n-1)(k-1)}{2L} \quad k = 1, 2, ...L \tag{1}$$

where

$$w(k) = \begin{cases} \frac{1}{\sqrt{L}} & k = 1 \\ \sqrt{\frac{2}{L}} & 2 \leq k \leq L \end{cases} \tag{2}$$

Figure 4 illustrates this decomposition procedure with an example. The top panel shows an example F0 contour. This contour can be decomposed into a set of four cosine components, shown in the four panels below.

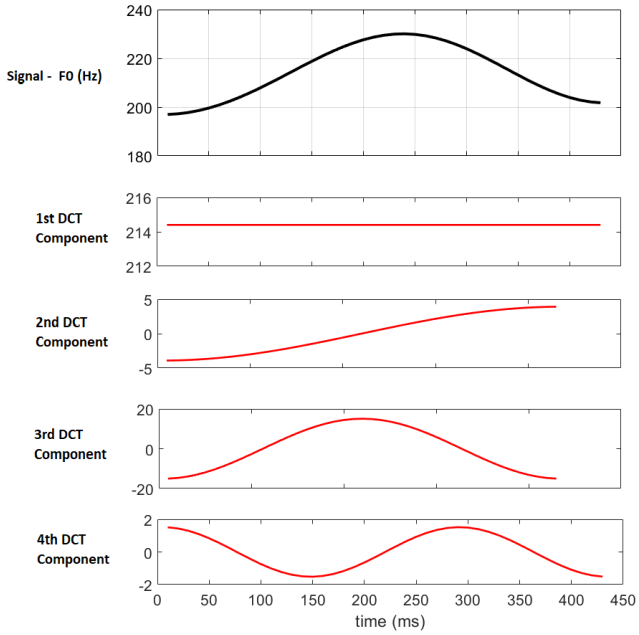

Figure 4: An example of a DCT analysis. The top panel represents an example F0 signal. The DCT components of the original contour are shown in the bottom four panels.

The F0 contour across Word3 and Word4 can be represented faithfully in the frequency domain using four DCT components. We determined this using the inverse function for DCT, iDCT, which transforms cosines (frequency domain) back to F0 trajectories in the time domain. The numerical expression of iDCT is given in formula (3)-(5).

$$y(k) \sim N(\mu(k), \sigma(k)) \tag{3}$$

$$x(n) = \sum_{n-1}^{L} w(k)y(k)\cos\frac{\pi(2n-1)(k-1)}{2L} \quad n = 1, 2, ...L \tag{4}$$

where

$$\begin{cases} \frac{1}{\sqrt{L}} & k = 1 \\ \sqrt{\frac{2}{L}} & 2 \leq k \leq L \end{cases} \tag{5}$$

Using iDCT, we simulated pitch trajectories from different numbers of DCT coefficients and compared the simulated trajectories to the actual trajectories. For the case at hand, using four DCT coefficients achieves higher than 95% fit between actual and simulated trajectories. Figure 5 illustrates the increase in correlation between actual F0 trajectories and simulated F0 trajectories

as a function of the number of DCT components. Comparisons between simulated contours and the actual contours, when we use four DCT components, are exemplified in Figure 6.

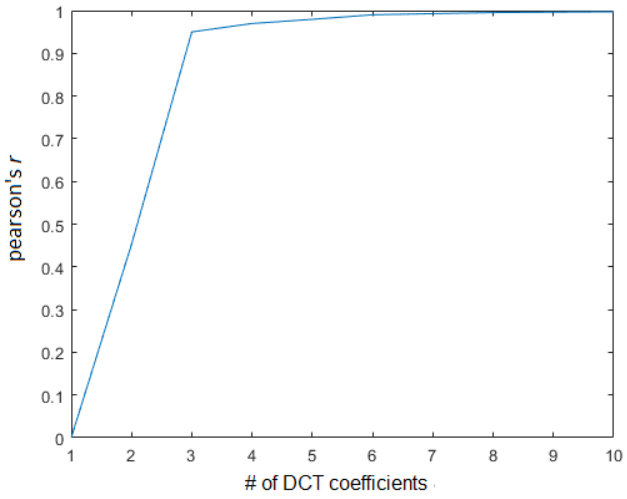

Figure 5: The correlation (Pearson's r) between real and simulated F0 trajectories based on increasing numbers of DCT coefficients.

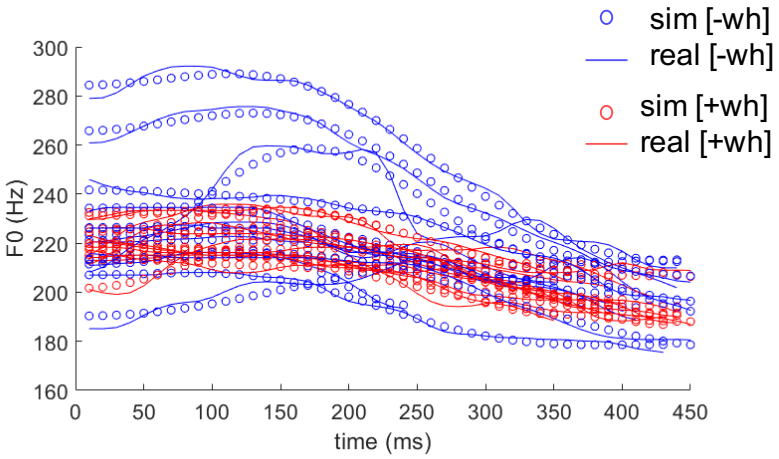

Figure 6: Examples of simulated contours using iDCT (lines consisting of circles), compared to actual F0 contours (solid lines).

Having verified that we can faithfully represent F0 trajectories (time domain) with four DCT coefficients (frequency domain), we proceeded to set up stochastic generators of our two competing phonological hypotheses, LHL and LΦL in Figure 3, in the frequency domain. Gaussian

distributions over DCT components for the control sentences (4) describe the LHL hypothesis. Both the mean and the standard deviation of the distributions were determined by the data.

Gaussian distributions for LΦL were defined with reference to linear interpolation between the two L tones in the test sentences (5), which corresponds to the straight line in Figure 3. The mean of the distributions was determined by the DCT components fit to the linear interpolation. The standard deviation was set to the same standard deviation for the LHL hypothesis. The LΦL is thus a realization of linear interpolation with the same level of variability as the control sentences (4). Contours created from the LΦL generator, using iDCT, simulate how the LΦL line in Figure 3 would be phonetically realized given a naturalistic amount of variability, i.e. the level of variability in the data.

Finally, we used the stochastic generators of our phonological hypotheses as a Bayesian classifier (formula (6)). The posterior probability (the left term of the equation) is the probability of the hypothesis (H_j) given the evidence and the prior probability of the hypothesis. In this case, we assumed that each hypothesis was equally likely, a priori, so that the posterior is dictated only by the evidence. The evidence is the set of DCT coefficients. The right side of the equation is the prior probability, $p(H_j)$, which is always 0.5 in our classifier, multiplied by the product of conditional probabilities of each coefficient given that hypothesis (i.e. "evidence"), normalized by the denominator, the product of the probabilities of the coefficients. This classifier assigned posterior probabilities to each test token, i.e. Word3 and Word4 following wh-elements in the test sentences (5). The posterior probability represents the likelihood that the token was generated by one phonological hypothesis or the other. Since the probabilities are complementary, we report, for each token, the posterior probability that it was generated by the LΦL hypothesis, indicating tone eradication.

$$p(H_j|Co_1, ..., Co_n) = \frac{p(H_j) \times \prod_{i=1}^{n} p(Co_i|H_j)}{\prod_{i=1}^{n} p(Co_i)} \tag{6}$$

2.2 Results

Figure 7 shows the posterior probability of eradication for Word3. Tokens on the right have a high probability of being generated by the LΦL model (=tone eradication, linear interpolation). Tokens on the left have a high probability of being generated by the LHL model, as in our control sentences (4) (=full target). Many speakers (Speakers 1, 2, 4, 5, 7, 8, 9) show at least some tokens that have a high posterior probability of eradication (right). These tokens thus support the view expressed by Deguchi and Kitagawa (2002); i.e. they instantiate tonal patterns deriving from the prosodic structure posited in (1) and directly support Richards' (2010) argument that wh-elements

and their licensers can be grouped within a single Minor Phrase in Japanese.

However, Speakers 6, 7, 8, and 9 show a large number of tokens that have a high probability of being generated by the LHL model, i.e. no different from tonal realizations in declarative sentences. These tokens show no or very little trace of reduction (near 0 probability of LΦL). We finally observe those tokens whose posterior probabilities are in the middle range (Speakers 2, 4, 5, 7). These tokens show properties that are intermediate between the two phonological hypotheses, and thus are best viewed as phonetically reduced, instantiating tonal patterns predicted by the prosodic structure in (3). For most speakers (all except Speaker 3), within-speaker variability is also evident.

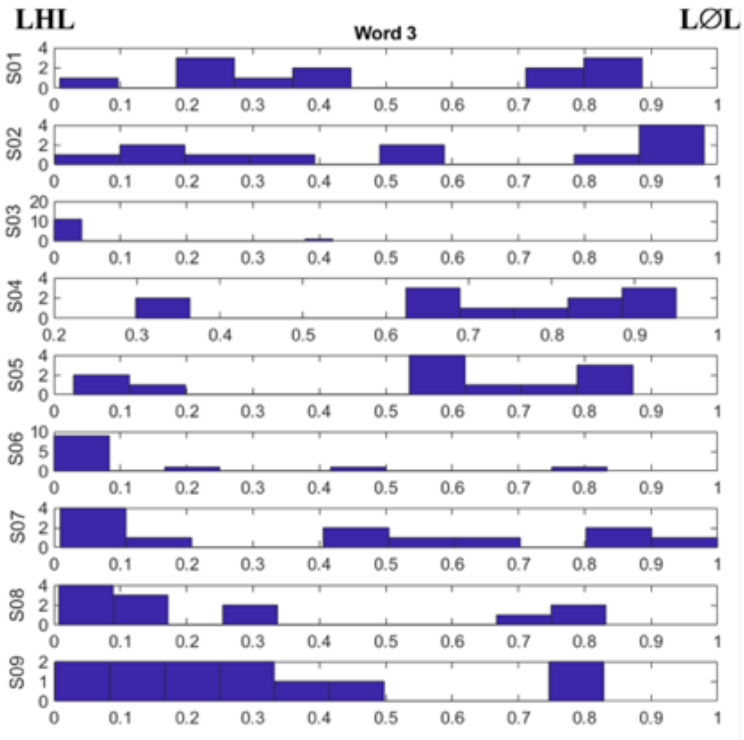

Figure 7: Posterior probabilities of tone eradication for Word3. For visibility, the scale of the vertical axis is optimized for each speaker.

Figure 8 shows the posterior probability of eradication for Word4. The structure of the figure is the same as Figure 4. Most speakers (all but Speakers 1 and 6) show several tokens of high eradication probability (right). Speakers 1, 3, 5, 6, and 8 also produced a number of tokens with full tonal targets (left) and there are many tokens that are phonetically reduced as well (middle). Again, just like Word3, we observe both inter- and intra-speaker variability.

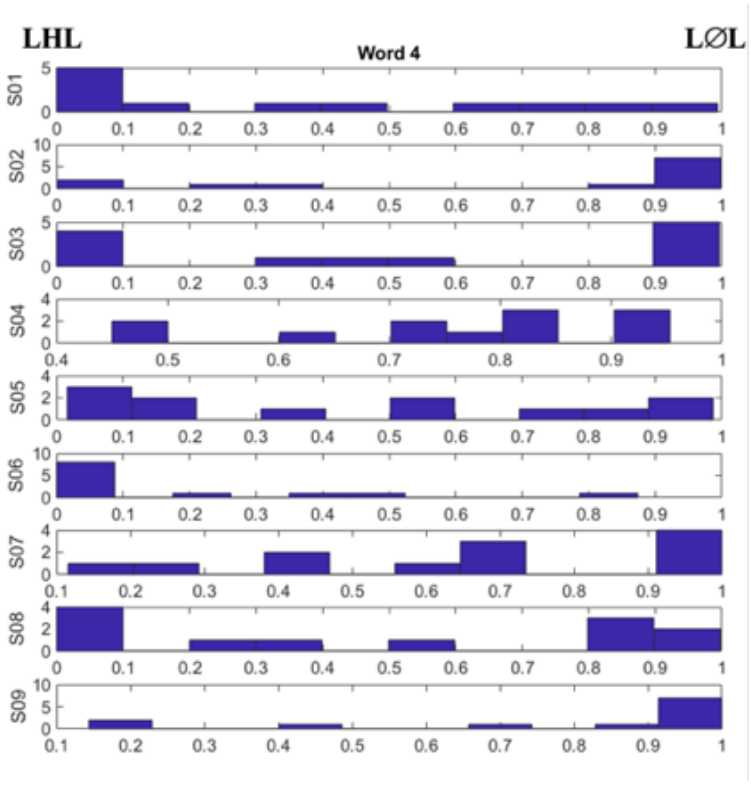

Figure 8: Posterior probabilities of tone eradication for Word4. For visibility, the scale of the vertical axis is optimized for each speaker.

At this point we would like to highlight the importance of analyzing each token separately, instead of just looking at averaged contours, as Ishihara (2011a) did. Take the case of Word4 for Speakers 3 and 4, for example. Speaker 4 shows reduction for many tokens; Speaker 3 on the other hand shows a bimodal distribution of full targets and eradication. If we were to be only looking at averages, we would have erroneously concluded that both speakers show reduction (Cohn 2006; Shaw and Kawahara 2018a). This comparison highlights the importance of analyzing each token separately. We would also like to note that our analyses do not grant any privileges to "magic moments" in intonational contours, such as F0 minima or maxima, but analyze the whole contours in their entirety, thus eschewing the danger of missing important aspects of dynamic speech (Cho 2016; Mücke, Grice, and Cho 2014; Vatikiotis-Bateson, Barbosa, and Best 2014).

3 General Discussion

3.1 Summary of the current results

Our analysis shows that most speakers do show some tokens in which %LH and H*+L tones associated with MiPs are eradicated (i.e. tokens which should be characterized as linear interpolation between %L and +L), supporting the prosodic structure in (1). Some other tokens were best viewed as showing reduction, suggesting the prosodic structure in (3). These tokens can be taken to offer direct support for Richards' (2010) theory. At the same time, however, no speakers consistently show eradication or reduction, and more importantly, there are tokens that show a high probability of LHL structure (i.e. no eradication or reduction). These particular tokens may be taken to present a challenge to Richards (2010), although as discussed in further detail below in §3.2, this challenge is not insurmountable.

Viewed from a slightly different perspective, the current results allow us to tease apart two interpretations of Richards' (2010) theory, laid out in the introduction: whether (i) each syntactic derivation is checked against its phonetic manifestations, or whether (ii) the relationship between wh-movement and the prosodic structure is more abstract. Our results are consistent only with the second interpretation, because there are tokens that show no phonetic sign of reduction or eradication. The first interpretation would predict that all sentences with wh-elements in-situ should show phonetic evidence for prosodic grouping between the wh-element and its complementizer. Our data show that this is not always the case, but there is instead phonetic variation. Nevertheless, with the possible exception of Speaker 6, who produced just two out of 24 tokens (one for Word3 and one for Word4) with a high (greater than .6) probability of LΦL, all of our speakers produced at least some tokens with a high probability of tone eradication. That is, prosodically grouping the wh-word and the complementizer is a strategy that is available to all speakers, even though it is not obligatory.

3.2 Variable mappings from phonology to phonetics

To the extent that Richard's (2010) theory is on the right track—and this is what many of the tokens in the current experiment do seem to suggest—it cannot be the case that each syntactic derivation is checked against its phonetic realization. Instead, it seems necessary that Japanese wh-sentences have either (1) or (3) as their prosodic structures; however, these prosodic grouping patterns do not necessarily result in eradication or reduction of pitch accent. To accommodate these observations, we would have to posit that the relationships between prosodic structure and the phonetic signal can be variable.[7]

[7]This possibility may have been anticipated by Richards, when he states (2010: 148) that "[w]hat kind of effect these wh-domains have on F0 is not part of the theory: wh-domains might involve F0 compression, a high tone, or

This postulation—positing variable phonetic realizations of a particular prosodic structure—is not particularly ad hoc, to the degree that mechanisms that can derive this sort of variations have been independently proposed in the literature. Here we briefly entertain three candidates for such mechanisms. The first possibility is to posit exemplar representations and a production-perception feedback loop (e.g. Pierrehumbert 2001; Wedel 2007). More specifically, it has been shown that the phonetic form of words takes on the characteristics of the environments in which they are typically produced, which has been shown for duration patterns in English (Seyfarth 2014). Another example comes from Mandarin, in which words that are typically produced in unpredictable environments and likely to receive focus are produced with higher maximum F0, greater intensity and longer duration, even when produced in predictable (unfocused) environments (Tang and Shaw 2020). If an accented word is typically produced with a lexical pitch accent in Japanese, as it would be in declarative sentences, then the phonetic detail associated with this word may retain that F0 increase even when produced in an environment that licenses eradication/reduction of the phonological tone (i.e. in wh-sentences).

The second possible mechanism, which is related to the first possibility just described above, is to resort to phonetic paradigm uniformity effects (Braver 2019; Steriade 2000; Yu 2007). Phonetic realizations in the wh-contexts, while being coerced to be reduced or eliminated, can nevertheless be realized with full accent because they are required to have the same phonetic realizations as those that are produced in declarative contexts.

The final possibility is to posit competition between allomorphs in production. The phonetic realization of a lexical item can be influenced by the presence of competitors in the lexicon, presumably due to cascading activation of lexical items in speech production (Baese-Berk and Goldrick 2009). In the case of the lexical items in our experiment, the target nouns can be understood to have an allomorph with a pitch accent and an allomorph without a pitch accent. If these allomorphs compete for selection and actuation in production, then residual activation of one allomorph (with a lexical accent) could influence the production of another (without a lexical accent).

It is not a task of the current paper to choose between these mechanisms. The point in a nutshell is that there is evidence from the speech production literature that lexical items can have multiple phonetic representations that influence speech production patterns; therefore, it is not unlikely that the canonical realization of the prosodic structure that Richards (2010) posits can be "disrupted" due to various mechanisms, which can affect actual speech production in systematic ways.[8] Notably, the mechanism responsible for perturbing canonical phonetic realizations of

(in principle) no prosodic effects at all." Richards (2010) thus does allow for the presence of a language that groups wh-elements and their licensers together, but does not overtly signal that grouping in any prosodic means. As we have seen, however, Japanese is a language that does signal wh-domains either by reduction or eradication; what we are finding is that not every token shows phonetic evidence for that grouping.

[8]This postulation implies that it is not necessarily the case that we can infer a particular prosodic structure from surface phonetic realizations alone. On the one hand, this is not a new observation: a particular syllable structure, for

prosodic structure must all be probabilistic in nature.[9]

3.3 Directions for future research

Before closing this paper, we would like to offer some discussion on possible directions for future research, opened up by the results of the current study. One clear remaining task is to apply the current computational method to even more naturalistic data (i.e. sentences that are spontaneously uttered), although we acknowledge that it is probably hard to find sentences consisting of words without many obstruents, which would perturb F0 contours. As the associate editor pointed out, words consisting of only sonorants may have atypical lexical frequencies, which can impact the intonation patterns. It would be nevertheless interesting to test how robust our method is given these additional layers of noise.

Another limitation of the current study is that our analysis is based on a controlled set of stimuli in order to address a specific set of hypotheses. As an anonymous reviewer pointed out, the sort of experimental design deployed by Ishihara (2011a) could have "encouraged [the participants] to imagine various information-structural contexts for these apparently de-contextualized phrases" and raises the question of "whether the different behavior...in reading the statements was connected with different assumptions about what (explicit or implicit) question they were responding to." A follow-up study with more varied set of stimuli, possibly controlling for factors such as information structure (see Roettger, Mahrt, and Cole 2019) may sharpen the results.

Finally, the current results open up a new research opportunity for studies of other languages, or even other dialects of Japanese (see e.g. Smith 2005 and 2011 and references cited therein for intonational patterns associated with wh-sentences in Fukuoka Japanese). Recall that Richard's (2010) theory is about the universal principle that should hold in all wh-sentences in all human languages. Given what the current research has revealed about how wh-question sentences in Tokyo Japanese are phonetically realized, we expect that the same method can and should be applied to examine how the prediction of Richard's theory holds up in other languages. Comparison of languages in which wh-movement is obligatory and those which allow wh-in-situ would be particularly informative.

---

example, may be difficult to infer from its surface phonetics, although there are proposals that syllabic organization does manifest itself in the phonetic signal (e.g. Browman and Goldstein 1988; Shaw, Gafos, et al. 2011). Phonetic evidence for foot structure in some languages (e.g. Japanese) is also notoriously hard to come by (Ota, Ladd, and Tsuchiya 2003). On the other hand, this postulation can raise an interesting challenge for studies of intonation in general, which generally assume that prosodic structure can be inferred from surface phonetic patterning (e.g. tonal distributions). Accepting this thesis therefore means that, in order to argue for a particular prosodic pattern, we need to take into consideration other possible influences, like the factors discussed in this section.

[9]The more extreme alternative interpretation of our results is that the phonetic signal reliably diagnoses prosodic phonological structure but that it is prosodic phrasing that is variable. On this view, some instances of wh-questions, i.e. those with full tonal realizations, would have the same prosodic structure as declaration sentence, a state of affairs which is not consistent with any interpretation of Richards (2010).

4 Conclusion

All in all, our results provide robust quantitative support from a reasonable sample of naturally elicited utterances that tone eradication/reduction, as assumed by Richards (2010), does indeed happen in Japanese. However, the process is variable. Therefore, it is not the case that a prosodic grouping pattern is checked for each syntactic derivation. That would prevent wh-in-situ just when the grouping between the wh-phrase and its licenser is not phonetically signalled. Instead we conclude that it is a more abstract prosodic feature of Japanese—the abstract phonological structure—which allows wh-phrases to stay in-situ.

Our results show that the same syntactic derivation can map to different surface phonetic outcomes. Specifically, the same syntactic derivation for Japanese questions is sometimes produced with tone eradication or reduction. In the growing literature on syntax-prosody interface, it is still unclear how such prosodic variation may influence syntactic derivations. Our results may suggest that the syntactic derivation can proceed without necessarily referencing the surface phonetic details of the utterance under construction. In this sense, the syntax may posit appropriate prosodic grouping, "trusting" that the derivation proceeds with "indicative phonetic outcomes," even though the phonetic implementation at times betrays this trust. To be clear, we are not presenting the current results as evidence against Richards' (2010) proposal. Language-specific syntax may indeed operate in the presence of language-specific prosody. Our conclusion regards the nature of this relation. The strongest hypothesis that follows from our results is that syntax does not vary to accommodate phonetic variation in prosody.

Declarations

Funding: NINJAL collaborative research project 'Cross-linguistic Studies of Japanese Prosody and Grammar.' Conflicts of interest/competing interests: NA. Availability of data and materials: The raw data can be made available upon request. Code availability: The Matlab scripts can be made available upon request.

References

Anttila, Arto, Matthew Adams, and Michael Speriosu (2010). "The role of prosody in the English dative alternation". In: *Language and Cognitive Processes* 25, pp. 946–981.

Baese-Berk, Melissa and Matthew Goldrick (2009). "Mechanisms of interaction in speech production." In: *Language and Cognitive Processes* 24, pp. 527–554.

Beddor, Patrice S. et al. (2018). "The time course of individuals' perception of coarticulatory information is linked to their production: Implications for sound change". In: *Language* 94.4, pp. 931–968.

Bennett, Ryan, Emily Elfner, and James McCloskey (2016). "Lightest to the right: An anomalous displacement in Irish". In: *Linguistic Inquiry* 47.2, pp. 169–234.
Braver, Aaron (2019). "Modeling incomplete neutralisation with weighted phonetic constraints". In: *Phonology* 36.1, pp. 1–36.
Breiss, Canaan and Bruce Hayes (2020). "Phonological markedness effects in sentential formation". In: *Language* 96, pp. 338–370.
Brickhouse, C.J. and Kate Lindsey (2020). *Investigating the phonetics-phonology interface with field data: Assessing phonological specification through acoustic trajectories.* Poster presented at the 96th meeting of the Linguistics Society of America.
Browman, Catherine and Louis Goldstein (1988). "Some notes on syllable structure in Articulatory Phonology". In: *Phonetica* 45, pp. 140–155.
Cho, Taehong (2016). "Prosodic boundary strengthening in the phonetics-prosody interface". In: *Language and Linguistic Compass* 10.3, pp. 120–141.
Chomsky, Noam (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.
Cohn, Abigail (2006). "Is there gradient phonology?" In: *Gradience in Grammar: Generative Perspectives*. Ed. by Gisbert Fanselow et al. Oxford: Oxford University Press, pp. 25–44.
Deguchi, Masanori and Yoshihisa Kitagawa (2002). "Prosody and wh-questions". In: *Proceedings of NELS* 32, pp. 73–92.
Hirotani, Masako (2005). "Prosody and LF Interpretation: Processing Japanese Wh-questions". Doctoral Dissertation. University of Massachusetts, Amherst.
Hombert, Jean-Marie, John Ohala, and William G. Ewan (1979). "Phonetic explanations for the development of tones". In: *Language* 55, pp. 37–58.
Igarashi, Yosuke (2015). "Intonation". In: *The Handbook of Japanese Language and Linguistics: Phonetics and Phonology*. Ed. by Haruo Kubozono. Mouton: Mouton de Gruyter, pp. 525–568.
Ishihara, Shinichiro (2000). *Scrambling and its interaction with stress and focus*. Ms. MIT.
— (2003). "Intonation and Interface Conditions". Doctoral Dissertation. MIT.
— (2011a). "Focus prosody in Tokyo Japanese wh-questions with lexical unaccented wh-phrases". In: *Proceedings of ICPhS* XVII, pp. 946–949.
— (2011b). "Japanese focus prosody revisited: Freeing focus from prosodic phrasing". In: *Lingua* 121.13, pp. 1870–1889.
— (2015). "Syntax-phonology interface". In: *The Handbook of Japanese Language and Linguistics: Phonetics and Phonology*. Ed. by Haruo Kubozono. Mouton: Mouton de Gruyter, pp. 569–618.
— (2016). "Japanese downstep revisited". In: *Natural Language and Linguistic Theory* 34, pp. 1389–1443.
Ito, Junko and Armin Mester (2012). "Recursive prosodic phrasing in Japanese". In: *Prosody Matters*. Ed. by Toni Borowsky et al. London: Equinox Publishing, pp. 280–303.
Jain, Anil K. (1989). *Fundamentals of digital image processing*. Englewood Cliffs: Prentice Hall.
Krivokapić, Jelena, Will Styler, and Benjamin Parrell (2020). "Pause postures: The relationship between articulation and cognitive processes during pauses." In: *Journal of Phonetics* 79.
Kubozono, Haruo (1988). "The Organization of Japanese Prosody". Doctoral dissertation. University of Edinburgh.
— (2007). "Focus and intonation in Japanese: Does focus trigger pitch reset?" In: *Interdisciplinary studies on information structure 9*. Ed. by Shinichiro Ishihara. Universitätsverlag Potsdam, pp. 1–27.

Ladd, D. Robert (1996). *Intonational Phonology*. Cambridge, UK: Cambridge University Press.
Lee, Sungbok, Dani Byrd, and Jelena Krivokapić (2006). "Functional data analysis of prosodic effects on articulatory timing". In: *Journal of the Acoustical Society of America* 119.3, pp. 1666–1671.
Maekawa, Kikuo (1994). "Is there 'dephrasing' of the accentual phrase in Japanese". In: *Ohio State University Working Papers in Linguistics*, pp. 146–165.
McCawley, James D. (1968). *The Phonological Component of a Grammar of Japanese*. The Hague: Mouton.
Mücke, Doris, Martine Grice, and Taehong Cho (2014). "More than a magic moment—Paving the way for dynamics of articulation and prosodic structure." In: *Journal of Phonetics* 44.1-7.
Ota, Mitsuhiko, D. Robert Ladd, and Madoka Tsuchiya (2003). "Effects of foot structure on mora duration in Japanese?" In: *Proceedings of the 15th International Conference on Phonetic Sciences*, pp. 459–462.
Pierrehumbert, Janet B. (2001). "Exemplar dynamics: Word frequency, lenition and contrast". In: *Typological studies in language, Vol. 45. Frequency and the emergence of linguistic structure*. Ed. by Joan Bybee and Paul Hopper. Amsterdam: John Benjamins, pp. 137–157.
Pierrehumbert, Janet B. and Mary Beckman (1988). *Japanese Tone Structure*. Cambridge: MIT Press.
Poser, William (1984). "The Phonetics and Phonology of Tone and Intonation in Japanese". Doctoral dissertation. MIT.
Richards, Norvin (2010). *Uttering Trees*. MIT Press.
— (2016). *Contiguity Theory*. MIT Press.
Roettger, Timo B., Tim Mahrt, and Jennifer Cole (2019). "Mapping prosody onto meaning – the case of information structure in American English". In: *Language, Cognition and Neuroscience*.
Seyfarth, Scott (2014). "Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation". In: *Cognition* 133, pp. 140–155.
Shaw, Jason, Adamantios Gafos, et al. (2011). "Dynamic invariance in the phonetic expression of syllable structure: A case study of Moroccan Arabic consonant clusters". In: *Phonology* 28.3, pp. 455–490.
Shaw, Jason and Shigeto Kawahara (2018a). "Assessing surface phonological specification through simulation and classification of phonetic trajectories". In: *Phonology* 35, pp. 481–522.
— (2018b). "The lingual gesture of devoiced [u] in Japanese". In: *Journal of Phonetics* 66, pp. 100–119.
Shih, Stephanie S and Vera Gribanova (2016). "Phonological influences in syntactic alternations". In: *The morphosyntax-phonology connection: Locality and directionality at the interfaces*. Oxford: Oxford University Press, pp. 223–254.
Shih, Stephanie S and Kie Zuraw (2017). "Phonological conditions on variable adjective-noun word order in Tagalog". In: *Phonological Analysis* 93.4, e317–e352.
Smith, Jennifer (2005). "On the *wh*-question intonational domain in Fukuoka Japanese: Some implications for the syntax-prosody interface". In: *UMOP 30: Papers on Prosody*. Ed. by Shigeto Kawahara. Amherst: GSLA, pp. 219–237.
— (2011). "[+wh] complementizers drive phonological phrasing in Fukuoka Japanese". In: *Natural Language and Linguistic Theory* 29, pp. 545–559.

Steriade, Donca (2000). “Paradigm uniformity and the phonetics-phonology boundary”. In: *Papers in Laboratory Phonology V: Acquisition and the Lexicon*. Ed. by Michael B. Broe and Janet B. Pierrehumbert. Cambridge: Cambridge University Press, pp. 313–334.

Sugahara, Mariko (2003). “Downtrends and Post-focus Intonation in Tokyo Japanese”. Doctoral Dissertation. University of Massachusetts, Amherst.

Tang, Kevin and Jason Shaw (2020). *Prosody leaks into the memories of words*. Ms. University of Florida and Yale University, arXiv:2005.14716.

Vatikiotis-Bateson, Eric, Adriano Vilela Barbosa, and Catherine T. Best (2014). “Articulatory coordination of two vocal tracts”. In: *Journal of Phonetics* 44, pp. 167–181.

Venditti, Jennifer (2005). “The ToBI model of Japanese intonation”. In: *Prosodic Typology: The Phonology of Intonation and Phrasing*. Ed. by Sun-Ah Jun. Oxford: Oxford University Press, pp. 172–200.

Venditti, Jennifer, Kikuo Maekawa, and Mary Beckman (2008). “Prominence marking in the Japanese intonational system”. In: *The Handbook of Japanese Linguistics*. Ed. by Shigeru Miyagawa and Mamoru Saito. Oxford: Oxford University Press.

Wedel, Andrew B. (2007). “Feedback and regularity in the lexicon”. In: *Phonology* 24.1, pp. 147–185.

Whalen, Douglas H and Andrea G. Levitt (1995). “The universality of intrinsic F0 of vowels”. In: *Journal of Phonetics* 23, pp. 349–366.

Yu, Alan (2007). “Understanding near mergers: The case of morphological tone in Cantonese”. In: *Phonology* 24, pp. 187–214.

Zahorian, Stephan A. and Hongbing Hu (2008). “A spectral/temporal method for robust fundamental frequency tracking”. In: *Journal of the Acoustical Society of America* 123.6, pp. 4559–4571.

Zhang, M., C. Geissler, and Jason Shaw (2019). “Gestural representations of tone in Mandarin: Evidence from timing alternations”. In: *Proceedings of the 19th International Congress of Phonetic Sciences*, pp. 1803–1807.