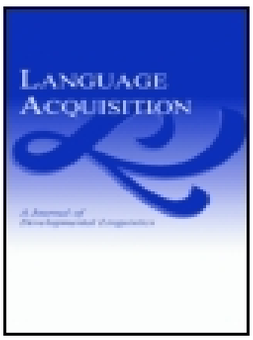
Coverbal speech gestures signal phrase boundaries: A production study of Japanese and English infant- and adult-directed speech

Irene de La Cruz-Pavía, Judit Gervain, Eric Vatikiotis-Bateson, Janet F. Werker



Language Acquisition

ISSN: 1048-9223 (Print) 1532-7817 (Online) Journal homepage: https://www.tandfonline.com/loi/hlac20

Coverbal speech gestures signal phrase boundaries: A production study of Japanese and English infant- and adult-directed speech

Irene de la Cruz-Pavía, Judit Gervain, Eric Vatikiotis-Bateson & Janet F. Werker

To cite this article: Irene de la Cruz-Pavía, Judit Gervain, Eric Vatikiotis-Bateson & Janet F. Werker (2019): Coverbal speech gestures signal phrase boundaries: A production study of Japanese and English infant- and adult-directed speech, Language Acquisition, DOI: 10.1080/10489223.2019.1659276

To link to this article: https://doi.org/10.1080/10489223.2019.1659276

Published online: 06 Sep 2019.

LANGUAGE ACQUISITION
https://doi.org/10.1080/10489223.2019.1659276

Routledge
Taylor & Francis Group

Coverbal speech gestures signal phrase boundaries: A production study of Japanese and English infant- and adult-directed speech

Irene de la Cruz-Pavía[a,b], Judit Gervain[a,b], Eric Vatikiotis-Bateson[c], and Janet F. Werker[c]

[a]Université Paris Descartes; [b]Centre National de la Recherche Scientifique; [c]University of British Columbia

ABSTRACT

The acoustic realization of phrasal prominence is proposed to correlate with the order of V(erbs) and O(bjects) in natural languages. The present production study with 15 talkers of Japanese (OV) and English (VO) investigates whether the speech signal contains coverbal visual information that covaries with auditory prosody, in Infant- and Adult-Directed Speech (IDS and ADS). Acoustic analysis revealed that phrasal prominence is carried by different acoustic cues in the two languages and speech styles, while analyses of motion showed that this acoustic prominence is not accompanied by coverbal gestures. Instead, the talkers of both languages produced eyebrow movements to mark the boundaries of target phrases within elicited utterances in combination with head nods. These results suggest that the signal might contain multimodal information to phrase boundaries, which could help listeners chunk phrases from the input.

ARTICLE HISTORY
Received 20 July 2018
Accepted 19 August 2019

1. Introduction

A wide range of visual information is available in face-to-face interactions. In speech processing, we benefit from seeing the movements of the visible articulators (lips, jaw, and tongue), as well as from coverbal gestures such as head or eyebrow movements. Determining how and to what extent this visual information accompanies speech at different linguistic levels is crucial toward understanding how humans produce and process the multimodal speech signal. The present investigation examines whether the signal contains visual information—in particular coverbal speech gestures—that covaries reliably with a specific aspect of auditory prosody at the phrase level, namely, phrasal prominence. The role of visual information in speech perception might be particularly important to listeners such as young infants, who are still building the lexicon and acquiring the regularities of the native language and hence cannot yet fully rely on this top-down linguistic knowledge. In particular, coverbal visual speech gestures might, as an aid for auditory prosody, help such listeners parse the continuous speech input and locate the boundaries of prosodic units such as phrases. Sensitivity to prosody emerges very early in development. From 6 months of age, infants use prosody to segment speech (Shukla, White & Aslin 2011), and phrase-level prosody helps prelexical infants associate new "words" with objects and constrains lexical access (Gout, Christophe & Morgan 2004; Johnson 2008). Moreover, prosodic phrases, in turn, correlate with underlying syntactic phrases (Nespor & Vogel 1986; Selkirk 1996), and prelexical infants are sensitive to this relationship (Jusczyk et al. 1992; Soderstrom et al. 2003). Thus, information that may help infants chunk prosodically relevant phrases from the speech input has the potential to help them discover syntactically relevant properties of phrases, such as word order, prior to having lexical knowledge.

CONTACT Irene de la Cruz-Pavía idelacruzpavia@gmail.com Integrative Neuroscience and Cognition Center (INCC - UMR 8002), CNRS-Université Paris Descartes, Paris, France.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hlac.

Both visual speech and coverbal facial gestures correlate with speech acoustics (Cavé et al. 1996; Munhall et al. 2004; Yehia, Kuratate & Vaitikiotis-Bateson 2002). Visual speech consists of the gestures that result from the production of speech sounds and plays an important role in speech perception from the earliest stages of language acquisition and into adulthood (see Soto-Faraco et al. 2012 for a review): It helps infants discriminate speech sounds and acquire the language's phonetic categories (Teinonen et al. 2008), allows infants and adults to discriminate languages (Soto-Faraco et al. 2007; Weikum et al. 2007, 2013), enhances speech intelligibility in adverse conditions (Sumby & Pollack 1954), etc. Coverbal gestures such as head and eyebrow motion co-occur with speech but do not directly result from the production of speech sounds (Cavé et al. 1996; Munhall et al. 2004). These gestures also influence speech perception and are the focus of the present work, given their role as markers of visual prosody (Dohen, Lœvenbruck & Hill 2006). Thus, in addition to facilitating speech intelligibility (Al Moubayed, Beskow & Björn 2010; Munhall et al. 2004), head nods and eyebrow movements enhance the perception of prosodic prominence and focus (House, Beskow & Granström 2001; Krahmer & Swerts 2007; Mixdorff, Hönemann & Fagel 2013; Prieto et al. 2015) and hinder it when visual and auditory prominence are incongruent (Swerts & Krahmer 2008). Further, the presence of head nods concurrent with phrasal auditory prominence facilitates parsing new input—i.e., an artificial language—into phrase-like units in adults (de la Cruz-Pavía et al. 2019).

A growing number of production studies aim to characterize the association between coverbal visual information and auditory prosody with both spontaneous and scripted speech. These studies evidence further the link between coverbal gestures and prosody observed in perceptual studies. Thus, talkers accompany phenomena such as prominence and focus with aligned eyebrow and head motion (French: Dohen, Lœvenbruck & Hill 2006; Dutch: Swerts & Krahmer 2010; English: Flecha-García 2010; Kim, Cvejic & Davis 2014; Catalan: Esteve-Gibert et al. 2017). Moreover, the production of these gestures can alter the realization of acoustic prominence (Krahmer & Swerts 2007). Interestingly, the use and frequency of coverbal speech gestures appears to be modulated by the strength of the prominent element: Dutch newscasters produce eyebrow raises and nods twice as frequently in pitch-accented words as compared with weakly accented words (Swerts & Krahmer 2010), whereas English talkers use both head and eyebrow movements to mark phrasal stress but only head motion to mark word stress (Scarborough et al. 2009). Phrasing also impacts the use of coverbal gestures. Catalan talkers spontaneously produce nods in focused words, and importantly, the alignment of their peaks is influenced by whether a phrase boundary precedes or follows the word (Esteve-Gibert et al. 2017). Further, Japanese talkers produce head nods in about 30%–40% of phrases containing strong boundaries, but only in 10%–15% of phrases with weak boundaries (Ishi, Ishiguro & Hagita 2014).

The aim of the current study is to examine the relationship between acoustic prominence and nonverbal visual information at the phrase level. Specifically, we aim to determine whether coverbal facial gestures—head nods and eyebrow movements—accompany a specific type of prosodic prominence, i.e., the phrasal prosodic patterns associated with word order.

Basic word order seems to be configured very early in development: Infants' first multiword utterances typically follow the word order rules of the language under acquisition (Bloom 1970), and infants at the one-word stage (16–18 months) can correctly interpret simple sentences that differ only in their word order (e.g., *Cookie Monster is tickling Big Bird* vs. *Big Bird is tickling Cookie Monster;* Hirsh-Pasek & Golinkoff 1996). Furthermore, around the same age, infants detect violations of word order in their native language (Shady 1996; Weissenborn et al. 1998) and are able to detect a change in word order in a four-word sequence already from birth (Benavides-Varela & Gervain 2017). Crucially, the signal contains specific types of statistical and prosodic information that correlate with the order of Verbs and Objects and that could potentially allow infants to build a rudimentary representation of word order in the first year of life: the frequency distribution of functors and content words (frequency-based information, Gervain et al. 2008) and the location and acoustic realization of phrasal prominence (prosodic information, Nespor et al. 2008).

Functors (determiners, adpositions, etc.: *the, behind*) are extremely frequent elements but typically acoustically nonprominent; content words (nouns, verbs, etc: *turtle, walk*) are much less frequent but phonologically more salient, e.g., they carry phrasal prominence. In languages with a V(erb)-O(bject) order (English, Spanish …) functors tend to occur phrase-initially (e.g., English: $in_{functor}\ London_{content\ word}$). In these languages, the phrasal prominence carried by the content word is realized through increased duration, resulting in an iambic or short-long pattern (*in* ***Lo:n****don*). In O(bject)-V(erb) languages (Japanese, Basque …), by contrast, functors tend to occur phrase-finally (e.g., Japanese: $Tokyo_{content\ word}\ ni_{functor}$ 'Tokyo-to'), and phrasal prominence is realized though higher pitch and/or intensity, resulting in a trochaic or high-low/loud-soft pattern (e.g., ^***To****kyo ni*) (Gervain & Werker 2013; Nespor et al. 2008).

Adults and prelexical infants are sensitive to this frequency-based and prosodic information: Both 8-month-old infants and adults track the frequency and the relative order of the elements in structurally ambiguous artificial languages and segment these unknown languages according to the order of functors and content words in their native languages (see Gervain et al. 2013; de la Cruz-Pavía et al. 2015 for adult studies, Gervain et al. 2008 for infant studies). Additionally, 7–8-month-old and adult monolinguals and bilinguals can use the acoustic realization of prominence to determine the order of the elements in these structurally ambiguous artificial languages (see de la Cruz-Pavía et al. 2019 for adult studies, Bernard & Gervain 2012; Gervain & Werker 2013 for infant studies), and 2-month-old infants and adults discriminate VO and OV languages based on the location of main prominence (Christophe et al. 2003). Importantly, infants start integrating frequency-based and prosodic information at least by 8 months of age (Bernard & Gervain 2012; Morgan & Saffran 1995), which potentially provides them with mounting information about the basic word order of the language(s) under acquisition. It has been proposed that visual speech might help infants segment words from speech by directing their attention to the relevant acoustic information available in the signal (Hollich, Newman & Jusczyk 2005). Indeed, infants can use visual speech or even a synchronous oscilloscope display to segment words from a novel stream (Hollich, Newman & Jusczyk 2005). A recent study has shown that seeing an avatar produce head nods aligned with phrasal acoustic prominence—i.e., the prosodic cue to word order—facilitated adults' segmentation of a structurally ambiguous artificial language (de la Cruz-Pavía et al. 2019). This result suggests that coverbal visual cues in the form of head nodding help adults parse new input into phrases. Visual information might similarly direct the infants' attention to the prosodic and frequency-based cues to phrase segmentation and basic word order.

Here, we investigate whether facial gestures systematically accompany the acoustic realization of phrasal prominence. Further, we examine whether cross-linguistic differences arise in the specific gestures produced by talkers of OV and VO languages. To that end, we present the results of a production study with native talkers of languages that have opposite basic word orders (English: VO, Japanese: OV) in which the measurement of facial gestures was conducted in parallel with the acoustic analysis of the utterances. Head and eyebrow motion were coded because they have been shown to enhance the perception of prosodic prominence. Japanese was chosen as it is one of the languages measured in Gervain & Werker (2013). It therefore allows us to replicate the results of their acoustic analysis and expand them to include English. Whether the predicted acoustic realization of phrasal prominence (i.e., a contrast in duration, Nespor et al. 2008) is observed in this VO language[1] has not been examined thus far and is a unique contribution of this work.

We examined the realization of phrasal prominence and of potential coverbal facial gestures accompanying it in Adult- and Infant-Directed Speech (IDS) in the two languages investigated. Previous literature has exclusively examined the realization of this particular aspect of phrasal prosody associated with word order in Adult Directed Speech (ADS, Gervain & Werker 2013; Nespor et al.

[1]Note that Japanese and English differ in a number of other suprasegmental aspects such as rhythmic class (Japanese: mora-based, English: stress-based), accent realization (Japanese: pitch-accented, English: stress-accented), or prosodic phrasing. These prosodic features are not the focus of the present study and are hence not discussed further.

2008). However, infants are often addressed in IDS; this speech style captures their attention more readily than ADS (Fernald 1985; Hayashi, Kametawa & Kimitani 2001; Pegg, Werker & McLeod 1992) and has been shown to help infants segment an unknown, artificial language (Thiessen, Hill & Saffran 2005). The literature comparing visual information in IDS and ADS is also very scarce and limited to visual speech but reveals interesting cross-linguistic differences. Vowels in IDS appear to be hyper-articulated—produced with exaggerated lip movements—in English IDS (Green et al. 2010; Shochi et al. 2009) but hypoarticulated—produced with reduced lip movements—in Japanese IDS (Shochi et al. 2009). To date, however, no one has examined coverbal speech gestures in IDS nor investigated how such gestures might correlate with acoustic cues to phrases. The present investigation seeks to fill that gap. To this end, English and Japanese participants were recorded producing the target stimuli in IDS and ADS. Importantly, we will not directly compare the productions of IDS and ADS but only determine whether the predicted patterns of phrasal prosody are present across speech styles and accompanied by coverbal gestures.

2. Methodology

2.1. *Participants*

Fifteen adults participated in this production study. Eight were native talkers of Japanese (mean age 40;06, range 34 to 47), and seven were native speakers of Canadian English (mean age = 36, range 28 to 42[2]). All participants were females and mothers of one or more infants and/or primary school-aged children. The English talkers reported no knowledge or exposure to OV languages at the time of the recording, while the Japanese talkers reported moving to North America in their 20s (three talkers) or 30s (five talkers). All participants provided informed consent and were financially compensated for their participation.

2.2. *Materials*

Stimuli consisted of target phrases containing a noun and a functor. The Japanese stimuli were the eight original phrases created by Gervain & Werker (2013)—four bi- or trisyllabic nouns followed by one of two bisyllabic functors (*made, niwa*). To have a greater number of tokens, 10—instead of eight—similar target phrases were created in English by combining one of two bisyllabic functors (*behind, beside*) with 10 bi- or trisyllabic nouns. Though English functors are typically monosyllabic, *behind* and *beside* were chosen to match in number of syllables the Japanese functors used by Gervain & Werker (2013). The target phrases were embedded in an invariant carrier sentence, as shown in (1).

(1). **Invariant carrier sentence** + *target phrase*
(a) English: **In English**, $behind_{\text{functor}}$ $cabinets_{\text{content word}}$ **is a phrase**.
(b) Japanese: **Nihon de**, $Mizuno_{\text{content word}}$ $niwa_{functor}$ **aru**.
Japan in Mizuno to exist
'In Japan, to Mizuno exists.'

These target sentences were intermixed with an equal number of filler sentences that varied in length to help avoid list intonation and facilitate the production of IDS (e.g., *Rabbits have long ears*; see Appendix A for all stimuli and fillers).

[2]One of the participants did not provide this piece of information. Mean age and age range were thus calculated on the remaining six participants.

2.3. *Procedure*

All recordings took place at Dr. Vatikiotis-Bateson's Communication Dynamics Laboratory (Department of Linguistics, University of Britsh Columbia) in Vancouver. Participants were videotaped using a Panasonic AJ-PX270 HD camcorder and a Sennheiser MKH-416 interference tube microphone. The stimuli were displayed as a PowerPoint presentation with each sentence on a separate slide on a MacBook Air placed in front of the talkers. Each sentence occurred twice during the presentation, yielding a total of 20 experimental sentences and 20 fillers in English and 16 experimental sentences and 16 fillers in Japanese. The talkers were asked to first read each sentence, then look up at the camera and say the sentence as a natural declarative sentence. They were first recorded uttering the sentences in IDS and then in ADS. To facilitate the production of IDS, a picture of a Caucasian or Asian baby was held above the camera, and participants were instructed to direct the utterances to the baby in the picture (see Fernald & Simon 1984 for similar characteristics in simulated IDS). During the ADS productions, participants were instead instructed to direct the utterances to the cameraman.

3. Analysis and results

A total of 1,072 utterances were recorded: 536 filler and 536 experimental sentences. Of the latter, 280 were utterances in English (10 sentences x 2 repetitions x 7 talkers x 2 speech styles, i.e., IDS and ADS), and 256 in Japanese (8 sentences x 2 repetitions x 8 talkers x 2 speech styles). A total of 110 utterances (20.52%) were excluded from the analysis due to low quality of the recording resulting from background noise (57 utterances, 10.63%), or disfluencies in production and/or mispronunciations (53 utterances, 9.89%). Three of the eight Japanese participants spontaneously produced the sentence-final particles *–yo/-dayo* (markers used to emphasize new information) and *–mas* (politeness marker) in most of their utterances (26 utterances in IDS, 34 in ADS), deviating from the original stimuli and potentially altering their intonational contour. Therefore, the Japanese utterances were submitted to two separate acoustic analyses, one containing all utterances (n = 426), and a second analysis retaining only the canonical utterances (n = 366), both reported in the following. Lastly, 12 productions from one of the Japanese talkers had to be discarded from the analysis of visual facial gestures (head nods and eyebrow movements) due to interference caused by the presence of another person. Thus, a total of 426 utterances were submitted to acoustic analysis (English IDS: 113, ADS: 99, Japanese IDS: 107, ADS: 107) and 414 to analysis of visual facial gestures (English IDS: 113, ADS: 99, Japanese IDS: 107, ADS: 95).

3.1. *Acoustic analysis*

The boundaries of the stressed vowels of the functor and the content word in the target phrase were marked manually (PRAAT, Boersma & Weenink 2008), and their respective duration, mean intensity, mean pitch, and pitch maximum were calculated and normalized to the respective measures of the sentence, eliminating effects of speech rate and other talker idiosyncrasies (Gervain & Werker 2013; Nespor et al. 2008).

The obtained results of the acoustic analysis are illustrated in Figure 1 and revealed differences in the realization of phrasal prominence in VO and OV languages. To establish that the target phrases contained prosodic prominence, i.e., a difference between the realization of the stressed vowel of functors and content words, the four dependent variables were analyzed using linear mixed effects models (lme4 package, R, Bates et al. 2015). All models included the fixed effects of Word (stressed vowel of functor vs. content word), Speech Style (IDS, ADS), and Language (Japanese, English)—all centered around 0—and the random factors Subject and Item. Only the results relevant to the analysis of phrasal prominence are reported, i.e., the fixed effects of Language, Word, and their potential interaction (see Table 1; the full results and details of the models are available in Appendix B). No direct comparison of the two speech styles was conducted here or in subsequent analyses. Theoretically, we have no strong predictions of

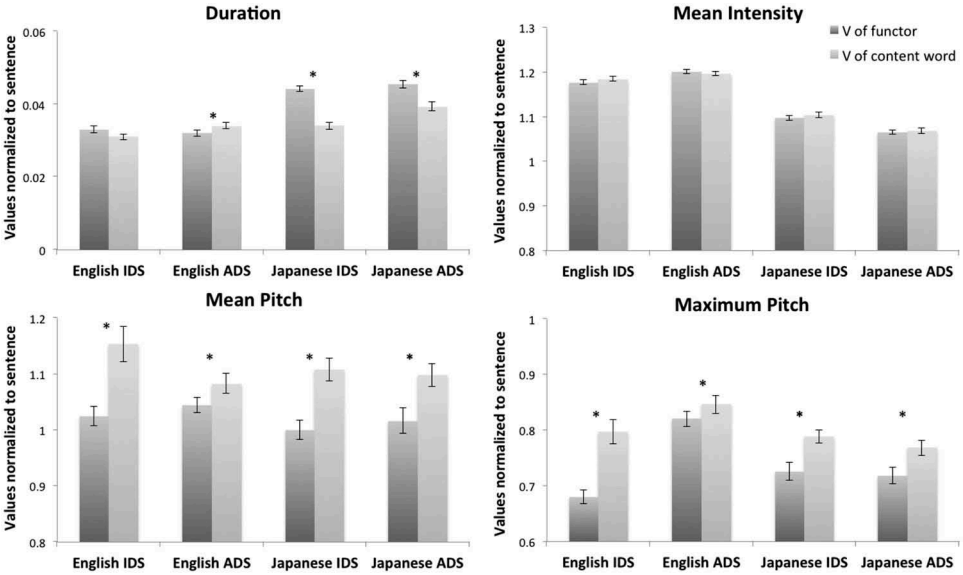


Figure 1. The four graphs show the duration, mean intensity, mean pitch, and pitch maximum of the stressed vowels of the functors (dark gray bars) and the content words (light gray bars). The *x* axes depict the language and speech style; the *y* axes depict the normalized values. Statistically significant differences are marked by asterisks. The values illustrated for Japanese correspond to the analysis of only the canonical utterances.

potential differences between IDS and ADS. Further, the different number of productions in the two styles renders pairwise comparison statistically inappropriate.[3]

A first set of models conducted on data from the English utterances and the Japanese canonical utterances revealed, as predicted, a significant fixed effect of Word for duration (p = .025), pitch maximum (p = .019), and an almost significant trend for mean pitch (p = .051). A significant fixed effect of Language was found for duration (p = .003) and for mean intensity (p < .001). The interaction between Language and Word only reached significance for duration (p = .019). A second set of models conducted over all English and Japanese utterances revealed similar fixed effects of Language (duration: p = .007, intensity: p < .001), as well as an interaction between Language and Word (duration: p = .027). Importantly, a significant fixed effect of Word was found only for duration (p = .034), in addition to a trend for pitch maximum (p = .055). To further explore the observed fixed effects of Word and Language and their interaction, the measures for the two word types were directly compared in two-tailed paired-sample *t*-tests, splitting them by language and speech style (Confidence Interval set to 98.75% to correct for multiple comparisons; see Figure 1 and Table 1).

In English ADS, the predicted contrast in duration was found (p = .048): The stressed vowel of the content word was longer than the stressed vowel of the functor. Surprisingly, this vowel also had higher mean pitch (p = .027) and pitch maximum (p = .046). In English IDS, a similar contrast in mean pitch and pitch maximum (both p < .001) was found, but not the predicted contrast in duration. Instead, a nonsignificant trend in the opposite direction was observed, i.e., longer duration in the vowel of the functor (p = .095).

The analysis of the canonical utterances in Japanese revealed the predicted pitch contrast—that is, higher pitch on the vowel of the content word—in ADS and IDS and both in mean and maximum

[3]Note, however, that potential effects of Speech Style and their interaction with Language and/or Word are reported in the full results presented in Appendix B. The results of these models show certain differences between ADS and IDS. These differences are clearly present in the acoustic analysis of the productions—particularly the canonical ones—and more limited in the analyses of coverbal gestures. These analyses reveal the need for exploring differences in speech styles in more naturalistic environments, i.e., in mothers' natural interactions with their infants.

Table 1. Acoustic analysis of English and Japanese utterances, using linear mixed models (upper part) and pair-wise *t*-tests (lower part). Here and in subsequent tables, asterisks and periods depict the following levels of significance: . = $p < .1$; * = $p < .05$; ** = $p < .01$, *** = $p < .001$.

		LINEAR MIXED MODELS	
		ENGLISH & JAPANESE CANONICAL UTTERANCES	ENGLISH & ALL JAPANESE UTTERANCES
DURATION	Language	$t(19.61) = 3.47$, $p = .003$**	$t(29.07) = 2.88$, $p = .007$**
	Word	$t(26.79) = 2.38$, $p = .025$*	$t(27.08) = 2.24$, $p = .034$*
	Language:Word	$t(26.79) = 2.50$, $p = .019$*	$t(27.08) = 2.34$, $p = .027$*
MEAN INTENSITY	Language	$t(18.70) = -6.93$, $p < .001$***	$t(18.69) = -6.89$, $p < .001$***
	Word	$t(19.66) = -0.17$, $p = .865$	$t(23.57) = 0.40$, $p = .695$
	Language:Word	$t(19.66) = 0.16$, $p = .874$	$t(23.57) = 0.56$, $p = .582$
MEAN PITCH	Language	$t(22.49) = -0.31$, $p = .759$	$t(22.07) = -0.28$, $p = .783$
	Word	$t(24.00) = -2.05$, $p = .051$.	$t(26.96) = -1.66$, $p = .109$
	Language:Word	$t(15.97) = 1.36$, $p = .193$	$t(26.96) = 0.27$, $p = .791$
PITCH MAXIMUM	Language	$t(16.47) = -0.47$, $p = .647$	$t(19.32) = -0.38$, $p = .708$
	Word	$t(21.59) = -2.54$, $p = .019$*	$t(26.20) = -2.01$, $p = .055$.
	Language:Word	$t(21.59) = 0.41$, $p = .684$	$t(26.20) = 0.80$, $p = .433$
		PAIRED SAMPLE *T*-TESTS (TWO-TAILED) CI SET TO 98.75%	
		ADS	IDS
ENGLISH	duration	$t(98) = -2.01$, $p = .048$*	$t(112) = 1.68$, $p = .095$
	mean pitch	$t(98) = -2.25$, $p = .027$*	$t(111) = -4.20$, $p < .001$***
	pitch maximum	$t(98) = -2.02$, $p = .046$*	$t(111) = -6.41$, $p < .00$ ***
	intensity	$t(98) = 1.45$, $p = .149$	$t(112) = -1.58$, $p = .117$
JAPANESE ALL UTTERANCES	duration	$t(106) = 5.22$, $p < .001$***	$t(106) = 9.63$, $p < .001$***
	mean pitch	$t(106) = -1.92$, $p = .058$	$t(106) = -2.67$, $p = .009$**
	pitch maximum	$t(106) = -1.52$, $p = .131$	$t(106) = -2.02$, $p = .045$*
	intensity	$t(106) = 1.93$, $p = .056$	$t(106) = 0.813$, $p = .418$
JAPANESE CANONICAL UTTERANCES	duration	$t(72) = 5.35$, $p < .001$***	$t(80) = 9.19$, $p < .001$***
	mean pitch	$t(72) = -2.42$, $p = .018$*	$t(80) = -4.02$, $p < .001$***
	pitch maximum	$t(72) = -2.12$, $p = .038$*	$t(80) = -3.58$, $p = .001$***
	intensity	$t(72) = -0.557$, $p = .579$	$t(80) = -0.79$, $p = .431$

pitch (ADS: mean pitch $p = .018$, pitch maximum $p = .038$; IDS: mean pitch $p < .001$, pitch maximum $p = .001$). A similar contrast was also found in the analysis of the complete set of Japanese IDS utterances (mean pitch $p = .009$, pitch maximum $p = .045$), but it remained a trend and only in mean pitch ($p = .058$) in the analysis of all ADS utterances. In addition, the vowel of the functor was significantly longer in both analyses and speech styles (all four $p < .001$), and a similar trend was observed for mean intensity in the analysis of all ADS utterances ($p = .056$).

These results thus replicate those obtained by Gervain & Werker (2013) in Japanese and extend them to IDS. Moreover, they suggest that the addition of sentence-final particles alters the intonational contour of the stimuli in Japanese ADS, while the intonational contour in IDS seems to be more stable and less vulnerable to change. In sum, the results of the acoustic analysis showed differences in the realization of phrasal prominence in VO and OV languages and suggest differences across speech styles based on the different pattern of results obtained.

3.2. *Analysis of eyebrow movements*

The eyebrow movements produced by the talkers were manually coded (Final Cut Pro 7), marking the onset of the movement, its maximum excursion (i.e., the peak or apex, see Figure 2), the end of the apex if sustained, and the movement end. A second coder marked about two-thirds of the productions. Analysis was restricted to any movements that took place—fully or partially—within the target phrases, and coders annotated whether each part of each movement occurred in the functor or the content word. If two movements took place within one single phrase, both movements were coded separately. To determine intercoder reliability, we compared the two coders' notation of

Figure 2. On the left: one of the English talkers in neutral position. On the right: maximum excursion, i.e., apex, of the eyebrows, produced in an IDS utterance.

every individual movement part. As failure to mark a given movement part (e.g., a movement start) is likely to impact marking of subsequent parts (e.g., a movement apex), the different movement parts were not analyzed separately. Disagreements were solved by discussion of the two coders or the intervention of a third coder if needed. We obtained a substantial level of intercoder reliability (Cohen's $\kappa = .645$).

As shown in Table 2, the Japanese and English talkers produced eyebrow movements with a similar frequency in ADS. In IDS, the number of eyebrow movements in the target phrase increased only in English. Analysis of the frequency and distribution of the eyebrow movements revealed that, contrary to our hypothesis, eyebrow movements did not accompany the element within the target phrase receiving phrasal prominence. Instead, onsets and offsets of eyebrow movements tended to occur at the beginning and end of the target phrases respectively. To determine potential differences across languages, four generalized linear mixed effects models were fitted (lme4, R) with probit link and using the numerical optimization algorithm BOBYQA (Powell 2009). The models included the binomial dependent variable Movement (presence vs. absence), the independent variables Speech Style (IDS, ADS), Language (English, Japanese), and Word (functor, content word), centered around 0 and Subject and Item as random factors. Two separate models examined the distribution of the onsets and apices. In both, only the interaction between Language and Word was significant (onsets: $z = -3.74$, $p < .001$; apices: $z = -2.92$, $p = .004$; see Appendix B for the full details and results). A similar trend was found in the

Table 2. Number and percentage of eyebrow movements per language and speech style and distribution of their onsets, apices, apex ends, and movement ends within the target phrase.

		JAPANESE				ENGLISH				
		IDS		ADS		IDS		ADS		ALL
Eyebrow motion in target phrase	Motion	43	39%	40	39%	83	61%	42	39%	
	No motion	68	61%	63	61%	53	39%	65	61%	
	All	$n = 111$		$n = 103$		$n = 136$		$n = 107$		$n = 457$
Movement onsets	Functor	8	24%	7	33%	31	78%	22	71%	
	Content Word	26	76%	14	67%	9	22%	9	29%	
	All	$n = 34$		$n = 21$		$n = 40$		$n = 31$		$n = 126$
Movement apices	Functor	12	29%	6	23%	37	71%	20	63%	
	Content Word	29	71%	20	77%	15	29%	12	38%	
	All	$n = 41$		$n = 26$		$n = 52$		$n = 32$		$n = 151$
Apex ends	Functor	5	42%	11	58%	15	33%	7	35%	
	Content Word	7	58%	8	42%	30	67%	13	65%	
	All	$n = 12$		$n = 19$		$n = 45$		$n = 20$		$n = 96$
Movement ends	Functor	9	50%	11	69%	23	43%	8	31%	
	Content Word	9	50%	5	31%	30	57%	18	69%	
	All	$n = 18$		$n = 16$		$n = 53$		$n = 26$		$n = 113$

analysis of the end of movements ($z = 1.70$, $p = .090$), whereas no significant fixed effect or interaction was found in the analysis of the ends of sustained apices.

To further examine these results, two-tailed binomial tests of proportions were conducted. Within the target phrases, movement onsets occurred significantly more often in the functor than in the content word in English IDS ($p = .001$) and ADS ($p = .029$), but more often in the content word than in the functor in Japanese IDS ($p = .003$). Similarly, movement apices occurred significantly more often in the functor in English IDS ($p = .003$) and in the content word in Japanese IDS ($p = .012$) and ADS ($p = .009$). By contrast, apex ends were more frequent in the content word in English IDS ($p = .036$), and a similar trend was found in the movement end in English ADS ($p = .076$).

In sum, the onset and apex of movements tended to occur in the functors in English and in the content words in Japanese, whereas apex and movement ends occurred more often in the content word in English. Note that in the English stimuli, functors precede content words within the target phrases, whereas the opposite order characterizes the Japanese stimuli. These results suggest that movement starts and apices tended to occur in the first element of the phrase in both languages. Thus, the data from all four sets of productions were collapsed and recoded as first versus second element of the phrase. New binomial tests of proportions confirmed the significantly higher proportion of movement starts and apices (both $p < .001$) in the first element of the target phrase, and apex ends ($p = .032$) and end of movements ($p = .038$) in the second element. In sum, onsets and offsets of eyebrow movements tended to occur at the beginning and end of the target phrases respectively.

3.3. *Optical flow analysis of head motion*

To examine head motion, Optical Flow analysis of the videos containing the talkers' productions was conducted using FlowAnalyzer, a program that can be applied to videos to quantify movement (FA, Barbosa, Yehia & Vatikiotis-Bateson 2008). Within a previously specified region(s) of interest (e.g., the talkers' head), the FA's algorithm compares pixel intensities in consecutive frames of a video and calculates the magnitude and direction of motion from one frame to the next of each pixel in the image (Barbosa, Yehia & Vatikiotis-Bateson 2008:1), yielding five output vectors: general magnitude of motion (Mag), horizontal direction (X) and horizontal magnitude of motion (XMag), vertical direction (Y) and vertical magnitude of motion (YMag) (see Figure 3). This technique has been shown to reliably measure motion—for instance, in the analysis of infants' response to language stimuli (Fais et al. 2012).

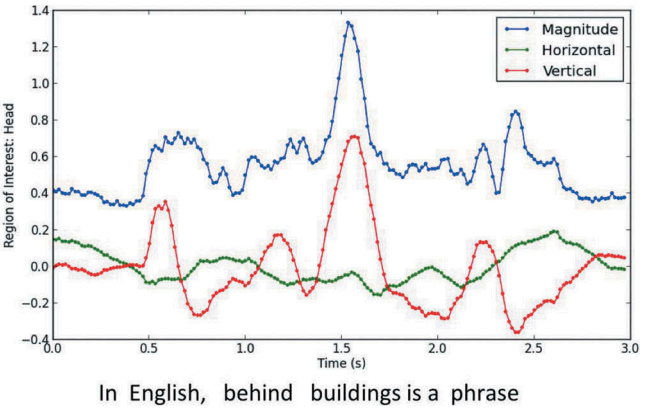


Figure 3. Sample measures showing time series analysis of motion in the area of interest—the head—using FlowAnalyzer. The figure depicts the talker's motion during an English IDS utterance. The red line depicts vertical motion, the green line horizontal motion, and the blue line magnitude (i.e., Euclidean distance). The peak observed in the red and blue lines results from a big head nod produced in the content word (*buildings*).

After submitting the talkers' videos to FA, the mean of each of the five output vectors was calculated within the time windows corresponding to the functor and the content word of the target phrases. To examine potential differences in motion, the five dependent variables were analyzed using linear mixed effects models (lme4, R). All models included the fixed effects of Word, Speech Style, and Language, centered around 0, and Subject and Item as random factors. Again, only the fixed effects of Word and Language and their potential interaction are discussed (see Appendix B for the full details and results). No significant fixed effect of Language was found in the models, but a fixed effect of Word was observed for general magnitude of movement, $t(17.09) = 3.28$, $p = .004$, and the vectors of horizontal direction and horizontal magnitude of movement, X: $t(16.80) = 2.84$, $p = .012$; XMag: $t(14.71) = 2.46$, $p = .027$. An interaction between Word and Language was found for horizontal magnitude of motion, XMag: $t(14.71) = 2.25$, $p = .040$, as well as for vertical direction and vertical magnitude of motion, Y: $t(16.71) = 2.65$, $p = .017$; YMag: $t(18.26) = 3.02$, $p = .007$.

To further explore these results, two-tailed paired-sample *t*-tests were conducted (CI set to 99% to correct for multiple comparisons; see Appendix B for the full results). Analysis of the English IDS productions revealed significantly greater movement in the content words than in the functors, limited to vertical direction ($p = .013$) and vertical magnitude ($p < .001$), which presumably results from the motion derived from head nods. Analysis of English ADS revealed a similar trend for vertical magnitude of motion ($p = .072$) and greater movement in the functors for general magnitude ($p = .001$) and horizontal direction ($p = .022$). Analysis of the Japanese productions revealed, in turn, significantly greater movement in the functors in all five vectors both in IDS and ADS (IDS: all $p \leq 0.014$; ADS: all $p \leq 0.002$).

A concurrent analysis was conducted that determined that the greater vertical motion found in the content words in English and the functors in Japanese resulted from the presence of head nods. The talkers' productions were thus visually inspected, and the presence of head nods in the functor and/or the content word was manually coded (see Table 3). An independent coder annotated just over 40% of the productions (intercoder reliability: Cohen's $\kappa = .661$). Analysis was restricted to any movements that took place within the target phrases. If a nod had its apex at the boundary between the functor and content word, and the head motion derived from the stroke and retraction phases spanned both words, the phrase was marked as having movement in both target words.

To determine if the frequency of occurrence and distribution of the head nods varied cross-linguistically, a generalized linear mixed effects model was fitted (lme4, R), with probit link using BOBYQA (Powell 2009). The model included the binomial dependent variable Movement (presence vs. absence of nod), the predictor variables Speech Style, Language, and Word—centered around 0—and Subject and Item as random factors. Only the interaction of Language and Word was significant ($z = 4.39$, $p < .001$; see Appendix B for the full details and results). To further explore this interaction, binomial tests of proportions were carried out, which revealed a significantly greater proportion of nods in the functor as compared with the content word in Japanese in both speech styles (IDS: $p = .007$; ADS: $p < .001$) but in the content word in English in both speech styles (both $p \leq 0.001$). Due to the opposite word orders of the two

Table 3. Number and percentage of head nods and head motion per language and speech style and their distribution within the target phrase.

	JAPANESE		ENGLISH	
	IDS	ADS	IDS	ADS
No nods in target phrase	33	44	62	45
	30.84%	46.32%	54.87%	45.45%
Nod in the functor	19	19	6	3
	17.76%	20.00%	5.31%	3.03%
Nod in the content word	5	2	25	20
	4.67%	2.11%	22.12%	20.20%
Nod in both target words	50	30	20	31
	46.73%	31.58%	17.70%	31.31%
Total number of productions	107	95	113	99

languages, these results suggest that the talkers produced more head nods in the second element of the target phrase. To confirm this, all productions were collapsed, and the independent variable was recoded as first versus second element in the phrase. A new binomial test confirmed the higher proportion of head nods occurring in the second element of the target phrase, as compared with the first ($p < .001$).

In sum, contrary to prediction, talkers did not produce head nods concurrent with the element receiving phrasal prominence. Instead, head nods occurred in the second and final element of the target phrase.

3.4. *Filler analysis*

To determine whether the pattern observed in the analysis of the experimental sentences was replicated in more variable contexts, we examined the occurrence of coverbal gestures at phrase boundaries in a subset of the filler sentences. Thus, we located phrases as similar as possible to the relevant syntactic structure, i.e., a bisyllabic functor combined with a content word. In English, we analyzed the phrase [*These cookies*], and in Japanese [*kudamono kara*]—'fruit from "from/of fruit."' We manually coded the occurrence of eyebrow movements and head nods in a total of a total of 58 productions,[4] following the same parameters as in the analysis of the experimental sentences. The full results are reported in Appendix C.

Across languages and speech styles, 36% to 44% of the filler phrases contained eyebrow movements. Movement starts and apices occurred overwhelmingly in the first element of the phrase—the functor in English and the content word in Japanese—whereas movement ends occurred more frequently in the second element of the phrase in English IDS and ADS and Japanese IDS. Head nods in Japanese occurred with a distribution similar to the one observed in the experimental phrases, as single nods occurred always in the second element of the phrase (i.e., the functor). However, occurrence of head nods in English was lower in the fillers than in the experimental phrases. Also, the distribution of single nods differed somewhat, with no single nods in IDS, and more nods occurring in the functor ($n = 2$) than the content word ($n = 1$) in ADS.

In sum, the pattern found in the filler phrases tallies greatly with the one obtained in the experimental sentences.

4. Discussion

In a production study with 15 talkers of Japanese (OV) and English (VO), we examined whether coverbal facial gestures accompanied the prosodic patterns of phrasal prominence associated with the order of verbs and objects in natural languages. We hypothesized that if facial gestures accompany the acoustic realization of phrasal prominence, cross-linguistic differences might arise in the specific gestures produced by talkers of OV and VO languages. The results of analysis of the talkers' head and eyebrow motion did not support this hypothesis. However, they revealed differences in the use of these two types of facial gestures.

Rather than as a visual marker of phrasal prominence, eyebrow movements were used to signal the boundaries of the target phrases. Thus, eyebrow raises and peaks were more likely to occur in the first element of the target phrase and ends of apices and movements in the second element. The fact that the target phrases were inserted in an invariant carrier sentence suggests that eyebrow movements were produced in both languages to signal the—informationally—most important part of the utterance, as previously observed by Ambrazaitis, Svensson Lundmark & House (2015), rather than a single prosodically prominent element (i.e., the content word).

The present materials had unusual semantics as a result of the presence of a constant carrier phrase. The pragmatic and informational contents may consequently not have been signaled in

[4]An utterance x 2 repetitions x 2 speech styles x 15 speakers = 60. Two of the productions in Japanese had to be discarded due to interference caused by the presence of another person in the frame.

typical ways. An analysis of spontaneous speech would allow us to determine whether or to what extent pragmatic and prosodic prominence coincide in naturally produced utterances. Importantly, the convergent pattern observed in the analysis of the fillers—in which target phrases were not embedded in an invariable carrier phrase—shows that our talkers produced eyebrow movements at the boundaries of phrases in variable contexts. This replication thus rules out the possibility that participants produced the observed coverbal gestures solely as an artifact of the experimental materials.

Interestingly, both the English and Japanese talkers produced a greater proportion of head nods in the second element of the phrase: the functor in Japanese and the content word in English. There are minimally two interpretations for this distribution. The simplest interpretation is that the English and Japanese talkers might produce nods to mark the ends of the target phrases which, when combined with eyebrow movements, would systematically signal both phrase boundaries. An alternative interpretation of the results is that nods might fulfill different functional roles in English and Japanese. Note that this second alternative is speculative, due to the structure of the present stimuli. English talkers might use them as a visual correlate to phrasal prominence, hence producing a greater proportion of nods in the prosodically prominent content word. The few available perceptual studies suggest that head nods are more informative than eyebrow movements in signaling prominence (Swedish: House, Beskow & Granström 2001; Catalan: Prieto et al. 2015), but further work is needed to see if this advantage is cross-linguistic or language-specific. Japanese talkers might instead use nods combined with eyebrow movements to signal the boundaries of phrases. Indeed, Ishi, Ishiguro & Hagita (2014) observed that Japanese talkers often produce nods at phrase boundaries and more frequently so in phrases with strong boundaries (30%–40%). In the present study, a greater proportion of ends of apices and movements in the second element of the phrase was indeed observed in the English productions and in the combined analysis of English and Japanese, but not in the separate analysis of Japanese productions. These results thus provide some support for Ishi, Ishiguro & Hagita's interpretation. The pattern found in the filler analysis could be interpreted as further evidence supporting this hypothesis. Thus, while distribution of nods in the Japanese fillers tallied with the one observed in the analysis of the experimental phrases, its frequency of occurrence and distribution differed in English. Analysis of eyebrow and head motion in additional languages or using different linguistic material (one in which prominence and phrasal boundaries do not coincide) is crucial to decide between these interpretations.

These results suggest that eyebrow movements—potentially in combination with head nods—might be used as a cue to prosodic boundaries across languages. In the present study, English participants were monolingual, whereas Japanese participants were recorded in Canada and also spoke English. Therefore, we cannot rule out a potential influence of the Japanese participants' L2, English, in their patterns of gestures. To the extent of our knowledge, no previous studies have compared the production of facial gestures in monolinguals and bilinguals of two spoken languages. Whether these gestural patterns tend to converge in the bilinguals' two languages remains an open question.

In line with Hollich et al.'s (2005) proposal that visual speech (i.e., articulatory gestures) might assist infants in word segmentation by directing their attention to the relevant acoustic information present in the signal, the current results suggest that coverbal facial gestures might fulfill a similar function at the phrase level. However, there is an important distinction between visual speech and coverbal gestures. Visual speech results necessarily from the production of speech sounds and is hence available in all face-to-face interactions. Meanwhile, the occurrence of coverbal gestures is less consistent, as evidenced by the fact that only one- to two-thirds of the target phrases in the present study contained eyebrow and/or head movements. Consequently, we consider that these coverbal gestures are not a direct signal of a given phonological unit such as phrases but rather a heuristic mechanism employed by talkers to mark phrase boundaries, which could in turn help perceivers parse the input.

Phrase boundaries are typically marked by pauses, final lengthening, and changes in fundamental frequency (F0). By 6 to 9 months of age, infants can use this information to segment speech. However, this ability appears to still be in a developing stage and crucially depends on the quantity and strength of the prosodic information (Soderstrom et al. 2003). Therefore, we speculate that signaling the boundaries of the informationally most relevant phrase(s) by means of visual information could facilitate infants' acquisition of the acoustic cues to phrase boundaries. Similarly, signaling phrase boundaries with the aid of facial gestures could allow the listener to attend to the information (prosodic, statistical, visual, etc.) available within the phrase. Phrasal prominence—the prosodic feature correlated with basic word order—could then be perceptually boosted by the combination of its acoustic realization (changes in duration and/or pitch) and the potential presence of head nods in certain languages (e.g., English) and speech styles. Indeed, as shown by House, Beskow & Granström (2001) and Prieto et al. (2015), head movements appear to have greater perceptual value in signaling prominence than eyebrow movements, and the occurrence of head nods concurrent with auditory prosody helps adults segment new input into phrases (de la Cruz-Pavía et al. 2019). Last, segmenting phrases from the input would allow the listener to attend to the elements that occur at their edges. This would in turn assist in the distributional analysis of the frequency of occurrence and relative order of functors and content words in the phrases, that is, the frequency-based information correlated with word order.

The present investigation examined the acoustic realization of phrasal prominence and the use of coverbal gestures accompanying it in Infant Directed Speech. Importantly, infants are often exposed to IDS and prefer it to ADS (Fernald 1985; Pegg, Werker & McLeod 1992), even if spoken in a language unknown to them (Werker, Pegg & McLeod 1994). No direct comparisons were conducted between the productions in ADS and IDS, due to the lack of theoretically motivated predictions and the fact that they differed in number. However, the different patterns of results observed suggest potential qualitative differences across these two styles. Firstly, eyebrow movements occurred more frequently in IDS than ADS in English. Further, the binomial tests of proportions showed that in Japanese IDS, onsets of movements occurred more often in the first than in the second element of the phrase. This pattern was not seen in Japanese ADS. Likewise, in English IDS apices occurred more often in the first than in the second element of the phrase, but again, this pattern was not seen in English ADS. Furthermore, the greater vertical motion in the content word as compared with the functor found in the Optical Flow analysis of the English talkers only reached significance in IDS, which suggests that the nods were larger in IDS than ADS. In sum, these patterns suggest that the talkers in both languages seem to have provided more frequent or more pronounced visual information in IDS than in ADS, though this interpretation is currently speculative.

While the current study investigated coverbal facial gestures, it is of interest to compare these results to previously reported findings with audiovisual articulatory speech. Here, it has been reported that visual speech (i.e., the visual information of the articulators) is exaggerated in English IDS (Green et al. 2010; Shochi et al. 2009) but reduced in Japanese IDS (Shochi et al. 2009). The difference between the earlier reported hypoarticulated visual-articulatory speech in Japanese and the more frequent presence of coverbal gestures (i.e., eyebrow movements) obtained in the present study suggests that these two aspects of the visual signal might fulfill different functions, and it begs further exploration.

The results of the acoustic analysis also suggest differences across styles. The pitch contrast observed in both speech styles in the analysis of the Japanese canonical productions was also found in the analysis of all Japanese utterances but remained only a trend (and only in mean pitch) in ADS. This result suggests that adding these very common sentence-final particles (*-mas, -yo*) altered the intonational contour of the target phrases. The realization of phrasal prominence in Japanese appears thus to be more fragile in ADS or rather more robust in the speech directed to listeners who are in the process of acquiring the language.

As in Gervain & Werker (2013), the stressed vowel of the functor had longer duration than the stressed vowel of the content word in Japanese, contrary to previous predictions (Nespor et al. 2008). The origin of this contrast remains to be determined, but we speculate that it might at least partially result from processes of deaccentuation. Specifically, two of the four content words in the target

phrases were frequently produced without their characteristic pitch accent (i.e., *hasu, Mizuno*), which was produced instead in the functors in some of the utterances.[5] Pitch accent has traditionally been associated exclusively with changes in F0 in the literature. However, a series of recent studies have shown that accented syllables can be longer than unaccented syllables (Kozasa 2004). Thus, this shift in pitch accent between the content words and functors might partially explain the contrast in duration observed.

The design of the stimuli might also have contributed to the observed duration contrast. The target phrases consisted of four content words and two functors combined exhaustively, resulting in eight different phrases that were then produced twice. As the talkers realized that all phrases contained one of only two possible functors, they might have used lengthening to contrast between the two possible combinations of a single content word (e.g., *hasu* ***made*** vs. *hasu* ***niwa***). To test this prediction, the talkers' productions were split into two blocks containing the first and second half of the utterances and examined to see whether the functors produced in the second half contained longer stressed vowels (normalized to the duration of the sentence) as compared with the first half, under the assumption that the hypothesized contrastive intonation builds up over time. The results of a linear mixed effects model (description and results in Appendix B) revealed no effect of block on the duration of the functors. The talkers thus did not increase the duration of their functors as the recording session progressed.

In the present study, we analyzed whether the pattern of phrasal prominence associated with VO languages (i.e., a contrast in duration) is observed in a previously unexamined language, namely, English. As predicted by Nespor et al. (2008), a contrast in duration between the functor and the content word was found in the analysis of the ADS utterances, in addition to an unexpected contrast in pitch: The stressed vowel of the content word was longer and had higher mean pitch and pitch maximum than the functor's stressed vowel. Interestingly, only the contrast in pitch was found in the IDS utterances. The presence of multiple markers of prominence, as found in ADS, is not unprecedented in the literature. Nespor et al. (2008) found longer duration in the content word, in addition to higher mean and maximum pitch, and greater intensity in Turkish, an OV language for which a contrast in pitch and/or intensity is predicted. Similarly, Molnar, Carreiras & Gervain (2016) found both higher intensity and longer duration in the content word in Basque, an OV language. Importantly, these combined markers occur, as predicted, in the content word and are therefore informative as to basic word order: Main prominence falls on the rightmost word of the phonological phrase in VO languages but on the leftmost word in OV languages (Nespor & Vogel 1986).

The observed pitch contrast might alternatively result from carryover effects from the IDS task, given that the sentences were always recorded first in IDS. To help rule out this possibility, an online naturalness test was conducted (Qualtrics, Provo, UT) with 31 participants, all of them parents of children under 24 months of age (27 had children under 12 months of age, 29 females). The participants listened to blocks of phrases excised from the talkers' recordings and judged how likely it was that the phrases were spoken to a baby (vs. to an adult). Each block contained six phrases from a single speaker and style separated by 500 ms pauses. Analysis of the responses (7-point likert scale: 1 = *very unlikely to be IDS*; 4 = *undecided*; 7 = *very likely to be IDS*) confirmed that all blocks containing ADS phrases were rated under 4 (mean rating 2.45, *SD* 0.57), and all blocks containing IDS phrases were rated around or above 4 (mean rating 5.13, *SD* 0.62).[6] Naïve listeners could thus appropriately classify the talkers' IDS and ADS utterances, diminishing the possibility of potential carryover effects from the IDS task, which in turn suggests that pitch might be a secondary marker of phrasal prominence in English ADS in addition to duration.

It could be argued that the differences across speech styles in phrasal prominence found in English might result from the fact that the phrases contained only two phonetically very similar functors (*behind, beside*). The talkers might have emphasized the production of the stressed syllable

[5]The authors wish to thank Dr. Mitsuhiko Ota (University of Edinburgh) for his invaluable input on this topic.
[6]The IDS blocks of two of the talkers were rated as 3.87 and 3.94 respectively.

of the functors—the portion of these two words that is dissimilar—to set them apart. This might have been particularly accentuated in IDS, as the infant is presumably still acquiring these prepositions. Lengthening the functors' stressed syllable might in turn have washed away a potential durational contrast in the content word. To test this hypothesis, the utterances were again split into two blocks (first vs. second half), and potential differences were analyzed in the normalized duration of the stressed vowel of the functors. No effect of block was found in a linear mixed effects model (see Appendix B for the full description and results), which suggests that the talkers did not increase the functors' duration as a means to contrast between the two words. It seems more likely that phrasal prominence in IDS is characterized by a pitch contrast and not by a contrast in duration. Pitch indeed plays a crucial role in several aspects of English phonology, and the presence of exaggerated pitch excursions is the most characteristic trait of English IDS (Fernald et al. 1989).[7] Crucially, these pitch excursions appear to drive the infants' preference for this speech style (Fernald & Kuhl 1987). Therefore, the differing patterns observed in English IDS and ADS highlight the need for contrasting these two speech styles in other (VO) languages.

The present analysis of phrasal prominence in IDS has implications for the ongoing discussion on the interplay between universal biases and language experience in prosodic grouping. The preference for a trochaic (strong-weak) grouping of sequences containing changes in pitch or intensity has been proposed to be a universal perceptual principle shared with other species and not specific to language, whereas the iambic (weak-strong) grouping preference of sequences containing changes in duration would instead be modulated by language experience (Bion, Benavides-Varela & Nespor 2011; de la Mora, Nespor & Toro 2013). Though the picture is far from clear, previous literature suggests that a general auditory mechanism may be in place from birth but simultaneously influenced by the phrasal and lexical prosodic properties of the language(s) from the earliest stages of language acquisition. Thus, newborns exposed to French, a language whose phonology makes extensive use of durational contrasts, show a bias for an iambic grouping of tones contrasted in duration (Abboub, Nazzi & Gervain 2016). Importantly, this bias is not found in English-learning 5.5- and 6.5-month-old infants—youngest ages examined to date—but seems to emerge between 7 and 9 months of age (Hay & Saffran 2012; Yoshida et al. 2010). The delay in the development of the iambic grouping bias in English might thus result from more limited exposure than previously thought to durational contrasts, particularly in the input received by the infants, that is, IDS.

5. Conclusions

The present study is the first to examine the available visual information to phrasal prominence. The analysis of the English and Japanese productions revealed that talkers did not produce coverbal gestures systematically accompanying acoustic phrasal prominence. Analyses revealed instead the presence, cross-linguistically, of reliable coverbal visual information signaling the boundaries of phrases. The Japanese and English talkers produced eyebrow raises and peaks systematically at the beginning of phrases, whereas the ends of apices and movements occurred instead in the final element of the phrase (i.e., the second element), though less frequently so. This second element was additionally characterized by the presence of head nods, both in the Japanese and English productions. As predicted, the content words contained in the target phrases carried phrasal prominence in Japanese and English (previously unexamined language), i.e., they were acoustically more salient than the functors.

The qualitative differences observed between IDS and ADS suggest that the speech directed to infants is characterized by the presence of more reliable and pronounced visual information both in English and Japanese as compared with ADS, as well as a more stable auditory prosody in Japanese. Due to the lack of direct comparison across speech styles, this interpretation needs to be taken with

[7]Fernald et al. (1989) showed that mothers who are speakers of American English have the most extreme prosodic contours as compared with mothers who are speakers of French, Italian, German, Japanese, and British English.

caution and points to the need for future research directly comparing these two speech styles. Further, the unexpected pitch contrast observed in English IDS (in contrast to the predicted duration contrast in English ADS), highlights the need to examine phrasal prominence across speech styles in other languages.

In conclusion, the results of this research suggest the presence of multimodal information in the talkers' productions—in ADS and IDS—that could help infants locate the boundaries of phrases and detect the prominent element within the phrase. Chunking the input into phrases could in turn help infants discover the basic word order of the language or languages under acquisition very early in development. However, it remains to be determined if this available visual information is also observed in the caretakers' natural interactions and indeed used by infants to segment speech.

Acknowledgments

We dedicate this paper to our friend and colleague Eric Vatikiotis-Bateson. You are sorely missed. We wish to thank Gorka Elordieta and Mitsuhiko Ota for their help with the acoustic analysis, Matteo Lisi for his help with the Linear Mix Models analyses, and Michael McAuliffe for his invaluable help with the recordings and OF analysis.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Agence Nationale de la Recherche [SpeechCode—ANR-15-CE37-0009-01]; FP7 People: Marie-Curie Actions [IOF 624972]; H2020 European Research Council [773202 ERC-2017-COG "BabyRhythm"]; Natural Sciences and Engineering Research Council of Canada [RGPIN-2015-03967]; French Investissements d'Avenir—Labex EFL [ANR-10-LABX-0083]; Social Sciences and Humanities Research Council of Canada [435-2014-0917].

References

Abboub, Nawal, Thierry Nazzi, and Judit Gervain. 2016. "Prosodic Grouping at Birth." *Brain & Language* 162:46–59. doi:10.1016/j.bandl.2016.08.002.

Al Moubayed, Samer, Jonas Beskow, and Björn Granström. 2010. "Auditory Visual Prominence: from Intelligibility to Behavior." *Journal of Multimodal User Interfaces* 3 (4):299–309. doi:10.1007/s12193-010-0054-0.

Ambrazaitis, Gilbert, Malin Svensson Lundmark, and David House. 2015. "Multimodal Levels of Prominence: A Preliminary Analysis of Head and Eyebrow Movements in Swedish News Broadcasts." In *Fonetik 2015, Working Papers in General Linguistics and Phonetics*, edited by Malin Svensson Lundmark, Gilbert Ambrazaitis and Jost van de Weijer, vol. 55, 11–16. Lunden: Lund University Publications.

Barbosa, Adriano V., Hani C. Yehia, and Eric Vatikiotis-Bateson. 2008. "Linguistically Valid Movement Behavior Measured Non-invasively." In *Proceedings of the International Conference on Auditory-Visual Speech Processing*, edited by Roland Göcke, Patrick Lucey and Simon Lucey, 173–77. Moreton Island, Australia: Causal Productions.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." *Journal of Statitiscal Software* 67 (1):1–48.

Benavides-Varela, Silvia, and Judit Gervain. 2017. "Learning Word Order at Birth: A NIRS Study." *Developmental Cognitive Neuroscience* 25:198–208. doi:10.1016/j.dcn.2017.03.003.

Bernard, Carline, and Judit Gervain. 2012. "Prosodic Cues to Word Order: What Level of Representation?" *Frontiers in Psychology* 3:451. doi:10.3389/fpsyg.2012.00451.

Bion, Ricardo A. H., Silvia Benavides-Varela, and Marina Nespor. 2011. "Acoustic Markers of Prominence Influence Infants' and Adults' Segmentation of Speech Sequences." *Language and Speech* 54 (1):123–40. doi:10.1177/0023830910388018.

Bloom, Lois. 1970. *Language Development: Form and Function in Emerging Grammars*. Cambridge, MA: MIT Press.

Boersma, Paul, and David Weenink. 2008, July 18. "Praat: Doing Phonetics by Computer." http://www.praat.org/.

Cavé, Christian, Isabelle Guaïtella, Roxane Bertrand, Serge Santi, Françoise Harlay, and Robert Espesser. 1996. "About the Relationship between Eyebrow Movements and F0 Variations." In *Proceedings of the 4th International*

Conference on Spoken Language Processing, edited by Timothy Bunnell and William Idsardi, Vol. 4, 2175–78. Delaware: Citation Delaware.

Christophe, Anne, Marina Nespor, Maria Teresa Guasti, and Brit van Ooyen. 2003. "Prosodic Structure and Syntactic Acquisition: the Case of the Head-direction Parameter." *Developmental Science* 6 (2):211–20. doi:10.1111/desc.2003.6.issue-2.

de la Cruz-Pavía, Irene, Gorka Elordieta, Nuria Sebastián-Gallés, and Itziar Laka. 2015. "On the Role of Frequency-based Cues in the Segmentation Strategies of Adult OVVO Bilinguals." *International Journal of Bilingual Education and Bilingualism* 18 (2):225–41. doi:10.1080/13670050.2014.904839.

de la Cruz-Pavía, Irene, Janet F. Werker, Eric Vatikiotis-Bateson, and Judit Gervain. 2019, April. "Finding Phrases: the Interplay of Word Frequency, Phrasal Prosody and Co-Speech Visual Information in Chunking Speech by Monolingual and Bilingual Adults." *Language and Speech*. doi:10.1177/0023830919842353.

de la Mora, Daniela M., Marina Nespor, and Juan M. Toro. 2013. "Do Humans and Nonhuman Animals Share the Grouping Principles of the Iambic - Trochaic Law?" *Attention, Perception, & Psychophysics* 75 (1):92–100. doi:10.3758/s13414-012-0371-3.

Dohen, Marion, Hélène Lœvenbruck, and Harold Hill. 2006. "Visual Correlates of Prosodic Contrastive Focus in French: Description and Inter-speaker Variability." In *Proceedings of the Third International Conference on Speech Prosody*, edited by Rüdiger Hoffmann and Hansjörg Midorff, 221–24. Dresden: TUDpress.

Esteve-Gibert, Núria, Joan Borràs-Comes, Eli Asor, Marc Swerts, and Pilar Prieto. 2017. "The Timing of Head Movements: the Role of Prosodic Heads and Edges." *The Journal of the Acoustical Society of America* 141 (6):4727–39. doi:10.1121/1.4986649.

Fais, Laurel, Janet F. Werker, Bronwyn Cass, Julia Leibowich, Adriano V. Barbosa, and Eric Vatikiotis-Bateson. 2012. "Here's Looking at You, Baby: What Gaze and Movement Reveal about Minimal Pair Word-object Association at 14 Months." *Laboratory Phonology* 3 (1):91–124. doi:10.1515/lp-2012-0007.

Fernald, Anne. 1985. "Four-month-old Infants Prefer to Listen to Motherese." *Infant Behavior and Development* 8:181–95. doi:10.1016/S0163-6383(85)80005-9.

Fernald, Anne, and Patricia Kuhl. 1987. "Acoustic Determinants of Infant Preference for Motherese Speech." *Infant Behavior and Development* 10:279–93. doi:10.1016/0163-6383(87)90017-8.

Fernald, Anne, and Thomas Simon. 1984. "Expanded Intonation Contours in Mothers' Speech to Newborns." *Developmental Psychology* 20 (1):104–13. doi:10.1037/0012-1649.20.1.104.

Fernald, Anne, Traute Taeschner, Judy Dunn, Mechthild Papousek, Bénédicte de Boysson-Bardies, and Ikuko Fukui. 1989. "A Cross-language Study of Prosodic Modifications in Mothers' and Fathers' Speech to Preverbal Infants." *Journal of Child Language* 16 (3):477–501.

Flecha-García, María L. 2010. "Eyebrow Raises in Dialogue and Their Relation to Discourse Structure, Utterance Function and Pitch Accents in English." *Speech Communication* 52 (6):542–54. doi:10.1016/j.specom.2009.12.003.

Gervain, Judit, and Janet F. Werker. 2013. "Prosody Cues Word Order in 7-month-old Bilingual Infants." *Nature Communications* 4:1490. doi:10.1038/ncomms2430.

Gervain, Judit, Marina Nespor, Reiko Mazuka, Ryota Horie, and Jacques Mehler. 2008. "Bootstrapping Word Order in Prelexical Infants: A Japanese-Italian Cross-linguistic Study." *Cognitive Psychology* 57:56–74. doi:10.1016/j.cogpsych.2007.12.001.

Gervain, Judit, Núria Sebastián-Gálles, Begoña Díaz, Itziar Laka, Reiko Mazuka, Naoto Yamane, Marina Nespor, and Jacques Mehler. 2013. "Word Frequency Cues Word Order in Adults: Crosslinguistic Evidence." *Frontiers in Psychology* 4:689. doi:10.3389/fpsyg.2013.00186.

Gout, Ariel, Anne Christophe, and James L. Morgan. 2004. "Phonological Phrase Boundaries Constrain Lexical Access II. Infant Data." *Journal of Memory and Language* 51 (4):548–67. doi:10.1016/j.jml.2004.07.002.

Green, Jordan R., Ignatius S. B. Nip, Erin M. Wilson, Antje S. Mefferd, and Yana Yunusova. 2010. "Lip Movement Exaggerations during Infant-directed Speech." *Journal of Speech, Language and Hearing Research* 53 (6):1529–42. doi:10.1044/1092-4388(2010/09-0005).

Hay, Jessica F., and Jenny R. Saffran. 2012. "Rhythmic Grouping Biases Constrain Infant Statistical Learning." *Infancy* 17 (6):610–41. doi:10.1111/j.1532-7078.2011.00110.x.

Hayashi, Akiko, Yuji Kametawa, and Shigeru Kimitani. 2001. "Developmental Change in Auditory Preferences for Speech Stimuli in Japanese Infants." *Journal of Speech, Language and Hearing Research* 44:1189–200. doi:10.1044/1092-4388(2001/092).

Hirsh-Pasek, Kathy, and Roberta M. Golinkoff. 1996. "The Preferential Looking Paradigm Reveals Emerging Language Comprehension." In *Methods for Assessing Children's Syntax*, edited by Dana McDaniel, Cecile McKee and Helen S. Cairns, 105–24. Cambridge, MA: MIT Press.

Hollich, George, Rochelle S. Newman, and Peter W. Jusczyk. 2005. "Infants' Use of Synchronized Visual Information to Separate Streams of Speech." *Child Development* 76:598–613. doi:10.1111/j.1467-8624.2005.00866.x.

House, David, Jonas Beskow, and Björn Granström. 2001. "Timing and Interaction of Visual Cues for Prominence in Audiovisual Speech Perception." In *Proceedings of EuroSpeech 2001*, edited by Paul Dalsgaard, Børge Lindberg and Henrik Benner, 387–90. Aalborg, Denmark: Aalborg Universitetsforlag.

Ishi, Carlos T., Hiroshi Ishiguro, and Norihiro Hagita. 2014. "Analysis of Relationship between Head Motion Events and Speech in Dialogue Conversations." *Speech Communication* 57:233–43. doi:10.1016/j.specom.2013.06.008.

Johnson, Elizabeth K. 2008. "Infants Use Prosodically Conditioned Acoustic-phonetic Cues to Extract Words from Speech." *Journal of the Acoustical Society of America* 123 (6):EL144–148. doi:10.1121/1.2875420.

Jusczyk, Peter, Deborah Kemler Nelson, Kathy Hirsh-Pasek, Lori J. Kennedy, Amanda Woodward, and Julie Piwoz. 1992. "Perception of Acoustic Correlates of Major Phrasal Units by Young Infants." *Cognitive Psychology* 24:252–93.

Kim, Jeesun, Erin Cvejic, and Chris Davis. 2014. "Tracking Eyebrows and Head Gestures Associated with Spoken Prosody." *Speech Communication* 57:317–30. doi:10.1016/j.specom.2013.06.003.

Kozasa, Tomoko. 2004. The Interaction of Duration and Pitch in Japanese Long Vowels. *Proceedings of the 30th Annual Meeting of the Berkeley Linguistics Society [BLS30.]*, 211–22. http://linguistics.berkeley.edu/bls/previous_proceedings/bls30.pdf.

Krahmer, Emiel, and Marc Swerts. 2007. "The Effects of Visual Beats on Prosodic Prominence: Acoustic Analyses, Auditory Perception and Visual Perception." *Journal of Memory and Language* 57:396–414. doi:10.1016/j.jml.2007.06.005.

Mixdorff, Hansjörg, Angelika Hönemann, and Sascha Fagel. 2013. "Integration of Acoustic and Visual Cues in Prominence Perception." In *Proceedings of the Auditory Visual Speech Processing Conference (AVSP) 2013*, edited by Slim Ouni, Frédéric Berthommier and Alexandra Jesse, 111–16. Annecy, France: Inria.

Molnar, Monika, Manuel Carreiras, and Judit Gervain. 2016. "Language Dominance Shapes Non-linguistic Rhythmic Grouping in Bilinguals." *Cognition* 152:150–59. doi:10.1016/j.cognition.2016.03.023.

Morgan, James L., and Jenny R. Saffran. 1995. "Emerging Integration of Sequential and Suprasegmental Information in Preverbal Speech Segmentation." *Child Development* 66 (4):911–36. doi:10.2307/1131789.

Munhall, Kevin G., Jeffery A. Jones, Daniel E. Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson. 2004. "Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception." *Psychological Science* 15 (2):133–37. doi:10.1111/j.0963-7214.2004.01502010.x.

Nespor, Marina, and Irene Vogel. 1986. *Prosodic Phonology*. Dordrecht, The Netherlands: Foris.

Nespor, Marina, Mohinish Shukla, Ruben van de Vijver, Cinzia Avesani, Hanna Schraudolf, and Caterina Donati. 2008. "Different Phrasal Prominence Realizations in VO and OV Languages." *Lingue e Linguaggio* 7 (2):1–29.

Pegg, Judith E., Janet F. Werker, and Peter J. McLeod. 1992. "Preference for Infant-Directed over Adult-Directed Speech: Evidence from 7-week-old Infants." *Infant Behavior and Development* 15:325–45. doi:10.1016/0163-6383(92)80003-D.

Powell, Michael J. D. 2009. "The BOBYQA Algorithm for Bound Constrained Optimization without Derivatives." Technical report, Department of Applied Mathematics and Theoretical Physics, Cambridge University. DAMTP 2009/NA06.

Prieto, Pilar, Cecilia Puglesi, Joan Borràs-Comes, Ernesto Arroyo, and Josep Blat. 2015. "Exploring the Contribution of Prosody and Gesture to the Perception of Focus Using an Animated Agent." *Journal of Phonetics* 49:41–54. doi:10.1016/j.wocn.2014.10.005.

Scarborough, Rebecca, Patricia Keating, Sven L. Mattys, Taehong Cho, and Abeer Alwan. 2009. "Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English." *Language and Speech* 51 (2/3):135–175. doi: 10.1177/0023830909103165.

Selkirk, Elisabeth. 1996. "The Prosodic Structure of Function Words." In *Signal to Syntax: Bootstrapping from Speech to Syntax in Early Acquisition*, edited by James L. Morgan and Katherine Demuth, 187–213. Hillsdale, USA: Erlbaum.

Shady, Michele. E. 1996. "Infants' Sensitivity to Function Morphemes." Doctoral dissertation, The State University of New York at Buffalo.

Shochi, Takaaki, Kaoru Sekiyama, Nicole Lees, Mark Boyce, Roland Göcke, and Denis Burnham. 2009. "Auditory-visual Infant Directed Speech in Japanese and English." In *International Conference on Auditory-Visual Speech Processing (AVSP) 2009*, edited by Barry-John Theobald and Richard Harvey, 107–12. Norwich, UK: University of East Anglia.

Shukla, Mohinish, Katherine S. White, and Richard N. Aslin. 2011. "Prosody Guides the Rapid Mapping of Auditory Word Forms onto Visual Objects in 6-mo-old Infants." *Proceedings of the National Academy of Sciences* 108 (15):6038–43. doi:10.1073/pnas.1017617108.

Soderstrom, Melanie, Amanda Seidl, Deborah G. Kemler Nelson, and Peter W. Jusczyk. 2003. "The Prosodic Bootstrapping of Phrases: Evidence from Prelinguistic Infants." *Journal of Memory and Language* 49:249–67. doi:10.1016/S0749-596X(03)00024-X.

Soto-Faraco, Salvador, Jordi Navarra, Whitney M. Weikum, Athena Vouloumanos, Núria Sebastián-Gallés, and Janet F. Werker. 2007. "Discriminating Languages by Speech-reading." *Perception & Psychophysics* 69 (2):218–31. doi:10.3758/BF03193744.

Soto-Faraco, Salvador, Marco Calabresi, Jordi Navarra, Janet F. Werker, and David Lewkowicz. 2012. "The Development of Audiovisual Speech Perception." In *Multisensory Development*, edited by Andrew J. Bremmer, David J. Lewkowicz and Charles Spence, 207–28. Oxford, UK: Oxford University Press.

Sumby, W. H., and J. Pollack 1954. "Visual Contribution to Speech Intelligibility in Noise." Journal of the Acoustical Society of America 26:212–215. doi: 10.1121/1.1907309.

Swerts, Marc, and Emiel Krahmer. 2008. "Facial Expressions and Prosodic Prominence: Comparing Modalities and Facial Areas." *Journal of Phonetics* 36 (2):219–38. doi:10.1016/j.wocn.2007.05.001.

Swerts, Marc, and Emiel Krahmer. 2010. "Visual Prosody of Newsreaders: Effects of Information Structure, Emotional Content and Intended Audience on Facial Expressions." *Journal of Phonetics* 38:197–206. doi:10.1016/j.wocn.2009.10.002.

Teinonen, Tuomas, Richard N. Aslin, Paavo Alku, and Gergely Csibra. 2008. "Visual Speech Contributes to Phonetic Learning in 6-month-old Infants." *Cognition* 108 (3):850–55. doi:10.1016/j.cognition.2008.05.009.

Thiessen, Erik D., Emily A. Hill, and Jenny R. Saffran. 2005. "Infant-Directed Speech Facilitates Word Segmentation." *Infancy* 7 (1):53–71. doi:10.1207/s15327078in0701_5.

Weikum, Whitney M., Athena Vouloumanos, Jordi Navarra, Salvador Soto-Faraco, Núria Sebastián-Gallés, and Janet F. Werker. 2007. "Visual Language Discrimination in Infancy." *Science* 316:1159. doi:10.1126/science.1142109.

Weikum, Whitney M., Athena Vouloumanos, Jordi Navarra, Salvador Soto-Faraco, Núria Sebastián-Gallés, and Janet F. Werker. 2013. "Age-related Sensitive Periods Influence Visual Language Discrimination in Adults." *Frontiers in Systems Neuroscience* 7:86. doi:10.3389/fnsys.2013.00039.

Weissenborn, Jürgen, Barbara Höhle, Dorothea Kiefer, and Damir Cavar. 1998. "Children's Sensitivity to Word-order Violations in German: Evidence for Very Early Parameter-setting." In *Proceedings of the 22nd Annual Boston University Conference on Language Development*, edited by Annabel Greenhill, Mary Hughes, Heather Littlefield and Hugh Walsh, Vol. II, 756–67. Somerville: Cascadilla Press.

Werker, Janet F., Judith E. Pegg, and Peter J. McLeod. 1994. "A Cross-language Investigation of Infant Preference for Infant-directed Communication." *Infant Behavior and Development* 17 (3):323–33. doi:10.1016/0163-6383(94)90012-4.

Yehia, Hani C., Takaaki Kuratate, and Eric Vaitikiotis-Bateson. 2002. "Linking Facial Animation, Head Motion and Speech Acoustics." *Journal of Phonetics* 30:555–68. doi:10.1006/jpho.2002.0165.

Yoshida, Katherine A., John R. Iversen, Aniruddh D. Patel, Reiko Mazuka, Hiromi Nito, Judit Gervain, and Janet F. Werker. 2010. "The Development of Perceptual Grouping Biases in Infancy: A Japanese-English Cross-linguistic Study." *Cognition* 115:356–61. doi:10.1016/j.cognition.2010.01.005.

Appendix A Stimuli and filler sentences

1. Japanese stimuli

4 nouns: 2 bisyllabic + 2 trisyllabic; 2 bisyllabic functors

Content words:

jisho = dictionary hasu = magnet

kabuto = cap, helmet Mizuno = Mizuno (name of a city)

Functors:

niwa = to made = till

Carrier sentence:

Nihon de [target phrase] aru.
Japan in exist.
'In Japan, [target phrase] exists.'

Target phrases:

jisho niwa jisho made

hasu niwa hasu made

Mizuno niwa Mizuno made

kabuti niwa kabuto made

2. English stimuli

10 nouns: 6 bisyllabic + 4 trisyllabic; 2 bisyllabic functors

Content words:

curtains furniture columns

buildings cabinets mountains

barriers restaurants

houses hospitals

Functors:

behind beside

Carrier sentence:

In English, [target sentence] is a phrase.

Target phrases:

behind furniture	beside columns
behind mountains	beside houses
behind barriers	beside cabinets
behind buildings	beside hospitals
behind curtains	beside restaurants

3. Japanese fillers

Saru wa banana ga daisukidesu. 'Monkeys love bananas.'

Usagi no mimi wa nagaidesu. 'Rabbit ears are long.'

Kore wa aisu kurimu desu. 'This is ice cream.'

Kame no namae wa Fu chan. 'The turtle's name is Fu.'

Kame wa yukkuri ugokimasu. 'Turtles move slowly.'

Susan wa sannin shimai de, choujo desu. 'Susan has two younger sisters.'

Kono kukkī wa ama sugidesu. 'This cookie is too sweet.'

Jamu wa kudamono kara tsukurimasu. 'Jam is made of fruits.'

4. English fillers

My turtle's name is Victoria.

Kittens are small and soft.

Susan has two little sisters.

Monkeys love bananas.

These cookies are very sweet.

This is chocolate ice cream.

Rabbits have long ears.

Jam is made from fruit.

Turtles move very slowly.

Puppies like to chew on things.

Appendix B Full results of the linear mixed models

1. Acoustic analysis

Results of the Linear Mixed Models used to analyze acoustic phrasal prominence. The lme4 package uses Satterthwaite approximations to degrees of freedom.

Table B1. Analysis of the English and Japanese canonical utterances. Style and Word were allowed to vary randomly by Subject in all models and by Item in the models analyzing duration, mean intensity, and maximum pitch. Only Word was allowed to vary randomly in the analysis of mean pitch. A model of greater complexity resulted in a convergence failure of the optimization algorithm.

		ENGLISH & JAPANESE CANONICAL UTTERANCES
DURATION	(Intercept)	β = 0.038, SE = 0.001, t(19.61) = 25.67, p < .001***
	Style	β = 0.003, SE = 0.001, t(13.40) = 2.49, p = .027*
	Language	β = 0.010, SE = 0.003, t(19.61) = 3.47, p = .003**
	Word	β = 0.004, SE = 0.002, t(26.79) = 2.38, p = .025*
	Style:Language	β = 0.002, SE = 0.002, t(13.40) = 1.01, p = .333
	Style:Word	β = − 0.004, SE = 0.001, t(11.85) = − 3.04, p = .010*
	Language:Word	β = 0.008, SE = 0.003, t(26.79) = 2.50, p = .019*
	Style:Language:Word	β = − 0.001, SE = 0.003, t(11.85) = − 0.47, p = .646
MEAN INTENSITY	(Intercept)	β = 1.131, SE = 0.008, t(18.70) = 146.71, p < .001***
	Style	β = − 0.013, SE = 0.004, t(15.61) = − 2.99, p = .009**
	Language	β = − 0.107, SE = 0.015, t(18.70) = − 6.93, p < .001***
	Word	β = − 0.001, SE = 0.008, t(19.66) = − 0.17, p = .865
	Style:Language	β = − 0.043, SE = 0.009, t(15.61) = − 4.83, p < .001***
	Style:Word	β = 0.006, SE = 0.009, t(15.00) = 0.66, p = .517
	Language:Word	β = 0.002, SE = 0.015, t(19.66) = 0.16, p = .874
	Style:Language:Word	β = − 0.017, SE = 0.017, t(15.00) = − 0.95, p = .357
MEAN PITCH	(Intercept)	β = 1.063, SE = 0.019, t(22.49) = 56.98, p < .001***
	Style	β = − 0.004, SE = 0.022, t(15.21) = − 0.17, p = .866
	Language	β = − 0.012, SE = 0.037, t(22.49) = − 0.31, p = .759
	Word	β = − 0.082, SE = 0.040, t(24.00) = − 2.05, p = .051 .
	Style:Language	β = 0.018, SE = 0.044, t(15.21) = 0.40, p = .692
	Style:Word	β = 0.055, SE = 0.040, t(15.97) = 1.36, p = .193
	Language:Word	β = − 0.008, SE = 0.080, t(24.00) = − 0.11, p = .917
	Style:Language:Word	β = − 0.053, SE = 0.081, t(15.97) = − 0.65, p = .523
PITCH MAXIMUM	(Intercept)	β = 0.772, SE = 0.018, t(16.47) = 42.34, p = < 0.001***
	Style	β = 0.034, SE = 0.021, t(15.94) = 1.75, p = .099 .
	Language	β = − 0.017, SE = 0.036, t(16.47) = − 0.47, p = .647
	Word	β = − 0.061, SE = 0.024, t(21.59) = − 2.54, p = .019*
	Style:Language	β = − 0.118, SE = 0.043, t(15.94) = − 2.74, p = .015*
	Style:Word	β = 0.050, SE = 0.034, t(14.09) = 1.50, p = .156
	Language:Word	β = 0.020, SE = 0.048, t(21.59) = 0.41, p = .684
	Style:Language:Word	β = − 0.067, SE = 0.067, t(14.0i) = − 0.99, p = .338

Table B2. Analysis of English and all Japanese utterances. Style and Word were allowed to vary randomly by Subject and Item in all models.

		ENGLISH & ALL JAPANESE UTTERANCES
DURATION	(Intercept)	β = 0.036, SE = 0.001, t(29.07) = 29.63, p < .001***
	Style	β = 0.001, SE = 0.001, t(14.36) = 0.60, p = .555
	Language	β = 0.007, SE = 0.002, t(29.07) = 2.88, p = .007**
	Word	β = 0.004, SE = 0.002, t(27.08) = 2.24, p = .034*
	Style:Language	β = − 0.001, SE = 0.003, t(14.36) = − 0.40, p = .692
	Style:Word	β = − 0.004, SE = 0.001, t(14.04) = − 2.80, p = .014*
	Language:Word	β = 0.007, SE = 0.003, t(27.08) = 2.34, p = .027*
	Style:Language:Word	β = − 0.001, SE = 0.003, t(14.04) = − 0.32, p = .754
MEAN INTENSITY	(Intercept)	β = 1.133, SE = 0.008, t(18.69) = 147.08, p < .001***
	Style	β = − 0.001, SE = 0.010, t(13.20) = − 0.10, p = .919
	Language	β = − 0.106, SE = 0.015, t(18.69) = − 6.89, p < .001***
	Word	β = 0.004, SE = 0.009, t(23.57) = 0.40, p = .695
	Style:Language	β = − 0.023, SE = 0.020, t(13.20) = − 1.16, p = .265
	Style:Word	β = 0.011, SE = 0.009, t(16.10) = 1.29, p = .217
	Language:Word	β = 0.010, SE = 0.018, t(23.57) = 0.56, p = .582
	Style:Language:Word	β = − 0.009, SE = 0.018, t(16.10) = − 0.48, p = .640
MEAN PITCH	(Intercept)	β = 1.063, SE = 0.018, t(22.07) = 58.51, p < .001***
	Style	β = 0.002, SE = 0.024, t(21.22) = 0.07, p = .946
	Language	β = − 0.010, SE = 0.036, t(22.07) = − 0.28, p = .783
	Word	β = − 0.067, SE = 0.040, t(26.96) = − 1.66, p = .109
	Style:Language	β = 0.026, SE = 0.048, t(21.22) = 0.53, p = .600
	Style:Word	β = 0.043, SE = 0.037, t(15.08) = 1.17, p = .262
	Language:Word	β = 0.022, SE = 0.081, t(26.96) = 0.27, p = .791
	Style:Language:Word	β = − 0.074, SE = 0.074, t(15.08) = − 1.00, p = .333
PITCH MAXIMUM	(Intercept)	β = 0.774, SE = 0.017, t(19.32) = 46.01, p < .001***
	Style	β = 0.045, SE = 0.019, t(16.90) = 2.39, p = .029*
	Language	β = − 0.013, SE = 0.034, t(19.32) = − 0.38, p = .708
	Word	β = − 0.051, SE = 0.025, t(26.20) = − 2.01, p = .055 .
	Style:Language	β = − 0.103, SE = 0.038, t(16.90) = − 2.73, p = .014*
	Style:Word	β = 0.042, SE = 0.032, t(16.61) = 1.32, p = .206
	Language:Word	β = 0.040, SE = 0.051, t(26.20) = 0.80, p = .433
	Style:Language:Word	β = − 0.084, SE = 0.063, t(16.61) = − 1.33, p = .203

2. Analysis of eyebrow movements

Table B3. Results of the Generalized Linear Mixed Models used to analyze the frequency and distribution of eyebrow movements. Style and Word were allowed to vary randomly by Subject in all models and by Item in the models analyzing movement starts and apex ends. Only Style was allowed to vary randomly in the analysis of apices, and only Word in the analysis of movement ends.

		ANALYSIS OF EYEBROW MOTION
START OF MOVEMENT	(Intercept)	$\beta = -1.434$, $SE = 0.181$, $z = -7.96$, $p < .001$***
	Style	$\beta = 0.004$, $SE = 0.191$, $z = 0.02$, $p = .983$
	Word	$\beta = -0.032$, $SE = 0.219$, $z = -0.15$, $p = .884$
	Language	$\beta = 0.010$, $SE = 0.350$, $z = 0.03$, $p = .978$
	Style:Word	$\beta = -0.390$, $SE = 0.400$, $z = -0.98$, $p = .330$
	Style:Language	$\beta = 0.313$, $SE = 0.298$, $z = 1.05$, $p = .294$
	Word:Language	$\beta = -1.443$, $SE = 0.385$, $z = -3.74$, $p < .001$***
	Style:Word:Language	$\beta = -0.751$, $SE = 0.665$, $z = -1.13$, $p = .259$
APEX OF MOVEMENT	(Intercept)	$\beta = -1.397$, $SE = 0.206$, $z = -6.78$, $p < .001$***
	Style	$\beta = 0.282$, $SE = 0.195$, $z = 1.45$, $p = .148$
	Word	$\beta = -0.409$, $SE = 0.304$, $z = -1.34$, $p = .179$
	Language	$\beta = -0.179$, $SE = 0.397$, $z = -0.45$, $p = .653$
	Style:Word	$\beta = 0.064$, $SE = 0.384$, $z = 0.17$, $p = .868$
	Style:Language	$\beta = 0.170$, $SE = 0.292$, $z = 0.58$, $p = .560$
	Word:Language	$\beta = -1.619$, $SE = 0.555$, $z = -2.92$, $p = .004$ **
	Style:Word:Language	$\beta = -0.356$, $SE = 0.591$, $z = -0.60$, $p = .546$
APEX END	(Intercept)	$\beta = -2.136$, $SE = 0.370$, $z = -5.78$, $p < .001$***
	Style	$\beta = -0.286$, $SE = 0.501$, $z = -0.57$, $p = .568$
	Word	$\beta = 0.533$, $SE = 0.522$, $z = 1.02$, $p = .306$
	Language	$\beta = 0.023$, $SE = 0.607$, $z = 0.04$, $p = .970$
	Style:Word	$\beta = 1.335$, $SE = 1.119$, $z = 1.19$, $p = .233$
	Style:Language	$\beta = -0.630$, $SE = 0.652$, $z = -0.97$, $p = .333$
	Word:Language	$\beta = 0.399$, $SE = 0.635$, $z = 0.63$, $p = .529$
	Style:Word:Language	$\beta = -0.824$, $SE = 1.636$, $z = -0.50$, $p = .614$
END OF MOVEMENT	(Intercept)	$\beta = -1.478$, $SE = 0.199$, $z = -7.44$, $p < .001$***
	Style	$\beta = 0.317$, $SE = 0.195$, $z = 1.63$, $p = .103$
	Word	$\beta = -0.197$, $SE = 0.220$, $z = -0.89$, $p = .371$
	Language	$\beta = -0.366$, $SE = 0.371$, $z = -0.98$, $p = .325$
	Style:Word	$\beta = 0.320$, $SE = 0.449$, $z = 0.71$, $p = .476$
	Style:Language	$\beta = -0.204$, $SE = 0.287$, $z = -0.71$, $p = .478$
	Word:Language	$\beta = 0.552$, $SE = 0.325$, $z = 1.70$, $p = .090$.
	Style:Word:Language	$\beta = -0.716$, $SE = 0.695$, $z = -1.03$, $p = .303$

3. Analysis of head nods

Table B4. Results of the Linear Mixed Models used in the Optical Flow analysis of head motion. The lme4 package uses Satterthwaite approximations to degrees of freedom. Style and Word were allowed to vary randomly by Subject in all models and by Item in all models except the model that analyzed magnitude of horizontal motion (Mean X Magnitude), where only Word was included.

		OPTICAL FLOW ANALYSIS
MEAN MAGNITUDE	(Intercept)	β = 0.232, *SE* = 0.010, *t*(18.94) = 22.46, *p* < .001***
	Style	β = − 0.015, *SE* = 0.008, *t*(15.77) = − 1.84, *p* = .085 .
	Language	β = − 0.023, *SE* = 0.021, *t*(18.94) = − 1.10, *p* = .287
	Word	β = 0.024, *SE* = 0.007, *t*(17.09) = 3.28, *p* = .004**
	Style:Language	β = − 0.032, *SE* = 0.016, *t*(15.77) = − 1.95, *p* = .070 .
	Style:Word	β = 0.001, *SE* = 0.011, *t*(30.51) = 0.10, *p* = .924
	Language:Word	β = 0.014, *SE* = 0.015, *t*(17.09) = 0.95, *p* = .356
	Style:Language:Word	β = − 0.022, *SE* = 0.021, *t*(30.51) = − 1.07, *p* = .294
MEAN X	(Intercept)	β = 0.237, *SE* = 0.011, *t*(17.86) = 22.20, *p* < .001***
	Style	β = − 0.012, *SE* = 0.009, *t*(15.39) = − 1.35, *p* = .196
	Language	β = − 0.027, *SE* = 0.021, *t*(17.86) = − 1.28, *p* = .216
	Word	β = 0.023, *SE* = 0.008, *t*(16.80) = 2.84, *p* = .012*
	Style:Language	β = − 0.038, *SE* = 0.018, *t*(15.39) = − 2.07, *p* = .055 .
	Style:Word	β = − 0.001, *SE* = 0.012, *t*(26.66) = − 0.09, *p* = .929
	Language:Word	β = 0.023, *SE* = 0.016, *t*(16.80) = 1.42, *p* = .174
	Style:Language:Word	β = − 0.022, *SE* = 0.024, *t*(26.66) = − 0.94, *p* = .353
MEAN X MAGNITUDE	(Intercept)	β = 0.242, *SE* = 0.011, *t*(17.49) = 21.77, *p* < .001***
	Style	β = − 0.013, *SE* = 0.009, *t*(14.38) = − 1.42, *p* = .177
	Language	β = − 0.027, *SE* = 0.022, *t*(17.49) = − 1.23, *p* = .234
	Word	β = 0.020, *SE* = 0.008, *t*(14.71) = 2.46, *p* = .027*
	Style:Language	β = − 0.039, *SE* = 0.019, *t*(14.38) = − 2.10, *p* = .054 .
	Style:Word	β = − 0.001, *SE* = 0.012, *t*(25.09) = − 0.10, *p* = .919
	Language:Word	β = 0.036, *SE* = 0.016, *t*(14.71) = 2.25, *p* = .040*
	Style:Language:Word	β = − 0.023, *SE* = 0.025, *t*(25.09) = − 0.93, *p* = .360
MEAN Y	(Intercept)	β = 0.247, *SE* = 0.012, *t*(18.67) = 21.26, *p* < .001***
	Style	β = − 0.015, *SE* = 0.011, *t*(16.26) = − 1.35, *p* = .195
	Language	β = − 0.020, *SE* = 0.023, *t*(18.67) = − 0.86, *p* = .401
	Word	β = 0.011, *SE* = 0.009, *t*(16.71) = 1.35, *p* = .194
	Style:Language	β = − 0.039, *SE* = 0.022, *t*(16.26) = − 1.82, *p* = .088 .
	Style:Word	β = − 0.001, *SE* = 0.014, *t*(21.23) = − 0.07, *p* = .944
	Language:Word	β = 0.046, *SE* = 0.017, *t*(16.71) = 2.65, *p* = .017*
	Style:Language:Word	β = − 0.035, *SE* = 0.027, *t*(21.23) = − 1.29, *p* = .212
MEAN Y MAGNITUDE	(Intercept)	β = 0.251, *SE* = 0.012, *t*(19.42) = 20.69, *p* < .001***
	Style	β = − 0.019, *SE* = 0.012, *t*(17.28) = − 1.64, *p* = .120
	Language	β = − 0.008, *SE* = 0.024, *t*(19.42) = − 0.34, *p* = .734
	Word	β = 0.003, *SE* = 0.009, *t*(18.26) = 0.33, *p* = .746
	Style:Language	β = − 0.032, *SE* = 0.023, *t*(17.28) = − 1.38, *p* = .184
	Style:Word	β = 0.003, *SE* = 0.014, *t*(21.23) = 0.22, *p* = .830
	Language:Word	β = 0.054, *SE* = 0.018, *t*(18.26) = 3.02, *p* = .007**
	Style:Language:Word	β = − 0.045, *SE* = 0.027, *t*(21.23) = − 1.65, *p* = .11

Table B5. Results of the two-tailed pair-sampled *t*-tests used in the Optical Flow analysis of head motion, comparing the mean of the five output vectors within the time windows corresponding to the functor and the content word of the target phrases. The Confidence Interval was set to 99% to correct for multiple comparisons.

		OPTICAL FLOW ANALYSIS
ENGLISH IDS	Mean Magnitude	$t(112) = 1.365$, $p = .175$
	Mean X	$t(112) = 0.658$, $p = .512$
	Mean X Magnitude	$t(112) = -0.697$, $p = .487$
	Mean Y	$t(112) = -2.518$, $p = .013$*
	Mean Y Magnitude	$t(112) = -4.358$, $p < .001$***
ENGLISH ADS	Mean Magnitude	$t(98) = 3.375$, $p = .001$***
	Mean X	$t(98) = 2.328$, $p = .022$*
	Mean X Magnitude	$t(98) = 0.795$, $p = .428$
	Mean Y	$t(98) = -0.616$, $p = .539$
	Mean Y Magnitude	$t(98) = -1.821$, $p = .072$.
JAPANESE IDS	Mean Magnitude	$t(106) = 3.391$, $p = .001$***
	Mean X	$t(106) = 3.206$, $p = .002$**
	Mean X Magnitude	$t(106) = 3.087$, $p = .003$**
	Mean Y	$t(106) = 2.871$, $p = .005$**
	Mean Y Magnitude	$t(106) = 2.500$, $p = .014$*
JAPANESE ADS	Mean Magnitude	$t(94) = 4.075$, $p < .001$***
	Mean X	$t(94) = 4.337$, $p < .001$***
	Mean X Magnitude	$t(94) = 4.768$, $p < .001$***
	Mean Y	$t(94) = 3.723$, $p < .001$***
	Mean Y Magnitude	$t(94) = 3.196$, $p = .002$**

Table B6. Results of the Generalized Linear Mixed Models used to analyze the frequency and distribution of head nods. Style and Word were allowed to vary randomly by Subject and Item.

	ANALYSIS OF HEAD MOTION
(Intercept)	$\beta = -0.397$, $SE = 0.328$, $z = -1.21$, $p = .226$
Style	$\beta = -0.206$, $SE = 0.286$, $z = -0.72$, $p = .471$
Word	$\beta = -0.039$, $SE = 0.170$, $z = -0.23$, $p = .817$
Language	$\beta = 0.243$, $SE = 0.655$, $z = 0.37$, $p = .710$
Style: Word	$\beta = -0.166$, $SE = 0.307$, $z = -0.54$, $p = .588$
Style: Language	$\beta = 0.906$, $SE = 0.553$, $z = 1.64$, $p = .101$
Word: Language	$\beta = 1.359$, $SE = 0.310$, $z = 4.39$, $p < .001$***
Style: Word: Language	$\beta = 0.112$, $SE = 0.572$, $z = 0.20$, $p = .845$

4. Analysis of the duration of the functors' stressed syllable in English and Japanese

Table B7. In order to analyze whether the normalized duration of the stressed vowel of the functors increased between the first and second halves of the recordings, linear mixed models (lme4, R) were applied that use Satterthwaite approximations to degrees of freedom. These models included the fixed effects of Speech Style (IDS, ADS) and Block (first half vs. second half of the utterances) and the random factors Subject and Item. Style and Bock were centered around 0 and allowed to vary randomly by Subject and Item. No effects of Style or Block were found in the duration of the functors, neither in English nor in Japanese.

		ANALYSIS OF FUNCTOR DURATION
ENGLISH	(Intercept)	$\beta = 0.032$, $SE = 0.002$, $t(12.14) = 18.08$, $p < .001$***
	Style	$\beta = -0.000$, $SE = 0.001$, $t(7.57) = -0.07$, $p = .949$
	Block	$\beta = -0.000$, $SE = 0.002$, $t(35.11) = -0.02$, $p = .987$
	Style: Block	$\beta = -0.001$, $SE = 0.003$, $t(8.15) = -0.46$, $p = .656$
JAPANESE	(Intercept)	$\beta = 0.043$, $SE = 0.002$, $t(12.81) = 26.57$, $p < .001$***
	Style	$\beta = -0.002$, $SE = 0.002$, $t(8.13) = -0.79$, $p = .452$
	Block	$\beta = -0.000$, $SE = 0.002$, $t(19.37) = 0.23$, $p = .822$
	Style: Block	$\beta = -0.002$, $SE = 0.002$, $t(14.19) = -1.06$, $p = .307$

Appendix C Full results of the filler analysis

1. Analysis of eyebrow movements

Table C1. Number and percentage of eyebrow movement types per language and speech style and their distribution within the filler phrase.

		JAPANESE				ENGLISH				
		IDS		ADS		IDS		ADS		ALL
Eyebrow motion in target phrase	Motion	7	44%	5	36%	5	36%	5	36%	
	No motion	9	56%	9	64%	9	64%	9	64%	
	All	$n = 16$		$n = 14$		$n = 14$		$n = 14$		$n = 58$
Movement onsets	Functor	0	0%	0	0%	2	100%	5	83%	
	Content Word	5	100%	4	100%	0	0%	1	17%	
	All	$n = 5$		$n = 4$		$n = 2$		$n = 6$		$n = 17$
Movement apices	Functor	0	0%	0	0%	4	80%	3	60%	
	Content Word	6	100%	5	100%	1	20%	2	40%	
	All	$n = 6$		$n = 5$		$n = 5$		$n = 5$		$n = 21$
Apex ends	Functor	1	25%	1	100%	1	50%	0	0%	
	Content Word	3	75%	0	0%	1	50%	1	100%	
	All	$n = 4$		$n = 1$		$n = 2$		$n = 1$		$n = 8$
Movement ends	Functor	3	75%	1	33%	0	0%	1	20%	
	Content Word	1	25%	2	67%	2	100%	4	80%	
	All	$n = 4$		$n = 3$		$n = 2$		$n = 5$		$n = 14$

2. Analysis of head nods

Table C2. Number and percentage of head nods per language and speech style and their distribution within the filler phrase.

	JAPANESE		ENGLISH	
	IDS	ADS	IDS	ADS
No nods in target phrase	6	8	12	9
	37.50%	57.14%	85.71%	64.29%
Nod in the functor	3	2	0	2
	18.75%	14.29%	0%	14.29%
Nod in the content word	0	0	0	1
	0%	0%	0%	7.14%
Nod in both target words	7	4	2	2
	43.75%	28.57%	14.29%	14.29%
Total number of productions	16	14	14	14