Relating Intonational Pragmatics to the Pitch Realizations of Highly Frequent Words in English Speech to Infants

Carolyn M. Quam (cquam@psych.upenn.edu)
Department of Psychology, University of Pennsylvania
3401 Walnut Street 408C, Philadelphia, PA 19104-6228

Jiahong Yuan (jiahong@babel.ling.upenn.edu)
Department of Linguistics, University of Pennsylvania
619 Williams Hall, Philadelphia, PA 19104-6305

Daniel Swingley (swingley@psych.upenn.edu)
Department of Psychology, University of Pennsylvania
3401 Walnut Street 411C, Philadelphia, PA 19104-6228

Abstract

Infant-directed speech (IDS) is characterized by exaggerated intonation patterns and short, simple phrases. Because these exaggerated intonation patterns frequently convey a small, stereotyped range of emotional signals, we might expect particular words, like *good* or *no*, to be realized with consistent pitch contours. This consistency in a word's pitch realization might facilitate word recognition, but in an intonation language like English, it could falsely suggest lexical *tones,* i.e., pitch variation signaling lexical contrast. The present work examines the speech input to the English-learning child to identify the amount, nature, and sources of pitch variation across about 3,300 tokens of 8 highly frequent words. We find two basic results. First, although intonation in IDS is prototypically exaggerated, about half the instances of frequently occurring, utterance-final words were flat in contour. Second, although each frequent word varied substantially in its intonation contours (e.g., rises versus rise-falls), there were large differences among words that seem to reflect the pragmatic categories typical of each word's use. For instance, *no* was generally flat or falling, and consistently low in pitch, reflecting its occurrence in prohibitive utterances; while *good* occurred more often with a rise-fall contour, reflecting its approbational meaning. Even the word *good*, however, still had more flat contours than rise-fall contours. This within-word variability in pitch realization could help the child rule out lexical tone as contrastive in English.

Keywords: intonation; prosody; infant-directed speech.

Introduction

Pitch is exploited in social interactions across the animal kingdom. Because larger organisms tend to produce lower sounds, many species use low or steeply falling pitch as an expression of dominance, and higher pitch for submissiveness or uncertainty (Ohala, 1994). Ohala (1984) argues that languages often capitalize on this link to express certainty (e.g., statements) through falling or low intonation, and uncertainty (e.g., questions) through high or rising intonation.

Pitch also plays an important role in early human development. The distinctive pitch characteristics of infant-directed speech (IDS) complement the infant's developing auditory system; the higher fundamental frequency (F0) mean and wider F0 range make the speech more interesting and easier for the developing auditory system to tune in to (Fernald, 1992). Infants prefer listening to IDS over adult-directed speech (ADS; Fernald, 1985), a preference driven primarily by IDS's pitch characteristics (Fernald & Kuhl, 1987; Katz, Cohn, & Moore, 1996). Pragmatic functions of speech are expressed more clearly in IDS than in ADS. Fernald (1989) elicited utterances from mothers, intended to (1) get their infant's attention, (2) show approval, (3) comfort the infant, or (4) prohibit the infant from touching an object. Comforting utterances and prohibitions were both low in pitch and falling, but prohibitions fell more sharply and were shorter and higher in amplitude. Attention-getting and approving utterances both had high mean F0 and a large F0 range, but attention-getting utterances were higher in amplitude. Considering the clarity of intonational meaning in IDS, it's not surprising that infants respond to the emotional information conveyed by pitch variation before they know many words (Moore, Spence, & Katz, 1997).

In spite of the early importance of intonation for capturing infants' attention and conveying emotions and intentions, infants learning English must disregard lexical pitch in order to successfully learn and recognize words. By 9 months, English learners fail to discriminate a Thai lexical tone contrast (Mattock et al., 2007). And by 30 months, English learners know that pitch cannot distinguish words in English (Quam & Swingley, 2007). But figuring this out could be difficult if the intonational and syntactic simplicity of IDS leads highly frequent words to be realized with one consistent pitch contour. Is this the case for English?

To answer this question, we examine the pitch contours of highly frequent words in mothers' speech to their preverbal infants. By measuring the F0 characteristics of these words across tokens, we attempt to determine how the pitch structure of English conveys the lack of lexical tones. We might expect the pitch patterns across tokens of words to be more variable in English than in a lexical-tone language, where the pitch contour is specified in the word representation. The amount of variability across tokens could thus tell the infant which type of language she is learning. Knowing whether the pitch realizations of words

like *good* and *no* in English IDS display variability or consistency requires distributional analysis of the input to children.

Distributional analyses of both partially scripted (Kuhl & Andruski, 1997; Werker et al., 2007) and synthetic (Maye, Werker, & Gerken, 2002) speech have shed light on the acquisition of vowels and consonants. Similarly, examining the input can tell us what cues children might use to learn the pitch structure of their language. Gauthier, Shi, and Xu (2007), for example, showed that an unsupervised learning algorithm acquired the lexical-tone categories of Mandarin using either the F0 contours or velocity profiles (first derivative of F0) of syllables. The biggest limitation of most existing distributional analyses is their reliance on small, laboratory-produced corpora that may exhibit limited variation relative to children's ordinary experience. Here, in contrast, we use an automatic method of locating word boundaries to investigate a large, naturalistically produced corpus of mothers' speech to their infants (Brent & Siskind, 2001).

Methods

The Brent corpus (Brent & Siskind, 2001) from the CHILDES database (MacWhinney, 2000) is an unusually large and rich dataset for analyzing the pitch patterns of highly frequent words in IDS. The corpus contains about 100 hours of speech produced, in a naturalistic setting, by 16 mothers to their young infants, aged 9 to 15 months. There is a word transcription for each utterance, including the utterance start and end times. To evaluate the pitch patterns of individual words, we located word boundaries by forced alignment using HTK[1] and the CMU pronunciation dictionary.[2] We downsampled the sound files to 22,000 Hz, because the files had used two different sampling rates: 24,000 and 22,050 Hz. Then we trained Gaussian Mixture Model–based, monophone Hidden Markov models (HMMs) on 39 Mel Frequency Cepstral Coefficients (MFCCs) extracted from the sound files. The HMMs were adapted to each speaker using only that speaker's data. We excluded utterances from the training and our analysis when they either contained an infrequent word not in the dictionary (although frequent out-of-dictionary words were added to the dictionary by hand), or had been transcribed as noisy, sung, or whispered. This excluded roughly 6,000 of the over 126,000 utterances.

Once the HTK word boundaries were sufficiently accurate, we extracted the F0 samples for each word using Praat (Boersma, 2001), and converted each sample to the Mel scale,[3] which approximates human pitch perception. Each token's pitch samples were z-normalized using the speaker's overall mean and standard deviation[4] to control for effects of the particular speaker's pitch characteristics. Outlying pitch samples in each token's pitch track (i.e., measurement error) were excluded.[5] Then, for all words in the corpus, we calculated the F0 mean, F0 range, and the location of the F0 maximum and minimum. Further analysis focused on word tokens, from a subset of word types, in utterance-final position in statements.

Results

We first consider the F0 patterns of 23 highly frequent content words. Even in lexical-tone languages, the realization of a word's pitch is distorted by context (Xu, 1994). To reduce this distortion, we restricted our analysis to words occurring in final position in statements. (Statements are defined here as utterances transcribed with a period, versus a question mark or exclamation point.) The number of remaining tokens for each word type ranged from 150 to 1650. Figure 1 illustrates the large variation across the 23 words in their mean F0 ranges (plotted in Hertz for interpretability). For example, *good* has a mean F0 range of 135 Hz, while *now* has a much lower mean range, 66 Hz.

To investigate the nature and sources of the F0 variation across words, we examined 8 of the 23 words in more detail. These 8 words—*good, no, up, down, ball*, *book*, *right*, and *okay*—have meanings and lexical/pragmatic contexts that lead to interesting predictions about their F0 realizations. (See Appendix for detail on lexical contexts.) *Good* usually expresses approval (of the child's behavior, a taste or smell, etc.), while *no* usually chastises or warns the child. Accordingly, we expect *good* to occur with higher mean F0 and more rise-fall contours, while *no* should be low and flat or falling. The different meanings of *up* and *down* might influence their F0 patterns. In a manner analogous to tone or word painting in music (where composers fit the melody to the words of a song, for instance, jumping to a high note on the word *up*), *up* might be uttered with a higher mean F0 than *down*. *Ball* and *book*, both concrete nouns, occur in similar lexical contexts that might suggest a predominance of rise-fall contours. *Right* should behave similarly to *good*, since the mother is usually expressing excitement or praise. Finally, *okay* usually appears in comforting utterances, e.g., "You're okay" or "It's okay," so we expect its F0 realization to be low and gently falling.

[1] http://htk.eng.cam.ac.uk/, version 3.3.

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] Mel = 1127 $log_e(1 + f(Hz)/700)$; Stevens & Volkmann, 1940

[4] Z-score = (M – F0) / SD. Mean F0s for the 16 mothers ranged from 207 to 280 Hz (mean: 250 Hz), and standard deviations ranged from 72 to 99 Hz (mean: 89 Hz).

[5] We excluded any pitch values falling outside the whiskers. Whiskers were calculated for each token using the following equations: Whisker 1 = Q1 – 1.5 * (Q3 – Q1); Whisker 2 = Q3 – 1.5 * (Q3 – Q1). Q1 and Q3 are the first and third quartiles (which define the interquartile range, the middle 50% of values).

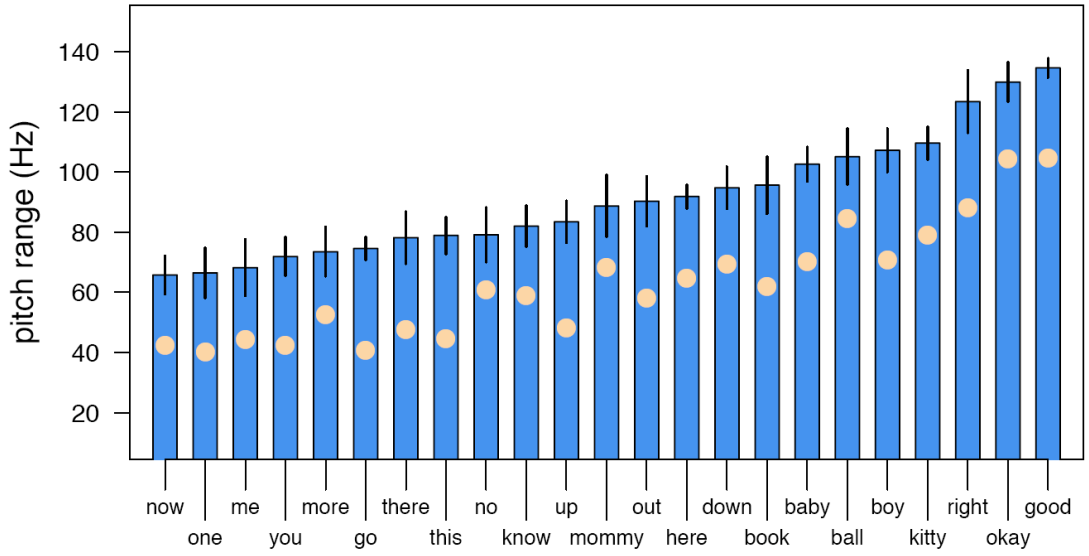

Figure 1: Pitch range for each word, plotted in Hertz for interpretability (though our analysis used the Mel scale). Means (bar heights), with their 95 % confidence intervals (vertical lines), and medians (circles) are plotted.

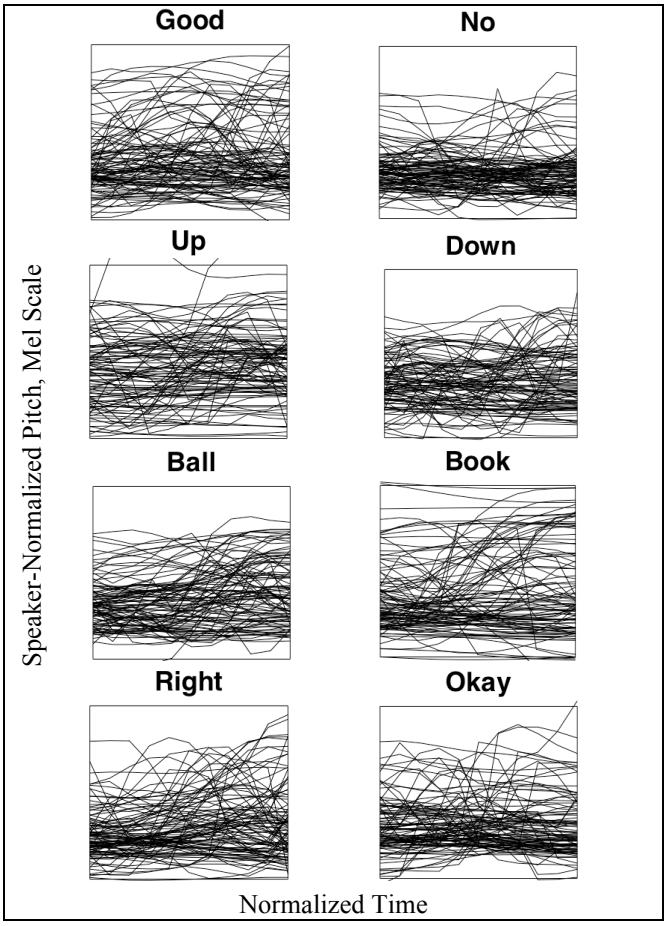

Figure 2: F0 plots (speaker-normalized Mels) for 8 highly frequent words. These tokens are all 0.3-0.4 seconds long, so they fall roughly in the middle 20%.

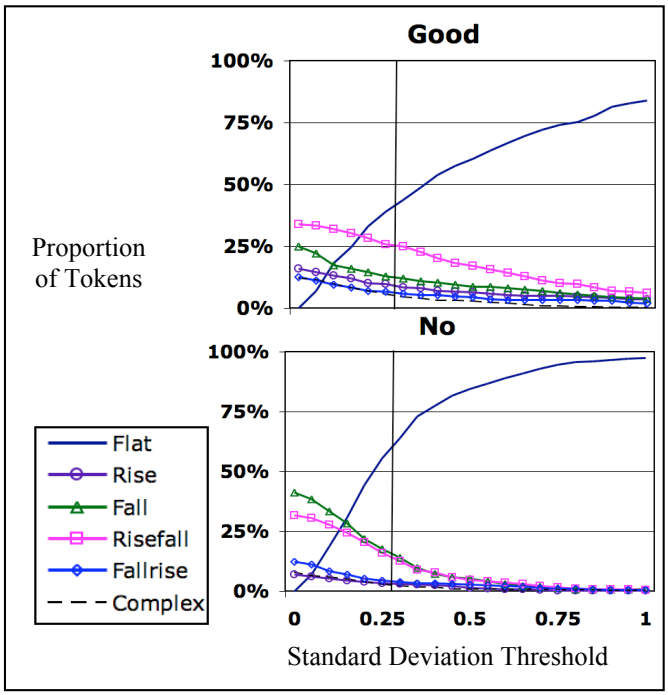

Figure 3: The standard deviation threshold. Varying the threshold affects the proportion of flat tokens for each word, but the differences between the words are evident within a large range of thresholds. We used the threshold 0.275, indicated by the vertical line.

The easiest way to get a first impression of a word's distribution of F0 realizations is to look at each word token's F0 samples plotted over time. Figure 2 shows pitch plots for a subset of the tokens that we analyzed (but similar results obtain for longer and shorter durations). For plotting purposes, each token's duration was normalized by taking

11 evenly spaced samples from the original pitch track. The F0 plots generally support our predictions. *Good* has more tokens with rise-fall contours, large F0 ranges, and high F0 means, while the *no* tokens are almost all low and flat. The *up* tokens have higher, more variable means than the *down* tokens. Surprisingly, *ball* appears to have slightly more rise-fall contours than *book.* Finally, *right* seems to have more rise-fall contours than *okay,* while *okay* has more falling tokens.

To quantify the differences between the 8 words, we first divided tokens into the categories *flat* versus *contoured*. We used the standard deviation (SD) of each token's F0 samples: if it fell above 0.275, the token was categorized as contoured; otherwise, it was characterized as flat. Figure 3 demonstrates that the particular SD threshold mainly affects the proportions of flat versus contoured tokens rather than the distributions of different contour types. At any particular SD threshold within a reasonable range, the same differences between words like *good* and *no* emerge. We then further categorized contoured tokens as *falls*, *rises, rise-falls*, *fall-rises*, or *complex*. We first normalized each token's list of pitch samples by its length, then divided the normalized duration into three regions: the start ($t \leq 3$); the middle ($3 < t \leq 7$), and the end ($t > 7$). If the F0 maximum occurred in the first region (near the start of the word) and the F0 minimum occurred in the third region (near the end), the token was categorized as *falling*. Conversely, *rises* had minima at the beginning and maxima at the end. *Rise-falls* had maxima in the middle and minima on either end, while *fall-rises* were the opposite: minima in the middle and maxima on either end. Finally, tokens that fell into none of these categories were deemed *complex*. Figure 4 illustrates the three regions and example contours of each type.

By describing each token as falling, rising, etc., we can compare the 8 words' distributions of contour types. Table 1 displays, for each word, the proportion of tokens categorized in each contour type, and the average mean F0 (converted to Mels, and z-normalized to control for each mother's pitch characteristics). The first thing to notice is the prevalence of flat tokens across all the word types, which is surprising considering that we excluded utterance-initial and utterance-medial tokens (which we would expect to be flatter than utterance-final tokens).

The contour-type distributions and F0 means in Table 1 mostly reflect the patterns observed in the F0 plots from Figure 2. *No* has more flat tokens, while *good* has more rise-falls and a higher mean F0; *up* has a higher mean F0 and more rises, while *down* has more falls; and *ball* has more rise-falls, while *book* has more flat tokens. *Right* looks strikingly similar to *good*, while *okay* has the most falling tokens of any word.

Unexpectedly, *up* has slightly more flat tokens than *no.* Since *no* occurs in prohibitive utterances, it should have more flat contours than the other words. In addition to a flat shape, however, we also expect *no* to have low mean F0s. The F0 plots in Figure 2, and the average F0 mean for each word (see Table 1), suggest the flat tokens of *up* have higher, more variable F0 means than the flat tokens of *no*. Figure 5 confirms this, comparing the distribution of F0 means for the flat tokens of *up* versus *no*. As predicted, *no*'s F0 means are more tightly clustered around lower values. Though *up* resembles *no* in its proportion of flat contours, *no* is unique in the consistency of its low mean F0.

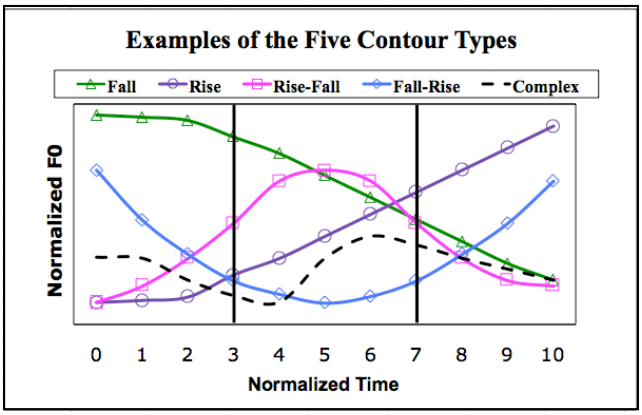

Figure 4: Examples of the five contour types. Each token is assigned a contour type using the location of its maximum and minimum F0 values.

Table 1: Contour-type distribution and average F0 mean (in Mels and z-normalized) for each word. For contour type, values greater than 0.15 are highlighted.

	Good	No	Up	Down	Ball	Book	Right	Okay
Flat	0.41	0.60	0.64	0.53	0.46	0.55	0.45	0.38
Rise	0.09	0.03	0.13	0.06	0.11	0.19	0.10	0.07
Fall	0.12	0.16	0.10	0.16	0.15	0.09	0.09	0.17
Risefall	0.26	0.14	0.07	0.15	0.22	0.07	0.26	0.16
Fallrise	0.06	0.04	0.05	0.06	0.05	0.06	0.06	0.16
Complex	0.05	0.03	0.01	0.04	0.02	0.03	0.04	0.06
F0 Mean	-0.09	-0.17	0.21	-0.21	-0.17	-0.07	0.02	0.11

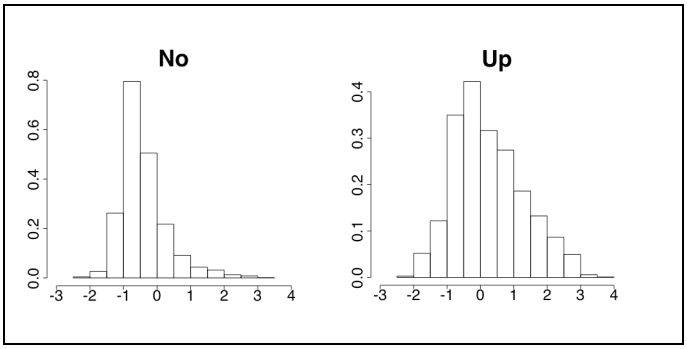

Figure 5: The distribution of mean F0 values for flat tokens of *no* versus *up*. The x-axes show mean F0 (in z-scored Mels). The y-axes show the frequency of occurrence. Flat tokens of *no*, used in prohibitions, are lower and more homogeneous in their F0 means.

Discussion

Though the 8 highly frequent words we investigated differed in their distributions of contour types, F0 means, and F0 ranges, the lack of one consistent pitch pattern within each word may cue the child that English word representations are not specified for tone.

The distributions of contour types for the words *good*, *no*, *up*, *down*, *ball*, *book*, *right*, and *okay* generally reflect the pragmatic functions of the utterances each word occurs in. *Good* and *right* often occur with a rise-fall contour, consistent with their approving function. In contrast, *no*, used in prohibitions, has predominantly flat or falling contours with low mean F0s. The opposing meanings of *up* and *down* are reflected in the higher proportion of rising contours for *up* and of falling contours for *down*, and in *up*'s higher mean F0s. Though *up,* surprisingly, had a slightly higher proportion of flat tokens than *no* did, the flat *no* tokens had lower and less variable F0 means. The higher proportion of rise-falls for *ball,* and of flat tokens for *book,* could reflect differences in the pragmatic contexts the words occur in; *book* may occur more frequently in calm, routine contexts, while *ball* may be uttered in more exciting, attention-getting contexts. For the word *okay,* we expected a large proportion of falling contours, given its comforting function. Though *okay* had the highest proportion of falling contours of any word, it was probably underestimated: the child's loud crying in comforting utterances often led to their exclusion.

Though the 8 words differ in their pitch characteristics, they also exhibit large within-word variability. Even for *right, good*, and *okay,* which occur in highly stereotyped contexts (see Appendix), the predominant contour is still flat, just as it is for *no*. The range of contour types within each word could cue the child that English word representations do not include tone. On the other hand, the consistent pitch realization of the word *no*—a crucial word to learn—probably facilitates recognition. (Changing the gender of the talker impairs young children's recognition of words, probably in large part because of the change in fundamental frequency; Singh, Morgan, & White, 2004.)

A natural next step for this research is to compare these results with the pitch contours of highly frequent words in the IDS of a lexical-tone language. Though the large within-word variability we found may cue the child that pitch is not used lexically, we do not yet know how the variability in English compares with variability in tone languages. Evidence for how reliably lexical tones are realized is unclear, with some results suggesting tones are not distorted by the exaggerated intonation of IDS (Liu, Tsao, & Kuhl, 2007; Kitamura et al., 2002), and others suggesting tones *are* distorted by IDS prosody (Papousek & Hwang, 1991). Further comparative study of multiple languages with different linguistic descriptions will help define the learning problem children face.

At present, scientific understanding of phonological development has proceeded almost entirely by empirically confirming children's gradual adaptation to language norms, with those norms described in very general terms. Such experiments testing the development of perception and production in children have revealed some of the extraordinary capabilities of infants to interpret and learn from the speech signal. But understanding the learning process in any detail will require moving beyond oversimplified, schematic descriptions of the information available to the learner. If we do not characterize the complexity and variation of the signal provided to children, we risk significantly underestimating children's ability, and distorting the nature of the developmental process. The present work provides a first step in furnishing the sort of quantitative description that will be needed for a full account of children's language learning.

Acknowledgements

We thank Kyle Gorman for his invaluable technical support and Python code, and members of the Phonetics lab and IRCS for their feedback and ideas. Funding was provided by NSF Graduate Research Fellowship and NSF IGERT Trainee Fellowship grants to C.Q., and NIH grant R01-HD049681 to D.S.

References

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5:9/10*, 341–45.

Brent, M. R. & Siskind, J. M (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81,* 31–44.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development, 8,* 181-195.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development, 60,* 1497–510.

Fernald, A. (1992). Meaningful melodies in mothers' speech to infants. In H. Papousek, U. Jurgens, & M. Papousek (Eds.), *Nonverbal vocal communication: Comparative and developmental approaches.* Cambridge: Cambridge University Press.

Fernald, A. & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development, 10,* 279–93.

Gauthier, B., Shi, R., & Xu, Y. (2007). Learning phonetic categories by tracking movements. *Cognition, 103,* 80–106.

Katz, G.S., Cohn, J.F., & Moore, C. A. (1996). A combination of vocal f0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. *Child Development, 67,* 205–17.

Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. (2002). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-

tonal language. *Infant Behavior and Development, 24,* 372–92.

Kuhl, P.K., & Andruski, J.E. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 227(5326),* 684–86.

Liu, H., Tsao, F., & Kuhl, P.K. (2007). Acoustic analysis of lexical tone in Mandarin infant-directed speech. *Developmental Psychology, 43(4),* 912–17.

MacWhinney, B. (2000). *The CHILDES database: Tools for analyzing talk, 3rd Edition. Vol 2: The database.* Mahway, NJ: Lawrence Erlbaum Associates.

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition*, *106(3)*, 1367-1381.

Maye, J., Werker, J.F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101–B111.

Moore, D.S., Spence, M.J., & Katz, G.S. (1997). Six-month-olds' categorization of natural infant-directed utterances. *Developmental Psychology, 33(6),* 980–89.

Ohala, J.J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica, 41,* 1–16.

Ohala, J.J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. In L. Hinton, J. Nichols, & J.J. Ohala (Eds.), *Sound symbolism*. Cambridge: Cambridge University Press.

Papousek, M. & Hwang, S.C. (1991). Tone and intonation in Mandarin babytalk to presyllabic infants: Comparison with registers of adult conversation and foreign language instruction. *Applied Psycholinguistics, 12,* 481–504.

Quam, C. & Swingley, D. (2007). Phonological knowledge trumps salient local regularity in 2-year-olds' word learning. *BUCLD 32,* Boston University.

Singh, L., Morgan, J.L., & White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language*, *51,* 173–89.

Stevens, S. & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *American Journal of Psychology, 53*, 329-353.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition, 103(1),* 147–62.

Xu, Y. (1994). Production and perception of coarticulated tones. *J. Acoust. Soc. Am., 95(4),* 2240–53.

Appendix: Typical lexical contexts

Good: "…very good" (106 tokens); "…so good" (46); "…that's good" (36); "…mmmm good" (29); "…it's good" (27).

No: "…no no" (607); "…oh no" (133).

Ball: "…the ball" (98); "…your ball" (35).

Book: "…this book" (32); "…the book" (28); "…a book" (25); "…your book" (23).

Up: "…it up" (60); "…you up" (54); "…stand up" (15); "…clean(ed) up" (23).

Down: "…fall/fell down" (57); "…sit down" (30); "…upside down" (20); "…get down" (17).

Right: "…that's right" (464); "you're right" (15).

Okay: "…it's okay" (147); "…you're okay" (41); "…that's okay" (32).