

## **Methodological questions in studying consonant acquisition**

JAN EDWARDS<sup>1</sup>, & MARY E. BECKMAN<sup>2</sup>

<sup>1</sup>*Department of Communicative Disorders, University of Wisconsin-Madison, WI, USA, and*

<sup>2</sup>*Department of Linguistics, The Ohio State University, Columbus, OH, USA*

*(Received 10 January 2008; accepted 6 June 2008)*

### **Abstract**

Consonant mastery is one of the most widely used metrics of typical phonological acquisition and of phonological disorder. Two fundamental methodological questions concerning research on consonant acquisition are (1) how to elicit a representative sample of productions and (2) how to analyse this sample once it has been collected. This paper address these two questions by reviewing relevant aspects of experience in evaluating word-initial consonant accuracy from transcriptions of isolated-word productions elicited from 2- and 3-year-olds learning four different first languages representing a telling range of consonant systems (English, Cantonese, Greek, Japanese). It is suggested that both researchers and clinicians should consider a number of different item-related factors, such as phonotactic probability and word length, when constructing word lists to elicit consonant productions from young children. This study also proposes that transcription should be supplemented by acoustic analysis and the perceptual judgements of naïve listeners.

**Keywords:** *Phonetic transcription, speech articulation tests, speech production measurement, phonological acquisition, child*

### **Introduction**

Children learn to talk in an extraordinarily short period of time. Over the first few years of life, they quickly progress from practicing the simple coos, squeals, and rudimentary syllables of early vocal play to saying words and longer utterances that contain recognizable forms of most of the sounds in their native language. Researchers have been investigating this developmental progression for more than a century, beginning with Taine (1876), Darwin (1877), and the more extended early 20<sup>th</sup> century diary studies that inspired Jakobson (1941/1968), but there is still much that we do not know even about *how* to study speech sound acquisition. The development of affordable portable technology for making permanent audio recordings in the middle of the 1960s and 1970s and of inexpensive digital signal analysis technology in the 1990s revolutionized our methods for studying children's productions by (among other things) allowing research teams to make multiple subsequent observations of the same set of productions. However, these technological

---

Correspondence: Jan Edwards, Department of Communicative Disorders, University of Wisconsin-Madison, Goodnight Hall, 1975 Willow Dr., Madison, WI 53706, USA. E-mail: jedwards2@wisc.edu

advances did not resolve two fundamental methodological questions: one is how to elicit a representative sample of speech sound productions and the second is how to analyse this sample once it has been collected.

This paper is a reconsideration of both questions, prompted by our experience with eliciting and analysing cross-sectional samples of word-initial lingual obstruents produced by preschool children in a series of studies of effects of phoneme frequency (Nicolaidis et al., 2003; Yoneyama, Beckman, and Edwards, 2003) and of phoneme-sequence frequency (Vodopivec, 2004; Edwards and Beckman, 2008) on word-initial consonant accuracy. The two questions became especially salient to us as we were designing the largest and most recent of these studies (Edwards and Beckman, 2008), because we wanted to compare consonant accuracy and error patterns of 2- and 3-year-old children learning four different first languages (English, Cantonese, Greek, and Japanese) in four different countries (US, Hong Kong, Greece, and Japan). We chose these four languages for several reasons: they all contain a rich set of lingual obstruents in word-initial position; they have some of these sounds in common (e.g. /s/, /t/, /k/); and online lexicons are available for all four languages so that we could compute phonotactic probabilities.

To ensure comparability across languages, we needed, first of all, to choose an elicitation method that could be used in the same way to get more or less identical samples across the different languages and cultures. More specifically, because we wanted to compare consonant accuracies in different following vowel contexts, we needed to be able to elicit a fairly prescribed set of target items in a way that would let us control for the following vowel. Given the relatively small vocabularies of 2- and 3-year-old children, this meant that it was difficult or even impossible to control for some other item-related factors, such as word length.

We also needed to choose a measure of consonant accuracy that could be applied in the same way across the languages. As a first rough measure, we chose to obtain and analyse transcriptions by native speakers who are trained phoneticians. Native-speaker transcription is an ecologically valid method of analysing a young child's productions in the sense that ultimately the child must produce sound patterns that are reliably interpreted in terms of the phoneme categories of the speech community in order to be intelligible to people outside the immediate family circle. However, expert transcription is not a direct measure of how the majority of listeners in the child's speech community will perceive a child's utterances, and we will suggest that it is time to rethink the status of transcription as an analytical tool.

The purpose of this paper, then, is two-fold. First, we would like to review the elicitation method that we chose and to provide some data to clinicians and researchers that suggest that there are influences of item-related factors such as word length on children's consonant productions. Such information should be considered by both the clinician and the researcher when constructing word lists to assess production accuracy. Second, we would like to prompt a discussion of the use of transcription as an analytic tool in both the clinic and the laboratory. Reviewing how we use transcription and how transcription fails to fully accomplish our purposes leads us to consider what other measures might be available to use instead of—or in addition to—transcription. In keeping with this two-fold purpose, the paper is divided into two sections, followed by an overall summary and conclusion section.

### **Eliciting a representative sample of productions**

As noted above, in the most recent of our studies, we elicited multiple productions from many young children using a cross-sectional design that was intended to evaluate the effects

of phoneme-sequence frequency (also called ‘phonotactic probability’) on word-initial consonant accuracy across four different languages. Our dependent measure, then, was consonant production accuracy. Therefore, we had to consider what aspects of the elicitation method might affect the accuracy of the children’s consonant productions. We intentionally varied some of these aspects and we chose not to vary others.

#### *Controlling degree of spontaneity*

One of the factors that we chose to hold constant was the degree of spontaneity of the sampled productions. This is a variable that has rarely been treated as an explicit control factor in previous research, although it does differentiate two broad classes of sampling method. The first and older method is the collection of completely spontaneous productions in a naturalistic setting (e.g. Leopold, 1949; McCurry and Irwin, 1953; Waterson, 1971; Ferguson and Farwell, 1975; Macken and Barton, 1980; Vihman, Macken, Miller, Simmons, and Miller, 1985; Shriberg, Austin, Lewis, McSweeney, and Wilson, 1997). This method has the strong advantage of being ecologically valid since, ultimately, children need to be able to talk intelligibly in connected speech about topics of interest to them and their conversational partners. However, natural speech samples have several disadvantages as well. A child may not produce all of the sounds of interest in a natural speech sample, and the segmental contexts cannot be controlled. Furthermore, if a child’s speech contains multiple articulation errors, it may be difficult to ascertain the target word.

An alternative sampling method is to elicit single words via a picture-naming or a word repetition task. This method has the advantage of allowing the researcher to control the phonetic context and to know what the target is. Because of the drawbacks involved in sampling spontaneous speech, the primary data for the majority of studies of consonant acquisition are elicited single words rather than connected speech. For example, McLeod (2007) reports on speech sound acquisition in 36 different languages or dialects. For 16 of these languages, data are available on age of acquisition of individual consonants, based on the results of 37 studies. Seventy-five per cent of these studies examined single-word productions only, an additional 14% examined a combination of single words and connected speech, and only 11% examined connected speech alone. In studies of consonant acquisition in English, scripted elicitation of single-word utterances has been used both in large-scale normative studies (e.g. Templin, 1957; Sander, 1972; Smit, Hand, Freilinger, Bernthal, and Bird, 1990; So and Dodd, 1995) and also in smaller studies that focus on the acquisition of a particular set of contrasts or on a particular phonological process (e.g. Stoel-Gammon and Stemberger, 1994; Brown and Matthews, 1997). The large-scale normative studies typically use a single word to elicit each target consonant in a specific word position, whereas the smaller studies frequently use multiple words to elicit several tokens of the types relevant to the contrast or phonological process under examination. In our study, we decided to use scripted elicitation of many different words because we wanted to ensure elicitation of all of the language’s lingual consonants in word-initial position in a variety of following vowel contexts from all of the children we recorded.

It is important to note that even within this broad class of scripted elicitation methods, there can be different degrees of spontaneity of single-word productions associated with different prompting protocols. The conventional wisdom is to differentiate between the more spontaneous productions that experimenters aim to elicit in picture-naming tasks and the purportedly more accurate imitative productions that experimenters elicit when they

use a recorded audio or live-voice verbal prompt in word-repetition tasks. (It should be noted that even when productions are elicited using picture-naming tasks, whether for an experiment or in standardized articulation tests, examiners usually use delayed or immediate imitation to prompt word productions in those instances when a child does not know the name for a particular picture.) However, the published literature on accuracy differences between spontaneous and imitative productions is inconclusive and, with the exception of one recent study on Spanish (Goldstein, Fabiano, and Iglesias, 2004), it is somewhat dated. Some of these studies have found that consonants are more accurate in imitated productions than in spontaneous ones (Kresheck and Socolofsky, 1972; Johnson and Somers, 1978), while others have found no difference between the two elicitation protocols (Templin, 1947; Paden and Moss, 1985; Goldstein et al., 2004).

In all likelihood, in cross-sectional studies such as ours, a combination of picture-naming and delayed/immediate imitation will result in a confounding age effect, with more words being produced spontaneously in response to pictures by older children who have larger vocabularies and more words being produced as imitations of a subsequent verbal prompt by younger children who have smaller vocabularies. Conservatively, we decided to use an explicitly imitative word-repetition task in our study so that we could ensure that all productions will be elicited by the same immediate pre-recorded audio prompt. We piloted this method with Greek-speaking children in Nicolaidis et al. (2003), and used it successfully in the subsequent cross-language study that is described in Edwards and Beckman (2008). More details are provided in the Appendix.

#### *Controlling for frequency of neighbouring phoneme context*

As noted earlier, experimenters use scripted elicitation methods to enable recording of a controlled sample of target forms. For example, in the cross-language studies described in Edwards and Beckman (2008), we used lists of words to elicit an even sampling of all lingual obstruents in a comparable variety of following vowel contexts in utterance-initial position. We wanted to elicit several words for each target consonant in each of five broadly defined vowel contexts because we were investigating the effects of a language-specific phoneme-sequence frequency (also termed phonotactic probability).

The effect of following vowel context frequency is one of several related effects that are of interest. For example, a number of researchers have compared consonant production accuracy across studies of children acquiring different languages to suggest that frequency of the target phoneme itself affects the time course of mastery of consonants in young children (e.g. Pye, Ingram, and List, 1987; Ingram, 1988; Yoneyama et al., 2003). Other recent studies also have shown that English-speaking children produce the same consonants more accurately in high-frequency consonant-vowel, vowel-consonant, and consonant-consonant sequences, both in real words and non-words (e.g. Edwards, Beckman, and Munson, 2004; Vodopivec, 2004; Zamuner, Gerken, and Hammond, 2004; Munson, Edwards, and Beckman, 2005; Munson, Kurtz, and Windsor, 2005). These results have been interpreted as direct effects of frequency in the input; however, an alternative interpretation of these results is that the influences of phoneme frequency and phonotactic probability are indirect—i.e. that phonetically difficult sounds and phonetically difficult sound sequences tend to be low frequency in the lexicon. To differentiate these two explanations, it is essential for one to compare phonetically similar sounds and sound sequences across languages. Therefore, phoneme-sequence frequency was the one item-related factor that we controlled most carefully in devising the lists of words to use in our

elicitation script. Because our methods for controlling phonotactic probability are described in detail in Edwards and Beckman (2008), we relegate a recap of that description to the Appendix and note here only those aspects of the results of this study that are relevant to our discussion of item-related effects that must be controlled in designing elicitation scripts.

To look at effects of context-specific frequency (phonotactic probability), we regressed the initial consonant accuracy against the word-initial CV frequency and found significant positive correlations in Cantonese [ $R^2=.16$ ,  $F(1,33)=6.225$ ,  $p<.05$ ], English [ $R^2=.46$ ,  $F(1,54)=45.2$ ,  $p<.001$ ], and Greek [ $R^2=.11$ ,  $F(1,56)=6.608$ ,  $p<.05$ ], as well as a trend in the same direction in Japanese [ $R^2=.07$ ,  $F(1,47)=3.587$ ,  $p=.06$ ]. There are number of possible reasons for the large cross-linguistic differences in the amount of variance accounted for, as discussed in detail in Edwards and Beckman (2008). For example, the phonotactic probabilities were calculated on the basis of adult lexicons, rather than child-directed speech lexicons, and there is some evidence that phonological characteristics of child-directed speech may differ in small but significant ways from that of the adult lexicon (cf. Hayashi, Yoshida, and Mazuka, 1998). It may be that the extent of these differences varies across languages.

The results of the phonotactic probability analyses are shown in Figure 1. In each of these plots, there are several plotting characters for each different target consonant, because the CV frequencies are not identical for different vowel categories. The plotting characters in bold for the three languages other than Greek illustrate why it is important to control for phonotactic frequencies in the clinic, as well as in the laboratory. These data points show the relationship between accuracy and phonotactic frequency for the following vowel contexts that are used in the words that test children's articulation in the clinic in the three languages for which we have at least one norm-referenced test of articulation. The points with parentheses around the plotting character in the plots for Cantonese and English are the predicted accuracies of the target consonants in words in the tests that use environments that were not included in our word list, namely /sœy<sup>35</sup>/ 'water' for the Cantonese consonant /s/ and *girl* and *quack* for the English consonants /g/ and /k<sup>wh</sup>/. That is, for these three data points, what we plot is not the observed accuracy for the consonant, which we cannot plot since these were not vocalic contexts elicited in our experiment, but rather the accuracy that is predicted by the regression function calculated over all of the CV sequences that we did elicit.

As the plots show, word-initial consonants elicited in items that provide a high-frequency following vowel environment tend to be more accurate than consonants elicited in items that provide a low-frequency following vowel environment. An alternative explanation for these differences in accuracy for a single consonant before different vowels is that they are due to coarticulatory effects that make a particular consonant easier to produce in one vowel context as compared to another. Whatever the explanation, an implication for researchers who are eliciting word productions to compare the relative accuracy and order of mastery of different consonants is that materials need to be chosen carefully so as not to have a confound from these effects of sequence frequency. An analogous implication for clinicians is that standardized tests of articulation (such as the *Goldman-Fristoe Test of Articulation-2* (*GFTA-2*: Goldman and Fristoe, 2000)) may not always be assessing the child's production accuracy for different consonants in fully comparable ways. For example, a child who is acquiring American English is transcribed as making an accurate production of the word-initial /k<sup>h</sup>/ in *cup* and *carrot*; however, this child misarticulates the /g/ in *girl*. Does this pattern imply better mastery of the voiceless velar, or is it simply because the voiced velar stop was elicited before the low frequency and difficult rhotic vowel?

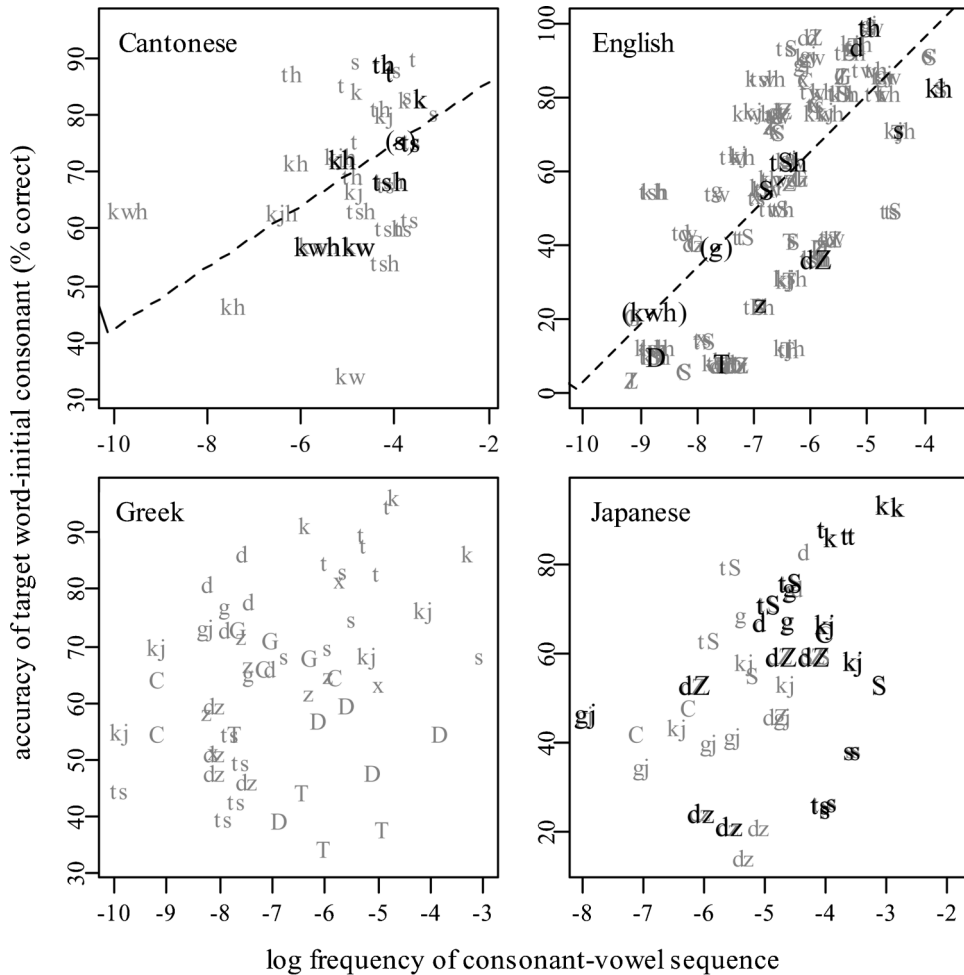


Figure 1. The four panels show the relationship between target consonant accuracy and word-initial consonant-vowel sequence frequency in the Cantonese-, English-, Greek-, and Japanese-speaking 2- and 3-year-old children recorded in Edwards and Beckman (2008). Dashed lines indicate regression curves for those languages where there was a significant relationship at the .05 level. Plotting characters in bold for the Cantonese, English, and Japanese plots are for the environments in which the word is elicited in word-initial position in the *Cantonese Segmental Phonology Test* (So, 1973), the *Goldman-Fristoe Test of Articulation-2* (Goldman and Fristoe, 2000), and the *Kōin kensa* (Nihon Chōin Gengo Hakasekai and Japan Society of Logopedics and Phoniatics, 1994). (There is no norm-referenced test in Greek, which is why no characters are emboldened in that panel.) The data point in parentheses in the Cantonese plot and the two data points in parentheses in the English plot are the predicted accuracies for the vocalic consonants for the words /soy<sup>35</sup>/ ‘water’ (which elicits /s/ in the Cantonese test) and *girl* and *quack* (which elicit /g/ and /k<sup>w</sup>/ in the English test).

*Controlling word length*

When we began to develop the target word lists for the four different languages, we also thought about which other item-related factors might influence consonant production accuracy. Although researchers have recognized that vowel context, word length, or stress pattern may have an influence on production accuracy (e.g. Kent, 1982), these factors are difficult to control in spontaneous picture-naming tasks, in which word familiarity and

pictureability severely constrain the choice of stimuli. Because we were using a word-repetition task, our choice of words was less limited, as there are many words that are familiar to children but which are not pictureable. For example, the words *thing* and *thinking* are likely to be familiar to many young English-speaking children and can be used in a word repetition task, but they would be difficult to represent in a picture-naming task.

The influence of item-related factors such as number of syllables was of particular concern to us because the four languages under consideration have very different word-internal prosodic characteristics. For example, in English the target /ka/ before a low back vowel might be elicited in a one-syllable word such as *cob* and *car*, or it might be elicited as the first part of a two-syllable word, such as *coffee*. However, we are not likely to elicit /ka/ in a longer word because most words that are longer than two syllables (e.g. *consecrate*, *congregation*) will not be familiar to such young children. Even in the vocabularies of first graders, there are seven times as many one- and two-syllable words as there are longer words (judging from counts based on the wordlist of Moe, Hopkins, and Rush, 1982). By contrast, words in Greek are typically longer than words in English. In fact, there are no one-syllable words in Greek, other than recent loan words such as /gol/ (*goal*). In Greek, the target /ka/ might be elicited in a two-syllable word such as /'kastro/ 'castle' or in a three-syllable word such as /kar'puzi/ 'watermelon'. Cantonese is similar to English in that most words familiar to children are one-to-two syllables in length, while Japanese is similar to Greek in that most familiar words are two-to-three syllables in length (two-to-four morae). We wondered whether word length would have an influence on consonant accuracy and, if so, whether this influence would vary across the different languages. Although we had not purposely varied word length in order to test its effects, there turned out to be enough variety to make a post-hoc comparison.

Specifically, in English and Cantonese, for each CV sequence, we could compare the accuracy of each target consonant on monosyllabic words to polysyllabic words. In Greek and Japanese, we could compare the accuracy of target consonants on 1- and 2-syllable words to accuracy on 3- and 4-syllable words. Figure 2 shows the effect of word length in

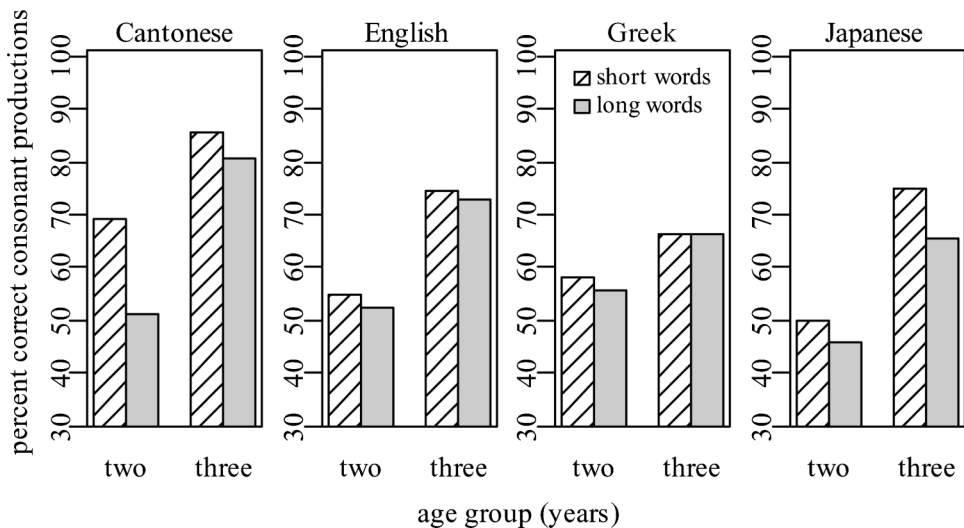


Figure 2. Mean percentage correct consonant productions for monosyllabic words versus polysyllabic words (in Cantonese and English) and for disyllabic and monosyllabic words versus trisyllabic and longer words (in Greek and Japanese).

number of syllables for each of the four languages. We had predicted that, if there is an effect, consonants beginning 'easier' shorter words would be more accurate. The difference was in the predicted direction in all four languages and was significant in all languages except Greek (Cantonese: mean difference=10.7%,  $t(21)=3.9$ ,  $p<.001$ ; English: mean difference=2.1%,  $t(22)=1.8$ ,  $p<.05$ ; Japanese: mean difference=6.2%,  $t(19)=2.7$ ,  $p<.01$ ).

It is unclear why the effect of word length was significant in Japanese but not in Greek, given that both languages have generally longer words. It may be that we had more variability in word length in Japanese than in Greek, so there were more item pairs to compare. These factors cannot be sorted out without an experiment that is explicitly designed to examine the effect of word length on consonant accuracy. Nevertheless, the fact that we did find a significant effect in three of the languages even though word length was not varied systematically in this experiment is suggestive. The implication for researchers who are eliciting word productions to compare relative accuracy or relative order of mastery of different consonants is that they need to control for word length either directly in the design of the materials to be elicited or statistically in a post-hoc way by entering word length as a factor in regression analyses. The implication for clinicians is comparable. If a Japanese-acquiring child whose speech is assessed using the *Kōin kensa* (Nihon Chōin Gengo Hakasekai and Japan Society of Logopedics and Phoniatics, 1994) produces the sibilant fricative /s/ correctly in the words *semi* 'cicada' and *sora* 'sky' but misarticulates the corresponding affricate at the beginning of *tsumiki* 'blocks', is the difference because the child has relatively less mastery of /ts/ or is it because the affricate is elicited in a longer word-form than the fricative in this standardized test of articulation?

#### *Controlling for other item-related effects*

The four languages also differ with respect to their use of stress and lexically-distinctive pitch patterns. Two of the four languages studied are stress-accent languages. That is, both English and Greek have contrasting patterns of syllable prominence specified in the lexicon such that, for example, the English words *collar*, *Jerry*, and *tomahawk* have stressed initial syllables, whereas in *collide*, *giraffe*, and *tomato* the initial syllable is unstressed. English and Greek differ in the frequency of position of the stressed syllable in the word. In English, most words have first-syllable stress, and this is especially true of words that are familiar to children. Greek, by contrast, has many more words with stress on the second syllable, including many words that are familiar to children such as /ma'ma/ 'mother', /ba'ba/ 'father', /jia'jia/ 'grandmother', /pa'pus/ 'grandfather', /mo'ro/ 'baby', /ka'lo/ 'good', and /ka'ko/ 'bad'. For English, we could control for stress by choosing only target words with initial syllable stress. We decided to do so because it is well known that young children learning English frequently delete initial unstressed syllables and we wanted to ensure that the word-initial target consonants were produced (e.g. Kehoe and Stoel-Gammon, 1997). By contrast, it is difficult to find enough familiar words of Greek to control for stress in the same way. A by-product of our study of phonotactic probability effects, therefore, is that we can evaluate the effect of stress on consonant accuracy ex-post facto. A question that we will address in this paper, therefore, is whether there are similar effects to the ones described for English. That is, even though second-syllable stress is common in Greek, we might expect word-initial consonant accuracy to be higher in words with initial stress because stressed syllables in Greek are typically more clearly articulated than unstressed syllables, with higher amplitudes and longer durations (Arvaniti, 2000).



Neither Cantonese nor Japanese has anything like the patterns of stress in English and Greek. However, the Cantonese lexical tone system does include a contrast that could plausibly have effects analogous to the effects of the contrast between shorter unstressed syllables and longer stressed syllables in English and Greek. This is the contrast between shorter ‘checked-tone’ syllables, with final /p, t, k/, and all other tone syllable types, with longer sonorant rhymes. Also, Japanese has a lexical pitch contrast that native speakers of Greek and English often assimilate to the stress patterns of their native languages. This is the contrast between words containing a pitch accent (a steep fall in pitch at a lexically designated syllable) and words that do not contain a pitch accent. We examined whether Cantonese-speaking children would produce word-initial consonants less accurately if the first syllable was a checked-tone syllable. We also examined whether Japanese-speaking children would produce consonants more accurately if the first syllable contained a high tone rather than a low tone, as this is the more usual tone pattern in child-directed words (Kubozono, 2003), although not in the adult lexicon.

Figure 3 shows the effect of stress in Greek in the left-most panel. We had predicted that consonants that are onsets of stressed syllables would be imitated more accurately than consonants that are onsets of unstressed syllables. A pairwise comparison across subject-by-subject means showed a trend in the predicted prediction (see the leftmost panel of Figure 3). The mean difference overall was 7.1% with a one-sided paired t-test yielding  $t(19)=2.8, p<.01$ .

The remaining three panels of Figure 3 show the influence of pitch accent and syllable structure on production accuracy in Japanese. (There was no effect of checked tone versus sonorant rhyme in Cantonese.) We had predicted that consonants that are onsets of syllables with an associated high tone might be more accurate than consonants that are onsets of first syllables with an associated low tone, as the former pattern seems to be much more frequent in child-directed speech than the latter and also is preferred by infants (Hayashi, 2003; Kubozono, 2003). A pairwise comparison across subject-by-subject means showed a trend in the predicted direction, as shown in the second panel of Figure 3 (mean

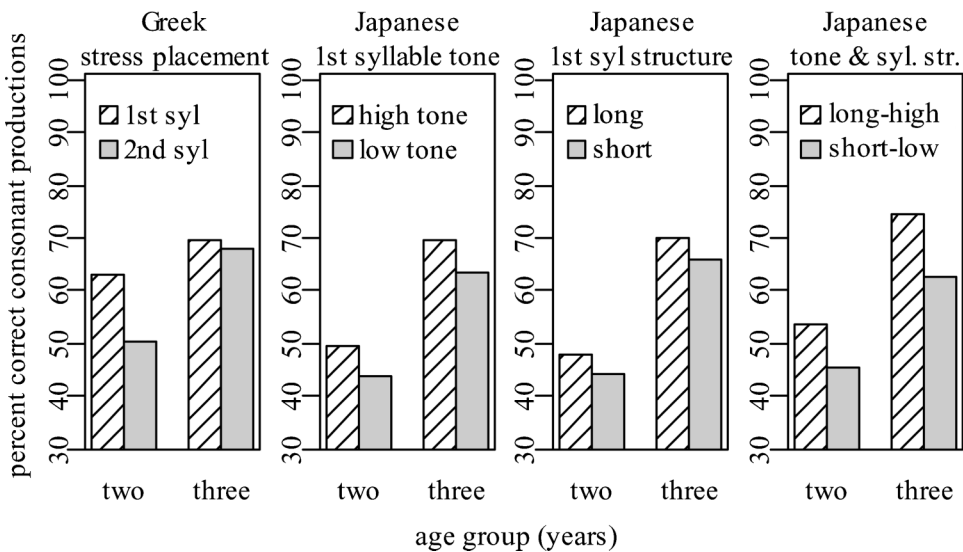


Figure 3. Mean accuracy by age group for initial consonant in pairs of words that contrasted in prosodic shape for Japanese and Greek.

difference overall=5.8%,  $t(19)=2.2$ ,  $p=.02$ ). For Japanese, we also compared consonants in short, monomoraic syllables with consonants in long, bimoraic syllables. There was a trend (albeit insignificant) in the predicted direction, as shown in the third panel of Figure 3. The mean difference in a paired comparison by subjects was 4.0% with a one-sided paired t-test yielding  $t(19)=2.0$ ,  $p=.03$ . We also predicted that any effect of high versus low tone would be magnified when it was combined with the effect of syllable structure. This prediction was borne out, as shown in the rightmost panel of Figure 3 (mean difference overall=9.7%,  $t(19)=2.7$ ,  $p<.01$ ).

The implications of these results for researchers are similar to those discussed above for the effects of word length in Cantonese and Japanese. More generally, it seems reasonable to conclude from the various item-related effects that we found either by testing for them directly or in post-hoc analyses, that clinicians should be aware of attested or potential item-related effects as they develop and evaluate assessment tools.

### Transcription and alternative analytic tools

The second issue of concern to us is the role of transcription in the analysis of children's productions. Transcription is, of course, a very familiar tool to both researchers and clinicians. In clinical settings and in most research studies, we rely on transcription by a phonetically-trained native speaker to determine whether a production is correct or incorrect. For example, in Yoneyama et al. (2003) we used three native speaker phoneticians to transcribe all targets and used only those tokens where at least two out of three agreed. In our most recent study, similarly, we used a native-speaker phonetician-transcriber for each of the four languages, with a second native speaker re-transcribing 10% of the data to provide a measure of reliability. In Nicolaidis et al. (2003), we had a highly-trained phonetician who is not a native speaker of Greek do the first-pass transcription, with checking and correction by a second highly-trained phonetician who is a native speaker.

Native-speaker transcription is an ecologically valid method of analysing the young child's productions in the sense that ultimately a child must produce sound patterns that are reliably interpreted in terms of the phoneme categories of the speech community in order to be intelligible to people outside the immediate family circle. However, in this paper we will suggest that we need to rethink the status of transcription as an analytical tool.

#### *The uses of transcription*

Transcription is traditionally used for two different purposes. First, transcription is used (in cross-sectional studies such as Smit et al. (1990), and in the measure of error rate in such standardized tests as the *GFTA-2*) as a *phonemic* measure of how the child will be perceived by the ambient speech community. This phonemic use of transcription asks the transcriber to decide if the child's production is correct or incorrect. In many clinical contexts, such as the administration of a standardized articulation test, this may even be done live, without the use of audiotape. Second, transcription is used (e.g. in studies such as Dinnsen, Gierut, and Chin (1987) and in 'process' analyses such as the *Khan-Lewis Phonological Analysis* (Khan and Lewis, 2000)) as a *phonetic* measure of production. This phonetic use of transcription asks the transcriber to provide a fairly narrow transcription of the child's production if it is incorrect (e.g. Shriberg, Kwiatkowski, and Hoffman, 1984; Louko and Edwards, 2001; Stoel-Gammon, 2001). Typically, the same person does both the phonemic and phonetic coding, whether for research purposes or in the clinic.

However, the kinds of transcription that are needed for the two different purposes are mutually incompatible. For the first purpose, the transcriber should be fairly naïve, and simply listen to the child, rather than look at the spectrogram or other acoustic measures. Also, the transcriber should not be asked to transcribe very much speech from any one child, lest there be a progressive accommodation to a child's habits as the transcriber becomes attuned to the child in the same way that members of the child's immediate circle typically are. For the second purpose, the transcriber should be a fairly sophisticated phonetician and rely on spectrograms and a close auditory inspection of the waveform. However, the documentation of phenomena such as covert contrasts and cross-language differences in cue-weighting suggests that transcription alone cannot accomplish this second purpose, unless it is supplemented by more systematic acoustic analysis.

Covert contrast is defined operationally as statistically reliable acoustic differences that are not perceptible to naïve listeners. Covert contrast has been observed in English in both typically developing children and children with phonological disorders for a number of different phonetic contrasts, including the voicing contrast for stop consonants, acquisition of final consonants, and the contrast between /s/ and /θ/ (e.g. Macken and Barton, 1980; Maxwell and Weismer, 1982; Baum and McNutt, 1990; Scobbie, Gibbon, Hardcastle, and Fletcher, 2000). Tsurutani (2004) also found some evidence of covert contrast in relation to the /s/-/ç/ contrast in the productions of Japanese-acquiring children in a small number of repetitions. Covert contrast is of interest to researchers, as it provides a finer-grained window into children's phonetic development and it is also relevant for clinicians with respect to prognosis and treatment decisions. For example, Tyler (1995) showed that children with phonological disorders who produced covert contrasts progressed through speech-sound therapy more quickly than children who did not differentiate between target sounds and their errors.

One important finding is that sometimes there is covert contrast because the child has latched onto a secondary cue and misinterpreted it as more important than the primary cue. This was the situation in the case studies described in Scobbie et al. (2000) and Frank (1998). A parallel finding in second-language acquisition is that some second-language contrasts may be particularly difficult because the cues to the contrast may also be harnessed for a first-language contrast, but they are weighted differently. For example, Yamada and Tohkura (1990) and Iverson, Kuhl, Akahane-Yamada, Diesch, Tohkura, Kettermann, and Siebert, (2003) show that Japanese speakers weight second- and third-formant transition cues differently from English speakers in the perception of approximants such as /w/, and this difference in cue-weighting seems to be strongly implicated in the notorious difficulty that Japanese speakers have in distinguishing English /ɹ/ and /l/. Researchers also have demonstrated cross-language differences in cue weighting for place of articulation contrasts for some of the word-initial obstruents that we are studying in our project (e.g. McGuire, 2007). These differences have important implications for the use of transcription in analysing children's productions, particularly in cross-language studies such as ours.

As Pye, Wilcox, and Siren (1988) pointed out, there is also a problem with how inter-transcriber disagreements are treated, whether the transcription is phonetic or phonemic. Typically, a second transcriber will independently transcribe a small proportion of the data and inter-rater disagreements are noted. Depending on the protocol, items on which transcribers disagree may be excluded (e.g. Yoneyama et al., 2003), or both transcribers may listen to these items together until they can agree on a transcription (Shriberg et al., 1984). The problem with both of these methodologies is that inter-transcriber disagreements are treated as 'noise' in the data. However, the results in Pye et al. (1988)

suggest that these disagreements are likely to be informative, since they tend to occur on children's productions that do not clearly belong to a single phoneme category.

### *Language transcriber effects*

In Nicolaidis et al. (2003), we noted several fairly systematic discrepancies between the first-pass transcriptions done by the third author (who was originally a first-language bilingual of English and Japanese) with the checked and corrected transcriptions of the first author (a native speaker of Greek). In our larger cross-language study, on the other hand, transcription was done by a single native-speaker trained phonetician for each language and we did not systematically analyse inter-rater reliability among native and non-native transcribers. However, the transcribers often listened to the productions from the other languages and informal discussion in the laboratory revealed systematic differences between native and non-native transcribers for some consonants and consonant-vowel sequences. Some of these differences were readily explicable. For example, the English transcriber classified the voiceless unaspirated stops /t/ and /k/ of Greek as /d/ and /g/. This discrepancy is easy to explain. In word-initial position, the voicing contrast in Greek is between the fully voiced stops /b, d, g/ and the voiceless unaspirated stops /p, t, k/, whereas the voicing contrast in English is between the often voiceless unaspirated stops (called /b, d, g/ in English) and the voiceless aspirated stops /p<sup>h</sup>, t<sup>h</sup>, k<sup>h</sup>/. Thus, it is not surprising that an English speaker would assimilate Greek voiceless stops to her voiced-stop category.

However, some of the cross-linguistic differences were initially more surprising. Two of these differences are illustrated in Figure 4. The top panel of Figure 4 shows the waveform and spectrogram for a production of the /k<sup>i</sup>e/-initial target word κέντρο 'centre' elicited from a Greek-speaking 2-year-old girl in the Nicolaidis et al. (2003) study. The bottom panel shows the waveform and spectrogram for the production of the /se/-initial word *senaka* 'back' elicited from a 3-year-old Japanese-speaking girl in the Edwards and Beckman (2008) study.

The Greek /k<sup>i</sup>e/ production in the top panel has been transcribed in at least three different ways. It is perceived as /dε/ or /t<sup>h</sup>ε/ by our English-speaking transcribers and by every other English-speaking listener to whom we have played the production. It was transcribed as /tʃe/ in the first-pass transcription by the originally bilingual third author, and this is typical of the perception of this (and other similar) targets by the Japanese-speaking listeners to whom we have played this token. However, the transcription was corrected by the native-Greek-speaking transcriber on the second pass, who perceived it to be a correct albeit somewhat affricated production of the target consonant. This matched the perception of the fourth author of the Nicolaidis et al. (2003) paper, as well as that of our current Greek-speaking transcriber. In the productions of this and other words that we have used to elicit the voiceless dorsal stops of Greek, there have been many of these very front Greek /k<sup>i</sup>i/ and /k<sup>i</sup>e/ productions that have been categorized in analogous ways—i.e. as correct by the Greek-speaking transcriber, but as an alveolar stop by English-speaking transcribers and as an alveolo-palatal by the second author of this paper as well as by other Japanese-speaking transcribers who are not first-language speakers of English.

A somewhat different pattern was observed for the Japanese /s/ production in the lower panel and similar productions of target /s/ for the language. This production of /se/ was transcribed as an *incorrect* /ʃ/-for-/s/ substitution by the native-Japanese-speaking transcriber, but was categorized as a correct /s/ by the native-English-speaking transcriber. The Japanese post-alveolar sibilant fricative /ʃ/ has a more palatal place of articulation than

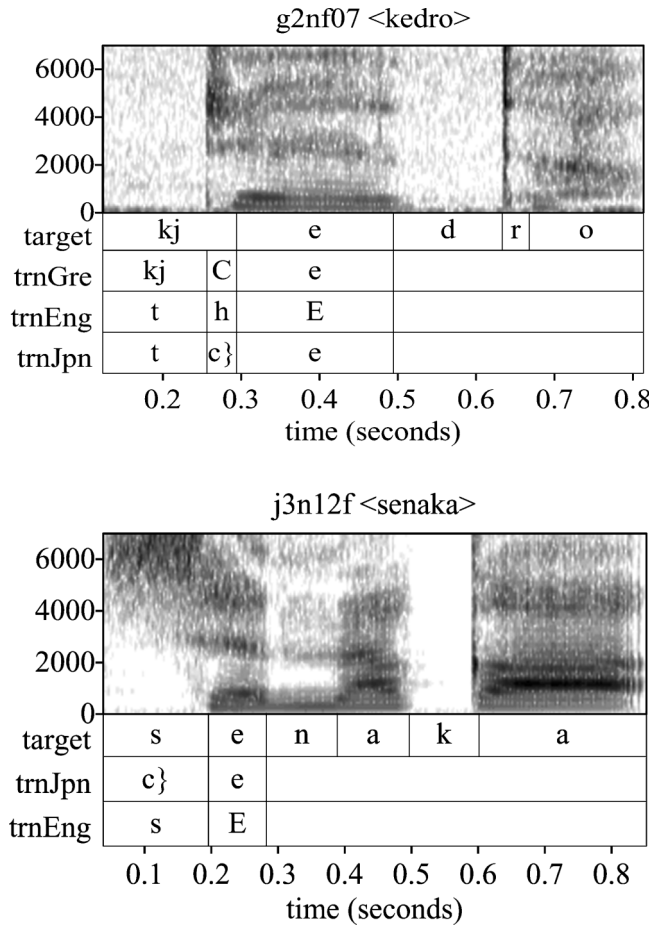


Figure 4. Spectrogram and waveform for production of /kedro/ by Greek-speaking 2-year-old girl (top plot) and production of /senaka/ by Japanese-speaking 3-year-old girl (bottom plot).

the English /ʃ/, but is readily assimilated to /ʃ/ by English speakers as in the loan word *sushi*. Again, there were many similar examples of ‘incorrect’ Japanese /s/ productions by this child and other children.

We suspect that these cross-linguistic transcriber differences are related to differences in fine phonetic detail for these phoneme categories across languages. Arbisi-Kelm, Beckman, and Edwards (2007) found that the peak amplitude frequency for the /kʲ/ burst before /i/ and /e/ was higher in Greek relative to English /kʰ/. This finding suggests that the voiceless dorsal stop before front vowels has a more anterior place of articulation in Greek as compared to English. This more front place of articulation would make the Greek dorsal stop intermediate between English /tʰ/ and English /kʰ/ and could explain the perception of a velar-fronting error by the English listeners. The Japanese /kʲ/ before /i/ also has a very high burst peak. However, unlike in Greek, the front dorsal stops in Japanese contrast with alveolo-palatal affricates.

The different categorizations of Japanese /s/ by Japanese and English listeners were of particular interest because of a reported difference in phoneme order of acquisition

between English- and Japanese-speaking children. English-acquiring children generally master /s/ before they master the postalveolar fricative /ʃ/ (Smit et al., 1990), while Japanese-acquiring children have the opposite pattern; they master the analogous postalveolar fricative /ç/ before they master /s/ (Nakanishi, Owada, and Fujiki, 1972). We wondered what the relationship was between this cross-linguistic difference in phoneme acquisition and the differing perceptions of our Japanese and English transcribers. The work of Li and colleagues (Li and Edwards, 2006; Li, Edwards, and Beckman, 2007; Munson, Li, Yoneyama, Hall, Beckman, Edwards, and Sunawatari, 2008) suggests some answers. Li and colleagues used both acoustic analysis of the children's productions and a speech perception task using the children's productions as stimuli with naïve Japanese and English listeners to study this question. The acoustic analysis revealed that the English /s/ occupies a larger acoustic space than the Japanese /s/ (as defined by measures from a spectral moment analysis, as well as F2 onset frequency). For the perception task, Li and colleagues asked 20 English and 20 Japanese naïve listeners to categorize English and Japanese children's productions of /s/ and /ʃ/ (or /ç/) as well as English [s]-for-/ʃ/ and Japanese [ç]-for-/s/ substitutions in a speeded response task. One interesting result was that naïve Japanese listeners, like the Japanese-trained native speaker/transcriber, had a narrower range for /s/ and a larger range for the post-alveolar fricative even in English: Japanese listeners rated English children's /ʃ/ productions as more accurate than their /s/ productions. Another important result emphasized our earlier concerns about transcription. Li and colleagues found that judgements of multiple naïve listeners uncovered gradience in listeners' judgements of children's phonetic accuracy for both languages. While phoneme-by-phoneme inter-rater reliability between two trained native speaker/phoneticians across the whole consonant set was 89% for Japanese and 90% for English, we found that agreement between the 20 naïve native speakers for each language and the trained phonetician ranged from a high of 94% (naïve English-speakers agreement with trained phonetician for incorrect /s/ productions) to a low of 64% (naïve Japanese-speakers agreement with trained phonetician for incorrect /s/ productions).

## Summary

This paper discussed two methodological issues related to the study of phonological acquisition in children. These issues were how best to elicit a representative sample and how to analyse this sample once it has been collected. Our exploration of these issues was prompted in large part by our experience of designing and conducting a large cross-sectional study comparing young children's single-word productions across four languages with substantial differences in phoneme inventory, phonotactic constraints, and prosody. We used our previously elicited samples to explore some of the item-related factors that affect the validity of results of scripted elicitation. We could only look at a small subset of the range of possible factors because all of the children's productions were single-word imitations of familiar words. Such a task is probably the simplest of all speech production tasks. Both picture-naming and spontaneous speech place greater cognitive and linguistic demands on the child than imitation of single words. The children recorded for our study simply had to produce a single word following an auditory model and a picture prompt. Further, the target words were familiar words that they had probably heard and said many times before. We might expect the observed item-related effects on initial consonant accuracy to be even greater in other contexts, such as picture-naming or spontaneous speech, where there are greater cognitive and linguistic demands on the child.

One point that we have begun to appreciate from looking across languages in this way is the value of doing controlled cross-language comparison. Even when fairly simple tasks are used in eliciting the sample, results are more generalizable than studies that focus on just one language. One item-related effect that we found was the effect of phonotactic probability (the frequency of the target consonant-vowel sequence relative to other consonant-vowel sequences in the ambient language). Low-frequency consonant-vowel sequences were produced less accurately than high-frequency sequences. This effect of phonotactic probability accounted for between 46% (for English) and 7% (for Japanese) of the overall variance in consonant accuracy across the four languages. This finding suggests that both researchers and clinicians should consider phonotactic probability when they are choosing stimulus items. For example, one widely-used norm-referenced articulation test is the *GFTA-2* (Goldman and Fristoe, 2000). This test, like most standardized articulation tests, elicits each consonant in English one time in each word position using a picture-naming task. The consonant /t<sup>h</sup>/ is elicited in word-initial position in a /t<sup>h</sup>ε/ context (in *telephone*), which is the most frequent word-initial consonant-vowel sequence for /t<sup>h</sup>/ in English. The sequence /t<sup>h</sup>ε/ begins 113 different words in the *Hoosier Mental Lexicon* (HML, Nusbaum, Pisoni, and Davis, 1984; Pisoni, Nusbaum, Luce, and Slowiaczek, 1985), while /t<sup>h</sup>Λ/ begins only 29 different words. By contrast, the consonant /k<sup>h</sup>/ is elicited in word-initial position in a /k<sup>h</sup>Λ/ context (in *cup*), one of the less frequent contexts for /k<sup>h</sup>/. The sequence /k<sup>h</sup>Λ/ begins 48 different words in the HML, as compared to /k<sup>h</sup>æ/ which begins 257. Unfortunately, we do not know of any norm-referenced articulation tests that control for phonotactic probability across different consonants, and it would probably be very difficult to design an articulation test that does so. An alternative to controlling phonotactic probability across different consonants would be to elicit consonants in several different vowel contexts. Such a method has the advantage of providing additional information to the clinician or researcher—namely, whether phonotactic probability influences consonant production accuracy for a particular child.

In our post-hoc analyses of the consonant accuracy in the Edwards and Beckman (2008) study, we found that word length and other prosodic factors also influenced word-initial consonant accuracy, although these factors had a smaller effect than the effect that we had targeted in designing this study. Moreover, the effect of these prosodic factors varied across languages. Nevertheless, these results suggest that researchers and clinicians might want to consider these factors also when constructing word lists to assess consonant accuracy. For clinicians, the primary consideration is to elicit a valid sample. This process involves an understanding of what the child can produce under less challenging and more challenging conditions. Insofar as possible, word lists should be constructed so that consonants are elicited in words that vary with respect to length and other prosodic factors, such as stress position, so that the influence of these factors on accuracy can also be examined.

The second question of interest to us was how to analyse a sample once it has been collected. Like most other clinicians and researchers, we have relied primarily on transcription in our research to date. However, the cross-linguistic differences that we observed in phoneme categorization emphasize some of our concerns about relying only on transcription to analyse children's productions. We found that the same consonant was categorized as belonging to a different phoneme category depending on the transcriber's experience (that is, her native language). Furthermore, when we compared the judgements of multiple naïve listeners to that of a single trained native speaker/transcriber, we found that there were sometimes significant discrepancies between these two measures. The conclusion that we draw from these discrepancies is that we cannot base clinical or research

decisions about consonant accuracy solely on transcriptions of consonants as correct or incorrect. As other researchers have suggested, transcription procedures can be modified to provide additional information if the transcriber includes information about intermediate productions (productions that sound as if they are in between two consonants or vowels), non-English sounds, and secondary articulations (e.g. Louko and Edwards, 2001; Powell, 2001; Stoel-Gammon, 2001). One recent study in our laboratory (Schellinger, Edwards, Munson, and Beckman, 2008) found that naïve listeners rated children's productions of /s/ that had been transcribed as intermediate between [s] and [θ] as less accurate than children's productions of /s/ that were transcribed as clear [s] for /θ/ substitutions. This observation provides some validation to using intermediate categories in transcription.

As researchers, we also need to put more emphasis on developing good acoustic measures, and these acoustic measures must be tailored to the language that the child is acquiring as well as being easy to use in the clinic. This need is particularly relevant now that there are several free and highly user-friendly waveform editors available to clinicians, such as Praat (Boersma, 2001) or WaveSurfer (Sjolander and Beskow, 2000). We also need to make sure that clinicians understand the importance of using these measures in addition to transcription. Furthermore, we need to validate the acoustic measures that we develop with cross-language comparisons of adult perception patterns. A good model for us is Li and colleague's work on acquisition of sibilant fricatives—work that combines transcription, acoustic analysis, and perception by naïve listeners.

### Acknowledgements

Supported by NIDCD grant R01 DC02932 to Jan Edwards. Thanks to Tim Arbisi-Kelm, Hyunju Chung, Junko Davis, Kiwako Ito, Eunjong Kong, Fangfang Li, Katerina Nicolaidis, Sarah Schellinger, Laura Slocum, Asimina Syrika, Georgios Tserdanelis, Peggy Wong, and Kiyoko Yoneyama for their work on data collection, native-speaker transcription, and acoustic analysis. Thanks also to the children who participated in the various studies, the parents who gave their consent, and the schools that let us use their facilities for testing.

**Declaration of interest:** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

### References

- Amano, S., & Kondo, T. (1999). *Lexical properties of Japanese*. Tokyo: Sansendo.
- Arbisi-Kelm, T., Beckman, M. E., & Edwards, J. (2007). Acquisition of stop burst cues in English, Greek, and Japanese. Poster presented at the Symposium for Research in Child Language Disorders, Madison, WI, June.
- Arvaniti, A. (2000). The acoustics of stress in modern Greek. *Journal of Greek Linguistics*, 1, 9–39.
- Baum, S. R., & McNutt, J. C. (1990). An acoustic analysis of frontal misarticulation of /s/ in children. *Journal of Phonetics*, 18, 51–63.
- Beckman, M. E., Yoneyama, K., & Edwards, J. (2003). Language-specific and language-universal aspects of lingual obstruent productions in Japanese-acquiring children. *Journal of the Phonetic Society of Japan*, 7, 18–28.
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott International*, 5, 341–345.
- Brown, C., & Matthews, J. (1997). The role of feature geometry in the development of phonemic contrasts. In S. J. Hannahs, & M. Young-Scholten (Eds.), *Focus on phonological acquisition* (pp. 67–112). Philadelphia: Benjamins.
- Chan, S. D., & Tang, Z. X. (1999). Quantitative analysis of lexical distribution in different Chinese communities in the 1990's. *Yuyan Wenzhi Yingyong [Applied Linguistics]*, 3, 10–18.
- Darwin, C. (1877). A biographical sketch of an infant. *Mind: A Quarterly Review of Psychology and Philosophy*, 2, 285–294.



- Dinnsen, D. A., Gierut, J. A., & Chin, S. B. (1987). Underlying representations and the differentiation of functional misarticulators. Paper presented at the Annual Convention of the American Speech-Language-Hearing Association, New Orleans, LA, November.
- Edwards, J., & Beckman, M. E. (2008). Some cross-linguistic evidence for modulation of implicational universals by language-specific frequency effects in phonological development. *Language, Learning, and Development*, 4, 122–156.
- Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, 47, 421–436.
- Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language*, 51, 419–439.
- Frank, S. A. (1998). The acquisition of the voicing contrast in word-initial alveolar and velar stop consonants: a longitudinal case study of one phonologically disordered child. Unpublished MS thesis, The Ohio State University, Columbus.
- Gavriliidou, M., Labropoulou, P., Mantzari, E., & Roussou, S. (1999). Prodiagrafes gia ena ipologistiko morphologiko lexiko tis Neas Ellinikis [Specifications for a computational morphological lexicon of Modern Greek]. In A. Mozer (Ed.), *Greek Linguistics*, 97, *Proceedings of the 3<sup>rd</sup> International Conference on the Greek Language* (pp. 929–936). Athens: Ellinika Grammata.
- Goldman, R., & Fristoe, M. (2000). *The Goldman-Fristoe Test of Articulation-2*. Bloomington, MN: Pearson Assessments.
- Goldstein, B., Fabiano, L., & Iglesias, A. (2004). Spontaneous and imitative productions in Spanish-speaking children with phonological disorders. *Language, Speech, Hearing Services in Schools*, 35, 5–15.
- Hayashi, A. (2003). Perception and acquisition of rhythmic units by infants. *Journal of the Phonetic Society of Japan*, 7, 29–34.
- Hayashi, A., Yoshida, K., & Mazuka, R. (1998). Baby-word rhythm preferences of Japanese infants. *Proceedings of the 16th International Congress on Acoustics and the 135th Meeting of the Acoustical Society of America*, Vol. 2, 2067–2068.
- Ingram, D. (1988). The acquisition of word-initial [v]. *Language and Speech*, 31, 77–85.
- Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, 87, B47–B57.
- Jakobson, R. (1941/1968). *Kindersprache, aphasie und allgemeine lautgesetzte*. [Published originally in 1941; citations are from the 1968 translation by A. R. Keiler, *Child language, aphasia, and phonological universals*. The Hague: Mouton.]
- Johnson, S., & Somers, H. (1978). Spontaneous and imitated responses in articulation testing. *British Journal of Disorders of Communication*, 13, 107–116.
- Kehoe, M., & Stoel-Gammon, C. (1997). Truncation patterns in English-speaking children's word productions. *Journal of Speech, Language, and Hearing Research*, 40, 526–541.
- Kent, R. D. (1982). Contextual facilitation of correct sound production. *Language, Speech, and Hearing Services in Schools*, 13, 66–76.
- Khan, L., & Lewis, N. (2000). *Khan-Lewis Phonological Analysis: 2*. Bloomington, MN: Pearson Assessments.
- Kresheck, J., & Socolofsky, G. (1972). Imitative and spontaneous assessment of 4-year-old children. *Journal of Speech and Hearing Research*, 15, 729–733.
- Kubozono, H. (2003). Acquisition of phonology and language universals. *Journal of the Phonetic Society of Japan*, 7, 5–17.
- Leopold, W. F. (1949). *Speech development of a bilingual child: A linguist's record*. Evanston, IL: Northwestern University Press.
- Li, F., & Edwards, J. (2006). Contrast and covert contrast in the acquisition of /s/ and /ʃ/ in English and Japanese. Poster presented at 10<sup>th</sup> Conference in Laboratory Phonology, Paris, June.
- Li, F., Edwards, J., & Beckman, M. E. (2007). Spectral measures for sibilant fricatives of English, Japanese, and Mandarin Chinese. *Proceedings of the XVIth International Congress of Phonetic Sciences*, 6–10 August, Saarbruecken, Germany.
- Louko, L. J., & Edwards, M. L. (2001). Issues in collecting and transcribing speech samples. *Topics in Language Disorders*, 21, 1–11.
- Macken, M. A., & Barton, D. (1980). The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. *Journal of Child Language*, 7, 41–74.
- Maxwell, E. M., & Weismer, G. (1982). The contribution of phonological, acoustic, and perceptual techniques to the characterization of a misarticulating child's voice contrast for stops. *Applied Psycholinguistics*, 3, 29–43.

- McCurry, W. H., & Irwin, O. C. (1953). A study of word approximations in the spontaneous speech of infants. *Journal of Speech and Hearing Disorders*, 18, 133–139.
- McGuire, G. L. (2007). Phonetic category learning. Unpublished doctoral dissertation, The Ohio State University, Columbus.
- McLeod, S. (2007). *The international guide to speech acquisition*. Clifton Park, NY: Thomson Delmar Learning.
- Moe, A. J., Hopkins, C. J., & Rush, R. (1982). *The vocabulary of first-grade children*. Springfield, IL: Charles C Thomas.
- Munson, B., Edwards, J., & Beckman, M. E. (2005). Relationships between nonword repetition accuracy and other measures of linguistic development in children with phonological disorders. *Journal of Speech, Language, and Hearing Research*, 48, 61–78.
- Munson, B., Kurtz, B. A., & Windsor, J. (2005). The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, 48, 1033–1047.
- Munson, B., Li, F., Yoneyama, K., Hall, K. C., Beckman, M. E., Edwards, J., & Sunawatari, Y. (2008). Language-specific production and perception of sibilant fricatives in Japanese and English. Paper presented at the 82nd Annual Meeting of the Linguistics Society of America, Chicago, IL, January.
- Nakanishi, Y., Owada, K., & Fujiki, N. (1972). Articulation test and its result. [In Japanese.] In Report of the Research Institute for the Education of Exceptional Children (1–41). Tokyo: Gakugei University.
- Nicolaidis, K., Edwards, J., Beckman, M., & Tserdanelis, G. (2003). Acquisition of obstruents in Greek. Proceedings of the 6th International Conference of Greek Linguistics, Rethymno, Crete.
- Nihon Chōin Gengo Hakasekai [Japan Hearing and Language Professionals Society], & Nihon Onsei Gengo Igakukai [Japan Society of Logopedics and Phoniatrics], editors. (1994). *Kōin kensa* [Articulation test], 1994 edition. Hiroshima: Success Bell Co., Ltd.
- Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier Mental Lexicon: measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report No. 10*, 357–376.
- Paden, E., & Moss, S. (1985). Comparison of three phonological analysis procedures. *Language, Speech, Hearing Services in Schools*, 16, 103–109.
- Pisoni, D., Nusbaum, H., Luce, P., & Slowiaczek, L. (1985). Speech perception, word recognition, and the structure of the lexicon. *Speech Communication*, 4, 75–95.
- Powell, T. W. (2001). Phonetic transcription of disordered speech. *Topics in Language Disorders*, 21, 52–72.
- Pye, C., Ingram, D., & List, H. (1987). A comparison of initial consonant acquisition in English and Quiché. In K. E. Nelson, & A. Van Kleeck (Eds.), *Children's Language*, Vol. 6 (pp. 175–190). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Pye, C., Wilcox, K. A., & Siren, K. A. (1988). Refining transcriptions: the significance of transcriber 'errors'. *Journal of Child Language*, 15, 17–37.
- Sander, E. (1972). When are speech sounds learned? *Journal of Speech and Hearing Research*, 37, 55–63.
- Schellinger, S., Edwards, J., Munson, B., & Beckman, M. E. (2008). Does 'close' count in transcription as well as in horseshoes? Poster presented at the International Child Phonology Conference, West Lafayette, IN, June.
- Scobbie, J. M., Gibbon, F., Hardcastle, W. J., & Fletcher, P. (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. *Papers in Laboratory Phonology*, 5, 194–207.
- Shriberg, L. D., Austin, D., Lewis, B. A., McSweeney, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric: extensions and reliability data. *Journal of Speech, Language, and Hearing Research*, 40, 708–722.
- Shriberg, L. D., Kwiatkowski, J., & Hoffman, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research*, 27, 456–465.
- Sjölander, K., & Beskow, J. (2000). WaveSurfer—an open source speech tool. In B. Yuan, T. Huang, & X. Tang (Eds.), *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 464–467). Beijing.
- Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders*, 55, 779–798.
- So, L. K. H. (1973). *Cantonese Segmental Phonology Test*. Hong Kong: Bradford Publishing Company.
- So, L. K. H., & Dodd, B. (1995). The acquisition of phonology by Cantonese-speaking children. *Journal of Child Language*, 22, 473–495.
- Stoel-Gammon, C. (2001). Transcribing the speech of young children. *Topics in Language Disorders*, 21, 12–21.
- Stoel-Gammon, C., & Stemberger, J. P. (1994). Consonant harmony and phonological specification in child speech. In M. Yavas (Ed.), *First and second language phonology* (pp. 63–80). San Diego, CA: Singular Publishing Group.
- Taine, H. A. (1876). De l'acquisition du langage chez les enfants et dans l'espèce humaine. *Revue philosophique de la France et de l'étranger*, 1, 3–23.

- Templin, M. C. (1947). Spontaneous versus imitated verbalization in testing articulation in pre-school children. *Journal of Speech Disorders*, 12, 293–300.
- Templin, M. (1957). *Certain language skills in children*. Minneapolis: University of Minnesota Press.
- Tsurutani, C. (2004). Acquisition of yo-on (Japanese contracted sounds) in L1 and L2 phonology in Japanese second language acquisition *Journal of Second Language*, 3, 27–48.
- Tyler, A. A. (1995). Durational analysis of stridency errors in children with phonological impairment. *Clinical Linguistics and Phonetics*, 9, 211–228.
- Vihman, M. M., Macken, M. A., Miller, R., Simmons, H., & Miller, J. (1985). From babbling to speech: a re-assessment of the continuity issue. *Language*, 61, 397–445.
- Vodopivec, S. (2004). The influence of phonotactic probability on consonant acquisition. Unpublished B.A. honor's thesis. Department of Speech and Hearing Science, The Ohio State University, Columbus.
- Waterson, N. (1971). Child phonology: a prosodic view. *Journal of Linguistics*, 7, 179–211.
- Yamada, R., & Tohkura, Y. (1990). Perception and production of syllable-initial English /r/ and /l/ by native speakers of Japanese. Proceedings of the First International Conference on Spoken Language Processing (pp. 757–760). Kobe, Japan.
- Yoneyama, K. (2002). Phonological neighborhoods and phonetic similarity in Japanese word recognition. Unpublished doctoral dissertation, The Ohio State University, Columbus.
- Yoneyama, K., Beckman, M. E., & Edwards, J. (2003). Phoneme frequencies and acquisition of lingual stops in Japanese. Unpublished manuscript. The Ohio State University, Columbus. [Results described in Beckman, Yoneyama, & Edwards, 2003.]
- Zamuner, T. S., Gerken, L. A., & Hammond, M. (2004). Phonotactic probabilities in young children's production of coda consonants. *Journal of Child Language*, 31, 515–536.

## Appendix

The results of a number of studies (e.g. Nicolaidis et al., 2003; Yoneyama et al., 2003; Edwards and Beckman, 2008) are discussed in this paper. All of the data presented in Figures 1–4 are from Edwards and Beckman (2008). In this appendix, we briefly describe the methodology that we used for this study so that readers do not have to look up this paper.

For each language, we collected data from ~ 10 2-year-olds and 10 3-year-olds. All children were typically developing, based on parent and teacher report, and had passed a hearing screening. All children were monolingual native speakers of the languages under consideration. Data were collected in Columbus, OH; Thessaloniki, Greece; Tokyo, Japan; and Hong Kong.

The stimuli were digital recordings of familiar real words of each language which were presented along with photographs for the children to repeat. We elicited the obstruents of each language in word-initial position in the vowel contexts /a, e, i, o, u/. These are the only vowels in Greek and Japanese. For English and Cantonese, we collapsed together vowels that have similar coarticulatory effects. For example, in English we included both lax and tense vowels in each vowel category where the tense/lax contrast is relevant (for example, both /i/ and /ɪ/ were included in the /i/ category) and we included all three low back vowels /ɑ, ʌ, ɔ/ in the /a/ category.

For each target CV sequence, we selected three words that we thought young children would be familiar with. However, we could not find appropriate words for all of the permissible CV sequences in each language (e.g. /gi/ in English), so there were some empty cells. Also, not all of the possible CV sequences are permissible across languages (e.g. Cantonese does not allow the vowel /u/ after alveolar consonants). We digitally recorded an adult female native speaker for each language producing all of the stimulus items with a child-directed speech intonation. Each stimulus item was paired with a culturally appropriate colour photograph of the named object, attribute, or action. All pictures were edited to fit on a fixed-size window on a laptop computer screen.

The testing took place in a quiet room at a preschool in the four countries. The pictures and sound files were presented simultaneously to each participant over a laptop with a 14 in. screen using a program written specifically for our purposes. The children were instructed to repeat each word exactly as they heard it. The children's responses were recorded directly onto a CD or a digital audiotape, using a high-quality head-mounted microphone. While the children sometimes produced multiple responses, only the first audible response was used in all of the analyses presented in Figures 1–4.

A native speaker who was also a trained phonetician listened to the response and examined the acoustic waveform for each repetition. The target consonants were coded as *correct* (e.g. /k<sup>h</sup>oŋ<sup>21</sup>/ for Cantonese /k<sup>h</sup>oŋ<sup>21</sup>/ 'poor' or /kek/ for English *cake*), *incorrect* (e.g. /hoŋ<sup>21</sup>/ for Cantonese /k<sup>h</sup>oŋ<sup>21</sup>/ or /tek/ for English *cake*), or *error of phonation type only* (e.g. unaspirated /koŋ<sup>21</sup>/ for Cantonese /k<sup>h</sup>oŋ<sup>21</sup>/ or /gek/ for English *cake*). In all subsequent analyses, we included the completely correct responses in calculating the percentage correct. For each language, a second native speaker, who was also a trained phonetician, blindly re-transcribed 20% of the data (repetitions of two 2-year-olds and two 3-year-olds). Phoneme-by-phoneme inter-transcriber reliability for accuracy was at or above 89% for all four languages (90% for English, 96% for Cantonese, 94% for Greek, and 89% for Japanese). The first native-speaker's transcriptions were used in the case of inter-rater disagreement.

We also calculated CV sequence frequency for all four languages. To do so, we counted the number of times this sequence occurred in word-initial position and divided this number by the total number of words in the database. We then took the log of the ratio, which effectively weights a percentage change at the low-frequency end of the distribution more heavily than the same percentage change at the high-frequency end. To find the number of times that each CV sequence occurred in each language, we used online lexicons. For English, we used the Hoosier Mental Lexicon (HML, Pisoni et al., 1985). For each of Cantonese and Greek, we used word frequencies for comparably large lexicons extracted from newspaper corpora. For Cantonese, the corpus was the Cantonese language portion of the *Segmentation Corpus* (Chan and Tang, 1999). For Greek, the corpus was the ILSP database (Gavrilidou, Labropoulou, Mantzari, and Roussou, 1999), from which we purchased a list of the 20,000 most frequent word-form types along with their associated token frequencies in the newspaper texts. For Japanese, we used a subset of words from the NTT database (Amano and Kondo, 1999). We used the subset of 78,801 words from this list that was also used by Yoneyama (2002) to calculate neighbourhood densities for Japanese.