The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition

Jan Edwards
Dept. of Speech and Hearing Science, Ohio State University
Mary E. Beckman
Dept. of Linguistics, Ohio State University
Benjamin Munson
Dept. of Communication Disorders, University of Minnesota

Please send inquiries to:
Jan Edwards
Dept. of Speech and Hearing Science
Ohio State University
Address until 7/15/03:
13 Al. Fleming St.
Thessaloniki 54642
Greece
Phone: 011 30 2310 839994
Email: edwards.212@osu.edu

Running title: Vocabulary size and phonotactic probability

Abstract

Adults' performance on a variety of tasks suggests that phonological processing of nonwords is grounded in generalizations about sublexical patterns over all known words. A small body of research suggests that children's phonological acquisition is similarly based on generalizations over the lexicon. To test this account, production accuracy and fluency were examined in nonword repetitions by 104 children and 22 adults. Stimuli were 22 pairs of nonwords, in which one contained a low-frequency or unattested two-phoneme sequence while the other contained a high-frequency sequence. For a subset of these nonword pairs, segment durations were measured. The same sound was produced with a longer duration (less fluently) when it appeared in a low-frequency sequence, as compared to a high-frequency sequence. Low-frequency sequences were also repeated with lower accuracy than high-frequency sequences. Moreover, children with smaller vocabularies showed a larger influence of frequency on accuracy than children with larger vocabularies. Taken together, these results provide support for a model of phonological acquisition in which knowledge of sublexical units emerges from generalizations made over lexical items.

Traditional models of grammar posit that phonological knowledge is instantiated in the form of rules or constraints operating on abstract mental representations of words. A fundamental assumption of these models is that the rules and constraints of phonology exist in a module of the grammar separate from the words whose structure they govern (e.g., Halle, 1985). This assumption is difficult to reconcile with a growing body of research which suggests that phonological processing in adult speakers of English is tightly coupled to the phonological structures of the words that they know. In particular, it is sensitive to the relative frequencies with which different sublexical sequences occur in the lexicon. These relative frequencies are often called phonotactic probabilities or transitional probabilities, reflecting the fact that they are usually expressed as the probability that a sequence of sounds will occur in a lexical item.

Sensitivity to phonotactic probability has been demonstrated using a variety of measures of implicit or procedural knowledge. For example, adults are faster to repeat nonwords that contain high-frequency consonant-vowel and vowel-consonant sequences (Vitevich, Luce, Charles-Luce & Kemmer, 1997, Vitevitch & Luce, 1999). Their speeded repetitions of nonwords containing high-frequency sequences also are more accurate, although this effect is not as robust or as consistently replicated across experiments as the effect on response time. Phonotactic probability also influences speech perception in adults. For example, listeners are biased to hearing acoustically ambiguous phonemes as members of high-probability sequences (Pitt & McQueen, 1998). Furthermore, when adults are asked to transcribe nasal-obstruent sequences embedded in nonwords, their transcription errors are more likely to "correct" a low-frequency sequence by writing a phonetically similar but more frequent sequence (Hay, Pierrehumbert, & Beckman, in press). Adults also have a better recognition memory for nonwords containing high-probability sequences of phonemes than for those containing low-probability sequences (Frisch, Large, & Pisoni, 2000).

Sensitivity to phonotactic probability is also reflected in explicit judgments of how well a nonword conforms to the phonological patterns attested in real words. When asked to judge how "wordlike" nonsense words are, adults give higher wordlikeness ratings to forms that contain phoneme sequences which are attested in many words. This result is extremely robust and has been seen in a large number of experiments (e.g., Pierrehumbert, 1994; Coleman & Pierrehumbert, 1997;Vitevitch et al., 1997; Frisch et al., 2000; Munson, 2001). Moreover, it interacts with vocabulary size (Frisch, 2001). Whereas adults with large vocabularies differentiate sequences with different low frequencies by assigning them different (low) wordlikeness ratings, adults with small vocabularies assign the same (lowest) wordlikeness rating to many low-frequency sequences, as if they were all equally unattested in the lexicon. Together, the results of these studies on implicit and explicit phonological knowledge support a view of grammar in which phonological rules or constraints emerge as the language user notices commonalities among the sound shapes of words in the lexicon.

The idea that the lexicon plays a key role in phonological development is not new. Almost thirty years ago, Ferguson and Farwell (1975) proposed that, "a phonic core of remembered lexical items and the articulations that produced them is the foundation of an individual's phonology, …even though it may be heavily overlaid or even replaced by phonologically organized acquisition processes in later stages " (p. 36). However, while there is a large body of research on adults' sensitivity to lexical factors in both perception and production, there is relatively little comparable research on young children. The few studies that have been done suggest that children, as well as adults, are sensitive to phonotactic probability. For example, Storkel (2001) found that three- to six-year-old children learned new words more rapidly when the words contained high probability sequences, as compared to low probability sequences. Neighborhood density is another measure of sequence frequency in the lexicon. (A word like *side* has a high neighborhood density because there are many words which differ from it by a single phoneme addition, substitution, or omission; the reverse is true for a word with a low neighborhood density such as *shine* [Pisoni, Nusbaum, Luce, & Slowicacek, 1985].) Several studies have shown that neighborhood density influences word recognition in children, although these effects are somewhat different from those seen in adults (Garlock, Walley, & Metsala, 2001; Metsala, 1997; Storkel, 2002).

Some indirect evidence suggests that children are also sensitive to phonotactic probability in production tasks. As discussed above, many researchers have found that adults give higher wordlikeness ratings to nonwords that contain high-frequency sequences (Pierrehumbert, 1994; Coleman & Pierrehumbert, 1997;Vitevitch, et al., 1997; Frisch et al., 2000). In addition, Gathercole, Willis, Emslie, and Baddeley (1991) found that four-, five-, and six-year-old children are more accurate at repeating nonwords that adults had judged to be more wordlike. These two findings – the finding that adults judge nonsense words with high-frequency sequences as more wordlike and the finding that children repeat nonwords with higher wordlikeness ratings more accurately – together suggest that phonotactic probability directly influences children's repetitions of nonwords.

A few recent studies, which have systematically controlled phonotactic probability in their nonword stimuli, have found this to be the case. Gathercole, Frankish, Pickering, and Peaker (1999) found that seven- and eight-year-old children repeat lists of nonwords more accurately in a serial recall task if the nonwords contain only high frequency consonant-vowel and vowel-consonant sequences. Using a less demanding immediate repetition task, Beckman and Edwards (2000a) found that children three to five years of age repeated low-frequency two phoneme sequences in nonwords less accurately than they repeated high-frequency two-phoneme sequences. Munson (2001) found an influence of phonotactic probability on production fluency as well as on accuracy. He used segment duration as a measure of fluency and found that children from three to eight years of age produced longer durations for the same segment when it was in a low-frequency consonant-consonant sequence, as compared to a high-frequency sequence.

In this paper, we continue to explore the influence of sublexical sequence frequency on production accuracy and fluency in children. A second focus of the paper is the relationship between the effect of sublexical sequence frequency and estimates of the child's vocabulary size. Specifically, we wanted to determine whether this effect of frequency, if observed, was mediated by vocabulary size. Gathercole et al. (1999) found an effect of vocabulary size on accuracy overall, but no interaction of high versus low vocabulary scores with high versus low transitional probabilities. However, the claim that children acquire a phonological system based on generalizations over the lexicon predicts that children with larger lexicons should have more robustly generalized phonological systems. Their representations of familiar sublexical patterns can be more quickly accessed and more flexibly reapplied to less familiar but analogous patterns. Children with smaller vocabularies, conversely, will know fewer words that exemplify any particular sequence in a variety of larger contexts as well as fewer words that exemplify the component segments in a variety of more or less similar sequences. Smaller vocabularies thus provide less support for abstracting knowledge about the acoustics and articulation of consonants and vowels away from the specific contexts in which they have been encountered. Representations of familiar sublexical patterns are more fragile, and cannot be reapplied as flexibly to form production routines for less familiar but analogous patterns. This effect might be particularly evident in younger children, where the same absolute difference in vocabulary size means a proportionally larger difference in experience — i.e., a proportionally larger difference in the support for a robust representation of the individual phonological components independent of specific contexts. This predicts that the effect of low transitional probability on a simpler repetition task might be especially pronounced in children with small vocabularies.

We tested these hypotheses using a nonword repetition task to measure production accuracy and fluency, and two standard clinical tests to estimate vocabulary size. Our approach differs from most previous research on children's nonword repetition accuracy in two respects. First, we systematically controlled the phonotactic probability of the sublexical sequences within the nonword stimuli, by matching each high-frequency sequence with a minimally different low-frequency sequence. Second, we measured both accuracy and fluency of production. This research also differs from our own previous work in that we tested a much larger group of children with a substantially larger set of stimuli. We found systematic effects of transitional probability on repetition accuracy and fluency, and a relationship between the accuracy effect and the size of the children's vocabularies.

Methods

Stimuli

An important concern with the three stimulus sets used in our earlier studies was that they were small — only six item pairs in each of the stimulus sets in Beckman and Edwards (2000a) and only eight pairs in Munson (2001). Therefore, we devised a new stimulus set that was designed to test a much larger range of segment types in several different syllable and word positions as well as a good range of transitional probabilities. In order not to make the stimulus set too large for the attention spans of our youngest participants, we kept the design of our earlier studies in which stimulus items are paired. One member of each nonword pair contained a low-frequency target that occurred in few or no words that would likely be familiar to children and the other member of the nonword pair contained a high-frequency target that occurred in many words familiar to children. The two sequences were placed in identical positions within similar nonwords. The final expanded set contained 22 nonword pairs, half of them disyllabic and half trisyllabic, with seven nonword pairs containing target CV sequences contrasting in low versus high transitional probability, seven pairs containing target VC sequences, and eight pairs containing CC sequences, with the last including word-initial onset clusters and word-final coda clusters as well as word-medial heterosyllabic clusters. The stimuli are listed in Table 1, along with wordlikeness ratings and two measures of the phonotactic likelihood of the target sequences.

Insert Table 1 about here

The sequences were developed using the MHR database, an on-line list of pronunciations of the 6366 most frequently occurring words in the spontaneous continuous speech of first grade children. This database was created by making an electronic version of the word list resulting from Moe, Hopkins, and Rush's (1982) study, and then extracting phonetic transcriptions for the words from the Carnegie Mellon University Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict), which gives pronunciations from the same general dialect region as the central Ohio varieties spoken by the children in our study. Each low-probability sequence occurred in either none or very few words in this database, while each high-probability sequence occurred in many words in this database. For example, one CC sequence pair was /ft/ and /fk/. The medial cluster /ft/ occurs in many words, such as *after*, *fifteen*, and *safety*, while /fk/ does not occur in any words at all. Sequences were then embedded in nonwords. For the two nonwords for each sequence pair, the sequence was placed in the same prosodic position in the two nonwords and the transitional probability of all other phoneme sequences within the two nonwords was matched as closely as possible.

We calculated the transitional probabilities of the target sequences based on the frequency of the segmental sequence in the target syllable position, adjusted by a factor representing the frequency of the sequence type. The adjustment factor was intended to capture the effect of prosodic context. That is, since phonological acquisition involves developing representations for prosodic structure as well as for the segments that can fill different prosodic positions, frequency of the sequence type should contribute to accuracy of a two-phoneme sequence independently of the frequency of the sequence itself. For instance, just as heterosyllabic /ft/ and /fk/ contrast in occurring in many versus no words, syllable-initial /ju/ and /jau/ contrast absolutely. The familiar sequence /ju/ occurs in many words such as *you*, *use* and *uniform*, whereas the novel sequence /jau/ occurs in no words at all. However, most English words contain at least one syllable-initial CV sequence, whereas heterosyllabic CC sequences are relatively more rare. For one thing, they cannot occur in monosyllabic forms. Thus, although /jau/ is no more frequent as a sequence than /fk/, it might be "easier" simply because CV sequences are more frequent than CC sequences. Therefore, the transitional probability of each sequence included two terms. For the first term, we counted the number of instances in which a target sequence occurred in the relevant syllable position (i.e., syllable-initial for CV; syllable-final for VC; and onset, medial heterosyllabic, or coda position for the different types of CC sequences), and divided this frequency count by the total number of two-phoneme sequences in all of the words in the MHR to get the raw transitional probability. For the second term, we counted the number of instances of the sequence type (e.g., the number of heterosyllabic CCs for sequences like /ft/ and /fk/), and divided that by the same denominator. The adjusted transitional probability was then the raw transitional probability of the two-phoneme sequence multiplied by the probability of the sequence type. As in other studies of the effects of frequency, we took the natural logarithm of this adjusted transitional probability. For sequences with a frequency of zero, we substituted a count of 0.5 for the numerator in the first term (the raw transitional probability of the sequence), since the natural log of 0 is undefined.

We calculated transitional probabilities first by counting occurrences in the MHR database for children, which was our source for the development of the low- and high-frequency sequences. We also calculated the transitional probabilities a second time, based on the Hoosier Mental Lexicon (HML, Pisoni et al., 1985), an on-line 19,000 word database that many researchers have used to compute transitional probability (e.g., Vitevitch, et al., 1997). We decided to include transitional probability counts based on the HML because we were concerned that the MHR database might underestimate children's productive vocabulary. Recall that the MHR database is a list of the 6000 most frequently occurring words in the speech of first grade children. The frequencies are based on number of occurrences in a corpus of 285,623 word tokens taken from spontaneous speech that includes both free-topic conversations between peers and more structured narratives elicited using prompts such as "Tell me about your favorite TV show." This database probably underestimates the expressive vocabulary of many 6-year-old children and necessarily underestimates that of older children and adults.

The frequency relationships in the HML were in accord with those in the MHR. Although many sequences that did not occur in any words in the MHR did occur in one or more words in the HML, paired comparison t-tests revealed that transitional probabilities were significantly different between the two sequences of each nonword pair in the HML, just as they were in the MHR (t[21] = 24.45, p < .001 for MHR; t[21] = 14.04, p < .001 for HML).

These sequences were embedded in larger "frames" that were matched in relevant aspects for each pair of low- and high-frequency targets. In particular, the frames for any pair were identical in prosodic structure and very similar in segmental content. We did not use segmentally identical frames because our previous studies showed that this induced a practice effect. Instead we controlled for any effect of the segments in the frame by matching for wordlikeness. We did this by creating a larger list of candidate nonwords for each pair and then choosing the final frames on the basis of a wordlikeness rating study.

Sixteen adults were presented with the larger list of nonwords over headphones in a sound-treated booth and were instructed to rate the nonwords on a 5-point scale, with 1 corresponding to "very unlike a real word" and 5 corresponding to "very like a real word." Five randomized blocks of the nonwords were presented to each adult. Insofar as possible, the final 44 nonwords were selected to minimize differences in wordlikeness ratings across the two members of each nonword pair. Analysis of the results showed that the participants used the entire scale. Moreover, ratings were fairly consistent from one block to the next, with no difference between the rating by any subject for first versus the last presentation of any word in 36% of cases and a difference of only one point on the 5-point scale in 38% of the cases. The wordlikeness ratings in Table 1, therefore, are the mean ratings averaged over all five trials for all subjects. Also, the difference between each subject's ratings for matched pairs on any trial clustered around 0, with no difference in 34% of the blocks and a difference of only one point in either direction in 39% of the blocks. Thus, we were fairly successful in controlling for wordlikeness across the two members of each pair. Nonetheless, the nonwords containing high-frequency sequences were judged on average to be slightly more wordlike than the paired nonwords containing low-frequency sequences. (Means are 2.98 for nonwords with high-frequency targets versus 2.65 for those with low-frequency targets, $t[21] = 2.07$, $p =.02$). Given that our purpose is to contrast transitional probabilities at the target sequence itself, we would expect some difference in wordlikeness.

The one remaining question then is whether this difference in mean wordlikeness rating is due to the contrasting transitional probabilities at the target sequence or to the uncontrolled difference in total transitional probability of the frame. Regression analyses showed the mean wordlikeness rating to be significantly correlated with the total transitional probability of the frame ($R^2 = .274$, $F[1,42] = 15.82$, $p < .001$), but not with the target sequence transitional probability, calculated from either the HML or from the MHR. These results replicate the analyses of Frisch et al. (2000), who showed that whole-form measures of goodness, such as the total log probability of all sequences in the nonword, were better predictors of wordlikeness than local measures of constraint violation, such as the transitional probability of the least likely sequence (which the target sequence is in the case of the low-frequency sequences in our stimulus set). At the same time, these results suggest that we need to be careful to correlate our accuracy results with wordlikeness, since Gathercole et al. (1991) found that young children are more accurate at repeating nonwords that adults have judged to be more wordlike in a test much like the one we report here.

Participants

The participants were 104 typically developing children ranging in age from 3;2 to 8;10 years;months and 22 young adults ranging in age from 21 to 34 years. All participants were part of a larger study on phonological knowledge deficits in phonological disorder and were monolingual speakers of English. Each of the 104 children met the following four criteria for typical development: (1) normal articulatory development, as evidenced by a score no more than one standard deviation below the mean on the *Goldman-Fristoe Test of Articulation* (GFTA, Goldman & Fristoe, 1986); (2) normal hearing, as evidenced by passing a hearing screening at 20 dB at 500, 1000, 2000, and 4000 Hz; (3) normal structure and function of the peripheral speech mechanism, as evidenced by a standard score no more than one standard deviation below the mean on the oral movement subtest of the *Kaufman Speech Praxis Test for Children* (KSPT, Kaufman, 1995); (4) normal non-verbal IQ, as evidenced by a standard score no more than one standard deviation below the mean on the *Columbia Mental Maturity Scale* (CMMS, Burgemeister, Blum, & Lorge, 1972). Each of the adult participants also passed a hearing screening, and had no reported history of speech, language, or hearing problems. Table 2 provides descriptive information for the different participant groups. The last two rows of the table report standard scores on measures of expressive and receptive vocabulary that were administered to all participants.

Insert Table 2 about here

Procedure

Three pseudo-randomized lists of the stimuli were created. For each list, all two-syllable words were presented before the three-syllable words, the two members of a nonword pair were always separated by at least two words, and an equal number of words containing high-frequency sequences were presented before their paired words containing low-frequency sequences as vice versa. The nonwords were played to the participants over two external speakers. The participants were instructed to repeat the nonwords as accurately as possible. Training prior to the experiment consisted of two practice words presented by live voice and then two additional digitized practice words presented over the speakers. Training with digitized practice word pairs then continued until the participant understood the task and repeated the two digitized practice words accurately. (No more than four practice trials with digitized practice word pairs was needed with any of the participants.) The participants' repetitions were recorded with a head-mounted microphone connected to a digital audio tape recorder.

Analysis

Transcription. As a first step in coding the responses for accuracy, the recording for each participant was transferred from the DAT to a digital file on a computer and the participant's responses were transcribed in the International Phonetic Alphabet at the level of a careful, broad phonemic transcription. That is, transcription was not done directly from the DAT, but using a waveform editor so that each nonword could be played as often as necessary without rewinding the tape and spectrograms could be examined in cases of doubt. All of the responses were transcribed by a single transcriber. A second transcriber independently transcribed 10 percent of the data, comprising all responses by four participants from the three-to-four-year-old group, four participants from the five-to-six-year-old group, three participants from the seven-to-eight-year-old group, and two adults. Phoneme-by-phoneme inter-rater reliability ranged from 86 to 99 percent for data from individual participants, with a mean of 94 percent across these 13 participants.

Coding. In coding responses on repetition tasks, researchers often use rather coarse-grained measures of accuracy, such as the number of tokens repeated without error in a string of seven repetitions of the target nonword (e.g., Gathercole et al., 1991) or the proportion of phonemes repeated accurately in the target sequence or syllable within the nonword (e.g., Beckman & Edwards, 2000a; Dollaghan, Biber, & Campbell, 1995; Fisher, Hunt, & Chambers, 2001; Munson, 2001). When coding responses from young children, such coarse-grained measures have several disadvantages. First, they do not distinguish between errors related to experimental conditions and "ordinary" mispronunciations that a very young child might make, such as the substitution of [θ] for /s/ or [d] for /g/. Second, they do not distinguish between small subtle errors such as the place feature substitution that perceptually "corrects" /mt/ to /nt/, and more drastic errors such as the deletion of the /t/ in the /mt/ cluster so that the /m/ is resyllabified as the onset of the following syllable. Our study covered an extremely large age range, and the larger study included children with phonological disorder, who have habitual age-inappropriate mispronunciations. Since the severity and type of error might be more informative of the nature of phonological generalization than the gross error rate, we decided to code the transcriptions using a finer-grained segmental accuracy score.

For this segmental accuracy score, each of the two phonemes in a target sequence was scored for accuracy on each of three features. For consonants, one point was awarded for correct place (labial, alveolar, or velar); one point was awarded for correct manner (stop, fricative, or glide); and one point was awarded for correct voicing (voiced or voiceless). For example, if the /k/ in the /kt/ sequence was produced as /s/, it would receive one point for correct voicing, but would lose two points, one for incorrect place and one for incorrect manner. For vowels, one point was awarded for correct production on the dimension front-back (front, central, or back), one point was awarded for correct vowel height (high, mid, or low), and one point was awarded for correct "length" (i.e., tense or lax for a monophthong target and monophthong or diphthong for a diphthong target). For example, an /u/ for /i/ substitution would receive two points, one for correct tenseness and one for correct height, but would lose one point for being a back rather than a front vowel. Thus, the maximum segmental accuracy score for any target sequence was six points, and the minimum score was 0.

Segment duration. We were also interested in whether fluency of production is related to sublexical sequence frequency. Following Munson (2001), we used segment duration as our measure of production fluency since duration is an acoustic measure of the speed with which a speech movement is executed. All other factors being equal, shorter segment durations should indicate greater fluency than longer durations. Duration measurements could be made for 9 of the 22 nonword pairs. These were pairs where the same sound occurred in the target sequence of both members of a nonword pair, and this sound (or this sound and an identical neighboring non-target phoneme) could be isolated on the waveform. The nonword pairs for which duration measurements could be made are indicated by listing the measured phoneme(s) in Table 1. Measurements were made from the waveform using conventional criteria for determining the onset and offset of each sound. Duration measurements were made only for productions that had completely correct segmental accuracy scores. Because of this restriction, the number of tokens per utterance type was not constant across types. Therefore, an utterance token was included in the statistical analysis only when the matched utterance token produced by the same participant also could be included.

Vocabulary size measures. Finally, we wanted to know whether differences in accuracy effects between younger and older participants reflect differences in typical vocabulary size across ages, as suggested above, or are due to some process of typical phonological development that is independent of vocabulary growth. To explore these two possibilities, we used two standardized tests to estimate vocabulary size. For receptive vocabulary size, the *Peabody Picture Vocabulary Test-III* (PPVT-III, Dunn & Dunn, 1997) was administered. This widely used measure was most recently revised and renormed in 1997 and this most recent version has been shown to be much less culturally biased than previous versions (Washington & Craig, 1999). We used the *Expressive Vocabulary Test* (EVT, Williams, 1997) to measure expressive vocabulary size. These two tests were co-normed for participants aged 2 through 90. It can be observed in Table 2 that, overall, the participants have larger than average vocabularies for their ages. Also, the four age groups are well matched for standard scores on the test of receptive vocabulary, but less so for the test of expressive vocabulary. A one-way ANOVA showed a significant effect of age on the EVT standard scores ($F[3, 122] = 10.69, p < .001, \eta^2 = .21$), with adults having significantly higher scores than any of the groups of children, and the 7-8 year olds having significantly lower scores than 3-4 year olds. In our analyses of the nonword repetitions, we used these scores as an independent variable in various regression analysis of each participant's mean segmental accuracy for high- versus low-frequency target sequences. Because the relationship between vocabulary size and age is exponential (that is, vocabulary growth levels off as age continues to increase), we used the natural log of the raw vocabulary scores in all analyses.

Results

Segmental accuracy scores by item

Accuracy scores were averaged over the 126 participants for each of the target sequences. A paired-comparison t-test on these scores for the 22 nonword pairs revealed a significant effect of frequency on accuracy ($t[21] = 2.89, p = .009$). That is, accuracy scores were significantly higher for the target sequences with high transitional probabilities, as compared to the sequences with low transitional probabilities ($M = 5.46$, $SD = .38$ for high-frequency sequences, $M = 5.16$, $SD = .45$ for low-frequency sequences). The difference between the two sequence types was somewhat more pronounced when the accuracy scores for the adults were not included in the analysis ($t[21] = 3.10, p = .005$, with $M = 5.34$, $SD = .38$, for high-frequency sequences, $M = 5.03$, $SD = .50$ for low-frequency sequences).

Figure 1 shows mean accuracy scores plotted against transitional probability based on each of the two databases, with the three sequence types (CV, VC, CC) represented by different symbols. The overall trend is for accuracy to be greater for sequences with higher transitional probabilities. Note also that the CV sequences are generally more accurate than would be predicted by transitional probability alone. This was so even though the transitional probabilities were adjusted to reflect the greater probability of the CV sequence type. There are also two outliers in these graphs, the low-frequency sequence /auk/ and the high-frequency sequence /aun/, both of which have lower accuracy scores than would be predicted by their transitional probabilities.

Insert Figure 1 about here

In order to determine whether this effect of transitional probability could be explained by the differences in wordlikeness due to the frame rather than by the transitional probabilities themselves, we correlated the mean accuracy scores for the target sequences with each of the three stimulus properties listed in Table 1. That is, we correlated mean accuracy of the sequences with their transitional probabilities as measured in the child-sized MHR database and in the adult-sized HML database, and we correlated the mean accuracy of the sequences with the mean wordlikeness scores of the nonwords in which they were embedded. Accuracy was significantly correlated with both measures of the target sequence probability ($r^2 = .18$, $p = .004$ for MHR, and $r^2 = .19$, $p = .003$ for HML), but not with wordlikeness scores ($r^2 = .07$, $p = .09$).

Segmental accuracy scores by subject

Figure 2 shows mean accuracy scores for the high frequency and low-frequency sequences for the four age groups. A two-way (frequency by age-group) repeated measures ANOVA showed a significant main effect of frequency (F[1,122] = 128.30, p< .001, $\eta^2 = .51$), a significant main effect of age group (F[3,122] = 23.30, $p < .001$, $\eta^2 = .36$), and a significant frequency by age-group interaction (F[3,122] = 6.56, $p < .001$, $\eta^2 = .14$). The interaction was due to the larger difference between low- and high-frequency sequences for the three groups of children, as compared to the adults. That is, post hoc tests of simple main effects found a significant main effect of sequence frequency for all four age groups. Measures of effect size, however, showed that target sequence frequency affected the segmental accuracy scores for adult repetitions less than it affected segmental accuracy for any of the three groups of children.

Insert Figure 2 about here

Duration analysis.

Since different segments have different inherent durations, the duration value for a particular segment produced by a particular participant was included in the analysis only if *both* the low-frequency and matched high-frequency target sequence containing the measured segment was produced correctly. For the younger age groups, therefore, this analysis necessarily over-represents productions by those participants who behaved more like older participants in terms of error rates. Table 3 shows the number of token pairs and the mean durations of each segment type in the low- versus high-frequency sequence for each age group. A segment in a low-probability sequence is generally longer than in a high-probability sequence. This tendency is more consistent for the younger groups and not evident in the means for the adults.

Insert Table 3 about here

The literature on segment durations in English suggests that nasals and voiced obstruents are inherently shorter than voiceless obstruents, which in turn should be shorter than sequences of two voiceless consonants or the two vowel targets of a diphthong. We therefore grouped [m], [n], [v] and [g] together as “short” segments and [pt] and [au] together as “double” segments, in a three-way ANOVA with factors segment type, age group, and sequence probability. As expected, there was a significant main effect of segment type (F[7,711] = 127.608, $p < .001$). There were also significant main effects of age group (F[1,711] = 4.701, $p = .03$) and sequence frequency (F[1,711] = 10.229, $p = .001$), as well as a significant age by frequency interaction (F[1,711] = 5.807, $p = .016$). The age by frequency interaction is due to the fact that duration generally decreases with age for the low-frequency sequences, but remains fairly constant for the high-frequency sequences, resulting in no difference in duration between the low- and high-frequency sequences for the adults.

Segmental accuracy scores and vocabulary size

Figure 3 shows mean accuracy scores for high- and low-frequency sequences plotted against the two measures of vocabulary size. For both plots, the regression line for the high-frequency sequences lies above the line for the low-frequency sequences, and the distance between the two lines is the effect of target sequence frequency. This distance decreases as vocabulary size increases. The participants with the largest vocabularies are the adults, for whom there is the smallest effect of frequency on accuracy, as indicated by the significant age by frequency interaction observed in the repeated measures ANOVA. To determine the quantitative relationship between vocabulary size and repetition accuracy more precisely, we correlated mean segmental accuracy scores for the low- and high-frequency sequences with our two measures of vocabulary size. These correlations were significant and were greater for low-frequency sequences, as compared to high-frequency sequences (for low-frequency sequences, $r^2 = .38$, $p < .001$ for PPVT-III, and $r^2 = .38$, $p < .001$ for EVT; for high-frequency sequences, $r^2 = .26$, $p < .001$ for PPVT-III, and $r^2 = .25$, $p < .001$ for EVT). When the adults were excluded from the analysis, the correlations were somewhat smaller, but still significant (for low-frequency sequences, $r^2 = .25$, $p < .001$ for PPVT-III, and $r^2 = .30$, $p < .001$ for EVT; for high-frequency sequences, $r^2 = .18$, $p < .001$ for PPVT-III, and $r^2 = .21$, $p < .001$ for EVT).

Insert Figure 3 about here

Accuracy is correlated both with vocabulary size and age. Furthermore, vocabulary size and age are highly correlated with each other. To tease apart the influence of these two factors, we performed two stepwise multiple regression analyses. This analysis is similar to ones in previous research examining the relative effects of age and vocabulary size on morphosyntactic development (e.g., Bates & Goodman, 1999). In both analyses, the independent variables were age, the natural log of the EVT raw score, and the natural log of the PPVT-III raw score, but the dependent variable differed. In the first analysis, it was mean segmental accuracy averaged across all items for each participant, and in the second it was the mean difference between the segmental accuracy scores for the high- versus low-frequency targets averaged across all item pairs. When the dependent variable was overall accuracy, the only significant predictor was PPVT-III raw score, accounting for 31 percent of the variance. When the dependent variable was the difference in accuracy between the high- versus low-probability sequences, the only significant predictor was EVT raw score, accounting for 17 percent of the variance. These analyses were performed a second time excluding the adult participants, with the same results. When the dependent variable was overall accuracy, the only significant predictor was again the measure of receptive vocabulary size, which accounted for 19 percent of the variance, whereas when the dependent variable was the mean difference in accuracy, the only significant predictor was the measure of expressive vocabulary size, accounting for 8 percent of the variance. The results of these regression analyses suggest that it is vocabulary size, rather than age *per se*, that accounts for the higher accuracy and the smaller effect of transitional probability on accuracy for older children and adults.

Discussion

We found that participants repeated consonants and vowels more accurately in the context of target sequences that occur in many real words. In pairs of productions of sequences containing an identical measurable phoneme segment where both the low- and the high-frequency sequence were produced accurately, participants also produced shorter durations in the high-frequency sequences. These effects of target sequence frequency on segmental accuracy and fluency were largest in productions by three- to four-year-old children and smallest in productions by adults. Given how much closer the young child is to the onset of lexical acquisition, it is not surprising that the child's representations of speech sounds are even more highly tied to the contexts in which these sounds occur in words in the lexicon. When the young child encounters a new word with a low-frequency sublexical pattern, there are fewer words in the lexicon that can be used by analogy to aid in the creation of acoustic and articulatory representations for the new word. This increased difficulty makes production of a new word less accurate and less fluent when it contains an infrequent phoneme, or a relatively frequent phoneme in an unfamiliar context.

An analysis of the relationships among target sequence frequency, age, and vocabulary size showed that the effect of frequency on segmental accuracy is related to the massive vocabulary growth that normally occurs during early childhood rather than to some other aspect of normal maturation that is independent of vocabulary size. These results support an account of acquisition in which the typically developing child gradually develops more and more robust phonological knowledge as a consequence of acquiring many words. More generally, these results support the view that symbolic knowledge at all levels of phonology emerges from each individual speaker's experience in acquiring and using words of the ambient language. In the mature language user, phonotactic constraints are patterns generalized over known words, which help the adult speaker pick out familiar words in connected speech and to recognize and remember new words. Similarly, at a younger age, phonemes, syllables, and the other symbolic structures specific to phonology emerge through interaction between the input forms that the child hears and the increasingly more complex hierarchy of representations that the child builds in order to recognize and produce words in connected speech.

Our results support a particular view of the relationship between grammatical knowledge and processing skills in general. Knowledge of more word forms is associated with more robustly generalized knowledge of how to learn to hear and say new word forms. This is consistent with a view of grammar as an emergent property of the history of interactions between the language user and the language events in the world (see, e.g., Allen & Seidenberg, 1999; Bates & Goodman, 1999; Beckman & Edwards, 2000b; Pierrehumbert, 2001; Werker, Corcoran, Fennell, & Stager, 2002). In this view, the relationship between knowledge of the phonological grammar and processing of phonological patterns is a symbiotic one. Knowledge feeds on processing, and processing feeds on knowledge. The more often a child has heard and said a word, the better the child knows the word. The child can fluently incorporate the word into unfamiliar prosodic structures in productions of novel sentences. In the same way, the more words the child has heard and said that contain a particular phonological pattern, the more basis the child has for abstracting away a generalized knowledge of the possible patterns, to quickly access the same or similar patterns in other words. As the child gains more experience with more words, and more specific instances of a pattern accumulate, fine-grained phonological knowledge becomes richer. At the same time, aspects of speech production and perception that are shared across sets of similar subparts of words and that contrast in analogous ways to subparts of other sets of words, can become practiced as a relational pattern at another higher level of representation. If we recast Ferguson and Farwell's (1975) idea of a "lexical core" in this view, it is not so much that a "pre-grammatical" foundation of knowledge of how to produce a small core of words is overlaid by phonological knowledge. Rather, phonological knowledge incrementally emerges from the initial layer of first-learned words to build an increasingly structured scaffolding, an increasingly rich set of alternative paths to hearing and reproducing a novel word-form.

Acknowledgements

This work was supported by NIDCD grant 02932 to Jan Edwards and by NIH training grant T32 DC0051 to Robert A. Fox. We thank the children who participated in the study, the parents who gave their consent, and the schools at which the data were collected. For assistance in stimuli preparation, data collection, and analysis, we thank Erin Casey, Lynn Carahaly, Lisa Draper, Melissa Epstein, Heidi Hochstetler, Maryann Holtschulte, Bridgett Isermann, Satako Katagiri, Laurie Vasicek, Amy Vitale, Pauline Welby, and David White.

References

Allen, E., & Seidenberg, M. S. (1999). On the emergence of grammar from the lexicon. In B. MacWhinney (Ed.), *The emergence of language* (pp. 115-151). Mahwah, NJ: Lawrence Erlbaum Associates.

Bates, J., & Goodman, J. C. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *The emergence of language* (pp. 29-79). Mahwah, NJ: Lawrence Erlbaum Associates.

Beckman, M. E., & Edwards, J. (2000a). Lexical frequency effects on young children's imitative productions. In M. Broe & J. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V* (pp. 207-217). Cambridge, UK: Cambridge University Press.

Beckman, M. E., & Edwards, J. (2000b). The ontongeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development, 71*, 240-249.

Burgemeister, B. B., Blum, L. H., & Lorge, I. (1972). *Columbia Mental Maturity Scale*. New York: Harcourt Brace Jovanovich, Inc.

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. *Proceedings of the 3rd Meeting of the ACL Special Interest Group in Computational Phonology*, pp. 49-56. Somerset, NJ: Assoc. for Computational Linguistics.

Dollaghan, C. A, Biber, M. E., & Campbell, T. F. (1995). Lexical influences on nonword repetition. *Applied Psycholinguistics*, *16*, 211-222.

Dunn, L. & Dunn, L. (1997). *Peabody Picture Vocabulary Test – III.* Circle Pines, MN: American Guidance Services.

Ferguson, C. A., & Farwell, C. B. (1975). Words and sound in early language acquisition; English initial consonants in the first fifty words. *Language*, *51*, 419-39.

Fisher, C., Hunt, C., & Chambers, K. (2001). Abstraction and specificity in preschoolers' representations of novel spoken words. *Journal of Memory and Language**, 45***, 667-687.

Frisch, S. (2001). Emergent phonotactic generalizations in English and Arabic. In J. Bybee & P. Hopper (Eds.) *Frequency and the emergence of linguistic structure* (pp. 159-179). Amsterdam: John Benjamins.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on processing of non-word sound patterns. *Journal of Memory and Language, 42*, 481-496.

Garlock, V., Walley, A., & Metsala, J. (2001). Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language, 45*, 468-492.

Gathercole, S. E., Frankish, C. R., Pickering, S., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 84-95.

Gathercole, S. E., Willis, C., Emslie, H, & Baddeley, A. D. (1991). The influences of number of syllables and wordlikeness on children's repetition of nonwords. *Applied Psycholinguistics*, *12*, 349-367.

Goldman, R. & Fristoe, M. (1986). *The Goldman Fristoe Test of Articulation*. Circle Pines, MN: American Guidance Services.

Halle, M. (1985). Speculations on the representation of words in memeory. In V. Fromkin,. (Ed.). *Phonetic Linguistics*. Orlando: Academic Press.

Hay, J., Pierrehumbert, J., & Beckman, M. (in press). Speech perception, well-formedness, and the statistics of the lexicon. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in Laboratory Phonology VI*. Cambridge, UK: Cambridge University Press.

Kaufman, N. (1995). *Kaufman Speech Praxis Test for Children*. Detroit, MI: Wayne State University Press.

Metsala, J. (1997). An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory and Cognition, 23*, 47-56.

Moe, S., Hopkins, M., & Rush, L. (1982). *A vocabulary of first-grade children*. Springfield, IL: Thomas.

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research, 44*, 778-792.

Pierrehumbert, J. (1994). Syllable structure and word structure: A study of triconsonantal clusters in English. In P. A. Keating (Ed.) *Papers in Laboratory Phonology III* (pp. 168-190). Cambridge, UK: Cambridge University Press.

Pierrehumbert, J. (2001) Why phonological constraints are so coarse-grained. *Language and Cognitive Processes*, ***16***, 691-698.

Pisoni, D., Nusbaum, H., Luce, P., & Slowiacek, L. (1985). Speech perception, word recognition, and the structure of the lexicon. *Speech Communication*, *4,* 75-95.

Pitt, M. & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39*, 347-370.

Storkel, H. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research*, *44*, 1321-1338.

Storkel, H. (2002). Restructuring of similarity neighborhoods in the developing mental lexicon. *Journal of Child Language, 29*, 251-274.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*, 374-408.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, *40*, 47-62.

Washington, J. A. & Craig, H. K. (1999). Performances of at-risk, African-American preschoolers on the Peabody Picture Vocabulary Test-III. *Language, Speech, and Hearing Services in Schools*, *30*, 75-82.

Werker, J. Corcoran, K. M., Fennell, C. T. & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, ***3***, 1-30.

Williams, K. (1997). *Expressive Vocabulary Test*. Circle Pines, MN: American Guidance Services.

Table 1. Nonword pairs, with the low- versus high-frequency target sequences underlined. The third column lists segments from pairs for which we measured the duration of one or both target phonemes, and subsequent column pairs show mean wordlikeness rating (on a scale from 1 to 5) and log transitional probabilities for the embedded target sequences calculated from the MHR database and from the HML database.

Phonetic form		Seg.	Wordlikeness		MHR		HML	
Low freq.	High freq.		Low	High	Low	High	Low	High
/jugoin/	/bogib/		3.06	3.30	-12.42	-9.71	-12.92	-10.84
/moipəd/	/mæbɛp/	[m]	2.96	2.76	-13.11	-8.09	-12.00	-7.81
/vugim/	/vɪdæg/	[v]	3.19	2.91	-13.11	-8.73	-12.92	-8.53
/bodəjau/	/medəju/		2.35	2.96	-13.11	-8.37	-14.30	-7.56
/vukɑtɛm/	/vɪtəgɑp/	[v]	2.96	2.65	-13.11	-8.73	-12.92	-8.53
/gaunəpek/	/gitəmok/		2.78	2.64	-12.42	-9.71	-11.82	-10.84
/nʊbəmən/	/nɪdəbɪp/	[n]	1.68	1.88	-13.11	-8.26	-10.84	-7.79
/motauk/	/petik/		3.38	3.50	-13.31	-9.48	-14.59	-9.77
/donug/	/bedæg/		3.08	3.50	-13.31	-9.79	-14.59	-9.62
/tedaum/	/podaud/		2.90	3.11	-13.31	-10.67	-14.59	-11.81
/auptəd/	/iptən/	[pt]	3.79	3.60	-13.31	-9.68	-14.59	-10.67
/dugnəted/	/tʌgnədit/	[g]	2.68	3.03	-13.31	-9.98	-14.59	-10.53
/aukpəde/	/ikbəni/		2.41	2.06	-13.31	-9.48	-14.59	-9.77
/auftəgɑ/	/auntəko/	[au]	2.43	3.11	-13.31	-8.56	-14.59	-8.96
/nəfæmb/	/mɪnæmp/		2.49	3.03	-13.57	-9.32	-15.73	-11.08
/pwɑgəb/	/twɛkɛt/		1.69	2.28	-13.88	-9.93	-13.55	-10.78
/bufkit/	/kiften/	[f]	2.61	3.68	-14.00	-11.11	-15.57	-11.79
/dogdet/	/tæktut/		2.76	3.38	-14.00	-9.75	-15.57	-9.45
/kɛdəwəmb/	/fɪkətæmp/		2.14	3.13	-13.57	-9.32	-15.73	-11.08
/pwɛnətɛp/	/twɛdəmɪn/		1.90	2.13	-13.88	-9.93	-13.55	-10.78
/næfkətu/	/gʌftədaɪ/	[f]	2.73	2.44	-14.00	-11.11	-15.57	-11.79
/dɛgdəne/	/tiktəpo/		2.43	2.54	-14.00	-9.75	-15.57	-9.45

Table 2. Sample size and number of males in each age group and mean age and test scores (with standard deviations in parentheses).

Group characteristics,	Age group			
including test scores	3-4-year-olds	5-6-year-olds	7-8-year-olds	Adults
Sample Size	43	38	23	22
Age in months	50 (6)	66 (5)	97 (6)	303 (42)
Gender	27 male	23 male	13 male	10 male
GFTA percentile ranking[a]	65 (24)	70 (22)	79 (19)	
CMMS standard score[a,b]	109 (10)	111 (12)	108 (10)	
EVT standard score [b]	111 (9)	110 (13)	102 (7)	120 (11)
PPVT-III standard score [b]	114 (11)	114 (13)	112 (16)	119 (12)

[a]GFTA and CMMS are not normed for adults. [b]Standard scores have a mean of 100 and a standard deviation of 15.

Table 3. Number of tokens and mean durations in ms (with standard deviations in parenthesis) for each measured segment type in low- versus high-frequency target sequences for each age group.

	Age Group											
	3-4 Years			5-6 Years			7-8 Years			Adults		
Seq	N	High-Freq	Low-Freq	N	High-Freq	Low-Freqy	N	High-Freq	Low-Freqy	N	High-Freq	Low-Freq
au	39	186 (54)	189 (63)	31	184 (38)	195 (71)	21	180 (38)	173 (39)	15	173 (33)	165 (30)
f	23	114 (57)	117 (51)	19	108 (47)	131 (57)	18	124 (40)	126 (38)	30	105 (25)	92 (31)
g	18	107 (48)	108 (59)	27	85 (43)	107 (89)	17	76 (36)	60 (37)	17	59 (32)	77 (58)
m	38	53 (33)	82 (72)	33	68 (37)	79 (58)	23	64 (27)	71 (44)	16	70 (26)	75 (37)
n	36	77 (57)	91 (72)	31	88 (48)	132 (12)	22	85 (51)	102 (38)	16	124 (12)	99 (40)
pt	23	206 (45)	197 (47)	26	211 (74)	108 (72)	11	187 (43)	212(122)	9	204 (28)	168 (17)
v	35	70 (47)	78 (48)	46	85 (79)	84(106)	38	64 (49)	87 (51)	41	73 (46)	82 (34)

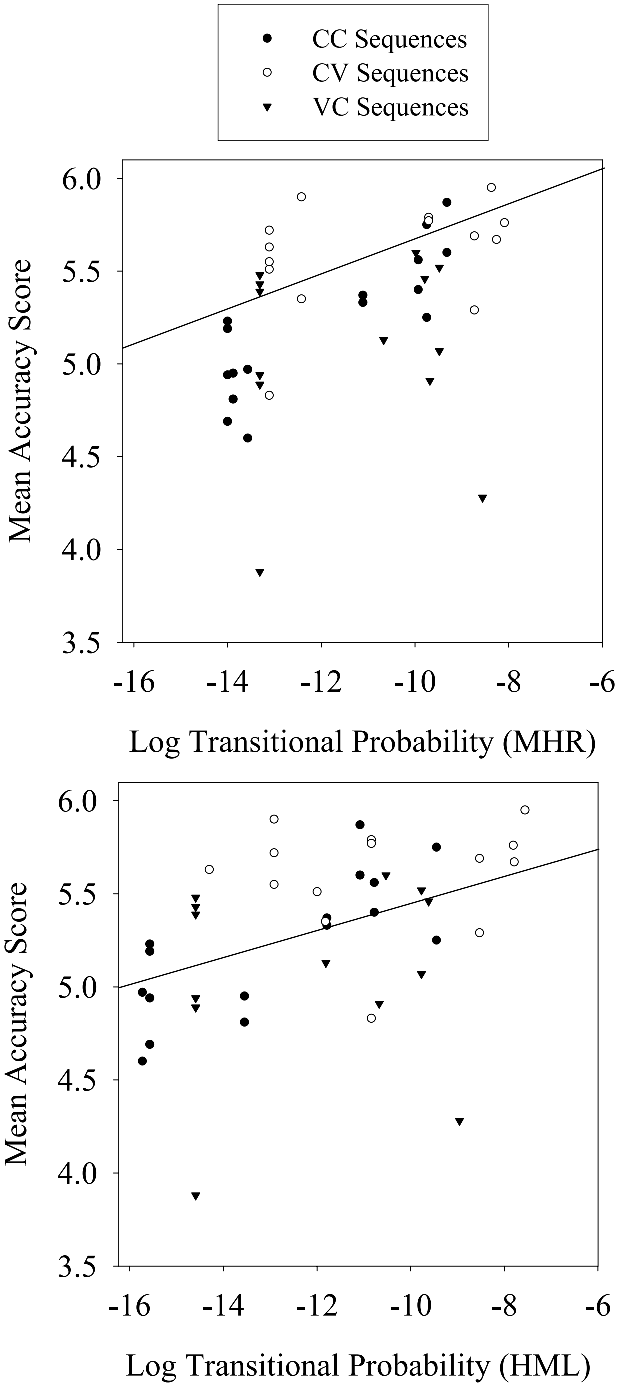

Figure 1. Mean accuracy for target sequence plotted against its transitional probability calculated from the MHR database (Fig. 1a, top plot) and from the HML database (Fig. 1b, bottom plot), for all 44 nonwords.

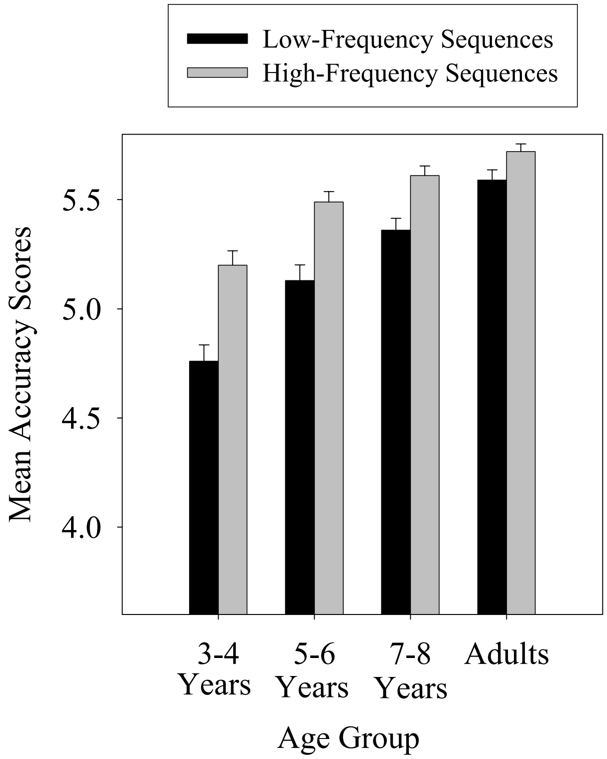

Figure 2. Mean accuracy scores (with standard errors) for the low- and high-frequency sequences for the four age groups.

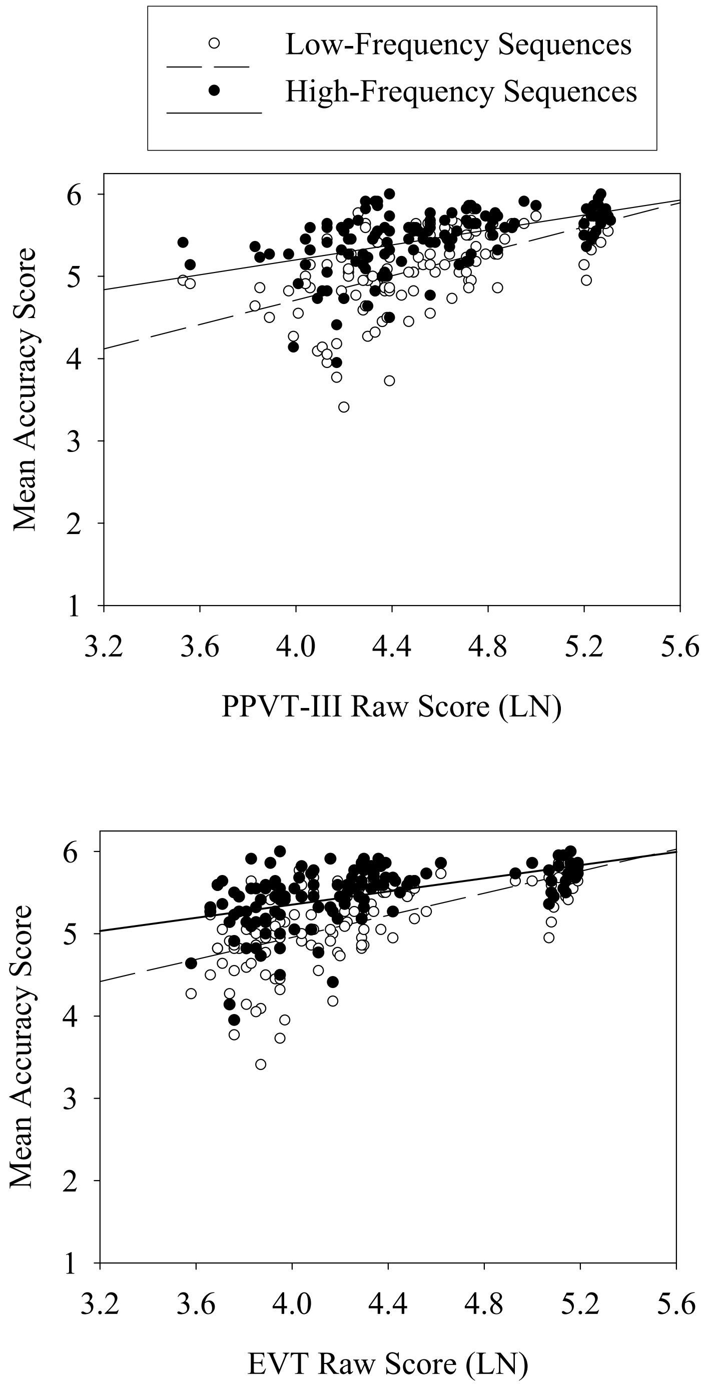

Figure 3. Mean accuracy scores for low- and high-frequency sequences plotted against receptive vocabulary size (PPVT-III, Fig. 3a, top plot) and expressive vocabulary size (EVT, Fig. 3b, bottom plot) for all participants.