J. Child Lang. **36** (2009), 291–321. © 2008 Cambridge University Press
doi:10.1017/S030500090800891X Printed in the United Kingdom

Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants*

HOLLY L. STORKEL
University of Kansas

(*Received* 14 *April* 2007. *Revised* 21 *December* 2007. *First published online* 2 *September* 2008)

ABSTRACT

The influence of phonological (i.e. individual sounds), lexical (i.e. whole-word forms) and semantic (i.e. meaning) characteristics on the words known by infants age 1;4 to 2;6 was examined, using an existing database (Dale & Fenson, 1996). For each noun, word frequency, two phonological (i.e. positional segment average, biphone average), two lexical (i.e. neighborhood density, word length) and four semantic variables (i.e. semantic set size, connectivity, probability resonance, resonance strength) were computed. Regression analyses showed that more infants knew (1) words composed of low-probability sounds and sound pairs, (2) shorter words with high neighborhood density, and (3) words that were semantically related to other words, both in terms of the number and strength of semantic connections. Moreover, the effect of phonological variables was constant across age, whereas the effect of lexical and semantic variables changed across age.

Three types of representations appear to play a role in word learning: phonological, lexical and semantic (e.g. Gupta & MacWhinney, 1997). PHONOLOGICAL REPRESENTATIONS refer to individual sounds (e.g. /k/, /æ/, /t/). LEXICAL REPRESENTATIONS refer to whole-word forms (e.g. /kæt/). SEMANTIC REPRESENTATIONS refer to the meaning or referent of a word (e.g. 'small furry four-legged pet'). The simplest example of word learning is

[*] This work was supported by the National Institutes of Health (DC08095; DC04781; DC005803; HD002528). David Slegers and Todd Little aided in the statistical analysis. Michael Vitevitch provided comments on an earlier version of this manuscript. Address for correspondence: Holly Storkel, PhD, Associate Professor, Department of Speech-Language-Hearing: Sciences and Disorders, University of Kansas, 3001 Dole Human Development Center, 1000 Sunnyside Avenue, Lawrence, KS 66045-7555. e-mail: hstorkel@ku.edu.

one in which a novel object labeled by a correctly perceived and articulated novel sound sequence is learned. In this case, it is assumed that when the novel word is encountered it activates existing phonological representations. These, in turn, activate lexical representations; however, a novel word will not exactly match an existing lexical representation. Likewise, a novel word will not exactly match an existing semantic representation. Thus, the formation of new lexical and semantic representations presumably is initiated. Word learning consists of creating new lexical and semantic representations, linking these new representations to one another, and integrating these new representations with existing phonological, lexical and semantic representations.

A limitation of current models of word learning is that many aspects of word learning are not fully specified, particularly as related to the phonological, lexical and semantic characteristics of the ambient language. That is, many models do not specify whether the influence of ambient characteristics on the formation of new representations is similar across phonological, lexical and semantic representations, and whether this influence might change over developmental time as more words are acquired. In part, these aspects of word learning are underspecified because the necessary data are limited. The goal of this paper is to provide preliminary evidence to address these three issues to promote elaboration of existing models as well as future research.

Phonological, lexical and semantic characteristics

There is evidence that phonological, lexical and semantic characteristics of the ambient language influence word learning. Beginning with phonological characteristics, there is evidence that the likelihood of occurrence of individual sounds and adjacent pairs of sounds in a language, termed PHONOTACTIC PROBABILITY, influences word learning. Phonotactic probability is considered a phonological characteristic because it measures characteristics of parts of words (i.e. single sounds and sound pairs) rather than whole words. Hollich, Jusczyk & Luce (2002) experimentally manipulated the characteristics of the ambient language through an exposure condition that occurred prior to a word learning task. In this study, infants age 1;5 were pre-exposed to non-words that contained similar or dissimilar sound sequences to the non-words to be learned. In the high-probability condition, 100% of the pre-exposure non-words contained similar sound sequences to the non-words to be learned. In the low-probability condition, only 25% of the pre-exposure non-words contained similar sound sequences to the non-words to be learned. In the word learning task, infants heard the to-be-learned non-word paired with a novel object. Results showed that infants learned the non-word in the high-probability condition but not in

the low-probability condition. Thus, the high-probability condition promoted word learning, whereas the low-probability condition did not.

Turning now to lexical characteristics, there is evidence that the number of words in a language that sound similar to a given word, namely NEIGHBORHOOD DENSITY, influences word learning. Neighborhood density is considered a lexical characteristic because it indexes whole-word similarity. The previously reported study by Hollich *et al.* (2002) also examined the influence of neighborhood density on word learning by infants age 1;5. As with phonotactic probability, neighborhood density was created in the experiment through pre-exposure to non-words that contained similar or dissimilar sound sequences. These non-words were repeated six times so that infants presumably would recognize them as whole words, rather than extracting just the parts of the non-words (i.e. individual sounds and adjacent pairs of sounds). In the high-density condition, 100% of the pre-exposure non-words were neighbors of the to-be-learned non-word. In the low-density condition, 25% of the pre-exposure non-words were neighbors of the to-be-learned non-word. Infants then heard the to-be-learned non-word paired with a novel object. Results showed that infants learned the non-word in the low-density condition but not the high-density condition. Thus, low density promoted word learning.

Finally, there is clear evidence that semantic characteristics of the ambient language bias young children to attend to certain semantic features. For example, infants appear to extend novel words referring to count nouns (i.e. individuated solid items such as 'car') on the basis of shape, presumably because they have extracted the regularity that count noun categories tend to refer to items sharing shape (see Smith, 2000, for review). Likewise, infants extend novel words referring to mass nouns (i.e. non-individuated non-solid items such as 'water') on the basis of material because it is precisely this feature that is shared across members of the same category (e.g. Soja, Carey & Spelke, 1991). Thus, infants appear to extract semantic regularities from known words and use these regularities to learn new words. Taken together, there is ample evidence that characteristics of the ambient language influence word learning, but it is unclear whether similar patterns of influence are observed across phonological, lexical and semantic characteristics, bringing us to the next issue.

Comparing across characteristics

There are two difficulties in making comparisons across phonological, lexical and semantic characteristics. The first relates to correlations between phonological and lexical characteristics, making it difficult to isolate the contribution of each to word learning. The second relates to differences in the types of characteristics studied across phonological/lexical characteristics

versus semantic characteristics, making it difficult to identify similarities and differences in the influence of each characteristic on word learning.

In terms of correlations between characteristics, phonotactic probability, a phonological characteristic, is positively correlated with neighborhood density, a lexical characteristic (Storkel, 2004b; Vitevitch, Luce, Pisoni & Auer, 1999). Specifically, whole words that are similar to many other words in the language tend to be composed of individual sounds and pairs of sounds that frequently occur in the language. Past studies of word learning have failed to differentiate the effects of each of these characteristics (Hollich *et al.*, 2002; Storkel, 2001, 2004a; Storkel & Rogers, 2000). That is, conclusions concerning the characteristic that influences word learning have been based on the direction of the observed effect (i.e. performance on high probability/density superior to low probability/density or performance on low probability/density superior to high probability/density) and the consistency of this effect with past studies of other types of spoken language processing by adults (e.g. recognition, production). Specifically, when performance is better for high probability/density than low probability/density, the effect is typically attributed to phonotactic probability, a phonological characteristic; whereas, when performance is better for low probability/density than high probability/density, the effect is attributed to neighborhood density, a lexical characteristic. This convention stems from recognition findings from adult research (e.g. Vitevitch & Luce, 1999).

In this way, the previously described study by Hollich and colleagues (2002) compared learning of high-probability/high-density non-words to low-probability/low-density non-words in both experiments. The only difference across the experiments was the amount of pre-exposure to the non-words. With minimal repetition of the pre-exposure non-words, high-probability/density non-words were learned better than low-probability/density non-words, and this was interpreted as an effect of phonotactic probability. With greater repetition of the pre-exposure non-words, low-probability/density non-words were learned better than high-probability/density non-words, and this was interpreted as an effect of neighborhood density. This inference is consistent with past interpretation of adult word recognition findings and is based on the assumption that greater repetition would lead to recognition of the pre-exposure non-words as whole-word units. Further evidence is needed to support this inference. In particular, it would be desirable to determine whether phonotactic probability and/or neighborhood density make unique contributions to the prediction of which words are learned by infants while the alternate characteristic is controlled.

A second problem is that the influence of phonological and lexical characteristics on word learning cannot be directly compared to those of semantic characteristics because of differences in the types of characteristics studied. The study of phonological and lexical influences on word learning

has focused on general properties, such as the number of related representations (i.e. neighborhood density). In contrast, many studies of semantics in word learning have focused on specific regularities that affect a subset of words (e.g. count nouns vs. mass nouns). However, there are semantic measures that parallel the lexical characteristic of neighborhood density, namely semantic set size.

SEMANTIC SET SIZE is the number of words that are meaningfully related to or frequently associated with a given word (i.e. the number of semantic neighbors) and is similar to neighborhood density in quantifying the number of similar items in the language. This characteristic has been extensively studied in the adult memory literature (see Nelson & Zhang, 2000; Nelson, Zhang & McKinney, 2001, for review). In this literature, semantic set size has typically been examined in a study-test paradigm where the participant is given a list of known words to study and later tested on recall of the items on the list or on recognition of the items on the list from a field of choices. Generally, recall is superior for words with a small set size (i.e. few semantic neighbors) than words with a large set size (e.g. Nelson, McKinney, Gee & Janczura, 1998; Nelson *et al.*, 2001). As with neighborhood density, the effect of semantic set size appears to vary by task. For example, in lexical decision of visually presented words, words with a large set size are responded to more quickly than words with a small set size (Buchanan, Westbury & Burgess, 2001). To our knowledge, the effect of semantic set size on word learning has not been investigated at any age and provides a promising avenue for examining similarities and differences between the influences of lexical versus semantic characteristics on word learning.

The adult memory literature yields several other characteristics related to more fine-grain aspects of the semantic structure of the ambient language. Generally, it is thought that these characteristics are important in the retrieval of a known target word because they affect the amount of activation that is spread to the target word from its semantic neighbors (Nelson, McKinney *et al.*, 1998). It is possible that these characteristics might influence word learning. The specific semantic characteristics include: CONNECTIVITY, the number of inter-connections between semantic neighbors in a given neighborhood; PROBABILITY RESONANCE, the number of bidirectional connections between the target word and its semantic neighbors; and RESONANCE STRENGTH, the weighting of the bidirectional connections between the target word and its semantic neighbors. Generally, memory research demonstrates superior recall and recognition for words with high connectivity, high-probability resonance and high-resonance strength when compared to those with low connectivity, low-probability resonance and low-resonance strength (e.g. Nelson, McKinney *et al.*, 1998; Nelson & Zhang, 2000; Nelson *et al.*, 2001). As with semantic set size, the role of these characteristics in word learning is unattested for any age group.

If these semantic characteristics influence word learning, then investigation of lexical analogs may be warranted.

Changes across development

The majority of studies examining these phonological and lexical factors in word learning have investigated only a single age group of children in the same study (Hollich *et al*., 2002; Storkel, 2001, 2004a; but see Storkel & Rogers, 2000). There is evidence that a child's phonological, lexical and semantic representations change across development as words are acquired, becoming stronger and/or more detailed, and that the processes that children use to learn new words change with word learning experience (e.g. Gershkoff-Stowe, 2002; Hollich *et al*., 2000). Developmental studies of other areas of language processing suggest different developmental patterns for phonological versus lexical characteristics. Specifically, the influence of phonotactic probability on language processing (i.e. oral naming) remains relatively constant with development (i.e. from adolescence to old adults), whereas the influence of neighborhood density on language processing may decrease with development (Newman & German, 2005). The consequence of these changes in representations and processing for word learning warrants further exploration.

Although several word learning studies are available that sample the effect of phonotactic probability and neighborhood density on word learning by infants (Hollich *et al*., 2002), preschool children (Storkel, 2001), school-age children (Storkel & Rogers, 2000) and adults (Storkel, Armbruster & Hogan, 2006), it is difficult to make comparisons across these studies because of methodological differences. For example, the infant study experimentally determined phonotactic probability and neighborhood density through pre-exposure, whereas the other studies defined phonotactic probability and neighborhood density relative to the ambient language. In addition, all of the child studies examined correlated phonotactic probability and neighborhood density, whereas the adult study differentiated the effects of phonotactic probability and neighborhood density by fully crossing the two characteristics. Different patterns in the effect of phonotactic probability and neighborhood density on word learning are observed across these studies but it is difficult to determine whether these differences are due to developmental or methodological changes. Thus, further examination of the developmental trajectory of the influence of phonotactic probability and neighborhood density on word learning is warranted.

Turning to semantics, there is clear evidence that changes in semantic representations and processes do have consequences for word learning. For example, the previously mentioned shape bias appears to emerge only after a number of count nouns have been learned (see Gershkoff-Stowe & Smith,

2004; Smith, 2000, for review). Developmental changes in the influence of semantic characteristics that parallel phonological and lexical characteristics (e.g. semantic set size) have not been examined. Therefore, the developmental trajectory of these semantic characteristics is not yet known.

Purpose of the current study

The main goal of the current study was to examine how phonological, lexical and semantic characteristics of the ambient language influence word learning by infants. Three specific objectives drive this study. The first objective was to differentiate effects of phonotactic probability, a phonological characteristic, and neighborhood density, a lexical characteristic, on word learning because most previous studies have examined these variables when correlated (Hollich *et al.*, 2002; Storkel, 2001; Storkel & Rogers, 2000; but see Storkel *et al.*, 2006). The second objective was to examine the role of a semantic characteristic that paralleled the lexical characteristic of neighborhood density to determine whether the influence of semantic and lexical characteristics on word learning was similar. In addition, the role of other semantic characteristics, namely connectivity, probability resonance and resonance strength, in word learning was explored to determine whether these warranted further investigation in the future. The final objective was to investigate how the influence of phonological, lexical and semantic characteristics on word learning changes during infancy to provide evidence of the developmental trajectory of each of these representations.

To accomplish these goals, numerous phonological, lexical and semantic variables were computed for a naturalistic database of words known by infants. A correlation analysis was conducted to determine whether variables that were hypothesized to index the same characteristic (i.e. phonological vs. lexical vs. semantic) could be combined to create a composite score. These composite scores were then entered as predictors of word learning in a regression analysis to address the study's objectives. Note that in a regression analysis, if a given predictor is significant, then it accounts for unique variance over and above that accounted for by the other predictors in the analysis. In terms of the first objective, it was hypothesized that both phonological and lexical characteristics would be significant predictors of word learning, indicating that both uniquely influence word learning. This finding would provide support to inferences from previous studies that both phonological and lexical characteristics influence word learning (Hollich *et al.*, 2002; Storkel, 2001; Storkel & Rogers, 2000) and would parallel findings from adult word learning (Storkel *et al.*, 2006). Turning to the second objective, semantic characteristics were hypothesized to be significant predictors of word learning based on past evidence that these characteristics influence memory in adults (Nelson & Zhang, 2000;

Nelson *et al.*, 2001) and based on the assumption that memory is a critical component of word learning. No prediction could be made concerning the similarity between lexical and semantic influences on word learning because past studies have shown that the effect of each of these characteristics varies by tasks, making it unclear what the effect might be on word learning. Finally, it was predicted that each characteristic would show a unique developmental trajectory based on past developmental studies of phonological, lexical and semantic influences on language processing and word learning (Gershkoff-Stowe & Smith, 2004; Newman & German, 2005; Smith, 2000). Specifically, across development it was predicted that phonological effects would be constant, lexical effects would diminish and semantic effects would strengthen.

METHOD

Database

The *MacArthur-Bates Communicative Development Inventory: Words and Sentences* (CDI) is an expressive vocabulary measure, consisting of a list of 680 words potentially known by children age 1;4 to 2;6 (Fenson *et al.*, 1993). On the inventory, parents indicate which words their child produces, regardless of the accuracy of that production. Previous studies indicate that the CDI is a valid and reliable measure of the words known by infants, demonstrating a high correlation with other vocabulary measures (e.g. Fenson *et al.*, 1993). Data are available from a national cross-sectional sample of 1,800 American children (Dale & Fenson, 1996). This database consists of the percentage of children from the normative sample who were reported to know each of the CDI words at 1-month age intervals between 1;4 and 2;6. Only nouns were examined in this analysis because of the previously reported discrepancies between noun and verb learning (e.g. Leonard *et al.*, 1982). This yielded 380 nouns for analysis.

Two outcome variables were derived from this data set: (1) the percentage of children reported to know a given word at each 1-month age interval from 1;4 to 2;6; (2) the 75% age-of-acquisition for each word. The 75% age-of-acquisition was defined as the earliest age when 75% (or more) of the children tested were reported to know the word. The analysis section below describes how each of these variables was used. In addition, phonological, lexical and semantic variables were computed for each word on the CDI.

Phonological predictor variables

The phonemic transcription of the nouns on the CDI was obtained from a 20,000-word computer readable dictionary (*Webster's Seventh Collegiate Dictionary*, 1967). Two measures of phonotactic probability were computed, positional segment average and biphone average, using an algorithm

from previous studies (e.g. Storkel, 2001, 2004a; Vitevitch & Luce, 1999). Note that these variables are computed using an adult dictionary. This adult dictionary is taken as being representative of the ambient language that an infant is exposed to. This assumption is supported by the finding that infants extract the phonotactic probability of their native language by 0;9 (Jusczyk, Luce & Charles-Luce, 1994). Likewise, similar values are obtained when phonotactic probability is calculated based on an adult versus child corpus, suggesting that the rank ordering of words from low probability to high probability is similar even though absolute values may differ across corpora (Jusczyk *et al.*, 1994).

Positional segment average. Positional segment average is the mean likelihood of occurrence of each sound in a given word position. This was computed for each sound in a word by summing the log frequency of all the words in the dictionary that contained the target sound in the same word position and dividing by the sum of the log frequency of all the words in the dictionary that contained any sound in the same word position. These positional segment frequencies of each sound in the word were then summed and divided by the number of sounds in the word to create an average (Storkel, 2004b). To illustrate, the calculation for the word 'cat' (i.e. /kæt/) would be the sum of the positional segment frequency for /k/ in the first position (i.e. 0·0927), /æ/ in the second position (i.e. 0·0794) and /t/ in the third position (i.e. 0·0660), divided by the number of sounds in the word (i.e. 3), yielding a positional segment average of 0·0794 (i.e. 0·2381/3). Positional segment averages were computed for all 380 nouns.

Biphone average. Biphone average is the mean likelihood of occurrence of each pair of adjacent sounds in a given word position. This was computed for each pair of sounds in a word by summing the log frequency of all the words in the dictionary that contained the target sound pair in the same word position and dividing by the sum of the log frequency of all the words in the dictionary that contained any sound in the same word position. These individual biphone frequencies were summed and divided by the number of biphones to create an average (Storkel, 2004b). Continuing the illustration with 'cat' (i.e. /kæt/), the calculation would be the sum of the biphone frequency for /kæ/ in the first position (i.e. 0·0122) and /æt/ in the second position (i.e. 0·0059), divided by the number of biphones in the word (i.e. 2), yielding a biphone average of 0·0091 (i.e. 0·0181/2). Biphone averages were computed for 379 of the 380 nouns. One of the nouns had no biphones (i.e. 'eye' /aɪ/).

Lexical predictor variables

Although neighborhood density was the main lexical variable of interest, past research has shown that neighborhood density is correlated with word

length, such that shorter words tend to have more neighbors and longer words tend to have fewer neighbors (e.g. Pisoni, Nusbaum, Luce & Slowiaczek, 1985). Thus, two lexical variables were computed for each noun on the CDI: neighborhood density and word length. As with the phonological variables, lexical variables were computed relative to an adult dictionary (*Webster's Seventh Collegiate Dictionary*, 1967). This approach is somewhat more controversial because others have reported that the size of lexical neighborhoods change over time (e.g. Charles-Luce & Luce, 1990; Coady & Aslin, 2003). However, this study focused on the characteristics of the ambient language rather than characteristics of individual children.

Neighborhood density. Neighborhood density was operationally defined as the number of words in the dictionary that differed from a target word by a one sound substitution, addition or deletion in any word position (Luce & Pisoni, 1998), regardless of meaning or syntactic class. For example, neighbors of /kæt/ (i.e. 'cat') include /*tʃ*æt/ (i.e. 'chat'), /*r*æt/ (i.e. 'rat'), /*m*æt/ (i.e. 'mat'), /*b*æt/ (i.e. 'bat'), /*p*æt/ (i.e. 'pat'), /*h*æt/ (i.e. 'hat'), /*f*æt/ (i.e. 'fat'), /*s*æt/ (i.e. 'sat'), /k*ɑ*t/ (i.e. 'cot'), /k*aɪ*t/ (i.e. 'kite'), /k*oʊ*t/ (i.e. 'coat'), /k*ɔ*t/ (i.e. 'caught'), /k*ʌ*t/ (i.e. 'cut'), /kæ*ʃ*/ (i.e. 'cash'), /kæ*tʃ*/ (i.e. 'catch'), /kæ*n*/ (i.e. 'can'), /kæ*p*/ (i.e. 'cap'), /kæ*b*/ (i.e. 'cab'), /kæ*s*t/ (i.e. 'cast'), /kæ*n*t/ (i.e. 'can't') and /æt/ (i.e. 'at'). Thus, the number of neighboring words for 'cat' is 21. Neighborhood density was computed for all 380 nouns.

Word length. Word length was computed by counting the number of phonemes in the dictionary transcription. Word length was computed for all 380 nouns.

Word frequency

Previous studies have shown that neighborhood density, a primary variable of interest, is correlated with word frequency, such that higher-frequency words tend to have more neighbors and lower-frequency words tend to have fewer neighbors (Landauer & Streeter, 1973). Moreover, word frequency indexes how often a person might encounter a given word, which will likely impact whether or not the word is learned (Rice, Oetting, Marquis, Bode & Pae, 1994). Finally, word frequency represents encounters with both the form of the word and the meaning of the word. Therefore, word frequency may not be a purely lexical or semantic variable. Because this was a post-hoc analysis of existing data, it was necessary to explore this potentially related variable to determine whether the effect of the other variables could be isolated from word frequency. Word frequency was obtained from Kucera & Francis (1967), an adult-based frequency count, which is assumed to represent the frequency of words in the ambient language. Similar to the previous variables, past work has shown that the Kucera & Francis

frequency counts are significantly correlated with child-based counts (Gierut & Dale, 2007). Word frequency values were obtained for 320 of the 380 nouns.

Semantic predictor variables

Four semantic variables were computed for each noun on the CDI: semantic set size, connectivity, probability resonance and resonance strength. These variables were obtained from a corpus of discrete association norms (Nelson, McEvoy & Schreiber, 1998). These data were collected by having adults write the first word that came to mind that was meaningfully related to or frequently associated with the target word. The four semantic variables were then derived by Nelson and colleagues from these data. Of the 380 nouns, 335 were found in this corpus. As with the other variables, an adult corpus was used for these semantic variables. This was necessary because no child corpora were available to provide such detailed measures of semantics. Like the lexical variables, we assume that adult-based semantic variables would be correlated with child-based semantic variables, although absolute values and specific neighbors might vary across adult and child counts (see Entwisle (1966) for data supporting this assumption for semantic set size).

Semantic set size. Semantic set size is the number of different words generated by two or more participants in response to a target word. These associate words can be thought of as the semantic neighbors of the target word. For example, 'cat' has three semantic neighbors: 'dog', 'mouse' and 'kitten'. Thus, semantic set size is the semantic analog of neighborhood density.

Connectivity. Connectivity refers to the mean number of connections among the semantic neighbors of a target word. Connectivity is calculated by gathering discrete association norms for the semantic neighbors of the target, counting the number of connections among these neighbors and dividing by the semantic set size. For example, each neighbor of 'cat' (i.e. 'dog', 'mouse' and 'kitten') would be presented in the discrete association task to determine its neighbors. In the case of 'dog', the semantic neighbors were 'cat', 'puppy', 'friend', 'animal' and 'house'. Then, it was determined whether the other neighbors of 'cat' (i.e. 'mouse' or 'kitten') had been reported as neighbors of 'dog'. In this case, they were not. Thus, the connectivity would be 0 for 'dog'. This was done for 'mouse' and 'kitten'. Then, the connectivity of each semantic neighbor was summed (i.e. 'dog' 0, 'mouse' 0, and 'kitten' 0) and divided by the total number of semantic neighbors (i.e. 3). For 'cat', none of the neighbors produced each other as neighbors, leading to a connectivity of 0 (i.e. $0/3 = 0$).

For a second illustration, 'aunt', like 'cat', has a semantic set size of three but a higher connectivity. The neighbors of 'aunt' are 'uncle', 'relative' and 'relation'. To determine connectivity, the neighbors of each neighbor

of 'aunt' are examined for overlap. Specifically, the neighbors of 'uncle' included 'relative', one of the neighbors of aunt. Thus, connectivity would be 1 for 'uncle'. The neighbors of 'relative' included 'uncle' and 'relation', which are both neighbors of 'aunt' (i.e. connectivity = 2). The neighbors of 'relation' included 'relative', a neighbor of 'aunt' (i.e. connectivity = 1). The connectivity of each 'aunt' neighbor is summed (i.e. 'uncle' 1, 'relation' 2 and 'relative' 1) and divided by the total number of semantic neighbors (i.e. 3), yielding a connectivity of 1·33 (i.e. 4/3). As seen in these two illustrations, connectivity represents the mean number of neighbor-to-neighbor connections, with higher numbers indicating greater overlap between neighbors of a given word (see Nelson, McKinney *et al.*, 1998: Figure 1, for another example).

Probability resonance. Probability resonance is a measure of the bidirectional connections between the target word and its semantic neighbors. This value is determined by obtaining discrete association norms for the semantic neighbors of a target word, counting the number of semantic neighbors of a target word that also produce the target word as a neighbor, and dividing by the semantic set size. For example, a neighbor of 'cat', such as 'dog', would be presented using the discrete association methodology to determine whether adults would produce 'cat' in response to 'dog'. In fact, when given 'dog', adults do produce 'cat' as a neighbor. Thus, there is a bidirectional connection between 'cat' and 'dog'. Specifically, 'dog' is a semantic neighbor of 'cat', and 'cat' is a semantic neighbor of 'dog'. In this way, it is not assumed that all semantic relationships are reciprocal, and probability resonance provides a means of testing for these reciprocal relationships. Probability resonance ranges in value from 0 (i.e. no bidirectional connections) to 1 (i.e. all target-neighbor connections are bidirectional). Returning to the example, all neighbors of 'cat' had bidirectional connections, so the probability resonance was 1 (i.e. 3 bidirectional connections/3 semantic neighbors = 1). Thus, resonance represents the neighbor-to-target bidirectional connections (see Nelson, McKinney *et al.*, 1998: Figure 1 for another example).

Resonance strength. Resonance strength relates to the weight of the bidirectional connections between target words and semantic neighbors. STRENGTH is the proportion of participants who produced a word as an associate of another word. Resonance strength is computed by multiplying the target-to-neighbor strength by the neighbor-to-target strength for each target–neighbor pair. These products are then summed. For example, when given 'cat', 80 of 145 adults produced 'dog' as a neighbor. Thus, the 'cat' to 'dog' strength is 0·51 (i.e. 80/145). When given 'dog', 104 of 156 adults produced 'cat' as a neighbor. Thus, the 'dog' to 'cat' strength is 0·67 (i.e. 104/156). These two values are multiplied (i.e. $0{\cdot}51 \times 0{\cdot}67$). This is done for each bidirectional connection and the results are summed (i.e. $(0{\cdot}51 \times 0{\cdot}67$ for cat–dog$) + (0{\cdot}26 \times 0{\cdot}54$ for cat–mouse$) + (0{\cdot}16 \times 0{\cdot}79$ for

cat–kitten), yielding a resonance strength of 0·61 for 'cat'. Like probability resonance, resonance strength is an index of the bidirectional connections between a target word and its neighbors. Values range, in theory, from 0 to 1, but values of 1 are seldom (if ever) obtained.

Analysis

Two types of analyses were performed: correlation and regression. The correlation analysis examined the relationships within and across the two phonological variables, the two lexical variables, word frequency and the four semantic variables to determine which variables could be combined into composite scores to reduce the number of predictors for the regression analysis. The general criteria for combining variables was: (1) the variables to be combined were significantly (i.e. $p < 0{\cdot}01$) and highly correlated (i.e. $r > 0{\cdot}50$, large effect (Cohen, 1988)) with each other; and (2) the variables to be combined were minimally correlated with other variables (i.e. $r < 0{\cdot}30$, medium effect (Cohen, 1988)). To compute composite scores, raw values for the selected variables were converted to z scores (i.e. $z = (\text{obtained raw value} - M)/SD$) and then averaged.

Two regression analyses using the composite scores were performed to address the objectives of the study. In the first regression analysis, the data were STACKED, meaning that CDI words were repeated for each 1-month age interval (i.e. 1;4–2;6) so that the percentage of children reported to know each word at each age (i.e. 1;4–2;6) could be analyzed to explore developmental trajectories. In this analysis, composite scores, age, and interactions between composite scores and age served as predictor variables, and the percentage of children reported to know a given word at each age served as an outcome variable. Because the data were stacked, the interactions between the composite scores and age are the predictors of greatest interest, whereas the analysis of the main effects is not quite valid because of the dependency caused by listing item data (i.e. composite scores) multiple times. To provide a more valid analysis of main effects, a second linear regression analysis was performed on UNSTACKED data, meaning composite scores for CDI words were only entered one time. Here, the composite scores were entered as predictors of the 75% age-of-acquisition criterion in a regression analysis. The results of the second analysis confirmed those of the first, and both are reported below. In both regression analyses, all composite scores were entered in the analysis so that the effect of each composite variable could be examined while the other composite variables were held constant. This is a critical feature of this analysis because of the previously reported correlations among variables.

A number of the variables were taken from existing corpora (i.e. word frequency, semantic set size, connectivity, probability resonance and

resonance strength). In some cases, words from the CDI were not available in the corpora used, leading to missing values. The pattern of missing data was as follows: 27 words (7%) were missing word frequency, 12 words (3%) were missing all four semantic variables, and 33 words (9%) were missing both word frequency and the four semantic variables. To avoid deleting these 72 words (19%) from analysis, missing values were statistically imputed (i.e. the missing value was replaced by statistically estimating a value based on the patterns observed in the non-missing data) prior to analysis, using the multiple imputation procedure in SAS and the EM algorithm (Yuan, 2002). Note that all patterns observed with the imputed data also were observed when the same analyses were performed without the imputed values. Thus, imputing the data did not alter the observed patterns but did increase power to detect statistically significant differences. In addition, the assumptions of each statistical analysis were checked prior to performing a given analysis. No significant violations were detected.

RESULTS

Correlation among variables

Table 1 shows the correlation among variables. As can be seen from this table, the highest correlation for phonological variables was observed for the within-domain correlation of positional segment average and biphone average ($r = 0{\cdot}67$). All other significant cross-domain correlations were much lower ($r \leqslant 0{\cdot}20$). This supports combining positional segment average and biphone average to create a composite phonological variable.

Turning to the lexical variables, the highest correlation was observed for the within-domain correlation of neighborhood density and word length ($r = -0{\cdot}69$), whereas the correlation with other variables was much lower ($r = 0{\cdot}28$ or less). This supports combining these two variables into a lexical composite score. Given the negative correlation (i.e. inverse relationship), word length had to be reverse coded (i.e. sign of the z-score reversed) to create a meaningful composite score. Thus, positive scores on this composite represent higher-density words with fewer phonemes and negative scores represent lower-density words with more phonemes.

As expected, word frequency was significantly correlated with several phonological (i.e. biphone average), lexical (i.e. neighborhood density, word length) and semantic variables (i.e. probability resonance, resonance strength) and all of these correlations were similar in magnitude (i.e. $0{\cdot}15 < r < 0{\cdot}23$, medium effect). Word frequency was not combined with any other variables to create a composite score because it appeared to cross-cut all three domains, supporting the initial hypothesis that word frequency indexes the number of encounters with a word's sound form (i.e. phonological and lexical characteristics) and meaning (i.e. semantic characteristic).

TABLE 1. *Correlations among predictor variables in the non-stacked data set*

		Phonological		Lexical			Semantic			
		1	2	3	4	5	6	7	8	9
Phonological	1. Positional segment average	—								
	2. Biphone average	0·67**	—							
Lexical	3. Neighborhood density	0·20**	0·10*	—						
	4. Word length	0·09	0·12*	−0·69**	—					
	5. Word frequency	0·07	0·15**	0·21**	−0·16**	—				
Semantic	6. Semantic set size	−0·02	0·02	−0·01	0·04	0·08	—			
	7. Connectivity	−0·02	−0·01	−0·09	0·09	−0·02	0·32**	—		
	8. Probability resonance	0·08	0·10	0·25**	−0·28**	0·22**	−0·11*	0·04	—	
	9. Resonance strength	0·05	0·11	0·20**	−0·24**	0·23**	−0·32**	−0·13*	0·51**	—

* $p < 0·05$, ** $p < 0·01$

Finally, the relationship among the semantic variables was less clear-cut. The a priori criteria failed to capture some observed patterns. Specifically, there appeared to be a possible separation between semantic set size and connectivity versus probability resonance and resonance strength, although this separation was somewhat ambiguous. Semantic set size showed a medium positive correlation with connectivity ($r=0{\cdot}32$) and negative correlations with probability resonance ($r=-0{\cdot}11$) and resonance strength ($r=-0{\cdot}32$). Turning to cross-domain correlations, semantic set size and connectivity were not significantly correlated with any of the non-semantic variables. On the other hand, probability resonance showed a large positive correlation with resonance strength ($r=0{\cdot}51$) and both probability resonance and resonance strength showed significant cross-domain correlations with lexical variables (i.e. neighborhood density, word length) and word frequency. Although the evidence for creating two separate semantic composite scores was weaker than for the other composite scores, it was decided that the exploratory nature of this study warranted maintaining the separation. That is, without prior evidence, it is possible that the effect of semantic set size and connectivity on word learning could differ from the effect of probability resonance and resonance strength on word learning. Combining all four semantic variables into one semantic composite could obscure this potential difference, which would not be desirable in the first study to address this issue. Furthermore, the two composite scores are cohesive theoretically. Specifically, the composite of semantic set size and connectivity captures global semantic structure, whereas the composite of probability resonance and resonance strength specifically captures bidirectional connections between a target and its semantic neighbors.

Regression analyses

The four composite scores (phonological, lexical, semantic neighbor, semantic bidirectional), word frequency, age and interactions between all variables and age were entered as possible predictors of the percentage of children reported to know a given word at a given age in the stacked regression analysis. Table 2 provides the regression results from the stacked data set. The phonological composite score was a significant predictor of the percentage of infants reported to know a given word. Specifically, fewer infants knew words composed of higher-probability sound sequences than words composed of lower-probability sound sequences. The effect of the lexical composite score was significant, with more infants knowing shorter words with many neighbors than longer words with few neighbors. Note the patterns observed for the phonological and lexical composites are the opposite of the patterns previously inferred by Hollich and colleagues (2002). The effect of word frequency was not significant but the trend

TABLE 2. *Regression analysis results from the stacked data set*

	β estimate	Standard error	t	p	r^2
Intercept	49·57	0·26	192·30	<0·0001	
Phonological composite	−2·68	0·29	−9·32	<0·0001	0·0043
Lexical composite	5·69	0·34	16·97	<0·0001	0·0452
Word frequency	0·55	0·29	1·87	0·065	0·0003
Semantic neighbor composite	2·24	0·44	5·03	0·0001	0·0019
Semantic bidirectional composite	4·39	0·50	8·84	<0·0001	0·0159
Age	4·80	0·06	82·15	<0·0001	0·5063
Age × phonological composite	0·08	0·06	1·27	0·203	0·0001
Age × lexical composite	−0·19	0·07	−2·83	0·005	0·0005
Age × word frequency	−0·06	0·06	−0·96	0·337	0·0001
Age × semantic neighbor composite	0·16	0·08	2·10	0·036	0·0002
Age × semantic bidirectional composite	0·17	0·08	2·32	0·021	0·0004

NOTE: For main effects, positive slope estimates indicate that the proportion of infants who knew a word increased as the variable increased, whereas negative slope estimates indicate that the proportion of infants who knew a word decreased as the variable increased. For interactions with age, positive slope estimates indicate that the steepness of the slope of the variable increased as age increased, whereas negative slope estimates indicate that the steepness of the slope of the variable decreased as age increased.

was for more infants to know high-frequency words than low-frequency words. The two semantic composite scores also were significant predictors of the percentage of infants reported to know a given word. Specifically, more infants knew words with many interconnected neighbors and many strong target-to-neighbor bidirectional connections than words with few interconnected neighbors and few weak target-to-neighbor bidirectional connections. Age was a significant predictor with older infants knowing more words than younger infants.

These main effects were further confirmed through the alternative non-stacked regression analysis where the four composite scores and word frequency were entered as predictors of the 75% age-of-acquisition criterion. The results from the non-stacked model are shown in Table 3. Note that the signs of the β estimates are the opposite of those reported for the stacked analysis because of the change in the dependent variable. A negative slope is interpreted as the age-of-acquisition of the word decreasing as the variable increases. The results of the non-stacked model replicate those of the stacked model exactly. Specifically, the non-stacked regression confirms that higher-probability words were learned at later ages than lower-probability words. Likewise, the non-stacked regression confirmed the

TABLE 3. *Regression analysis results from the non-stacked data set*

	β estimate	Standard error	t	p	r^2
Intercept	28·67	0·32	89·57	<0·0001	
Phonological composite	0·84	0·35	2·37	0·0178	0·008
Lexical composite	−1·62	0·37	−4·39	<0·0001	0·076
Word frequency	−0·29	0·34	−0·86	0·3901	0·008
Semantic neighbor composite	−0·95	0·42	−2·28	0·0228	0·007
Semantic bidirectional composite	−1·56	0·43	−3·58	0·0004	0·035

NOTE: Negative slope estimates indicate that the 75% age-of-acquisition decreased as the variable increased, whereas positive slope estimates indicate that the 75% age-of-acquisition increased as the variable increased.

findings for the lexical and semantic composite scores, namely words with many lexical neighbors and few phonemes, many interconnected semantic neighbors and many strong target-to-neighbor bidirectional connections were learned at earlier ages than words with low values on those same composites. Finally, the effect of word frequency was not significant but in the same direction as in the stacked analysis with high-frequency words being learned at earlier ages than low-frequency words.

These main effects must be interpreted with caution due to significant interactions with age in the stacked model (see Table 2). To illustrate both non-significant and significant interactions with age, figures were constructed (see Figures 1–5). The composite score of interest (i.e. phonological, lexical, semantic neighbor, semantic bidirectional) was partitioned into four intervals: scores less than −1·00, scores between −1·00 and 0 (including −1·00 and 0), scores between 0 and +1·00 (excluding 0 but including +1·00) and scores greater than +1·00. In general, these intervals resulted in a similar distribution of data across composite scores with approximately 12% of the data in the first interval, 41% of the data in the second interval, 34% in the third interval and 13% in the last interval. The exception to this was the word frequency data, where the majority of the data (i.e. 80%) was clustered in the second interval. Thus, a decision was made to use different intervals for word frequency to provide a better distribution of the data. Specifically, the intervals for word frequency were: scores less than −0·25, scores between −0·25 and 0 (including −0·25 and 0), scores between 0 and +1·00 (excluding 0 but including +1·00) and scores greater than +1·00. For each interval, the mean percentage of children reported to know the words within that interval was computed for each age (i.e. from 1;4 to 2;6 in 1-month intervals). Only even-numbered ages are displayed on the figures for readability but the odd-numbered ages followed the same general pattern as those displayed. In addition, the slope of the line for each age is displayed

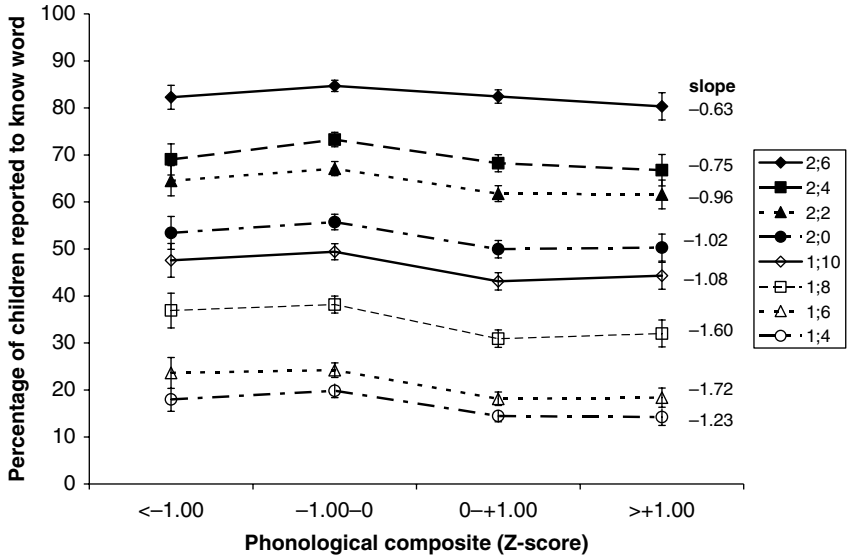

Fig. 1. The mean percentage of children at each age reported to know words on the *MacArthur-Bates Communicative Development Inventory* within a given phonological composite z-score interval.

on the figures so that the change in the relationship between the composite score and the mean percentage of children reported to know the words could be discerned more readily.

As shown in Table 2, the phonological composite did not show a significant interaction with age, indicating that the phonological effect was relatively stable across this infant period. This is illustrated in Figure 1. Recall that the main effect of phonology was that fewer infants knew words composed of higher-probability sound sequences (i.e. positive z-score) than words composed of lower-probability sound sequences (i.e. a negative z-score), leading to a negative slope. Figure 1 shows that children from 1;4 to 2;6 all displayed a negative slope of relatively similar magnitude (i.e. $M = -1{\cdot}10$; $SD = 0{\cdot}38$). This indicates that the magnitude of the effect of phonotactic probability was relatively similar across ages (i.e. the steepness of the slope did not change appreciably and/or consistently with age).

Turning to the lexical composite score, a significant interaction with age was observed (see Table 2). Recall that the main effect for the lexical composite score was that more infants knew shorter words with many neighbors (i.e. a positive z-score) than longer words with few neighbors (i.e. a negative z-score), leading to a positive slope. As shown in Figure 2, the steepness of the slope increases from age 1;4 to 1;8 but then decreases steadily from age 1;8 to 2;6.

https://doi.org/10.1017/S030500090800891X Published online by Cambridge University Press

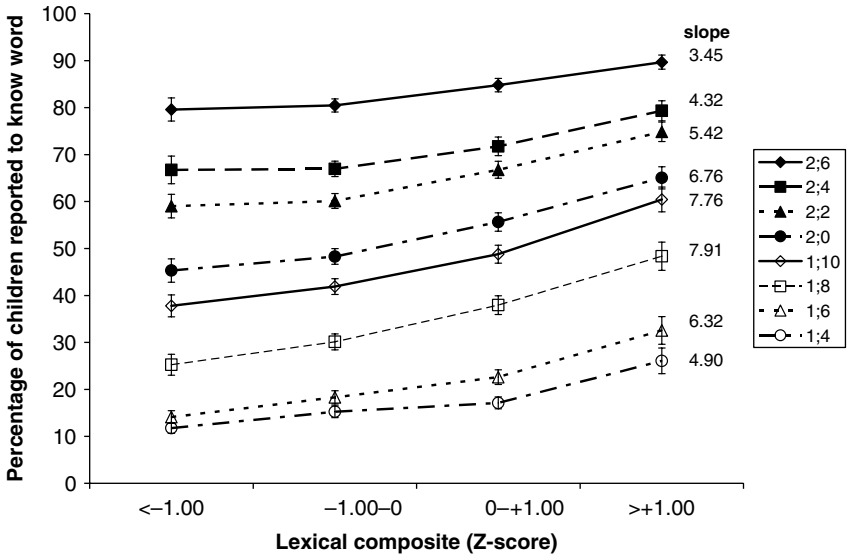

Fig. 2. The mean percentage of children at each age reported to know words on the *MacArthur-Bates Communicative Development Inventory* within a given lexical composite z-score interval.

Word frequency did not show a significant interaction with age (see Table 2). As previously described, the main effect of word frequency was not significant but the trend was for more infants to know high-frequency words (i.e. a positive z-score) than low-frequency words (i.e. a negative z-score), yielding a positive slope. Figure 3 illustrates that the magnitude of the word frequency effect was not predicted by age.

The interaction between the two semantic composites and age was significant. Interestingly, this interaction was the reverse of that found for the lexical composite. Examining the semantic neighbor composite first, recall that the main effect showed that more infants knew words with many interconnected neighbors (i.e. a positive z-score) than words with few interconnected neighbors (i.e. a negative z-score), yielding a positive slope. Figure 4 illustrates that this effect varied by age. Specifically, children aged 1;4 to 1;8 demonstrated the reverse effect with fewer infants knowing words with many interconnected neighbors than words with few interconnected neighbors (i.e. a negative slope). In contrast, more children aged 1;10 and older knew words with many interconnected neighbors than words with few interconnected neighbors (i.e. a positive slope). Moreover, the magnitude of the positive slope generally increased with increasing age from 1;10 to 2;6, although not necessarily in a linear fashion.

https://doi.org/10.1017/S030500090800891X Published online by Cambridge University Press

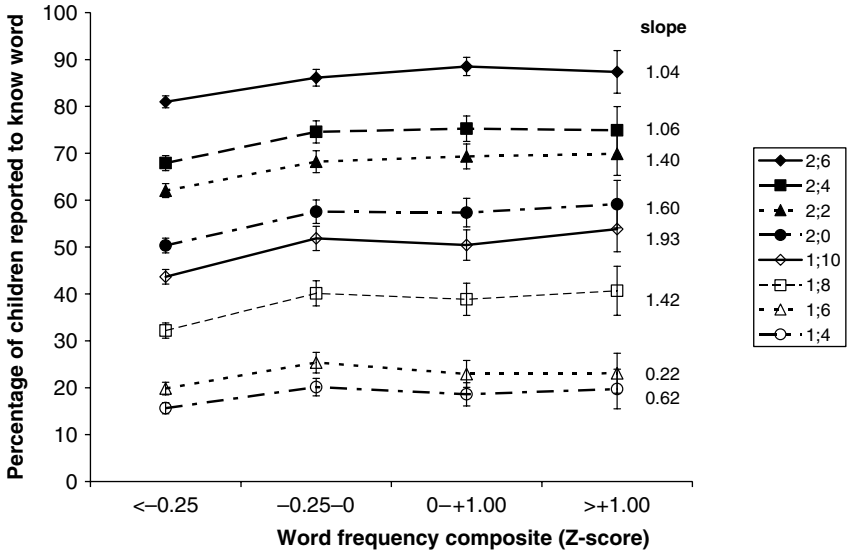

Fig. 3. The mean percentage of children at each age reported to know words on the *MacArthur-Bates Communicative Development Inventory* within a given word frequency z-score interval.

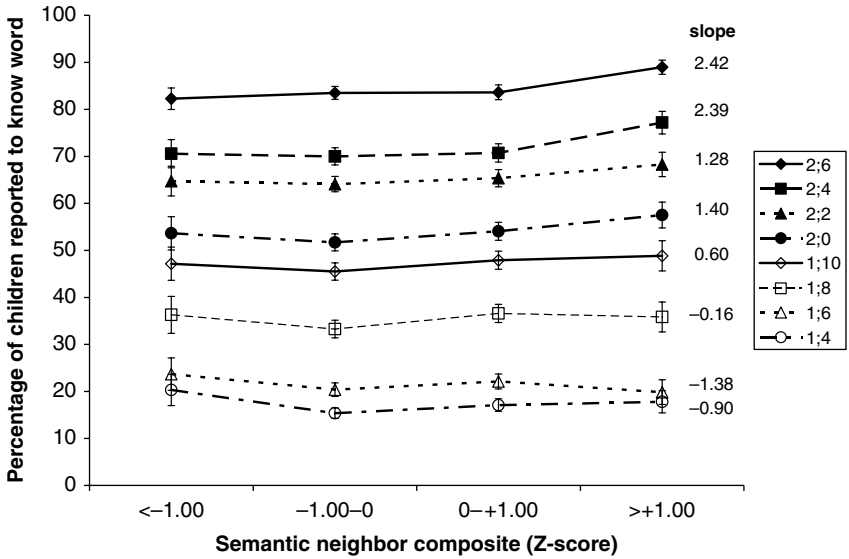

Fig. 4. The mean percentage of children at each age reported to know words on the *MacArthur-Bates Communicative Development Inventory* within a given semantic neighbor composite z-score interval.

https://doi.org/10.1017/S030500090800891X Published online by Cambridge University Press

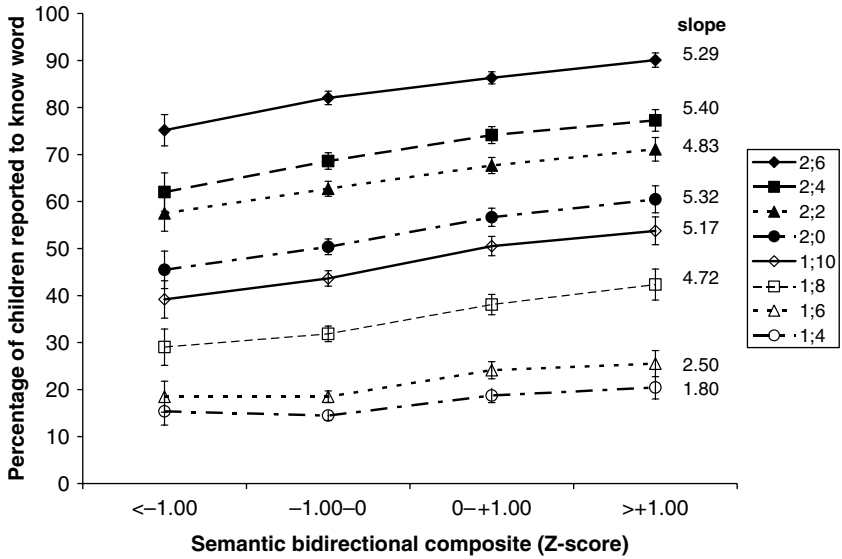

Fig. 5. The mean percentage of children at each age reported to know words on the *MacArthur-Bates Communicative Development Inventory* within a given semantic bidirectional composite z-score interval.

Turning to the semantic bidirectional composite, a significant interaction with age was also observed (see Table 2). The main effect described previously was for more infants to know words with many and stronger target-to-neighbor bidirectional connections (i.e. a positive z-score) than words with fewer and weaker target-to-neighbor bidirectional connections (i.e. a negative z-score), yielding a positive slope. As shown in Figure 5, all ages demonstrated this effect; however, the magnitude of the effect increased with increasing age, as demonstrated by the increasing steepness of the slopes, although not necessarily in a linear fashion.

DISCUSSION

The goal of this study was to differentiate the influence of phonological, lexical and semantic characteristics on word learning and to explore differences in developmental trajectories within and across these three characteristics. Results of the correlation analysis supported the creation of composite scores consisting of one phonological (i.e. positional segment average and biphone average), one lexical (i.e. neighborhood density and word length) and two semantic scores (i.e. semantic set size and connectivity vs. probability resonance and resonance strength). The two semantic factors appeared to relate to overall structure of semantic sets (i.e. number and connectivity of

https://doi.org/10.1017/S030500090800891X Published online by Cambridge University Press

neighbors) versus target-to-neighbor bidirectional connections. In addition, word frequency was correlated with several variables from each domain to some degree, indicating that word frequency indexes multiple characteristics. Most importantly, regression analyses provided evidence that each type of characteristic influences word learning and has a unique developmental trajectory.

Role of phonological, lexical and semantic characteristics in word learning

Results of the regression analysis provided evidence that all three characteristics influence word learning, although the direction of this influence varied. With respect to phonological characteristics, fewer infants knew words composed of high-probability than low-probability sound sequences. This is consistent with findings from a study of adult word learning, where phonotactic probability was differentiated from neighborhood density, and adults were observed to learn fewer high-probability than low-probability sound sequences in an experimental word learning task (Storkel *et al*., 2006). With respect to lexical characteristics, more infants knew words with many lexical neighbors and few phonemes than words with few lexical neighbors and many phonemes. This finding also is consistent with findings from a study of adult word learning where phonotactic probability was differentiated from neighborhood density (Storkel *et al*., 2006). Semantic characteristics also influenced word learning with more infants knowing words with many interconnected semantic neighbors and many strong target-to-neighbor bidirectional connections. Unfortunately, there is no previous research examining the influence of semantic neighbors on word learning. However, the findings of this study suggest that the number of semantic neighbors influences word learning in a manner similar to the number of lexical neighbors. Further investigation of these semantic characteristics warrants study, particularly in laboratory-controlled experiments. Lastly, word frequency was not a significant predictor of word learning but trends were in the expected direction with more infants knowing higher-frequency than lower-frequency words. Results of the correlation analysis supported the a priori hypothesis that word frequency cross-cuts domains by indexing the number of encounters with a word's sound form (i.e. phonological and lexical characteristics) and meaning (i.e. semantic characteristic). Word frequency may not have been a significant predictor of word learning in this study because other variables that more purely indexed phonological, lexical and semantic characteristics were included.

The findings for the phonological and lexical variables require integration with the previous results from Hollich and colleagues (2002). The Hollich and colleague results are both similar and counter to those of the current

study. In terms of similarities, the data are similar across studies. Specifically, Hollich and colleagues provide evidence that low-probability/low-density non-words can be learned more readily than high-probability/high-density non-words. This is similar to the effect of phonological characteristics in the current study (i.e. low-probability words learned at an earlier age than high-probability words). Likewise, Hollich and colleagues provide evidence that high-probability/high-density non-words can be learned more readily than low-probability/low-density non-words. This is similar to the effect of lexical characteristics in the current study (i.e. short high-density words learned at an earlier age than long low-density words). The difference between studies lies in the interpretation. Hollich and colleagues assumed that the low-probability/low-density advantage was attributable to neighborhood density because this effect occurred in the experiment with the greatest pre-exposure, which was hypothesized to tap lexical representations. Similarly, Hollich and colleagues attributed the high-probability/high-density advantage to phonotactic probability because this effect occurred in the experiment with the least pre-exposure, which was hypothesized to tap phonological representations. In the current study, the effect of phonological and lexical characteristics was disentangled through regression analysis rather than interpretation, and this yielded the finding of a low-probability advantage for phonological characteristics and a high-density (short-word) advantage for lexical characteristics.

Although it may seem parsimonious to conclude that Hollich and colleagues' (2002) assumption about the influence of exposure on phonological versus lexical representations was incorrect, this is likely not warranted. There are a number of methodological differences across the two studies that could have lead to differing effects of phonological and lexical characteristics. Specific differences include examination of short-term versus long-term word learning, investigation of the process of word learning versus the products of word learning, and defining phonological and lexical characteristics within the experiment versus within the ambient language. Ultimately, additional data are needed from studies that systematically vary these factors while examining learning of words fully crossed in phonotactic probability and neighborhood density to more clearly determine when and how each variable influences word learning by infants.

Turning to the potential interpretation of the results of this study, one interesting finding is that phonological characteristics had an influence on word learning that differed from that of lexical and semantic characteristics. One possible interpretation of this difference is that the role of phonological characteristics may differ from that of lexical and semantic characteristics. One important step in the word learning process is the initiation of the creation of new representations. When listening to running speech, one must determine whether the words being presented are known or novel.

Presumably, a known word would invoke different processes than a novel word. Specifically, for a known word, the existing lexical and semantic representations of the word would be accessed to support language comprehension (i.e. word recognition processes). For a novel word, a new lexical and semantic representation must be created so that the word can be learned (i.e. word learning processes). If known words and novel words did not invoke different processes, then one would be forced to treat all words as known or novel, likely resulting in inefficiency. Thus, it is possible that there are cues to indicate when a word is novel so that word learning is initiated. One possible cue is phonotactic probability (see also Storkel *et al.*, 2006). A high-probability sound sequence will be more word-like than a low-probability sound sequence (e.g. Vitevitch, Luce, Charles-Luce & Kemmerer, 1997). For this reason, children may be slower to recognize that a high-probability sound sequence is novel and thus slower to initiate the creation of new lexical and semantic representations. In contrast, a low-probability sound sequence may stand out as unique, facilitating recognition that the sound sequence is novel and immediately initiating the creation of new lexical and semantic representations. Thus, phonological characteristics may influence how quickly a sound sequence is detected as novel and how soon word learning is initiated. It is likely that other cues also exist that facilitate this initiation of word learning (e.g. uniqueness of the referent). Further investigation of the characteristics that trigger word learning is warranted.

A second important step in word learning is the integration of new lexical and semantic representations with existing representations. That is, it is thought that relationships between words in the lexicon are indexed by connections among similar lexical representations and among similar semantic representations. It is not enough to create a lexical and semantic representation of a word. One also must learn how the new representation relates to existing representations and must establish the relevant connections between the new and existing representations. Based on the current results, one might infer that establishing connections with many existing representations, as would occur when there are many lexical or many semantic neighbors, strengthens the new lexical or semantic representation through spreading activation (see also Storkel *et al.*, 2006).

Moreover, the inclusion of additional semantic variables provides evidence that other aspects of semantic structure, in addition to the number of related representations, may strengthen new semantic representations. Specifically, having known neighbors that were connected to many other known neighbors may strengthen the new semantic representation of the novel word. Likewise, reciprocal connections between the novel word and the known neighbors also appeared to influence word learning. These types of relationships warrant further investigation in the lexical domain to determine whether analogous patterns are observed.

Developmental changes in the role of phonological, lexical and semantic characteristics

The regression analysis of developmental patterns also provided evidence that the influence of phonological characteristics differed from that of lexical and semantic characteristics. Specifically, phonological characteristics appeared to exert a relatively constant influence on word learning from 1;4 to 2;6, whereas the influence of lexical and semantic characteristics changed during infancy. Interestingly, different developmental patterns were observed across lexical and semantic characteristics. Beginning with the lexical developmental pattern, the influence of lexical characteristics on acquisition increased from age 1;4 to 1;8 but then diminished from 1;8 to 2;6. Turning to the semantic developmental pattern, different patterns were observed for semantic neighbors as compared to semantic bidirectional connections. In particular, semantic neighbor characteristics showed a change in the direction of the effect on acquisition between 1;8 and 1;10. That is, from 1;4 to 1;8 the slope was negative, whereas from 1;10 to 2;6 the slope was positive and generally increasing. In contrast, the influence of semantic bidirectional connections on word learning increased across 1;4 to 2;6. These findings are globally similar to those of other studies that have shown constant phonological effects, diminishing lexical effects and increasing semantic effects with increasing age (Gershkoff-Stowe & Smith, 2004; Newman & German, 2005; Smith, 2000).

The role of phonological characteristics in word learning appears to be established early in development and changes only minimally throughout the infant period. Children extract the phonotactic probability of the ambient language by about 0;9 (e.g. Jusczyk *et al.*, 1994). The results of this study further suggest that this sensitivity to phonotactic probability is harnessed for word learning relatively early in development and continues to be used throughout infancy with minimal change (i.e. 1;4 to 2;6). Combining this observation with the previous hypothesis that phonological characteristics aid in initiating word learning suggests the additional hypothesis that the cues relevant to the initiation of word learning may be among the first to be established and may change minimally as age increases.

Previously, it was hypothesized that establishing connections with many existing lexical or semantic neighbors may strengthen a newly created lexical or semantic representation, facilitating word learning. The observed developmental patterns suggest that this process changes across age in different ways for lexical versus semantic representations. In terms of lexical representations, it appears that the facilitatory effect of creating many lexical connections increases from 1;4 to 1;8 but then slowly diminishes. Interestingly, the age of 1;8 also marks a change in the influence of

semantic connections on word learning. Specifically, there is an inhibitory effect of creating many semantic connections from 1;4 to 1;8 (i.e. negative slope for semantic neighbor composite), followed by an increasingly facilitatory effect of creating many semantic connections from 1;10 to 2;6 (i.e. increasing positive slope for semantic neighbor composite). Other studies also have noted changes in lexical and semantic representations around age 1;8. Specifically, infants aged 1;2 experience difficulty learning phonologically similar words, whereas infants aged 1;5 can learn phonologically similar words and this ability continues to improve through age 1;8 (Werker, Fennell, Corcoran & Stager, 2002). Likewise, semantic errors in naming appear to increase dramatically from approximately age 1;4 to 1;7 (Gershkoff-Stowe, 2002). Across both types of representations, it is assumed that word learning processes and/or representations are fragile during this period prior to 1;8, leading to difficulties learning and accessing similar words (Gershkoff-Stowe, 2002; Werker *et al.*, 2002). Applying this hypothesis to the current data, the benefit of forming many connections with similar lexical representations may increase as word learning processes and lexical representations strengthen. In a similar vein, the benefit of forming many connections with similar semantic representations may not emerge until word learning processes and semantic representations are strong.

In terms of continued change after age 1;8, it is possible that changes in vocabulary size may be relevant. That is, as the number of known words increases, the number of lexical and semantic connections that need to be formed between the representation of a new word and the representations of existing known words increases. It is possible that this increase in the number of connections that needs to be formed between new and existing representations may have a different impact on lexical versus semantic representations. For lexical representations, there may be an asymptote on the benefit that can be derived from the number of connections formed with existing lexical representations. That is, connections with a certain number of existing lexical representations may strengthen the new lexical representation, but adding more existing representations does not provide any added strength. In contrast, the reverse pattern may be true for semantic representations. Specifically, adding connections with more existing semantic representations may further strengthen the new semantic representation with no upper limit.

The developmental pattern for semantic bidirectional connections was somewhat similar to that of semantic neighbor characteristics. Specifically, the effect of semantic bidirectional connections on word learning generally increased from 1;4 to 2;6 and the largest change in slope occurred between age 1;6 and 1;8. Therefore, a similar account of developmental changes in the effect of semantic bidirectional connections may be applicable. However, it is important to note that the change at 1;8 is somewhat less

clear than for lexical characteristics and semantic neighbor characteristics, warranting further exploration.

Limitations

While this study does provide insights into the influence of phonological, lexical and semantic characteristics on word learning by infants, there are several limitations that are important to keep in mind. The first limitation is that the CDI reflects the outcome of word learning (i.e. the words that a child currently knows) rather than the word learning process itself (i.e. the course of acquiring those words). Yet, preliminary inferences were made about the word learning process based on this examination of the products of word learning. These inferences should be viewed as tentative given that the word learning process was not directly observed. However, these tentative hypotheses provide suggestions for future research directed towards the word learning process itself. In addition, the CDI represents naturalistic data. As such, other variables that were not measured or analyzed in the current study likely varied and may have influenced the number of children reported to know a word at a given age. For example, past studies have demonstrated that word learning is influenced by phonological development (e.g. Schwartz & Leonard, 1982), yet this was not examined in the current study. It is possible that the effects reported here might be attenuated if other variables were included in the regression analysis.

The second limitation is that the CDI does not provide information about an infant's underlying representation of words. This is important because the underlying representation of both mispronounced and correctly pronounced words is unclear. Some argue that infants have target appropriate representations of words they mispronounce (e.g. Smith, 1973), whereas others argue that underlying representations of mispronounced words may not be target appropriate (e.g. Macken, 1980). Similar arguments also are levied against correctly pronounced words where some argue that infants encode phonetic detail (e.g. Bailey & Plunkett, 2002; Swingley & Aslin, 2002), whereas others argue that underlying representations may be holistic, lacking phonetic detail (e.g. Jusczyk, Goodman & Baumann, 1999). Likewise, it has also been suggested that the quality of semantic representations may change over time (e.g. Gershkoff-Stowe, 2002). At issue here is whether computations of the characteristics of the ambient language really reflect the characteristics that the child represents underlyingly. Stated in this way, the issue is primarily methodological. One might argue that since significant effects were obtained in this study, use of ambient language characteristics at least partially indexes the child's underlying knowledge of the language. This would be consistent with other studies examining the influence of ambient language characteristics in preverbal infants (e.g. Jusczyk *et al.*,

1994). A more theoretical approach to the issue would be to ask what data children use to determine language characteristics such as phonotactic probability, neighborhood density and semantic set size: their own internalized knowledge or the input they receive from the environment. To date, this issue has only been considered for phonotactic probability in preschool children, and the results suggest that phonotactic probability is based on input rather than internalized knowledge (Storkel, 2004a). Clearly, this would be a fruitful area for future inquiry.

A third limitation is that the computation of all predictor variables was based on adult corpora. There were two reasons for this. One was that the research questions related to the influence of ambient language characteristics on word learning. The second was that it is rather difficult to obtain discrete association norms from infants, and these were needed to compute the semantic variables. Thus, all predictor variables were based on adult corpora for uniformity across variables. Past work suggests a significant correlation between adult-based calculations and child-based calculations for these predictor variables, indicating that the ranking of the words by these calculations would likely be similar across adult and child counts. Therefore, the patterns identified in the regression analyses, which relied on this type of ranking rather than absolute values (as would be examined in other statistical analyses such as *t*-tests), may still hold if child counts were used. However, this method clearly introduces some amount of error into the data. Thus, the effects reported might be strengthened or attenuated if child counts were used for all variables.

CONCLUSION

Phonological characteristics appear to influence word learning in a way that is distinct from lexical and semantic characteristics and does not appear to change across infancy. Thus, models of word learning need to consider what mechanism could account for this unique influence of phonological characteristics. One possibility is that phonological characteristics may be critical in initiating the learning process, whereas lexical and semantic characteristics may influence the strength of newly created word representations. Moreover, the influence of lexical and semantic characteristics varied across development with different patterns being observed prior to versus after 1;8. This suggests that models of word learning need to address how learning words might alter the word learning process and how to account for differences in lexical versus semantic influences on word learning. In addressing this issue, it will be important to consider how changes in the quality of representations within the lexicon and in the size of the lexicon may alter the influence of existing lexical and semantic representations on word learning.

REFERENCES

Bailey, T. M. & Plunkett, K. (2002). Phonological specificity in early words. *Cognitive Development* **17**, 1265–82.

Buchanan, L., Westbury, C. & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin and Review* **8**, 531–44.

Charles-Luce, J. & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language* **17**, 205–215.

Coady, J. A. & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child Language* **30**(2), 441–70.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dale, P. S. & Fenson, L. (1996). Lexical development norms for young children. *Behavioral Research Methods, Instruments & Computers* **28**, 125–27.

Entwisle, D. R. (1966). *Word associations of young children*. Baltimore, MD: The John Hopkins Press.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., Pethick, S. & Reilly, J. S. (1993). *The MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego: Singular Publishing Group.

Gershkoff-Stowe, L. (2002). Object naming, vocabulary growth, and development of word retrieval abilities. *Journal of Memory and Language* **46**, 665–87.

Gershkoff-Stowe, L. & Smith, L. B. (2004). Shape and the first hundred nouns. *Child Development* **75**(4), 1098–1114.

Gierut, J. A. & Dale, R. A. (2007). Comparability of lexical corpora: Word frequency in phonological generalization. *Clinical Linguistics & Phonetics* **21**(6), 423–33.

Gupta, P. & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language* **59**, 267–333.

Hollich, G., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., Hennon, E. & Rocroi, C. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development* **65**, v–123.

Hollich, G., Jusczyk, P. W. & Luce, P. A. (2002). Lexical neighborhood effects in 17-month-old word learning. In B. Skarabela, S. Fish & A. H.-J. Do (eds), *Proceedings of the 26th annual Boston University Conference on Language Development*, Vol. 1, 314–23. Sommerville, MA: Cascadilla.

Jusczyk, P. W., Goodman, M. B. & Baumann, A. (1999). Nine-month-olds' attention to sound similarities in syllables. *Journal of Memory and Language* **40**, 62–82.

Jusczyk, P. W., Luce, P. A. & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* **33**, 630–45.

Kucera, H. & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University.

Landauer, T. K. & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior* **12**, 119–31.

Leonard, L. B., Schwartz, R. G., Chapman, K., Rowan, L. E., Prelock, P. A., Terrell, B., Weiss, A. L. & Messick, C. (1982). Early lexical acquisition in children with specific language impairment. *Journal of Speech and Hearing Research* **25**, 554–64.

Luce, P. A. & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing* **19**, 1–36.

Macken, M. A. (1980). The child's lexical representation: The 'puzzle-puddle-pickle' evidence. *Journal of Linguistics* **16**, 1–17.

Nelson, D. L., McEvoy, C. & Schreiber, T. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Retrieved from: www.usf.edu/FreeAssociation/.

Nelson, D. L., McKinney, V., Gee, N. & Janczura, G. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review* **105**, 299–324.
Nelson, D. L. & Zhang, N. (2000). The ties that bind what is known to the recall of what is new. *Psychonomic Bulletin and Review* **7**, 604–617.
Nelson, D. L., Zhang, N. & McKinney, V. (2001). The ties that bind what is known to the recognition of what is new. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **27**, 1147–59.
Newman, R. S. & German, D. J. (2005). Life span effects on lexical factors in oral naming. *Language and Speech* **48**(2), 123–56.
Pisoni, D. B., Nusbaum, H. C., Luce, P. A. & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication* **4**, 75–95.
Rice, M. L., Oetting, J. B., Marquis, J., Bode, J. & Pae, S. (1994). Frequency of input effects on word comprehension of children with specific language impairment. *Journal of Speech and Hearing Research* **37**, 106–122.
Schwartz, R. G. & Leonard, L. B. (1982). Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language* **9**, 319–36.
Smith, L. B. (2000). Learning how to learn words: An associative crane. In R. M. Golinkoff, K. Hirsh-Pasek, N. Akhtar, L. Bloom, G. Hollich, K. Plunkett, L. Smith, M. Tomasello & A. Woodward (eds), *Becoming a word learner: A debate on lexical acquisition*, 51–80. London: Oxford Press.
Smith, N. V. (1973). *The acquisition of phonology: A case study*. Cambridge: Cambridge University Press.
Soja, N. N., Carey, S. & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. *Cognition* **38**, 179–211.
Storkel, H. L. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research* **44**, 1321–37.
Storkel, H. L. (2004a). The emerging lexicon of children with phonological delays: Phonotactic constraints and probability in acquisition. *Journal of Speech, Language, and Hearing Research* **47**(5), 1194–212.
Storkel, H. L. (2004b). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language and Hearing Research* **47**(6), 1454–68.
Storkel, H. L., Armbruster, J. & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research* **49**(6), 1175–92.
Storkel, H. L. & Rogers, M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition. *Clinical Linguistics & Phonetics* **14**, 407–425.
Swingley, D. & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science* **13**, 480–84.
Vitevitch, M. S. & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory of 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.
Vitevitch, M. S., Luce, P. A., Pisoni, D. B. & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language* **68**, 306–311.
Webster's Seventh Collegiate Dictionary (1967). Los Angeles: Library Reproduction Service.
Werker, J. F., Fennell, C. T., Corcoran, K. M. & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy* **3**, 1–30.
Yuan, Y. C. (2002). *Multiple imputation for missing data: Concepts and new development*. Rockville: SAS Institute Inc.