The company that words keep: comparing the statistical structure of child- versus adult-directed language

THOMAS HILLS

Journal of Child Language / Volume 40 / Issue 03 / June 2013, pp 586 - 604
DOI: 10.1017/S0305000912000165, Published online: 14 May 2012

J. Child Lang. **40** (2013), 586–604. © Cambridge University Press 2012
doi:10.1017/S0305000912000165

The company that words keep: comparing the statistical structure of child- versus adult-directed language*

THOMAS HILLS
University of Warwick

(*Received* 25 *August* 2010 – *Revised* 24 *November* 2011 – *Accepted* 10 *April* 2012 – *First published online* 14 *May* 2012)

ABSTRACT

Does child-directed language differ from adult-directed language in ways that might facilitate word learning? Associative structure (the probability that a word appears with its free associates), contextual diversity, word repetitions and frequency were compared longitudinally across six language corpora, with four corpora of language directed at children aged 1;0 to 5;0, and two adult-directed corpora representing spoken and written language. Statistics were adjusted relative to shuffled corpora. Child-directed language was found to be more associative, repetitive and consistent than adult-directed language. Moreover, these statistical properties of child-directed language better predicted word acquisition than the same statistics in adult-directed language. Word frequency and repetitions were the best predictors within word classes (nouns, verbs, adjectives and function words). For all word classes combined, associative structure, contextual diversity and word repetitions best predicted language acquisition. These results support the hypothesis that child-directed language is structured in ways that facilitate language acquisition.

A central problem of child language acquisition is determining what words mean. Closely related to this problem is determining whether or not adults, when speaking to children, alter the structure of their language in ways that might facilitate children's learning of meaning. Prior research has found that child-directed language differs from adult-directed language, for example, in terms of phonology, grammatical complexity, number of word

[*] I thank Linda Smith, Josita Maouene, Brian Riordon, Peter Todd, Thorstun Pachur and Sara Hills for suggestions and comments on the research and manuscript. I also thank Fionniain Hills for the banjo–violin example. This work was supported by a grant from the Swiss National Science Foundation (100014 130397/1). Address for correspondence: Thomas Hills, University of Warwick, Department of Psychology, Gibett Hill Road, Coventry CV4 7AL, UK. tel: +44-(0) 24-7652-3183. e-mail: t.t.hills@warwick.ac.uk

repetitions, and the use of lexical substitution such as saying *choo-choo* for *train* (e.g. Ferguson, 1964; Hayes & Ahrens, 1988; Newport, Gleitman & Gleitman, 1977; Snow, 1972). These have been discussed using terms such as 'motherese', 'parentese' and 'baby talk', referring to the potential for language directed to children to better facilitate language acquisition, as compared with language directed to other adults (e.g. Newport *et al.*, 1977). It is unclear, though, how the above documented changes in child-directed language communicate meaning. For example, repeating a word may make the word more salient, but it does not necessarily help disambiguate a word's intended meaning from the vast number of possibilities. Moreover, previous analyses documenting changes in child-directed language have focused primarily on the learning of individual words, and thus it is further unclear to what extent the large-scale statistical structure of child-directed language changes in relation to the words being learned.

Investigating the statistical structure of language is critical to understanding how meaning is learned because dominant theories of the construction of meaning rely on interrelatedness between concepts. For example, STRUCTURAL LINGUISTICS is based on the claim that meaning is created via relations between concepts. According to Saussure (1916/1959), ''language is a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others'' (p. 116).[1] Thus, the concept of **Hot** takes part of its meaning from its relationship with concepts like **Cold**, **Sun**, **Stove** and **Burn**, which in turn depend on still other concepts. Similar views on the inter-relatedness between concepts are well represented in cognitive science, e.g. in philosophy (Block, 1999), computer science (Lenat & Feigenbaum, 1991), and psychology (Goldstone, Steyvers & Rogosky, 2003; Jones & Mewhort, 2007). In each case, meaning is embedded in the structural relationships among concepts.

Though the structural basis of meaning is well represented in cognitive science, it has yet to be fully explored in the realm of language acquisition. The goal of the current study is to investigate the linguistic structure around early-learned words across multiple corpora of language directed at individuals who range from age 1;0 to adults. This structure is evaluated with respect to statistics traditionally associated with language acquisition (frequency and word repetitions) as well as attributes that may contribute to acquiring the meaning of children's earliest learned words, specifically ASSOCIATIVE STRUCTURE and CONTEXTUAL DIVERSITY (e.g. Hills, Maouene, Riordan & Smith, 2010). Associative structure is a measure of how often a word appears near its associates (as measured by adult free association norms; see below) in natural language. For example, how often do

[1] The title for this article is inspired by a quote with a similar meaning from J. R. Firth (1957: 11): ''You shall know a word by the company it keeps!''

associates of the word *hot* – like *sun*, *stove* and *burn* – appear near the word *hot* in natural language. Contextual diversity, on the other hand, measures how many unique word types a word appears near in natural language; this measures the linguistic diversity in which a word is embedded.

The present study aims to provide a longitudinal perspective on how the statistical structure of language changes – with respect to frequency, repetitions, associative structure and contextual diversity – as language is directed at progressively older individuals. Before describing this study in more detail, the relationships between associative structure, contextual diversity and meaning in child language acquisition will be briefly reviewed.

STRUCTURE AND MEANING

Linguistic structure has been shown to provide meaning in at least two ways. The first of these is in terms of the associative relations between words. When two concepts frequently co-occur, one concept comes to predict the presence of the other. This is the basis of semantic space theories, in which semantic relationships are formed using word co-occurrence in natural language. Words either appear together or appear with similar OTHER words and thus come to share a similar semantic role (e.g. Jones & Mewhort, 2007). Early theories of semantic networks share a similar logic, where word meaning is acquired via relations with other words (Collins & Quillian, 1969). Nodes are represented by concepts and relationships between them are represented by links, such that words like *bird* become associated with *feathers* and *animal* by links such as **Has** and **Is-a**.

There is growing evidence that these associative relations may be related to early language acquisition. One way to measure associative relations is via free association norms, where one word is provided (the cue) and participants are asked to produce the first word that comes to mind (the target). Several studies have now shown that the number of cue words that produce a given target word in free association norms is correlated with children's age of acquisition of that target word (Hills, Maouene, Maouene, Sheya & Smith, 2009; Hills *et al.*, 2010; Steyvers & Tenenbaum, 2005). In other words, if, in response to a broad range of cues, the word *ball* was produced more often as a target than *zebra*, then *ball* will tend to have an earlier age of acquisition. Though the causal direction here is undetermined, it is interesting to note that if adult caregivers amplify associative structure around a word, this may indicate that they are, either prior to or as a result of their child's learning, elaborating on the meaning of these earliest learned words. This increase in associative structure would thus represent an additional feature of motherese.

A second way in which structure contributes to meaning is via contextual diversity, which may help establish exclusive meanings for words. In any

given instance where a word is used, the mapping between the word (the signifier) and the thing to which it refers (the signified) is potentially unlimited. Quine (1960) described this as THE PROBLEM OF INDETERMINACY. Indeterminacy leads to common errors of mismapping and generalization, such as overextension. As an example of overextension, a four-year-old child recently asked me if he could have a banjo that you play ''like this'', and he made the motion for playing a violin. According to Saussure (1916/1959), co-occurrence helps to solve such problems of indeterminacy by establishing what a word does NOT mean, that is, words are ''negatively defined''. Saussure asserted that ''concepts are purely differential and defined not by their positive content but negatively by their relations with other terms of the system. Their most precise characteristic is in being what the others are not'' (p. 117). Thus, if *banjo* and *violin* appear near one another in language, then they are unlikely to mean the same thing.

Explanations similar to Saussure's have been proposed for child language development, such as the PRINCIPLE OF CONTRAST (Clark, 1990) and MUTUAL EXCLUSIVITY (Markman, 1984), which state that no two words may share the same meaning. The proposal is that children learn to assign new words that they hear to objects for which they do not already know a name. Indeed, children do tend to map novel words onto novel objects in the environment (e.g. Mather & Plunkett, 2012; Vincent-Smith, Bricker & Bricker, 1979).

Contextual diversity potentially results in learning across contexts, sometimes called CROSS-SITUATIONAL LEARNING, which is a time-iterated form of the above exclusion principle. Here potential targets can be excluded because they do not appear in all contexts in which the word is heard; words that appear in different contexts can only map onto objects that are consistent across those contexts (Figure 1).

For children to benefit from the disambiguating power of cross-situational learning, they need to remember potential word–meaning mappings across situations. Using an artificial word learning task, Smith and Yu (2008) demonstrated that children as young as 1;0 have this capacity. By presenting children with pairs of objects and pairs of novel words, Smith and Yu (2008) demonstrated that children could learn to map these words onto their appropriate objects by noting how words consistently appeared with certain objects. In this case, diversity across contexts (i.e. contextual diversity) allowed the children to disambiguate the true word–object mapping from its background of other possible mappings.

A host of studies have established that both high and low contextual diversity can enhance word learning in children and adults. For example, increased contextual diversity has been shown to facilitate word segmentation (e.g. Hayes & Clark, 1970; Newman, 2008; Saffran, Aslin & Newport, 1996; Saffran, Newport & Aslin, 1996), artificial language learning (Gillette, Gleitman, Gleitman & Lederer, 1999; Kachergis,

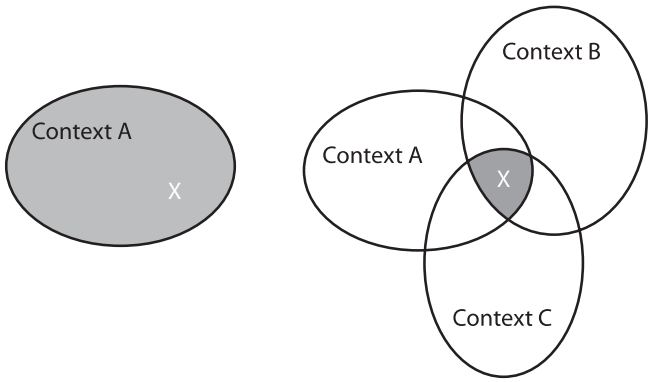

Fig. 1. Contextual diversity disambiguates potential word–meaning mappings. Hearing the word X always in Context A does not assist in eliminating what in context A is not an appropriate target for X (appropriate targets are represented by the area in grey). However, if word X is heard in multiple contexts, then only those targets that are common to all contexts persist as potential word–meaning mappings.

Yu & Shiffrin, 2012; Plaut & Kello, 1999; Recchia, Johns & Jones, 2008; Yu & Smith, 2007), and rapid and diverse vocabulary acquisition (e.g. Hoff & Naigles, 2002; Hurtado, Marchman & Fernald, 2008; Huttenlocher, Haight, Bryk, Seltzer & Lyons, 1991; Naigles & Hoff-Ginsberg, 1998; Rowe, 2008). On the other hand, consistency (i.e. low contextual diversity) across contexts has been shown to play an important role for learning verbs and adjectives (e.g. Brown, 2008; Waxman & Klibanoff, 2000). In principle, consistency may facilitate word learning among other word classes as well.

The source of this discrepancy between consistency and diversity is not clear. It may represent peculiarities in language use for different words or even different languages (e.g. Maouene, Laakso & Smith, 2011). There may also be a trade-off between the contextual diversity needed to facilitate the identification of phonemic and semantic boundaries (i.e. the negative definition), and the consistency needed to facilitate appropriate word usage in early language learners. Prior work investigating the statistical structure of child-directed language corpora found that diversity in word co-occurrence was a predictor of order of acquisition of children's earliest learned words (Hills *et al.*, 2010). Words embedded in more linguistically diverse speech were learned earlier. However, though this research did compete contextual diversity against frequency in a language-learning model, it failed to address the extent to which language structure was altered in child-directed speech relative to what one would expect in a completely unstructured (i.e. random) language. Thus, while contextual diversity may facilitate language acquisition to some degree, adults may nonetheless increase the consistency of the earliest learned words when

speaking to children, relative to the consistency of these same words if they were spoken at random.

The two roles of structure addressed above (associative structure and contextual diversity) are hypothesized to be related in child-directed language. In particular, Hills *et al.* (2010) have shown that the contextual diversity of words in child-directed language is highly correlated with the frequency with which a word is an associative target in the University of South Florida Free Association Norms (Nelson, McEvoy & Schrieber, 1998). This relationship between the structure of associates in our heads, their co-occurrence in our speech, and their predictive power for early word acquisition led to the prediction of a potential ASSOCIATIVE STRUCTURE IN CHILD-DIRECTED LANGUAGE (Hills *et al.*, 2010). The claim is that early word learning may be driven in part by contextual diversity in child-directed speech, which is in turn driven by the cue–target structure in adult free associations. However, to say that child-directed language is associative is to imply that it is MORE associative than adult-directed language, and specifically in ways that correlate with early language acquisition. Testing this hypothesis is one of the focal aims of the present study.

THE PRESENT STUDY

The present study evaluated the structure of language across multiple language corpora, composed of language directed at different aged individuals, ranging from children to adults. Using a co-occurrence analysis, a window was moved word-wise across each corpus to evaluate statistics such as contextual diversity around words, associative structure (i.e. the probability that a word appeared with its free associates), word repetitions and frequency. To remove the effects of different word frequency distributions in the different corpora, word statistics were adjusted using randomly shuffled versions of each corpus. These data allow us to address which features of child-directed language (associative structure, contextual diversity, repetitions and frequency) are altered relative to adult-directed language. Further, by comparing word statistics with their age of acquisition, these data allow us to determine whether or not differences in the structure of child- and adult-directed language are consistent with child-directed language better facilitating early language acquisition.

METHOD

Corpora and free associates

A corpus taken from the American section of the CHILDES database (MacWhinney, 2000), representing transcripts of caregiver speech directed

to children aged 1;0 to 5;0 (provided by Riordan & Jones, 2007; see also Riordan & Jones, 2011), were divided into four consecutive twelve-month periods. These four corpora consisted of approximately 500,000 words each, consisting of only adult speech. Two adult-directed corpora were chosen to capture the diversity of adult-directed language. One was the Santa Barbara Corpus of Spoken American English (SBC), parts 1–4, which contained approximately 250,000 words, taken from sixty discourse recordings between adults, ranging from personal conversations to university lectures (Du Bois & Englebretson, 2000–2005). The SBC corpus is similar to the CHILDES corpus in that it represents largely conversational speech in American English. The other corpus was the written text corpus of Touchstone Applied Science Associates (TASA) used by Landauer and Dumais (1997), which contains approximately 10 million words of fiction and non-fiction, from over 6,000 textbooks used in schools in the United States. The SBC and TASA corpora, therefore, represent potential extremes in the usage patterns of American English directed to individuals above the age of early language acquisition.

The free associations used were the University of South Florida Free Association Norms (FAN), consisting of approximately 5,000 word cues, for which participants were asked to provide the first word that came to mind (Nelson *et al.*, 1998). For the present analysis, the same words were used as targets. Nelson *et al.* (1998) did not select the cue words in the free association norms systematically; there was no single criterion for their inclusion. However, many of these cue words were selected because they were target words of other cues. Thus, the list of cue words provides 75% of the targets for these cue words. The FAN cue words represent 79% (Age 1), 78% (Age 2), 79% (Age 3), 77% (Age 4), 71% (SBC), and 60% (TASA) of the words produced in each of the corpora. All FAN words that appeared in a given corpus were analyzed for that corpus. This is because we are interested in how often words appear with their associates, not necessarily how often words spoken dominantly to one group appear with their associates when spoken to another group. Should the SBC and TASA corpora be more similar to one another than to child-directed speech, they offer a degree of confidence in the relative properties of adult- and child-directed language. The FAN was used to contruct a free association matrix, $\boldsymbol{F}$, where a 1 is in cell ij if the word j is a target of the cue word i in the FAN.

The age of acquisition analysis focused on words from the FAN that overlapped with the MacArthur-Bates Communicative Development Inventory (MCDI: Dale & Fenson, 1996), Toddler version. The MCDI includes data on the normative productive vocabularies of children – the words children say – in one-month increments from age 1;4 to 2;6. The 562 words used were 337 nouns, 96 action words (verbs), 59 descriptive words

TABLE 1. *Sample co-occurrence matrix, containing each unique word once, for the sentence "That dog is friends with that dog" using a window of size* 3

	that	dog	is	friends	with
that	0	2	1	0	0
dog	0	0	1	1	0
is	0	0	0	1	1
friends	1	0	0	0	1
with	1	1	0	0	0

(adjectives), and 70 function words consisting of pronouns, quantifiers, articles, helping verbs and connecting words. The age of acquisition for each cue word was set to the first month at which the word was produced by more than 50% of the children in the normative tables of the MCDI. For each of the corpora, the MCDI words represent 60% (Age 1), 59% (Age 2), 60% (Age 3), 57% (Age 4), 49% (SBC), and 37% (TASA) of all words produced.

ANALYSIS OF STRUCTURE

A method similar to the Hyperspace Analogue to Language (HAL) (Lund & Burgess, 1996) and the word co-occurrence detector (Li, Farkas & MacWhinney, 2004) was used to produce a measure of associative structure, word repetitions and contextual diversity. Each of these measures was computed with respect to a constant window size that moved word-wise through the corpus. Based on previous work (Hills *et al.*, 2010), the window size was set at 5 for the analyses presented here. Using a window size of 10 was found to produce similar results. The ASSOCIATIVE STRUCTURE of a word (the cue) represents the probability that a target associate followed that cue word within the window. REPETITIONS indicate the probability that a word was repeated within the window. The CONTEXTUAL DIVERSITY of a word represents the number of unique word types that followed a word within the window per appearance of that word (see Hills *et al.*, 2010).

In what follows, the formal definition of each measure is provided and then an example using the sentence in Table 1, "That dog is friends with that dog," using a window size of 3. Words are repeated in the sentence to demonstrate that the matrix does not contain the same number of word tokens as the corpus, but instead contains the number of unique word types in the corpus. As an example of the window size, for the word *dog*, the words *is* and *friends*, follow within two words, and are therefore in a three-word window with *dog*.

For each corpus a matrix, $\boldsymbol{C}$, (of the same dimensions as the free association matrix, $\boldsymbol{F}$, above) was formed, where each cell, *ij*, was filled according

to the following rule: a moving window of size five moved word-wise through the corpus, with the INITIAL word, *i*, adding a unit of 1 to cell *ij* if the word *j* was in the window simultaneously with *i*.

The associative structure was computed by taking the Hadamard product (entry-wise matrix multiplication) of the co-occurrence matrix, $\boldsymbol{C}$, with the FAN matrix, $\boldsymbol{F}$, producing an associative co-occurrence matrix, $A = F \circ C$. In words, non-zero entries in $\boldsymbol{A}$ only remain if they represent cue–target relations in the free association norms. Because associative structure is the probability that a target follows the occurrence of the cue, the associative structure for word *i* is the row plus column sum from the associative co-occurrence matrix, $\boldsymbol{A}$, divided by the frequency of occurrence of the word, *i*, in the corpus. For example, in Table 1, the word *dog* would have an associative structure of 0·5 because the cell (**Dog, Friends**), corresponding to its only associate in the free association norms (**Friends**), contains a 1, which is one half of the number of times the word *dog* appeared.

REPETITIONS for a word, *i*, were computed by taking the diagonal, $\boldsymbol{C}_{i,i}$, of the co-occurrence matrix – which is the number of times a word followed itself within the window of observation – and dividing by the total frequency of occurrences of the word *i*. This provides the probability that a word, once occurring, was repeated in the window. In Table 1 none of the words are repeated within the three-word window, which can be seen by noting that the diagonal of the matrix is everywhere zero.

The contextual diversity was computed as the sum of the row for a given cue word in $\boldsymbol{C}$, after constraining all non-zero cells to one, divided by the frequency of occurrence of the cue word. For example, in Table 1, the word *that* appeared twice and was followed by two other word types, *dog* and *is*, and would have a contextual diversity of 1 in Table 1. Using a symmetric matrix shows the same pattern of results as those reported below.

Child-directed speech is potentially composed of more high-frequency words than adult-directed speech (see Hayes & Ahrens, 1988). This change in the distribution of words spoken to adults and children may lead to artificial changes in the associative structure, repetitions and contextual diversity. To control for the effects of frequency, the statistics for associative structure, repetitions and contextual diversity were each adjusted by subtracting out the proportion of their effect that was generated in a randomized (i.e. shuffled) corpus. To do this, each of the original corpora was shuffled so that words appeared in a random order. Then statistics for each of the above variables was computed using the shuffled corpora. The results of these computations were then subtracted from the results obtained from the non-shuffled corpora. All of the analyses presented below use the adjusted statistics, except where otherwise noted.

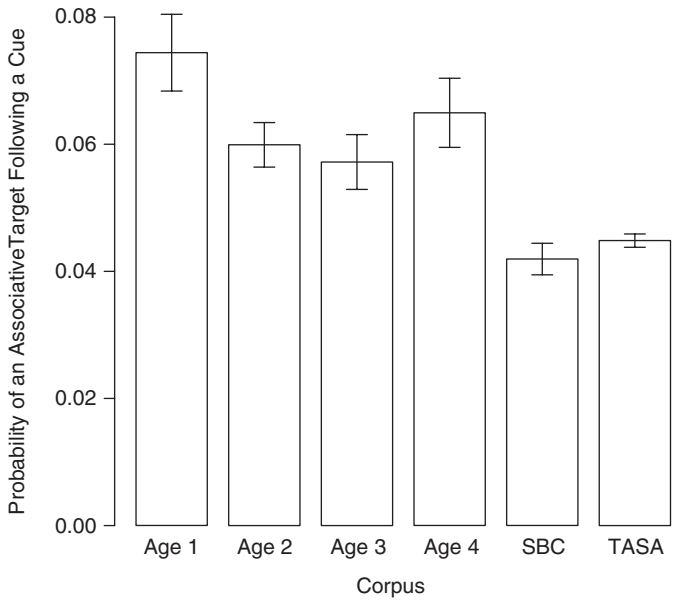

Fig. 2. The probability of producing an associative target of a cue within a five-word window following the production of that cue in child- or adult-directed language. Bars show the probability after subtracting the probability of producing a target following a cue in a random (i.e. shuffled) corpus. Child-directed language is represented by CHILDES. Adult-directed speech is represented by the Santa Barbara Corpus (SBC). Adult-directed written language is represented by the TASA corpus. Error bars are standard error of the mean.

RESULTS

Do child- and adult-directed language differ in associative structure?

The analysis of the associative structure of the six corpora confirms that child-directed language is more associative than adult-directed language (Figure 2). A one-way ANOVA predicting associative structure as a function of child- or adult-directed language reveals a significant effect ($F(1, 18441) = 43{\cdot}41$, $p < 0{\cdot}001$). In more detail, the first year corpus is significantly higher in associative structure than the Santa Barbara Corpus (difference $= 0{\cdot}03$, $t(3194) = 5{\cdot}85$, $p < 0{\cdot}001$, $\eta^2 = 0{\cdot}21$) and the TASA corpus (difference $= 0{\cdot}03$, $t(4975) = 7{\cdot}41$, $p < 0{\cdot}001$, $\eta^2 = 0{\cdot}21$). The other three CHILDES corpora show similar significant differences with the adult corpora ($p < 0{\cdot}001$). However, restricting the analysis to the CHILDES corpora, a repeated-measures ANOVA reveals a non-significant effect of age on associative structure ($F(1, 10270) = 1{\cdot}37$, $p = 0{\cdot}24$). This supports the hypothesis that associative structure is enhanced in child-directed language

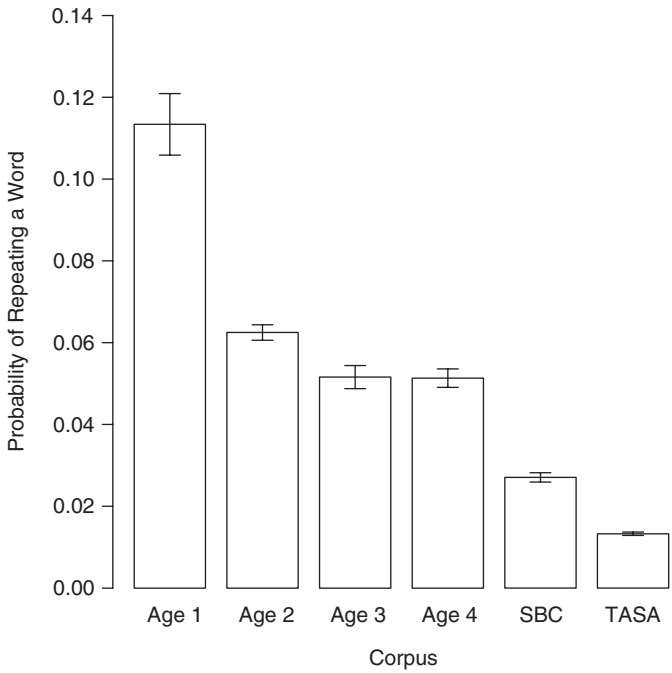

Fig. 3. The proportion of word repetitions in a five-word window within each corpus. Bars show the probability of a repetition after subtracting the probability of repeating words in the shuffled corpus. Corpora are as in Figure 2. Error bars are standard error of the mean.

relative to adult-directed language, but indicates that enhancement of associative structure may persist over the first four years.

Note that the associative structure produced by the shuffled corpora is higher for the CHILDES corpora ($M = 0{\cdot}011$) than for the adult-directed corpora ($M = 0{\cdot}004$; $F(1, 18441) = 18{\cdot}52$, $p < 0{\cdot}001$). This indicates that the correction for the shuffled corpora is warranted, because child-directed language uses words that are, in general, more likely to be associates. However, though child-directed language would be more associative than adult-directed language, even if words were randomly chosen, the analysis of the adjusted data indicate that the observed structure of child-directed language further enhances these associative relations in comparison with the structure of adult-directed language.

Do child- and adult-directed language differ in word repetitions?

Similar to associative structure, word repetitions are also found to be more likely in language directed at the youngest children (Figure 3). A one-way

ANOVA predicting the probability of word repetition as a function of child- or adult-directed language reveals a significant effect ($F(1, 18441) = 217{\cdot}13$, $p < 0{\cdot}001$). Focusing on the first year, the probability of a repetition within the age 1 corpus is significantly higher than both the Santa Barbara Corpus (difference $= 0{\cdot}10$, $t(3194) = 8{\cdot}94$, $p < 0{\cdot}001$, $\eta^2 = 0{\cdot}32$) and the TASA corpus (difference $= 0{\cdot}11$, $t(4975) = 12{\cdot}65$, $p < 0{\cdot}001$, $\eta^2 = 0{\cdot}36$). Similar significant effects are found between all pairs of child- and adult-directed corpora ($p < 0{\cdot}001$). Further restricting the analysis to the CHILDES corpora, a repeated-measures ANOVA reveals a significant effect of age on the probability of repetition ($F(1, 10270) = 52{\cdot}08$, $p < 0{\cdot}001$), with the most repetitive speech directed at the youngest listeners.

Note that the probability of a word repetition in the shuffled corpora did not significantly differ between the child- and adult-directed corpora ($F(1, 18441) = 2{\cdot}16$, $p = 0{\cdot}14$). The mean probability of a repetition across all shuffled corpora was 0·01. Taken together, the above results suggest that the probability of a repetition is higher for child- than adult-directed language, but also reveals that this pattern decays over the first four years.

Do child- and adult-directed language differ in contextual diversity?

Child-directed language was found to be significantly less diverse than adult-directed language. The negative numbers in Figure 4 show the reduction in novelty, in number of novel word neighbours per word appearance, relative to the shuffled corpora. Comparing child- with adult-directed language, a one-way ANOVA predicting contextual diversity reveals a significant difference ($F(1, 18441) = 9{\cdot}74$, $p = 0{\cdot}001$). Further restricting the analysis to the CHILDES corpora, a repeated-measures ANOVA reveals a significant effect of age on contextual diversity ($F(1, 10270) = 38{\cdot}48$, $p < 0{\cdot}001$), with language directed to younger children being less contextually diverse than language directed to older children. A repeated-measures ANOVA reveals that the contextual diversity is not significantly different across the four child-directed corpora when shuffled ($F(1, 10270) = 1{\cdot}38$, $p = 0{\cdot}23$).

Comparing individual corpora, we find that the age 1 corpus is less contextually diverse than the SBC corpus (difference $= -0{\cdot}24$, $t(3194) = -11{\cdot}56$, $p < 0{\cdot}001$, $\eta^2 = -0{\cdot}41$) and the TASA corpus (difference $= -0{\cdot}06$, $t(4975) = -4{\cdot}02$, $p < 0{\cdot}001$, $\eta^2 = -0{\cdot}11$). The age 4 corpus is also less contextually diverse than the SBC corpus (difference $= -0{\cdot}12$, $t(3194) = -9{\cdot}58$, $p < 0{\cdot}001$, $\eta^2 = -0{\cdot}34$), but is more contextually diverse than the TASA corpus (difference $= 0{\cdot}06$, $t(4975) = 7{\cdot}63$, $p < 0{\cdot}001$, $\eta^2 = -0{\cdot}22$). A comparison of the unadjusted corpora produces a similar result. A one-way ANOVA comparing the unadjusted corpora reveals that the CHILDES corpora ($M = 3{\cdot}00$) are significantly less diverse than the

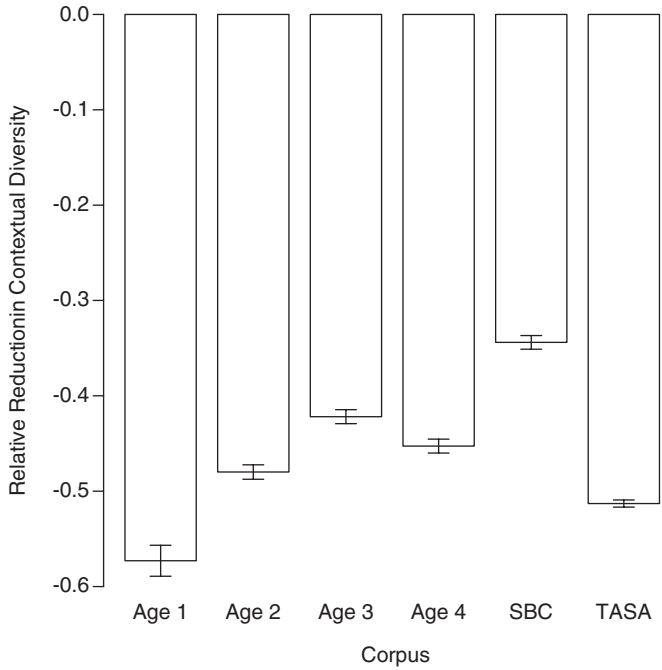

Fig. 4. Reduction in number of novel neighbouring words per appearance of a word, relative to the shuffled corpora. Corpora are as in Figure 2. Error bars are standard error of the mean.

SBC corpus ($M = 3{\cdot}53$, $F(1, 13465) = 139{\cdot}5$, $p < 0{\cdot}001$). A similar comparison between the unadjusted CHILDES and TASA corpora is also significant ($F(1, 15246) = 1099{\cdot}6$, $p < 0{\cdot}001$), but the TASA corpus is less contextually diverse ($M = 1{\cdot}93$) than the CHILDES corpora ($M = 3{\cdot}00$).

Thus, for contextual diversity, results are mixed for language directed at older children, but as with associative structure and repetitions, the largest (and most consistent) differences between child- and adult-directed language are found when comparing adult-directed language with the age 1 corpus. Here again we find evidence for statistical changes in language structure, with child-directed language being less diverse than adult-directed language.

When are the statistical properties of language best correlated with age of acquisition?

To evaluate the potential role of associative structure, word repetitions, contextual diversity and frequency in facilitating early language acquisition,

correlations were computed between these variables and age of acquisition across the different corpora. Because previous work has observed that frequency is an incomplete predictor of age of acquisition, with most of its predictive power within word classes, and much more limited predictive power across word classes (Goodman, Dale & Li, 2008; Hills *et al.*, 2010), correlations were evaluated both within word classes (i.e. nouns, verbs, adjectives and function words) and over all word classes combined. To evaluate the influence of changes in associative structure, word repetitions and contextual diversity relative to the shuffled corpora, the correlations with age of acquisition were computed against the log ratio of the observed value of the statistic over the shuffled value. For frequency, the log of the observed frequency was used. These variables were also entered into a step-wise multiple regression, to identify which variables explained a significant amount of the variance after incorporating the other variables. This further provides an overall measure of how much variance remains to be explained beyond the statistical properties of language investigated here. Table 2 presents the results of this analysis, for which the major findings are outlined below.

The critical observation relative to the current study is that the earliest learned words are significantly more likely to be repeated, to appear with their associates, to have a higher frequency of usage, and to be used more consistently than words learned later (correlations in bold in Table 2). Note that for contextual diversity, the positive correlation indicates more diverse words relative to the shuffled corpora are learned later. In addition, in most all cases except function words and adjectives, the correlations are reduced as the language is directed at older individuals. That is, including an interaction effect for data grouped by child- or adult-directed corpora is significant (see the rows marked Δ child/adult). Thus, in general, child-directed language relative to adult-directed language is more associative, more repetitive, has higher word frequencies, and shows increased consistency in usage specifically among words that children are mostly likely to learn when young.

Table 2 also supports previous work indicating differences between word classes (Goodman *et al.*, 2008; Hills *et al.*, 2010). For example, frequency is a strong predictor of word acquisition within some word classes, but has limited predictive power across all word classes combined (compare individual word classes with 'All word classes'). Across all word classes combined, the importance of associative structure, repetitions and contextual consistency becomes more evident, as they all show a significant contribution towards predicting word acquisition in child-directed speech in the multiple regression (see the asterisks following the correlation coefficients).

Finally, though previous research has shown that the most contextually diverse words are learned earlier (Hills *et al.*, 2010), Table 2 shows that

TABLE 2. *Correlations and regression results on age of acquisition for word classes from the MCDI using word statistics from child- and adult-directed language*

	Ass. Struct.	Diversity	Repetitions	Frequency	R^2
Nouns					
Childes Age 1	**−0·36**	**0·27****	**−0·57*****	**−0·62*****	0·39***
Childes Age 2	**−0·37**	**0·24**	**−0·53****	**−0·60*****	0·34***
Childes Age 3	**−0·29**	**0·17***	**−0·53*****	**−0·53*****	0·31***
Childes Age 4	**−0·29**	**0·24**	**−0·47****	**−0·50*****	0·24***
Δ child/adult	***	***	***	***	
SBC	**−0·24***	0·04	**−0·16**	**−0·23****	0·07***
TASA	**−0·17**	**0·12**	**−0·30*****	**−0·24**	0·09***
Verbs					
Childes Age 1	**−0·39****	**0·40**	**−0·46****	**−0·45*****	0·35***
Childes Age 2	**−0·25**	**0·40****	**−0·40**	**−0·36****	0·26***
Childes Age 3	**−0·29****	**0·24**	**−0·24**	**−0·28****	0·15**
Childes Age 4	**−0·25***	**0·30***	−0·14	**−0·21**	0·14***
Δ child/adult	*	n.s.	*	***	
SBC	−0·11	**0·21***	−0·07	**−0·19**	0·03*
TASA	−0·14	**0·25***	−0·09	**−0·18**	0·05*
Adjectives					
Childes Age 1	−0·01	**0·31**	**−0·46*****	**−0·40**	0·20***
Childes Age 2	−0·05	**0·31**	**−0·32**	**−0·34***	0·13**
Childes Age 3	−0·11	0·20	**−0·36****	**−0·26**	0·11**
Childes Age 4	−0·05	**0·29**	**−0·24**	**−0·27**	0·08*
Δ child/adult	n.s.	n.s.	n.s.	***	
SBC	0·00	0·03	−0·11	−0·09	−0·06
TASA	−0·19	**0·28***	−0·21	0·04	0·06
Function words					
Childes Age 1	−0·18	0·14	0·03	**−0·40**	0·06
Childes Age 2	**−0·30****	0·04	0·09	−0·34*	0·12**
Childes Age 3	−0·11	−0·02	**0·31****	−0·26	0·13**
Childes Age 4	−0·05	0·01	0·19	−0·27	0·04
Δ child/adult	n.s.	n.s.	*	n.s.	
SBC	−0·09	0·06	0·11	**−0·09***	0·05*
TASA	−0·04	0·07	−0·10	0·04	−0·05
All word classes					
Childes Age 1	**−0·31****	**0·37*****	**−0·42****	**−0·22**	**0·22*****
Childes Age 2	**−0·31*****	**0·36*****	**−0·34*****	**−0·14**	**0·19*****
Childes Age 3	**−0·27*****	**0·25*****	**−0·27*****	−0·05	**0·12*****
Childes Age 4	**−0·26*****	**0·27*****	**−0·18***	−0·01*	**0·12*****
Δ child/adult	***	***	***	***	
SBC	**−0·16*****	0·04	**0·12**	**0·16****	**0·07*****
TASA	**−0·28****	**0·26***	**−0·30*****	**0·12*****	**0·14*****

NOTE: R^2 is the adjusted R^2 value when including all the variables. Bold correlations are significant at $p<0·01$. * $p<0·05$, ** $p<0·01$, *** $p<0·001$ indicate degree of the significant effect of the partial regressors (for the individual statistics), the overall regression (for the R^2) following a stepwise multiple regression initiated with all the variables, or the significance of including an interaction effect for the child and adult directed corpora in a regression over all corpora for a given structural statistic (Δ child/adult). n.s. indicates that the result of the interaction was non-significant at $p>0·05$.

words that are held most consistent relative to what they would be in shuffled language are learned earliest. Indeed, there is a negative correlation between the observed contextual diversity and the adjusted contextual diversity of a word (e.g. in the age 1 corpus, $r = -0{\cdot}41$, $p < 0{\cdot}001$). In other words, while the most contextually diverse words are learned earlier (see Hills *et al.*, 2010, for a detailed analysis of this effect using the CHILDES corpus), these same words show the greatest reduction in contextual diversity from what one would expect based on their frequency of usage (i.e. in the shuffled corpora). Though somewhat unsatisfying, this may be interpreted to support previous findings for both consistency (e.g. Brown, 2008; Waxman & Klibanoff, 2000) and contextual diversity (Hills *et al.*, 2010). Future research will be needed to disentangle these variables.

CONCLUSIONS

This study reports the results of a comparative study of the statistical structure of natural language in relation to language acquisition. The results strongly support the central claim of 'motherese' (Newport *et al.*, 1977). That is, language is structured differently when directed to early language learners than when directed to more fluent speakers, and specifically in ways that appear to be correlated with early language learning. In particular, these structural changes correspond to changes in associative structure, contextual diversity, repetitions and frequency, all of which correlate with words that are being learned during the earliest years.

One of the goals of the present study is to understand why adult-generated free associates are correlated with the order of acquisition in early language learning (see Hills *et al.*, 2009; Steyvers & Tenenbaum, 2005). The observation that child-directed language is more associative than adult-directed language offers two potential reasons. First, the associates themselves may facilitate the development of meaningful semantic and syntactic roles for these words (e.g. Recchia *et al.*, 2008). This is similar to the way the semantic and syntactic bootstrapping hypotheses suggest that words rely on the contexts in which they are used (Gleitman, 1990; Grimshaw, 1981). Second, the likelihood that a target word is the object of a cue is correlated with the target's contextual diversity in the language environment (Hills *et al.*, 2010), which reduces the class of possible mappings, facilitating cross-situational learning.

In both cases, associative structure and contextual diversity can facilitate the acquisition of meaning in multiple ways, and the results presented here provide correlational evidence that when adults direct speech to children they alter their patterns of language production to make these potential paths to learning more easily available. They do this both by embedding

to-be-learned words among neighbours that contribute to their meaning, and by reducing the contextual diversity of the earliest learned and most contextually diverse words. These results further support the notion that meaning is acquired through the company that words keep – either through association or exclusion. In each case, structure matters.

The bird's-eye view provided by corpus analysis is unlikely to be sufficient to tease all the structural variables associated with language learning apart. Nonetheless, though corpus analyses cannot evaluate the causal role of the structural variables outlined here, what it can establish is that child-directed language amplifies (relative to adult-directed language) the associative structure, contextual consistency, frequency and repetitions of the earliest-learned words. This may in turn facilitate the learning of the semantic boundaries and usage patterns that define these words in our common usage.

REFERENCES

Block, N. (1999). Functional role semantics. In R. A. Wilson & F. C. Keil (eds), *MIT encyclopedia of the cognitive sciences*, 331–32. Cambridge, MA: MIT Press.

Brown, P. (2008). Verb specificity and argument realization in Tzeltal child language. In M. Bowerman & P. Brown (eds), *Cross linguistic perspectives on argument structure: Implications for learnability*, 167–90. New York: Oxford University Press.

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language* **17**, 417–31.

Collins, A. M. & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior* **8**, 240–47.

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

Du Bois, J. W. & Englebretson, R. (2000–2005). *Santa Barbara corpus of spoken American English, Parts 1–4*. Philadelphia: Linguistic Data Consortium.

Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist* **66**, 103–114.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis*. Special volume of the Philological Society, 1–32. Oxford: Philological Society.

Gillette, J., Gleitman, H., Gleitman, L. & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition* **73**, 135–76.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition: A Journal of Developmental Linguistics* **1**, 3–55.

Goldstone, R. L., Steyvers, M. & Rogosky, B. (2003). Conceptual interrelatedness and caricatures. *Memory & Cognition* **31**, 169–80.

Goodman, J. C., Dale, P. S. & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language* **35**, 515–31.

Grimshaw, J. (1981). Form, function and the language acquisition device. In C. L. Baker & J. McCarthy (eds), *The logical problem of language acquisition*, 163–82. Cambridge, MA: MIT Press.

Hayes, D. P. & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of Child Language* **15**, 395–410.

Hayes, J. R. & Clark, H. H. (1970). Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (ed.), *Cognition and the development of language*, 221–34. New York: Wiley.

Hills, T., Maouene, M., Maouene, J., Sheya, A. & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science* **20**, 729–39.

Hills, T., Maouene, J., Riordan, B. & Smith, L. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language* **63**, 259–73.
Hoff, E. & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development* **73**, 418–33.
Hurtado, N., Marchman, V. A. & Fernald, A. (2008). Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish-learning children. *Developmental Science* **11**, F31–F39.
Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M. & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology* **27**, 236–48.
Jones, M. N. & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review* **104**, 1–37.
Kachergis, G., Yu, C. & Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word–referent mappings. *Psychonomic Bulletin & Review* **19**, 317–24.
Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* **104**, 211–40.
Lenat, D. B. & Feigenbaum, E. A. (1991). On the thresholds of knowledge. *Artificial Intelligence* **47**, 185–250.
Li, P., Farkas, I. & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks* **17**, 1345–62.
Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavioral Research Methods, Instruments, and Computers* **28**, 203–208.
MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*, 3rd edn. Mahwah, NJ: Erlbaum.
Maouene, J., Laakso, A. & Smith, L. B. (2011). Object associations of early-learned light and heavy English verbs. *First Language* **31**, 109–132.
Markman, E. M. (1984). The acquisition and hierarchical organization of categories by children. In N. C. Sophian (ed.), *Origins of cognitive skills: The 18th annual Carnegie symposium on cognition*, 276–406. Hillsdale, NJ: Lawrence Erlbaum Associates.
Mather, E. & Plunkett, K. (2012). The role of novelty in early word learning. *Cognitive Science*. [Advance online publication. doi:10.1111/j.1551-6709.2012.01239.x]
Naigles, L. R. & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of Child Language* **25**, 95–120.
Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved from http://w3.usf.edu/FreeAssociation.
Newman, R. (2008). The level of detail in infants' word learning. *Current Directions in Psychological Science* **17**, 229–32.
Newport, E. L., Gleitman, H. & Gleitman, L. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (eds), *Talking to children: Language input and acquisition*, 109–150. Cambridge: Cambridge University Press.
Plaut, D. C. & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (ed.), *Carnegie Mellon symposium on cognition, May 1997, Pittsburgh, PA, US*, 381–415. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.
Recchia, G., Johns, B. T. & Jones, M. N. (2008). Context repetition benefits are dependent on context redundancy. In V. Sloutsky, B. Love & K. McRae (eds), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 267–72. Mahwah, NJ: Lawrence Erlbaum.
Riordan, B. & Jones, M. N. (2007). Comparing semantic space models using child-directed speech. In D. S. McNamara & J. G. Trafton (eds), *Proceedings of the 29th conference of the Cognitive Science Society*, 599–604. Austin, TX: Cognitive Science Society.

Riordon, B. & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science* **3**, 303–345.
Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language* **35**, 185–205.
Saffran, J. R., Aslin, R. N. and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* **274**, 1926–28.
Saffran, J. R., Newport, E. L. & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language* **35**, 606–621.
Saussure, F. de (1959). *Course in general linguistics*, trans. W. Baskin. New York: McGraw-Hill. (Original work published in 1916).
Smith, L. & Yu, C. (2008). Infants rapidly learn word–referent mappings via cross-situational statistics. *Cognition* **106**, 1558–68.
Snow, C. (1972). Mothers' speech to children learning language. *Cognitive Development* **43**, 549–65.
Steyvers, M. & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science* **29**, 41–78.
Vincent-Smith, L., Bricker, D. & Bricker, W. (1974). Acquisition of receptive vocabulary in the child. *Child Development* **45**, 189–93.
Waxman, S. R. & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology* **36**, 571–81.
Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science* **18**, 414–20.