The semantic structure of children’s vocabulary predicts future word learning

Young children are bombarded by words they do not know. Over time, children begin to use these words in their own speech. To understand how words go from novel to meaningful, researchers have often focused on studying the relationship between word-forms and a child’s external environment (e.g., Markman, 1990). For example, a child may observe the word “cup” being used in the presence only certain objects, and thereby narrow her hypothesis-space to the objects that are cups (Smith & Yu, 2008). Researchers have also studied social factors such as where a caretaker is looking during naming (e.g. Baldwin, 1991, 1993). These two strands of research have in common a focus on relationships between the words and the world. Another less prominent strand of research has focused on language-internal factors such as the linguistic context in which a word appears (e.g., syntactic bootstrapping; Gleitman, 1990), and the relationship between the word being learned and a child’s existing vocabulary (Hills, 2009; Hills, et al., 2009; Beckage, Smith, & Hills, 2011). Here, we take a novel approach to a study of language-internal factors by examining whether the coherence of a child’s vocabulary at time1 predicts their language knowledge at time2.

We analyzed the transcripts from English-speaking children in the CHILDES corpus (MacWhinney, 2000) of 56 children whose production vocabularies overlapped with the words in other data sources (described below) and who had transcripts available at two distinct age bins:

time1 (M=22.6m; range 20.0-24.5); time2 (M=27.1m, range 25.0-29.3). For each child at each timepoint, we quantified their linguistic knowledge using two measures: a measure of textual lexical diversity that controls for variability in transcript length (MTLD; McCarthy & Jarvis, 2010; mean MTLD time1=11.8; time2=15.3; Fig. 1), and mean length of utterance (MLU; mean MLU time1=8.20; time2=20.40). We then predicted each child’s MTLD and MLU at time2 based on semantic properties of the child’s vocabulary at time1. For each child, we measured the overall semantic similarity of their vocabulary using published concept feature norms (McRae, et al., 2005). A word was high on similarity to the extent that it shared features with many other word meanings (e.g., “car” which is similar to many other vehicles; see Table 1). We predicted children’s MTLD and MLU while controlling for MTLD, MLU, mean utterance length, mean word frequency, transcript length, age, and proportion nouns in vocabulary at time1.

Mean word similarity at time1 was a strong predictor of a child’s MTLD at time2 (*B*=0.25; SE=.07; *t* = 3.98) suggesting that children who know more semantically cohesive words at time1 go on to learn more words *and* to have a more productive vocabulary. Semantic similarity did not reliably predict MLU at time2 (*B*=0.11; SE=.09; *t*=1.33). Although our finding is correlational, it hints at a possible causal factor: a well-structured vocabulary at a younger age may enable children to more effectively learn future words.

|  |  |
| --- | --- |
| **Low similarity words** | **High similarity words** |
| clock | orange |
| corn | car |
| fence | elephant |
| rocket | socks |
| bow | house |
| belt | cup |

Table 1: Examples of words produced by children in CHILDES that are low (left) and high (right) on semantic similarity.

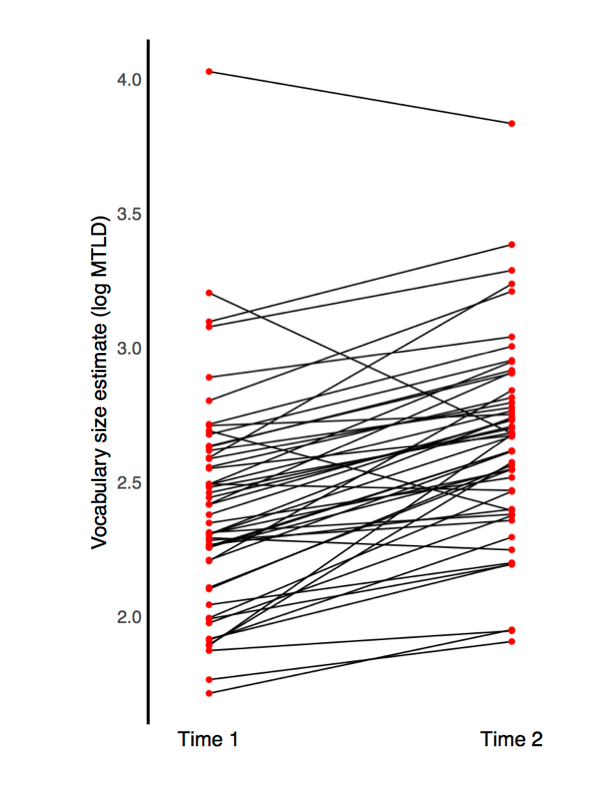


Figure 1: Log MLTD at estimated for children in CHILDES at time1 (mean age= 22.6m) and time2 (mean age = 27.1m).