Supplemental Material S2. Additional information about training and long-term testing protocol.

Table 1. Words that were retrieved and produced correctly in response to the Free Recall task at the beginning of each session. These words did not receive training during Training Blocks 1-3 on the day indicated. All words were trained for all children on Day 1.

Participant	Group	Day 2	Day 3	Day 4	Day 5	Day 6
15C0194	TD	/gramɚ/	/bɪnɪp/ /gramɚ/ /kinɪt/ /sibl̩/	/bɪnɪp/ /dob/ /gramɚ/ /kinɪt/ /sibl̩/ /topɪn/	/bɪnɪp/ /dob/ /gramɚ/ /kinɪt/ /mep/ /nedɪg/ /plun/ /sibl̩/ /topɪn/	
18J0549	TD	/kinɪt/	/gramɚ/ /kinɪt/ /sibl̩/	/gramɚ/ /kinɪt/ /sibl̩/	/gramɚ/ /kinɪt/ /sibl̩/	/gramɚ/ /kinɪt/ /sibl̩/
18J0576	TD	/dob/	/dob/ /sibl̩/	/dob/ /gramɚ/ /nedɪg/ /sibl̩/ /topɪn/	/dob/ /gramɚ/ /mep/ /nedɪg/ /sibl̩/ /topɪn/	/dob/ /gramɚ/ /nedɪg/ /sibl̩/ /topɪn/
18J0580	TD	/dob/	/dob/ /gramɚ/	/dob/ /mep/ /nedɪg/ /plun/		
19J0584	TD		/nedɪg/	/nedɪg/	/nedɪg/	/bɪnɪp/ /nedɪg/ /plun/
19J0603	TD		/gramɚ/	/dob/	/dob/	/dob/ /topɪn/
19J0611	TD	/gramɚ/ /mep/ /nedɪg/ /plun/				
19J0620	TD	/dob/	/kinɪt/ /sibl̩/ /topɪn/	/bɪnɪp/ /dob/ /gramɚ/ /kinɪt/ /nedɪg/ /plun/ /sibl̩/		
19J0673	TD	/dob/	/dob/ /kinɪt/ /plun/	/dob/ /nedɪg/ /plun/	/dob/ /mep/ /nedɪg/ /plun/ /sibl̩/	/dob/ /mep/ /nedɪg/ /plun/ /topɪn/

Participant	Group	Day 2	Day 3	Day 4	Day 5	Day 6
18J0474	DLD		/nedɪg/	/dob/ /nedɪg/	/nedɪg/	/nedɪg/ /plun/
18J0483	DLD		/mep/	/mep/ /nedɪg/	/dob/ /mep/ /nedɪg/	/dob/ /mep/ /nedɪg/ /plun/ /topɪn/
18J0511	DLD			/gramɚ/ /sibḷ/	/bɪnɪp/ /gramɚ/ /mep/ /plun/	/bɪnɪp/ /gramɚ/ /kinɪt/ /mep/ /plun/ /sibḷ/
18J0513	DLD			/dob/ /gramɚ/		
18J0550	DLD	/dob/	/dob/ /gramɚ/ /sibḷ/	/dob/ /gramɚ/ /kinɪt/ /sibḷ/	/gramɚ/ /sibḷ/	
18J0554	DLD			/kinɪt/		/gramɚ/ /nedɪg/ /plun/
18J0563	DLD	/dob/	/dob/ /gramɚ/ /plun/	/dob/ /gramɚ/ /plun/	/dob/ /gramɚ/ /nedɪg/ /plun/ /topɪn/	/bɪnɪp/ /dob/ /gramɚ/ /plun/
18J0581	DLD				/gramɚ/ /sibḷ/	/dob/ /gramɚ/
19J0587	DLD	/plun/	/dob/ /kinɪt/ /plun/	/dob/ /kinɪt/ /plun/ /sibḷ/	/dob/ /gramɚ/ /kinɪt/ /mep/ /nedɪg/ /plun/ /sibḷ/ /topɪn/	/dob/ /gramɚ/ /kinɪt/ /mep/ /nedɪg/ /plun/ /topɪn/

How words were divided into three sets within child for post-training assessments

Within child, the nine words were divided into three sets of three to be tested one-week (Set A) or two-weeks (Set B) after the last training day, or to be tested one-month (Set C) after the last training day prior to re-testing Set A and Set B. The one-week and two-week testing sessions included three testing blocks with one-minute stretch breaks between each block. During the first block, the examiner presented each object one at a time and administered a free recall prompt immediately followed by a cued recall prompt if needed. During the second block, the examiner administered the 4AFC dot task for all objects. The third testing block was identical to the first

testing block however children were provided with feedback to their responses, similar to the feedback given during training.

Within child, words were organized based on relative performance throughout training as a proxy measure for how well each word was learned. Response to the free recall probe at the beginning of each session was the most challenging measure of word learning. Thus, words were first sorted based on the percentage of times the child produced each word correctly in response to the Free Recall probes administered at the beginning of sessions. Response to the free recall probe at the end of each session was the next most challenging measure of word learning. Thus, words that were produced correctly the same percentage of times at the beginning of each session were sorted based on percentage of correct productions in response to the free recall at the end of each session. We continued sorting words based on responses to end of session cued recall prompts, immediate repetition prompts, and then response to the 4AFC dot task respectfully (see Table 2 below for an example). After all words were sorted, we identified the three words with the best relative performance and randomly assigned each word to the three sets. Similarly, we randomly assigned the three words with the middling performance to each set and the three words with the worse performance to each set. In this way, each set tested at the one-week, two-week, and one-month test included one word with a high performance during training, one word with a middling performance during training, and one word with poor performance during training.

Table 2. The sorting of words based on one child's performance during training and assignment to set. Percentages reflect the percentage of correct responses to child gave to that word in that task during training

	Free Recall Beginning	Free Recall End	Cued Recall End	Repetition	4AFC Dot Task	Performance across Training	Set
/dob/	50%					High	C
/topɪn/	17%					High	B
/grɑmɚ/	17%					High	A
/plun/	0%	50%				Middle	A
/kinɪt/	0%	0%	100%			Middle	C
/mep/	0%	0%	83%	100%	100%	Middle	B
/sibl̩/	0%	0%	83%	100%	80%	Low	B
/nedɪg/	0%	0%	83%	83%		Low	A
/bɪnɪp/	0%	0%	16%			Low	C