Using ‘overlap’ as a measure of young children’s syntactic knowledge

Hao Wang and Toben H. Mintz
University of Southern California
contact: haowang@usc.edu, tmintz@usc.edu

Recent corpus studies assessed young children’s knowledge of the determiner category by examining overlap in contexts in which different determiners occurred (1,2,3). For example, the overlap of determiners \(a\) and \(the\) is computed as the number of nouns used with both \(a\) and \(the\) divided by the number nouns used with either \(a\) or \(the\). The presence of an abstract determiner category is argued to enable acquired knowledge of one determiner to be available to other category members. Children with the category were expected to use multiple determiners with a noun and thus show similar overlap as adults. However, prior analyses came to different conclusions as to whether children’s overlap compared to their mothers’ indicated a determiner category.

Different sample sizes may have contributed to this discrepancy (3,4,5). Valian et al. speculated that small samples would lead to significant underestimates of overlap (3), but did not directly test this hypothesis. In our Analysis 1, we tested the hypothesis that overlap is positively correlated with sample size. For each of seven mothers in CHILDES (6) (children’s age<3 and MLU<4), samples of different sizes were created by successively adding recording sessions. To avoid any effect of mothers adjusting the complexity of their speech according to children’s development, the adding was done working backwards from the last session (e.g., sample 2 included utterances from the final two sessions). Overlap was computed for each sample. The mean overlap for the largest sample (31%, SD=0.06) was significantly higher than the 11% overlap obtained in (3). Overlap and the number of utterances were significant correlated for all corpora (mean r=0.90, p<0.01). Thus, previous studies may have underestimated adults’ and children’s actual overlap because of the small samples used. Our analysis demonstrates how critical sample size is for interpreting overlap scores.

Even samples of the same size may show different overlap because overlap is correlated with the frequency distribution of nouns and determiners in sample—less frequent nouns have fewer opportunities to show overlap (3). Thus, interpreting overlap scores may be further improved by comparing actual overlap to expected values that take into account noun frequency. Assuming determiners have equal opportunities to occur with a noun, the probability of overlap for a word of a given frequency is given in Equation 1. Using the actual noun frequency distribution of the Brian corpus (7), we computed 1,000,000 Monte Carlo trials using this formula to calculate a probability density function of overlap (Figure 1). The mother’s actual overlap (43%) was 40 SDs below the mean expected overlap (63%). We thus quantify what has been recognized—even with an abstract determiner category, determiners do not freely distribute with nouns. This fact, along with the sample size issue, means that we should expect children’s overlap scores to be quite low, even if they have a determiner category. As a first step, using adults’ deviation from the expected values generated by our formula, as opposed to raw overlap, could provide a baseline for evaluating children’s overlap in a more meaningful way.
\[ P(\text{overlap}) = \frac{d^f - 2}{d^f} \]

Equation 1. Probability of an overlap for a particular noun occurring \(f\) times, when there are \(d\) possible determiners. Overlap occurs when the noun occurs with at least two different determiners. To compute the probability density of expected overlap for a corpus sample, we considered only the determiners \(a\) and \(the\) \((d=2)\) and ran Monte Carlo trials that yielded whether each noun in a given Monte Carlo run showed overlap or not, based on this formula. Thus, each Monte Carlo trial yielded a number of nouns that showed overlap, which could be directly converted to an overlap score.

Figure 1. The probability density function of overlap computed from 1,000,000 Monte Carlo trials using Equation 1. (The total number of noun types for the mother is 2312.)

References: