Introduction

• Grammatical categories (like noun, verb, determiner) are building blocks of grammar.
• Syntactic rules and structures are stated in terms of grammatical categories.
• Acquiring grammatical categories is crucial for acquisition of syntax.
• Determiner
  • One of the major and early-acquired grammatical categories
  • Carry less meaning than content words, difficult to learn from semantic cues.
  • Consistent distributional regularities in input
  • Errors are rare in young children’s production
• Studying acquisition of determiner may reflect general principles underlying development of grammatical categories.
• Sample size is critical for corpus studies (Rowland, Fletcher, & Freudenthal, 2008; Tomasello & Stahl, 2004).

Deterior Overlap as an Indicator for Determiner Category

• Presence of an abstract determiner category will result in the use of multiple determiners with nouns: e.g., a dog, the dog, etc.
• The use of multiple determiners with a noun has been coined overlap, and is quantified in the equations below.

\[
\text{Overlap}_{a,b} = \frac{\# \text{ nouns occurring with both } a \text{ and the }}{\# \text{ nouns occurring with either } a \text{ or the }}
\]

Equation 1. Overlap between determiner \( a \) and the

\[
\text{Overlap}_{a,b} = \frac{\# \text{ nouns occurring with any two determiners}}{\# \text{ nouns occurring with any determiner}}
\]

Equation 2. Overlap between all determiners

• Used to assess young children’s knowledge of determiner category (Pine & Martindale, 1996; Pine & Lieven, 1997; Valian et al., 2009): if children have a determiner category, they should have similar overlap scores as adults.
• Different conclusions from prior studies regarding status of determiner category in young children
  • Abstract determiner category (Valian et al., 2009)
  • Limited scope formula (Pine & Martindale, 1996; Pine & Lieven, 1997)
• Sample sizes may have been too small to reliably detect overlap (400 - 800 per child)
• Adults’ and children’s overlap may have been underestimated

Present Research Questions

• What is the relation between overlap and sample size?
• How to quantify children’s knowledge of determiner category using overlap, giving overlap is correlated with sample size and noun frequency distribution in a sample?

Corpora Analyzed

• Spontaneous speech of eight monolingual English-speaking children and their mothers from CHILDES database (MacWhinney, 2001)
• Only utterances when children’s age < 3 and MLU < 4 included
• All corpora were automatically tagged with morphosyntactic information such as part-of-speech.

Table 1. Descriptive statistics of the eight corpora analyzed

<table>
<thead>
<tr>
<th>Child</th>
<th>File</th>
<th>Begin Date</th>
<th>End Date</th>
<th>Begin MLU</th>
<th>End MLU</th>
<th>MLU</th>
<th>Utterances</th>
<th>Utterances/month</th>
<th>MLU/month</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brian</td>
<td>619</td>
<td>1-22.18</td>
<td>2-03.12</td>
<td>11.28</td>
<td>14.29</td>
<td>2107</td>
<td>119907</td>
<td>929</td>
<td>16460</td>
</tr>
<tr>
<td>Nina</td>
<td>1-38</td>
<td>1-11.05</td>
<td>2-12.20</td>
<td>2.07</td>
<td>3.58</td>
<td>2114</td>
<td>1922</td>
<td>2296</td>
<td>1963</td>
</tr>
<tr>
<td>Anne</td>
<td>1-38</td>
<td>1-11.05</td>
<td>2-12.20</td>
<td>2.07</td>
<td>3.58</td>
<td>2114</td>
<td>1922</td>
<td>2296</td>
<td>1963</td>
</tr>
<tr>
<td>Alex</td>
<td>1-38</td>
<td>1-11.05</td>
<td>2-12.20</td>
<td>2.07</td>
<td>3.58</td>
<td>2114</td>
<td>1922</td>
<td>2296</td>
<td>1963</td>
</tr>
<tr>
<td>Aidan</td>
<td>1-38</td>
<td>1-11.05</td>
<td>2-12.20</td>
<td>2.07</td>
<td>3.58</td>
<td>2114</td>
<td>1922</td>
<td>2296</td>
<td>1963</td>
</tr>
<tr>
<td>Lani</td>
<td>1-38</td>
<td>1-11.05</td>
<td>2-12.20</td>
<td>2.07</td>
<td>3.58</td>
<td>2114</td>
<td>1922</td>
<td>2296</td>
<td>1963</td>
</tr>
<tr>
<td>Nina</td>
<td>1-38</td>
<td>1-11.05</td>
<td>2-12.20</td>
<td>2.07</td>
<td>3.58</td>
<td>2114</td>
<td>1922</td>
<td>2296</td>
<td>1963</td>
</tr>
<tr>
<td>Lily</td>
<td>1-45</td>
<td>1-11.05</td>
<td>2-12.20</td>
<td>2.07</td>
<td>3.58</td>
<td>2114</td>
<td>1922</td>
<td>2296</td>
<td>1963</td>
</tr>
</tbody>
</table>

Analysis 1

• Hypothesis: overlap is positively correlated with sample size.
• For each of the seven mothers, we created samples of different sizes by successively adding recording sessions.
• To avoid effect of mothers adjusting speech to the development of their children, we added utterances in reverse chronological order, from the latest to the earliest session (see Table 2).
• A computer program automatically searched for NPs with a determiner.
• Mother’s overlap between \( a \) and the was computed on each sample using Equation 1.

Results

• Mean overlap for the largest sample (31%, SD=0.06) was significantly higher than the 11% overlap obtained in Valian et al. (2009).
• Overlap and number of utterances were significant correlated for all corpora (mean r=0.90, p<0.01) (see Figure 1 for an example).
• Previous studies may have underestimated adults’ and children’s actual overlap because of the small samples used.
• This analysis demonstrates how critical sample size is for interpreting overlap.

Analysis 2

• Hypothesis: to quantify overlap the following needs to be taking into account:
  • Sample size (Pine & Martindale, 1996; Valian et al., 2009)
  • Frequency distribution of nouns in sample (less frequent nouns have fewer opportunities to show overlap) (Valian et al., 2009).
  • Expected overlap of a sample can be computed using Equation 3 assuming determiners have equal opportunities to occur with a noun.
  • Probability density function of expected overlap is computed from Monte Carlo simulation of 1,000,000 Bernoulli trials using actual noun frequency distribution of Brian corpus (Lieven et al., 2009) and Equation 3.
  • Actual overlap for Brian’s mother is computed using Equation 1.

Results

• Figure 2 shows the probability density function of expected overlap.
• Mother’s expected overlap is 63% (1450/2312) and children’s overlap is 47% (63%) below the mean expected overlap.
• Even adults, who have a determiner category, show considerably less productivity in the distribution of determiners with nouns than might be expected.
• Thus, absolute overlap scores might not be a helpful metric for measuring productivity.
• Adults’ deviation from expected overlap may be a better indicator for what to expect from children who have a determiner category.

Conclusions

• Overlap is positively correlated with sample size, in particular due to the “opportunities” for nouns to occur with different determiners (Valian et al., 2009).
• Previous studies have underestimated adults’ and maybe children’s overlap.
• We suggest that a more meaningful comparison of overlap would be between the actual and expected values of a sample that takes into the account the sample size and noun frequency distribution.
• Children’s knowledge of determiner category could be measured by comparing his/her deviation from expected overlap to adults’ deviation.

References

Erlbaum Associates.