Redundancy in Children’s Free-Reading Choices

MARGARET HUBBARD JONES AND EDWARD C. CARTERETTE

University of California, Los Angeles, California

Information theory has provided quantitative measures of redundancy which can be used to evaluate the difficulty of text (Shannon, 1951; Garner, 1962). Application of these statistics to adult text gave evidence of a very regular function of constraint among letters in English at various distances from the first, and also good discrimination of text presumed to be different in difficulty (Newman and Waugh, 1960). Carterette and Jones (1963) applied similar measures to children’s graded readers. These also gave excellent discrimination of difficulty and showed that a first reader is considerably more redundant than simple adult text. The questions then arise as to what level of redundancy occurs in reading material children prefer, in the language they speak, and in the material, both written and aural, which they learn most easily. The present paper is addressed to the first of these questions.

METHOD

Statistical Procedure. The method used for the calculation of letter redundancy was a modification of Newman and Waugh’s method, giving more precise estimates because of more extensive use of tables of \( \log_2 \) and the addition of the Miller-Madow (1954) bias correction for sample size. Contingency tables were constructed showing the number of times a given symbol was followed at step \( n \) to the right by every possible symbol, where \( n \) takes on the values 2, 3, . . . , 12, 30, 60, 120. The number of symbols, \( a \), was 28: the 26 letters of the English alphabet plus word mark and sentence mark. From these contingency tables, calculations were made of

\[
H(1) = - \sum_{i=1}^{a} p(i) \log_2 p(i)
\]

in which \( H(1) \) is the average amount of information per letter in bits and \( p(i) \) is the probability associated with occurrence of each of the symbols. The relative sequential constraint contained in pairs of letters for a sequence of \( n \) letters is given by

\[
C_n = \frac{\sum_{k=2}^{n} T(1,k)}{H(1)}
\]

where

\[
T(1,k) = 2H(1) - H(1,k)
\]

These calculations yielded the constraints between all pairs of letters at steps 2, 3, . . . , 12, 30, 60, and 120 to the right of the first letter. Finally, letter counts, word counts, sentence counts, mean word length, and mean sentence length were computed. A more complete discussion of the statistical treatment will be found in Carterette and Jones (1963).

Selection of Material. In order to select books which children enjoy, the free-reading choices were taken from among the top-rated books in Wilson’s Children’s Catalogue (West and Shor, 1961). These books are rated by a wide sampling of children’s librarians all over the country, and by specialists in children’s literature, as books they could not do without. Each rating is accompanied by a consensus of proper grade level. We chose at random from among those books rated in the top one-ninth and graded K to 2 or 3 as our Level 1, those graded 2 or 3 to 5 as our Level 3, and those graded 4 to 6 as our Level 5. At a single level, we used all or part of three or four separate books, drawing equally from them to reduce the effects of idiosyncratic lan-
guage, aiming at approximately ten-thousand-word samples. Although there is undoubtedly some adult bias in these ratings due to pedagogical urges, they are also dependent upon circulation and enjoyment in "story hours." It is difficult to believe that a book which is unpopular with children could remain in the list year after year, regardless of adult bias. And circulation figures alone are no more free of adult interference than these. At least these ratings have the advantage of representing a national sample and a large one.

The various statistics were computed for the three levels of free-reading choices in order to determine whether there was a decrease in sequential constraint as the reading level of the texts increased, whether the 5th level approached adult text in redundancy, and what relation the three levels bore to Readers of approximately the same level of reading skill. Carterette and Jones (1963) found a regular decrease in constraints from a First Reader through a Third Reader to a Fifth Reader. The redundancy of the Fifth Reader approximated average, adult text, as calculated by Newman and Waugh (1960).

Since we were somewhat concerned about the influence of style and lexicon upon these measures of constraint, we compared the statistics for three complete texts at Level 5. The three were chosen at random from among the books rated in the top one-ninth of books at Level 5 in the Wilson's Children's Catalogue. Level 5 was chosen for this because the texts are longer, approaching the ten-thousand-word level, whereas at Levels 1 and 3, the individual texts are too short to yield dependable results. A second concern arose regarding the effect of sample size. In order to investigate this, an unusually long text at Level 5, Alice in Wonderland, was broken down into separate samples of about 1000, 3000, 6000, 12,000, and 18,000 words. Finally, because word length and sentence length are sometimes considered to be reliable indices of difficulty, a peculiar text rated as easy (Level 3) but having very long words and sentences was run to discover whether the measure of redundancy would follow the word and sentence length or the rated level of difficulty.

### RESULTS AND DISCUSSION

#### Levels

Table 1 presents the statistics for the three levels of free-reading choices. It will be seen that mean word length gives no evidence of regularity here. Sentence length shows regular growth. \( H(1) \), the average amount of information in the letters, is not different for the three levels, but \( C_2 \) and \( C_{11} \), the relative sequential constraint for pairs of letters adjacent (\( C_2 \)), and at the two ends of a sequence of 11 letters (\( C_{11} \)), do show some differences. Reference to Fig. 1 will make this clearer. The regular nature of the function, \( C_n \), is obvious. Level 1 lies above the other two, which are indistinguishable. Therefore, Level 1 is slightly more redundant than the others —42% as opposed to 38%—but Levels 3 and 5 lie together. These results can be compared with those obtained for the graded Readers of the Ginn Series reported by Carterette and Jones (1963). The same regularity of the function, \( C_n \), was observed there. Level 1 is less redundant than the First or Second Readers, approximating the redundancy of the Third Reader and simple adult text. Levels 3 and 5 are very similar to the Fifth Reader and to average adult text.

Children's free-reading choices are thus less
redundant than their Readers. It seems probable that at least at Level 1, and to some extent at Level 3, these books are, for the most part, read to the child. The point is not whether he can read them, given our current techniques of language teaching, but rather what he enjoys. He appears to prefer language which is less redundant than that presented to him in his Readers. Writers of Readers assume that a very high level of redundancy is optimal. The child himself prefers less redundancy. It remains to be seen which is the optimal level for learning. It is also of interest to inquire into the relation between the child’s reading preferences and his own use of language.

Style

The effects of style and differences in lexicon can be estimated from the results of the analysis presented in the first three lines of Table 2 and in Fig. 2. The smooth curves of growth of relative sequential constraint are similar to those found for levels and for other texts. \( H(1) \) does not differ among the individual texts, nor did it among levels. It is apparently descriptive of the English language rather than of minor differences in its use. The differences in sequential constraint among the three texts are slight, although Alice is somewhat less redundant than the other two. This may well reflect a genuine

**Table 2**

<table>
<thead>
<tr>
<th>Source</th>
<th>Level</th>
<th>( C_2 )</th>
<th>( C_{11}^a )</th>
<th>Mean word length</th>
<th>Mean sentence length (words)</th>
<th>Number of words in total text</th>
<th>( H(1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Doolittle</td>
<td>5</td>
<td>0.216</td>
<td>0.423</td>
<td>3.97</td>
<td>15.38</td>
<td>8246</td>
<td>4.09</td>
</tr>
<tr>
<td>Mary Poppins</td>
<td>5</td>
<td>0.210</td>
<td>0.414</td>
<td>4.15</td>
<td>14.45</td>
<td>9930</td>
<td>4.14</td>
</tr>
<tr>
<td>Alice in Wonderland</td>
<td>5</td>
<td>0.207</td>
<td>0.374</td>
<td>4.01</td>
<td>18.05</td>
<td>18,000</td>
<td>4.09</td>
</tr>
<tr>
<td>The Emperor’s New Clothes</td>
<td>3</td>
<td>0.224</td>
<td>0.446</td>
<td>4.20</td>
<td>19.31</td>
<td>1970</td>
<td>4.07</td>
</tr>
</tbody>
</table>

\(^a\) \( C_{11} \times 100 \) approximates the usual measure of per cent redundancy.
Table 3

<table>
<thead>
<tr>
<th>Sample size (words)</th>
<th>$C_2$</th>
<th>$C_{11}$</th>
<th>Mean word length</th>
<th>Mean sentence length</th>
<th>$H(1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1395</td>
<td>0.202</td>
<td>0.366</td>
<td>4.037</td>
<td>17.885</td>
<td>4.091</td>
</tr>
<tr>
<td>2859</td>
<td>0.204</td>
<td>0.363</td>
<td>3.991</td>
<td>16.062</td>
<td>4.086</td>
</tr>
<tr>
<td>5719</td>
<td>0.204</td>
<td>0.365</td>
<td>3.977</td>
<td>17.597</td>
<td>4.083</td>
</tr>
<tr>
<td>6433</td>
<td>0.206</td>
<td>0.372</td>
<td>3.956</td>
<td>19.733</td>
<td>4.087</td>
</tr>
<tr>
<td>11,314</td>
<td>0.206</td>
<td>0.374</td>
<td>3.998</td>
<td>17.117</td>
<td>4.089</td>
</tr>
<tr>
<td>18,000</td>
<td>0.207</td>
<td>0.374</td>
<td>4.006</td>
<td>18.054</td>
<td>4.090</td>
</tr>
</tbody>
</table>

$C_{11} \times 100$ approximates the usual measure of percent redundancy.

$\text{b This sample consists of the entire first third of running text.}$

difference in difficulty, comparable to that found by Newman and Waugh (1960) between passages from *William James* (36% redundant) and from *The Atlantic Monthly* (32% redundant). The comparison of the three complete texts at Level 5 suggests that mere differences in style or lexicon do not affect the results markedly. *Dr. Doolittle* and *Mary Poppins* are in these respects very different, but the redundancy and curves of sequential constraint are similar. This we believe to be due to similar difficulty, whereas *Alice in Wonderland* is somewhat more difficult. The effect of stylistic differences is, therefore, not as great as was suspected previously.

**Sample Size**

The effects of sample size are shown in Table 3. Samples were drawn randomly in blocks of approximately 250 consecutive words forming a set of complete sentences. Very small samples do not reflect properly the constraints in a total text. A sample size of about 6000 words seems to be adequate for an accurate estimate of constraints among pairs of letters. Probably it is also adequate for an estimate of sentence length. We interpret the relative large mean sentence length in the sample of 6433 words to reflect the change from narrative in the first part of the book to conversation in the last part. The effect of sample size has been of some concern because Seibel's (1962) analysis of long texts showed that, for N-grams greater than 3 or 4, 50 to 80 thousand letters was not enough. For our statistics, where significant constraints do not appear to extend further than 8 or 9 steps at most, 6400 words (about 32,000 characters, including word and sentence marks) are required for stable results. The samples for both levels and for individual texts at Level 5 are thus adequate.

**Word Length**

The results for the single text at Level 3, *The Emperor's New Clothes*, are also given in Table 2. Here the mean word length is greater than that of any other children's text, including the previous study of Readers (Fifth Reader, mean word length = 4.106) but the redundancy is nevertheless high, lying between that of the First and Second Readers with their extremely restricted vocabularies. Sentence length is equally astonishing for so redundant a text. The correlation between difficulty of text and mean sentence length breaks down, as did the similar correlation with mean word length. In spite of long words and long sentences, the text is evaluated both by $C_{11}$ and by expert opinion as not very difficult. If the lexicon size is limited, it is not necessary to restrict it to short words or words of high frequency. Similarly sentences may be long if not too involved in construction. Thus certain of the restrictions placed upon Readers may be unnecessary as, indeed, one might suspect from the fact that Rinsland (1945) found 25,632 different words used in writing by elementary school children,
and commented that almost any word might be used by children (p. 20).

**Summary**

Sequential constraints of pairs of letters were computed for free-reading choices of children at approximately first, third and fifth grade levels. When these were compared with the constraints found in First, Third and Fifth Readers, it was found that free-reading choices are less redundant than Readers. The child apparently prefers reading material which lies closer to the redundancy level of adult text. Investigation of three individual texts at the fifth level showed little effect of mere stylistic differences. Neither word length nor sentence length is a reliable index of the difficulty of a text. Sample size should be about 6000 words for computations of this sort.

**References**


(Received May 30, 1963)