Lexical Diversity in the Spontaneous Speech of Children With Specific Language Impairment: Application of D

The lexical diversity of children with specific language impairment (SLI) (ages 3 years 7 months to 7 years 3 months) was compared to that of normally developing same-age peers and younger normally developing children matched according to mean length of utterance in words (MLUw). Lexical diversity was calculated from spontaneous speech samples using D, a measure that uses repeated calculations of type-token ratio (TTR) to estimate how TTR changes as the speech samples increase in size. When D computations were based on 250-word samples, developmental differences were apparent. For both children with SLI and typically developing children, older subgroups showed higher D scores than younger subgroups, and subgroups with higher MLUws showed higher D scores than subgroups with lower MLUws. Children with SLI did not differ from same-age peers. At lower MLUw levels, children with SLI showed higher D scores than younger typically developing children matched for MLUw. The developmental sensitivity of D notwithstanding, comparisons using 100-utterance samples, in which the number of lexical tokens varied as a function of the children’s MLUws, and comparisons between 250- and 500-word samples revealed the possible influence of sample size on this measure. However, analysis of the effect sizes using smaller and larger samples revealed that D is not affected by sample size to the degree seen for more traditional measures of lexical diversity.

KEY WORDS: specific language impairment, vocabulary, lexical diversity, lexicon, language sample

A recent focus of research has been the development of a profile of the language abilities of children with specific language impairment (SLI). Attempts have been made to assess all areas of these children’s language, including phonology, morphosyntax, pragmatics, and vocabulary development. Often these children’s grammatical abilities are considered to be weaker than their lexical abilities, although even the latter appear to be below age-level expectations (see Leonard, 1998, for a recent review). This interpretation has been based on comparisons using formal tasks of both production and comprehension and on comparisons using spontaneous speech measures.

Most of the measures used to assess children’s vocabulary abilities in spontaneous speech are estimates of their lexical diversity. These measures are intended to reflect the variety and specificity of words that a child chooses to use and, less directly, the breadth of topics that can be
discussed. One might consider such measures an estimate of how many words a child knows well enough to use in everyday conversation. Lexical diversity does not refer to all the words a child can comprehend, nor does it address abilities assessed by standardized tests of vocabulary, such as the ability to provide definitions, synonyms, and analogies, or the ability to produce a name rapidly when presented with the referent, as in confrontation naming tasks. Nevertheless, because lexical diversity measures are based on spontaneous speech, they constitute an indispensable gauge of lexical ability. Lexical diversity measures are based on spontaneous speech, rapid when presented with the referent, as in confron-tonyms, and analogies, or the ability to produce a name rapidly when presented with the referent, as in confrontation naming tasks. Nevertheless, because lexical diversity measures are based on spontaneous speech, they constitute an indispensable gauge of lexical ability.

Measures of Lexical Diversity

The most straightforward measure of lexical diversity is a simple count of the number of different words (NDW) in a sample (Miller, 1991). However, this measure does not provide a value that can be compared across children, since NDW, calculated for a sample standardized by time or number of utterances, is influenced by sample size (number of lexical tokens) and consequently also covaries with mean length of utterance (MLU) (Klee, 1992). For example, children with a higher MLU will produce a larger number of words in a 100-utterance sample, and therefore will have a larger total number of words (TNW), than children with a lower MLU. Because NDW should be influenced by TNW, the former measure conflates lexical skills and MLU when comparing a standard number of utterances. Likewise, TNW is related to talkativeness when the sample is constrained by a specific unit of time. Although NDW may be an effective measure for discriminating typically developing children from children with language disorders (e.g., Klee, 1992; Watkins, Kelly, Harbers, & Hollis, 1995), this does not provide a measure of lexical diversity, since it confuses volubility with lexical skills.

Additionally, NDW will increase as the size of the sample increases until the speaker’s total active vocabulary has been exhausted. The rate at which NDW increases will slow across the sample; however, the absolute value of NDW reported will continue to increase. This creates a measure that is difficult to compare in a developmental context or across studies, because, without a specification of sample size, the value of NDW is difficult to interpret. In order to use NDW in a study or clinical comparison, one must shorten all available samples to match the smallest sample size, as measured in lexical tokens, to avoid sample size confounds. Cutting samples based on number of utterances is not sufficient due to the confound with MLU.

One of the best-known measures of lexical diversity is type-token ratio (TTR) (e.g., Fairbanks, 1944; Johnson, 1944), a measure of vocabulary diversity. TTR attempts to correct for some of the defects inherent in the NDW measure. It takes the number of different words (NDW, or types) and compares it to the total number of words (TNW, or tokens) to yield a ratio that serves as a measure of lexical diversity (Richards, 1987). The larger the TTR, the less frequent the repetition of words. Using Templin’s (1957) data, Miller (1981) found that TTR for the first 50 utterances of a sample yielded a ratio of approximately .45, regardless of age, in the age range of 3 to 8 years. Klee (1992) reported similar results in a study of children ranging from 2 to 4 years old. Theoretically, dividing by TNW reduces the effects of sample size; however, it actually produces a new set of difficulties. Because certain closed-class words are frequently repeated within a sample, the larger the sample size is, the smaller the TTR measure will be, provided the speaker is beyond the telegraphic stage of language development.

In addition, if the speaker remains on the same topic and reuses content words related to that topic, TTR will also continue to fall. The literature offers several possible ways of adjusting TTR to accommodate the variation based on sample size, including using a standard number of utterances (Templin, 1957), employing separate measures for open- and closed-class words (Richards, 1987), and using a standard number of tokens (Klee, 1992).

Each of these corrections provides a less than satisfactory result. Relying on a standard number of utterances allows the denominator, number of tokens, to vary. This allows for the possibility that the resulting TTR will vary less as a function of lexical diversity than as a function of MLU, as was discussed above for NDW. Separate measures for open- and closed-class words produce two new problems: in this case, two measures (open-class words, closed-class words) will vary as a function of sample size, rather than one; and the delineation of what is an open-class, as opposed to closed-class, lexical item comes into debate (Richards, 1987). For instance, does one include prepositions, or semi-auxiliaries (e.g., hafta), as closed- or open-class items? Using a standardized number of tokens is the most likely way to overcome difficulties with sample size. However, researchers cannot agree on this standardized sample size, resulting in different values across studies. In addition, researchers must standardize the number of tokens used within each study, possibly leading to the exclusion of children from whom a sufficient number of tokens was not obtained. TTR for a standard number of tokens does reflect changes over age (Klee, 1992), which is a positive attribute.

Recently, Richards, Malvern, and McKee (Richards
& Malvern, 1997; McKee, Malvern, & Richards, 2000) have proposed a new method of measuring lexical diversity, which they refer to as D (using VOCD, dedicated software, to calculate D). The approach for calculating D is based on the probability of introducing new vocabulary into progressively longer language samples. It uses the repeated calculation of TTR over a range of tokens (35–50) to show how TTR changes in relation to sample size. This relationship is then compared to a mathematical model of TTR, and that comparison yields the D score. D therefore provides a means of measuring lexical diversity that “is not a function of the number of words in a sample, uses all of the data available, and is more informative [than TTR] because it represents how the TTR varies over a range of token size” (see MacWhinney, 2000, p. 128). Although D is not measured in units that are as transparent as NDW, the higher the D value, the greater the lexical diversity.

Because D involves a sampling of the lexical tokens and represents the shape of the curve of TTR plotted against number of tokens rather than a single point on that curve (as is the case for TTR), D is assumed to be independent of sample size effects. McKee et al. (2000) demonstrated this via split-half reliability procedures. Using archival data from the CHILDES database (MacWhinney, 2000), they compared D scores for the odd words only, the even words only, and the mean of odd/even words against D scores for all the words in 38 samples. They found that the results for the halved samples and the combined means of the samples were not significantly different from the results for the whole sample. Thus, they concluded that there is no effect of the number of tokens available on the results for D (McKee et al., 2000).

The possible contributions of D to the study of language development are already being explored by researchers. Silverman and Bernstein Ratner (2000) recently compared D and TTR in a study of young children who stuttered and their typically developing, same-age peers. Only D distinguished the two groups of children. Furthermore, D scores were correlated with performance on an expressive vocabulary test, whereas TTR was not.

**Lexical Diversity and Specific Language Impairment**

Although there have been numerous studies of the lexical abilities of children with SLI, relatively few have focused on lexical diversity. Klee (1992) considered the usefulness of several analyses of spontaneous language as developmental and diagnostic measures. The measures considered included TNW in 20 minutes, NDW in the first 50 utterances, TTR for 100, 200, and 400 tokens, and MSL (a measure similar to MLU). According to regression analyses, only MSL, NDW, and TNW served as good diagnostic measures, although TTR for 100 and 200 words were also correlated with age in both the normal and the language-impaired groups.

Watkins et al. (1995) measured TTR and NDW for children with SLI and typically developing children. They compared age- and MLU-matched groups of typically developing children with a group of children with SLI on measures of TTR calculated on 50- and 100-utterance samples and NDW calculated on 100- and 200-token samples. They found that TTR did not differentiate typically developing children from children with SLI for either sample size. NDW, however, did differentiate the age control children from the children with SLI, although the children with SLI were similar to MLU-matched control children.

Similar results were found in the Leonard, Miller, and Gerber (1999) study. NDW calculated from 100-utterance samples was compared for typically developing children and children with SLI ranging in age from 2:2 (years;months) to 6;11. A significantly lower NDW was found for the children with SLI than for the typically developing children of the same age. Similar results were found when the two groups of children were compared strictly on the basis of the number of different verbs and the number of different nouns they produced. However, as these comparisons were based on 100-utterance samples rather than on samples based on the number of lexical tokens, it is difficult to draw conclusions. Children with higher MLUs will have a greater number of tokens in their samples, and this may serve to inflate the NDW reported for typically developing children.

In a study of the growth of lexical diversity in 9 preschoolers with SLI, Goffman and Leonard (2000) found that the NDW used by these children in a 50-utterance sample was equivalent to or exceeded by the number used by younger typically developing children at the same MLU levels. This proved true for the children with SLI who began the study with a low MLU, as well as for those who began with a higher MLU. The lexical diversity of all 9 children with SLI remained consistent with expectations based on MLU during the entire 24-month period of the study.

Thus, there are hints that children with SLI may have lower lexical diversity than typically developing children of the same chronological age. These children’s lexical diversity may instead resemble that of younger normally developing children exhibiting similar MLUs. However, due to constraints in the measures used to compare the typically developing children and the children with SLI, definitive conclusions cannot be drawn. In the present study we compare children with SLI and typically developing children using D.
Method

Participants

Archival spontaneous language samples from typically developing children and children with SLI provided the data for this study. Participants were drawn from a pool of 144 children, ranging in age from 2;2 to 7;3. All of the children included in this study had previously participated in earlier studies conducted by Leonard and colleagues (Leonard, Bortolini, Caselli, McGregor, & Sabbadini, 1992; Leonard et al., 2000; Leonard, Eyer, Bedore, & Grela, 1997; Leonard, Eyer, Bedore, & McGregor, 1993; Leonard, Sabbadini, Leonard, & Volterra, 1987). Each child participating in the study passed a hearing screening and scored above 85 on an age-appropriate nonverbal intelligence test, either the Arthur adaptation of the Leiter International Performance Scale (Arthur, 1952), the Columbia Mental Maturity Scale (Burgemeister, Blum, & Lorge, 1972), or the Wechsler Preschool and Primary Scale–Revised (Wechsler, 1989). In addition, all children passed an oral mechanism exam and demonstrated the ability to use word-final /s, z, t, d/. Based on parental report and examiner observation, there was no history of neurological impairment or seizures, and no symptoms of psychological disturbances or autism/PDD were present.

Fifty-three of these children had been diagnosed as exhibiting SLI. These children ranged in age from 3;7 to 7;3. They scored more than 1 standard deviation below the mean for their age on a standardized expressive language measure. Although the cutoff for inclusion in the SLI group was a score greater than 1 standard deviation below the mean on a standardized language test, many of these children were more than 1.5 or 2.0 standard deviations below the mean. Tests included the expressive composite/subtests of the then-current standardization of the Test of Language Development–Primary: 2 (Newcomer & Hammill, 1991), the Test of Early Language Development–2 (Hresko, Reid, & Hammill, 1991), the Preschool Language Scale–3 (Zimmerman, Steiner, & Pond, 1992), and the Reynell Developmental Language Scales: U.S. Edition (Reynell & Gruber, 1990). All of these children were enrolled in treatment at the time of the study.

The 91 remaining children were considered normally developing (ND). They ranged in age from 2;2 to 6;1. These children scored within one standard deviation or higher on the language measures administered to the children with SLI.

Data Collection

Each language sample was obtained as a component of a diagnostic battery. To obtain the language samples, the child and an examiner played with age-appropriate toys. As they interacted with and discussed the toys, the examiner focused on encouraging the child to initiate and participate in the conversation. The sessions were audiorecorded and later transcribed in SALT format (Miller & Chapman, 1986–2000) and coded for grammatical morpheme use. Coding decisions, including how to segment utterances, were made based on the criteria described in the SALT manual. All child utterances were transcribed; however, only spontaneous, complete, and intelligible utterances were included in the calculations. Interjudge transcription reliability was computed for 15 children (8 ND and 7 SLI). Word-for-word percentage agreement between the original investigator and an independent judge ranged from 89% to 95% (Leonard et al., 1987, 1992, 1997). Complete transcripts were converted from SALT to CHAT/CLAN (MacWhinney, 2000) format and checked for errors using the CHECK command in CLAN.

D

VOCD is a program that runs within the CLAN program (MacWhinney, 2000) on transcripts that are in CHAT format, yielding the measure D. It is calculated by drawing an empirical curve that is compared to a model curve of TTR derived from the work of Sichel (1986). The model curve of TTR × tokens adopted for D applies with the greatest accuracy at low numbers of tokens, so the empirical sampling occurs over the range of 35–50 tokens. Random sampling within the transcript is used in order to best match a single-parameter model, which is based on the multi-parameter probabilistic model proposed by Sichel, and to avoid artificially deflating the score when the speaker remains on the same topic. Thus, the procedure for calculating D involves beginning at a token size, N, randomly selecting N tokens without replacement from within the sample, and calculating TTR for those N tokens. The tokens are then replaced within the sample, and this process is repeated 100 times. The average TTR is calculated for that N, and then the next number of tokens (N + 1) is chosen and the entire procedure begins again. Once these calculations have been accomplished for the entire range of tokens, 35–50, a value D is calculated. D is the parameter that is adjusted to estimate the best fit of the empirical curve to the model curve through the least squares method. The value D is calculated three times through this procedure, and the average is then reported as the final optimum D value for that transcript. Since repeated random sampling is used, the entire transcript is sampled, despite the fact that the curve is only estimated for 35–50 tokens. Additionally, it is important to realize that a slightly different value of D will be reported each time VOCD is run, since the value is based on averages of random sampling (McKee et al., 2000; Richards & Malvern, 1997).
Analysis 1: 100-Utterance Samples

Method

Transcripts

All available transcripts were cut to 100 utterances in length. For each of these samples, MLU based on words (MLUw) and D were calculated for all complete and intelligible utterances (-sxxx, -syyy excludes unintelligible utterances). Morphological endings were excluded from both MLUw and D measures through the use of a switch available in CLAN (+s"*-%"); thus, the words jump, jumps, and jumping were all counted equally in these measures as one word type or morpheme. Inclusion of these bound morphemes would potentially inflate NDW and thus could also inflate D. However, irregular words, such as went (go) and children (child), were counted as separate types (but single morphemes) in these measures with the assumption that they were segmented as a single morpheme by the child. No differentiation was made for homographs. Nonwords (e.g., um, uh, hmmm) and onomatopoeia were excluded via mazes and other transcription codes in SALT (Miller & Chapman, 1986–2000) when the language samples were originally transcribed. The commands used within CLAN to accomplish these analyses are included in the Appendix. The MLUw and D values were entered into a database, which also included information on age and diagnostic category, from which matched groups of children were chosen.

Age Matching

One type of comparison involved SLI and ND groups matched for chronological age. Children in each group were divided into younger and older subgroups, with at least 6 months separating the two subgroups. The younger children ranged in age from 42 to 54 months. The ND cohort (n = 20, M = 48.85 months) did not differ significantly in age from the SLI cohort (n = 16, M = 49.88 months), p = .43. The older children ranged in age from 60 to 71 months. Once again, the ND cohort (n = 18, M = 64.61 months) did not differ significantly in age from the SLI cohort (n = 17, M = 64.29 months), p = .76. This information is summarized in Table 1.

MLUw Matching

A second type of comparison involved SLI and ND groups matched for MLUw. Children in each group were divided into lower and higher MLUw subgroups, with at least 0.50 morphemes separating the two subgroups. This form of matching provided a means to test the status of lexical diversity as a relative strength or weakness in children with SLI compared to children with similar MLUws and allowed us to consider the relation of lexical development to another measure of language development. The MLUw of the low-MLUw children ranged from 2.70 to 3.50 morphemes per utterance. The ND cohort (n = 13, M = 3.14) did not differ significantly in MLUw from the SLI cohort (n = 20, M = 3.07), p = .50. The SLI low-MLUw cohort had a mean age of 59.90 months, and the ND low-MLUw cohort had a mean age of 36.00 months. The MLUw of the high-MLUw children ranged from 4.00 to 5.00 morphemes. Once again, the ND cohort (n = 32, M = 4.34) did not differ significantly in MLUw from the SLI cohort (n = 9, M = 4.32), p = .87. The SLI high-MLUw cohort had a mean age of 67.88 months, whereas the ND high-MLUw cohort had a mean age of 50.25 months. For a summary, see Table 2.

Results

Age Matching

The children's D scores were compared using a Type I sum of squares (SS) analysis of variance (ANOVA) with age (older, younger) and diagnostic group (SLI, ND) as between-subjects factors. A significant difference was found for both age, F(1, 67) = 14.99, p < .0001, and diagnostic group, F(1, 67) = 9.39, p < .01. The older children (M = 63.77, SD = 11.11) showed higher scores than the younger children (M = 53.54, SD = 13.26), and the ND children (M = 62.48, SD = 12.30) exhibited higher scores than the children with SLI (M = 54.09, SD = 12.94). There was no age by diagnostic group interaction (p = .37). A summary appears in Table 3.
MLUw Matching

The Type I SS ANOVA for this comparison resembled the previous analysis, with the exception that MLUw level (higher, lower) served as a between-subjects factor along with diagnostic group. A significant difference was observed for MLUw level, $F(1, 70) = 31.89$, $p < .0001$, but not for diagnostic category, $F(1, 70) = 0.001$, $p = .97$. Higher D scores were seen for the children with higher MLUws ($M = 58.84$, $SD = 13.67$) than for the children with lower MLUws ($M = 43.24$, $SD = 12.81$). Although D scores were generally similar for the SLI ($M = 51.82$, $SD = 13.78$) and ND ($M = 51.92$, $SD = 16.42$) groups across MLUw levels, the MLUw level by diagnostic group interaction was marginally significant, $F(1, 70) = 2.80$, $p = .0985$. As can be seen in Table 4, the children with SLI and the ND children displayed similar D scores at the higher MLUw levels; however, differences favoring the children with SLI were rather clear at the lower MLUw levels.

Interpretation

The findings based on 100-utterance samples can certainly be interpreted as reflecting developmental differences. Specifically, older children are likely to have stronger language skills, and hence greater lexical diversity, than younger children. Children with higher MLUws are likewise probably more linguistically capable than those with lower MLUws, a factor that should also be reflected in lexical diversity. Many children with SLI are known to be below age level in their vocabulary skills (see review in Leonard, 1998); therefore, it is not surprising that their lexical diversity falls short of that seen for same-age peers.

| Table 3. Means and standard deviations for D scores, calculated on 100 utterances for age-matched comparisons. |
|-------|-------|-------|-------|
| Older | Younger |       |       |
|       | $M$  | $SD$ | $M$  | $SD$ |
| ND    | 66.83| 9.65 | 58.57| 13.32|
| SLI   | 60.52| 11.89| 47.26| 10.48|

Table 4. Means and standard deviations for D scores, calculated on 100 utterances for MLUw-matched comparisons.

| Table 4. Means and standard deviations for D scores, calculated on 100 utterances for MLUw-matched comparisons. |
|-------|-------|-------|-------|
| Higher | Lower |       |       |
|       | $M$  | $SD$ | $M$  | $SD$ |
| ND    | 58.53| 14.00| 35.67| 8.87 |
| SLI   | 59.91| 13.17| 48.17| 12.72|

For most of the differences observed, the group with the higher D scores also had samples containing a larger number of lexical tokens than those of the group with the lower D scores. A 100-utterance sample with a higher MLUw—especially when bound morphemes are excluded, as in this study—will by necessity contain more tokens than a 100-utterance sample with a lower MLUw.

According to McKee et al. (2000), an important property of D is that it is not influenced by sample size. Although they demonstrated this through the use of split-half reliability measures, it seemed useful to establish that sample size played no role in our own findings. Accordingly, we repeated the analyses with the same children for both age matching and MLUw matching, but employed only the first 250 lexical tokens of each sample rather than 100 utterances. The use of samples with the same number of lexical tokens ensured that any effects of sample size would be neutralized.

Analysis 2: 250-Word Samples

Method

Transcripts

The transcripts were cut using the CLAN command to shorten transcripts to 250 words. Once again, morphological endings were excluded from MLUw and D calculations. The commands used to accomplish these analyses within CLAN are included in the Appendix. The MLUw and D values were entered into the database from which the matched groups were drawn.

Matched Groups

The same groups described for the analyses involving 100-utterance samples were used for the analyses using the 250-word samples. Thus, for age matching, there were older and younger SLI and ND cohorts. Similarly, for MLUw matching, there were higher and lower MLUw cohorts of both ND children and children with SLI.

Results

Age Matching

The Type I SS ANOVA employed age (older, younger) and diagnostic group (SLI, ND) as between-subjects variables. As in the age-matching analysis using 100-utterance samples, a significant difference was found for age, $F(1, 67) = 21.33$, $p < .0001$. The older children ($M = 59.91$, $SD = 12.44$) exhibited higher D scores than the younger children ($M = 48.15$, $SD = 9.23$). However,
Unlike the earlier age-matching analysis with 100 utterances, there was no difference according to diagnostic group, $F(1, 67) = 1.99$, $p = .16$. The D scores of the ND children ($M = 55.65$, $SD = 13.24$) were not statistically greater than those of the children with SLI ($M = 51.99$, $SD = 11.14$). It should also be noted that, for both groups, the D scores based on 250 words were numerically smaller than the D scores based on 100 utterances (see above). There was no age by diagnostic group interaction ($p = .73$). These results are summarized in Table 5.

**MLUw Matching**

The Type I SS ANOVA for MLUw matching with 250-word samples yielded results similar to those found for 100-utterance samples. A significant difference was seen for MLUw level, $F(1, 70) = 14.91$, $p < .001$. Children with higher MLUws ($M = 51.30$, $SD = 12.68$) earned higher D scores than children with lower MLUws ($M = 42.27$, $SD = 13.94$). A main effect for diagnostic group was not found, $F(1, 70) = 1.31$, $p = .26$. D scores for the children with SLI ($M = 49.33$, $SD = 12.76$) were no higher than those for the ND children ($M = 45.94$, $SD = 14.59$). However, a significant MLUw level by diagnostic group interaction was also observed, $F(1, 70) = 4.93$, $p = .03$. Tukey HSD post hoc tests for unequal Ns indicated that at the higher MLUw levels, the two diagnostic groups were very similar ($p = .99$); however, at the lower MLUw levels, the children with SLI had significantly higher D scores than the ND children ($p = .02$). For the ND group, children at the higher MLUw levels showed significantly larger D scores than the children at the lower MLUw levels ($p < .001$). These results are summarized in Table 6.

**Interpretation**

The findings for samples of 250 words continue to provide strong evidence for developmental differences. Even though analysis was restricted to the same number of words, the older children had higher D scores than the younger children, and ND children with higher MLUws earned higher D scores than ND children with lower MLUws. It is clear, then, that D scores are capable of detecting developmental differences that cannot be attributed to the number of lexical tokens in a sample.

However, the analyses of the 250-word samples also leave room to suspect that sample size might also be a factor in D. When the children with SLI and the ND children matched for age were compared, differences favoring the ND children were quite striking for 100-utterance samples, but no differences were found for 250-word samples. The 100-utterance samples of the ND children were considerably larger in number of lexical tokens than were the 100-utterance samples of the children with SLI.

For the 250-word samples, of course, the two groups’ sample sizes were the same. It can also be seen from the D values reported above that for older and younger children, as well as for children with SLI and ND children matched for age, the D scores for the 100-utterance samples were larger than the D scores for the 250-word samples.

The results of the ANOVAs for the ND versus SLI comparisons involving MLUw matching were very similar for the 100-utterance and 250-word samples. This is to be expected; when MLUw matching is performed, the number of lexical tokens for the children being compared will be quite similar. However, it can be seen that for both the ND children and the children with SLI, the D scores were somewhat lower for the 250-word samples than for the 100-utterance samples.

When children are divided, as they were in this study, into two MLUw levels, the children with higher MLUws will have 100-utterance samples that contain a substantially larger number of lexical tokens than their 250-word samples. It is noteworthy, then, that these are the children whose D scores differ the most—proportionately as well as in absolute terms—when the 100-utterance and 250-word samples are compared (see Tables 4 and 6).

**Analysis 3: Sample Size Influences on D, NDW, and TTR**

**Method**

The lingering possibility that sample size played a role in the D scores prompted us to conduct a third

---

### Table 5. Means and standard deviations for D scores, calculated on 250 words for age-matched comparisons.

<table>
<thead>
<tr>
<th></th>
<th>Older</th>
<th></th>
<th>Younger</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
</tr>
<tr>
<td>ND</td>
<td>62.37</td>
<td>13.57</td>
<td>49.60</td>
</tr>
<tr>
<td>SLI</td>
<td>57.31</td>
<td>10.91</td>
<td>46.35</td>
</tr>
</tbody>
</table>

### Table 6. Means and standard deviations for D scores, calculated on 250 words for MLUw-matched comparisons.

<table>
<thead>
<tr>
<th></th>
<th>Higher</th>
<th></th>
<th>Lower</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
</tr>
<tr>
<td>ND</td>
<td>51.13</td>
<td>13.09</td>
<td>33.15</td>
</tr>
<tr>
<td>SLI</td>
<td>51.87</td>
<td>11.82</td>
<td>48.19</td>
</tr>
</tbody>
</table>

---
analysis. This analysis included all children in our database whose samples consisted of at least 500 words. Transcripts were cut using the CLAN command to shorten samples to 500 words, and D and MLUw calculations were conducted as before. We then compared the children’s D scores for the first 250 words with their D scores for 500 words. Separate comparisons were made for the ND children (n = 78) and the children with SLI (n = 41). See Table 7 for a summary of age and MLUw information for these children. The same comparisons between the 250-word and 500-word samples were also made using the more traditional measures of lexical diversity, TTR and NDW. For each measure and for each group of children, pairwise t tests were conducted to compare values for 250 words and 500 words. We then calculated the effect size (Cohen, 1988) for each comparison. These effect sizes were used to compare the three measures in terms of the magnitude of the sample size differences they produce. According to Cohen, a d value of 0.2 constitutes a small effect size, 0.5 represents a medium effect size, and 0.8 can be interpreted as a large effect size.

Results

A summary of the findings can be seen in Table 8. For the ND children, a difference between the scores for the two sample sizes was found for all three measures [D t(77) = 8.02; TTR t(77) = 30.42; NDW t(77) = 47.08, p < .0001]. The effect size for D was small to medium (d = 0.43), whereas the effect sizes for both TTR (d = 1.86) and NDW (d = 3.18) were very large. The value d can also be used to determine the degree of overlap that exists in the distribution of scores for the 250- and 500-word samples. For D there was a 70% overlap. In contrast, for TTR and NDW, the overlap was only 21% and 6%, respectively.

A similar difference was found for the SLI group [D t(40) = 3.34, p < .01; TTR t(40) = 20.64; NDW t(40) = 31.03, p < .0001]. Comparable effect sizes were also found. The effect size for D was small to medium (d = 0.35). The effect sizes for TTR and NDW were very large (TTR d = 2.41; NDW d = 3.92). For D there was a 76% overlap between the distributions of the 250- and 500-word samples. For TTR and NDW the overlap was only 12% and 3%, respectively.

General Discussion

The primary aim of this study was to determine whether a lexical diversity measure, such as D, would reflect developmental differences. In order to be capable of reflecting developmental differences, a measure should be capable of reflecting differences in developmental levels. For 250-word samples, as well as for 100-utterance samples, higher D scores were found for older children compared to younger children and ND children with higher MLUws compared to ND children with lower MLUws. Thus, linguistic capabilities seem to be reflected in D scores across a relatively wide range of children.

Since D appears to reflect developmental differences, it is interesting to consider the performance on this measure of children with SLI compared to typically developing children matched for age and MLUw. The children with SLI performed similarly to age-matched peers at both older and younger ages when samples were controlled for number of tokens. This was a somewhat surprising finding. Although the lexical abilities of children with SLI are often stronger than their phonological and morphosyntactic abilities, most studies have reported below-age-level performance by these children on lexical tasks. With a few exceptions (e.g., Dollaghan, 1987),
children with SLI seem to perform more poorly than same-age peers on tasks ranging from fast mapping to word recall (Leonard, 1998). Indeed, because some of the children we included had language test scores that were only slightly lower than one standard deviation below the mean, it is possible that our sample of children represented a “mild” form of SLI. More strict identification criteria might have revealed a gap between the SLI group and the ND group. However, a recent study of lexical diversity by Thordardottir and Ellis Weismer (2001) suggests that our findings may not have been an aberration. These investigators compared children with SLI to age-matched ND peers on both number of different verbs and NDW in a 315-word sample. No significant differences were found. Combining this finding with the results of the present study, evidence from two lexical diversity measures—NDW and $D$—suggests that many children with SLI may have only subtle lexical deficits when measures are taken from spontaneous speech.

These results should not be interpreted to mean that the lexical abilities of children with SLI warrant only minimal attention during language assessment and intervention. We suspect that our findings were influenced by the fact that, in spontaneous speech, the children had considerable control of the topic of conversation and, therefore, the lexical items to be produced. Faced with a task in which their lexical use is determined by others, the children may appear decidedly weaker in their lexical abilities. Nevertheless, these findings seem to reinforce the view that many children with SLI are more proficient in using the lexicon in spontaneous speech than in using morphosyntax (e.g., Goffman & Leonard, 2000).

Thordardottir and Ellis Weismer (2001) also reported no differences between children with SLI and a group of younger MLU-matched peers. This finding is consistent with our finding of no main effect for diagnostic group when MLUw matching was performed. However, we did find a significant interaction indicating higher $D$ scores for the children with SLI at the lower MLUw levels.

Why were group differences seen for the lower but not the higher MLUw levels? One possibility is that the age difference between the children with SLI and the younger ND children matched for MLUw may have had a greater impact at the lower MLUw levels. The children with SLI at the lower MLUw levels were 24 months older, on average, than their MLUw control counterparts, whereas the children with SLI at the higher MLUw levels were only an average of 18 months older than their MLUw-matched comparison group. More importantly, a difference of 24 months in the younger age range of the low MLUw children (36 to 60 months) probably constitutes a larger developmental difference than a difference of 18 months in the older age range of the high MLUw children (50 to 68 months). These larger developmental differences could be reflected in larger differences in world knowledge, which, in turn, could influence lexical diversity.

Somewhat different findings were obtained for samples of 100 utterances. Most notably, a diagnostic group difference was found for $D$ scores when age matching was employed. However, as ND children are known to have a higher MLU than children with SLI of the same age, there is a confound between the sample size in tokens and the $D$ score, making these results difficult to interpret. No diagnostic group differences were found for the 100-utterance samples of the children matched according to MLUw. This is notable because, by matching for MLUw, we controlled the number of tokens in the sample, providing an internal replication of the results found for 250 words.

It would be ideal for a measure such as $D$ to be independent of the size of the language sample. If a measure of lexical diversity were free of sample size fluctuations, it would allow comparison across projects with different sample sizes, thus enabling researchers to use all of the available data (Richards & Malvern, 1997; McKee et al., 2000). Unfortunately it appears that $D$ does not entirely avoid the problem of sample size influence. The possibility of a relationship between $D$ scores and the number of tokens was evident from the discrepancy in the results found in the analyses of 100-utterance and 250-word samples. Additionally, pairwise $t$ tests demonstrated that for 500-word samples, $D$ scores were higher than for 250-word samples.

The manner in which lexical items are computed in VOCD renders our findings of sample size effects rather surprising. In this method, after each computation of TTR, the lexical tokens are replaced and random sampling begins anew (a process repeated 100 times). Our findings of possible sample size effects, then, suggest that the size of the population from which words are randomly sampled must in some way shape the value of $D$.

McKee et al. (2000) showed that sample size does not have an effect on $D$ scores by calculating $D$ for the odd, even, and total words in a sample. Thus, they compared half the words present in a sample to all the words present in a sample, as we did in the 250-word versus 500-word comparison. We would speculate that the differences in the findings of the two studies are related in part to the distribution of the words in the sample and to the introduction of new themes—a possibility also raised independently by B. Richards (personal communication, June 2001). When McKee et al. compared the odd (or even) words against the total words in the sample, they were comparing two sample sizes; yet each
theme of the conversation in the larger sample was also represented in the smaller sample. However, given that the 250-word samples we used were the first 250 words of the 500-word sample, there is the possibility that themes appeared in the second half of the sample that had not appeared in the first half. These new themes would, by necessity, introduce new content words, altering the D scores of the children towards a slightly higher score. If this speculation that D scores are affected by the distribution of new words through the sample is accurate, use of D to compare samples of varying sizes would benefit from experimental control of the topics included in the child’s speech.

Although D may not be free of sample size effects, it is clearly less vulnerable to the influence of sample size than the traditional measures of NDW and TTR. The effect sizes for the latter two measures were extremely large, whereas D showed only small to medium effect sizes. Because D seems as sensitive to developmental differences as other measures and is not as heavily influenced by sample size, it deserves careful consideration as a tool in research and clinical settings.

Acknowledgments

The research reported here was supported in part by research grant 5 R01 DC00458 from the National Institute of Deafness and Other Communication Disorders, National Institutes of Health, and by training grant 5 T32 DC00030 from the National Institutes of Health. The authors thank Brian Richards, University of Reading, for offering comments on particular details of this study.

References


Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research, 38*, 1349–1355.


Received December 3, 2001
Accepted April 9, 2002
DOI: 10.1044/1092-4388(2002/075)
Contact author: Amanda J. Owen, Audiology and Speech Sciences, Heavilon Hall, Purdue University, West Lafayette, IN 47907. E-mail: ajowen@purdue.edu

**Appendix. CLAN Commands.**

**Commands to import from SALT**
1. saltin *.str +f
2. cd c:\childes\clan\salt\normal\ndconvert\n
**Commands to cut the transcript to approximately 100 utterances and approximately 250 and 500 words**

**100 utterances:**
1. cd c:\childes\clan\salt\normal\ndconvert\n
**250 words:**
1. cd c:\childes\clan\salt\normal\ndconvert\n
**500 words:**
1. cd c:\childes\clan\salt\normal\ndconvert\n
**Commands to provide the analysis**
1. freq +/-"CHI" +s"-%%%" *.cha +f
2. mlu +/-"CHI" +s"-%%%" *.cha +f
3. vocd +/-"CHI" +s"-%%%" a*.cha +f (where approximately 5 transcripts are run at once)