Applied Linguistics 25/2: 220–242

Developmental Trends in Lexical Diversity

[1]PILAR DURÁN, [1]DAVID MALVERN, [1]BRIAN RICHARDS, and [2]NGONI CHIPERE

[1]University of Reading, UK, [2]University of the West Indies, Barbados

This article discusses issues in measuring lexical diversity, before outlining an approach based on mathematical modelling that produces a measure, *D*, designed to address these problems. The procedure for obtaining values for *D* directly from transcripts using software (*vocd*) is introduced, and then applied to thirty-two children from the Bristol Study of Language Development (Wells 1985) at ten different ages. A significant developmental trend is shown for *D* and an indication is given of the average scores and ranges to be expected between the ages of 18 and 42 months and at 5 years for these L1 English speakers. The meaning attributable to further ranges of values for *D* is illustrated by analysing the lexical diversity of academic writing, and its wider application is demonstrated with examples from specific language impairment, morphological development, and foreign/second language learning.

MEASURING LEXICAL DIVERSITY

Issues in measuring lexical diversity

Although lexical diversity indices have been used in many research topics in language, measuring it is not a straightforward task. Diversity has something to do with the range of vocabulary displayed, and consequently in one sense a transcript with $T + t$ different words (types) can be said to be more diverse than one with just T. Counting the number of different words (NDW), therefore, is the most obvious method of measuring diversity, but is also the most obviously flawed. It is indisputable that a transcript containing 35 types (say) shows a greater *range* of vocabulary than one with 25. Doubts occur, however, if the transcripts differ in the total number of words used (tokens)—for example, 35 types within a transcript of 100 tokens is less clearly more diverse than 25 types in a transcript of 50 tokens. This problem of sample size when measuring relative range is common and in general there have been two approaches to solving it in language development research.

The first is to standardize the size of the samples. Samples of the same number of utterances have been compared (for example Klee 1992), but differences in the mean length of utterance (MLU) mean that the comparison is between samples of differing token size. Counting NDW from recordings over a standard time has also been suggested (Snow 1996), but that

confounds diversity with volubility and fluency. It is better to standardize on the number of tokens compared. To do so, however, requires a way of determining both the standard number of tokens in the sub-samples used and the method of selecting them (for example truncation, choosing a sequence at random, random selection of individual words, and so on). Both can be problematic (for instance, simply truncating a transcript wastes the data in the remainder), and furthermore, to effect comparison with other studies, all researchers would have to agree on both.

The second method is to consider a ratio. Here the ratio is between the number of different words (types) and the total number of words (tokens), and dividing types by tokens gives the Type Token Ratio or TTR. At first, TTR seems to be an improvement on NDW. For the two transcripts in our example TTR = 35/100 (0.35) and TTR = 25/50 (0.5), which seems to establish their comparison on a fairer basis that takes their differing size into account. A ratio provides better comparability than the simple raw value of one quantity when the quantities in the ratio come in fixed proportion regardless of their size (for example the density of a substance, that is mass/volume, remains the same regardless of the volume from which it is calculated). Language production is not like that, however. Adding an extra word to a language sample always increases the token count (N) but will only increase the type count (T) if the word has not been used before. Consequently, the token count in the denominator increases at a faster rate than the type count in the numerator and the TTR (T/N) inevitably falls. A graph of TTR against N tokens shows a monotonically descending curve, with negative slope that becomes less and less steep, but continues to fall towards zero (Figure 1).

There have been various attempts to overcome this problem. Some have standardized the sample, calculating TTR on a standard number of utterances for example (Stickler 1987), but, as with NDW, this confounds lexical diversity with utterance length, as more advanced children produce longer utterances. A better standardization is to calculate TTRs from an agreed number of tokens, but this involves the same problems as counting NDW on sub-samples of fixed size. Others have used algebraic transformations of TTR, for instance, Root TTR ($T/\sqrt{N}$) (Guiraud 1960), Corrected TTR ($T/\sqrt{2N}$) (Carroll 1964), Log TTR ($\log T/\log N$) (Herdan 1960) and a more recent method (Yoder *et al.* 1994) that divides the number of types by the number of utterances. These do little other than to alter the scale of the TTR, however, without overcoming its basic flaw. Root TTR reduces mathematically to $\sqrt{N} \times$ TTR, for example, and if L is the average number of tokens in the utterances used, Yoder's method, simplifies to $L \times$ TTR. In other words, both these are simply TTR multiplied by a scaling factor. (For more detailed discussion, see Malvern and Richards 1997; Tweedie and Baayen 1998; Vermeer 2000.)

It is of more interest to accept the relationship between TTR and token size. A key to understanding the general principle is that lexical diversity is about more than vocabulary range. Alternative terms, 'flexibility', 'vocabulary richness' (Read 2000), 'verbal creativity' (Fradis, Mihailescu, and Jipescu

1992), or 'lexical range and balance' (Crystal 1982), indicate that it has to do with how vocabulary is deployed as well as how large the vocabulary might be. Several investigations into literary studies and author stylistics are based on rank frequencies (the number of words occurring in a text once, twice, three times and so on). They have yielded indices such as Michéa's Constant and Yule's characteristic K (see Tweedie and Baayen 1998).[1] Others have tackled the problem more directly and produced models of the probability of introducing new types into language samples of increasing tokens. Perhaps the most complete is Sichel's type–token characteristic, an equation that models the relationship between number of types (T) and tokens (N) in terms of two parameters (b and c) (Sichel 1986):

$$T = \frac{2}{bc}\left[1 - e^{\left(-b\left(\{1+cN\}^{\frac{1}{2}}-1\right)\right)}\right]$$

Such indices and models have been extensively studied, particularly by Baayen, and for an examination of a number of them applied to textual similarity between and within authors, see Tweedie and Baayen (1998). These investigations tend to be carried out on very large samples of tens or hundreds of thousands of tokens, however, whereas many language applications, such as L2 or child language research, have to work with very much smaller samples.

What does emerge from this discussion is a clarification of what is required from a suitable measure of lexical diversity. It needs to:

- take into account:
  - the range of vocabulary,
  - the way it is deployed—for example the amount of repetition,
  - that TTR is a function of token size, N;
- use all the data;
- have the same sampling method for all transcripts;
- be appropriate for samples of tens to a few hundred words as well as larger samples;
- ensure all users employ the same method.

A mathematical model for lexical diversity

Surprisingly the very flaw of TTR provides the key to a valid approach to measuring lexical diversity. That TTR decreases as the token count increases, and that it does so in a regular way, suggests that TTR is a function of N. As noted above, plotting a graph of TTR against N for a transcript results in a falling curve, whose slope is at first steep but decreases with larger N. Transcripts, then, can be represented by curves in the plane which has TTR as the y-axis and N as the x-axis (see Figure 1).

All transcripts start at the point (1,1), as an utterance of one word will have TTR = 1. Eventually, the subject will exhaust his or her active vocabulary and

although more and more tokens can still continue to be uttered, no new types will be introduced. So, all curves will tend towards zero TTR for very large numbers of tokens. In between, however, the curves will differ for different subjects depending on how many types are introduced and how often they occur. Transcripts will produce curves of the same general shape, but the curve for one transcript may lie above or below that of another. Which of these represents greater diversity can be identified by considering the extreme cases (see Figure 1). The *least diverse* language sample consists of saying the same word over and over again. The number of types in such a case will remain constant at 1 while the number of tokens grows and grows, and the curve will have the equation TTR = 1/*N*. The *most diverse* instance imaginable, on the other hand, is introducing a new word every time a word is uttered. In this case the number of types is always equal to the number of tokens and the equation for such a curve is TTR = 1, a straight line parallel to the x-axis. Curves for real cases must be somewhere in between.

We now have a basic mathematical model that represents language samples and defines their lexical diversity as the combination of properties which locate the curve in the plane bounded by the two extremes—the higher it lies,

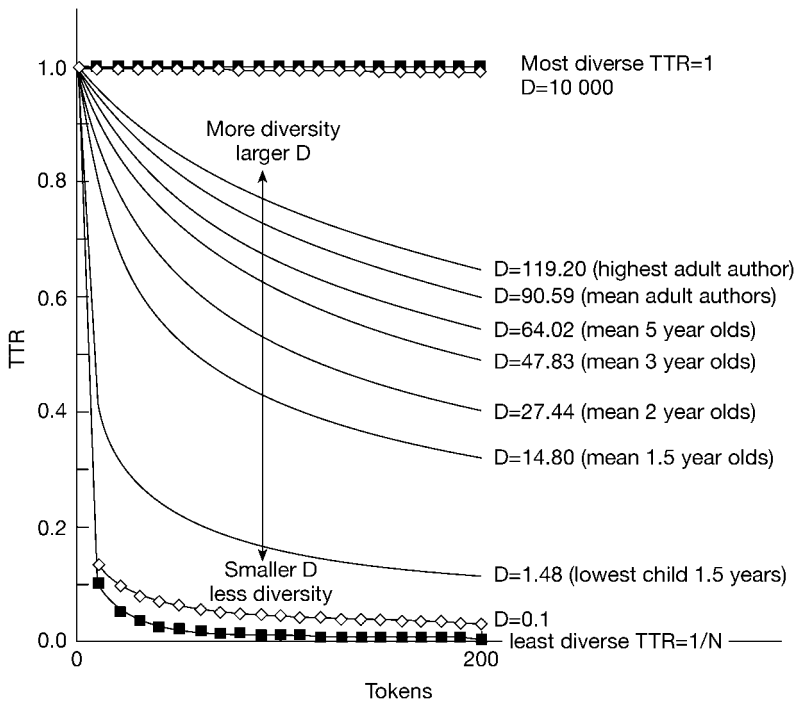

Figure 1: Model TTR plotted against samples of increasing length for different values of D

the greater the diversity (see Figure 1). To use this model for measurement requires a standard way of plotting the curve for a transcript, and a method of measuring its location (high or low).

Measuring lexical diversity

The first task is to plot the curve of TTR against *N* from a real transcript. To do this requires determining, for each value of *N*, the value of TTR which best represents sub-samples of *N* words across the whole transcript. This can be achieved by taking a number of sub-samples at random and averaging them. So that every researcher employs the same method, the values of *N* for which the curve is plotted, the number of sub-samples used for the average at each point and the method of random selection should all be fixed.

The location of a particular curve in the mathematical model can be found by comparing it with a probabilistic model for such curves. We have found that for the relatively small samples found in child language and other similar applications, Sichel's type–token characteristic provides a good starting point, but it contains two parameters (*b*, *c*), which introduces the possibility of ambiguity with two alike curves arising from different combinations of values for *b* and *c*. What is needed is an equation that has only one parameter on which the location of the curve depends. To achieve this, we have derived from Sichel's original equation, shown above, a simplification which applies to small samples:

$$TTR = \frac{D}{N}\left[\left(1 + 2\frac{N}{D}\right)^{\frac{1}{2}} - 1\right]$$

From extensive testing, we have found that for small transcripts (of the order of a few tens to a few hundred tokens) this equation fits data well. Jarvis (2002) has also tested our equation on narratives written by adolescent Swedish and Finnish students learning English as a foreign language, and found that there was a good fit between it and the data in all but five cases from 276 subjects (98 per cent). The equation describes a family of curves of the right general shape, which occupy the plane in the model with one curve lying above another throughout. How high a curve lies is determined by the value of the parameter *D* in the simplified equation—the larger *D*, the higher the curve and the greater the diversity (Figure 1). *D* itself, then, is a measure of diversity for these theoretical curves. The curve for a real transcript can now be compared to this family of theoretical curves, and be assigned the value of *D* for the particular theoretical curve to which it fits best. To do this we have devised a program, '*vocd*', which standardizes the procedures used.[2]

Vocd and *D*

Vocd is included in the CLAN suite of programs (available on the CHILDES website at *http://childes.psy.cmu.edu*) and reads language samples transcribed in

(or converted to) CHAT (see MacWhinney 2000a). Originally devised for first language research, CLAN and CHAT are increasingly being used in other fields; for example, Florence Myles and colleagues at the University of Southampton are using and adapting them to address second language issues (Marsden *et al*. 2003). ASCII files can be converted to CHAT format using the *textin* program in the CLAN suite. By use of various switches, the user can determine details of the analysis including selection of text. Having read the whole file and selected the text for analysis according to these settings, *vocd* begins with 35 tokens and plots the first point on the transcript's curve by undertaking 100 trials of randomly sampling 35 tokens from throughout the text without replacement and calculating their average TTR. The number of tokens is then increased to 36 and the calculation repeated, and so on up to 50 tokens. In all, therefore, 16 points are plotted from $N = 35$ to $N = 50$.

As the basis for the method of selecting these sub-samples, various reasons can be put forward for choosing either random sampling of individual tokens, or strings of words in sequence. The argument for sequential sampling is that it preserves the structure of the language, whereas random sampling does not (see Jarvis 2002; Tweedie and Baayen 1998). The underlying probabilistic model, however, assumes that there is a distribution of frequencies for the occurrence of words which is such as to permit two or more words to have the same or similar frequency. This automatically takes account of collocations, for example, because the contribution of such structures to the frequencies of their constituent types will be the same for each and be reflected in the random sampling with equal likelihood. The theoretical underpinning, therefore, does not require that the word sequence need be preserved in the sampling. In some transcripts, the flow of communication causes clusters of the same vocabulary item at certain points. Where that is so, sequential sub-samples would include some with such vocabulary clusters (and hence lower diversity) and others without (and of higher diversity). Although this difference would tend to be washed out by their being averaged, the reliability is decreased because the variation among sub-samples would increase. Random sampling is chosen as the default for *vocd* because, first, it matches the assumptions underlying the probabilistic model and, second, it avoids reliability problems caused by any clustering of the same vocabulary items at particular points in the transcript. None the less, for particular applications there may be good reason to require sequential sampling, and there is a version of *vocd* which calculates *D*s for sequential sub-samples.

The default values for the range of points ($N = 35$ to $N = 50$) and the number of sub-samples from which TTRs were averaged at each value of N (100 sub-samples) arose from extensive trials of alternatives. We sought the best combination to obtain reliable results while adhering to the mathematical assumptions underlying the model, including those implicit in the simplifications made in deriving the equation to be applicable to small language samples.

Finally, the transcript's curve is compared to the family of theoretical

curves, and, by varying *D* in the equation, the best fit is found between the two by the least square difference method (minimizing the sum of squares of the deviation of each data point from the curve). Extensive testing confirmed that the best-fit procedure was reliably finding a *unique* minimum at the least square difference. As the points on the curve are average TTRs obtained from random samples of word tokens, a slightly different value of *D* is obtained each time the program is run. These differences are relatively small, but reliability is optimized by calculating *D* three times by default and giving the average value as output (see McKee *et al*. 2000, for a complete flow chart). To test the stability of *D* using this method, we ran 38 L1 transcripts of children aged 32 months through *vocd* twice. Not only was there no significant difference between the two sets of scores, but there was also a perfect correspondence in their rank order.

It is important to understand that *vocd* is a measuring device. Reliability is crucial, therefore, and the combination of the default values for the sampling method, the number of sub-samples taken for each point, the standardization of the segment of the curve that is fitted and the averaging of three runs, are there to ensure that the measurement is always made in the same way for all transcripts. The calculation can make use of all the words available in the transcript by randomly sampling from the whole transcript, and it maximizes reliability by focusing on how the TTR versus Token curve is falling, rather than on any one value TTR happens to have at any single point.

Figure 1 illustrates that increasing *D* leads to higher curves indicative of greater diversity, as follows. The curve for a very small value of *D* (0.1) is plotted to demonstrate that values of *D* approaching zero lie close to the extreme of low diversity (TTR = 1/*N*). That high values of *D* lead to curves that approach the high extreme (TTR = 1) is also shown by plotting the curve for an arbitrarily high value of *D* (10,000). In practice, however, comparing *D*s to such extremes is of little use, as real language samples approach neither. In between, therefore, selected curves for different values of *D* are drawn. The values chosen correspond to key values from the smallest (*D* = 1.48 for one 18-month-old child) to the largest (*D* = 119.20 for an adult academic author) found in the cohorts discussed below.

The next sections of this article address the following questions. First, what are typical values and ranges of *D* in children acquiring their first language between 18 months and 5 years, and how do these compare with values obtained from other corpora? Second, does *D* show a consistent and significant developmental trend across this age range? Third, how well does *D* correlate with other measures of language production and comprehension at each age? Fourth, to what extent does *D* vary according to how word types are defined?

INDICATIVE VALUES OF *D*

An indication of the values and ranges to be expected of *D* will be shown, first from a detailed analysis of carefully collected language samples from a representative cohort of children, and second, but more briefly and by way of comparison, various other corpora which illustrate potential applications of *D*.

The Bristol cohort

Between 1973 and 1978, Gordon Wells directed a normative study of pre-school children in the city of Bristol in the west of England (Wells 1985). Their families were representative of the local urban population, but children whose parents did not speak English as their first language, those in full-time day care, multiple births and children with known handicaps were not included. For each child, ten recordings were made between the ages of 15 and 42 months. The children were recorded in their homes without the presence of an observer. For the current study we used the transcripts of 32 children, available in CHAT format from the CHILDES database (MacWhinney 2000b), recorded at three-monthly intervals from 18 months to 42 months. For 15 of them there is a further transcript made at 5 years. The distribution of the sample by sex and family background is shown in Table 1.

The Wells corpus has many advantages for the study reported here. These are:

- the representative sample of children;
- naturalistic recordings obtained in the home without the presence of researchers;
- a longitudinal design and regular sampling over an extended period; and
- lack of missing data.

Table 1: Distribution of sample by sex and social class

	Sex		
Family background	Girls	Boys	Total
Group 1 (highest)	4	4	8
Group 2	5	2	7
Group 3	3	4	7
Group 4 (lowest)	4	6	10
Total	16	16	32

The preparation of transcripts

It is essential that there is compatibility between the software, the transcription system and the researchers' definition of what counts as a word and what is defined as a *different* word. After scrutinizing all transcripts and obtaining complete word lists, it became clear that to analyse them in their raw form would compromise the validity of the measure, and the transcripts were edited in five ways. First, inconsistencies of spelling ('doggy' versus 'doggie') and phonetic variants of the same word ('yes', 'yeah', etc.) were standardized. Second, homographs ('may': the month of May versus the modal verb) were tagged so that they would be treated as different words. Third, self-repetition of words and phrases, for example after a false start, was coded so it could be excluded from analyses (see MacWhinney 2000a: 76–7). Fourth, the word lists had revealed a large number of non-words (laughter, pause markers, etc.) that needed to be omitted. These were listed in an 'exclude file' and filtered out by *vocd*. Finally, boundaries for inflectional morphemes were marked. This allows a choice of counting, for example, 'fall', 'falls', and 'fell' in three ways:

- at their face value as three types, or
- to strip off the regular inflections and base the count on stem forms (that is two types), or
- to remove inflections *and* treat 'fused forms' ('fell') as the root form or lemma ('fall') (that is one type).

The version chosen for the indicative values given below was the second, stem forms, as the version least likely to confound lexical diversity and the development of morphology.

Values for *D* were calculated for each child on each possible occasion. Because *vocd* carries out its curve-fitting procedure on a curve segment of 35–50 tokens, a minimum of 50 valid words are needed to supply all 16 data points. *D* could not therefore be obtained for those children who produced fewer than 50 words. The effect of this in reducing the overall sample size can be seen from Table 2. As is to be expected, the greatest loss of data is in the earliest two recordings. Inevitably, it tends to be the children who are linguistically least developed who produce the fewest words, so it is likely that at these ages the mean values for *D* are an overestimate of the mean for the whole population.

The *D* values were scrutinized at each age for sex and social background differences using Mann–Whitney *U* tests and the Kruskal-Wallis one-way analysis of variance respectively. Non-parametric statistics were used here and elsewhere in this article because the data were not normally distributed at all ages. No significant results were obtained, and consequently we can treat the cohort as a single homogenous sample representing children of these ages.

Results for the Bristol cohort

Table 2 shows descriptive statistics for the cohort at each age and Figure 2 illustrates them, showing the mean and selected percentiles against age. The steady increase in mean values is accompanied by corresponding and comparable increases in the median except between 30 and 33 months. Standard deviations initially rise and then fall, with the lowest value occurring at 60 months. Minimum and maximum values also tend to increase over time although there is a surprisingly high maximum value at 24 months.

It must be remembered that lexical diversity will be influenced by factors such as the number of topic changes and that these will, in turn, be influenced by a number of contextual factors such as the physical surroundings, nature and variety of activities, and the number and age of the participants in the interactions. In the Bristol study these were deliberately left uncontrolled, the aim being to sample a typical day in the life of each child. This is a factor that is likely to contribute to the high ranges of scores and may result in relatively high values compared to more constrained laboratory or clinical contexts.

It is a useful triangulation, therefore, to compare the Bristol *D*s with those from laboratory transcripts, such as the 38 children (mean age 30.3 months) in the 32-month directory of the New England Corpus (Dale *et al*. 1989; Snow 1989) in the CHILDES database. This contains equal numbers of boys and girls of whom 17 are working class and 21 middle class. The recordings took place in a laboratory and they sample parent and child playing with the contents of four boxes presented in succession. The descriptive statistics for Bristol scores at 30 months and the New England cohort at 30 months are respectively: means = 41.53 and 39.51; *SDs* = 16.9 and 14.12; medians = 45.59 and 41.80;

Table 2: Descriptive statistics for D *(stem forms) at each age*

Age (months)	*N*	Mean	Std. Dev.	Median	Min.	Max.
18	18	14.80	10.31	13.60	1.48	36.99
21	20	21.49	16.70	19.09	2.60	67.24
24	28	27.44	20.52	25.41	2.50	84.64
27	29	34.77	17.70	31.16	7.48	65.76
30	29	41.53	16.93	45.59	4.05	69.67
33	29	43.67	15.45	45.47	10.38	73.88
36	29	47.83	13.97	47.14	13.26	69.95
39	30	49.48	15.41	49.08	11.22	80.78
42	29	53.12	13.55	53.80	10.57	73.54
60	15	64.02	8.46	63.48	50.83	83.30

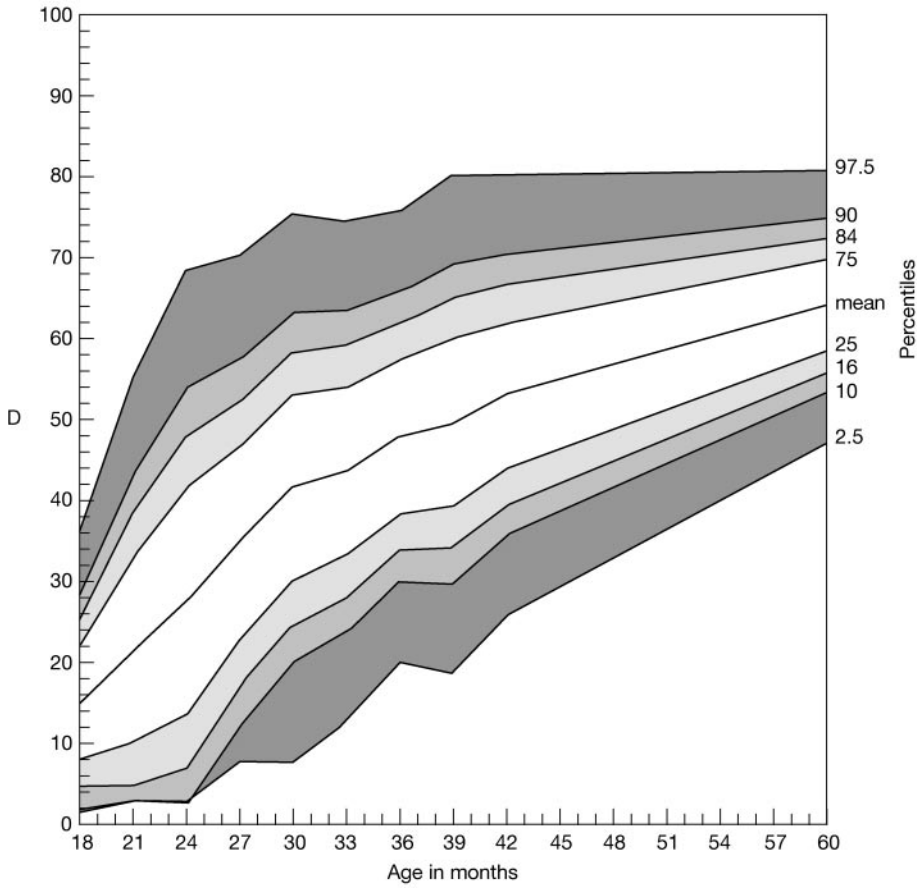

Figure 2: Values of D *against age for selected percentiles derived from the Bristol cohort*

min. = 4.05 and 4.69; max. = 69.67 and 63.37. Although the Bristol sample has the higher average scores and greater range, the means for the two at 30 months are very close, despite the differences in geographical location and context, and the other descriptive statistics are comparable. A Mann–Whitney test showed no significant difference between the two samples ($U = 500.0$; $N = 67$; $p = 0.5$).

D AS AN INDICATOR OF DEVELOPMENT

Intuitively, as a subject's facility with language and the active vocabulary expands, we would expect lexical diversity to increase. No consistent developmental trend is shown by overall values of TTR, however. For this corpus, the mean overall TTRs actually consistently fall with age, from 0.37 at

18 months to 0.25 at 60 months, because as subjects' language capacity increases, they produce more language and the samples obtained in the set time contain more tokens. TTR is a decreasing function of the number of tokens in the sample, however, and the overall TTRs of older children are therefore depressed. This is a direct consequence of the flaw in using overall TTR, and one reason why trends in lexical diversity have not emerged in earlier research (for example Fletcher 1985; Miller 1981).

In contrast, inspection of Figure 2 indicates a continual improvement in *D* with advancing age. A Friedman analysis of variance by ranks (Siegel and Castellan 1988) shows a highly significant effect for age (Chi-square = 51.06, $df = 9$, $p < .001$), and a Page test for ordered alternatives across the first nine recordings (Siegel and Castellan 1988) indicates the trend is statistically significant ($L = 7405$; $k = 9$; $N = 32$; $z_L = 1.708$; $p < .05$). To see to what extent the trend held true for individual children, we computed Spearman rank order correlations between *D* and age for each child separately. The values for Spearman's *rho* were statistically significant for 26 out of the 30 children who supplied sufficient data points to compute correlations. Unlike overall TTR, then, *D* shows the obvious characteristic of a developmental measure, being significantly associated with age for the cohort as a whole and for individuals in 87 per cent of cases. We believe that this is the first time that a statistically significant, clear, consistent, and continuous trend has been demonstrated for lexical diversity for early language development across these ages.

A further advantage of using the Wells corpus is the availability for the first nine recording points (ages 18–42 months) of three language measures used in the Bristol project. They are the Mean Length of Structured Utterances (MLUS), the Bristol Language Development Scale (BLADES) and, at 39 months, the English Picture Vocabulary Test (EPVT). These can be correlated with *D* to test its capacity to reflect development, and each will be considered in turn.

The Mean Length of Structured Utterances (MLUS) was developed by Wells as an alternative to Brown's (1973) MLU. Wells' MLUS scores were chosen in preference to calculating traditional MLU, because the original project team were meticulous in checking their reliability and MLUS has been well validated and favourably compared with the more traditional MLU (see Wells 1985: 120–5).

The overall correlation between *D* and MLUS for the children's data pooled across the first nine recording occasions is 0.752 ($N = 241$; $p < .001$). Separate correlations at each age are shown in Table 3. After the age of 18 months these are, in most cases, moderate to strong and are highly significant. Correlations for individual children were statistically significant for 24 out of the 30 children who provided a sufficient number of data points for a correlational analysis.

The correlations between *D* and MLUS, then, are highly significant, moderate to strong at most ages and significant for most (80 per cent) individual children. The exception is at 18 months, the age with the smallest

Table 3: Rank order correlations between D *(stem forms) and other language measures at each age*

Age (months)	MLUS	Scale Score	*N*
18	.340	.381	18
21	.585**	.636***	20
24	.673***	.738***	28
27	.689***	.692***	29
30	.755***	.747***	29
33	.473**	.546***	29
36	.333*	.422*	29
39	.558***	.615***	30
42	.544***	.255	29

* $p < .05$; ** $p < .01$; *** $p < .001$

sample of children, which reduces the power of the study to detect significant effects. Moreover, both MLUS and *D* have their lowest ranges and standard deviations at these ages (see Table 2). It is possible, therefore, that the rank orders are less reliable in this case.

The second measure adopted from the Bristol study is the Bristol Language Development Scales (BLADES) (Gutfreund *et al.* 1989). BLADES contains profiles of syntactic, semantic, and pragmatic development, each divided into ten levels of attainment and these profiles are combined into a single ordinal scale. The items included in each scale, and their levels, are derived from an analysis of the transcripts of the 60 younger children in the younger cohort of the Bristol study. Criteria for the inclusion of items included their saliency (the ease with which they could be identified in spontaneous speech), frequency (items should occur frequently once they emerge in child speech), and, most importantly, order of emergence (items at each level should be strongly and statistically significantly ordered in relation to each other) (Wells 1985).

The overall correlation between *D* and the scale for the data pooled across the first nine occasions is 0.765 ($N = 241$; $p < .001$). Separate correlations at each age are shown in Table 3. These follow much the same pattern as MLUS, except at 42 months where the result is not significant. As with MLUS, 24 out of 30 correlations for individual children are statistically significant.

Again, there is no significant correlation at 18 months, but in addition to the points made above, an examination of the frequency distribution of scale scores for the 18 children who went into this analysis shows that eight were tied on level two and five were tied on level three. It seems that for this sub-sample of the Bristol cohorts, the scale fails to discriminate well at this age.

The other age at which *D* failed to correlate with the scale was 42 months, in spite of the larger sample of 29. *D* correlates well with MLUS at this point, so the reason seems likely to lie in the distribution of scale scores rather than with *D*. The fact that 22 out of the 29 children are clustered at levels 7 and 8 of the scale supports this interpretation.

The third and final measure adopted is the English Picture Vocabulary Test (EPVT) (Brimer and Dunn 1963). EPVT is a standardized test of receptive vocabulary that was administered at 39 months. The correlation between *D* at 39 months and standardized EPVT was computed, therefore. This was weak and non-significant (*rho* = .218; *N* = 26; ns). It may seem surprising that there was no correlation with the EPVT as, unlike MLUS and the Bristol scale, it is specifically targeted towards vocabulary. One reason for this lack of correlation may in part lie in the contrasting situations in which the data were collected—spontaneous speech sampling versus a relatively formal testing situation. Wells (1985: 333) found that correlations between measures of spontaneous speech and EPVT were relatively low at both 39 and 60 months for the Bristol cohorts and he comments on the greater likelihood of children in the lower family background group to be more ill at ease in the testing situation. This view is supported by a higher correlation for family background with EPVT than with spontaneous speech measures (Wells 1985: 459), and raises concerns about the reliability and validity of formal testing with very young children. More immediately, however, *D* measures how diversely vocabulary resources are *deployed* and is likely to be related more to productive language than to assessments of receptive vocabulary like EPVT. This is supported by an analysis of the New England data referred to above. A statistically significant correlation was found between *D* at 30 months and the MacArthur Communicative Development Inventory (CDI) (Fenson *et al.* 1993) *production* scores at 14 months but there was no correlation with *receptive* vocabulary (Richards and Malvern 2003). (For further discussion of early dissociations between comprehension and production in L1, see Bates *et al.* 1988).

D, then, correlates significantly with MLUS and BLADES over the ages where they can be relied on as general indicators of development. That these correlations do not approach unity, however, shows that *D* is measuring something which differs from both, and the EPVT and CDI results suggest that this is related more to productive capability. This is precisely what is to be expected of a measure that specifically reflects the extent to which the active vocabulary is employed and how richly it is deployed. These results, taken together, provide powerful support for *D*'s usefulness as a developmental measure.

ILLUSTRATIVE APPLICATIONS OF *D*

To illustrate potential applications of *D* and to provide further benchmarks for the interpretation of values of *D*, brief presentations of data will be given from

four other linguistic fields: academic writing, Specific Language Impairment (SLI), morphological development, and foreign/second language learning.

Adult academic writing

Some indication of the upper range of *D* can be offered, by comparing the values of *D* for the Bristol children with those obtained from adult academic writing. The mean lexical diversity for the spoken language of children ranged from mean *D* = 14.8 at 18 months to 64.02 at 5 years. We would obviously expect adults writing academic text to use language richly, avoiding unnecessary repetition, and generally to display a diversity substantially higher than that of an average 5 year old. To make this comparison, 23 samples of academic writing were analysed. *D* was found to range from 69.74 to 119.20 with a mean of 90.59 and a standard deviation of 10.79, fulfilling that prediction. Figure 5 combines these data with the Bristol results to summarize indicative comparison groups for ranges of values for *D*.

Specific language impairment (SLI)

Data for three language-matched pairs of siblings, one younger normally developing child and one older with SLI, were obtained from the Conti-2 Corpus of the CHILDES database (Conti-Ramsden and Dykins 1991). There are seven recordings for each, gathered over a period of two years. As is to be expected, TTR fails to correlate with age for any of the subjects and does not distinguish between the normal group and those with SLI. There are overall trends in lexical diversity as measured by *D*, however, and significant correlations between *D* and age for all but one child (with SLI). Two of the children with SLI (SLI 1 and SLI 2) produce *D*s entirely within the range found for the Bristol children at 18 months, with their highest values approximating to the equivalent of means for 23 months and 26 months (estimated from Figure 2), but they achieve these at the ages of 78 and 93 months respectively—that is 4.5–5.5 years later. The third child with SLI (SLI 3) has *D*s within the range found for Bristol children at 21 months, with his highest score achieved at age 73 months, approximating to the Bristol mean for 42 months, that is 2.5 years later. The first six recording points for this child fall within the age range of the Bristol data, when his *D*s all fall below the 2.5 percentile as shown in Figure 2. Figure 3 plots *D* for these three children with SLI against age alongside the means for the Bristol corpus, exposing these delays.

Morphological development

As described above, in preparing the Bristol transcripts, boundaries for inflectional morphemes were marked. This means that *D*s could be obtained for three versions of the transcripts. The first took the words in the files at their face value, that is in their fully transcribed form. Therefore, 'fall', 'fall-s',

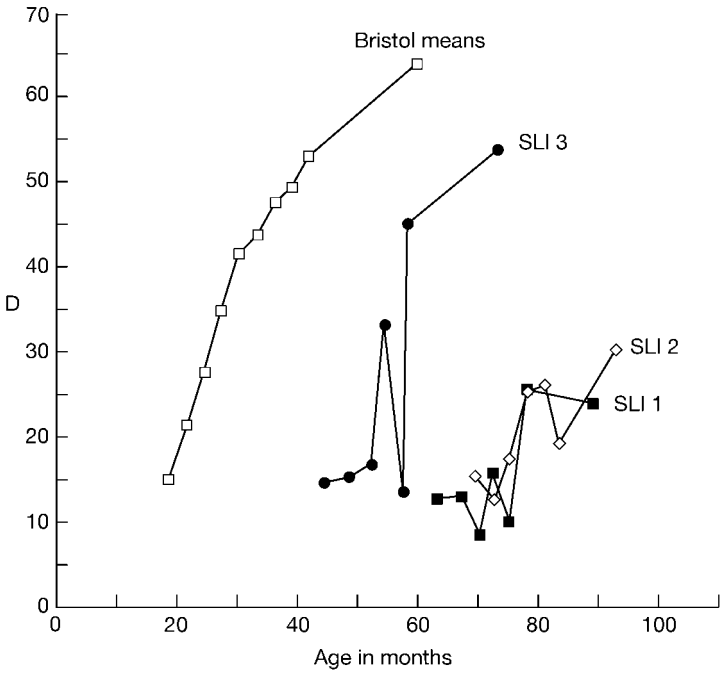

Figure 3: D *against age, for three children with SLI and means for the Bristol cohort*

and 'fell' were counted as three tokens and three types. This version will be referred to as '*D* (inflected forms)'. The second, referred to as '*D* (stem forms)', strips inflections from the ends of nouns and verbs, and would count three tokens and two types (that is 'fall' and 'fell'). The third, '*D* (root forms)', additionally reduced irregular verbs and nouns to their root form, giving us three tokens and one type for the example above (that is 'fall'). It can be confidently predicted that, because *D* (inflected forms) produces the most word types from a given number of tokens this will give rise to the highest values. Conversely, the *D* (root forms) gives the fewest word types from the same token count and will produce the lowest, with *D* (stem forms) lying in between. As the differences between the three forms are caused by the use of inflections, they would be increased both by extending the range of inflections used and by applying them to a greater range of stems. It is to be expected, therefore, that the differences will grow with the development of inflectional morphology.

Figure 4 plots the mean for the three versions of *D* against age for the Bristol children and demonstrates a consistent upward trajectory for all three versions. The relative position of the three curves in Figure 5 bears out the predictions made above, with the inflected forms being highest, and the root forms the lowest. These differences are barely discernible at 18 months and

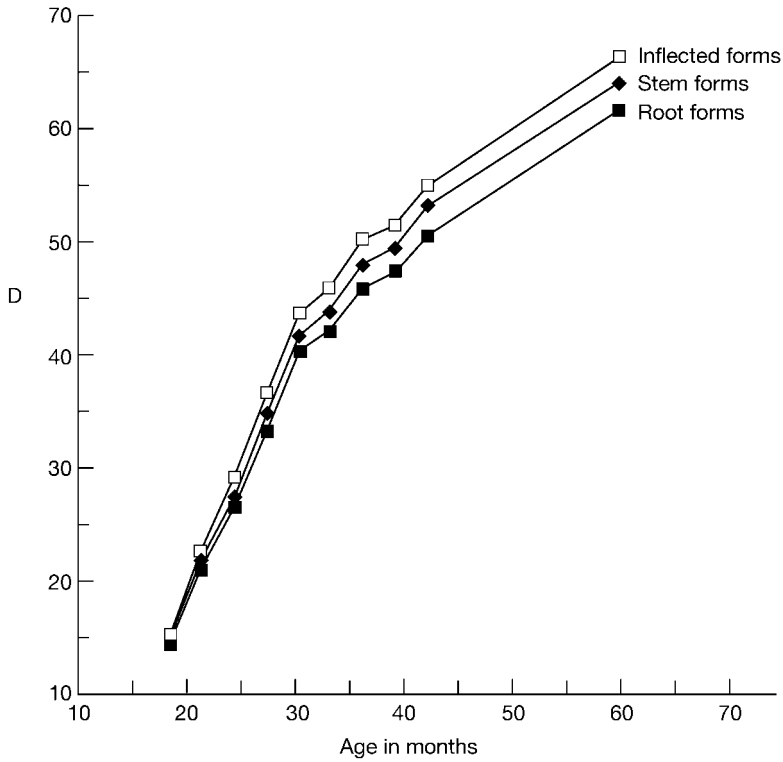

Figure 4: Mean D *for inflected forms, stem forms, and root forms against age*

are relatively small at 21 months, when little inflectional morphology is apparent. Over time, however, the gap between the three versions widens as an increasing repertoire of inflections is applied to a broader range of verbs and nouns. A Wilcoxon signed ranks test was used at each age to test whether the three measures differed significantly from each other. There were no differences at 18 months and the difference between the stem form and the root form was not significant at 21 months. From then on, however, all differences between all three versions were highly significant at $p < .001$. The widening gap between these versions of *D* suggests that difference scores, particularly those between inflected forms and stem forms, are valid indicators of productivity in early morphological development (Richards and Malvern 2004).

Foreign and second language learners

Transcripts from a group of second language learners were made available to us by Peter Skehan and Benny Teasdale. The sample consists of 32 pre-intermediate level learners of English from a variety of L1 backgrounds, aged 18–30 and attending classes for six hours per week. The transcripts were

obtained from recordings of pairs of students conducting an oral decision making task (Foster and Skehan 1996). Self-repetition and back channels were excluded from the analysis. Although the minimum *D* was 35.78 and the maximum 91.99, the mean *D*, 56.58, is slightly higher than that of the 42-month-old first language English speakers (53.12) in the Bristol cohort, and with a standard deviation of 12.10, the spread from 10th to 90th percentiles is very similar to that age range.

D can, of course, be calculated for languages other than English, and the software has been used by the Spencer Foundation Project 'Developing Literacy in Different Contexts and in Different Languages' for Dutch, French, Hebrew, Icelandic, Spanish, and Swedish (Berman and Verhoeven, 2002; Strömqvist *et al*. 2002), and others have used it with Cantonese (Klee *et al*. submitted) and even a polysynthetic language such as Inuktitut (Allen 1998), for which the analysis was conducted on morphemes rather than words. Work is in progress with Finnish, German, and Russian. By way of contrast with the ESL learners, then, a set of transcripts from 16-year-old British learners of French as a foreign language were obtained from recordings of their General Certificate of Secondary Education (GCSE) oral interviews. This examination is taken after 5 years of study and consists of 'free conversation' with the teacher. For a full analysis, comparing student *D*s with those of the teachers and a wide range of other language measures, see Malvern and Richards (2002), but for this purpose, and even though one must exercise caution in making cross-linguistic comparisons of language measures, the descriptive statistics will suffice to show that the results are comparable to those of the ESL cohort. Twenty-seven students produced sufficient tokens to calculate *D* in the same way as for the ESL transcripts, and their values ranged from 29.64 to 87.35 with a mean of 56.28 and standard deviation of 14.87. To illustrate their similarity, these two cohorts are included in Figure 5, which provides a summary by plotting the means and sub-ranges between the 10th and 90th percentiles for selected ages from the Bristol cohort and the adult writing. This helps to give meaning to any particular value of *D* by reference to the sub-range in which it lies.

IN CONCLUSION

Lexical diversity measures are applied to a wide variety of topics in applied linguistics. The problems with its measurement, however, have meant that results have often been difficult to interpret and, at times, confusing. The methodological approach to its measurement described in this article addresses those problems and produces a valid index, *D*, which can be calculated with good reliability even for short transcripts. The range of potential applications is extensive. Although only a few examples can be given here, the kind of work which would benefit from having such a measure can be illustrated by reference to recent studies which have employed lexical diversity.

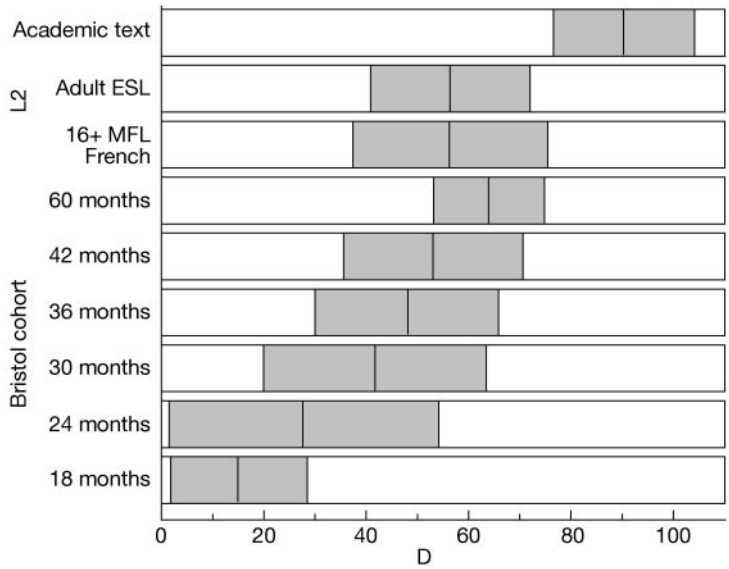

Figure 5: Means and sub-ranges (10th–90th percentiles) of D *for various cohorts*

As an indicator of development, *D* can be applied to first and foreign or second language learners. It can be useful in studies of children's writing development and comparisons of different genres, both in children developing normally and in children with learning disabilities (for example Scott and Windsor 2000). Berman and Verhoeven (2002) have already applied *vocd* in a crosslinguistic study of seven languages that compared four ages, two genres, and two modalities (speech versus writing). In second language research, lexical diversity can be used as an outcome measure in comparing the influence of task type and task conditions on language performance (for example Foster and Skehan 1996), as a criterion for assessing the validity of vocabulary tests (Ukrainetz and Blomquist 2002), for investigating the richness of input in foreign language classrooms (Shortreed 1993) and for exploring relationships between language use by students and examiners during oral production (Malvern and Richards 2002).

D also has potential as a dependent variable in experimental research on language impairments. Lexical diversity has been measured in such work as the treatment of stuttering (for example Onslow, Ratner, and Packman 2001) and language delays (for example Ellis Weismer *et al*. 1993). In particular, lexical diversity measures have been applied to comparisons between normal and impaired populations leading to description and diagnosis (for example Le Normand and Cohen 1999, for language delay in pre-term children; Ouellet *et al*. 2000, for developmental dysphasia; Delaney-Black *et al*. 2000, for prenatal exposure to cocaine; Bucks *et al*. 2000, for the speech of patients with

Dementia of Alzheimer type; Geers, Spehar, and Sedey 2002, for the progress of children with cochlear implants; Stokes and Fletcher 2000, for children with Specific Language Impairment; Holmes and Singh 1996, for aphasia in adults).

The capacity to select the text in transcripts from which *D* is calculated, permits separate analyses for different speakers and for separate word classes, which would facilitate multicontextual or crosslinguistic first language research. An example is the analysis of diversity of different word classes in early child speech and maternal input in order to address the issue of the universality of a noun bias in early vocabulary composition (for example Tardif *et al*. 1999).

In sum, lexical diversity measures are widespread in applied linguistic research and practice, ranging across work as diverse as First and Second Language Development, Language Impairment, Literacy, Education, Authorship Studies, and Forensic Linguistics. *D* offers a robust metric of 'vocabulary range and balance' for such research and for applications where quantification of lexical diversity is required.

Final version received December 2003

ACKNOWLEDGEMENTS

The research was supported by grants from The University of Reading and the Economic and Social Research Council (R000221995; R000238260). We should like to thank Gerard McKee for writing the *vocd* program and Brian MacWhinney and Leonid Spektor for incorporating it in CLAN. We are also grateful to Gordon Wells for his support and permission to use the Bristol data for this project. The ESL data originated from Peter Skehan's project 'Syntactic and pragmatic mode in task-based foreign language learning' (Foster and Skehan 1996) and the analysis performed in collaboration with Benny Teasdale. We would like to express our thanks to Suzanne Graham, Mair Richards and three anonymous reviewers for their comments on a previous version of this article.

NOTES

1 Michéa noticed what appeared to be a constancy in the ratio of the number of word types occurring twice and the number of types in a text. In fact, the ratio is not constant, but rises with increasing word tokens and then falls very slowly giving an apparent constancy only for a range of word tokens. Yule's characteristic K, a measure of repetition based on factors such as the probability that two tokens chosen at random will be the same type, is also an apparent constant.

2 *Vocd* was written by Gerard McKee of The University of Reading Computer Science Department. The basic mathematical model consists of a set of curves with TTR as the y-axis and *N* as the x-axis, which fall from the point (1,1) with decreasing slope within the space between the horizontal line TTR = 1 and the curve TTR = 1/*N*. In the model, each curve represents a language sample and lexical diversity is defined as the combination of properties which locate a curve in the space bounded by the two extremes—the higher the curve, the greater the diversity. *Vocd* operationalizes the model using our equation (as given in the text) so that all the properties which contribute to diversity (range of vocabulary, its deployment, amount of repetition, and so on) are

combined in a single number, *D*, calculated in a standardized way.

Other appropriate equations could be used in our model. Tweedie and Baayen (1998) considered a number of lexical diversity measures, some of which (including Sichel's original equation) could yield appropriate equations, and found them all wanting over a text of 20,000 words. For short texts (of 200–300 tokens), however, Jarvis (2002) found that out of five possibilities he tested, two performed well in fitting curves to real language data, namely our equation and one other ($TTR = N^{\frac{-1}{U}(\log N)}$) derived from the definition of *U*, the Uber index. *Vocd* gives an indication of the goodness of fit for our equation, by including the minimum least square value (MLSV) in the output for each of the three iterations of curve fitting (which are averaged for the final *D*). MLSV is the sum of the squared difference between the TTR calculated by sampling from the actual text and that predicted by the optimum (best fit) value of *D* for each of the points on the curve.

REFERENCES

Allen, S. E. 1998. 'Linguistic change in Inuktitut narratives across ages.' Paper presented at the winter meeting of the Society for the Study of the Indigenous Languages of the Americas, New York, January 1998.

Bates, E., I. Bretherton, and **L. Snyder.** 1988. *From First Words to Grammar: Individual Differences and Dissociable Mechanisms*. Cambridge: Cambridge University Press.

Berman, R. A. and **L. Verhoeven.** 2002. 'Cross-linguistic perspectives on the development of text-production abilities: Speech and writing,' *Written Language and Literacy* 5: 1–43.

Brown, R. 1973. *A First Language: The Early Stages*. London: Allen & Unwin.

Brimer, M. A. and **L. Dunn.** 1963. *English Picture Vocabulary Test*. Windsor: NFER.

Bucks, R. S., S. Singh, J. M. Cuerden, and **G. K. Wilcock.** 2000. 'Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance,' *Aphasiology* 14: 71–91.

Carroll, J. B. 1964. *Language and Thought*. Englewood Cliffs, NJ: Prentice-Hall.

Conti-Ramsden, G. and **J. Dykins.** 1991. 'Mother–child interactions with language impaired children and their siblings,' *British Journal of Disorders of Communication* 26: 337–54.

Crystal, D. 1982. *Profiling Linguistic Disability*. London: Edward Arnold.

Dale, P., E. Bates, S. Reznick, and **C. Morisset.** 1989. 'The validity of a parent report instrument of child language at twenty months,' *Journal of Child Language* 16: 239–49.

Delaney-Black, V., C. Covington, T. Templin, T. Kershaw, B. Nordstrom-Klee, J. Ager, N. Clark, A. Surendran, S. Martier, and **R. J. Sokol.** 2000. 'Expressive language development of children exposed to cocaine prenatally: Literature review and report of a prospective cohort study,' *Journal of Communication Disorders* 33: 463–81.

Ellis Weismer, S., J. Murray-Branch, and **J. F. Miller.** 1993. 'Comparison of two methods for promoting productive vocabulary in late talkers,' *Journal of Speech and Hearing Research* 36: 1037–50.

Fenson, L., P. S. Dale, J. S. Reznick, D. Thal, E. Bates, P. P. Hartung, S. Pethick, and **J. S. Reilly.** 1993. *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual*. San Diego: Singular Publishing Group.

Fletcher, P. 1985. *A Child's Learning of English*. Oxford: Blackwell.

Foster, P. and **P. Skehan.** 1996. 'The influence of planning and task type on second language performance,' *Studies in Second Language Acquisition* 18: 299–323.

Fradis, A., L. Mihailescu, and **I. Jipescu.** 1992. 'The distribution of major grammatical classes in the vocabulary of Romanian aphasic patients,' *Aphasiology* 6: 477–89.

Geers, A., B. Spehar, and **A. Sedey.** 2002. 'Use of speech by children from total communication programs who wear cochlear implants,' *American Journal of Speech-Language Pathology* 11: 50–8.

Guiraud, P. 1960. *Problèmes et Méthodes de la Statistique Linguistique*. Dordrecht: D. Reidel.

Gutfreund, M., M. Harrison, and **C. G. Wells.**

1989. *Bristol Language Development Scales*. Windsor: NFER-Nelson.

Herdan, G. 1960. *Type-Token Mathematics: A Textbook of Mathematical Linguistics*. The Hague: Mouton.

Holmes, D. I. and **S. Singh.** 1996. 'A stylometric analysis of conversational speech of aphasic patients,' *Literary and Linguistic Computing* 11: 133–40.

Jarvis, S. 2002. 'Short texts, best-fitting curves and new measures of lexical diversity,' *Language Testing* 19: 57–84.

Klee, T. 1992. 'Developmental and diagnostic characteristics of quantitative measures of children's language production,' *Topics in Language Disorders* 12: 28–41.

Klee, T., S. F. Stokes, A. M. Y. Wong, P. Fletcher, and **W. J. Gavin.** submitted. 'Utterance length and lexical diversity in Cantonese-speaking children with and without SLI.'

Le Normand, M.-T. and **H. Cohen.** 1999. 'The delayed emergence of lexical morphology in preterm children: The case of verbs.' *Journal of Neurolinguistics* 12: 235–46.

McKee, G, D. D. Malvern, and **B. J. Richards.** 2000. 'Measuring vocabulary diversity using dedicated software,' *Literary and Linguistic Computing* 15: 323–38.

MacWhinney, B. 2000a. *The CHILDES Project: Tools for Analyzing Talk. Volume I: Transcription Format and Programs* 3rd edn. Mahwah, NJ: Erlbaum.

MacWhinney, B. 2000b. *The CHILDES Project: Tools for Analyzing Talk. Volume II: The Database* 3rd edn. Mahwah, NJ: Erlbaum.

Malvern, D. D. and **B. J. Richards.** 1997. 'A new measure of lexical diversity' in A. Ryan and A. Wray (eds): *Evolving Models of Language*. Clevedon: Multilingual Matters, pp. 58–71.

Malvern, D. D. and **B. J. Richards.** 2002. 'Investigating accommodation in language proficiency interviews using a new measure of lexical diversity,' *Language Testing* 19: 85–104.

Marsden, E., F. Myles, S. Rule and **R. Mitchell.** 2003. 'Using CHILDES tools for researching second language acquisition' in S. Sarangi and T. van Leeuwen (eds): *Applied Linguistics and Communities of Practice*. London: BAAL/Continuum.

Miller, J. F. 1981. *Assessing Language Production: Experimental Procedures*. London: Arnold.

Onslow, M., N. B. Ratner, and **A. Packman.** 2001. 'Changes in linguistic variables during operant, laboratory control of stuttering in children,' *Clinical Linguistics and Phonetics* 15: 651–62.

Ouellet, C., H. Cohen, M.-T. Le Normand, and **C. Braun.** 2000. 'Asynchronous language acquisition in developmental dysphasia,' *Brain and Cognition* 43: 352–7.

Read, J. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Richards, B. J. and **D. D. Malvern.** 2004. 'Investigating the validity of a new measure of lexical diversity for root and inflected forms' in K. Trott, S. Dobbinson, and P. Griffiths (eds): *The Child Language Reader*. London: Routledge, pp. 81–9.

Scott, C. M. and **J. Windsor.** 2000. 'General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities.' *Journal of Speech, Language and Hearing Research* 43: 324–39.

Shortreed, I. M. 1993. 'Variation in foreigner talk input: the effects of task and proficiency' in G. Crookes and S. M. Gass (eds): *Tasks and Language Learning: Integrating theory and practice*. Clevedon: Multilingual Matters, pp. 96–122.

Sichel, H. S. 1986. 'Word frequency distributions and type-token characteristics,' *Mathematical Scientist* 11: 45–72.

Siegel, S. and **J. N. Castellan, Jr.** 1988. *Nonparametric Statistics for the Behavioral Sciences* 2nd edn. New York: McGraw-Hill.

Snow, C. E. 1989. 'Imitativeness: A trait or a skill?' in G. E. Speidel and K. E. Nelson (eds): *The Many Faces of Imitation in Language Learning*. New York: Springer-Verlag, pp. 73–90.

Snow, C. E. 1996. 'Change in child language and child linguists' in H. Coleman and L. Cameron (eds): *Change and Language*. Clevedon: Multilingual Matters, pp. 75–88.

Stickler, K. R. 1987. *Guide to Analysis of Language Transcripts*. Eau Claire, WI: Thinking Publications.

Stokes, S. F. and **P. Fletcher.** 2000. 'Lexical diversity and productivity in Cantonese-speaking children with specific language impairment,' *International Journal of Language and Communication Disorders* 35: 527–41.

Strömqvist, S., V. Johansson, S. Kriz, H. Ragnarsdóttir, R. Aisenman, and

D. Ravid. 2002. 'Toward a crosslinguistic comparison of lexical quanta in speech and writing,' *Written Language and Literacy* 5: 45–67.

Tardif, T., S. Gelman, and **F. Xu.** 1999. 'Putting the ''noun bias'' in context: A comparison of English and Mandarin,' *Child Development* 70: 620–35.

Tweedie, F. J. and **R. H. Baayen.** 1998. 'How variable may a constant be? Measures of lexical richness in perspective,' *Computers and the Humanities* 32: 323–52.

Ukrainetz, T. A., and **C. Blomquist.** 2002. 'The criterion validity of four vocabulary tests compared with a language sample,' *Child Language Teaching and Therapy* 18: 59–78.

Vermeer, A. 2000. 'Coming to grips with lexical richness in spontaneous speech data,' *Language Testing* 17: 65–83.

Wells, C. G. 1985. *Language Development in the Pre-school Years*. Cambridge: Cambridge University Press.

Yoder, P. J., B. Davies, and **K. Bishop.** 1994. 'Adult interaction style effects on the language sampling and transcription process with children who have developmental disabilities,' *American Journal on Mental Retardation* 99: 270–82.