Measuring Vocabulary Diversity Using Dedicated Software

Article *in* Literary and Linguistic Computing · September 2000

DOI: 10.1093/llc/15.3.323

Gerard T. Mckee
Baze University Abuja
94 PUBLICATIONS **603** CITATIONS

Brian James Richards
University of Reading
47 PUBLICATIONS **1,397** CITATIONS

Measuring Vocabulary Diversity Using Dedicated Software

Gerard McKee, David Malvern, and Brian Richards
The University of Reading, Reading, UK

Abstract

This paper describes software (*vocd*) that implements a solution to problems encountered in quantifying vocabulary diversity. Researchers in various fields of linguistic enquiry have calculated vocabulary diversity using the ratio of different words (Types) to total words (Tokens)—the Type–Token Ratio (TTR)—or measures derived from it. Such measures are flawed, however, because the values obtained are related to the number of words in the sample. The paper shows how the relationship between TTR and sample size can be described by a new mathematical model, which in turn leads to an innovative method of measuring vocabulary diversity. The software automates measurement from transcripts prepared in a widely used computer-readable set of conventions: the CHAT format of the CHILDES project. Options in *vocd* are described to show how the user can determine which linguistic items will count as valid types and tokens in the analysis. The new measure is calculated by, first, randomly sampling words from the transcript to produce a curve of the TTR against Tokens for the empirical data. Then the software finds the best fit between this empirical curve and theoretical curves calculated from the model by adjusting the value of a parameter. The parameter, *D*, is shown to be a valid and reliable measure of vocabulary diversity without the problems of sample size found with previous methods.

1 Introduction

Measurements of vocabulary diversity are used in a wide range of linguistic research including child language development, language impairment, foreign and second language learning, the development of literacy, authorship studies, forensic linguistics, stylistics, studies of schizophrenia, and many other areas. Many commonly used measures have been based on the ratio of different words (Types) to the total number of words (Tokens), known as the Type–Token Ratio (TTR). TTRs are output by a number of widely available computer programs dedicated to linguistic analysis, such as Systematic Analysis of Language Transcripts (SALT) (Miller and Chapman, 1993), the Oxford Concordance Program (Hockey, 1988), and the Computerized Language Analysis (CLAN) programs (MacWhinney and Snow, 1990; MacWhinney, 1995). Unfortunately,

Correspondence:
Professor B. J. Richards,
The University of Reading,
School of Education,
Bulmershe Court,
Reading RG6 1HY, UK.
E-mail:
b j richards@reading.ac.uk

such measures, including mathematical transformations of the TTR such as Root TTR (Guiraud, 1960) and Corrected TTR (Carroll, 1964), are functions of the number of tokens in the transcript or language sample—samples containing larger numbers of tokens give lower values for TTR and vice versa (for a demonstration, see Richards, 1987; Richards and Malvern, 1997a; Tweedie and Baayen, 1998). The reason for this is very simple—as longer and longer samples of language are produced, more and more of the active vocabulary is likely to be included and the available pool of *new* word types that can be introduced steadily diminishes. It is obvious that once a sample is large enough to have included all of the subject's active vocabulary, any further sampling of tokens can only result in a hyperbolic decline in the values for TTR. But, it is also the case that however small the sample is, as more and more tokens are taken, the likelihood is that (because of repetition of previously ıncluded types) the cumulative number of types will increase at a slower rate than the number of tokens and the TTR values inevitably fall. This problem has frequently distorted research findings (Richards and Malvern, 1997b), but, unfortunately, research is still being published where TTR has been used without any attempt to control for variation in the size of language samples (see Tweedie and Baayen, 1998, for examples). Previous attempts to overcome the problem, for example by standardizing the number of tokens to be analysed from each subject, have failed to ensure that measures are comparable across researchers who use different baselınes of tokens, and inevitably waste data ın reducing analyses to the size of the smallest sample.

The *vocd* program was developed to overcome these problems as part of the project 'A new research tool: mathematical modelling in the measurement of vocabulary diversity'. The approach is based on an analysis of the probability of new vocabulary being introduced into longer and longer samples of speech or writing. This yields a mathematical model of how TTR varies wıth token size. By comparing the mathematical model with empirical data in a transcript, it provıdes a new measure of vocabulary diversity that we refer to as *D*. The measure has three advantages: (1) it is not a function of the number of words in the sample; (2) it uses all the data available; (3) it is more informative because, as opposed to a single value of TTR, it represents how the TTR varies over a range of token size for each speaker or writer (i.e. it is based on the TTR versus token curve calculated from data for the transcrıpt as a whole rather than a particular TTR value on ıt). *D* has been shown to be superior to previous measures in both avoiding the inherent flaw in raw TTR with varying sample sizes and in discriminating across a wide range of language learners and users (Richards and Malvern, 1998; Malvern and Richards, 2000).

2 Origin of the Measure, *D*

TTRs inevitably fall with increasing size of token sample and consequently any single value of TTR lacks reliability, as it will depend on the length in words of the language sample used. A graph of TTR against tokens (*N*) for a transcript will lie in a curve beginning at the point (1.1) and falling with a negative gradient that becomes progressively less steep (see Malvern and Richards, 1997a). All language samples will follow this trend, but transcripts from speakers or writers with high vocabulary diversity will produce curves that lie above those with low diversity (see Fig. 1). That TTR falls in a predictable way as the token size increases provides the basis for our approach to finding a valid and reliable measure. Various analyses have been carried out to model the relationship between growth in types and tokens, and a useful summary of many of these has been given by Tweedie and Baayen (1998). In particular, they investigated various transformations of TTR, including those proposed or cited by Guiraud (1954), Herdan (1960), Maas (1972), Brunet (1978), Tuldava (1978), and Dugast (1979); and probabilistic measures based on an assumed frequency distribution of different types, such as the methods of Yule (1944), Good (1953), Sichel (1975), and Honoré (1979). The analysis of Tweedie and Baayen (1998) shows that all these measures have problems with very large sample sizes. This is perhaps not surprising. As has been pointed out above, once the active vocabulary is exhausted the total number of types remains constant, thereafter the type–token characteristic curve falls hyperbolically and all these models no longer apply exactly. Very large samples (and Tweedie and Baayen used samples of the

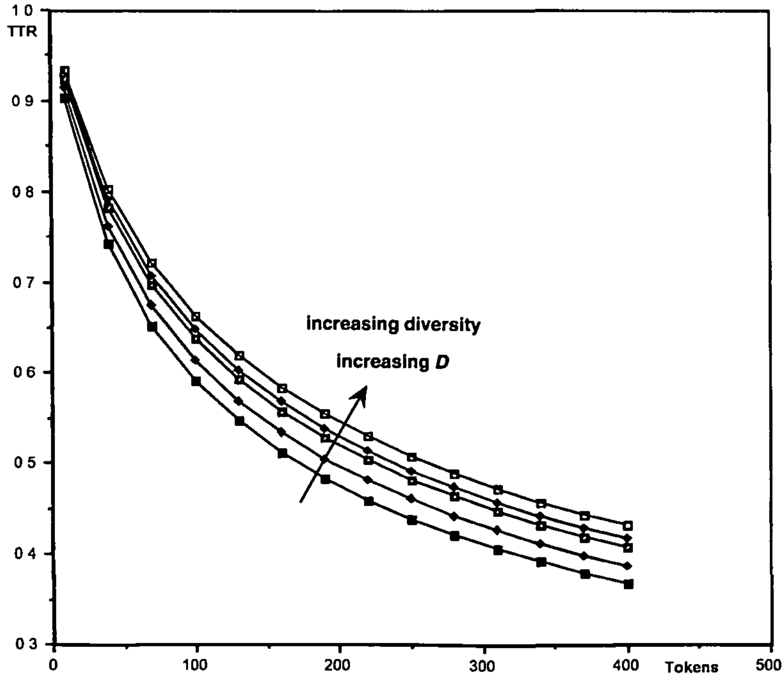

Fig. 1 Family of curves showing increasing diversity with increasing values of *D*.

order 26,500–120,000 tokens) will either reach this state or, for a large proportion of the sample, approximate to it as the late introduction of the occasional rare type will make only a small difference compared wıth the token count.

In many applications, however, much smaller language samples are typically used and in some cases, that of investıgations of normal or impaired child language for example, often only very small samples are available. The method developed for *vocd*, then, builds on previous theoretıcal analyses, notably by Brainerd (1982) and in particular Sichel (1986), which model the TTR versus token curve mathematically with reasonable accuracy for relatively small language samples of up to a few thousand tokens. Unfortunately, neither Brainerd's nor Sichel's model is mathematically simple and both need two parameters to describe an individual's lexical diversity.

Various probabilistic models were developed and ınvestıgated to arrive at a model whose characteristics yield a valid measure of vocabulary diversity by containing only *one* parameter, which increases with increasing diversity, and falls into a range suitable for discriminating among the range of transcripts found in various language studies. The model chosen is derived from a simplification of Sichel's (1986) type–token characteristic curve and is in the form of the following equation, where N is the number of tokens and D is a parameter:

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

This equation yields a family of curves (Fig. 1) all of the same general and appropriate shape, with different values for D distinguishing different members of this famıly (see also Malvern and Richards, 1997). In the model, D itself is used directly as an index of lexical diversity with predicted values of the order 10^1–10^2.

To calculate D from a transcript, the *vocd* program first plots the empirical TTR versus token curve for the speaker. It derives each point on the curve from an average of 100 trials on subsamples of words of the token size for that point. The subsamples are made up of words randomly chosen (without replacement) from throughout the transcript. The program then finds the best fit between the theoretical model and the empirical data by a curve-fittıng procedure that adjusts the value of D in the equation until a match is obtained between the actual curve for the transcript and the closest member of the family of curves represented by the mathematical model. This value of D for best fit ıs the index of lexical dıversity. High values of D reflect a high level of lexical diversity and lower diversıty produces lower values of D. It should be emphasized that, being based on a probabilistic model that predicts *expected* values, the calculating processes are stochastic, and averaging is intrinsic at all stages (see details of the software below).

The validity of D has been the subject of extensive investigation (Malvern and Richards, 1997, 2000; Richards and Malvern, 1997a, 1998) on samples of child language, children with Specific Language Impairment (SLI), teenagers learning French as a foreign language and the language of their teachers, children's narratives, adult learners of English as a second language, and academic writing. In these validation trials, the empirical TTR versus token curves for a total of 162 transcripts from five corpora covering ages from 24 months to adult, four languages and a variety of settings, all fitted the model. The model produced consistent values for *D*, which, unlike TTR and even Mean Segmental TTR (the average TTR for segments of a given token size), correlated well with other well-validated measures of language (see Richards and Malvern, 1997a, pp. 35–8). These five corpora also provide useful indications of the scale for *D*.

3 Calculation of *D*

The program *vocd* is written in C and exists in UNIX, PC, and Macintosh versions. It operates on ASCII files of transcripts set out and coded according to the Codes for the Human Analysis of Transcripts (CHAT) system developed by Brian MacWhinney as part of the Child Language Data Exchange System (CHILDES). The goal of the CHILDES project is to facilitate the study of child language development through a common transcription format, the sharing of transcript data and the provision of computerized tools for analysis (MacWhinney and Snow, 1990; MacWhinney, 1995). The text-handling features of *vocd* have been modelled on MacWhinney's Computerized Language Analysis (CLAN) programs, particularly a program called *freq* whose function is to make frequency counts of words or codes. *Freq* and *vocd* are compatible with regard to the items they include in word counts by default. As with *freq*, however, the use of various switches and exclude files, in combination with the careful use of CHAT transcription conventions, allows linguistic items to be selected or rejected in a way that corresponds to researchers' theoretical perspective on what should count as a word and what should count as a *different* word. A description of the text-handling options is given in Section 4.

In calculating *D*, *vocd* uses random sampling (without replacement) of tokens in plotting the curve of TTR against increasing token size for the transcript under investigation. Random sampling has two advantages over sequential sampling. First, it matches the assumptions underlying the probabilistic model. Secondly, unlike sequential sampling, it avoids the problem of the curve being distorted by the clustering of the same vocabulary items at particular points in the transcript.

In practice, each empirical point on the curve is calculated from averaging the TTRs of 100 trials on subsamples consisting of the number of tokens for that point, drawn at random from throughout the transcripts (see Fig. 2). This default number was found by experimentation and balanced the wish to have as many trials as possible with the desire for the program to run reasonably quickly. The run time has not been reduced at the expense of reliability, however, as it was found that taking

100 trials for each point on the curve produced consistency in the values output for *D* without unacceptable delays.

Which part of the curve is used to calculate *D* is crucial. First, the final value of *N* (the number of tokens in a subsample) cannot be larger than the transcript itself and, to have subsamples to average for the final point

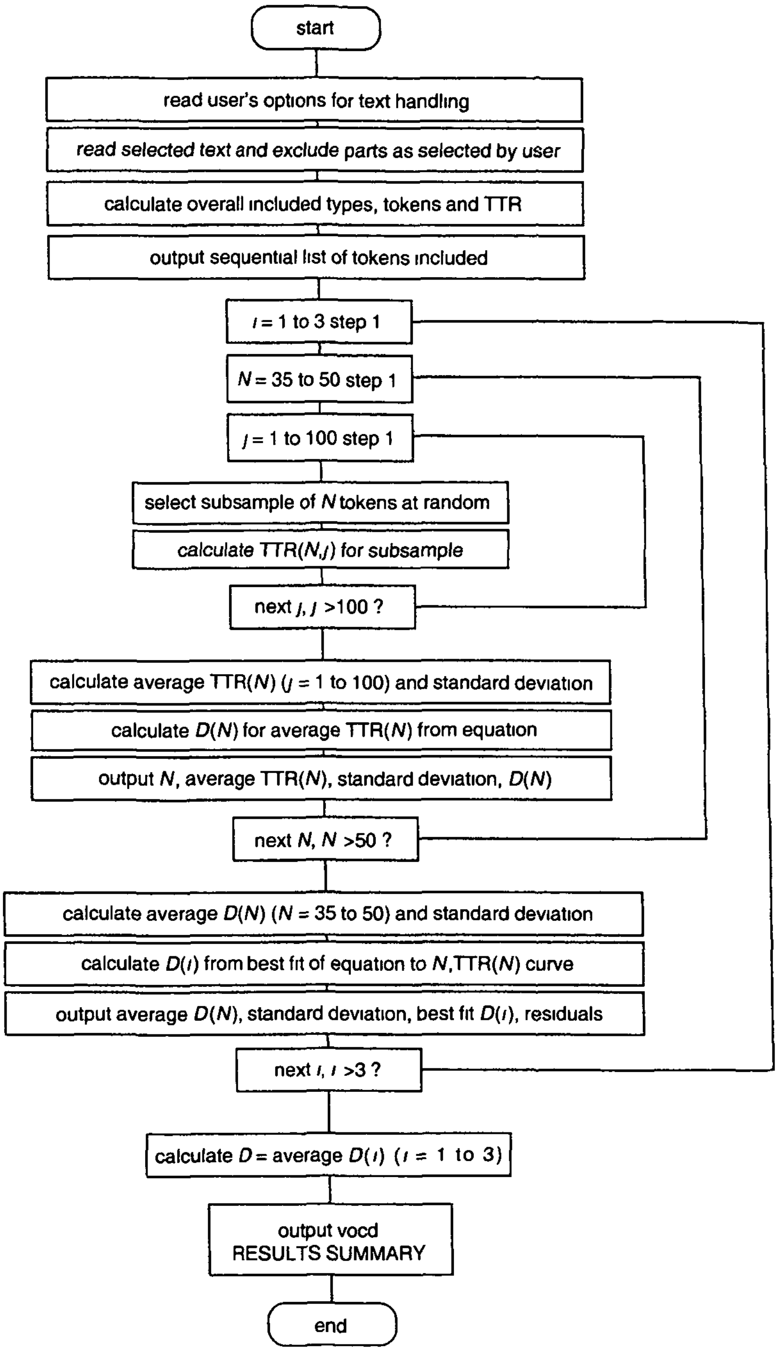
start
read user's options for text handling
read selected text and exclude parts as selected by user
calculate overall included types, tokens and TTR
output sequential list of tokens included
i = 1 to 3 step 1
N = 35 to 50 step 1
j = 1 to 100 step 1
select subsample of N tokens at random
calculate TTR(N,j) for subsample
next j, j >100 ?
calculate average TTR(N) (j = 1 to 100) and standard deviation
calculate D(N) for average TTR(N) from equation
output N, average TTR(N), standard deviation, D(N)
next N, N >50 ?
calculate average D(N) (N = 35 to 50) and standard deviation
calculate D(i) from best fit of equation to N,TTR(N) curve
output average D(N), standard deviation, best fit D(i), residuals
next i, i >3 ?
calculate D = average D(i) (i = 1 to 3)
output vocd
RESULTS SUMMARY
end

Fig. 2 Flow chart for producing values of *D* from transcripts.

on the curve, needs to be smaller. Moreover, although the transcripts and texts to be analysed will vary hugely in total token count, the range chosen for the users of *vocd* must be applicable to investigations such as child language studies, where typically transcripts may contain only 100 to a few hundred tokens in total. Secondly, the equation is derived by *approximation* from Sichel's (1986) model, and the approximations apply with greater accuracy at lower numbers of tokens (and note that Tweedie and Baayen (1998) demonstrated that all the models they investigated, including Sichel's, fail for large token sizes). We conducted extensive trials calculating *D* over different parts of the curve to find a portion for which the approximation held good and averaging worked well. As a result of these trials the default is for the curve to be drawn and fitted for $N = 35$ to $N = 50$ tokens in steps of one token. We should emphasize that, as noted above, each of these points is calculated from averaging 100 subsamples, each drawn from the *whole* of the transcript, so that, although only a relatively small part of the curve is fitted, it uses all the information available in the transcript. This also has the advantage of calculating *D* from a standard part of the curve for all transcripts regardless of their total size, further providing for reliable comparisons between subjects and between the work of different researchers.

The procedure then depends on finding the best fit between the empirical and theoretically derived curves by the least square difference method. Extensive testing confirmed that the best-fit procedure was valid and was reliably finding a *unique* minimum at the least square difference.

Finally, as the points on the curve are averages of random samples, a slightly different value of *D* is to be expected each time the program is run. Tests showed that with the chosen defaults these differences are relatively small, but consistency was improved by *vocd* calculating *D* three times by default and giving the average value as output (see Fig. 2).

As noted above, the software plots the TTR versus token curve from thirty-five tokens to fifty tokens, and each point on the curve is produced by random sampling *without* replacement. The software therefore requires a minimum of fifty tokens to operate. We should point out, however, that the matter of minimum sample size needs further investigation: the fact that the software will satisfactorily output a value of *D* from a sample as small as fifty tokens does not guarantee that values obtained from such small samples will be reliable (see Section 5).

4 Text-handling options

The software can be used to investigate lexical diversity in single or multiple files of speech or writing provided that they are transcribed accurately in legal CHAT format. This can be ascertained by running them through the CHILDES *check* program (MacWhinney, 1995), which will also help to track down typographical and spelling errors, and by using *freq* (see above) to create a complete word list that can be scanned for further errors.

Various switches and options can be entered into the UNIX command line to select text for analysis and to determine what is counted as a word and a different word. Examples of these functions are listed below.

- Specification of speaker.
- Morphemicization. By default, *vocd* treats inflected words (transcribed as 'go', 'go-es', 'go-ing') as different word types. This switch allows them to be treated as tokens of a single word type.
- Inclusion or exclusion. Single words, a list of words in a separate ASCII file, or whole utterances whose codings are selected by the user can be excluded or included in the analysis. Exclude files contain a list of items to be omitted (for example, hesitation markers, laughter, etc.). Include files consisting of only those words one wishes to be included. This would be useful, for example, if a lexical diversity value for just closed-class items were required.
- Split-half reliability. This switch allows separate analysis of odd and even words in the transcript. Results can then be fed into a split-half reliability analysis.
- Include capitalized words only in the analysis.
- Exclude retracing. This switch will remove self-repetitions and self-corrections.
- Upper case/lower case. By default, *vocd* would treat 'may' (modal verb) and 'May' (the month) as the same word type. This switch allows them to be counted as different types.
- Non-completed words. Words frequently occur in a shortened form such as 'till' for 'until' or 'cos' for 'because'. Sometimes a speaker will alternate between the full and shortened form, which might distort analyses if *vocd* treated them as different word types. There are three options for dealing with, for example, a word transcribed as '(be)cause': it can be processed as 'because' (the default), as '(be)cause', or as 'cause'.
- Text replacement. In the utterance 'it goed [: went] in there' a child's 'goed' is followed by the adult form in square brackets. By default preceding text is replaced by the text in square brackets, so the form processed would be 'went'. This allows alternative realizations of a word (e.g. yes, yeah, yep) to be treated as the same word type. This facility can also be overridden so that 'goed' would be returned rather than 'went'.

5 Command line and output from vocd

The minimum command line for *vocd* would be *vocd filename.cha*, where *vocd* executes the program and *filename.cha* is the transcript in CHAT format. However, when analysing speech data it would be usual to specify a speaker tier: *vocd +t"*CHI" filename.cha*, where *+t"*CHI"* selects the child tier. A typical command line used by researchers can be illustrated by the following example. In an analysis of oral interviews with students learning French as a foreign language (Richards and Malvern, 1998), we wished to compare the lexical diversity of teachers with that of their students. The analysis was carried out on the Richards and Chambers 'Reading' corpus in the CHILDES database using for the teachers the command *vocd +t"*TEA" –s@ttrexclu +r6 –s"[+ bch] " w01.cha*, where:

vocd executed the program;
*+t"*TEA"* limited analysis to the teacher's speaker tier;
–s@ttrexclu filtered out a list of items held in an exclude file called 'ttrexclu';
+r6 removed retraced material;
–s"[+ bch] " filtered out utterances coded as 'back channels';
w01.cha was the filename for the first interview.

Table 1 shows a sample of output from another example of a typical analysis—this time on one file from a 32-month-old child in the New England Corpus of the CHILDES database (MacWhinney, 1995). Here the command line *vocd +t"*CHI" +r6 –s@exclude +s"*-%%" d096132.cha* specifies analysis of the child speaker tier (*+t"*CHI"*), exclusion of self-repetition and self-correction (*+r6*), removal from analysis of words contained in an exclude file (*–s@exclude*), and treatment of inflected forms ('book'/'books') as the same word type (*+s" *-%% "*). The output consists of:

- A sequential list of utterances by the speaker selected containing *only those tokens that will be retained for analysis.*
- Tables showing the data that produce the empirical curve: number of tokens for each point on the curve from $N = 35$ to $N = 50$, average TTR and the standard deviation for each point and the value of D obtained from the equation for each point. Three such tables appear, one for each time the program takes random samples and carries out the curve fitting.
- At the foot of each table is the average of the Ds obtained from the equation and their standard deviation, the value for D that provides the best fit, and the residuals.
- Finally, a Results Summary repeats the command line and file name, and provides Type and Token information for the lexical items retained for analysis, and gives the three optimum values of D and their average.

6 Validation

To test for possible effects of sample size on the values obtained for D, and to gauge the reliability of the measure, thirty-eight child language transcripts from the New England Corpus (Dale *et al.*, 1989; Snow, 1989) of the CHILDES database were subjected to analysis. The children ranged in age from 27 to 33 months with a mean of 30.3 months, and they produced a mean of 316 word tokens with a standard deviation of 141. Three analyses were conducted using *vocd*: first, on all words in the transcripts; second, on even-numbered words only; third, on odd-numbered words. This allowed effects of sample size to be gauged by comparing mean values for the whole sample of words with those for half the sample.

Table 1 Example of output from *vocd*

Command line:

```
vocd +t"*CHI" +r6 -s@exclude +s"*-%%" d096132 cha
```

Output:

```
readıng exclude file <exclude>
UTTERANCES: (vocd<d096132.cha>)
where are the toy
yeah
yeah
book
book
book
book
book
two
where the two booko@
ı want two booko@
okay
baby
bug
duck
there
yeah
what
what
yeah
comb
comb
that
comb
brush
brush
soap
yeah
purple
that yellow
no purple
purple
milk there
no
yeah
me
yeah
yeah
there
purple
yeah
get more book
puppet show
puppet
puppet
mommy
no more cookıe
no
hello
hello
yeah
hello
```

Table 1 (*Continued*)

no
yeah
yeah
that box
yeah
mommy color
mommy color too
three color
here mommy
purple
that purple
yeah
here mommy
help me
help
help
help me
yeah
eye
the nose
all done
house
house
the house
me
down
go down
out
there
car
car
car
car
nope
want to put ıt away mommy
ı want ıt away
yeah
no

tokens	samples	ttr	st.dev	*D*
35	100	0 6951	0.053	27.739
36	100	0.6925	0.053	28 072
37	100	0.6905	0.052	28 507
38	100	0.6887	0.058	28 946
39	100	0.6790	0.056	28 003
40	100	0.6703	0.061	27.247
41	100	0.6566	0.051	25.735
42	100	0.6605	0.054	26.981
43	100	0.6567	0 050	27.016
44	100	0 6466	0.050	26.026
45	100	0 6531	0.050	27.667
46	100	0.6461	0.052	27.128
47	100	0.6336	0.048	25.751
48	100	0.6225	0.051	24 636
49	100	0.6214	0.044	24.992
50	100	0.6248	0.049	26 011

D: average = 26.903, std dev. = 1 221
*D*_optımum <26.85; min least sq val = 0.001>

Table 1 (*Continued*)

```
tokens     samples     ttr         st.dev      D
35         100         0.6937      0.058       27 496
36         100         0.6897      0 061       27.598
37         100         0.6876      0 062       27.993
38         100         0.6705      0 055       25.928
39         100         0.6833      0.060       28 754
40         100         0.6672      0.059       26 760
41         100         0.6671      0.061       27 400
42         100         0.6633      0.052       27.446
43         100         0.6612      0.049       27.737
44         100         0.6541      0.048       27.211
45         100         0.6504      0.058       27.232
46         100         0.6450      0 046       26.954
47         100         0.6381      0 047       26.437
48         100         0.6388      0 054       27.106
49         100         0.6198      0 042       24.754
50         100         0.6290      0 047       26.660
D average = 27.092; std dev. = 0.874
D_optimum <27.06; min least sq val = 0.000>
tokens     samples     ttr         st dev      D
35         100         0.6966      0 055       27.984
36         100         0 6786      0 059       25 792
37         100         0 6832      0.059       27 264
38         100         0 6897      0.059       29 133
39         100         0 6764      0.059       27.571
40         100         0.6700      0.051       27.206
41         100         0.6732      0 055       28.424
42         100         0.6607      0 059       27.020
43         100         0.6588      0.050       27 355
44         100         0.6593      0.044       28 071
45         100         0.6544      0 046       27 888
46         100         0.6504      0 048       27.836
47         100         0.6387      0 051       26.537
48         100         0.6323      0 047       26.094
49         100         0 6271      0.049       25.844
50         100         0.6262      0 045       26.226
D: average = 27.265; std dev  = 0.937
D_optimum <27 24, min least sq val = 0.001>
VOCD RESULTS SUMMARY
====================
Command line           vocd +t*CHI +r6 -s@exclude +s*-%% d096132.cha
File name:             d096132.cha
Types,Tokens,TTR.      <52,129, 0.403101>
D_optimum values:      <26.85, 27.06, 27.24>
D_optimum average:     27.05
```

Table 2 shows the mean values of *D* for each analysis, and for the average of the odd- and even-numbered words. It can be seen that there is little difference between the values for the whole transcripts and half the transcripts. A series of two-tailed paired *t*-tests confirmed that none of the three comparisons was statistically significant (even versus all: t = 0.310, d.f. = 37, p = 0.310; odd versus all: t = –0.464, d.f. = 37, p = 0.646; mean of odd and even versus all: t = 1.314, d.f. = 37, p = 0.197).

Table 2 Mean values of *D* for different portions of the transcript

	Mean *D*	SD	*N*
All words	46.26	17.77	38
Even-numbered words	47.54	20.51	38
Odd-numbered words	45.79	17.88	38
Mean of odd and even	46.67	18.03	38

From this we can conclude that there is no detectable effect of the number of tokens on the values obtained for *D*.

Reliability, in the form of internal consistency, was estimated using the split-half method. Values of *D* for even-numbered words across the thirty-eight children were correlated with those for odd-numbered words. The result was a Pearson's coefficient of 0.763 ($p = 0.000$). For a half transcript consisting of 158 words on average, this compares well with the reliability of previous measures—on the basis of an extensive examination of a range of lexical diversity measures, Hess *et al.* (1986) suggested that a minimum of 350 tokens were required to obtain a reliability coefficient of 0.70.

7 Conclusion

It has been shown that a mathematical model of the curvilinear relationship between the size of a language sample and the range of vocabulary it contains can be simplified and used as a way of quantifying vocabulary diversity. Furthermore, this simplified model has been implemented in software in a way that makes it accessible to other language researchers and allows them flexibility in defining what is meant by a Type and a Token.

The program, *vocd*, plots the empirical curve for TTR against tokens and fits the theoretical curve of the model to it by adjusting a single parameter, *D*. The value of *D* that gives the best fit of the theoretical curve to the empirical curve is taken as the measure of vocabulary diversity. This procedure allows all the valid word tokens in a transcript of speech or a sample of writing to be included in the analysis, and produces a measure that has acceptable levels of reliability when applied to typical child language transcripts and is not discernibly affected by the size of language sample.

Acknowledgements

The *vocd* program was written by Gerard McKee of the Department of Computer Science, The University of Reading. The research project 'A new research tool: mathematical modelling in the measurement of vocabulary diversity' was funded by grants awarded to Brian Richards and David Malvern from the Research Endowment Trust Fund of The University of Reading and by the Economic and Social Research Council, UK (No. R000221995). We would also like to acknowledge the help and advice of Raymond Knott, who acted as consultant to the project.

References

Brainerd, B. (1982). The type–token relation in the works of S. Kierkegaard. In Bailey, R. W. (ed.), *Computing in the Humanities.* Amsterdam: North-Holland, pp. 97–109.

Brunet, E. (1978). *Vocabulaire de Jean Giraudoux: Structure et Evolution.* Geneva: Slatkine.

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

Dale, P., Bates, E., Reznick, S., and Morrisset, C. (1989). The validity of a parent report instrument of child language at 20 months. *Journal of Child Language,* 16: 239–49.

Dugast, D. (1979). *Vocabulaire et Stylistique. I Théâtre et Dialogue. Travaux de Linguistique Quantitative.* Geneva: Slatkine–Champion.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika,* 40: 237–64.

Guiraud, H. (1954). *Les Caractères Statistiques du Vocabulaire.* Paris: Presses Universitaires de France.

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

Herdan, G. (1960). *Quantitative Linguistics.* London: Butterworth.

Hess, C. W., Sefton, K. M., and Landry, R. G. (1986). Sample size and type–token ratios for oral language of pre-school children. *Journal of Speech and Hearing Research,* 29: 129–34

Hockey, S. (1988). *Micro-OCP.* Oxford: Oxford University Press.

Honoré, A. (1979). Simple measures of richness of vocabulary. *Association for Literacy and Linguistics Computing Bulletin,* 7(2): 172–9.

Maas, H.-D. (1972). Zusammenhang zwischen Wortschatzumfang und Lange eines Textes. *Zeitschrift fur Literaturwissenschaft und Linguistik,* 8: 73–9.

MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk,* 2nd edn. Hillsdale, NJ: Erlbaum.

MacWhinney, B. and Snow, C. E. (1990). The Child Language Data Exchange System: an update. *Journal of Child Language,* 17: 457–72.

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

Malvern, D. D. and Richards, B. J. (2000). Validation of a new measure of lexical diversity. In Beers, M , v. d. Bogaerde, B., Bol, G., de Jong, J., and Rooijmans, C. (eds), *From Sound to Sentence: Studies on First Language Acquisition.* Groningen: University of Groningen, Centre for Language and Cognition.

Miller, J. F. and Chapman, R. (1993). *SALT: Systematic Analysis of Language Transcripts, Version 3.0.* Baltimore, MD: University Park Press.

Richards, B. J. (1987). Type/token ratios: what do they really tell us? *Journal of Child Language,* 14: 201–9.

Richards, B. J. and Malvern, D. D. (1997a). *Quantifying Lexical Diversity in the Study of Language Development.* New Bulmershe Papers. Reading: University of Reading.

Richards, B. J. and Malvern, D. D. (1997b). *Type–Token and Type–Type Measures of Vocabulary Diversity and Lexical Style. an Annotated Bibliography.* Reading:

Faculty of Education and Community Studies, University of Reading. (Also available at: http://www.rdg.ac.uk/~ehsrichb/)

Richards, B. J. and Malvern, D. D. (1998). *A New Research Tool· Mathematical Modelling in the Measurement of Vocabulary Diversity* (Award reference no. R000221995). Final Report to the Economic and Social Research Council, Swindon, UK.

Sichel, H. S. (1975). On a distributive law for word frequencies. *Journal of the American Statistical Association*, **70**: 542–7.

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

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

Tuldava, J. (1978). Quantitative relations between the size of the text and the size of vocabulary. *SMIL Quarterly, Journal of Linguistic Calculus*, **4**: 28–35.

Tweedie, F. J. and Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, **32**: 323–352.

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.