Investigating accommodation in language proficiency interviews using a new measure of lexical diversity

David Malvern and **Brian Richards** *The University of Reading*

Lexical diversity is an important indicator of language learners' active vocabulary and how it is deployed. Traditionally it has been measured by the Type–Token Ratio (TTR), the ratio of different words to total words used. Unfortunately, TTR is a function of sample size: larger samples of words will give a lower TTR and even commonly used measures derived from TTR which are claimed to be independent of sample size are problematic. To overcome this, the authors have developed an innovative measure of vocabulary diversity, D, based on mathematically modelling how new words are introduced into larger and larger language samples, and have produced software (*vocd*) to calculate it.

Previous research by the authors into language proficiency interviews (Richards and Malvern, 2000) investigated linguistic and discourse accommodation of teacher-testers using a wide range of student and teacher variables. In a study of teenage learners of French, the aspect of teachers' language in oral interviews that was most responsive to the ability of their students was lexical diversity. The analysis reported here focuses on this finding in greater depth using the new measure, D. The relationship between D and other measures of foreign language proficiency is investigated, the Ds of students and teachers are compared and the correlations between teachers' D and students' proficiency are computed.

Results firstly demonstrate the validity of D as a measure of vocabulary diversity and the effectiveness of *vocd* as a tool to analyse language data. Secondly, with regard to accommodation processes in oral testing, the two teachers did not finely tune their vocabulary diversity to the proficiency of individual students. Instead, each teacher roughly adjusted his or her language to the ability of the class they examined.

I Introduction

There has been much discussion of the validity of the oral interview as a means of assessing second language proficiency. A major factor in these discussions is the extent to which, in theory and practice, the interview resembles natural conversation (for an introduction and overview of the issues, see He and Young, 1998). Van Lier (1989) claimed that the language proficiency interview lacked features of real

Address for correspondence: B.J. Richards, The University of Reading, School of Education, Bulmershe Court, Reading, RG6 1HY, UK; email: b.j.richards@reading.ac.uk

Language Testing 2002 19 (1) 85–104 10.1191/0265532202lt221oa

conversation because of the power differential between participants and the precedence accorded to eliciting language. Subsequent analyses of interviews suggested that, while interviews do indeed show structural similarities to conversations (Lazaraton, 1992) or are 'authentic instances of talk-in-interaction' (Moder and Halleck, 1998: 144), there is an asymmetry of control over, and contribution to, the interaction (Young and Milanovic, 1992). For example, it is the testees who show greater conversational contingency or 'reactiveness' to their interlocutor, and the testers who show more goal orientation (Young and Milanovic, 1992). The interviewers have greater influence over the choice of topic (Johnson and Tyler, 1998; Moder and Halleck, 1998) and management of turn-taking differs from natural conversations (Lazaraton, 1992; Johnson and Tyler, 1998; Moder and Halleck, 1998).

By contrast, some researchers have turned their attention to accommodation in language proficiency interviews as a feature of authentic conversation. Accommodation theory attempts to account for processes by which the speech of participants in linguistic interaction converges or diverges in a systematic way, i.e. how the speech of one person becomes more similar to, or different from, that of a conversational partner. Convergent accommodation can be the result of a desire for social approval or the need to improve efficiency of communication (see Thakerar *et al*., 1982). In the latter category convergent accommodation encompasses the simplification and discourse adjustments made to young children acquiring their first language in the so-called 'motherese' or 'child-directed speech' register (for an overview, see Pine, 1994). This would include the 'fine-tuning hypothesis' (e.g., Cross, 1977) whereby maternal language is claimed to be optimally matched to the stage of children's linguistic and communicative development. In second language research, accommodative processes in conversations between native speakers and non-native speakers and in language classrooms have been identified as 'foreigner talk' or 'language teacher talk' (see Wesche, 1994; Gass and Varonis, 1985).

Foreigner talk modifications have also been found to be a characteristic of language proficiency interviews (Ross, 1992; Ross and Berwick, 1992; Lazaraton, 1996). Lazaraton (1996) argues that the kinds of linguistic and interactional support she identified are conversational features, but that in the context of the test, their impact on candidates' ratings are unclear. Little is known about the factors in the candidate that trigger such adjustments in the tester. Ross (1992) and Ross and Berwick (1992) were able to demonstrate that the frequency and extent of accommodation was related – i.e., (finely) tuned – to the proficiency level of the interviewees. These authors suggested

that the degree of interviewer accommodation could be used as an additional dimension in the assessment of candidates. Furthermore, they argued that major threats to the validity of the interview test would be, first, a lack of appropriate accommodation to the proficiency of students on the part of the teacher-examiner and, second, over-accommodation that fails to allow candidates to demonstrate the full extent of their proficiency.

In an earlier study into discourse and linguistic accommodation in oral interviews with 34 teenage learners of French, Richards and Malvern (2000) investigated whether teachers who were not native speakers accommodated their language to the proficiency of students. They found that, while on some teacher variables such as various categories of teacher repetition of the student, accommodation to individual students does occur, other aspects of their language are more grossly tuned to the general level of ability of their language class. One particularly large effect involved the lexical diversity of teachers. Of all the measures of teachers' language it was their vocabulary diversity that was most strongly predicted by the average ability of the class they taught. Sixty per cent of the variance in teachers' vocabulary diversity was explained by which of two language classes their students belonged to.

Measures of vocabulary diversity are used in a wide range of educational and linguistic research (Richards and Malvern, 1997; for a recent example, see Vermeer, 2000). They reflect the variety of active vocabulary deployed by a speaker or writer and – together with lexical density (the ratio of content words to function words), precision of expression (use of rare words) and lack of errors of lexical choice – they can be regarded as a component of lexical richness in second language assessment (Read, 2000: 200–05). Unfortunately lexical diversity is notoriously difficult to quantify reliably (Malvern and Richards, 1997; Richards and Malvern, 1997; Vermeer, 2000). Measurements are based on a comparison between the number of different words (types) and the total number of words (tokens). The best known of these, the Type–Token Ratio (TTR), is problematic because it is a function of sample size; large numbers of tokens in a sample produce lower TTRs than small samples (Chotlos, 1944; Hess *et al*., 1986; Richards, 1987). It is invalid therefore to compare overall TTRs calculated from speakers or writers who have produced different sizes of language sample. Some measures that are mathematical transformations of TTR – e.g., Carroll's (1964) Corrected TTR, Guiraud's (1960) Root TTR or Herdan's (1960) Bilogarithmic TTR – are claimed to be independent of sample size. Nevertheless, they have all been shown to be a function of the number of tokens (Ménard,

1983; Arnaud, 1984; Hess *et al*., 1986; 1989; Malvern and Richards, 1997; Tweedie and Baayen, 1998).

In their study of oral interviews, the authors (Richards and Malvern, 2000) addressed the problem of calculating lexical diversity from varying sample sizes by using the Mean Segmental Type–Token Ratio (MSTTR), an index that appears to have been originally recommended by Johnson (1944). Since then it has been used in many different kinds of linguistic investigation, including normal spoken language (Fairbanks, 1944), students' L1 writing (Mann, 1944), schizophrenia (Manschreck *et al*., 1981), aphasiology (Wachal and Spreen, 1973), historical documents (Carpenter and Hersh, 1985), and foreign language learning (Meara, 1978). Richards and Malvern (1997: 35) describe MSTTR as 'the average TTR for successive segments of text containing a standard number of word tokens'. For the teachers in the oral interview study, their transcripts were divided into segments of 100 words (MSTTR-100). Many of the students, however, contributed fewer than 100 words to a five-minute conversation, and their standard segment had to be as low as 30 words (MSTTR-30).

There are two clear advantages to MSTTR: it removes the problem of variation in sample size and it also wastes less data than if all analyses were performed on a standard number of words (i.e., reducing all transcripts to the length of the shortest). Nevertheless, at least five problems remain:

1) MSTTRs calculated from different sizes of standard segment are not directly comparable, because larger segments will tend to give lower TTRs.
2) Very short segments (even those of 100 tokens) are likely to distort results because they are not sensitive to repetition of words beyond the boundary of their own segment.
3) Transcripts do not usually divide exactly into standard-sized segments. This results in at least some loss of data and a consequent reduction in reliability.
4) As will be explained more fully below, the relationship between number of types and number of tokens for any individual sample of speech or writing is a dynamic one. That is to say, an MSTTR value represents only a single point on a curve representing the way in which TTR falls with increasing token size for that sample.
5) It is worth noting that in the Richards and Malvern (2000) study the variation in MSTTR for both teachers and students was very small compared with other measures, raising the possibility that this might have attenuated correlations.

Since carrying out the analyses described in Richards and Malvern (2000), the authors have developed a new measure of lexical diversity that overcomes all the disadvantages of MSTTR. The solution is based on the observation that the way in which TTR falls with sample size is systematic and that this means that the probability of new vocabulary being introduced into longer and longer samples of speech or writing can be mathematically modelled. The model is a mathematical equation that relates TTR to token size (N) in terms of a third variable, a parameter referred to as 'D':

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

The equation, which is a simplification of Sichel's (1986) type–token characteristic curve, applies to a family of curves that plot TTR against the number of tokens. All of these fall as the token size increases, but the curves of speakers or writers with high lexical diversity will lie above those with low lexical diversity (see Figure 1). It is the parameter D in the equation that reflects the relative height of

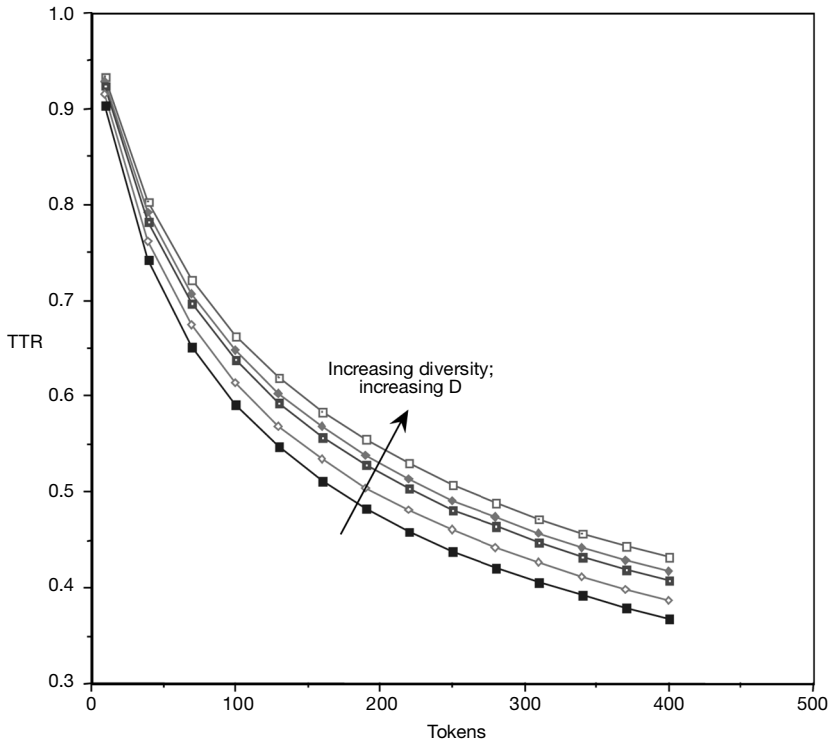

Figure 1 Family of curves showing increasing diversity with increasing values of D

these TTR by token curves and hence the degree of lexical diversity contained in a transcript. The full mathematical derivation of the equation is beyond the scope of this article but it is explained in detail in Richards and Malvern (1997) and a full rationale is contained in McKee *et al*. (2000).

The method for obtaining D values from transcripts depends on producing a graph of the way the TTR in a given transcript falls with increasing token size within the language sample, and comparing this empirical graph with the theoretical curves obtained from the mathematical model, i.e., from the equation. The best fit between the two, obtained by adjusting the value of D until the theoretical curve matches the empirical curve as closely as possible, yields a measure of the person's vocabulary diversity represented by the value of D for optimum fit. A higher D represents greater diversity and values have been found to range from D = 5 for a five-year-old language impaired child to D = 120 for a sample of academic writing (Richards and Malvern, 1999).

The procedure outlined above has been implemented in the form of a computer program (*vocd*), which operates on language samples transcribed in the standard CHAT format (Codes for the Human Analysis of Transcripts) of the CHILDES project (Child Language Data Exchange System) (MacWhinney and Snow, 1990). *Vocd*, which was written by Gerard McKee, is freely available to other researchers as part of the CLAN (Computerized Language Analysis) programs (MacWhinney, 2000a) from the CHILDES web site (http://childes.psy.cmu.edu). The software automates the process of calculating and averaging TTRs for 100 random trials of subsamples from the transcript of a given token size increasing in increments of one token from 35 tokens to 50 tokens. This produces an empirical TTR by token curve consisting of 16 points for the segment of the overall curve that ranges from 35–50 tokens. To do so, *vocd* uses random sampling (without replacement) of word tokens. A curve-fitting procedure using the equation provides the index of lexical diversity (D) as described above. The software contains various text-handling options that, for example:

- allow the speaker to be specified;
- exclude parts of the text such as self-repetition;
- exclude non-words (hesitation markers, laughter, etc.);
- enable analysis of root forms only.

It also contains a split-half reliability function that allows comparison between values of D obtained for even-numbered words vs. odd-numbered words. Recently the software has been adapted further to

include a wider range of options including diversity measures for separate word classes, and currently a method is being developed to allow the calculation of type/type measures of lexical style, such as noun/verb ratios.

The measure D overcomes the disadvantages of other measures, including MSTTR, first, because it is independent of sample size, thus allowing valid comparisons between speakers or writers who produce varying quantities of linguistic data. Second, because *vocd* takes numerous random samples from the whole of a transcript, it takes account of both long-distance and short-distance repetition, and no data remain unused. Finally, it is more informative because it is representative of the whole of the TTR vs. token curve rather than just a single point on it. Extensive research has been carried out on the reliability and validity of D. It compares well with other measures on split-half reliability (McKee *et al*., 2000) and in first language development it correlates strongly and significantly with age and with other well-validated measures of linguistic progress such as Mean Length of Utterance (MLU) (Malvern and Richards, 2000).

Thanks to the development of the new measure, D, and the software that automates its calculation directly from transcripts, it was now possible to apply these new procedures to the original transcripts of the 34 learners of French as a foreign language. These analyses have two purposes. First, they allow the properties of D to be explored and further validated on a new type of data. Secondly, it provides a more powerful tool than MSTTR to investigate whether or not non-native speaking teacher-examiners accommodate their lexical diversity to individual students, or whether in the Richards and Malvern (2000) study MSTTR simply lacked the sensitivity to detect this.

II Method

The data consist of the audio-tapes of 34 British secondary school students taking their oral examination in French for the General Certificate of Secondary Education (GCSE). The GCSE is a national examination taken by school students in Britain at the age of 16. Oral interviews (described in the documentation of the examining group as 'free conversation'), which averaged just over five minutes in length, had been conducted by two teachers of French both of whom tested their own students. Unlike many other summative oral language tests conducted at key points in students' education, the GCSE regulations require that candidates are both tested and their performance simultaneously scored, not by a stranger, but by their own class teacher. While this, inevitably, reduces the generalizability of the

research reported here, the fact that this situation exists in a national examination system makes it an important context for investigation.

The teachers had learnt French as a foreign language but were experienced and well qualified, and both had a successful record of preparing students for the GCSE examination and of conducting and assessing the oral interview. As noted above, each teacher examined the students in his or her own French class. The interviews were transcribed in CHAT format (MacWhinney, 2000a) by a native French speaker (Francine Chambers) who was also an experienced teacher of French to English-speaking children of this age, and were re-transcribed by the second author of this article. Discrepancies were resolved with the assistance of a near-native speaker of French who was also an experienced teacher of French as a foreign language in Britain. Where discrepancies could not be resolved, the utterance or word was coded as 'unintelligible' and excluded from the data. In addition to allowing analysis by *vocd*, CHAT format also enabled further computer-assisted analyses to be carried out by the other CLAN programs of the CHILDES project (MacWhinney, 2000a). The 34 transcripts are available to other researchers by selecting 'reading.zip' in the CHILDES database (http://childes.psy.cmu.edu/win/biling) (for further details, see MacWhinney, 2000b).

1 The students

The students had been learning French for five years, receiving four 35-minute lessons a week. Their proficiency varied from those who made very little contribution to the conversation to one student who performed at a level comparable with native-speakers. An indication of the huge variation between the students is given in Table 1 by the standard deviation (166.5 for a mean of 183.6 words) for the number of words they produced during the interview. These ranged from 35 to 808 words. The student who performed at near-native-speaker level scored the maximum possible number of points for listening, speaking, reading and writing in the GCSE examination. This student obtained extreme values on various other measures and, for this reason, and because of the restricted range of some of the student measures reported below, non-parametric statistics are used for all statistical tests which include him. Twelve students were interviewed by Teacher A and 22 by Teacher B. The school operates a policy of grouping students according to ability from the first year of foreign language study with the regular possibility of promotion or demotion into higher or lower ability groups. Students had been assigned to Teacher A's or Teacher B's class on this basis.

Table 1 Student variables ($n = 34$)

Student measure	Mean	sd
Measures taken from transcripts		
Number of words	183.6	166.5
Number of different words	85.1	56.2
MSTTR-30	22.8	1.7
Mean length of utterance (MLU words)	4.6	2.1
Utterances per turn (MLT)	1.1	0.1
Percentage unintelligible words	0.02	0.02
Words per minute	29.8	16.8
Type–token ratio (TTR)	0.5	0.1
GCSE examination results		
Score for oral examination (out of 7)	4.1	1.3
GCSE points (out of 28)	18.1	4.6
Mean ratings from 24 teachers of French		
Range of vocabulary (0–7)	2.8	1.7
Fluency (0–7)	2.7	1.6
Complexity of structure (0–7)	2.2	1.5
Content (0–3)	1.5	0.8
Accuracy (0–3)	1.2	0.6
Pronunciation (0–3)	1.3	0.6

2 *Student variables*

Three categories of student variable and their means and standard deviations are listed in Table 1. First, there are the objective measures extracted from transcripts using the CLAN software. These include MSTTR-30 as a measure of lexical diversity (see above). Secondly, there are the final results of the GCSE examination itself. The examining group had converted the students' scores on each of the four skills to a mark out of seven. Here we report the score out of 7 for the oral examination and the total examination score out of 28.

Thirdly, six further measures were obtained from the mean ratings of the tape recordings by 24 experienced teachers of French. Range of vocabulary, Fluency and Complexity of structure were rated on eight-point scales (0–7) and Content, Accuracy and Pronunciation on four-point scales (0–3). These measures were chosen because they were all included in the oral examination criteria for the GCSE examination groups (Chambers and Richards, 1992). Full details of the scales and the procedures used can be found in Richards and Chambers (1996).

In addition, using *vocd*, values of D were obtained for both the teachers (one D value for each student they tested, totalling 34 Ds) and the students. Seven students produced fewer than 50 word tokens in their oral interview and for these no D values could be calculated.

Sample size for analyses involving the students' D is therefore limited to 27. For all other analyses the sample is 34. For the 27 students the mean value for D is 56.9 (sd = 16.3); for the 34 Ds calculated from the two teachers the mean value for D is 44.9 (sd = 9.6). These will be compared and discussed in Section III, Subsection 2 below, but first we address the convergent and discriminant validity of D by examining the correlation between student Ds and other measures of their language.

III Results

1 The students

Rank order correlations between students' D and other student variables are presented in Table 2. It had been predicted that, as a valid measure of vocabulary diversity which is independent of the quantity of language produced, D would correlate most strongly with the other vocabulary measures except for the overall TTR which is invalid when sizes of language sample vary. In Table 2 it can be seen that this is indeed the case: the correlation with MSTTR-30 stands apart even from other significant correlations as the most powerful in the

Table 2 Spearman rank order correlations between student D and other student measures for all students for whom D was calculable ($n = 27$)

Student measure	*rho*
Measures taken from transcripts	
Number of words	.18
Number of different words	.35*
MSTTR-30	.59**
Mean length of utterance (MLU words)	.23
Utterances per turn (MLT)	.09
Percentage unintelligible words	.02
Words per minute	.23
Type–token ratio (TTR)	.20
GCSE examination results	
Score for oral examination (out of 7)	.34*
GCSE points (out of 28)	.31
Mean ratings from 24 teachers of French	
Range of vocabulary	.31
Fluency	.33*
Complexity of structure	.31
Content	.30
Accuracy	.31
Pronunciation	.23

Notes: *$p < .05$; **$p < .01$ (one-tailed tests)

set, the other significant relationships being with number of different words, the oral score and fluency. Importantly, D correlates with the number of different words rather than with the total number of words. These results provide further evidence of D's validity; therefore, D is sensitive to vocabulary and, to a lesser extent, to broader aspects of language proficiency. As expected, there is no correlation with overall TTR.

A more surprising result is the lack of any significant correlation between D and teachers' ratings of Range of vocabulary. Even though the value for *rho* (.31) is positive and approaches significance, it is very weak compared with the correlation between D and MSTTR-30 (.59). This raises the question of the validity of the subjective ratings. To investigate this further, the full matrix of Spearman intercorrelations between the ratings of the 24 teachers of French was computed. It can be seen from Table 3 that Range of vocabulary correlates extremely highly with the other scales; all the intercorrelations in the matrix are above .900. The highest figure is between Range of vocabulary and Content at .996. Unlike the more objective measures, therefore, the teachers' ratings do not discriminate between vocabulary deployment and other areas of proficiency. The rating of Range of vocabulary is likely to be heavily contaminated by halo effects. This result may also reflect the sheer difficulty of the task of rating range of vocabulary while listening to a tape recording. Whereas values of D are adjusted for length of conversation, it is unlikely that teacher raters would even attempt to do this. Instead they are likely to respond to other aspects of lexical richness, such as the use of low-frequency words, or, at least, words which are less common in the foreign language classroom. To throw further light on this finding one additional correlation was computed: that between Range of vocabulary and student MSTTR-30. Interestingly, this is close to zero (*rho* = –.08). Such results raise the wider issue of the extent to which

Table 3 Spearman rank order intercorrelations between the ratings provided by 24 teachers of French ($n = 34$)

Scale	1	2	3	4	5	6
1. Range of vocabulary	–					
2. Fluency	.987*	–				
3. Complexity of structure	.988*	.977*	–			
4. Content	.996*	.985*	.982*	–		
5. Accuracy	.974*	.970*	.979*	.970*	–	
6. Pronunciation	.922*	.918*	.920*	.911*	.946*	–

Note: *$p < .01$ (one-tailed tests)

raters are able to assess particular aspects of performance independently using analytical as opposed to holistic rating scales.

2 The relationship between teacher D and student measures

The aim of the next analysis was, first, to compare D values of students and teachers and, secondly, to assess whether the teachers' deployment of vocabulary was finely tuned to the language proficiency of the students. A positive correlation between teachers' D and student variables would be indicative of accommodation strategies.

The comparison between the average D for teachers and students is revealing and suggests over-accommodation. The means and standard deviations of the 34 Ds of the teachers and for the 27 student Ds reported at the end of Section II showed a lower lexical diversity and less variance for the teachers than for the students. Even with the extreme case excluded, and confining the analysis to the remaining 26 teacher and student scores for whom D could be calculated, the average for the students (mean = 55.1; sd = 13.8; $n = 26$) is still higher than for the teachers (mean = 46.7; sd = 7.5; $n = 26$). This difference is statistically significant on a paired samples t-test: $t = -2.92$; df = 25; $p < .01$. D values are also higher for the students than the teacher in each class when analysed separately, although statistical significance can be shown only for the 22 students of Teacher B (because all seven students for whom D could not be calculated were in Teacher A's group, the degrees of freedom fall to 4, making the detection of significant differences unlikely). It should be noted that the reason that teacher D is lower than student D on average does not lie in teachers giving students the floor in order to get them to talk. Interestingly, the mean and median number of words is substantially higher for the teachers (mean = 295.11; median = 285) than for the students (mean = 183.56; median = 153.5), even when the extreme case in included. Besides, the way D is computed would adjust for quantity of speech.

It is also interesting that the range for D (29.6–77.8) and the standard deviation (13.8) are also higher for the students than for the teachers (29.9–63.9 and 7.5). So the teachers are operating both at a lower level and within a narrower band. A Spearman rank order correlation between the two sets of Ds is not significant (*rho* = .24; $n = 27$; ns). Correlations computed separately between each teacher and his or her students are also non-significant. There is, therefore, no evidence of accommodation in teachers' lexical diversity in response to variation in individual student D.

By contrast, teachers' Ds do enter into significant, positive correlations with 12 out of 14 measures of the 34 students' language. These are shown in Table 4. At first sight it would appear, therefore, that the teachers, in spite of a general tendency to over-accommodate, are using greater lexical diversity with students whose language is more proficient, thus engaging in a form of tuning to the individual.

Nevertheless, the above interpretation would be correct only if each teacher could be shown to be behaving in the way suggested by their pooled data. Separate correlations for the Ds of Teacher A and Teacher B with the language measures of their students show that this is, however, far from being the case. These two sets of correlations are shown in Table 5. Teacher A shows no evidence of accommodation in the predicted direction. In fact, there is one negative correlation in particular, between Teacher D and GCSE points, that might suggest that the teacher uses greater diversity with weaker students. Although unexpected, this finding might be accounted for by processes of discourse accommodation such as the need for the teacher to reformulate questions, provide synonyms or paraphrases for weaker candidates, or to change topic more frequently if students provide little or no response. A similar lack of relationships is found for Teacher B for whom the correlations are predominantly negative, although none is statistically significant.

Table 4 Spearman rank order correlations between Teacher D and measures of students' language ($n = 34$)

Student measure	*rho*
Measures taken from transcripts	
Number of words	.53*
MSTTR-30	.11
Mean length of utterance (MLU words)	.53*
Utterances per turn (MLT)	.53*
Percentage unintelligible words	.01
Words per minute	.54*
GCSE examination results	
Score for oral examination (out of 7)	.50*
GCSE points (out of 28)	.59*
Mean ratings from 24 teachers of French	
Range of vocabulary	.50*
Fluency	.46*
Complexity of structure	.46*
Content	.49*
Accuracy	.47*
Pronunciation	.42*

Note: *$p < .01$ (one-tailed tests)

Table 5 Spearman rank order correlations between measures of students' language and Teacher A ($n = 12$) and Teacher B ($n = 22$)

Student measure	*rho* with Teacher A ($n = 12$)	*rho* with Teacher B ($n = 22$)
Measures taken from transcripts		
Number of words	−.13	−.08
MSTTR-30	.47	.26
Mean length of utterance (MLU words)	−.18	−.04
Utterances per turn (MLT)	.48	−.23
Percentage unintelligible words	.48	.01
Words per minute	.12	−.21
GCSE examination results		
Score for oral examination (out of 7)	−.22	−.10
GCSE points (out of 28)	−.57	−.03
Mean ratings from 24 teachers of French		
Range of vocabulary	.02	−.22
Fluency	−.04	−.28
Complexity of structure	−.01	−.30
Content	−.01	−.26
Accuracy	−.06	−.26
Pronunciation	−.05	−.24

Note: No relationships are statistically significant at .05 (one-tailed tests)

The striking result that there is a positive correlation for pooled data and yet no relationship or a tendency towards a negative relationship when the analysis is performed for each teacher separately is explained by the difference in ability between the two classes. This can be demonstrated by concentrating on the strongest correlation in the pooled data, that between the Teachers' D and students' GCSE points. Figure 2 shows a scatterplot of this correlation in which the two teachers are indicated separately. From this it can be seen that Teacher A's students score lower in the GCSE French examination than those of Teacher B. The median number of GCSE points for Teacher A's students is 13 compared with 20.5 for those of Teacher B. The difference is statistically significant on a Mann–Whitney U test ($U = 4.00$; $n = 34$; $p < .001$). Similarly, the median D for Teacher A is 35.1 compared with 50.2 for Teacher B. This difference is also significant ($U = 10$; $n = 34$; $p < .001$). The separation between the two groups can be seen even more starkly in Figure 3, which shows the median D plus and minus two semi-interquartile ranges (SIQR) plotted against the students placed in ascending order of their GCSE points. From the students' order, it can be seen that all but one student in Teacher B's class were of higher ability than those of Teacher A. Immediately, it can be seen that the upper bound (median plus

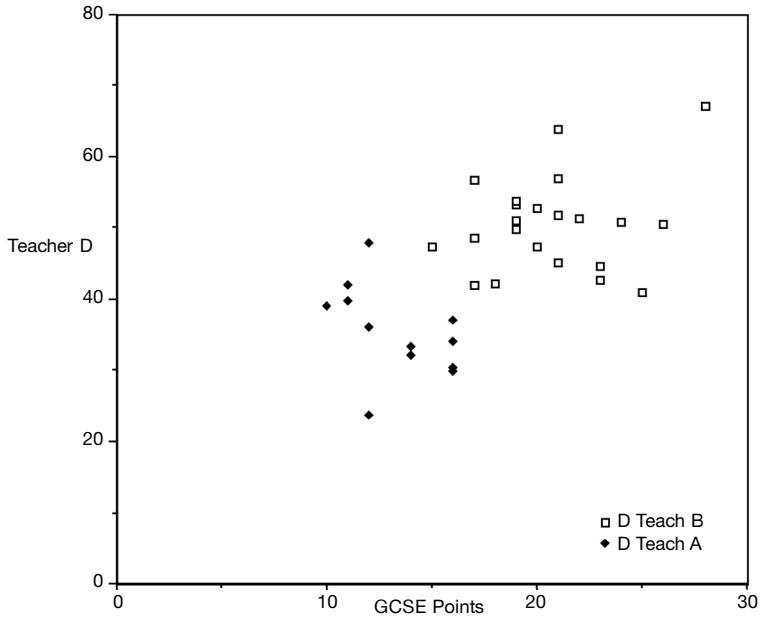

Figure 2 Scatterplot of D values for each teacher against students' GCSE points

$2 \times$ SIQR) for Teacher A virtually coincides with the lower bound (median minus $2 \times$ SIQR) for Teacher B. Within each group, the values of D form no particular pattern with respect to students' order, but the two bands formed by the median plus or minus two semi-interquartile ranges of D for each teacher hardly overlap. Only four of Teacher B's Ds fall within the range for Teacher A, and only one of Teacher A's Ds fall within the range of Teacher B. Although in terms of lexical diversity the teachers are not accommodating to individuals, each teacher is, therefore, pitching the general level and range of the diversity of their vocabulary to the collective proficiency of his or her own teaching group.

IV Conclusions

With regard to the first aim of the investigation, the findings reported above provide further evidence of the validity of mathematically modelling the relationship between TTR and token size to assess vocabulary diversity. As predicted for the student measures, D correlated with another measure of vocabulary diversity, MSTTR-30, rather than with measures of general language proficiency. As expected, there was no correlation between D and overall TTR, and D was significantly correlated with the number of different words as opposed to

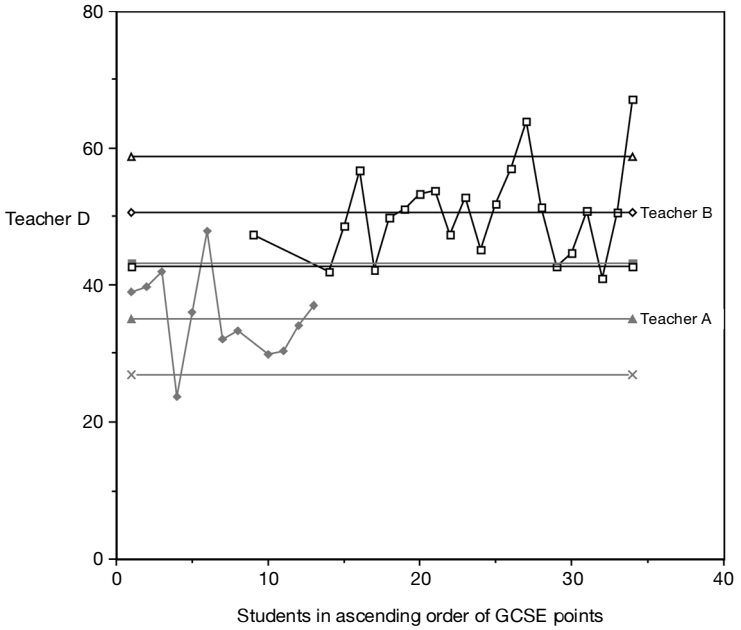

Figure 3 D for each teacher against students in ascending order of GCSE points, showing median D and the interval plus and minus two semi-interquartile range for each teacher

the total number of words. Contrary to predictions, however, D was not related to the ratings of 24 experienced teachers of 'range of vocabulary', but it seems likely that teachers are simply unable to assess lexical diversity independently of other factors, particularly when attempting to do so impressionistically from audio-taped recordings. This interpretation is supported by the extremely strong intercorrelations (all greater than .9) between the factors they rated, and by the failure of Range of vocabulary to correlate with MSTTR-30.

The second aim of this research was to investigate whether variation in teachers' vocabulary diversity was itself a form of accommodation to the linguistic proficiency of their students. At first sight this seemed to be the case: there was a significant correlation between the Ds of the teachers and a wide range of language measures for the students. Closer investigation, however, showed that this overall effect was not replicated for either teacher when their data were analysed separately. It had been brought about by a significant difference in the ability of each class which corresponded with a significant difference in the average D for each teacher. There was little overlap between the Ds of the teacher of each class. What appeared to be happening was that, while the language of each teacher was not finely

tuned to the ability of the individual students, they were pitching their language at a level appropriate to the ability of the class as a whole. Whether this general adjustment is in direct response to the input and interaction during the interviews themselves, or to previous perceptions and expectations based on knowledge of the groups derived from teaching them in class, can only be addressed through a parallel study using interviewers who had no previous acquaintance with the candidates. This is a question for future research.

It will be recalled, however, that there was far more variation in the D values for the students than for the teachers. There appears to be a tendency, therefore, in the context of a public examination conducted by non-native speakers, for each teacher to provide an approximately standard level of language across all the students he or she is testing. This may well reflect the need for public examination to be reliable and fair. It does, however, introduce the very threats to validity identified by Ross and Berwick (1992), the first of which is the absence of appropriate accommodation. The evidence here is that, although accommodation to individual students may well be absent at a finely-tuned level, there is a general adjustment to match the student's level of language, but this adjustment is kept within relatively narrow limits (see Figure 3). Clearly such general adjustment is appropriate, for without it students with low to average ability would find it more difficult to display even the proficiency they have. For these students, therefore, the validity of the test may survive the demand of reliability that the teacher-examiner behaves in a broadly similar way for each candidate. It is more questionable whether or not validity survives the second threat referred to by Ross and Berwick (1992), namely that of inappropriate accommodation which fails to stretch students. Given that lexical diversity is higher on average for candidates than for interviewers, and noting from Figure 3 that for six of the top seven candidates the teacher D is at, or well below, the median for the more able group, the evidence on the relative degree of accommodation is that beyond the general adjustment to the ability of the class as a whole, there is no systematic increase in teacher Ds as the candidates' ability rises.

Acknowledgements

We are most grateful to Ngoni Chipere, Pilar Durán, Suzanne Graham, Mair Richards and three anonymous reviewers for *Language Testing* for their comments on previous drafts of this article. We should also like to thank Gerard McKee for writing the *vocd* program, Francine Chambers for her help with data collection and transcription

and Mair Richards for checking the transcripts. The research was supported by a grant from The University of Reading Research Endowment Trust Fund (Oral assessment in modern languages) and two grants from the Economic and Social Research Council (A new research tool: mathematical modelling in the measurement of vocabulary diversity: R000221995; Mathematically modelling vocabulary diversity and lexical style: R000238260).

V References

Arnaud, P.J.L. 1984: The lexical richness of L2 written productions and the validity of vocabulary tests. In Culhane, T., Klein Bradley, C. and Stevenson, D.K., editors, *Practice and problems in language testing: papers from the International Symposium on Language Testing*. Colchester: University of Essex, 14–28.

Carpenter, R.H. and **Hersh, R.E.** 1985: A stylistic index of deteriorating military morale: using form in correspondence for intelligence purposes. *Language and Style* 18, 185–91.

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

Chambers, F. and **Richards, B.J.** 1992: Criteria for oral assessment. *Language Learning Journal* 6, 5–9.

Chotlos, J.W. 1944: Studies in language behaviour: IV. A statistical and comparative analysis of individual written samples. *Psychological Monographs* 56, 75–111.

Cross, T.G. 1977: Mothers' speech adjustments: the contribution of selected child listener variables. In Snow, C.E. and Ferguson, C.A., editors, *Talking to children: language input and acquisition*. Cambridge: Cambridge University Press, 151–88.

Fairbanks, H. 1944: Studies in language behavior: II. The quantitative differentiation of samples of spoken language. *Psychological Monographs* 56, 19–38.

Gass, S. and **Varonis, E.** 1985: Variation in native speaker speech modification to non-native speakers. *Studies in Second Language Acquisition* 7, 37–57.

Guiraud, P. 1960: *Problèmes et méthodes de la statistique linguistique.* Dordrecht: Reidel.

He, A.W. and **Young, R.** 1998: Language proficiency interviews: a discourse approach. In Young, R. and He, A.W., editors, *Talking and testing: discourse approaches to the assessment of oral proficiency*. Amsterdam: John Benjamins, 1–24.

Herdan, G. 1960: *Quantitative linguistics.* London: Butterworth.

Hess, C.W., Haug, H.T. and **Landry, R.G.** 1989: The reliability of type-token ratios for the oral language of school age children. *Journal of Speech and Hearing Research* 32, 536–40.

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

Johnson, M. and **Tyler, A.** 1998: Re-analyzing the OPI: how much does it look like natural conversation? In Young, R. and He, A.W., editors, *Talking and testing: discourse approaches to the assessment of oral proficiency*. Amsterdam: John Benjamins, 28–51.

Johnson, W. 1944: Studies in language behavior: I. A program of research. *Psychological Monographs* 56, 1–15.

Lazaraton, A. 1992: The structural organization of a language interview: a conversation analytic perspective. *System* 20, 373–86.

—— 1996: Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 13, 151–72.

McKee, G., Malvern D.D. and **Richards, B.J.** 2000: Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing* 15, 323–37.

MacWhinney, B. 2000a: *The CHILDES Project: tools for analyzing talk. Volume I: Transcription format and programs*. 3rd edition. Mahwah, NJ: Erlbaum.

—— 2000b: *The CHILDES Project: tools for analyzing talk. Volume II: The database*. 3rd edition. Mahwah, 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., editors, *Evolving models of language*. Clevedon: Multilingual Matters, 58–71.

—— 2000: Validation of a new measure of lexical diversity. In Beers, M., v. d. Bogaerde, B., Bol, G., de Jong, J. and Rooijmans, C., editors, *From sound to sentence: studies on first language acquisition*. Groningen: Centre for Language and Cognition, 81–96.

Mann, M.B. 1944: Studies in language behavior: III. The quantitative differentiation of samples of written language. *Psychological Monographs* 56, 41–74.

Manschreck, T.C., Maher, B.A. and **Ader, D.N.** 1981: Formal thought disorder, the type-token ratio, and disturbed voluntary movement in schizophrenia. *British Journal of Psychiatry* 139, 7–15.

Meara, P. 1978: Schizophrenic symptoms in foreign language learners. *UEA Papers in Linguistics* 7, 22–49.

Ménard, N. 1983: *Mesure de la richesse lexicale.* Geneva: Slatkine.

Moder, C.L. and **Halleck G.B.** 1998: Framing the language proficiency interview as a speech event: native and non-native speakers' questions. In Young, R. and He, A.W., editors, *Talking and testing: discourse approaches to the assessment of oral proficiency*. Amsterdam: John Benjamins, 117–46.

Pine, J. 1994: The language of primary caregivers. In Gallaway, C. and Richards, B.J., editors, *Input and interaction in language acquisition*. Cambridge: Cambridge University Press, 15–37.

Read, J. 2000: *Assessing vocabulary.* Cambridge: Cambridge University 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 **Chambers, F.** 1996: Reliability and validity in the GCSE oral examination. *Language Learning Journal* 14, 28–34.

Richards, B.J. and **Malvern, D.D.** 1997: *Quantifying lexical diversity in the study of language development*. Reading: The University of Reading New Bulmershe Papers.

—— 1999: The application of a new measure of lexical diversity to preschool children. Paper presented at the 8th International Congress for the Study of Child Language, The University of the Basque Country, San Sebastian-Donostia, Spain.

—— 2000: Accommodation in oral interviews between foreign language learners and teachers who are *not* native speakers. *Studia Linguistica* 54, 260–71.

Ross, S. 1992: Accommodative questions in oral proficiency interviews. *Language Testing* 9, 173–86.

Ross, S. and **Berwick, R.** 1992: The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14, 159–76.

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

Thakerar, J.N., Giles, H. and **Cheshire, J.** 1982. Psychological and linguistic parameters of speech accommodation theory. In Fraser, C. and Scherer, K.R., editors, *Advances in the social psychology of language*. Cambridge: Cambridge University Press, 205–55.

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, 323–52.

van Lier, L. 1989: Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly* 23, 489–508.

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

Wachal, R.S. and **Spreen, O.** 1973: Some measures of lexical diversity in aphasic and normal language performance. *Language and Speech* 16, 169–81.

Wesche, M. 1994: Input and interaction in second language acquisition. In Gallaway, C. and Richards, B.J., editors, *Input and interaction in language acquisition*. Cambridge: Cambridge University Press, 219–49.

Young, R. and **Milanovic, M.** 1992: Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition* 14, 403–04.