7. An investigation of the lexical dimension of the IELTS Speaking Test

Authors
John Read
University of Auckland

Paul Nation
Victoria University of Wellington

This study investigates vocabulary use by candidates in the IELTS Speaking Test by measuring lexical output, variation and sophistication, as well as the use of formulaic language.

ABSTRACT

This is a report of a research project to investigate vocabulary use by candidates in the current (since 2001) version of the IELTS Speaking Test, in which *Lexical resource* is one of the four criteria applied by examiners to rate candidate performance. For this purpose, a small corpus of texts was created from transcriptions of 88 IELTS Speaking Tests recorded under operational conditions at 21 test centres around the world. The candidates represented a range of proficiency levels from Band 8 down to Band 4 on the nine-band IELTS reporting scale. The data analysis involved two phases: the calculation of various lexical statistics based on the candidates' speech, followed by a more qualitative analysis of the full transcripts to explore, in particular, the use of formulaic language. In the first phase, there were measures of lexical output, lexical variation and lexical sophistication, as well as an analysis of the vocabulary associated with particular topics in Parts 2 and 3 of the test.

The results showed that, while the mean values of the statistics showed a pattern of decline from Band 8 to Band 4, there was considerable variance within bands, meaning that the lexical statistics did not offer a reliable basis for distinguishing oral proficiency levels. The second phase of the analysis focused on candidates at Bands 8, 6 and 4. It showed that the sophistication in vocabulary use of high-proficiency candidates was characterised by the fluent use of various formulaic expressions, often composed of high-frequency words, perhaps more so than any noticeable amount of low-frequency words in their speech. Conversely, there was little obvious use of formulaic language among Band 4 candidates. The report concludes with a discussion of the implications of the findings, along with suggestions for further research.

AUTHOR BIODATA

JOHN READ

John Read is an Associate Professor in the Department of Applied Language Studies and Linguistics, University of Auckland, New Zealand. In 2005, while undertaking this research study, he was at Victoria University of Wellington. His research interests are in second language vocabulary assessment and the testing of English for academic and professional purposes. He is the author of *Assessing Vocabulary* (Cambridge, 2000) and is co-editor of the journal *Language Testing*.

PAUL NATION

Paul Nation is Professor of Applied Linguistics in the School of Linguistics and Applied Language Studies, Victoria University of Wellington, New Zealand. His research interests are in second language vocabulary teaching and learning, as well as language teaching methodology. He is the author of *Learning Vocabulary in Another Language* (Cambridge, 2001) and also the author or co-author of widely used research tools such as the Vocabulary Levels Test, the Academic Word List and the Range program.

CONTENTS

1   INTRODUCTION

The revised Speaking Test for the International English Language Testing System (IELTS), introduced in 2001, involved various changes in both the way that a sample of speech is elicited from the candidates and in the criteria used to rate their performance. From our perspective as vocabulary researchers, a number of issues stimulated our interest in investigating the test from a lexical perspective. An obvious one is that, whereas examiners previously assessed each candidate on a single global scale incorporating various descriptors, the rating is now done more analytically with four separate scales, one of which is *Lexical resource*. Examiners are required to attend to the accuracy and range of the candidate's vocabulary use as one basis for judging his or her performance. A preliminary study conducted by Cambridge ESOL with a pilot version of the revised test showed a very high correlation with the grammar rating scale, and indeed with the fluency one as well (Taylor and Jones, 2001), suggesting the existence of a halo effect, and perhaps a lack of salience for the examiners of lexical features of the candidates' speech. Thus, there is scope to investigate characteristics of vocabulary use in the Speaking Test, with the possible outcome of guiding examiners in what to consider when rating the lexical resource of candidates at different proficiency levels.

A second innovation in the revised test was the introduction of the Examiner Frame, which largely controls how an examiner conducts the Speaking Test, by specifying the structure of the interaction and the wording of the questions. This means that the examiner's speech in the test is quite formulaic in nature. We were interested to determine if this might influence what the candidates said. Another possible influence on the formulaic characteristics of the candidates' speech is the growing number of IELTS preparation courses and materials, including at least one book (Catt, 2001) devoted just to the Speaking Test. The occurrence of formulaic language in the test would not in itself be a problem. One needs to distinguish here between purposeful memorising of lexical phrases specifically to improve test performance – which one might associate with less proficient candidates – and the skilful use of a whole range of formulaic sequences which authors like Pawley and Syder (1983) see as the basis of fluent native-like oral proficiency.

More generally, the study offered an opportunity to analyse spoken vocabulary use. As Read noted (2000: 235-239), research on vocabulary has predominantly focused on the written language because – among other reasons – written texts are easier to obtain and analyse. Although the speaking test interview is rather different from a normal conversation (cf van Lier, 1989), it represents a particular kind of speech event which is routinely audiotaped, in keeping with the operational requirements of the testing program. As a result, a large corpus of learner speech from test centres all around the world is available for lexical and other analyses once a selection of the tapes has been transcribed and edited. Thus, a study of this kind had the potential to shed new light on the use of spoken vocabulary by second language learners at different levels of proficiency.

2   LITERATURE REVIEW

Both first and second language vocabulary research have predominantly been conducted in relation to reading comprehension ability and the written language in general. This reflects the practical difficulties of obtaining and transcribing spoken language data, especially if it is to be "natural", ie, unscripted and not elicited. The relative proportions of spoken and written texts in major computer corpora such as the Bank of English and British National Corpus maintain the bias towards the written language, although a number of specialised spoken corpora like the CANCODE (Cambridge and Nottingham Corpus of Discourse in English) and MICASE (Michigan Corpus of Academic Spoken English) are now helping to redress the balance.

To analyse the lexical qualities of texts, scholars have long used a range of lexical statistics. Here again, for practical reasons, the statistics have, until recently, been applied mostly to written rather than spoken texts. Nevertheless, they potentially have great value in allowing us to describe key features of spoken vocabulary in a quantitative manner that may provide useful comparisons between test-takers at different proficiency levels. Read (2000: 197-213), in an overview of the statistical procedures, identifies the main qualities which the statistics are designed to measure: lexical density; lexical variation; and lexical sophistication.

Lexical density is operationalised as the proportion of content words in a text. It has been used to distinguish the relative denseness of written texts from that of oral ones, which tend to have lower percentages of nouns, verbs and adjectives. In a language testing context, O'Loughlin (1995) showed that candidates in a "direct" speaking test, in which they interacted with an interviewer, produced speech with a lower lexical density than those who took a "semi-direct" version of the test, which required test-takers to respond on audiotape to pre-recorded stimulus material with no interviewer present.

Lexical variation, which has traditionally been calculated as the type-token ratio (TTR), is simply the proportion of different words used in the text. It provides a means of measuring what is often referred to as "range of vocabulary". However, a significant weakness of the TTR when it is used to compare texts is the sensitivity of the measure to the variable length of the texts. Various unsatisfactory attempts have been made over the years to correct the problem through algebraic transformations of the ratio. Malvern and Richards (Durán, Malvern, Richards and Chipere, 2004) argue they have found a solution with their measure, D, which involves drawing multiple word samples from the text and plotting the resulting TTRs on a curve that allows the relative lexical diversity of even quite short texts to be determined. In a study which is of some relevance to our research, Malvern and Richards (2002) used D to investigate the extent to which teachers, acting as examiners in a secondary school French oral examination, accommodated their vocabulary use to the ability level of the candidates.

Lexical sophistication can be defined operationally as the percentage of low-frequency, or "rare", words used in a text. One such measure is Laufer and Nation's (1995) Lexical Frequency Profile (LFP), which Laufer (1995) later simplified to a "Beyond 2000" measure – the percentage of words in a text that are not among the most frequent 2000 in the language. Based on the same principle, Meara and Bell (2001) developed their program called P_Lex to obtain reliable measures of lexical sophistication in short texts. It calculates the value lambda by segmenting the text into 10-word clusters and identifying the number of low-frequency words in each cluster. As yet, there is no published study which has used P_Lex with spoken texts.

Apart from the limited number of studies using lexical statistics, recent work on spoken vocabulary has highlighted a number of its distinctive features, as compared to words in written form. One assumption that has been widely accepted is that the number of different words used in informal speech is substantially lower than in written language, especially of the more formal kind. That is to say, a language user can communicate effectively through speaking with a rather smaller vocabulary than that required for written expression. There has been very little empirical evidence for this until recently. In their study of the CANCODE corpus, Adolphs and Schmitt (2003) found a vocabulary of 2000 word families could account for 95% of the running words in oral texts, which indicates that learners with this size of vocabulary may still encounter quite a few words they do not know. The authors suggest that the target vocabulary size for second language learners to have a good foundation for speaking English proficiently should be around 3000 word families, which is somewhat larger than previously proposed.

But perhaps the most important area in the investigation of spoken vocabulary is the use of multi-word lexical items. This represents a move away from the primary focus on individual word forms and word families in vocabulary research until now. Both in manual and computer analysis, it is simpler to count individual forms than any larger lexical units, although corpus linguists are now developing sophisticated statistical procedures to identify collocational patterns in text.

The phenomenon of collocation has long been recognised by linguists and language teaching specialists, going back at least to Harold Palmer (1933, cited in Nation, 2001: 317). What is more recent is the recognition of its psycholinguistic implications. The fact that particular sequences of words occur with much greater than chance probability is not simply an interesting characteristic of written and spoken texts, but also a reflection of the way that humans process natural language. Sinclair (1991) distinguishes two approaches to text construction: the open-choice principle, by which language structures are generated creatively on the basis of rules; and the idiom principle, which involves the building of text from prefabricated lexical phrases. Mainstream linguistics has tended to overlook or undervalue the significance of the latter approach.

Another seminal contribution came from Pawley and Syder (1983), who argued that being able to draw on a large memorised store of lexical phrases was what gave native speakers both their ability to process language fluently and their knack of expressing ideas or speech functions in the appropriate manner. Conversely, learners reveal their non-nativeness in both ways. According to Wray (2002: 206), first language learners focus on large strings of words and decompose them only as much as they need to, for communicative purposes, whereas adult second language learners typically store individual words and draw on them, not very successfully, to compose longer expressions as the need arises. This suggests one interesting basis for distinguishing candidates at different levels in a speaking test, by investigating the extent to which they are able to respond fluently and appropriately to the interviewer's questions.

Applied linguists are showing increasing interest in the lexical dimension of language acquisition and use. In their research on task-based language learning, Skehan and his associates (Skehan, 1998; Mehnert, 1998; Foster, 2001) have used lexical measures as one means of interpreting the effects of different task variables on learners' oral production. As part of his more theoretical discussion of the research, Skehan (1998) proposes that the objective of good task design is to achieve the optimum balance between promoting acquisition of the rule system (which he calls syntacticisation) and encouraging the fluent use of lexical phrases (or lexicalisation).

Wray's (2002) recent book on formulaic language brings together for the first time a broad range of work in various fields and will undoubtedly stimulate further research on multi-word lexical items. In addition, Norbert Schmitt, Zoltan Dornyei and their associates at the University of Nottingham have just completed a series of studies on factors influencing the acquisition of multi-word lexical structures by international students at the university (Schmitt, 2004).

Another line of research relevant to the proposed study is work on the discourse structure of oral interviews. Studies in this area in the 1990s included Ross and Berwick (1992), Young and Milanovic (1992) and Young and He (1998). Lazaraton (2001), in particular, has carried out such research on an ongoing basis in conjunction with UCLES, including her recent analysis of the new IELTS Speaking Test (Lazaraton, 2000, cited in Taylor, 2001).

In one sense, a lexical investigation gives only a limited view of the candidates' performance in the speaking test. It focuses on specific features of the spoken text rather than the kind of broad discourse analysis undertaken by Lazaraton and appears to relate to just one of the four rating scales employed by examiners in assessing candidates' performance. Nevertheless, the literature cited above gives ample justification to explore the Speaking Test from a lexical perspective, given the lack of previous research on spoken vocabulary and the growing recognition of the importance of vocabulary in second language learning.

3   RESEARCH QUESTIONS

Based on our reading of the literature, we set out to address the following questions:

1.   What can lexical statistics reveal about the vocabulary of a corpus of IELTS Speaking Tests?
2.   What are the distinctive characteristics of candidates' vocabulary use at different band score levels?
3.   What kinds of formulaic language are used by candidates in the Speaking Test?
4.   Does the use of formulaic language vary according to the candidate's band score level?

Formulaic language is used here as a cover term for multi-word lexical items, following Wray (2002: 9), who defines a formulaic sequence as:

> *a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.*

4   METHOD

4.1   The format of the IELTS Speaking Test

As indicated in the introduction, the IELTS Speaking Test is an individually administered test conducted by a single examiner and is routinely audiotaped. It takes 11–14 minutes and consists of three parts:

- Part 1: Interview (4–5 minutes)
  The candidate answers questions about himself/herself and other familiar topic areas.
- Part 2: Long Turn (3–4 minutes)
  After some preparation time, the candidate speaks for 1–2 minutes on a topic given by the examiner.
- Part 3: Discussion (4–5 minutes)
  The examiner and candidate discuss more abstract issues and concepts related to the Part 2 topic.

The examiner rates the candidate's performance on four nine-band scales: *Fluency and coherence; Lexical resource; Grammatical range and accuracy;* and *Pronunciation*. The four criteria have equal weighting and the final score for speaking is the average of the individual ratings, rounded to a whole band score.

4.2   Selection of texts

The corpus of spoken texts for this project was compiled from audiotapes of actual IELTS tests conducted at various test centres around the world in 2002. The tapes had been sent to Cambridge ESOL as part of the routine monitoring process to ensure that adequate standards of reliability are being maintained. The Research and Validation Group of Cambridge ESOL then made a large inventory of nearly 2000 tapes available to approved outside researchers. The list included the following data on each candidate: centre number; candidate number; gender; module (Academic or General Training); Part 2 task number; and band score for Speaking.

The original plan was to select the tapes of 100 candidates for the IELTS Academic Module according to a quota sample. The first sampling criterion was the task (or topic) for Part 2 of the test. We wanted to restrict the number of tasks included in the sample because we were aware that the topic would have quite an influence on the candidates' choice of vocabulary and we wanted to be able to reveal its effect by working with just a restricted number of tasks. Thus, the sample was

limited to candidates who had been given one of four Part 2 tasks: Tasks 70, 78, 79, 80. The choice of these specific tasks was influenced by the second criterion, which was that the band scores from 4.0 to 8.0 should be evenly represented, to allow for meaningful comparisons of the lexical characteristics of candidate speech at different proficiency levels, and in particular at Bands 4.0, 6.0 and 8.0. Since there are relatively fewer IELTS candidates who score at Band 4.0 or Band 8.0, compared to the scores in between, it was important to select tasks for which there was an adequate number of tapes across the band score range in the inventory. The four tasks chosen offered the best coverage in this sense.

The score that we used for the selection of candidates was the overall band level for Speaking, rather than the specific rating for *Lexical resource* (which was also available to us). We decided that, for the purpose of our analyses, it was preferable to classify the candidates according to their speaking proficiency, which was arguably a more reliable and independent measure than the *Lexical resource* score. In practice, though, the two scores were either the same or no more than one point different for the vast majority of candidates.

Where there were more candidates available than we required, especially at Bands 5.0, 6.0 and 7.0, an effort was made to preserve a gender balance and to include as many test centres in different countries as possible.

However, it was not possible to achieve our ideal selection. Ours was not the first request for the speaking tapes to be received from outside researchers by Cambridge ESOL and thus a number of our selected tapes were no longer available or could not be located. Thus, the final sample consisted of 88 recorded Speaking Tests, as set out in Table 1.

The sample included 34 female and 54 male candidates. The tests had been administered in Australia, Cambodia, China, Colombia, Fiji, Hong Kong, India, Ireland, Libya, New Zealand, Peru, Pakistan, Sudan and the United Kingdom. This meant that a range of countries were included. Although the original intention was to select only Academic Module candidates, the sample included eight who were taking the General Training Module. This was not really a problem for the research because candidates for both modules take the same Speaking Test.

	Task 70	Task 78	Task 79	Task 80	Totals
Band 8	4	4	4*	3	15
Band 7	5	4	6	4	19
Band 6	5	5	5	4	19
Band 5	5	5	5	6	21
Band 4	4	4	2	4	14
Totals	23	22	22	21	88

*One of these tapes turned out to have a different Part 2 task. It was thus excluded from the analyses by task.

Table 1: The final sample of IELTS Speaking Test tapes by band score and Part 2 task

4.3   Preparation of texts for analysis

The transcription of the tapes was undertaken by transcribers employed by the Language in the Workplace Project at Victoria University of Wellington. They had been trained to follow the conventions of the Wellington Archive of New Zealand English transcription system (Vine, Johnson, O'Brien and Robertson, 2002), which is primarily designed for the analysis of workplace discourse. Since the transcribers were mainly Linguistics students employed part-time, the transcribing took nearly nine months to complete.

For the qualitative analyses, the full transcripts were used. To produce text files for the calculation of lexical statistics for the candidates' speech, the transcripts were electronically edited to remove all of the interviewer utterances, as well as other extraneous elements such as pause markings and notes on speech quality which had been inserted into the transcripts in square brackets. The resulting files were saved as plain text files and then manually edited to delete the hesitations *um*, *er* and *mm*; back-channelling utterances such as *mm*, *mhm*, *yeah*, *okay* and *oh*; and false starts represented by incompletely articulated words and by short phrases repeated verbatim. In addition, contracted forms were separated (*it'll* → *it 'll*, *don't* → *do n't*) and multi-word proper nouns were linked as single lexical items (*Margaret_Thatcher*, *Lord_of_the_Rings*).

5   STATISTICAL ANALYSES

5.1   Analytical procedures

To investigate the words used by the candidates, a variety of lexical statistics were calculated, using four different computer programs.

1. *WordSmith Tools* (Smith, 1998). This is a widely used program for analysing vocabulary in computer corpora. The *Wordlist* tool was used to identify the most frequently occurring content words, both in the whole corpus and in the texts for each of the four Part 2 tasks. It also provided descriptive statistics on the lexical output of candidates at the five band score levels. A second WordSmith tool, *Keyword*, allowed us to identify words that were distinctively associated with each of the tasks and with the whole corpus.

2. *Range* (Nation and Heatley, 1996). This program produces a profile of the vocabulary in a text according to frequency level. It includes three default English vocabulary lists – the first 1000 words, the second 1000 words (both from West, 1953) and the Academic Word List (Coxhead, 2000). The output provides a separate inventory of words from each list, plus words that are not in any of the lists. There are also descriptive statistics which give a summary profile and indicate the relative proportion of high and lower frequency words in the text. The Range program was used to produce profiles not for individual candidates but for each of the five band score levels represented in the corpus.

3. *P_Lex* (Meara and Bell, 2001). Whereas Range creates a frequency profile, P_Lex yields a single summary measure, lambda, calculated by determining how many non-high frequency words occur in every 10-word segment throughout the text. A low lambda shows that the text contains predominantly high-frequency words, whereas a higher value indicates the use of more lower-frequency vocabulary.

4. *D_Tools* (Meara and Miralpeix, 2004). The purpose of this pair of programs is to calculate the value of D, the measure of lexical diversity devised by Malvern and Richards. D values range from a maximum of 90 down to 0, reflecting the number of different words used in a text.

6   STATISTICAL RESULTS

6.1   Lexical output

Let us first review some characteristics of the overall production of vocabulary by candidates in the test. In Table 2, candidates have been classified according to their band score level and the figures show descriptively how many word forms were produced at each level.

	TOTALS		MEANS (standard deviations)	
	Tokens	Types	Tokens	Types
BAND 8 (n=15)	22,366	2374	1491.0 (565.9)	408.1 (106.0)
BAND 7 (n=19)	21,865	2191	1150.7 (186.7)	334.6 (46.0)
BAND 6 (n=19)	18,493	1795	937.3 (261.4)	276.7 (48.2)
BAND 5 (n=21)	15,989	1553	761.4 (146.7)	234.2 (35.5)
BAND 4 (n=14)	6931	996	475.8 (216.9)	166.6 (48.6)

Table 2: Lexical output of IELTS candidates by band score level (WordSmith analysis)

Since there were different numbers of candidates in the five bands, the mean scores in the third and fourth columns of the table give a more accurate indication of the band score distinctions than the raw totals. There is a clear pattern of declining output from top to bottom, with candidates at the higher band score levels producing a much larger amount of vocabulary on average than those at the lower levels, both in terms of tokens and types. It is reasonable to expect that more proficient candidates would have the lexical resources to speak at greater length than those who were less proficient. However, it should also be noted that all the standard deviations were quite large. That is to say, there was great variation within band score levels in lexical production, which means that number of words used is not in itself a very reliable index of the quality of a candidate's speech. For example, the range in length of the edited texts for Band 8 candidates was from 728 to 2741 words. Thus, high proficiency learners varied in how talkative they were and in the extent to which the examiner allowed them to speak at length in response to the test questions.

It would be possible to calculate type-token ratios (TTRs) from the figures in Table 2 – and in fact, the WordSmith output includes a standardised TTR. However, as noted above, the TTR is a problematic measure of lexical variation, particularly in a situation like the present one where candidate texts vary widely in length.

6.2   Lexical variation

To deal with the TTR problem, Malvern and Richards' D was calculated by means of D_Tools. The D values for the texts in our corpus are presented in Table 3. As noted in the table, there may be a small bug in the program, because seven texts yielded a value above 90, which is not supposed to happen. An inspection of the seven texts suggested the possibility that the use of rare or unusually diverse vocabulary by some more proficient candidates may tend to distort the calculation, but this will require further investigation. Leaving aside those anomalous cases, the pattern of the findings for lexical variation is somewhat similar to those for lexical output. The mean values for D decline as we go down the band score scale, but again the standard deviations show a large dispersion in the values at each band level, and particularly at Bands 7 and 6.

As a general principle, more proficient candidates use a wider range of vocabulary than less proficient ones, but D by itself cannot reliably distinguish candidates by band score.

	D (LEXICAL DIVERSITY)			
	Mean	**SD**	**Maximum**	**Minimum**
BAND 8 (n=11)*	79.0	4.9	87.5	72.0
BAND 7 (n=17)*	71.8	18.2	89.5	61.2
BAND 6 (n=18)*	67.2	16.0	81.4	57.0
BAND 5 (n=21)	63.4	11.3	86.7	39.5
BAND 4 (n=14)	60.7	11.4	76.1	37.5
* Seven candidates with abnormal D values were excluded				

Table 3: Summary output from the D_Tools Program, by band score level

6.3　Lexical sophistication

The third kind of quantitative analysis used the Range program to classify the words (in this case, the types) into four categories, as set out in Table 4. Essentially, the figures in the table provide Laufer and Nation's (1995) Lexical Frequency Profile for candidates at the five band score levels represented in our corpus.

If we look at the List 1 column, we see that overall at least half of the words used by the candidates were from the 1000 most frequent words in the language, but the percentage rises with decreasing proficiency, so that the high-frequency words accounted for two-thirds of the types in the speech of Band 4 candidates. Conversely, the figures in the fourth column ("Not in Lists") show the reverse pattern. Words that are not in the three lists represent less frequent and more specific vocabulary, and it was to be expected that the percentage of such words would be higher among candidates at Bands 8 and 7. In fact, there is an overall decline in the percentage of words outside the lists, from 21% at Band 8 to about 12% at Band 4.

	TYPES				
	List 1	List 2	List 3	Not in Lists	Total
BAND 8 (n=15)	1270 53.7%	347 14.7%	243 10.3%	504 21.3%	2364 100%
BAND 7 (n=19)	1190 54.6%	329 15.1%	205 9.4%	455 20.9%	2179 100%
BAND 6 (n=19)	1060 59.5%	266 14.9%	179 10.0%	277 15.5%	1782 100%
BAND 5 (n=21)	958 62.1%	222 14.4%	119 7.7%	243 15.8%	1542 100%
BAND 4 (n=14)	677 68.5%	132 13.3%	58 5.9%	122 12.3%	989 100%

KEY	
List 1	First 1000 words of the GSL (West, 1953)
List 2	Second 1000 words of the GSL
List 3	Academic Word List (Coxhead, 2000)
Not in Lists	Not occurring in any of the above lists

Table 4: Analysis by the Range program of the relative frequency of words (lemmas) used by candidates at different band score levels

The patterns for the two intermediate columns are less clear-cut. Candidates at the various band levels used a variable proportion of words from the second 1000 list, around an overall figure of 13–15%. In the case of the academic vocabulary in List 3, the speech of candidates at Bands 6–8 contained around 9–10% of these words, with the percentage declining to about 6% for Band 4 candidates. If we take the percentages in the third and fourth columns as representing the use of more "sophisticated" vocabulary, we can say that higher proficiency candidates used substantially more of those words.

Another perspective on the lexical sophistication of the speaking texts is provided by Meara and Bell's (2001) P-Lex program, which produces a summary measure – lambda – based on this same distinction between high and low-frequency vocabulary use in individual texts. As noted above, a low value of lambda shows that the text contains mostly high-frequency words, whereas a higher value is intended to indicate more sophisticated vocabulary use.

In Table 5, the mean values of lambda show the expected decline from Band 8 to 4, confirming the pattern in Table 4 that higher proficiency candidates used a greater proportion of lower-frequency vocabulary in their speech. However, the standard deviations and the range figures also demonstrate what was seen in Tables 2 and 3; except to some degree at Band 6, there was a great deal of variation within band score levels.

LAMBDA				
	Mean	**SD**	**Maximum**	**Minimum**
BAND 8 (n=15)	1.10	0.22	1.50	0.77
BAND 7 (n=19)	1.05	0.26	1.49	0.60
BAND 6 (n=19)	0.89	0.17	1.17	0.55
BAND 5 (n=21)	0.88	0.24	1.38	0.33
BAND 4 (n=14)	0.83	0.33	1.48	0.40

Table 5: Summary output from the P-Lex Program, by band score level

To get some indication of why such variation might occur, it is interesting to look at candidates for whom there is a big mismatch between the band score level and the value of lambda. There were four cases of Band 8 candidates with lambdas between 0.77 and 0.86. An inspection of their transcripts suggests the following tentative explanations:

- Candidate 62 may have been overrated as Band 8, based on the simple language used and the apparent difficulty in understanding some of the examiner's questions in Part 3 of the test.

- Candidate 19 spoke fluently in idiomatic English composed largely of high-frequency words.

- Three of them used relatively few technical terms in discussing their employment, their study and the Part 2 task.

- Candidate 76 used quite a lot of technical terminology in talking about his employment history but switched to a much greater proportion of high-frequency vocabulary in the rest of the test.

On the other hand, four Band 4 candidates had lambdas between 1.16 and 1.48. There is an interesting contrast between two Band 4 candidates who said relatively little in the test (their edited texts are both around 300 words) but who had markedly different lambdas. Candidate 78 responded in simple high-frequency vocabulary, which produced a value of 0.44, whereas Candidate 77 used quite a few somewhat lower frequency words, often ones that were repeated from the examiner's questions (*available, transport, celebrating, information, encourage*), and thus obtained a lambda of 1.48. The other Band 4 candidates with high lambdas also appeared to produce a good proportion of words outside the high-frequency vocabulary range, relative to their small lexical output. Another factor with some Band 4 candidates was that poor pronunciation reduced their intelligibility on tape, with the result that it was difficult for the transcriber to make a full record of what they said – and this may have affected high-frequency function words more than phonologically salient lexical items.

Some of these lexical characteristics of performance at the different band score levels are considered further below in the qualitative analysis of the transcripts.

6.4 Key words in the four tasks

To investigate the vocabulary associated with particular topics, the texts were classified according to the four Part 2 tasks represented in the corpus. There were about 21–23 texts for each task. Table 6 lists the most frequently occurring content word forms in descending order, according to the WordSmith word lists. The lists have been lemmatised, in the sense that a stem word and its inflected forms (*cook, cooking, cooked*; *book, books*) were counted as a single unit, or lemma.

TASK 70 Eating out (n=23)		TASK 78 Reading a book (n=22)		TASK 79 Language learning (n=21)		TASK 80 Describing a person (n=21)	
food	269	book	333	English	315	people	340
think	190	read	224	think	226	think	229
people	187	think	195	learn	175	know	157
like (vb)	177	people	130	language	130	famous	148
restaurant	151	like (vb)	126	like (vb)	129	good	88
time	125	time	82	people	105	name	87
good	117	friend	81	know	79	like (vb)	85
friend	104	name	80	start	79	person	79
eat	96	good	69	speak	72	friend	63
fast	86	work	61	school	67	work	59
place	79	different	49	friend	62	country	55
home	60	child	48	different	60	time	55
work	58	study	46	time	57	year	45
cook	54	story	44	country	55	day	43
know	54	life	43	study	54	life	40
country	50	television	43	important	53	family	39
travel	49	problem	42	good	51	help	39
family	47	write	38	year	50	study	37
year	41	family	37	difficult	48	city	36
nice	39	find	34	start	47	important	34
city	38	important	32	class	44	public	33
spend	38	interesting	32	music	42	different	32
name	38	country	29	work	42	example	30
traditional	36	help	28	listen	41	way	29
talk	31	learn	28	name	39	problem	29
different	30	city	26	word	38	history	25
find	30	love	25	teach	37	transport	25
study	28			teacher	35		
course	27			place	34		
prefer	26			write	34		
enjoy	26			grammar	31		
dish	25			interesting	31		
prepare	25			mean	31		
				travel	31		
				family	28		
				talk	27		
				new	26		
				university	26		
				foreign	25		

Note: Some high-frequency verbs which occurred fairly uniformly across the four tasks have been excluded: *get*, *go*, *make*, *say*, *see*, *use/used to* and *want*.

Table 6: The most frequent content words used by candidates according to their Part 2 topic (WordSmith Wordlist analysis)

The lists represent, in a sense, the default vocabulary for each topic – the mostly high-frequency words one would expect learners to use in talking about the topic. As such, these words will almost certainly not be salient for the examiners in rating the learners' lexical resource, except perhaps in the case of low-proficiency candidates who exhibit uncertain mastery of even this basic vocabulary.

It should be remembered that these lists come from the full test for each candidate, not just Parts 2 and 3, where the designated topic was being discussed. This helps to explain why words such as

friend, *people*, *family*, *study* and *country* tend to occur on all four lists because of the frequency of these words in Part 1 of the test, where candidates talked about themselves and their background, including in particular, a question about whether they preferred to socialise with family members or friends.

Words were selected for the four lists down to a frequency of 25. It is interesting to note some variation between topics in the number of words above that minimum level. The longest list was generated by Task 79, on language learning. This indicates that, from a lexical point of view, the candidates discussed this topic in similar terms, so that a relatively small number of words, including *English*, *learn*, *language*, *study*, *listen*, *word* and *talk*, recurred quite frequently. That is to say, their experience of language learning had much in common from a vocabulary perspective. By contrast, for Task 78 the list of frequently repeated words is noticeably shorter, presumably because the books that the candidates chose to discuss had quite varied characteristics. The same would apply to the people that candidates who were assigned Task 80 chose to talk about.

TASK 70 Eating out (n=23)		TASK 78 Reading a book (n=22)		TASK 79 Language learning (n=21)		TASK 80 Describing a person (n=21)	
food	463.1	read	342.8	English	713.1	he	346.5
restaurant	327.8	books	309.2	language	233.6	famous	270.4
fast	184.0	book	358.9	learn	251.1	people	115.2
eat	104.8	reading	102.2	speak	99.4	him	110.6
foods	90.0	story	66.4	learning	76.8	person	76.0
eating	86.7	children	57.2	languages	74.7	his	60.2
go	76.1	internet	38.4	school	72.4	public	53.0
cook	74.3	television	38.4	class	69.7	admire	51.5
like	58.8	girl	36.8	grammar	62.2	who	50.6
home	57.7	men	36.8	communicate	56.2	known	48.5
traditional	52.0	writer	35.1	foreign	52.1	media	45.7
restaurants	47.0	boy	29.7	started	40.5	become	42.0
dishes	45.3	this	28.6	words	37.7	she	39.0
cooking	45.3	hear	28.5	speaking	34.9	chairman	24.2
nice	42.2	women	27.4	teacher	33.8	president	24.2
out	40.0	fiction	24.3	difficult	32.4		
McDonalds	32.0			communication	29.3		
meal	31.3			listening	27.5		
delicious	29.3						
shop	26.6						
healthy	24.0						

Table 7: Results of the WordSmith Keyword analysis for the four Part 2 tasks

Another facility offered by WordSmith is a Keyword analysis, which identifies words occurring with high frequency in a particular text – or set of texts – as compared with their occurrence in a reference corpus. For this purpose, the texts associated with each of the four Part 2 tasks were collectively analysed by reference to the corpus formed by the texts on the other three tasks. The results can be seen in Table 7, which lists the keywords for each of the four tasks, accompanied by a keyness statistic, representing the extent of the mismatch in frequency between the words in the texts for a particular task and in the rest of the corpus.

The keyword results show more clearly than the previous analysis the semantically salient words associated with each task. From a lexical point of view, it is the vocabulary needed for the Part 2 long turn and the Part 3 discussion which dominates each candidate's Speaking Test.

7   QUALITATIVE ANALYSES

To complement the statistical analyses, a subset of the test transcripts was selected for a more qualitative examination. There were two aims in this part of the study:

1.   to identify lexical features of the candidate speech which might help to distinguish performance at different band score levels

2.   to seek evidence of the role that formulaic language might play in the Speaking Test.

7.1   Procedures

The approach to this phase of the analysis was exploratory and inherently subjective in nature. As we and others have previously noted (Wray 2002, Schmitt and Carter 2004, Read and Nation 2004), there is a great deal of uncertainty about both how to define formulaic language in general, and how to identify particular sequences of words as formulaic.

Our initial expectations were that formulaic language could potentially take a number of different forms in the IELTS Speaking Test:

1.   The examiner's speech in the test is constrained by a "frame", which is essentially a script specifying the questions that should be asked, with only limited options to tailor them for an individual candidate. This might give the examiner's speech a formulaic character which would in turn be reflected in the way that the candidate responded to the questions.

2.   In the case of high-proficiency candidates who were fluent speakers of the language, one kind of evidence for their fluency could be the use of a wide range of idiomatic expressions, ie, sequences of words appropriately conveying a meaning which might not be predictable from knowledge of the individual words. This would make their speech seem more native-like than that of candidates at lower band score levels.

3.   Conversely, lower-proficiency candidates might attempt such expressions but produce ones that were inaccurate or inappropriate.

4.   At a low level, candidates might show evidence of using (or perhaps *over*using) a number of fixed expressions that they had consciously memorised in an effort to improve their performance in the test. It could be argued that the widespread availability of IELTS preparation courses and materials might encourage this tendency.

In order to highlight contrasts between score levels, the transcripts at Bands 8, 6 and 4 for each of the four tasks were selected for analysis. Our strategy was to read each of the selected transcripts carefully, marking words, phrases and longer sequences that seemed to be lexically distinctive in the following ways:

- individual words that we judged to be of low frequency, whether or not they were accurately or appropriately used
- words or phrases which had a pragmatic or discourse function within the text
- sequences of words which could in some sense be regarded as formulaic.

At this point, it is useful to make a distinction between formulaic sequences which could be recognised as such on the basis of native speaker intuition, and sequences that were formulaic for the individual learner as a result of being stored and retrieved as whole lexical units, regardless of how idiomatic they might be judged as being by native speakers. One indication that a sequence was formulaic in the latter sense was that it was produced by the candidate with little if any pauses, hesitation or false start. Another was that the same sequence – or a similar one – was used by the candidate more than once during the test.

8   QUALITATIVE RESULTS

8.1   Band 8

As noted in the results of the statistical analyses, the candidates at Band 8 produced substantially more words as a group than did those at lower proficiency levels. However, the quality of their vocabulary use was also distinctive. This was reflected partly in their confident use of low frequency vocabulary items, particularly those associated with their employment or their leisure interests. Several of the Band 8 candidates in the sample were medical practitioners and here for example is Candidate 01 recounting the daily routine at his hospital:

> *..and after that um I should er go back to the <u>ward</u> to check <u>patients</u> and check if there's any <u>complication</u> from receiving the <u>drugs</u> er usually after er er giving the drugs er some drugs may cause <u>side effects</u> which need to my <u>intervention</u> ...*

The underlined words are obviously more or less technical terms in medicine and one would expect a doctor to have command of them.

Similarly, Candidate 48 described her favourite movie actor in this way:

> *... he is a very, very <u>versatile</u> actor like he's er he has got his own <u>styles</u> and <u>mannerisms</u> in a very short <u>span</u> er in two <u>decades</u> or two and a half decades he has <u>established himself</u> as a very good actor in the (<u>cine</u> field) ...*

The use of *styles* and *span* here may not be entirely "native-like", but the candidate was able to give a convincing description of the actor.

Thus, high-proficiency candidates have available to them a wide range of low frequency words that allow them to express more specific meanings than can be conveyed with more general vocabulary.

However, it is important to emphasise that such lower-frequency vocabulary does not necessarily occur with high density in the speech of Band 8 candidates. The sophistication of their vocabulary ability may also be reflected in their use of formulaic sequences – made up largely or entirely of high-frequency words – which give their speech a native-like quality. Here are some excerpts from the transcripts of Band 8 candidates, with some of the sequences that we consider to be formulaic underlined:

> *one of the main reasons [why he became a doctor] was both my parents are doctors so naturally <u>I got into that line</u> but I was also interested in this medicine, as such, of course <u>the money factors come into play</u> (Candidate 54)*

> *it's quite nice [describing a restaurant] it's er its er Japanese er <u>all type of food</u> but <u>basically what I like there is</u> the sushi, <u>I love sushi</u> so <u>I just enjoy going there</u> and when you go in they start shouting <u>and stuff,</u> very Japanese culture type of restaurant which is very good (Candidate 19)*

> *[after visiting a new place] ...I like to remember everything later on and er <u>I don't know</u> it's <u>a habit</u> I <u>just keep picking up</u> these small things like er + um if I go to the northern areas that is if I go to [place name] <u>or some place like that,</u> I'll be picking up these small pieces and them um on the way back when I look at them <u>I was like God,</u> I cannot explain why I got this, there's <u>just this weird stuff</u> that I've picked up ... (Candidate 72)*

A related feature of the speech of many Band 8 candidates was the use of short words or phrases functioning as pragmatic devices, or what Hasselgren (2002) has termed "smallwords". These include *you know*, *I mean*, *I guess*, *actually*, *basically*, *obviously*, *like* and *okay*. These tend to be semantically "empty" and as such might be considered outside the scope of a lexical analysis, but

nevertheless they need to be included in any account of formulaic language use. Here are some examples from Candidate 47, who possibly overdoes the use of such devices:

> *I'm a marine engineer <u>actually</u> so er I work on the ship and er <u>basically</u> ... we have to go wherever the ship <u>you know</u> goes and so <u>obviously</u> we are on the ship so <u>basically</u> I am taking care of the machinery and <u>that's it</u> so + er <u>well</u> I've travelled quite a lot <u>you know</u> <u>I mean</u> all around the world ...*

Another distinctive user was Candidate 38:

> *[I prefer America] er because um <u>to be frank</u> <u>like</u> er um people were nice <u>I mean</u> they were not biased or <u>you know</u> they didn't show any coloured preference <u>or whatever</u> <u>yeah</u> they were more friendly ...*

In most cases, these pragmatic devices did not occur as frequently as in these two excerpts, but they were still a noticeable feature of the speech in many of the Band 8 transcripts that we examined.

Another kind of device was the use of discourse markers to enumerate two or three points that the candidate wanted to make:

> *if you compare er my language with er English ... it's completely different ... because .. er <u>firstly</u> we write from right to left and in English you write from right to left ... um <u>another thing</u> the grammar our grammar it's not like English grammar ...*
> *(Candidate 62)*

> *my name has two meanings there's one um it's actually a Muslim name so there's two meanings to that <u>one</u> is that it means a guardian from heaven and <u>the second meaning</u> it's er second name it was given to a tribe of people that were lofty and known for their arrogance*
> *(Candidate 71)*

These discourse markers were not so common in candidate speech, which is perhaps a little surprising, particularly in relation to the Part 2 long turn, when the candidates were given a minute or so to prepare what they were going to say.

It is important not to overstate the extent to which the features identified so far can be found in the speech of all the candidates at Band 8. In fact, they varied in the extent to which their speech appeared to be formulaic, in the sense of containing a lot of idiomatic expressions, pragmatic devices and so on. Here is a candidate who expresses her opinion about the importance of English in a relatively plain style:

> *er I think the English language is very important now + at first it didn't used to be, actually it has been strong for the last er fifty years but importance was not given to it + now in every organisation in every school in every college, er basically at the university level everything is taught in English basically so you need to understand the language, I think we students are better off because we are studying from a younger age we understand the language but a big problem we have here is that + people don't communicate but now teachers encourage the students to speak in English and um + it is very important (Candidate 83)*

Apart from the words *actually* and *basically*, plus a phrase such as *are better off*, there is not much in this excerpt which could be considered formulaic in any overt way.

Another example is this candidate talking about the kind of friends he prefers:

> *er normally I prefer one or two very close friends so that I can discuss with them if I have any problems or things like that, I can have more contact close contact with them instead of having so many friends, but I have so many friends I make friends as soon as I see I see*

people for the first time it's like when I came here today I talk to a number of people here ... but I have I prefer to have just one or two friends who are very close to me (Candidate 64)

On the face of it, these opinions are expressed in very simple vocabulary without any idiomatic expression. It should be noted that the phrase *prefer to have one or two very close friends* was part of the preceding question asked by the examiner and thus the opening statement is formulaic in the sense that it echoes what the examiner said. On closer inspection, there are other phrases that could be formulaic, such as *or things like that*, *close contact with them*, *I make friends* and *it's like when*.

8.2   Band 6

As compared to Band 8 candidates, those at Band 6 had some similar features but overall they showed a more limited range of vocabulary and a smaller amount of idiomatic expression.

One tendency among Band 6 candidates was either to use an incorrect form of a word or to reveal some uncertainty about what the correct form was. For instance, Candidate 09 said *... if I go by myself maybe some <u>dangerous</u> or something* and *it's more e- <u>economy</u> if I travel with other <u>peoples</u>*. Similarly, Candidate 69 made statements such as *when I was <u>third</u> year old*, *the differences between health and <u>dirty</u>* and, in perhaps an extreme case of uncertainty, *... then my parents brang bring bringed me branged me here ...*.

One noticeable characteristic of many candidates at this level was the occurrence of a mixture of appropriate and inappropriate expression, both in individual word choice and in the longer word sequences which they produced. Here are some examples:

I think adventurous books are really good for um pleasure time where you can sit and you can think and read those books and really come into real world ... (Candidate 36)

No people rarely do [change their names] especially because er first of they're proud of their names and proud of their tribes if you ever ever er go through the history of those people ... they would think themselves like a very proudy person and most of the people don't change their name (Candidate 84)

Mm I think train is better because it's fast and convenient but sometimes when in the weekend there's many people who are travel by train to somewhere else + so I think that time is very busy (Candidate 10)

These examples illustrate how Band 6 candidates were able to communicate their meaning effectively enough, even though they made some errors and did not express themselves in the more idiomatic fashion that Band 8 candidates were capable of. Here is one further example, which includes low-frequency vocabulary such as *relaxation*, *dwelling*, *cassette recorder* and *distract*, as well as the formulaic expression (*it's) (just) a matter of...*, but in other respects it is not very idiomatic:

I usually listen to music as a relaxation time after duties at my dwelling it's just a matter of relaxation ( ) cassette recorder certain cassettes I have picked ( ) I am travelling from town to town in the recorder of my car I used to put it on just a matter of you can distract ( ) going by your thoughts ( ) cannot sleep ( ) so it's a matter of relaxation ( ) something to distract me ( ) also I enjoy it very much. (Candidate 87)

Candidates at Band 6 did not generally use pragmatic devices such as *actually*, *you know* and *I mean* with any frequency. Candidate 69 is a clear exception to this but the other transcripts contained few, if any examples, of such devices.

8.3   Band 4

First, it should be noted that there were some practical difficulties for the transcribers in accurately recording what candidates at this level said, both because of the intelligibility of their accent and because their answers to questions might not be very coherent, particularly when the candidate had not properly understood the question.

Although candidates at this level used predominantly high-frequency vocabulary, they often knew some key lower-frequency words related to familiar topics, which they would use without necessarily being able to incorporate them into well-formulated statements, as in this response by Candidate 77 about transport in his city:

> *Transport problems <u>locally</u> there is a problem of these ( ) and er <u>rickshaws</u> <u>motor rickshaws</u> a lot of problems of making <u>pollution</u> and er problems*

Here is another example from Candidate 18:

> *Er in my case I have a working holiday <u>visa</u> yes (before) I I worked as <u>salesperson</u> in <u>convenience</u> shop*

A third example is a description of a local festival by Candidate 73:

> *Er is er our <u>locality</u> is very famous we're <u>celebrating</u> [name] <u>festivals</u> and we er too celebrate with our er <u>relatives</u> and there's a big <u>gathering</u> there and we always er make <u>chitchat</u> and we <u>negotiate</u> and <u>deal</u> of our personal <u>characters</u> in such kind of + festi-festivals*

There was not a great deal of evidence of formulaic language among the candidates at the Band 4 level. In some respects, the most formulaic section of the test was at the very beginning, as in this exchange:

IN:	Can you tell me your full name please?
CM:	My full name is [full name]
IN:	Okay and um and what shall I call you?
CM:	Um you can call me [name]
IN:	[Name] okay [name] er can you tell me where you're from [name]?
CM:	Er I'm from [place] in [country] (Candidate 65)

Of course, this introductory part was formulaic to varying degrees for candidates at all levels of proficiency because examiners are required to go through the routine at the beginning of every test.

There was one Band 4 candidate who gave an unusually well-formulated response which seems quite formulaic in the sense of being perhaps a rehearsed explanation for her decision to study medicine:

> *I want to be a doctor because I think this is a meaningful job to use my knowledge to help others and also to contribute to the society (Candidate 45)*

More typically, the responses by Band 4 candidates to questions that they had understood were not nearly as well-formed as this. For example, Candidate 78 responded thus to a question about English teaching in her country:

> *Er in my school is very good I can er I'll er + read there er two years last + nine ten matric than I'll leave the school go to college + and there's no good English in colleges*

For the most part, there were only certain limited sequences which we could identify as in any way formulaic in the speech of these low-proficiency candidates. For instance, Candidate 80 used the formula *Yes of course* six times. Other phrases such as *most of the time*, *in my opinion*, *first of all*,

I don't know I'm not sure and *I like music very much* occur sporadically in the transcripts we examined.

Particularly in Part 3, which is designed to be the most challenging section of the test, the Band 4 candidates had difficulty in understanding the examiner's questions, let alone composing an adequate answer. However, even here they mostly did not have formulaic expressions to express their difficulty and to request a repetition of the question. Some used *pardon*, *please* or (*I'm*) *sorry*, or else just struggled to respond as best they could. Exceptions were *I do not understand* (Candidate 80) and *sorry I don't exactly understand what you're ( ) can you repeat please* (Candidate 45).

9   DISCUSSION

In this study we used a variety of statistical tools, as well as our own judgement, to explore the lexical characteristics of oral texts produced by IELTS candidates in the Speaking Test.

We decided to conduct most of the analyses using the band scores for speaking which had been assigned to the candidates' performance by the examiners in the operational situation. For research purposes, it might have been desirable to check the reliability of the original scores by having the tapes re-rated by two certificated examiners. On the other hand, the fact that the recording quality of the audiotapes was quite variable, and that rating of tapes is a different experience from assessing candidates live, meant that the re-ratings would not necessarily have produced more valid measures of the candidates' speaking ability.

Classifying the candidates by band score, then, we found that the lexical statistics revealed broad patterns in the use of individual word forms which followed one's general expectations:

- Higher proficiency candidates gave more extended responses to the questions and thus produced more vocabulary than lower proficiency candidates.

- Candidates with higher band scores also used a wider range of vocabulary than those on lower band scores.

- The speech of less proficient candidates contained a higher proportion of high-frequency words, particularly the first 1000 most frequent words in the language, reflecting the limitations of their vocabulary knowledge.

- Conversely, higher proficiency candidates used greater percentages of lower frequency words, demonstrating their larger vocabulary size and their ability to use more specific and technical terms as appropriate.

It is important, though, that all of these findings should be seen as tendencies of varying strengths rather than defining characteristics of a particular band score level, because in all cases, there was substantial variation within levels. Thus, for instance, some Band 8 candidates gave relatively short responses and used predominantly high-frequency word forms, whereas those at Band 4 often produced quite a few low-frequency words, which could form a substantial proportion of their lexical output. Another point worth reiterating here is that, following Nation (2001: 13-16), we are defining "high-frequency" as occurring among the 2000 most frequent words in English – and, in the case of the P_Lex analysis, even more narrowly as the first 1000 words. As Nation (pp 19) also notes, the distinction between high and low is a somewhat arbitrary one and many very familiar words are classified as low-frequency by this criterion. However, the division still seems to provide a useful basis for evaluating the lexical quality of these oral texts.

No particular analysis was conducted of technical terms used by these IELTS candidates. The test questions are not really intended to elicit much discussion of the candidate's field of study or employment, particularly since the same test material is used with both Academic and General Training candidates. Within the short time-span of the test, the examiner cannot afford to let the candidate speak at length on any one topic. Even the Part 2 "long turn" is supposed to be restricted to 1–2 minutes. Nevertheless, some more proficient candidates who were well-established professionals in medicine, finance or engineering did give relatively technical accounts of their professional experience and interests in Parts 1 and 2 of the test.

The WordSmith analyses of the four Part 2 tasks clearly showed the influence of the topic that was the focus of Parts 2 and 3 of each candidate's test. The distinctive, frequently occurring content words were mostly those associated with the Part 2 task, which then led to the more demanding follow-up questions in Part 3. One interesting point to emerge from the analysis of the four topics was that they varied in terms of the range of content vocabulary that they elicited. Task 79, which concerned the candidates' experience of learning English, was the most narrowly focused in this regard. In other words, the candidates who talked on this topic tended to draw on the same lexical set related to formal study of the language in a classroom. On the other hand, Tasks 78 (a book) and 80 (a person) required some generic terms, but also more specific vocabulary to talk about the particular characteristics of the book or person.

The qualitative analysis was exploratory in nature and the findings must be regarded as suggestive rather than in any way conclusive. As noted in the literature review, there are no well-established procedures for identifying formulaic language, which indeed can be defined in several different ways. We found it no easier than previous researchers to confidently identify multi-word units as formulaic in nature on the basis of a careful reading of the transcripts. The comparison of transcripts within and across Bands 4, 6 and 8 produced some interesting patterns of lexical distinction between candidates at these widely separated proficiency levels. However, we were also conscious of the amount of individual variation within levels, which of course was one of the findings of the quantitative analysis as well. It should also be pointed out that the candidates whose tapes we were working with comprised a relatively small, non-probabilistic sample of the IELTS candidates worldwide – another reason for caution in drawing any firm conclusions.

The simple fact of working with the transcripts obliged us to shift from focusing on the individual word forms that were the primary units of analysis for the statistical procedures to a consideration of how the forms combined into multi-word lexical units in the candidates' speech. This gave another perspective on the concept of lexical sophistication. In the statistical analyses, sophistication is conceived in terms of the occurrence of low frequency words in the language user's production. The qualitative analysis, particularly of Band 8 texts, highlighted the point that the lexical superiority of these candidates was shown not only by their use of individual words but also their mastery of colloquial or idiomatic expressions which were often composed of relatively high-frequency words.

10   CONCLUSION

In the first instance, this study can be seen as a useful contribution to the analysis of spoken vocabulary in English, an area which is receiving more attention now after a long period of neglect. Within a somewhat specialised context – non-native speakers performing in a high-stakes proficiency test – the research offers interesting insights into oral vocabulary use, both at the level of individual words and through multi-word formulaic units. The texts are incomplete in one sense, in that the examiner's speech has been deleted, but of course the primary focus of the assessment is on what the candidate says (and discourse analytic procedures such as those used by Lazaraton (2002) are more appropriate for investigating the interactive nature of the Speaking Test). Although oral

texts like these are certainly not as tidy as written ones, it appears that lexical statistics can provide an informative summary of some key aspects of the vocabulary they contain.

From the perspective of IELTS itself, it is important to investigate vocabulary use in the Speaking Test as part of the ongoing validation of the IELTS test, particularly as *Lexical resource* is one of the criteria on which the candidates' performance is assessed. Our findings suggest that it is not surprising if examiners have some difficulty in reliably rating vocabulary performance as a separate component from the other three rating criteria. Whereas broad distinctions can be identified across band score levels, we found considerable variation in vocabulary use by candidates within levels. Ideally, research of this kind will, in the longer term, inform a revision of the rating descriptors for the *Lexical resource* scale, so that they direct the examiners' attention to salient distinguishing features of the different bands. However, it would be premature to attempt to identify such features on the basis of the present study.

One fruitful area of further research would be to ask a group of IELTS examiners to listen to a sample of the Speaking Test tapes and discuss the features of each candidate's vocabulary use that were noticeable to them. Their comments could then be compared with the results of the present study to see to what extent there was a match between their subjective perceptions and the various quantitative measures. However, it should also be remembered that, in the operational setting, examiners need to be monitoring all four rateable components of the candidate's performance, thus restricting the amount of attention they can pay to *Lexical resource* or any one of the others. It may well be that it is unrealistic to expect them to reliably separate the components. Moreover, the formulaic nature of oral language, as we observed it in our data particularly among Band 8 candidates, calls into question the whole notion of a clear distinction between vocabulary and grammar. Thus, while as vocabulary researchers we emphasise the importance of the lexical dimension of second language performance, we also recognise that it represents one perspective among several on what determines how effectively a candidate can perform in the IELTS Speaking Test.

REFERENCES

Adolphs, S and Schmitt, N, 2003, 'Lexical coverage of spoken discourse', *Applied Linguistics,* vol 24, pp 425-438

Ball, F, 2001, 'Using corpora in language testing' in *Research Notes 6*, EFL Division, University of Cambridge Local Examinations Syndicate, Cambridge, pp 6-8

Catt, C, 2001, *IELTS Speaking – preparation and practice*, Catt Publishing, Christchurch

Durán, P, Malvern, D, Richards, B and Chipere, N, 2004, 'Developmental trends in lexical diversity', *Applied Linguistics,* vol 25, pp 220-242

Foster, P, 2001, 'Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers' in *Researching pedagogic tasks: Second language learning, teaching and testing,* eds M Bygate, P Skehan, and M Swain, Harlow: Longman, pp 75-93

Hasselgren, A, 2002, 'Learner corpora and language testing: Smallwords as markers of oral fluency' in *Computer learner corpora, second language acquisition and foreign language teaching,* eds S Granger, J Hung and S Petch-Tyson, John Benjamins, Amsterdam, pp 143-173

Hawkey, R, 2001, 'Towards a common scale to describe L2 writing performance' in *Research Notes 5,* EFL Division, University of Cambridge Local Examinations Syndicate, Cambridge, pp 9-13

Laufer, B, 1995, 'Beyond 2000: A measure of productive lexicon in a second language' in *The current state of interlanguage,* eds L Eubank, L Selinker and M Sharwood Smith, John Benjamins, Amsterdam, pp 265-272

Laufer, B and Nation, P, 1995, 'Vocabulary size and use: Lexical richness in L2 written production', *Applied Linguistics*, vol 16, pp 307-322

Lazaraton, A, 2002, *A qualitative approach to the validation of oral language tests*, Cambridge University Press, Cambridge

Malvern, D and Richards, B, 2002, 'Investigating accommodation in language proficiency interviews using a new measure of lexical diversity', *Language Testing*, vol 19, pp 85-104

McCarthy, M, 1990, *Vocabulary*, Oxford University Press, Oxford

McCarthy, M, 1998, *Spoken language and applied linguistics*, Cambridge University Press, Cambridge

Meara, P and Bell, H, 2001, 'P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts', *Prospect*, vol 16, pp 5-24

Meara, P and Miralpeix, I, 2004, *D_Tools* computer software, Lognostics (Centre for Applied Language Studies, University of Wales Swansea), Swansea

Mehnert, U, 1998, 'The effects of different lengths of time for planning on second language performance', *Studies in Second Language Acquisition,* vol 20, pp 83-108

Nation, ISP, 2001, *Learning vocabulary in another language*, Cambridge University Press, Cambridge

Nation, P and Heatley, A, 1996, *Range* computer program, English Language Institute, Victoria University of Wellington, Wellington

Pawley, A and Syder, FH, 1983, 'Two puzzles for linguistic theory: Native-like selection and native-like fluency' in *Language and communication,* eds JC Richards and RW Schmidt, Longman, London, pp 191-226

Read, J, 2000, *Assessing vocabulary*, Cambridge University Press, Cambridge

Read, J and Nation, P, 2004, 'Measurement of formulaic sequences' in *Formulaic sequences: Acquisition, processing and use*, ed N Schmitt, John Benjamins, Amsterdam, pp 23-35

Ross, S and Berwick, R, 1992, 'The discourse of accommodation in oral proficiency interviews', *Studies in Second Language Acquisition,* vol 14, pp 159-176

Schmitt, N (ed), 2004, *Formulaic sequences: Acquisition, processing and use*, John Benjamins, Amsterdam

Sinclair, J, 1991, *Corpus, concordance, collocation*, Oxford University Press, Oxford

Skehan, P, 1998, *A cognitive approach to language learning*, Oxford University Press, Oxford

Smith, M, 1998, *WordSmith tools*, version 3.0, computer software, Oxford University Press, Oxford

Taylor, L, 2001, 'Revising the IELTS Speaking Test: Developments in test format and task design' in *Research Notes, 5*, EFL Division, University of Cambridge Local Examinations Syndicate, Cambridge, pp 2-5

Taylor, L and Jones, N, 2001, 'Revising the IELTS Speaking Test' in *Research Notes, 4*, EFL Division, University of Cambridge Local Examinations Syndicate, Cambridge, pp 9-12

van Lier, L, 1989, 'Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversation', *TESOL Quarterly,* vol 23, pp 489-508

Wray, A, 2002, *Formulaic language and the lexicon*, Cambridge University Press, Cambridge

Young, R and He, AW (eds), 1998, *Talking and testing: Discourse approaches to the assessment of oral proficiency,* John Benjamins, Amsterdam

Young, R and Milanovic, M, 1992, 'Discourse variation in oral proficiency interviews', *Studies in Second Language Acquisition,* vol 14, pp 403-424