Applied Linguistics 24/4: 425–438

Lexical Coverage of Spoken Discourse

SVENJA ADOLPHS and NORBERT SCHMITT

University of Nottingham

The Schonell *et al*. (1956) study of Australian oral English found that 2,000 word families provided around 99 per cent lexical coverage of spoken discourse. Based on this, scholars have accepted that around 2,000 word families provide the lexical resources to engage in everyday spoken English discourse. This study analysed a modern spoken corpus (the CANCODE corpus) and found that 2,000 word families made up less than 95 per cent coverage. A second analysis was performed on the CANCODE and the spoken component of the British National Corpus, which found that around 5,000 individual words were required to achieve about a 96 per cent coverage figure. These results suggest that more vocabulary is necessary in order to engage in everyday spoken discourse than was previously thought. The implication is that a greater emphasis on vocabulary development is necessary as part of oral/aural improvement.

INTRODUCTION

Teachers and learners have always known that vocabulary is an essential foundation to language learning, and a flurry of scholarly work in the last two decades has reintroduced its importance to the academic world (e.g. Coady and Huckin 1997; Huckin *et al*. 1993; Nation 1990, 2001; Schmitt 2000; Schmitt and McCarthy 1997; Singleton 1999). One strand of research in this vocabulary revival concerns determining the number of words which are required to achieve the things a learner might want to do. A common initial goal for many learners is to speak conversationally in the L2. In English, the consensus is that around 2,000 word families provide the lexical resources to engage in this sort of everyday chat (e.g. Nation and Meara 2002; Schmitt 2000). This 2,000 figure is based on the main large-scale study into spoken discourse carried out to date—Schonell *et al*. (1956). It studied the verbal interaction of Australian workers and found that 2,000 word families covered nearly 99 per cent of the words used in their speech. This figure is key because it represents a major vocabulary size target for ESL/EFL learners, and thus has direct implications for language pedagogy in general and vocabulary instruction in particular. If the figure was appreciably higher or lower, the amount of vocabulary which learners needed to learn in order to enable competent oral skills would rise or fall accordingly. It is therefore important for the field to have a sound figure based on the best available information. Unfortunately, the Schonell *et al.* study is now half a century old, focused on a narrow group of informants, and by today's standards, relied on a relatively small corpus. Today, a much larger and more diverse spoken corpus is

available, the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), and it seems opportune to analyse this modern resource to confirm or revise the Schonell figures.

BACKGROUND

There have been a number of studies which explore the lexical coverage of written texts (see Nation 2001, and Nation and Waring 1997, for surveys of this area). However, we are aware of only one significant study into the lexical coverage of adult spoken general English: Schonell *et al*. (1956). In this study a team of researchers recorded around 1,723 male and 1,095 female Australian semi-skilled and unskilled workers, collecting a total of 512,647 words. The tapes were then manually transcribed and a word count manually tallied. The published report consists of a description of the methodology, a listing of the most frequent words in the corpus, and a brief discussion of the text coverage that these words provided. The study was a monumental effort, and provided the main evidence for the behaviour of spoken English up until recently, when computerized analyses of spoken corpora began to appear (e.g. McCarthy and Carter 1997; Leech *et al*. 2001). However, the fact that it was completed nearly fifty years ago means that it inevitably has serious shortcomings when viewed from the perspective of today's research methodology. The corpus focused on a very specific segment of society, namely semi-skilled and unskilled workers. Also, the subjects were all Australian. In order to gather a sufficient amount of data, only about half consisted of spontaneous speech; the other half was derived from interviews with researchers, which potentially differs from more spontaneous forms of speech. The main shortcoming is that the complete corpus collected totalled only around half a million words. While this is still a respectable size for specialized corpora, it is considered quite small for a general corpus of English. Current corpora with a written focus are counted in the hundreds of millions of words (British National Corpus [BNC]—100 million, Bank of English—500 million), while the specialized Michigan Corpus of Academic Spoken English [MICASE] contains 813,684 words (as of 25 April 2002). Spoken corpora are obviously more expensive and difficult to compile than written corpora, because the speech needs to be recorded and professionally transcribed into electronic format. However, the relatively small size of the Schonell *et al.* corpus must be considered a serious limitation.

A more current spoken corpus is the CANCODE corpus, which consists of around 5 million words of transcribed conversations which were recorded across a wide variety of settings across the UK and Ireland.[1] Most of the data were collected between 1994 and 1999 with a focus on gathering conversations from a variety of discourse contexts and speech genres. The collected conversations are divided into five categories which reflect differences in the relationships that hold between the speakers: intimate, sociocultural, professional, transactional, and pedagogic. These categories can

be placed on a continuum which reaches from the very private to the very public (see Adolphs and Schmitt, under review, for more information on these categories in relation to lexical coverage). Conversations were carefully selected to include speakers from a variety of socioeconomic backgrounds, ages and levels of education (for a comprehensive description of the CANCODE corpus see McCarthy 1998). A comparison of the Schonell *et al.* corpus and the CANCODE corpus is given in Table 1.

Now that the modern CANCODE corpus is available, it makes sense to determine whether its most frequent 2,000 word families provide a similar level of coverage (about 99 per cent) as the Schonell *et al*. corpus.

PROCEDURE

Schonell *et al.*/CANCODE analysis

The procedure of the current study essentially sets out to replicate the Schonell *et al*. word count analysis, but using the CANCODE corpus. The first step involved creating a frequency list of the words in the CANCODE. As the CANCODE is not lemmatized, or coded for word class, the word lists generated were based on individual word forms. In contrast, the Schonell *et al*. analysis was carried out in headwords, which effectively equates to word families. There are good practical reasons for carrying out a word count in word family units, primarily because it avoids the problem of counting related words such as *pay*, *pays*, *paying*, and *payments* as separate words. There are also good pedagogic reasons for analysing vocabulary in terms of word families, mainly because learners seem to mentally handle the members of a word family as a group (Nagy *et al*. 1989). Therefore, it was necessary to convert the frequency list consisting of individual word forms into a list of word families. This entailed manually going down the list and tallying the frequencies of all members of a word family and inserting the resulting word count figure under the headword for the specific word family. However, one problem with using word families lies in their definition; beyond the base word and its inflections,

Table 1: Comparison of the Schonell et al. and CANCODE corpora

Schonell	CANCODE
512, 647 words	5,001,978 words
Semi-skilled and unskilled workers	All segments of society
Australian	British and Irish
Manually tabulated	Computerized
Approx 1/2 speech among workers	Spontaneous conversations/ interactions
Approx 1/2 interview with researchers	

there is no consensus of exactly which derivational variants and compounds should be included and which should not. The criteria for inclusion into a word family in this study have been set as follows:

- all inflections (for example, for the word family based around *enjoy*, the inflections *enjoyed*, *enjoying*, and *enjoys* were included);
- derivatives involving suffixation (*enjoyable*, *enjoyment*);
- prefixed derivatives were not included;
- transparent contractions included under their 'full forms' (*'ll* = will, *n't* = not, *'m* = am, *'re* = are, and *'ve* = have);
- opaque contractions included under one potential 'full form' (*'d* represents either *had* or *would*, while *'s* represents *has* or *is* [possessive form was disregarded]. The total tally for these two contractions was placed under one of their potential forms, e.g. the complete tally for *'d* was placed under *would*);
- compounds were not included except those on the baseword list forming the basis of Paul Nation's RANGE program (Nation 2002);
- homographs were counted under the same word family (*bank* [financial institute] and *bank* [side of the river] were counted under the same headword).

These criteria were set to be as similar as possible to the criteria used by Schonell *et al*. in their study. The main differences stem from the different methodologies employed in the respective studies. Schonell *et al*. manually tabulated each instance of a word form's occurrence, and so were able to split contractions into their separate components by analysing the context of each occurrence (e.g. *I'd* = *I* + *would*). Our study, on the other hand, is based on a computer-generated frequency list, and consequently we were only able to derive a total frequency figure for contractions such as *I'd*. In cases where the contraction was unambiguous (*n't* = not), we included the number of occurrences of that contraction in the 'full form' of the contraction (e.g. instances of *n't* were tallied under *not*). In the ambiguous case of *'d*, we simply added its total tally to the word *would*. Although in many cases it actually represented *had*, for the purposes of our calculations, this did not matter. Both *had* and *would* were among the most frequent words, and so allotting the total tally solely to *would* rather than separating it into the *would* and *had* categories makes no difference in our calculations of the amount of lexical coverage provided by the contraction. Similarly, in the case of 's, we added the tally to *has*, while disregarding all 'possessive s' forms. We adjusted the total size of the corpus in our calculations to account for the contractions which were 'added' as full form words.

By recording each word manually, the researchers who worked on the Schonell *et al.* study were also able to determine which meaning sense each homograph represented. While this is clearly desirable, we simply lacked the resources to do this manually, and so were forced to count all identical word forms as part of the same family for purely practical reasons. Moreover, we did

not include prefixed members of a word family. Another difference concerns the common backchannel verbalizations which do not normally qualify as words, such as *eh*, *uh huh*, *mmm*, and *Oh!*. Biber *et al.* (1999) show that these items convey a great deal of meaning and are an important feature of spoken discourse. We thus decided to include them in our count, combining all of the various forms under one category labelled 'Backchannels'.

These methodological differences mean that the Schonell *et al*. count and ours are not strictly equivalent. On the one hand, some of the Schonell *et al*. word family counts will be higher than ours because they included marginally more compounds than we did, but those compounds were generally very low in frequency and so would not change the overall frequency figures to any substantial degree. They also included a few prefixed members in their word family counts. Offsetting this, they counted homographs under different headwords, which would lead to more, but smaller, word families in such cases. Thus the differences in counting methodology will tend to cancel themselves out. More importantly, both the number of homographs divided into multiple meaning senses and the number of prefixed forms included in word families were very small in the Schonell *et al.* word lists. Overall, we are confident that the counting methodologies from the Schonell *et al*. study and ours are sufficiently comparable.

Once the list of word families and their frequency of occurrence was entered into our spreadsheet, we simply divided various frequency levels by the total number of words in the corpus to arrive at a percentage of text coverage. For example, to derive the coverage figure for the most frequent 2,000 word families, we divided the total number of words occurring in those 2,000 families (5,003,727 words) by the total words in the corpus after it was adjusted for our contraction additions (5,280,558). This led to the calculation 5,003,727 ÷ 5,280,558 = 94.757 per cent.

CANCODE/BNC analysis

In order to supplement the Schonell *et al.* analysis, we also examined a further large corpus of general spoken English. The 100 million word BNC includes a spoken component of around 10 million words, of which approximately 4.2 million words are of a conversational nature similar to the CANCODE. The conversational data were collected by having 124 adults (aged 15+ years) tape their conversations for a period of two to seven days. The adults were chosen to provide a balanced sampling of age, sex, geographical location, and social class. Additional recordings were made by younger subjects as part of the University of Bergen COLT Project. Overall, around 700 hours of recordings were collected, and the total number of speakers appearing in the recordings number over 1,000. (See Burnard 1995, and Aston and Burnard 1998, for details.)

In addition to this conversational data, we also selected a small amount of BNC data which was not strictly spontaneous but which still reflected

everyday language use, e.g. meetings, lectures, and sermons. The CANCODE contains some of this material, and so we included it in our BNC sample to make the corpora more comparable, both in size and in content. The total size of the BNC corpus analysed was 4,505,462 words.

Because both the BNC and CANCODE corpora are compiled with individual word forms as the basic unit, we decided to carry out the comparative analysis in terms of word forms instead of word families. This would better ensure comparability across the two corpora, and would also provide a complementary perspective to the word families analysis outlined above. Other than skipping the word family compilation stage, the method of analysis was the same as outlined above. Because we did not need to manually compile word families with these data, we analysed these word counts up to the 5,000 frequency level.

RESULTS

The main purpose of the study was to ascertain whether the 2,000 word family coverage figure from the Schonell *et al*. corpus would be confirmed by a similar analysis of the newer and more representative CANCODE corpus. The results are illustrated in Table 2. The figures in the table reflect the figures reported in the original Schonell *et al*. study, with the equivalent CANCODE figures, as well as CANCODE coverage figures at additional frequency points. The CANCODE figures indicate that 2,000 word families cover only around 95 per cent of general spoken discourse, rather than the 98–99 per cent figure reached by Schonell *et al*. We continued our analysis to the most frequent 3,000 word families and found that they covered nearly 96 per cent of spoken discourse. These results suggest that a wider range of vocabulary is necessary to engage in everyday verbal communication than was previously thought.

The analysis of the BNC and CANCODE data inevitably showed lower coverage figures because the analysis is based on word forms rather than word families. English speakers would have to have a vocabulary approaching 5,000 individual words in order to achieve the 96 per cent coverage figure which is almost realized by 3,000 word families in Table 2. The results are illustrated in Table 3.

DISCUSSION

The Schonell *et al*. study found that 2,000 word families provided around 99 per cent lexical coverage of the spoken discourse of their subjects. Our analysis of the CANCODE corpus indicates a much lower percentage of lexical coverage, namely 94.76 per cent. Although some of this discrepancy may be due to the fact that we analysed corpora of different sizes (see below), the greatest portion is likely to be attributable to the CANCODE containing a wider range of speakers discussing a more diverse range of topics. We would argue that our figures are likely to be more representative of the kind of

Table 2: Vocabulary coverage of spoken discourse in the Schonell et al. and CANCODE corpora (word families)

Word families	Schonell *et al.* percentage	CANCODE percentage
89	71.22	71.96
145	78.69	77.23
209	83.44	80.60
451	91.21	86.57
674	94.22	89.23
990	96.38	91.52
1,281	97.48	92.85
1,623	98.31	93.93
2,000	—	94.76
2,279	99.17	95.20
2,500	—	95.48
3,000	—	95.91

— figures not available for the Schonell *et al.* corpus

Table 3: Vocabulary coverage of spoken discourse in BNC and CANCODE Corpora (individual word forms)

Words	BNC conversational percentage	CANCODE percentage
25	39.76	34.38
50	53.09	48.20
100	66.39	62.09
500	84.29	83.01
1,000	89.25	88.24
1,500	91.75	90.72
2,000	93.30	92.26
2,500	94.35	93.35
3,000	95.13	94.16
5,000	96.93	96.11

spoken discourse the typical native speaker or L2 learner would be in contact with, simply because the CANCODE corpus is a larger, more modern, and more diverse sample of general spoken English. The validity of the CANCODE data is also supported by the BNC/CANCODE comparison. These two, large, modern, carefully-constructed spoken corpora agree, within around 1 per cent, about the lexical coverage at frequency levels above 1,000 words. This suggests that a BNC analysis in terms of word families would also agree closely with our CANCODE word family results. We are inclined to believe that the congruent results from our two modern corpora better reflect the reality of today's English verbal communication than the Schonell *et al.* results.

If we accept that the CANCODE figures are more accurate estimates of the lexical coverage of general English as it is spoken today, then it is questionable whether 2,000 word families provide enough lexical resources for everyday spoken discourse. Two thousand word families only provide a little under 95 per cent coverage, so the next question is whether this will allow speakers to operate in a verbal environment without major lexical problems. Most scholars would accept that 99 per cent lexical coverage, as was derived from the Schonell *et al*. study, is adequate for language use, as this translates to only one word in a hundred being unknown. However, it is not nearly as clear whether 95 per cent lexical coverage (one unknown word in twenty) is sufficient. Unfortunately, there is almost no research which explores the percentage of words which need to be known in order to operate successfully in a spoken environment. The only direct study we are aware of found no absolute lexical percentage 'threshold'. Japanese EFL learners in this study who knew less than 80 per cent of the words in the target texts almost always had poor comprehension, while most learners required more than 95 per cent coverage for good comprehension (Bonk 2000). Beyond this one orally-based study, we must turn to research in the written mode for guidance. This type of research suggests that knowledge of between 95–99 per cent of the words in written discourse is necessary to process it adequately, which is congruent with Bonk's results. Laufer (1989) found that learners who knew 95 per cent of the words in text were more likely to be successful readers and had better comprehension scores. Hu and Nation (2000) discovered that at 95 per cent coverage, some learners could achieve adequate comprehension of a fiction text, and a lesser number were even able to do this at 90 per cent. However, most learners were not able to attain adequate comprehension at these levels, although almost all were able to at 98 per cent coverage. Hirsh and Nation (1992) suggest that in order for texts to be easy enough to read for pleasure, 98–99 per cent coverage is desirable. Carver (1994) found that, for native speakers, a known vocabulary level of around 99 per cent leads to materials that are neither too easy nor too difficult.

It must be stressed that coverage figures derived from reading can only be indicative, because operating in a spoken environment is different from operating in a written one. Interlocutors in an oral context can use a variety of communication strategies which can help them make efficient use of

whatever vocabulary knowledge they have. For instance, they can seek clarification if they have misunderstood the message, they can negotiate meaning, and they can use contextual cues (e.g. gestures, intonation, mutually observable aspects of the physical environment) to help them understand meaning, even if their vocabulary is partially deficient. All these factors should facilitate oral discourse with fewer different words than written discourse, and indeed a common finding is that spoken discourse uses a smaller variety of word types than written discourse. This line of reasoning suggests that a *lower* percentage of lexical coverage might be required for spoken discourse than written discourse. On the other hand, spoken discourse must be processed on-line with all the time constraints this entails. This, in turn, could indicate that a *higher* percentage of lexical coverage might be required. Thus, the on-line processing constraints and the facilitative effect of face-to-face contextualization/communication strategy use are in tension with each other, and at the moment it is difficult to say how they interact in terms of the amount of vocabulary required. Indeed, some of Bonk's (2000) subjects achieved good comprehension with vocabulary coverage in the 80–95 per cent range, possibly through the use of communication strategies, while other subjects knew 95–100 per cent of the words and still had poor comprehension.

It is possible that various oral situations may favour one factor or the other, and so different percentages of lexical coverage might be necessary in different situations. For example, a lecture situation where the instructor is speaking quickly about technical details with no visual support and no questions allowed would presumably put a premium on a listener's quick and accurate decoding of vocabulary, and so require relatively more of that vocabulary to be known. Conversely, a situation where two friends are leisurely putting together a chair with all of the pieces in plain sight, with an on-going negotiating process accompanying the task, would presumably require much less vocabulary to be known. In reality, the amount of lexical coverage required probably varies on a cline affected by both facilitative and constraining factors. We feel this is a key area to be explored in future vocabulary research.

Since there may be no one lexical coverage figure which would supply an adequate amount of vocabulary in every situation, we can only speculate on what our figures mean for pedagogy. More precise statements will have to wait until we know more about the percentage(s) of lexical coverage necessary to operate in a spoken environment. However, on balance, because our 94.76 per cent figure is at the low end of the range indicated in research of written text coverage, and because most of Bonk's (2000) subjects required 95 per cent coverage for good comprehension, we would tentatively suggest that 2,000 word families should provide the lexical resources necessary to *begin* to adequately engage in everyday communication, but that 3,000 word families (providing coverage of nearly 96 per cent) is a better goal if learners wish to minimize their lexical gaps.[2] Of course more vocabulary is better, but

at these frequency levels the percentage of coverage gained per 1,000 word families learned falls off sharply. For example, the net gain in coverage achieved by including the 1,000 word families between the 2,000 and 3,000 frequency levels is only 1.15 per cent (94.76–95.91 per cent), and the gain between the 3,000 and 4,000 levels would be less than this due to the decreasing frequency of occurrence.

The necessity of learning larger numbers of word families implies that orally-focused language programmes need to include a significant vocabulary component that is maintained over an extended period of time. An increased lexical focus in language programmes is being increasingly advocated (e.g. Nation 2001; Nation and Meara 2002; Schmitt 2000), and this study provides evidence that a more intensive lexical focus is also necessary in programmes that are orally-focused. The results from this study were based on an L1 language corpus, but they suggest that if L2 learners wish to converse in a manner similar to L1 speakers, they would require similar levels of vocabulary. However, the only way to be truly sure about the level of vocabulary necessary for L2 speakers to converse successfully would be to compile a corpus of proficient L2 language users, and to subject it to similar analyses.

Table 3 highlights the importance of the fact that our main discussion has been in terms of *word families*. If coverage is discussed in terms of individual words, 2,000 word forms provide a coverage percentage which is unlikely to be adequate, only 92.26–93.30 per cent. In real terms this means that around seven out of every one hundred words would be unknown. It also shows the importance of helping students to learn the various members of a word family instead of just single words. Schmitt and Zimmerman (2002) found that their learners typically knew some, but not all, members of a word family. They conclude that learning the various derivative members of a word family is not an easy task, and that learners could benefit from more explicit instruction and awareness-building in this area. They suggest that when presenting a new word to students, teachers could also introduce its derivative forms. This would help learners to begin thinking in terms of word families instead of individual words. They also suggest the explicit teaching of derivative suffixes.

In terms of the individual word form percentages of coverage, the BNC and the CANCODE are in relatively close agreement, with the BNC showing about a 1 per cent greater coverage overall. The lower frequency levels show a greater advantage, but this is probably due to the different way contractions are counted in the two corpora. In the BNC, the contractions (e.g. *'ll*) were represented as separate items, and generally occurred near the top of the frequency count. This pushed the percentage of coverage up at the highest frequency levels. The CANCODE left the contractions attached to their 'partner words' (e.g. *I'll* or *we'll*) and we counted them as they occurred. This means that the contractions in the CANCODE were spread further down the frequency count, resulting in a lower percentage of coverage in the highest frequency levels.

Readers might find it surprising that relatively few high-frequency word types make up such high percentages of lexical coverage. For example, only 89 different word families make up more than 70 per cent of the spoken discourse in the CANCODE corpus, and around 1,000 word families make up over 90 per cent. In fact, this is a well-attested phenomenon, with a relatively small number of very frequent words (or word families depending on your unit of analysis) typically making up the vast majority of the running words in a corpus, and a large number of lower-frequency words occurring only once or a few times (see Nation and Waring 1997, for more discussion). One ramification of this frequency distribution is that the last few percentage points of lexical coverage between 95 per cent and 100 per cent are made up of a great number of word types. This is illustrated by the fact that 15,779 word types in the CANCODE only occurred once, and together they made up only 0.0032 per cent of the total corpus. Based on this frequency distribution of vocabulary, it becomes clear that mastery of the most frequent words (or word families) is essential because they provide high lexical coverage, while mastery of low-frequency words is less critical because they are encountered relatively rarely in discourse.

One technical aspect of this study is worth discussing in more detail. We compared two corpora of different sizes (the Schonell *et al*. corpus of half a million words and the CANCODE with 5 million words), and it is important to know whether this size difference alone could have affected our results, as would be the case if we were doing a type/token analysis.[3] In a study of the effect of genre on the lexical coverage of spoken discourse (Adolphs and Schmitt, under review), we explored whether using genre-based sub-corpora from the CANCODE, with varying numbers of tokens (smallest = 456,177 tokens; largest = 1,709,598 tokens), would have an effect on the coverage percentages obtained. We analysed one of the larger sub-corpora (Transactional—1,166,825 words) and a subsample of that same sub-corpus (434,128 words), and found only small differences in lexical coverage at

Table 4: Comparison of lexical coverage between the Transactional sub-corpus and a smaller subsample of the transactional sub-corpus (individual word forms)

	Percentage of coverage		
	2,000 words	4,000 words	5,000 words
Transactional subsample (434,128 tokens)	94.39	97.45	98.20
Transactional full sub-corpus (1,166,825 tokens)	94.30	97.14	97.82

three frequency points (Table 4). Following up on this, we extracted a representative subsample of 497,658 tokens (similar in size to the Schonell *et al.* corpus) from the CANCODE which drew from all genre categories and passage lengths and compared the subsample and the full CANCODE. We found that differences between the two were not substantial, with the largest discrepancy being .65 percentage points at the 2,000 level (Table 5).[4] We also examined the five genre sub-corpora (of varying sizes) and found no obvious relationship between corpus size and lexical coverage (Adolphs and Schmitt, under review). We conclude from this that, in contrast to type/token ratios, lexical coverage figures are not greatly affected by corpus size, at least not for the size of corpora discussed here. As regards this study, corpus size probably has some influence on our figures, but the difference in lexical coverage potentially attributable to corpus size dissimilarity (.65 percentage points or less) does not appear to be large when compared to the differences in lexical coverage found between the Schonell *et al.* and CANCODE corpora, as shown in Table 2.

Table 5: Comparison of lexical coverage between the CANCODE and a sub-sample similar in size to the Schonell et al. corpus (individual word forms)

	Percentage of coverage		
	2,000 words	4,000 words	5,000 words
CANCODE subsample (497,658 tokens)	91.61	95.35	96.35
Full CANCODE (5,001,978 tokens)	92.26	95.32	96.11

CONCLUSION

Overall, it seems clear from the CANCODE and BNC results that more vocabulary is necessary in order to engage in everyday spoken discourse than was previously thought. However, until further research determines the percentage(s) of lexical coverage required for spoken discourse, it is difficult to set firm vocabulary size targets. What we can say, however, is that a greater emphasis on vocabulary development is probably necessary as part of the push to improve oral skills. Teachers and learners have always known vocabulary is important; this study merely indicates that we might need more of it than we had previously thought.

(Revised version received October 2002)

ACKNOWLEDGEMENTS

This study was inspired by John Read who first suggested that a new study into spoken text coverage based on modern corpora would be valuable. Thanks also to Ron Carter who commented on an earlier draft of this paper and to the three anonymous *Applied Linguistics* reviewers.

NOTES

1 CANCODE stands for Cambridge and Nottingham Corpus of Discourse in English and is a collaborative project between Cambridge University Press and the University of Nottingham. The corpus was funded by Cambridge University Press with whom sole copyright resides.

2 A reviewer insightfully commented that word families appearing in the 2,000–3,000 frequency band are relatively infrequent in the CANCODE, and are probably highly influenced by the specific topics, purposes, and settings of the CANCODE itself. This means that a different 5-million word corpus might produce a different set of words at this frequency level. While acknowledging this problem, it appears likely that learners need to know more than 2,000 words to function in spoken discourse, and so teachers, administrators, and materials writers still need principled means of deciding which words to address at this level. Although modern spoken corpora are not infallible, they remain the best source of information available at present.

3 Type/token ratios are very sensitive to text length, because 'as authors write more, they use fewer and fewer words that they have not already used earlier in the text' (Read 2000: 201–2). Thus the number of types does not normally rise as quickly as the number of tokens, and consequently a 'tailing off' of the type/token ratio occurs. Although type/token ratios have mainly been used to analyse written discourse, it is reasonable to expect a similar effect in spoken discourse.

4 The discrepancies between the lexical coverage percentages at any particular level in Tables 4 and 5 are probably due to the different types of language contained in the respective corpora. Table 5 gives figures for the full CANCODE corpus and a representative subsample of it, which are both made up of a wide variety of speech contexts. Table 4 reflects the language involved in the much narrower context of transactional speech, which is made up in part of the largely ritualized language of service encounters, as well as interviews with a narrow topic focus. Because this type of language is more routinized than general English, we would expect less vocabulary diversity. This is reflected in higher lexical coverage figures in Table 4 compared to Table 5 (see Adolphs and Schmitt, under review, for a more detailed description of the transactional category).

REFERENCES

Adolphs, S. and **N. Schmitt.** (Under review) 'Vocabulary Coverage According to Spoken Discourse Context', in P. Bogaards and B. Laufer (eds): *Vocabulary in a Second Language: Selection, Acquisition and Testing*. Amsterdam: John Benjamins.

Aston, G. and **L. Burnard.** 1998. *The BNC Handbook*. Edinburgh: Edinburgh University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, and **E. Finegan.** 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bonk, W. J. 2000. 'Second language lexical knowledge and listening comprehension.' *International Journal of Listening* 14: 14–31.

Burnard, L. 1995. *British National Corpus: Users Reference Guide*. Oxford: Oxford University Computing Services.

Carver, R. P. 1994. 'Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for

instruction.' *Journal of Reading Behavior* 26/4: 413–37.

Coady, J. and **T. Huckin.** (eds) 1997. *Second Language Vocabulary Acquisition*. Cambridge: Cambridge University Press.

Hirsh, D. and **I. S. P. Nation.** 1992. 'What vocabulary size is needed to read unsimplified texts for pleasure?' *Reading in a Foreign Language* 8/2: 689–96.

Hu, M. and **P. Nation.** 2000. 'Vocabulary density and reading comprehension.' *Reading in a Foreign Language* 13/1: 403–30.

Huckin, T., M. Haynes, and **J. Coady.** (eds) 1993. *Second Language Reading and Vocabulary Learning*. Norwood, NJ: Ablex.

Laufer, B. 1989. 'What percentage of text-lexis is essential for comprehension?' in C. Lauren and M. Nordman (eds): *Special Language: From Humans to Thinking Machines*. Clevedon: Multilingual Matters. pp. 316–23.

Leech, G., P. Rayson, and **A. Wilson.** 2001. *Word Frequencies in Written and Spoken English*. Harlow: Longman.

McCarthy, M. 1998. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M. and **R. Carter.** 1997. 'Written and spoken vocabulary' in N. Schmitt and M. McCarthy (eds): *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge: Cambridge University Press. pp. 20–39.

Nagy, W., R. C. Anderson, M. Schommer, J. A. Scott, and **A. C. Stallman.** 1989. 'Morphological families in the internal lexicon.' *Reading Research Quarterly* 24/3: 262–82.

Nation, I. S. P. 1990. *Teaching and Learning Vocabulary*. New York: Newbury House.

Nation, I. S. P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, P. 2002. RANGE Word Analysis Software. Available at *http://www.vuw.ac.nz/lals. Accessed 5 March 2002.*

Nation, P. and **P. Meara.** 2002. 'Vocabulary' in N. Schmitt (ed.): *An Introduction to Applied Linguistics*. London: Arnold.

Nation, P. and **R. Waring.** 1997. 'Vocabulary size, text coverage, and word lists' in N. Schmitt and M. McCarthy (eds): *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge: Cambridge University Press. pp. 6–19.

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

Schmitt, N. 2000. *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Schmitt, N. and **M. McCarthy.** (eds) 1997. *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge: Cambridge University Press.

Schmitt, N. and **C. B. Zimmerman.** 2002. 'Derivative word forms: What do learners know?' *TESOL Quarterly* 36/2: 145–71.

Schonell, F. J., I. G. Meddleton, and **B. A. Shaw.** 1956. *A Study of the Oral Vocabulary of Adults*. Brisbane: University of Queensland Press.

Singleton, D. 1999. *Exploring the Second Language Mental Lexicon.* Cambridge: Cambridge University Press.