CHAPTER 3

Vocabulary coverage according to spoken discourse context

Svenja Adolphs and Norbert Schmitt
University of Nottingham

Abstract

In 1956, Schonell *et al.* found that 2,000 word families provided around 99% lexical coverage of spoken discourse. Based on this, it has been generally accepted that approximately 2,000 word families provide the lexical resources to engage in everyday spoken discourse. However, we recently conducted a study of spoken discourse based on more modern corpora which found that 2,000 word families provide less than 95% coverage rather than 99%, suggesting that a wider range of vocabulary is required in speech than previously thought (Adolphs & Schmitt 2003). These results were for unscripted spoken discourse in general, but we know that spoken discourse is not a homogenous phenomenon; rather it varies to some extent according to a number of factors, such as degree of familiarity between interlocutors and purpose of the discourse. This chapter reports on a follow-up study which explored whether the percentage of lexical coverage also varies depending on the context in which the spoken discourse is embedded, or whether it remains constant regardless of the context. Our results show that in order to reach a vocabulary coverage in the mid-90% range, a larger number of word forms is required in contexts where interlocutors have intimate or friendship-based relationships compared to ones in which the interlocutors have a professional or business-based relationship. This indicates that the percentage of coverage is affected by the spoken discourse context.

1. Introduction

The study of vocabulary is an essential part of language learning and the question of how much vocabulary a learner needs to know to achieve a particular purpose remains an important area of research and discussion. Schonell *et al.*'s (1956) study of the verbal interaction of Australian workers found that 2,000 word families covered nearly 99% of the words used in their speech. This was a landmark study, but limitations of the time inevitably meant that their hand-compiled corpus would be limited in diversity and size. Miniaturization in tape recorder technology now allows spoken data to be gathered unobtrusively in a wide range of naturally-occurring environments. Likewise, modern technology in corpus linguistics allows the use of far larger corpora than in the past. Using a current 5 million word spoken corpus (compared to the Schnell *et al.* 512,000 word corpus), Adolphs & Schmitt (2003) found that 2,000 word families supply lexical coverage for less than 95% of spoken discourse. This indicates that a wider range of vocabulary is necessary to engage in spoken discourse than previously thought.

However, a limitation of the Adolphs & Schmitt study is that it looked at spoken discourse in general, treating all types of speech the same. This clearly oversimplifies the situation, as different spoken contexts have at least somewhat different characteristics. For example, Stenström (1990) compared the frequency and function of a range of discourse markers across different situations of speaking and found considerable differences in the use of these items between the situations. The spoken contexts she compared were a casual conversation between a couple and a narrative delivered to an audience. As the levels of interactivity in the two speaking situations differed, so did the frequency of certain discourse markers. McCarthy (1998) discovered similar variation when he looked at a number of other linguistic features, such as the frequency of deictic items and the use of full lexical words, while Carter & McCarthy (1995) found that different types of spoken discourse had different profiles of grammatical features. Spoken discourse also varies according to the discourse setting, for instance, 'language in action' contexts where interlocutors are doing something together at the moment (such as when mutually assembling a piece of furniture) typically produce spoken discourse with a substantial number of deictic items (*that, there, here, it*) and discourse markers (*I see, okay*) which mark the stages of the process they are trying to complete (Carter & McCarthy 1997).

Because context affects spoken discourse in these ways, one might expect that the diversity of the lexis contained in the discourse would also be affected by different speech contexts. This study explores this issue, by analysing five corpora which differ according to interlocutor relationship and purpose of discourse. It will explore whether any difference in percentage of lexical coverage can be found between the corpora, and if so, what size of vocabulary is necessary to reach viable levels of coverage in each of the context types.

2. Spoken discourse contexts and the CANCODE corpus

It is now accepted that the context in which discourse takes place has considerable influence over that discourse. Halliday & Hasan (1985) suggested that the *field* (the environment of the discourse), *tenor* (who is taking part in the discourse and their relationship to one another), and *mode* (the role the language plays in the context) all shape the nature of the discourse, and this work has been taken forward by scholars working in areas such as corpus linguistics (e.g. Biber *et al.* 1999) and genre theory (e.g. Martin & Rothery 1986; Hammond & Deriwianka 2001). However, there is a relatively small amount of research which explores how spoken discourse differs according to context, compared with written discourse. One consequence of this is that the aspect of vocabulary coverage in different spoken contexts has received little attention. This may be a result of the lack of spoken corpora or the lack of corpora that are suitably categorised.

The categorisation applied to conversations in the CANCODE corpus can be used to examine vocabulary coverage in different contexts.[1] The careful categorisation and annotation of this corpus makes it a valuable resource for such comparisons despite its relatively small size compared to other modern corpora, such as the British National Corpus and the Bank of English for example. The main phase of data collection took place between 1994 and 1999 with a focus on gathering conversations from a variety of discourse contexts and speech genres. In order to ensure a wide demographic and socio-economic representation, conversations were carefully selected to include adult speakers of different ages, sex, social backgrounds and levels of education.

Traditional divisions between formal and informal have been used as general guidelines for achieving diversity in the corpus. The framework adapted for the CANCODE corpus distinguishes between five different context-types. Our research is based on these carefully established categories the validity of which has been illustrated in previous studies (see McCarthy 1998). Other ways of categorising the CANCODE corpus, such as by topic for example, are possible but would require a complete re-organisation of the text files. However, it is unlikely that such a re-organisation would be sensible if we consider the vast diversity of topics and topic shifts that may occur in any one conversation.

In the current model the context-type axis of categorization reflects the relationship that holds between the participants in the dyadic and multi-party conversations in the corpus. These types of relationships fall into five broad categories which were identified at the outset: *Intimate, Socio-cultural, Professional, Transactional* and *Pedagogic*. These categories were found to be largely exclusive while being comprehensive at the same time. They are described in turn below.

Intimate: In this category, the distance between the speakers is at a minimum and is often related to co-habitation. Only conversations between partners or close family qualify for this category in which participants are linguistically most 'off-guard'. At times the *Intimate* category is difficult to distinguish from the *Socio-cultural* one. Alongside the criteria agreed by the corpus compilers as to what shall qualify for which category, it was decided to let participants judge which category they felt they belonged to. All participants in a conversation have to fall under this category for the conversation to be classified as *Intimate*.

Socio-Cultural: This category implies the voluntary interaction between speakers who seek each other's company for the sake of the interaction itself. Most of the texts that did not fall into any of the other categories turned out to be *Socio-cultural*. The relationship between the speakers is usually marked by friendship and is thus not as close as that between speakers in the *Intimate* category. Typical venues for this type of interaction are social gatherings, birthday parties, sports clubs, and voluntary group meetings.

Professional: This category refers to the relationship that holds between people who are interacting as part of their regular daily work. As such, this category only applies to interactions where all speakers are part of the professional context. Thus a conversation between two shop assistants would be classed as *Professional*, while the interaction between a shop-assistant and a customer would be classed as Transactional. Talk that is not work related but occurs between colleagues in the work-place has still been classified as *Professional*, based on the observation that the participants retain their professional relationship even when the topic of the conversation is not work related.

Transactional: This category embraces interactions in which the speakers do not previously know one another. The 'reason' for transactional conversations is usually related to a need on the part of the hearer or the speaker. As such, the conversations aim to satisfy a particular transactional goal. This category has traditionally been referred to as 'goods-and-services' (Ventola 1987), having the exchange of goods as the main aim of the interaction.

Pedagogic: This fifth category was set up to include any conversation in which the relationship between the speakers was defined by a pedagogic context. A range of tutorials, seminars and lectures were included. As the emphasis was on the speaker relationship rather than the setting, conversations between lecturers as well as academic staff meetings were classified as *Professional* rather than *Pedagogic*. At the same time, training sessions in companies were classified as *Pedagogic* rather than *Professional*. Perhaps a better label for this category would have been *Academic* or *Training* language, since the type of interaction recorded under this category included a large proportion of subject specific lectures and seminars. In addition, the language was entirely authentic L1 academic discourse; there was no simplified ESL pedagogic material included. It is also likely that *Pedagogic* is the category that comes closest to including scripted and technical language.

From the classification above we can see a 'cline' of distance emerging between the speakers in four of the categories (*Intimate, Socio-cultural, Professional, Transactional*) which allows for a corpus-based analysis of linguistic choice in those contexts, with the *Intimate* category being the most private, and the *Transactional* the most public. If percentage of lexical coverage varies according to the distance between speakers, then a comparison between the four corpora should demonstrate this. Although the *Pedagogic* category does not fit into the scale of public versus private, it provides an example of a different type of discourse context, and so will be analysed as well.

3. Methodology

The procedure of the current study is different to that used by Schonell *et al.* (1956) and Adolphs & Schmitt (2003) in that it considers individual word forms rather than word families in the calculation of vocabulary coverage. While there are good pedagogic reasons to analyse vocabulary in terms of word families, such as the observation that learners seem to mentally handle the members of a word family as a group (Nagy *et al.* 1989), unfortunately it is still impossible to program a computer to identify word families automatically. Current software which counts word families, such as Nation's RANGE program (Internet resource), do so by referring to baseline lists of word family members which have already been compiled. The only way to identify the members of a word family reliably for such baseline lists and other purposes is to do it manually. Both Schonell *et al.* and Adolphs & Schmitt used this time-consuming method. On the other hand, concordancers can quickly and automatically count individual word forms. Because the purpose of the present study (comparing the degree of vocabulary coverage across various spoken genres) can be achieved just as well using individual word forms rather than word families as the unit of measurement, we decided to use the more computer-automated approach in order to avoid possible errors in manual tabulation.

The first step in the research involved creating a frequency list of the words in the categories of CANCODE outlined above. The CANCODE is not lemmatised or coded for word class, therefore the word lists generated were based on individual word forms. Any corpus specific codes or annotation markers were deleted from the list. Backchannel verbalisations which do not normally qualify as words, such as *eh, uh uh, mmm, and Oh!,* were included in the count since these items convey a great deal of meaning and are an important feature of spoken discourse (see Biber *et al.* 1999). Once the lists of words and their frequency of occurrence were set in our spreadsheet, we simply divided various frequency levels by the total number of words in the category to arrive at a percentage of text coverage. For example, to derive the coverage figure for the most frequent 2,000 words in the *Transactional* category, we divided the total number of tokens for each of the 2,000 word forms by the total number of tokens in the transactional sub-corpus. The resulting figure was multiplied by 100 to arrive at a percentage of coverage for each form. These numbers were added up for the first 2,000, 3,000, 4,000 and 5,000 word forms respectively.

Whereas the five different categories in the CANCODE were made up of varying numbers of running words (smallest = *Pedagogic* with 456,177 tokens; largest = *Socio-cultural* with 1,709,598 tokens), it was important to ensure that any differences in lexical coverage were not merely an artefact of the different sizes of the categories.

In order to ascertain whether the differences in overall word count within the various corpus categories would effect the degree of coverage, we carried out an analysis that set out to test the relationship between overall number of words in a corpus and vocabulary coverage provided by the first 2,000, 4,000 and 5,000 word forms. For this analysis we used the transactional category which is one of the larger corpus categories with an overall word count of 1,166,825 words. We extracted a set of files with varying word counts from the transactional corpus to form a sub-sample of 434,128 words, which is similar in size to the smallest category — *Pedagogic.* We then carried out a procedure to determine vocabulary coverage as outlined above for the sub-sample and the full transactional sub-corpus. We found that there were small differences in vocabulary coverage based on the size of the corpus (see Table 1), and this fact will have to be considered in the analysis of the study.

Table 1. Differences in lexical coverage of a sub-sample and full version of the transactional sub-corpus

	2,000 word forms % of coverage	4,000 word forms % of coverage	5,000 word forms % of coverage
Transactional sub-sample (434,128 tokens)	94.39	97.45	98.20
Transactional full sub-corpus (1,166,825 tokens)	94.30	97.14	97.82

4. Results and discussion

The results of our analysis summarised in Table 2 show noticeable differences in vocabulary coverage according to spoken discourse context. The differences between the category with the highest coverage (*Transactional*) and the least coverage (*Pedagogic*) ranges from between approximately 1.7 percentage points at the 5,000 word level to almost 4 percentage points at the 2,000 word level. While the percentage differences between categories do not seem large in simple terms, (and are not statistically significant in terms of a Chi-squared analysis: χ^2, p>.05), they become very substantial when translated into the number of word forms involved. Let us take the 2,000 word level where the difference is greatest as an example. The difference is 3.94 percentage points (*Transactional* 94.30% — *Pedagogic* 90.36% = 3.94%). We then counted the number of additional word forms required to raise the coverage figure in the

Pedagogic category from 90.36% to 94.30%. We found that it took 1,608 word forms. Thus, with 2,000 word forms you can achieve 94.30% lexical coverage in the *Transactional* category, but to achieve the same percentage of lexical coverage in the *Pedagogic* category, you would need 3,608 word forms. At the 5,000 level, even though the difference in percentage of coverage is smaller, it actually takes more word forms to make up the difference due to the effects of decreasing frequency. To raise the *Pedagogic* coverage figure from 96.11 to 97.82 (equivalent to the *Transactional* figure at the 5,000 level) would require an additional 2,307word forms, or a total of 7,307 forms. Overall, the various spoken contexts have noteworthy differences in terms of lexical coverage and number of word forms required.

Using the CANCODE classification system which groups texts according to the relationship that holds between the speakers, the results seem to suggest a 'cline' in the degree of vocabulary coverage which is generally at its lowest in the more private/interactional spheres and increases towards the more public/transactional spheres. The cline is not completely consistent across the categories however. In fact, the lexical coverage figures for the *Intimate* and *Socio-cultural* categories are quite similar, with the *Socio-cultural* figures being the lowest at all frequency levels. Thus, in terms of the diversity of vocabulary required, there does not seem to be much difference between truly *Intimate* interlocutors and those who are merely friends. A greater difference occurs when the categories move to the more goal-oriented discourse of *Professional* and *Transactional* encounters. In these categories, the lexical coverage figures are notably higher, indicating that a narrower range of vocabulary is required to engage in transactional and professional interaction than in more casual conversation. We could speculate that the reason for this result is to be found in the wide range of topics discussed in the more private situations as opposed to the more transactional ones which tend to have more focused topics and follow more predictable patterns of language use.

Table 2. Percentage of lexical coverage of five speech genre categories

Category (Total tokens in sub-corpus)	2,000 word forms % of coverage	3,000 word forms % of coverage	4,000 word forms % of coverage	5,000 word forms % of coverage
Pedagogic (456,177)	90.36	93.15	94.90	96.11
Intimate (957,192)	92.81	94.70	95.84	96.63
Socio-cultural (1,709,598)	92.43	94.34	95.51	96.31
Professional (480,627)	93.28	95.28	96.51	97.35
Transactional (1,166,825)	94.30	96.10	97.14	97.82

It is interesting to note in this context that the *Pedagogic* category, which does not fit into the original classification scheme of private versus public discourse, displays the lowest degree of vocabulary coverage. The defining feature of this category is the academic/training nature of the discourse context, and so it should contain a relatively high percentage of a more formal, academic type of discourse. Thus the lower percentage of coverage in this category provides evidence for what teachers have always known: that learners need a wider vocabulary to cope with academic or training discourse than to cope with everyday conversation. The figures also argue for the inclusion of a significant vocabulary component in English for Academic Purposes courses, in order to help learners deal with the more diverse vocabulary found in this type of discourse.

In the Methodology section, we explored whether the size of the sub-corpora would affect the lexical coverage percentages. We compared a sub-sample and the full version of the *Transactional* sub-corpus and found that sub-corpus size made only a small difference in lexical coverage. The magnitude of difference in lexical coverage percentage between the context categories in Table 2 are clearly far greater than that found due to corpus size in Table 1, which suggests that any differences in lexical coverage found in this study should mainly be attributable to contextual differences rather than to the different sizes of the CANCODE categories. An examination of Table 2 also reveals no obvious relationship between corpus size and the magnitude of lexical coverage. This supports the case that corpus size, at least with the size of corpora under discussion, does not affect lexical coverage to any great degree. The trend that does emerge is that the rank order of the context categories is the same at each frequency level (in the order: *Transactional* > *Professional* > *Intimate* > *Socio-cultural* > *Pedagogic*), indicating that the influence of context is consistent across the frequency bands. It is useful to note however, that our analysis between corpora of different sizes was based only on individual word forms, and a similar comparison based on word families remains to be carried out. In sum, although corpus size probably has a small influence, the differences in lexical coverage in the table appear to be a result primarily of context category.

5. Conclusion

Just as the spoken discourse context affects speech in terms of frequency and function of discourse markers (Stenström, 1990), the frequency of deictic items (McCarthy, 1998), and the propensity towards various grammatical features (Carter & McCarthy, 1995), this study shows that the spoken context also has an effect on the diversity of words typically used. Spoken discourse among intimates or friends typically contains a greater range of vocabulary than spoken discourse which is used for more transactional roles. Taken together with Adolphs & Schmitt (2003), the two CANCODE-based studies indicate that operating in a spoken English environment requires more vocabulary than previously thought, and the amount required depends on the spoken context.

Note

[1] CANCODE stands for Cambridge and Nottingham Corpus of Discourse in English and is a joint project between Cambridge University Press and the University of Nottingham. The corpus was funded by Cambridge University Press with whom sole copyright resides. For a comprehensive description of the corpus, see McCarthy 1998.

References

Adolphs, S. and Schmitt, N. 2003. "Lexical coverage of spoken discourse". *Applied Linguistics 24*, 4: 425–438.

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

Carter, R. A. and McCarthy, M. J. 1995. "Grammar and the spoken language". *Applied Linguistics 16*, 2: 141–158.

Carter, R. and McCarthy, M. 1997. *Exploring Spoken English.* Cambridge University Press.

Halliday, M. A. K. and Hasan, R. 1985. *Language , Context, and Text: Aspects of Language in a Socio-semiotic Perspective.* Oxford: Oxford University Press.

Hammond, J. and Deriwianka, B. 2001. "Genre". In *The Cambridge Guide to Teaching English to Speakers of Other Languages,* R. Carter and D. Nunan (eds), 186–193. Cambridge: Cambridge University Press.

Martin, J. R. and Rothery, J. 1986. "Writing Project Report No. 4". *Working Papers in Linguistics,* Linguistics Department: University of Sydney.

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

Nagy , W., Anderson, R. C., Schommer, M., Scott, J. A., and Stallman, A. C. 1989. "Morphological families in the internal lexicon". *Reading Research Quarterly* 24: 262–282.

Nation, P. RANGE vocabulary analysis program. Available free of charge at <http://www.vuw.ac.nz/lals/>.

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

Stenström, A.-B. 1990. "Lexical items peculiar to spoken discourse". In *The London-Lund Corpus of Spoken English: Description and Research,* J. Svartvik (ed), 137–175. Sweden: Lund University Press.

Ventola, E. 1987. *The Structure of Social Interaction.* Pinter: London.