Rosamund MOON, Cobuild and the University of Birmingham

On using spoken data in corpus lexicography[1]

Abstract

Corpora are increasingly used in lexicography in order to provide good evidence for dictionary statements: the inclusion of spoken data in corpora is generally considered important. This paper raises some issues connected with the use of spoken data. It points out that the extensive differences between written and spoken language have great consequences for dictionary-making. It argues that the repercussions have not yet been fully thought through, and suggests that new models for the lexicographical description of spoken language may have to be developed.

Keywords: Corpora, dictionary-making, spoken data, English, meaning

The purpose of this paper is to set out some issues concerning the use of data drawn from spoken interaction in lexicographical work, with particular reference to English and to monolingual dictionaries. The last two decades have seen rapid expansions both in the size of computer-held corpora of language (see, for example, Leech 1991 for a discussion of this), and in the extent to which they have become used in lexicography. The exact usage praxis, of course, varies from institution to institution and from project to project: some projects are entirely or virtually corpus-driven, and others use corpora mainly to augment or complement extensive, more focussed, collections of citations. The crucial point here is that corpora **are** used, and are now widely accepted as valuable, arguably essential, resources in serious linguistic description of any kind: a wealth of papers in IJL and in EURALEX proceedings shows the extent to which linguists, lexicographers, and lexicologists have taken advantage of them. But this does not mean that corpora can be used uncritically or automatically. There are undoubtedly problems as well as benefits.

The principal benefit for lexicography is that corpora provide convenient evidence of the formal usage of lexical items: the associated syntactic structures, phraseological patterns, and collocations; contexts of use; frequencies and distributions in terms of variety, genre, and register; and, where diachronic corpora are available, evidence of changes in currency and usage. The principal benefit for theory is that corpus evidence can be used to prove, test, and adjust models: for example, with respect to semantics, corpora can be used to examine indeterminacy or discreteness of meanings in polysemous items, and to assess the extent to which real ambiguity truly exists.

Problems with corpus use include practical ones: there is too much data for some items, too little for others. Besides, corpus research can only be as effective and robust as retrieval tools or statistical and other pre-analysis tools allow. For example, if a part-of-speech tagger is not very near 100% accurate, we cannot make accurate statements about word class incidence and parsing. We cannot make accurate statements about collocational statistics or even the frequency of a single form, let alone lemma, unless we know that the corpus is 100% 'correct', without misprints, misspellings, or typos (though misspellings are arguably parts of the lexicon in their own right). Type-token counts of the same corpus can vary according to whether each hyphenated word form is counted as a single compound token or as two tokens.

Word frequency counts are affected by the status given to spelling variants: although *colour* and *color* are different forms or types, they are the same 'word'. All of this has a bearing on the objective 'authoritativeness' which corpora are believed to bestow. Moreover, while corpus-based and corpus-driven analyses are quasi-scientific, they must inevitably involve and be tempered by intuition. Once the formal 'facts' about an item have been set down, the analyst has to begin using intuition and inference: *'this is a separate meaning, this isn't'*, *'this is the register, or the relationship between the speakers'*, or *'this is the intended pragmatic effect'*. Lexicography cannot be a mechanistic, automated, purely left-brain process; it needs to be creative and intuitive, a right-brain process too. The ascertaining of the corpus subtext and context of situation becomes an activity and skill like the reading of any literary or artistic text, a matter of judgement and perception, not just objective observation of the 'facts'. There is what might be termed a possible problem of presumption here — analysts being presumptuous enough to guess at speakers' intentions and likely implicatures and contexts, and not necessarily guessing right.

If we look at the constitutions of major English language corpora, and we leave aside deliberately genre-focussed corpora, for example of business English or academic writing, we can see that the dominant trends are towards variety and balance. The early Brown and Lancaster-Oslo-Bergen corpora comprise balanced swathes of written data; similarly with the later International Corpus of English and the British National Corpus, which both include spoken interaction as well as written data. The Bank of English (Cobuild/University of Birmingham) is comparatively unbalanced and indiscriminate, but at various stages of its development, it has included up to 20% spoken data, both scripted and unscripted. Written data is easier to collect than spoken, but nobody involved with developing general corpora these days would want to exclude spoken data, in spite of the huge practical, technical, financial, and ethical problems associated with the acquisition of spoken data. Its linguistic importance and value is attested quantitatively by its inclusion in such general corpora, and qualitatively by some pioneering studies into its lexicology and grammar: for example, by Altenberg and Eeg-Olofsson (1990), Aijmer (1996), and Carter and McCarthy (1995). Nobody seriously believes that corpora reflect normal or average language experience — it would be impossible, anyway, to determine what 'normal' or 'average' means — but it is likely that normal or average language experience would include proportionately far more exposure to and use of spoken language than the meagre representation in the British National Corpus and The Bank of English.

Phonological Problems

Various aspects of spoken data need to be dealt with when using it: features of conversational performance such as false starts and hesitation, or backchannel, and unique grammatical structures (see Carter and McCarthy 1995). However, something that cannot be ignored is that studying spoken data in corpora is actually a strange thing to do, since it is studying transcriptions, spoken-into-written form. The very transcription process will have necessitated interference, for example the imposition of punctuation or pauses; the loss of pacing or speed, which is meaningful in its right; the loss (in some corpora) of information about intonation; the conversion of strings of phonemes into conventionalized graphological forms, that is, spelled in relatively conventional ways. Note that while non-standard usages of words tend to be preserved in transcription, some common pronunciation forms do not. For example, the phrase *be going to*, expressing intention or probability (see Palmer 1990: 142 ff.), is often

reduced in pronunciation, and this is conventionally represented graphologically as *(be) goin' to* or *(be) gonna*. The Bank of English shows frequencies for *gonna* of 100 per million words of conversation, and 10 per million in fiction; *goin' to* has a frequency of just under 4 per million in fiction, but never occurs in the subcorpus of transcribed conversation. This might be chance, but it might also reflect processes of normalization and convention. This can also be seen in some other cases of common reductions. The auxiliary *have* is often reduced to a clitic, or loses vowel quality. The conventionalized written representation is *'ve*, only rarely *of*, although this is a natural and common 'misspelling' found in children's writing: again, the latter is commoner in Bank of English fiction than in its transcribed conversation. So too with forms such as *kinda* and *sorta*, which show up in fiction or written journalism, but only rarely in transcribed speech, suggesting again that the transcriptions may not reflect precisely the original pronunciations. This might seem to be a problem of a particular transcription system or even individual transcriber, but it recurs in different corpora and different systems. It is perhaps more likely to reflect the fact that in order to transcribe meaningfully, some normalization must be imposed.

Phonology leads into another point. Pawley and Syder (1983) and Nattinger and DeCarrico (1992) argue the case forcefully for 'lexical phrases': semifixed lexicalizations which are used to structure discourse or to prepackage responses, evaluations, and so on, for example *Come to think of it, It doesn't bear thinking about, I thought you knew, What I'd like to do today is..., I think the most interesting one is....* It is accepted that there is a very large number of these, yet they are difficult to identify or inventorize. In fact, this may only be possible if their phonological realizations are considered, rather than their graphological forms or syntactic structures: it seems more likely that they could be identified by looking at the tone unit structures of utterances (see Brazil 1985): then lexical phrases can be distinguished as recurrent realizations forming complete tone units, rather than by being syntactic units or semantically/pragmatically non-compositional strings. From the dictionary point of view, the traditional inventory of the lexical units of a language may not be wholly appropriate for spoken language. Should dictionaries even attempt to be repositories of spoken discourse?

Facts and Figures

Corpus data shows that there is no simple and clear distinction between written and spoken modes: neither mode is monolithic, and each comprises many genres and subgenres with their own structural patterns and conventions. There does, however, appear to be a discontinuity between conversation and other spoken modes, such as monologues, lectures, screenplays, semiscripted broadcast journalism, and so on. This can be demonstrated by considering just the ten most frequent forms in a series of subcorpora of British English in The Bank of English, as in Table 1. These subcorpora are ordered crudely according to formality, and they consist of data drawn from respectively the semitechnical journal The *New Scientist* (6 million words); the broadsheet newspaper The Times (21 million); a collection of fictional and non-fictional, non-journalistic prose works (42 million); the now defunct tabloid newspaper Today (27 million); transcriptions of semiscripted broadcasts from the BBC World Service (19 million); and transcriptions of unscripted conversation (20 million).

WHOLE CORPUS	NEW SCIENTIST	TIMES	(NON)FICTION	TODAY	BBC	CONVERSATION
323m	6m	21m	42m	27m	19m	20m
the 17845265	the 413923	the 1239015	the 2437460	the 1245943	the 1361436	the 622430
of 8321813	of 208346	of 582453	of 1209683	to 650599	of 588161	i 604954
to 8034738	to 155617	to 526890	and 1145056	a 637893	to 501131	and 486960
and 7582198	a 134613	a 491773	to 1110648	and 540729	in 447313	you 477917
a 7061412	and 131837	and 458213	a 908778	of 525255	and 378166	it 406675
in 5854481	in 122268	in 391069	in 770569	in 433579	a 367870	to 399270
that 3323182	is 78159	is 217689	was 484938	's 269715	that 218720	that 365205
's 3119315	that 76751	that 198810	that 458502	for 248426	's 190076	a 350211
is 2970503	for 52682	for 194677	i 446049	is 240819	is 165570	's 338557
it 2957913	it 47014	's 179779	it 443300	was 223724	for 159051	of 305183

Table 1

There is a lot of overlap between the first five subcorpora: much less between them and the conversation subcorpus. Compare Nelson (1997), who draws attention to contrasts between different subgenres of written and spoken English: for example, *the* and *I* have very different relative frequencies in debates and in phonecalls. The Bank of English also demonstrates the relative paucity of the lexicon of conversation. Table 2 shows the number of types occurring in each subcorpus, the number of types occurring more than once (non-hapaxes), and the number of types occurring at least once per million words of the particular subcorpus.

	NEW SCIENTIST	TIMES	(NON)FICTION	TODAY	BBC	CONVERSATION
	6m	21m	42m	27m	19m	20m
types	84266	154305	213938	137058	114351	67233
1+	55489	100227	129086	87947	69780	45580
1/million+	31220	33347	31436	28636	21813	14033

Table 2

The more informal genres generally have fewer lexical types, but conversation has dramatically fewer.

The detail of the data

Once we start looking at the detail of the data, contrasts between conversation and other modes are striking. This can be demonstrated by looking at the evidence for three words, *well*, *appreciate*, and *strange*, in the subcorpus of conversation and comparing them with the evidence in the subcorpus of The *Times*. It is not in the least surprising that the uses of *well* showing up in these subcorpora are very different. The main uses in The *Times* are of the adverb 'very successfully or satisfactorily; skilfully; to a great extent' and of the phrasal preposition *as well as*:

But his grassroots support has held up well in his native Lombardy, where opinion polls show he could win.

Pugh knows rugby needs to be improved as a spectacle, but will be well aware of the concerns of rugby diehards, who fear the game will be tampered with to suit 'outsiders'.

Jenkins, who is non-executive director of such organisations as Rank, Thorn EMI and Gartmore, as well as being the chairman of Hambros Falcon property trust, should not get too comfortable, however. The bank is planning to move the occupants of its Brentwood office down to Tower Hill.

Main uses in conversation include *well* as a discourse marker, and in the adverbial *as well* (without the following *as*):

Well I'm out Wednesday all day 'cos we're going to er Symonds Yat and Ross-on-Wye from here.

Yeah. I've got well I've got everything and I've got cucumbers bananas... erm cakes.

He's the transport manager and he organizes the organizes the buses as well.

The lines for *appreciate* from The *Times* could mostly be glossed as 'understand (the necessity for or rationale behind)' or 'recognize the value of':

Through having taught the British to appreciate better food, cooked with skill, but simply and without pretension and to appreciate the food of other cultures...

This kind of meaning can be seen amongst the lines taken from conversation, but there is also evidence of the thanking formula *I appreciate it*, and the use of *appreciate* in acknowledging politely another speaker's opinion, while signalling one's own disagreement:

That's great. Thanks a lot for your help anyway... I do appreciate it.

I mean I mean I appreciate you you need the profits to be able to invest but I still think a hundred and five pounds a second is an awful lot of money in anybody's book.

It might be assumed that the usage of *strange* would not differ very widely in the corpora; yet in The *Times* subcorpus, *strange* is often followed by fully lexical nouns, whereas in the conversation subcorpus there is greater evidence of its being followed by 'general' or superordinate nouns. Semantically and pragmatically, there are slight differences: in written English, *strange* seems to be used to indicate 'otherness' as a trait:

An Edwardian country church seems a strange choice of environment.

Tiny black fish with strange curative powers have become the star attraction of a remote Turkish region.

In spoken English, *strange* seems to be used in more overtly evaluative structures, to acknowledge or negotiate 'otherness' interactively:

A: *You know who she is?*
B: *Yeah. Yeah. She's one of the secretaries of the...*
A: *Yes.*
B: *Yeah.*

A: *That's a little bit strange.*

It's strange really that a place where so many people live has so few shops, isn't it?

Other distinctions include the tendencies of *strange* to be premodified in conversation, or preceded by epistemic *seem* or *find it*, whereas in written data, *seem* rather than *find* precedes, and there is also a lot of evidence of a pattern *strange and ADJECTIVE*. Such observations point up the fact that there are different phraseologies, different meaning distributions, and different pragmatic subtexts in conversational data. We can argue that conversation has a restricted range of vocabulary choices and an enhanced range of pragmatic devices.

Conversation is lexically specialized, and in this way it can be compared to other genres such as technical writing or literary writing, where particular items have deviant frequencies or deviant forms, in comparison with general English. Ironically, it is not only the extreme features of spoken language which are problematic for lexicographical description but also the lacunae: the words and senses in the lexicon which are apparently not used in ordinary conversation, and ellipsis and extralinguistic reference. Moreover, it is generally believed that many 'colloquialisms', slang terms, and more 'informal' items such as phrasal verbs and idioms occur more commonly in spoken interaction than in written English, and therefore cannot be expected to occur frequently in predominantly written corpora. Yet in many cases, conversational data does not support this belief: see Moon (1998: 66-8, 72-4) for further discussion of the frequencies of idioms and proverbs in spoken language.

Dictionaries

In the final part of this paper, I want to consider what dictionaries do with spoken data, and then to speculate about what they could do. Traditionally, dictionaries are considered guardians of prescribed usage, in spite of any declared intentions on their part to be descriptive. The deviant or marked is acknowledged as deviant or marked, for example with indications of restricted usage or register: 'colloquial', 'informal', 'non-standard', and so on, just as some usages are indicated as 'journalistic' or 'literary'. No real attempt, however, has been made to indicate that words are not used in particular genres, although pedagogical EFL dictionaries go further than native-speaker dictionaries in this respect. The *Collins Cobuild English Dictionary* indicates the orality of the phrase *if you like* in its definition metalanguage:

"You say **if you like** when you are making or agreeing to an offer or suggestion in a casual way. *You can stay here if you like... 'Shall we stop talking about her?'—'If you like.'* "

Yet it is not always possible to see such metalanguage as indicating a restriction to spoken interaction. For example, in

"You say **if you like** when you are expressing something in a different way, or in a way that you think some people might disagree with or find strange. *This is more like a downpayment, or a deposit, if you like.*"

'say' implies something about pragmatic intention, rather than the mode in which the item is used. The third edition of the *Longman Dictionary of Contemporary English* goes a long way

towards describing some of the striking differences between spoken and written language. It marks the 3,000 commonest words in speech and writing, and shows through graphics the relative frequencies of certain structures and other usages in speech and writing. The dictionary also contains a wealth of phraseologies associated with headwords, and the ways in which they are used in conversational English: see Summers (forthcoming) for further discussion of this radical approach to treating spoken language in dictionaries.

There is a problem for lexicographers who want to exemplify dictionary entries with authentic chunks of conversational data: much of it is so heavily context-bound, with intonation of significance that is almost impossible to reproduce meaningfully. Dictionaries appear to privilege written data in selecting examples, but it is actually because conversational data rarely provides autonomous snippets (to use sanitized fictional dialogue would of course be to falsify natural conversational patterns), as can be seen in the following:

A: *So how would you describe the running of it now I mean back when You came here it was very victorian. Is it*
B: *Yeah*
A: *more like What is it*
B: *Oh no ... it's er it's it's it's run like ... a a proper office now....*
A: *... O - of what we'd describe as a perhaps an industrial concern is it?*
B: *Yes yes yes. I mean they're trying to make it ... a a money-making concern aren't they.*
A: *Well they are yes*
B: *I mean one time ... i - i - it was learning for the sake of learning wasn't it.*
A: *Yes.*
B: *Now it's learning with an object. [laughs]*
A: *It is right. I think that's a very good contrast. [pause] And you've got to make money.*
B: *Well yes you see the*
A: *Or at least bring money in from outside.*

This is by no means an atypical stretch, yet little of it could be selected as exemplification material, at least not without extensive and distorting editing.

Spoken corpus data is wonderfully rich and exciting linguistically, but using it for a dictionary, especially a pedagogical dictionary is another matter altogether. The question must be asked: what use is it in a pedagogical dictionary? how would a user use it? this is encoding information rather than decoding. Yet nobody encodes real time conversation with a dictionary — nor even with specialized phrasebooks, not with any success. Furthermore, the range of conversational implicatures in such formulae as *if you like* or *XX would like* can scarcely be described in a linguistic monograph, let alone within the constrained space and vocabulary of explication in a dictionary.

I am suggesting that the constraints of the conventional dictionary, with its emphasis on the individual lexical item and objective meaning, make it difficult if not impossible to deal with the distinguishing features of spoken language properly and fully. This does of course tilt the conventional dictionary towards the written and normative (which is not a problem for users, who tend to want information about norms). I could argue, though without complete conviction, that it probably does not in fact matter a great deal if spoken data is ignored when writing about many items, though of course a distorted picture would be given if entries for discourse markers and so on were based on fictional dialogue or other forms of written

English. (It might be thought that *alas* was really an exclamation in everyday life or that *hold your horses* was really a recommendation to delay, when in fact spoken data shows that *alas* is only ever used in speech as a disjunct or sentence adverb, meaning 'unfortunately', and that *hold your horses* is only found in novels, screenplays, and so on.)

I am also suggesting, however, that if it is considered desirable for spoken English to be described in a lexicon, as part of the detailed description and inventory of the language, then that lexicon must be structured very differently. The units of lexical description may need to be defined phonologically rather than graphologically; units of lexical description must involve and include their phraseological patternings, collocations, and colligations; syntax at anything higher than part of speech may need to observe new models, and even at part-of-speech level the traditional handful of labels may need to be expanded; meaning may need to be described in terms of pragmatics rather than semantic components (accepting that concrete terms and terminology do not differ so much in usage between speech and writing); and register must be much more carefully described because of the intrinsic importance of hierarchies and power relationships in conversation.

Spoken data cannot be seen as simply added value in a corpus — it is intractable in its present format — but it could be seen as giving an opportunity for the development of an innovative lexicon, a fascinating teaching and research resource. It would offer the opportunity to identify, establish, and then dissect the norms of standard spoken English: note that corpus linguists have barely begun the work of looking at dialects and slang or antilanguage. It remains to be seen how findings and insights would eventually feed back into conventional dictionaries of normative written English, and what the implications are for bilingual dictionaries.[2]

Notes

[1] This paper draws on my contributions to a paper presented jointly with Rosalind Combley at a BAAL symposium in September 1997. I am indebted to Rosalind Combley for her comments. I would also like to acknowledge Henri Béjoint, for comments made on the first draft of this paper.

[2] Corpus data here is drawn from The Bank of English corpus, created by Cobuild at Birmingham University.

References

Collins Cobuild English Dictionary (1995, 2nd edition). HarperCollins, London and Glasgow.

Longman Dictionary of Contemporary English (1995, 3rd edition). Longman, London.

Aijmer, K. (1996) *Conversational Phrases in Spoken English*. Longman, London.

Altenberg, B. and Eeg-Olofsson, M. (1990) 'Phraseology in spoken English: presentation of a project' in J. Aarts, W. Meijs (eds.) *Theory and Practice in Corpus Linguistics*, pp. 1—26. Rodopi, Amsterdam.

Brazil, D. (1985) *The Communicative Value of Intonation in English* (English Language Research Monographs). University of Birmingham, Birmingham.

Carter, R. and McCarthy, M.J. (1995) 'Grammar and the spoken language', *Applied Linguistics*, Vol. 16/2, pp. 141—158.

Leech, G.N. (1991) 'The state of the art in corpus linguistics' in K. Aijmer, B. Altenberg (eds.) *English Corpus Linguistics*, pp. 8—29. Longman, London.

Moon, R. (1998) *Fixed Expressions and Idioms in English: A Corpus-based Approach* (Oxford Studies in Lexicography and Lexicology). Oxford University Press, Oxford.

Nattinger, J.R. and DeCarrico, J.S. (1992) *Lexical Phrases and Language Teaching*. Oxford University Press, Oxford.

Nelson, G. (1997) 'A study of the top 100 wordforms in ICE-GB text categories', *International Journal of Lexicography*, Vol. 10/2, pp. 112—134.

Palmer, F.R. (1990) *Modality and the English Modals* (2nd edition). Longman, London.

Pawley, A. and Syder, F.H. (1983) 'Two puzzles for linguistic theory; nativelike selection and nativelike fluency' in J.C. Richards and R.W. Schmidt (eds.) *Language and Communication*, pp. 190—225. Longman, London.

Summers, D. (forthcoming) 'Real language, taught language, and spoken language in relation to the Longman Dictionary of Contemporary English' in Th. Herbst, K. Popp (eds.) *The Perfect Learner's Dictionary*. Niemeyer, Tübingen.