This that and the other: Multi-word clusters in spoken English as visible patterns of interaction

Michael McCarthy
University of Nottingham and University of Limerick

Ronald Carter
University of Nottingham

Abstract

This paper investigates multi-word strings automatically retrieved from a 5-million word corpus of conversational English from Britain and Ireland. Many such strings have neither syntactic nor semantic integrity, for example *at the, it was a, what do you*. However, many strings display pragmatic integrity, encoding interactive functions such as hedging, vagueness, discourse marking, etc. Examples include *and that sort of thing, you know, a couple of*. We identify the most common pragmatically integrated clusters and discuss their functions, and compare their frequency with single words, illustrating that many clusters are more frequent than single words accepted as belonging to the core vocabulary of English. The clusters also contrast with the low frequency of opaque idiomatic expressions. High-frequency clusters raise issues around the distinction between lexis and grammar, and support a synthetic view of language production and storage, with implications for the understanding of notions such as fluency and idiomaticity.

Introduction

The single word

In the study of the lexicon, the single word has remained, until recently, relatively unchallenged as the basic unit of meaning and as the focus in the study of lexical acquisition in second and foreign languages. This is not without good reason: single words form a substantial part of the lexicon of English and are perceived in pedagogy as the central units to be acquired. Other units consisting of more than one word, such as phrasal verbs, compounds, and idioms,

are often thought of as items belonging to higher levels of achievement. There are, of course, exceptions to this: greetings and other phatic expressions (e.g. *How's it going?*, *See you soon*, *Thanks a lot*), specialized functional phrases (e.g. *Happy birthday*, *Good luck*), basic prepositional phrases (e.g. *in the morning*, *at home*), and common compounds (e.g. *car park*, *check-in*) are often taught and/or acquired even at elementary level.

Collocation

Recent developments in the study of lexis have generated new applications within lexicography and language teaching, offering the possibility of a better understanding of the nature of the lexicon, especially multi-word phenomena. The most important of these developments can be seen in the Neo-Firthian approach to word meaning. Firth (1935) famously proposed that the meaning of a word was as much a matter of how the word combined in context with other words (i.e. its collocations) as any inherent properties of meaning it possessed of itself: *dark* is part of the meaning of *night*, and vice-versa, through their high probability of co-occurrence in texts (Firth 1951/1957). Collocations are not absolute or 100 percent deterministic, but are the probabilistic outcomes of repeated combinations created and experienced by language users. We talk of being *madly in love* in preference to (but not in absolute exclusion or prohibition of) being *crazily in love*; tea is usually *strong*, but cars are *powerful*; and so on. Key discussions of the implications of Firth's theory of collocation appear in Halliday (1966) and Sinclair (1966). Both Halliday and Sinclair foresaw in those papers the development of the computational analysis of lexis using large amounts of text. Collocation studies show, most importantly, that a good deal of semantically transparent vocabulary is to a greater or lesser degree fossilized into restricted patterns (see also Aisenstadt 1981). The notion of collocation shifts the emphasis from the single word to pairs of words as integrated chunks of meaning, and collocation has become an uncontroversial element in a good deal of language description and pedagogy.

Words in corpora

The growth of corpus linguistics (see McCarthy 1998: Chapter 1 for a brief historical sketch) has convinced linguists that vocabulary is

much more than the 'unordered list of all lexical formatives' which Chomsky (1965:84) referred to it as. Pioneering studies of large corpora by linguists such as Sinclair (1991) have shown lexis to be a far more powerful influence in the basic organization of language and of meaning than was ever previously conceived. Corpora reveal the regular, patterned preferences for modes of expression of language users in given contexts, and show how large numbers of users separated in time and space repeatedly orient towards the same language patterns when involved in comparable social activities. Corpora reveal that much of our lexical output consists of multiword units; language occurs in ready-made chunks to a far greater extent than could ever be accommodated by a theory of language insistent upon the primacy of syntax.

Sinclair (1987, 1991), based on his lexicographic studies of collocation in the Birmingham Collection of English Text (later known as The Bank of English), sees two fundamental principles at work in the creation of meaning. These he calls the *idiom principle* and the *open choice principle*. The *idiom principle* is the central one in the creation of text and meaning in speech and writing, and works on the basis of the speaker/writer having at his/her disposal a large store of ready-made lexico-grammatical chunks. Syntax, far from being primary, is only brought into service occasionally, as a kind of 'glue' to cement the chunks together.

Sinclair (1996) sees form and meaning as complementary: different senses of a word will characteristically be realized in different structural configurations. This extends the original notion of collocation to encompass longer strings of words and includes their preferred grammatical configurations or *colligations* (see also Mitchell 1971). The unitary consequences of collocation and colligation produce meaningful strings or chunks which are stored in memory (see also Bolinger 1976) and which substantiate the idiom principle.

Corpus-based work on grammar has had similar consequences, especially in the research of Biber and his associates (Biber et al. 1999). Biber et al. examine a wide range of recurrent expressions, even though many of them are not 'idiomatic' in the sense of being semantically opaque, and even though they may be syntactically incomplete (see the discussion below), and they term such strings *lexical bundles* (see also Biber and Conrad 1999). Significant recurrence is defined by establishing frequency cut-off points, for example, that a string must occur at least 10 times per million

words of text (or 20 times in the case of Cortes 2002), and must be distributed over a number of different texts. This means that a bundle might consist of a syntactically incomplete but meaningful string such as *to be able to* or *a lot of the*, examples offered by Cortes (2002), along with more obviously semantically- and pragmatically-integrated expressions such as *as a result of* and *on the other hand*. Those investigating lexical bundles generally argue that the bundles operate as important structuring devices in texts and are register- (or genre-) sensitive. Oakey (2002) demonstrates that common recurring strings such as *it has been [shown/observed/argued/etc] that*, which are used to introduce external evidence in writing, are differently distributed across three genres. Furthermore, the presence (or absence) of lexical bundles in second-language learner output has been considered a useful measure of comparison and evaluation of learner competence vis-à-vis native speaker competence (see de Cock 1998, 2000; see also Granger 1998).

Phraseology and idiomaticity

Developments arising from corpus-based studies have been paralleled, over the years, by non-corpus-based research into multi-word lexical units. The general field of phraseology and the study of idiomaticity have contributed to our understanding of multi-word phenomena, both in the West and (at the same time but often unknown to Western linguists) in the former Soviet Union (see Kunin 1970, Benson and Benson 1993). Such linguists have long worked within frameworks not dominated by syntax.

In the literature, discussion usually centres upon the semantics, the syntax, the cross-linguistic differences, and the universality of opaque idiomatic expressions (Makkai 1978, Fernando and Flavell 1981), which, by and large, are relatively rare in occurrence in everyday conversation. But there has also been useful and illuminating research into everyday conversational routines, gambits, and discourse markers which has involved a recognition of the multi-word nature of such items (see Coulmas 1979, 1981a, b). However, few idiomatologists have gone so far as to examine idiom use in naturally-occurring spoken data, an exception being Strässler (1982), and more recently Powell (1992).

McCarthy (1998) lists different formal and functional types of idiomatic expression which were found through manually searching

the CANCODE spoken corpus, the corpus on which the present paper is based (see below). McCarthy's purpose in that categorization was to show that a wide range of idiomatic fixed expressions occur in everyday conversation, both formally and functionally, perhaps wider than that suggested by the traditional emphasis on *verb + object* idioms (e.g., *kick the bucket, pass the buck*) in language pedagogy.

The study of multi-word units has also focused on how they have developed pragmatic specialisms in regular contexts of use (e.g. Bolinger 1976, Cowie 1988, Nattinger and deCarrico 1992, Lewis 1993, and Howarth 1998). Multi-word expressions have additionally come under the scrutiny of sociolinguists and conversation analysts, where the purpose is to judge the social significance of the moment of placement and use of particular items. Drew and Holt (1998), for instance, show that idiomatic expressions occur regularly at places of topic-transition and as summaries of gist. Such work underlines the non-random use of idiomatic expressions and strengthens the claims of the present paper that investigating multi-word phenomena can tell us much about the nature of interaction.

Different terminology has been used to describe the phenomena of interest to us here, including *lexical phrases* (Nattinger and deCarrico 1992), *prefabricated patterns* (Hakuta 1974) *routine formulae* (Coulmas 1979), *formulaic sequences* (Wray 2000, 2002), *lexicalized stems* (Pawley and Syder 1983), *chunks* (De Cock 2000), as well as the more conventionally-understood labels such as *(restricted) collocations, fixed expressions, multi-word units/ expressions, idioms*, etc. Whatever the terminology, multi-word phenomena seem to be central to a wide range of linguistic and applied linguistic preoccupations. 'Off-the-peg' vocabulary enables fluent production in real time, and would seem to be at least as significant as the single-word elements that compose texts when it comes to investigating either the semantics or the pragmatics of language. Indeed, one can hardly imagine language not being (at least in part) produced ready-assembled (see Bolinger 1976).

In pedagogical terms, an over-emphasis in language teaching on single words out of context may leave second language learners ill-prepared both in terms of the processing of heavily-chunked input such as casual conversation, as well as in terms of productive fluency. Wray, whose recent work on what she calls *formulaic sequences* (which include idioms, collocations, and institutionalized

sentence frames), stresses that both formally and functionally, formulaic sequences circumvent the analytical processes associated with the interpretation of open syntactic frames in terms of both encoding and decoding (see Wray 2000, 2002). She also notes, with relevance to the present paper, that utterances may be formulaic 'even though they do not need to be' (Wray 2000: 466), in the sense that they can be generated by the rules of open syntax and the lexicon (she gives as an example *It was lovely to see you*). Their formulaic nature resides in their recurrence and established lexico-grammatical patterns in alliance with their pragmatically-specialized functions (in the case of *it was lovely to see you*, a follow-up message after spending pleasurable time with someone).

The present paper attempts to shift the balance away from the more semantically opaque multi-word expressions and seeks to tease out some of the most common sequences in everyday talk. As with most high-frequency phenomena, their recurrence is typically subliminal and not immediately accessible to the intuition of the native speaker. This paper therefore allows the first steps in the process of examining recurrent everyday multi-word strings to be effected automatically, by a computer count of recurring characters and spaces. This has both advantages and disadvantages, as the next section will show.

Data and method for the present study

Data and analytical procedure

This paper uses the 5-million word CANCODE spoken corpus. CANCODE stands for 'Cambridge and Nottingham Corpus of Discourse in English'. The corpus was established at the Department of English Studies, University of Nottingham, and is funded by Cambridge University Press. The corpus consists of five million words of transcribed conversations. The corpus recordings were made non-surreptitiously in a variety of settings including private homes, shops, offices, and other public places, in non-formal settings across the islands of Britain and Ireland, with a wide demographic spread. The CANCODE corpus forms part of the larger Cambridge International Corpus. For further details of the CANCODE corpus and its construction, see McCarthy (1998).

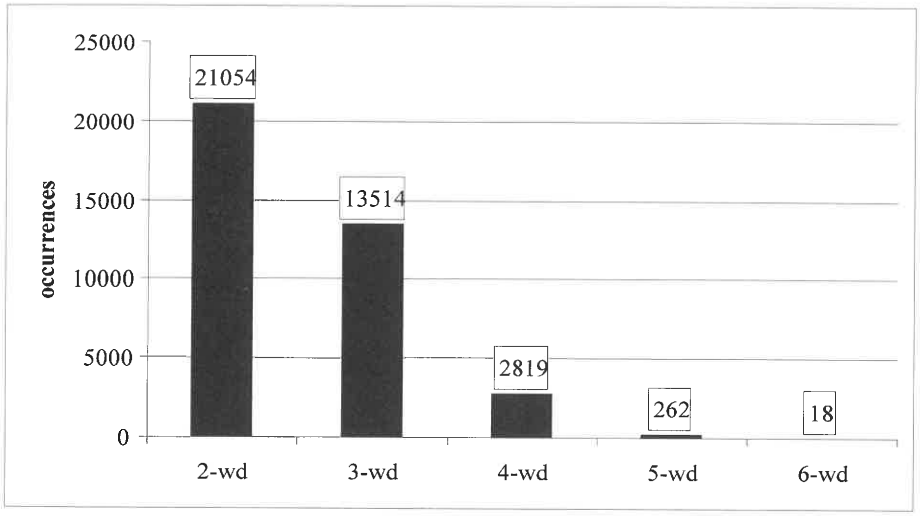
The analytical software used for the present paper (Wordsmith Tools, Scott 1999) is capable of automatically retrieving recurrent strings of characters and spaces (words for our intents and purposes) and giving a count for their occurrence. The user sets the number of words for the recurrent strings (e.g. two-word strings, three-word strings) and any cut-off points for frequency (e.g., minimum 10/50/100 occurrences). This necessarily means that the software will retrieve strings which in many cases lack any syntactic or semantic integrity, as well as strings that display integrity of one or both kinds.

Computers in their present state cannot distinguish between strings which recur but which have no psychological status as units of meaning (e.g., the fragmentary string *to me and* occurs more than 100 times in the CANCODE corpus) and those units which have a semantic unity and syntactic integrity, even though they may be less frequent (e.g., the discourse-marker phrase *as far as I know* occurs with less than half the frequency of *to me and*). This difficulty has led some researchers (e.g., Altenberg 1998, De Cock 2000) to incorporate fragmentary strings into their definition of chunks even where these include sub-phrasal and sub-clausal strings (De Cock offers as examples *in the* and *that the*), alongside pragmatically adequate sentence-frames such as *it is true that*. In the present paper we wish to focus on those items in the automatically extracted strings which display pragmatic integrity regardless of their syntax or lack of semantic wholeness, a task which necessitates manual inferencing and interpretation of the automatically generated data (see below).

The procedure for extracting the recurrent strings was to generate rank-order frequency lists of two-, three-, four-, five-, and six-word sequences for the entire 5-million word corpus. For practical reasons, a frequency cut-off point had to be established, and for the present purposes, an occurrence of at least 4 times per million words was the criterion for inclusion (in other words 20 times in the 5-million word corpus). This compares with Biber et al's (1999) figure of 10 times per million and Cortes' (2002) figure of 20 per million. Our figure is more liberal mainly because of the low occurrence of six-word clusters (only 18 being generated at the necessary 20 or more occurrences in five million words). Six-word recurrent clusters are of very low frequency in CANCODE, and it does seem that six is a practical cut-off point beyond which recurrent clusters seem to be extremely rare. Only one cluster of seven words occurs

more than 20 times: *but at the end of the day* (on the 'magic' number of seven as a psychological limit, see Miller, 1956). The lists for the smaller combinations were, predictably, much longer. Figure 1 shows the comparative distribution of two-, three-, four-, five-, and six-word clusters in excess of 20 occurrences, and it can be seen that there is a very sharp fall-off between three-word clusters and four-word clusters, and an even sharper drop between four- and five-word clusters. It should be noted that, in these counts, contracted forms such as *it's* and *don't* are considered as one 'word', since the computer is counting characters and spaces only.

Figure 1: Distribution of clusters in excess of 20 occurrences



Results

Tables 1 to 5 show the top 20 items in each cluster list for 2-5 word clusters, and all of the 6-word clusters.

Table 1:Top 20 two-word clusters

	Word	Frequency
1	*You know*	28,013
2	*I mean*	17,158
3	*I think*	14,086
4	*In the*	13,887

5	*It was*	12,608
6	*I don't*	11,975
7	*Of the*	11,048
8	*And I*	9,722
9	*Sort of*	9,586
10	*Do you*	9,164
11	*I was*	8,174
12	*On the*	8,136
13	*And then*	7,733
14	*To be*	7,165
15	*If you*	6,709
16	*Don't know*	6,614
17	*To the*	6,157
18	*At the*	6,029
19	*Have to*	5,914
20	*You can*	5,828

Table 2: Top 20 three-word clusters

	Word	Frequency
1	*I don't know*	5,308
2	*A lot of*	2,872
3	*I mean I*	2,186
4	*I don't think*	2,174
5	*Do you think*	1,511
6	*Do you want*	1,426
7	*One of the*	1,332
8	*You have to*	1,300
9	*It was a*	1,273
10	*You know I*	1,231
11	*You want to*	1,230
12	*You know what*	1,212
13	*Do you know*	1,203
14	*A bit of*	1,201
15	*I think it's*	1,189
16	*But I mean*	1,163
17	*And it was*	1,148
18	*A couple of*	1,136
19	*You know the*	1,079
20	*What do you*	1,065

Table 3: Top 20 four-word clusters

	Word	Frequency
1	*You know what I*	680
2	*Know what I mean*	674
3	*I don't know what*	513
4	*The end of the*	512
5	*At the end of*	508
6	*Do you want to*	483
7	*A bit of a*	457
8	*Do you know what*	393
9	*I don't know if*	390
10	*I think it was*	372
11	*A lot of people*	350
12	*Thank you very much*	343
13	*I don't know whether*	335
14	*And things like that*	329
15	*Or something like that*	328
16	*What do you think*	312
17	*I thought it was*	303
18	*I don't want to*	296
19	*That sort of thing*	294
20	*You know I mean*	294

Table 4: Top 20 five-word clusters

	Word	Frequency
1	*You know what I mean*	639
2	*At the end of the*	332
3	*Do you know what I*	258
4	*The end of the day*	235
5	*Do you want me to*	177
6	*In the middle of the*	102
7	*I mean I don't know*	94
8	*This that and the other*	88
9	*I know what you mean*	84
10	*All the rest of it*	76
11	*And all that sort of*	74
12	*I was going to say*	71

13	*And all the rest of*	68
14	*And that sort of thing*	68
15	*I don't know what it*	63
16	*All that sort of thing*	61
17	*Do you want to go*	61
18	*To be honest with you*	59
19	*An hour and a half*	56
20	*It's a bit of a*	56

Table 5: The six-word clusters (all)

	Word	Frequency
1	Do you know what I mean	236
2	At the end of the day	222
3	And all the rest of it	64
4	And all that sort of thing	41
5	I don't know what it is	38
6	But at the end of the	35
7	And this that and the other	33
8	From the point of view of	33
9	A hell of a lot of	29
10	In the middle of the night	29
11	Do you want me to do	24
12	On the other side of the	24
13	I don't know what to do	23
14	And all this sort of thing	22
15	And at the end of the	22
16	If you see what I mean	22
17	Do you want to have a	21
18	If you know what I mean	21

The tables exclude repetitions such as *you, you, you*, which often occur as stutter starts (although we recognize that these may indeed have importance in some kinds of analysis) and non-lexical phenomena such as hesitation markers (e.g., *er, er*). The lists were then used as the basis for analysis and interpretation, firstly in terms of identifying integrated units, and then in terms of what such units reveal about conversational interaction.

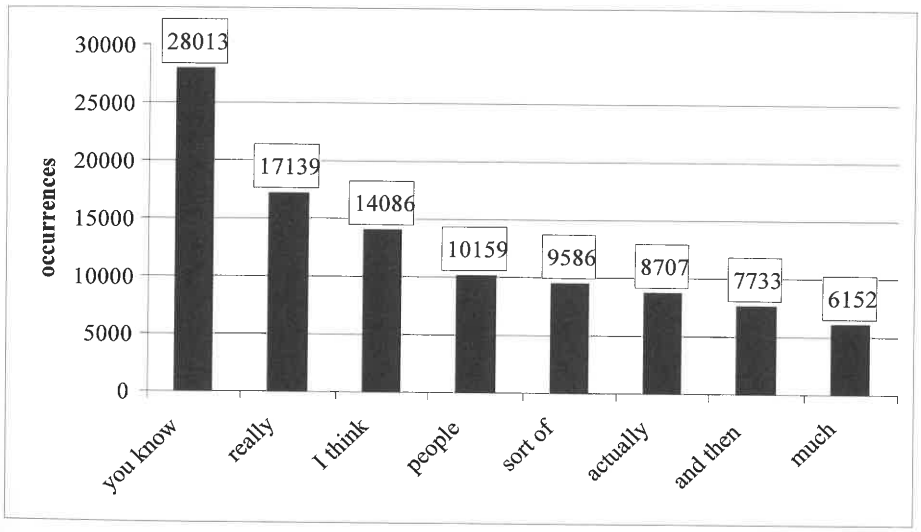
Clusters and single words

It is useful to gain a perspective on how the high frequency clusters relate to the distribution of single words in the corpus. An exhaustive count is beyond the scope of this paper, but some indicative examples are offered to assist the overall understanding of the place of clusters in a corpus-based description of the lexicon.

Only 33 items in the single-word rank order frequency list for CANCODE occur more frequently than the most frequent cluster (i.e., more frequently than the number one *you know*, which occurs 28,013 times). Clearly then, *you know* is one of the most frequent items in the English lexicon.

A selection of two-word clusters which occur with greater frequency than some common, everyday single words is given in Figure 2.

Figure 2: Two-word clusters and common single words

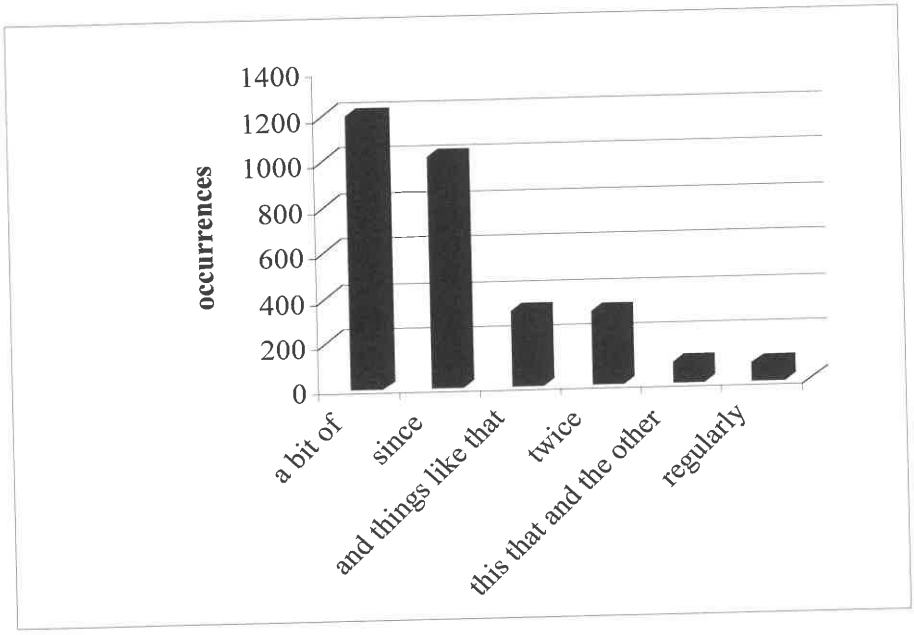


Individual clusters will be commented on below. Figure 3 below shows examples of three- and four-word clusters which occur more frequently than some common everyday words.

The graphs suggest that word lists which focus only on single words risk losing sight of the fact that many high-frequency clusters are more frequent and central to communication than even very frequent words. However, the question remains whether the clusters

in the tables and figures presented here should be considered as units of any kind or simply as statistical phenomena reflecting inevitable recurrence of a finite number of words in the vocabulary. In short, should something like *and then* be merely viewed as a co-occurrence arising from the extremely high frequency and weak collocability of its component words and their inevitable repeated collision in the corpus, or do such co-occurrences reveal anything about how we converse with one another?

Figure 3: Three- and 4-word clusters and common single words



Clusters as units of interaction

Pragmatic integrity

Many of the recurrent clusters present in the tables and graphs above are syntactic fragments, i.e., they do not constitute complete syntactic elements at phrasal or clausal levels. These include *in the, and I, of the*, and *do you* in the two-word list; *one of the* and *I think it's* in the three-word list; *the end of the* and *a bit of a* in the four-word list; and so on. Conventional grammars would certainly label these as incomplete in terms of structural units. That is not to say

that all models of grammar would reject such phenomena: emergent grammar, as epitomized in the work of Hopper (1998), considers fragments to be important clues as to how interaction unfolds and how meaning emerges rather than being pre-determined in linguistic units. And there is no obvious reason why one should exclude syntactically fragmentary strings from consideration when evaluating their *interactive* role. For instance, *I think it's* is indicative of the ubiquity of *I think* as a hedge prefacing evaluations of situations likely to be referred to by pro-form *it*. *I think* is number 3 in the two-word list, occurring more than 14,000 times. *A bit of a* may be considered similarly: speakers routinely downtone utterances with *a bit (of a)* (e.g., *it's a bit late, it was a bit of a mess*), evidenced by the fact that *a bit* occupies rank number 24 (with a frequency of 5,341) in the two-word cluster list. Thus although an expression like *a bit* may be semantically delexicalized (in other words fairly lexically 'empty'), and although it may be syntactically dependent in its role as a modifier, it is pragmatically specialized as a downtoner, and exhibits pragmatic adequacy and integrity. Other clusters seem less pragmatically motivated (e.g., *it was, what do you, in the middle of the*) and their occurrence is probably due to the regularity and stability of the content-world itself. For example, the cluster *an hour and a half* is number 19 in the five-word list; this may simply reflect the fact that people frequently make references to time and duration. We would argue, then, that it is in pragmatic categories rather than syntactic or semantic ones that we are likely to find the reasons why many of the strings of words are so recurrent. By pragmatic categories here we mean those which embrace the creation of speaker meanings in context. Such categories include discourse marking, the preservation of face and the expression of politeness, and the acts of hedging and purposive vagueness, all of which create the speaker-listener world rather than the content- or propositional world.

Discourse marking

Some of the most frequent clusters have discourse-marking functions. These include: *You know, I mean, And then, But I mean, You know what I mean, Do you know what I mean, At the end of the day,* and *If you see what I mean.*

You know, as the most frequent cluster of all, is an important token of projected shared knowledge between speaker and listener,

as well as being a topic-launcher (Östman 1981 and Erman 1987); it is ubiquitous in everyday informal talk, as extract (1) shows. (All corpus extracts indicate the different speakers as <$1>, <$2>, etc. The equals sign (=) indicates a truncated word or turn. The plus sign (+) indicates that an incomplete turn continues after an interruption by another speaker.)

(1) <$1> *You know*, our Gregory he's only fifteen but he wants to be a pilot.
<$2> Does he?
<$1> Now he couldn't get in this year to go to Manchester, *you know*, on that erm course that they do, experience course thing.
<$2> Work experience.
<$1> But he's going for next we= next year.
<$2> Oh yeah.
<$1> Work+
<$3> Oh yeah.
<$1> +experience yeah. And this time he's been to erm Headingley, coaching, doing a bit of coaching with the young kids *you know*.

The extended clusters *(do) you know what I mean* have a similar function of checking shared knowledge. Separately, *I mean* is used when shared knowledge is not inferred or when the speaker needs to reformulate (Erman 1987):

(2) [In a sports equipment shop]
<$1> Are there any tennis racquets you'd recommend? Erm I need the medium price range.
<$2> Medium price.
<$1> Yeah.
<$2> What are you looking= What sort of price range are you looking at?
<$1> Erm well not too expensive.
<$2> *I mean*, they start at m= about fifteen pounds and they go up anywhere to about three hundred quid.
<$1> Oh right. Probably under a hundred pounds cos it's not+
<$2> Okay.
<$1> +professional

<$2> Is it for yourself?
<$1> Yeah.
<$2> *I mean*, the decent racquets, you've got you've got a Head seventy nine.
<$1> Yeah.

The overlap of components within *(do) you know (what) (I mean)* partly account for the extreme high frequency of *you know* and *I mean*, but above all it is their core function in the monitoring of the state of shared knowledge which gives them the pragmatic integrity which qualifies them for consideration as units. Likewise, *and then* is extremely frequent in narrative as a marker of temporal sequence, while *at the end of the day* typically has a summarizing function.

Face and politeness

Speakers use indirect forms to perform speech acts such as directives and requests in order to protect the face of their receivers, and the clusters reveal common everyday frames for such acts. Indirectness is also important in the polite and non-face-threatening expression of attitude, opinion, and stance. Speakers work hard to protect the face of their interlocutors, wishing neither to demean them or coerce them (see Brown and Levinson 1987). Clusters in this category include: *Do you think, Do you want (me) (to), I don't know if/whether, What do you think, I was going to say.*

Extracts (3) and (4) show these in action:

(3) [Discussing the priorities for preserving lives in the British National Health Service, and whether age should be a factor]
<$2> I thought it was shocking.
<$1> Mm. *Do you think it* would have made any difference if she was say eighty years of age instead of a teen-ager?
<$2> Well I think that er anyone's attitude should be to save life irrespective of age

(4) [At a travel agent's]
<$3> Did you want to take out insurance?

<$1> Erm I'd like to ask about it but *I don't know if* I want to do that today.
<$3> Okay.

The utterances containing the clusters can be perfectly well-formed with more direct assertions (e.g., *Would it have made any difference ...?; I don't want to do that today*) but the presence of the clusters plays a significant role in the mutual protection of face and the smooth, sensitive, and polite progression of the talk. Once again, it is pragmatic integrity rather than syntactic or semantic wholeness which is most relevant.

Another important aspect of face-protection and politeness is hedging. Some of the most frequent clusters have a hedging function, i.e., they modify propositions to make them less assertive and less open to challenge or refutation. These include: *I think, Sort of, A bit (of a), I don't know, I don't think,* and To *be honest with you.* Extracts (5) and (6) illustrate these functions.

(5) <$1> That's fine Jess. Are there many to do?
<$2> No.
<$1> No. I've got an appointment in Healdham at five fifty so I'm going to have to leave you know *sort of* shortly after three.

(6) <$1> I went to college in the spring
<$2> Mm.
<$2> and sat the exam in June and passed it.
<$1> Mm.
<$2> But it was basically er an E-E-C update on the new regulations. *To be honest with you* it was pret= pretty easy I thought but you know s= some people have to fail I suppose and some do it you know.

Vagueness and approximation

Equally apparent in the high frequency clusters are markers of purposive vagueness and approximation. Vagueness is central to informal conversation, and its absence can make utterances blunt and pedantic, especially in such domains as references to number and quantity, where approximations are the norm in conversation. Vagueness also enables speakers to refer to semantic categories in

an open-ended way which calls on shared cultural and real-world knowledge to fill in the category members referred to only obliquely (see Chafe 1982, Powell 1985, and Channell 1994). Such tokens include: *A couple of, And things like that, Or something like that, (And) that sort of thing, (And) this that and the other, All the rest of it*, and *(And) all this/that sort of thing*. Examples from the corpus show the clusters in action.

(7) [At a travel agent's]
<$1> And what about er local taxis *and things like that*? Are they included or are they extra?
<$2> Er everything is included apart from any sort of top up insurance you may want.

(8) <$1> She said, "We've just come out here. We've just bought an apartment here".
<$2> Mm.
<$1> And she said, "We've come out to furnish it and buy the furniture *and this that and the other*".

In extracts (7) and (8) it would be clearly conversationally inappropriate to list all the items implied by the vague tokens; speakers need only allude to the shared cultural knowledge and may assume their listeners can fill in the detail. Once again, the vague tokens exhibit pragmatic integrity and play central interactive roles, even though their syntax is incomplete and dependent.

Discussion and conclusion

Not all of the clusters can or need to be accounted for in terms of pragmatic integrity. For example, clusters such as *on the*, *it was a*, and so on are probably best explained either by their semantics (e.g. core spatio-temporal notions) or by the frequency of acts such as describing location or narrating the past. However, by exploring the uses of the clusters in the corpus, it does seem that amongst the most frequent (the top 20 in each case), there seem to be a considerable number which achieve wholeness as units when their pragmatic functions are adduced. What such clusters show is the all-pervasiveness of interactive meanings in everyday conversation and the degree to which speakers constantly engage on the interactive

plane as well as the transactional or content plane. Their addition to the vocabulary list of any language is not an optional extra, since the meanings they create are extremely frequent and necessary in discourse, and are fundamental to successful interaction. The units support Sinclair's notion of the idiom principle at work, with the clusters best viewed as being evidence of single linguistic choices rather than assembled at the moment of speaking. They make fluency a reality.

A final word needs to be said about the status of such units vis-à-vis the more opaque idiomatic units that have traditionally been studied. In the absence of corpus evidence it is difficult to introspect on what one says. It is much easier to introspect on what one writes, and additionally, introspection is more likely to light upon the colourful, the curious, the rare, precisely because such items are psychologically salient. Hence it should not surprise us that, with few exceptions, pre-corpus studies of multi-word units focussed on idioms, phrasal verbs, compounds, and so on, either as colourful curiosities or, in the pedagogic domain, a difficult characteristic of English for learners to struggle with. Meanwhile the banal, hidden, subliminal patterns of the everyday lexicon stubbornly resisted exposure. Corpus analysis enables us to circumvent our difficulties in retrieving such patterned occurrences, but the automatic retrieval of recurrent strings is only the beginning, and a good deal of inferential analysis is still necessary to see meaning in the mechanical and dispassionate statistics spewed out by the computer.

References

Aisenstadt, Esther. 1981. Restricted collocations in English lexicology and lexicography. *ITL Review of Applied Linguistics* 53: 53-61.

Altenberg, Bengt. 1998. On the phraseology of spoken English: the evidence of recurrent word combinations. *Phraseology: Theory Analysis and Applications,* ed. by Anthony Cowie, 101-122. Oxford: Oxford University Press.

Bazell, Charles, John Catford, Michael Halliday, and Robert Robins, eds. 1966. *In Memory of J. R. Firth.* London: Longman.

Benson, Morton and Evelyn Benson. 1993. *Russian-English Dictionary of Verbal Collocations (REDVC)*. Amsterdam: Benjamins.

Biber, Douglas and Susan Conrad. 1999. Lexical bundles in conversation and academic prose. *Out of Corpora: Studies in Honor of Stig Johansson*, ed. by Hilde Hasselgard and Signe Oksefjell, 181-190. Amsterdam: Rodopi.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

Bolinger, Dwight. 1976. Meaning and memory. *Forum Linguisticum,* 1: 1-14.

Brown, Penelope and Stephen Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. *Spoken and Written Language: Exploring Orality and Literacy*, ed. by Deborah Tannen, 35-53. Norwood, NJ: Ablex Publishing Corporation.

Channell, Joanna. 1994. *Vague Language.* Oxford: Oxford University Press.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Cortes, Viviana. 2002. Lexical bundles in freshman composition. In Reppen et al. (2002), 131-145.

Coulmas, Florian. 1979. On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics* 3: 239-66.

Coulmas, Florian, ed. 1981a. *Conversational Routine.* The Hague: Mouton.

Coulmas, Florian. 1981b. Idiomaticity as a problem of pragmatics. *Possibilities and Limitations of Pragmatics*, ed. by Herman Parret, Marina Sbisà, and Jef Verschueren, 139-51. Amsterdam: John Benjamins.

Cowie, Anthony. 1988. Stable and creative aspects of vocabulary use. *Vocabulary and Language Teaching*, ed. by Ronald Carter and Michael McCarthy, 126-39. London: Longman.

De Cock, Sylvie. 1998. A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics* 3: 59-80.

De Cock, Sylvie. 2000. Repetitive phrasal chunkiness and advanced EFL speech and writing. *Corpus Linguistics and Linguistic Theory. Papers from ICAME 20 1999*, ed. by Christian Mair and Marianne Hundt, 51-68. Amsterdam: Rodopi.

Drew, Paul and Elizabeth Holt. 1998. Figures of speech: figurative expressions and the management of topic transition in conversation. *Language in Society* 27: 495-522.

Erman, Britt. 1987. *Pragmatic Expressions in English: A Study of "you know," "you see," and "I mean" in Face-to-face Conversation*. Stockholm: Almqvist & Wiksell.

Fernando, Chitra and Roger Flavell. 1981. *On Idiom: Critical Views And Perspectives*. Exeter: University of Exeter.

Firth, John Rupert. 1935. The technique of semantics. *Transactions of the Philological Society*: 36-72.

Firth, John Rupert. 1951/1957. *Papers in Linguistics*. Oxford: Oxford University Press, 190-215.

Granger, Sylviane. 1998. Prefabricated writing patterns in advanced EFL writing: collocations and formulae. *Phraseology: Theory, Analysis and Applications*, ed. by Anthony Cowie, 145-160. Oxford: Clarendon Press.

Hakuta, Kenji. 1974. Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning* 24: 287-298.

Halliday, Michael. 1966. Lexis as a linguistic level. In Bazell et al. (1966), 148-162.

Hopper, Paul. 1998. Emergent grammar. *The New Psychology of Language*, ed. by Michael Tomasello, 155-175. Hillsdale, NJ: Lawrence Erlbaum Associates.

Howarth, Peter. 1998. Phraseology and Second Language Proficiency. *Applied Linguistics* 19 (1): 24-44.

Kunin, Aleksandr. 1970. *Anglijskasa frazeologija*. Moscow: Izdat'elstvo 'Vysšajaškola'.

Lewis, Michael. 1993. *The Lexical Approach: The State of ELT and a Way Forward*. Hove UK: LTP.

Makkai, Adam. 1978. Idiomaticity as a language universal. *Universals of Human Language, Volume 3: Word Structure*, ed. by Joseph Greenberg, 401-448. Stanford, CA: Stanford University Press.

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

Miller, George. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63: 81-97.

Mitchell, Terence. 1971. Linguistic 'goings-on': collocations and other lexical matters arising on the linguistic record. *Archivum Linguisticum* 2: 35-69.

Nattinger, James and Jeanette DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Oakey, David. 2002. Formulaic language in English academic writing. In Reppen et al. (2002), 111-129. Amsterdam: John Benjamins.

Östman, Jan-Ola. 1981. *You Know: A Discourse Functional Approach.* Amsterdam: John Benjamins.

Pawley, Andrew and Frances Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. *Language and Communication*, ed. by Jack Richards and Richard Schmidt, 191-226. New York: Longman.

Powell, Mava. 1985. Purposeful vagueness: an evaluative dimension of vague quantifying expressions. *Journal of Linguistics* 21: 31-50.

Powell, Mava. 1992. Semantic/pragmatic regularities in informal lexis: British speakers in spontaneous conversational settings. *Text* 12 (1): 19-58.

Reppen, Randi, Susan Fitzmaurice, and Douglas Biber, eds. 2002. *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins.

Scott, Michael. 1999. *Wordsmith Tools*. Software. Oxford: Oxford University Press.

Sinclair, John. 1966. Beginning the study of lexis. In Bazell et al. (1966), 410-430. London: Longman.

Sinclair, John. 1987. Collocation: a progress report. In Steele and Threadgold (1987), 319-331.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. 1996. The search for the units of meaning. *Textus* IX: 75-106.

Steele, Ross, and Terry Threadgold, eds. 1987. *Language Topics: An International Collection of Papers by Colleagues, Students and Admirers of Professor Michael Halliday to Honour him on his Retirement*, Vol. II. Amsterdam: John Benjamins.

Strässler, Jörg. 1982. *Idioms in English: a Pragmatic Analysis*. Tübingen: Gunter Narr Verlag.

Wray, Alison. 2000. Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, 21 (4): 463-489.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.