What Is Advanced-Level Vocabulary?  
The Case of Chunks and Clusters

Ronald Carter  
University of Nottingham  
Nottingham, England

Using suggestions and examples made by Michael McCarthy (see his paper in this symposium), we move from the notion of advanced vocabulary as a set of words to the notion of advanced vocabulary as sets of words in combination. Once again, corpus analysis will be employed to help us search for patterns and frequencies. However, when we expand our search criteria to look for groupings of more than one word, things become more complicated, and there are clear lessons to be learned about how we describe the vocabulary of a language, as well as implications for what teachers teach in their vocabulary lessons and how learners approach the task of acquiring vocabulary and developing fluency. Throughout this paper we work from, but also hope to challenge, the understanding of many teachers, researchers, and learners that vocabulary means no more than all the single words of a language.

FROM WORD TO COLLOCATION TO CLUSTER

The single word has served us well and will continue to do so. However, units consisting of more than one word, such as phrasal verbs, compounds, and idioms, are often taught only at higher levels of proficiency. Exceptions include basic greetings and everyday expressions (e.g., How are things? See you tomorrow), functional phrases (e.g., Happy New Year, Good luck), prepositional phrases (e.g., on the weekend, in May), and compounds (e.g., cell phone, bookstore). Is this a problem?

Linguists have for a long time been interested in how words combine as pairs in collocations and how groupings of more than one word often have unitary meanings and specialised functions (Biber, Johansson, Leech, Conrad, & Finegan, 1999, chap. 13; Bolinger, 1976; Cowie, 1988; Lewis, 2000; Pawley & Syder, 1983). The advent of corpus linguistics has enabled linguists to verify these earlier, mainly intuition-based notions in actual, attested language use on a large scale.

Collocation

The meaning of a word is as much a matter of how it combines with other words in actual use (i.e., its collocations) as it is of the meaning it possesses in itself. So, in this view, lean is part of the meaning of meat, and vice versa, because of their high probability of co-occurrence in texts. Lean and meat thus keep regular company with each other and collocate significantly: slim and meat are not likely to do so to any significant extent, though slim and figure might. Collocations are not absolute or deterministic but are probabilistic patterns resulting from repeated combinations used and encountered by the speakers of any language. We say amicable divorce (in preference to friendly divorce); tea is usually strong (but not powerful), and cars are powerful (though not strong), while arguments can be both powerful and strong, and so on.

Some 40 years ago, both Halliday (1966) and Sinclair (1966) foresaw the development of computational analysis of texts as a way of getting at the common collocations of a language,
and both authors, in different ways, have fulfilled that vision, most notably Sinclair (1991, 2004). However, it is computers that have been very good at teasing out the collocations of the everyday words that are most difficult to arrive at by intuition alone. Common verbs such as get, go, turn, and so on, also display distinct preferences for what they combine with. Things turn or go grey, brown, white; people go (but not turn) mad, insane, bald, blind. The notion of collocation therefore shifts the emphasis from the single word to pairs of words as integrated chunks of meaning and usage, and collocation has now become an accepted aspect of vocabulary description and pedagogy (e.g., Lewis, 2000; McCarthy & O’Dell, 2005; Rudzka, Channell, Putseys, & Ostyn, 1982). Clearly, for the learner of any second or foreign language, learning the collocations of that language is not a luxury if anything more than a survival level mastery of the language is desired because collocation permeates even the most basic, frequent words.

**Idioms and Phrases**

Idioms and phraseology have always been deeply embedded within the pedagogy of vocabulary teaching. Publishers are aware of this and offer materials specially devoted to learning idioms, and there are good dictionaries of idioms available for English, including corpus-based ones (e.g., Cowie & Mackin, 1975). Teachers and learners, especially advanced-level learners, are often fascinated by idioms, and everyone has their own favourite. *It’s raining cats and dogs* (It’s raining very hard) or *I smell a rat* (I am very suspicious) are often quoted as examples, though searches in multimillion-word corpora show that such examples only rarely occur in actual usage. Until recently, very little analytical work has been done on how idioms function in naturally occurring discourse (however, see McCarthy, 1998, chap. 7; McCarthy & Carter, 1994, chap. 3; and). The existence of idioms does, however, indicate that vocabulary and knowledge of vocabulary go beyond single words into larger structures and that these structures are probably learned as whole units and are formulaic in organisation. Whether idioms count as chunks is another question that needs to be posed alongside the question: What is the best point at which pedagogically to focus on the development of idiomatic knowledge? Further exploration of corpora will be undertaken to examine such questions. But first, on to formulaic language.

**THE PHENOMENON OF FORMULAIC LANGUAGE: FIXED WORDS IN FIXED PLACES?**

A considerable amount of work has been done on formulaic language, but the research tends to be widely scattered across a number of fields (child L1 acquisition, psychology, corpus linguistics). This diffusion is illustrated by the wide variety of terminology (see Carter & McCarthy, 1988; Carter, 1998, chap. 4; Wray, 2002, p. 9; Schmitt, 2004) found for the various sorts of formulaic language:

- chunks
- collocation
- conventionalized forms
- clusters
- formulaic speech
- formula
- holophrase
- lexical bundles
- multiword units
- prefabricated routines
- ready-made utterances

Schmitt and Carter (2004) also illustrate this diversity:

formulaic sequences can be long (You can lead a horse to water, but you can’t make him drink) or short (Oh no!), or anything in between. They are commonly used for
different purposes. They can be used to express a message or idea (The early bird gets the worm = do not procrastinate), functions ([I'm] just looking [thanks] = declining an offer of assistance from a shopkeeper), social solidarity (I know what you mean = agreeing with an interlocutor), and to transact specific information in a precise and understandable way (Wind 28 at 7 = in aviation language this formula is used to state that the wind is 7 knots per hour from 280 degrees). They realize many other purposes as well, as formulaic sequences can be used for most things society requires of communication through language. These sequences can be totally fixed (Ladies and Gentlemen) or have a number of “slots” which can be filled with appropriate words or strings of words, for example [someone/thing, usually with authority] made it plain that _[something as yet unrealized was intended or desired]_. (p. 3)

What descriptions of these various forms reveal is that fixedness is a key feature of formulaic language, in that at least some of its components are fixed, allowing them to be memorized and used as wholes, rather than being newly created for each use. On first inspection, one might assume that all formulaic language is completely fixed, but this is not the case. Of course, some is, and idioms are usually cited as examples. For instance, corpus evidence shows that to have forty winks occurs almost exclusively as that exact phrase and not as variations such as to have thirty-nine winks, to have a short wink. In other words, if we want to use an idiom to express the notion to have a short sleep we can only use the intact idiom to have forty winks and not some variation. However, much formulaic language is not completely fixed and, in fact, allows for a surprising amount of flexibility. Carter (1998, chap. 4, 2004) and especially Moon (1997, p. 53) illustrate this:

British vs. American English

not touch someone/something with a bargepole
not touch someone/something with a ten foot pole

Varying a lexical component

burn your boats
burn your bridges

Verb variation

cost an arm and a leg
pay an arm and a leg
spend an arm and a leg
charge an arm and a leg

Truncation

every cloud has a silver lining
silver lining

Transformation

break the ice
ice-breaker
ice-breaking

In fact, it seems that once a piece of formulaic language becomes well known in a speech community, it can be creatively adapted and still be comprehensible. It is worth expanding on Moon’s example of truncation. For example, as Schmitt (2006) illustrates, the well-known
idiom *Every cloud has a silver lining* occurs in the 100-million-word British National Corpus (BNC) seven times. But the phrase *silver lining* occurs in the corpus 75 times representing the same meaning. Clearly, people prefer to use a shortened version of the idiom in practice, often in highly creative ways (examples from Schmitt, 2006, pp. 8–9):

This proved to be much more than a search for the proverbial silver lining -may contain a silver lining for the consumers. 
reformed shopaholics almost always speak of a silver lining to the cloud which hung over their lives (and bank accounts).

Beyond idioms, it is not surprising that many other types of formulaic language also contain variation, for it is an advantage in much of language use to allow more flexibility of meaning. For example, if we wish to express the notion that some activity or achievement is unusual, unexpected, or exceptional, then we can use phrases like *Mike thinks nothing of running 5 miles before breakfast* or *They think nothing of driving 100 miles per hour on the freeway*. The underlying structure to these sentences is ______ thinks nothing of ______, which allows the flexibility to express the unexpected notion in many different situations (see Schmitt, 2006, for more examples).

**Words in Corpora**

A corpus can reveal the regular, patterned preferences of the language users represented in it, speaking and writing in the contexts in which the corpus was gathered. A big, general corpus can show how large numbers of language users, separated in time and space, repeatedly orientate toward the same language choices when involved in comparable social activities. And corpora do reveal that much of our linguistic output consists of multiword units rather than just single words. Language is available for use in ready-made chunks and clusters to a far greater extent than could ever be accommodated by a theory of language which rested upon the primacy of syntax, as the transformational-generative (TG) tradition did. One could reasonably posit that an overemphasis 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 their own productive fluency.

In this paper, we want to shift the balance away from the more semantically opaque multiword expressions, the traditional *idioms*, in other words, and will focus instead on some of the most common *chunks* in English. (Although the term *cluster* captures the way words combine and coalesce, the term *chunk* is preferred here as a way of indicating the necessary preformulated and flexibly fixed nature of this phenomenon.)

The main examples will be drawn from spoken English, not least because issues of fluent production and reception are foregrounded in talk. As with most high frequency phenomena, their core contribution to language use is subliminal and not immediately accessible to the intuition of the native speaker or fluent user. In this paper, therefore, we allow the first steps in the process of examining recurrent everyday chunks to be done automatically, by a computer count of recurring characters and spaces. This has both advantages and disadvantages, as we have already argued and as the next section will show with concrete examples. This is where a corpus comes in.
LOOKING AT CORPUS DATA
Using a 4.7-million-word sample of North American English conversation from the Cambridge International Corpus (CIC), and applying corpus analytical software to obtain a frequency count for recurrent chunks, the following totals emerge for chunks occurring more than 20 times:

- two-word chunks: 19,509
- three-word chunks: 12,681
- four-word chunks: 2,953
- five-word chunks: 385

Tables 1 and 2 show the top 10 items in the list of chunks for two- and four-word items.

### Table 1. Top 10 Two-Word Chunks in the Cambridge International Corpus

<table>
<thead>
<tr>
<th>Chunk</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOU KNOW</td>
<td>45,873</td>
</tr>
<tr>
<td>I DON’T</td>
<td>17,708</td>
</tr>
<tr>
<td>I THINK</td>
<td>17,046</td>
</tr>
<tr>
<td>IN THE</td>
<td>13,979</td>
</tr>
<tr>
<td>AND I</td>
<td>13,757</td>
</tr>
<tr>
<td>OF THE</td>
<td>12,040</td>
</tr>
<tr>
<td>I MEAN</td>
<td>11,735</td>
</tr>
<tr>
<td>IT WAS</td>
<td>11,271</td>
</tr>
<tr>
<td>A LOT</td>
<td>10,174</td>
</tr>
<tr>
<td>KIND OF</td>
<td>9,962</td>
</tr>
</tbody>
</table>

### Table 2. Top 10 Four-Word Chunks in the Cambridge International Corpus

<table>
<thead>
<tr>
<th>Chunk</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>I DON’T KNOW IF</td>
<td>999</td>
</tr>
<tr>
<td>A LOT OF PEOPLE</td>
<td>759</td>
</tr>
<tr>
<td>I DON’T KNOW WHAT</td>
<td>709</td>
</tr>
<tr>
<td>OR SOMETHING LIKE THAT</td>
<td>570</td>
</tr>
<tr>
<td>A LOT OF THE</td>
<td>560</td>
</tr>
<tr>
<td>AND THINGS LIKE THAT</td>
<td>499</td>
</tr>
<tr>
<td>I DON’T WANT TO</td>
<td>479</td>
</tr>
<tr>
<td>I DON’T KNOW HOW</td>
<td>466</td>
</tr>
<tr>
<td>THERE’S A LOT OF</td>
<td>448</td>
</tr>
<tr>
<td>WHAT DO YOU THINK</td>
<td>442</td>
</tr>
</tbody>
</table>

Chunks and Single Words

Only 14 items in a single-word frequency list occur more often than the most frequent chunk (i.e., you know, which occurs 45,873 times). Of the first 100 items in the overall frequency list, 11 are two-word chunks, including I think and I mean. By the time we reach 500 items, there are 177 two-word chunks and 7 three-word chunks, that is, 35% of the most frequent items are chunks, not single words. A selection of chunks with greater frequency than some common single words is given in Table 3.
Table 3. High-Frequency Chunks And Single Words in the Cambridge International Corpus

<table>
<thead>
<tr>
<th>Chunk</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOU KNOW</td>
<td>45,873</td>
</tr>
<tr>
<td>REALLY</td>
<td>20,838</td>
</tr>
<tr>
<td>I THINK</td>
<td>17,046</td>
</tr>
<tr>
<td>PEOPLE</td>
<td>11,984</td>
</tr>
<tr>
<td>KIND OF</td>
<td>9,962</td>
</tr>
<tr>
<td>AND THEN</td>
<td>8,971</td>
</tr>
<tr>
<td>I DON’T KNOW</td>
<td>8,074</td>
</tr>
<tr>
<td>WHERE</td>
<td>7,851</td>
</tr>
<tr>
<td>THEIR</td>
<td>6,487</td>
</tr>
<tr>
<td>SOMETHING LIKE THAT</td>
<td>1,027</td>
</tr>
<tr>
<td>FRIEND</td>
<td>1,014</td>
</tr>
<tr>
<td>I DON’T KNOW IF</td>
<td>999</td>
</tr>
<tr>
<td>A LOT OF PEOPLE</td>
<td>759</td>
</tr>
<tr>
<td>UNDER</td>
<td>743</td>
</tr>
</tbody>
</table>

Table 3 suggests that many high-frequency chunks are more frequent and more central to communication than even very frequent words. However, the question remains whether something like and then arises merely from the high frequency and weak collocability of its component words and their inevitable repeated clustering in the corpus, or whether such co-occurrences reveal something about conversation.

**Chunks as Units of Interaction**

**Pragmatic Integrity**

Many of the chunks above are syntactic fragments, that is, they are not complete phrases or clauses, for example, in the, and I, and I think it’s. However, they do have an interactive identity. I think it’s is indicative of the ubiquity of I think as a hedge prefacing evaluations of situations referred to by it. Other chunks seem less pragmatically integrated (e.g., it was), and their occurrence is probably due to the regularity of the content world itself. It is in pragmatic categories rather than syntactic or semantic ones that we are likely to find the reasons that many chunks are so frequent. Pragmatic categories refer to the creation of speaker meanings in context, including such functions as discourse marking; the preservation of face; and the expression of politeness, hedging, and purposive vagueness, which create the speaker-listener world rather than the content- or propositional-world.

**Discourse Marking**

Some of the most frequent chunks are discourse markers, for example, You know, I mean, I guess, (Do) You know what I mean. You know, the most frequent chunk, is an important token of projected shared knowledge between speaker and listener. I mean is also of high frequency, used when speakers need to paraphrase or elaborate. Extract 1 shows both chunks at work:

(1) Like I remember when I went to public school in Jersey and not that it wasn’t that bad. I mean I’m from a middle class town. You know we had people that you know . . . We had kids that whose family made you know a hundred and hundred fifty thousand dollars a year and people that generally didn’t make anything at all. You know.
The extended chunk (do) you know what I mean has a similar function of checking shared
knowledge:

(2) He’s totally like, you know what I mean, it’s like he’s very liberal. I mean, he’s
open minded. He doesn’t care. You know and so . . .

Face and Politeness
Speakers use indirect forms to perform speech acts such as directives (e.g., commands,
requests, suggestions) to protect the face of their addressees, and the chunks reveal common
frames for such acts. Indirectness is also important in the polite and non-face-threatening
negotiation of attitude and stance. Chunks in this category include Do you think, I don’t know
if, What do you think, I was going to say:

(3) [Someone describing how to keep food fresh.] I have the plastic bags that are
supposed to keep everything . . . I don’t know if you’ve seen it where they put the
vacuum cleaner on top and it sucks out all the air.

Some of the most frequent chunks have a hedging function, that is, they modify propositions
to make them less assertive and less open to refutation: I think, kind of, I don’t know, I don’t
think, a little bit.

Vagueness and Approximation
Equally apparent are chunks expressing purposive vagueness and approximation. Vagueness
is central to informal conversation, and its absence can make utterances blunt and pedantic,
especially in references to number and quantity. Vagueness also enables speakers to refer to
semantic categories in an open-ended way, which calls on shared knowledge to fill in
category members referred to obliquely. Such tokens include a couple of, and things like that,
and stuff like that:

(4) [Someone talking about hobbies] I do enjoy baking and I guess I always liked
making, uh, cookies and bars and things like that, that was more my specialty.

(See Carter and McCarthy, 2006, Appendix 1, for further discussion.)

CONCLUSIONS AND IMPLICATIONS

Not all of the recurrent strings listed here can or need to be accounted for in terms of
pragmatic integrity. For example, repeated strings such as on the, it was a, and so on, are
probably best explained either by their semantics (e.g., core spatiotemporal notions) or by the
frequency of acts such as describing location or narrating the past. However, by exploring the
uses of the chunks in the corpus, it is apparent that among the most frequent (the top 20 in
each case), there are a considerable number which have clear, common pragmatic functions
in the organisation and management of conversation and the speaker-listener relationship.
What the chunks show is the all-pervasiveness of interactive meaning-making in everyday
conversation and the degree to which speakers constantly engage with each other on the
interactive plane. The addition of these chunks to the vocabulary list of any language should
not be seen as an optional extra, since the meanings they create are extremely frequent and
necessary in discourse and are fundamental to successful interaction. The chunks are best
viewed as being evidence of single linguistic choices rather than assembled piece by piece at
the moment of speaking. Fluency is a generally undertheorised and unexplored category. Such examples make fluency more of a reality.

**Chunks in the Classroom**

There are thus some lessons to be learned about how vocabulary is organised through our analyses of common chunks. But what about the other type of lesson, what we do in class, and how students can be helped to learn and use chunks in a natural way? Some of the issues raised by this paper include the following:

- Chunks seem to be a badge of native-speaker identity. Why should learners who do not necessarily wish to sound like native speakers bother with them?
- If the use of ready-made chunks is central to fluency, how can they be presented and practised in language classrooms and teaching materials?
- How do learners typically process chunks when they encounter them?
- How can learners become aware of chunks and recognise potential chunks when they listen or read?

**Chunks as a Mark of the Native Speaker**

Research by Prodromou (2005) suggests that the speech of native speakers can be distinguished from the speech of advanced nonnative users of English (Prodromou refers to such users as successful users of English, or SUEs) by, among other things, the presence or absence of common chunks. Prodromou argues very persuasively that core chunks such as sort of and you know membership speakers within cultural communities and project a “deep commonality” (p. 134) among interlocutors which the learner or even the highly successful (nonnative) user may not wish to claim nor has any reason to claim. Prodromou is not advocating the enforced metamorphosis of expert users into native speakers. The lesson here may be that receptive mastery is more important than productive repertoire. But the issue is twofold: First, it is argued here that those students who do wish to push forward toward near-native fluency should be given appropriate exposure to and practice in the use of chunks. Certainly in terms of social integration (e.g., students living and attempting to integrate in the L2 environment), it would seem that those who integrate more successfully are likely to acquire and use chunks more naturally. But, second, even those whose espoused goal is to be themselves, and not simply to ape native speakers, may wish to consider the implications of engaging in conversation without the use of the highly interactive tools which the common chunks represent— it may be that we end up precisely not being ourselves in the target language and may be presenting quite a false image of ourselves and a stereotyped image of our culture. It is important to air such issues in the language classroom so that students can make informed choices and not prejudge them.

**Chunks and Fluency**

One of the features of chunks not yet discussed, in which the evidence has, of necessity, been the purely printed evidence of corpus output on a computer screen, is that chunks have phonological unity; put simply, they need to be said fast and all in one go. Typically, chunks occupy a single intonation unit (or tone unit, separated here by //, characterised by one strong stressed syllable, marked here in bold capitals), and the rest of the chunk is much reduced:

// he’s SHY // you know what i MEAN //
// they sell JEWellery // and THAT sort of thing //
// the ROOM was // a BIT of a // MESS actually //
Choral or private repetition, increasing the speed at each repetition, with practice in reducing the nonstressed syllables, can be a useful way of drilling chunks so that they become imprinted in the memory as musical items. Then, in actual use, it can be stressed that it does not matter how slowly and carefully the rest of the utterance is, or needs to be, constructed. Provided the chunk is said fast, the utterance will sound natural; the opposite, a fast message with a slow chunk, will sound completely unnatural. The appropriate use of a smooth, quickly uttered chunk can transform even a lower level speaker’s fluency. The challenge of saying chunks at ever increasing speed can also be an enjoyable interlude in a vocabulary lesson.

Although chunks can be drilled for speed in isolation, naturally, it is also a good idea to incorporate them into sentences and longer utterances for more sustained practice. Presentation of chunks in spoken language can most naturally be done by raising awareness of them through listening and noticing activities. Practice can also take the form of re-inserting chunks into dialogues from which they have been removed.

**Processing Chunks**

There is evidence that the use of chunks frees up the cognitive processing load so that mental effort can be allocated to other aspects of production, such as discourse organisation and successful interaction (Girard & Sionis, 2004). In that sense, chunks liberate the learner and allow a degree of automaticity to take over in both comprehension and production. Wray (2002) stresses the nonanalytical nature of formulaic language in native speaker competence. Attempts by teachers and textbooks to encourage the analysis of chunks by learners are, in Wray’s words, “pursuing native-like linguistic usage by promoting entirely unnative-like processing behaviour” (p. 463, emphasis in original). This is certainly the case. However, Spöttl and McCarthy (2004) offer two counterweights to this: (a) There is psycholinguistic evidence that, even among native speakers, at least some degree of literalness or at least metaphoric awareness is retained in the processing of figurative expressions (Gibbs, 1994; Gibbs & O’Brien, 1990). Learners may be even more inclined to analyse and may see it as an important part of the learning process. Receptive mastery may indeed gain from an occasional analytical approach; and (b) One might also add that the more the learner has successfully acquired a repertoire of chunks, the easier it becomes to reflect and analyse them at a later stage, so that certain aspects of grammatical acquisition may flow from the knowledge and use of chunks, rather than vice versa.

It is also worth noting that chunks may not necessarily be acquired in an “all-or-nothing” manner (Schmitt & Carter, 2004, p. 4); in other words, the absorption and learning of the meaning and appropriate use of a chunk may be gradual and only apparent over time and after a number of exposures, just as with grammatical structures or single words.

**Chunks and Writing**

Although the focus in this paper has been on the spoken language and on the relationship between chunks and fluency, it is important to understand that chunks occur in all kinds of written contexts too. Fluent writing is of a different order to fluent speaking (and would require a separate paper), but writing, too, depends on the writer being able to draw on a range of ready-made, off-the-peg items that act as structuring devices, such as prefaces, or as conclusers. In this respect, impersonal passive voice structures with *it* are common and are likely to be recalled and written as whole chunks, for example,
Awareness Raising

But back to speech: The most salient chunks, because of their curiosity and rarity, are the low frequency idioms, and learners often find it easier to recognize these rather than some of the more transparent, high frequency ones. Underlining or colour-highlighting patterns which are frequently repeated in texts and dialogues (e.g., those collected in Carter & McCarthy, 1997) may be one way of raising awareness of useful chunks, and encouraging students to record whole chunks in their vocabulary notebooks may raise awareness of their usefulness as frames that can be used with a potentially large number of utterances. Listening activities are perhaps the best way of raising awareness, especially, since in naturalistic listening passages, common chunks will be spoken rapidly and will punctuate content. Several listenings to the same passage can be carried out--some for content, others purely for noticing chunks.

SUMMARY

The chunks illustrated in this paper show the all-pervasiveness of interactive meaning-making in conversation. The addition of chunks to the vocabulary syllabus is not an optional extra, since their meanings are extremely frequent, necessary, and fundamental to successful interaction. What descriptive and pedagogical lessons should we draw from all this? We offer the following:

- The term *chunks* is preferable to the term *clusters*.
- High frequency chunks are often more frequent than core single words.
- The most frequent chunks, like the most frequent single words, perform core communicative functions in everyday interaction.
- Fluency must involve the ability to call upon a vocabulary of ready-assembled chunks.
- We should not assume, however, that high-frequency chunks should be obligatory components of the learner’s productive repertoire. It may be that receptive mastery is more important than productive repertoire.
- Chunks are chunks: Analysing them and taking them apart may not be of use, and some researchers argue that they should only be processed and retrieved holistically (Wray, 2002).

Conversation materials should, where possible, incorporate useful high-frequency chunks as attested in everyday use (see McCarthy, McCarten, & Sandiford, 2005–2006; O’Keeffe, McCarthy, & Carter, in press; Schmitt, 2006).

A FINAL WORD, OR WORDS, OR CHUNKS

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 on the colourful, the curious, the rare, precisely because such items are psychologically salient. Hence it should not surprise us that, with few exceptions, precorpus studies of multiword units focussed on idioms, phrasal verbs, compounds, and so on, either as colourful curiosities or, in the pedagogic domain, a perverse and difficult characteristic of English for learners to struggle with. Meanwhile the hidden, subliminal patterns of the everyday lexicon stubbornly resisted exposure. Corpus analysis enables us to circumvent many of the difficulties in retrieving such patterned occurrences, but the automatic retrieval of recurrent strings is only the beginning, and a good deal of further analysis, discussion, and classroom-based exploration is still necessary to see meaning in the lists generated by the computer.
REFERENCES


TESOL Symposium on Vocabulary
March 27, 2006
Dubai Men’s College
Dubai, United Arab Emirates

Teachers of English to Speakers of Other Languages, Inc.
700 South Washington Street, Suite 200
Alexandria, VA 22314 USA
Tel: 703-836-0774
Fax: 703-836-7864
E-mail: tesol@tesol.org
http://www.tesol.org/

Copyright © 2006 by Teachers of English to Speakers of Other Languages, Inc. (TESOL)

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, or any informational storage and retrieval system, without permission from the author. Permission is hereby granted for personal use only.