Word Sequences for English Language Learners

Patrick Griffiths

Introduction

This paper discusses sequences of two or more words which happen to occur together often in English usage and their place in EFL (English as a foreign language) syllabuses. The two-word sequence *come on* is used more than 300 times per million words in natural English conversation, which, given average speech rates, means that *come on* tends to occur slightly more often than once per 30 minutes of natural English conversation (Biber, Johansson, Leech, Conrad and Finnegan 1999: 39, 410). *Come on* is used in conversation with three main functions: as an exhortation "Come on, let Andy do it", as a summons to move "Come on, let's go", or meaning 'to start', as in "The heating didn't come on this morning" (authentic examples from Biber et al 1999: 411). Students of English as a foreign language who have learnt basic meanings for the separate words *come* and *on* might not understand these uses of the sequence *come on*. The sequence is therefore a candidate to be taught as a lexical item in its own right.

Language learning strategies are the basis for the recommendations that will be presented towards the end of the paper. The term *language learning strategy* covers the varied activities that have to be undertaken in the long haul of learning a foreign language. (Oxford 2001 provides a concise survey.) TEFL syllabuses should be written to engage with learning strategies. Two categories of learning strategy are important for my purposes:

* 'cognitive' strategies, which are learners' ways of gaining and structuring knowledge of the language

* 'communication' strategies, ways of using their current – and perhaps inadequate – language knowledge and skill to communicate

Frequently-occurring sequences of words are a significant part of the language knowledge that a learner has to gain, so I argue that syllabuses, up to intermediate level, should include controlled inventories of high-frequency sequences. Learners should also be encouraged to explore the extent to which there are patterns even in naturally idiosyncratic combinations like *come on*; for instance, *come* is also found in other frequently occurring 'intransitive phrasal verbs', such as *come off* and *come up*, and *on* in common combinations like *go on* and *get on* (Biber et al 1999: 413). I also argue that, from intermediate level onwards, even relatively infrequent word sequences have to be accepted into teaching materials, to cater for students' need to develop communication strategies. There are almost certainly too many word sequences for them all to be learnt on TEFL courses. It is hard to say how many there are because the total depends on how often a sequence has to occur to be counted as frequent, but Ellis (2003: 69) is not alone in suggesting that there may be hundreds of thousands.

Recurring word-sequences are certainly a feature of authentic usage. Biber et al (1999: 25, 990, 993) report that 25% of the 6½ million words of British and American conversation that they analysed came in the form of three-word sequences (each sequence occurring at least ten times per million running words in the transcripts). Common two-word sequences are more numerous than three-word sequences, and including them would quite likely allow us to say that more than half of natural conversation is made up out of such lexical bundles. A range of examples should be experienced by students, so that they can work out ways of coping with them in communication.

Different types of recurring word-sequence are surveyed next.

Considered after that are some TEFL syllabus implications – essentially ones culled from the applied linguistics literature.

Different kinds of sequence

Labelling the phenomenon

Lexical bundle is the term used by Biber et al (1999: 990) for a sequence of words that occurs frequently and occurs across a range of different texts. Psycholinguists and cognitive theorists concerned with how we use our brains to store, assemble and decode language tend to favour the term *chunk* for the same thing (see for instance the title of Ellis' 2003 paper). Other approximately synonymous terms are *multi-word unit* (Grant and Bauer 2004), *fixed expression, formulaic expression* and *prefabricated expression* ('*prefab*' for short).

The term *lexical phrase* used by Koprowski (2005) is too restrictive for my purposes, because it implies that the sequence forms a syntactic unit, and some of the frequently-occurring sequences that are relevant to learners straddle phrase boundaries, for instance the sequence *I don't know whether*, which Biber et al (1999: 1002) found in conversations at a rate of more than 40 occurrences of per million words, is an incomplete constituent, though (as they note, page 1003) useful as an 'utterance launcher'. A learner who has memorised that sequence can begin an utterance fluently while simultaneously thinking out how to express whatever is going to be said after *whether*.

Collocation, a "statistical tendency of words to co-occur" (Hunston 2002: 12) is a category that includes the terms just listed but is more general, since words which collocate do not necessarily appear in a fixed sequence. The collocation of *break* and *heart* is seen in both of the orders *a broken heart* or *a heart-breaking experience* (where the underlined

sequences would count as two different lexical bundles). The tendency of collocates to occur together can be detected even when they are separated by another word or a few others, as in *be careful not to break your best friend's heart*. To keep the scope of this article manageable it is deliberately restricted to collocations with fixed order, that is to word sequences.

Idioms

Quite a number of languages have borrowed the compound word *hot dog*, for a 'sausage served with onion in a soft roll'. Non-English-speakers who already know *hot dog* in this meaning may be surprised if they begin to study English and find that the meanings of the parts, *hot* and *dog*, do not contribute in any straightforward way to the meaning of the whole expression. For a similar reason, English language students who have learnt the words *fat* and *cat*, need to learn the phrase *fat cats* as a whole item for one of its possible meanings: 'company directors who pay themselves generously', because that sense is not transparently based on the meanings of the two words that make it up.

Grant and Bauer (2004) present an analysis of the notion idiom. An idiom is:

* a sequence of two or more words
* which has to be committed to memory because the meaning of the whole cannot fully be derived from the meanings of the parts
* and cannot be figuratively explained either

There is a cline of idiomaticity. *Hot dog* is almost fully idiomatic. *Fat cat* in the 'self-rewarding director' sense is figuratively motivated, a metaphor from the smug selfishness of a well-fed cat. Although its meaning might not be obvious to someone who (already knowing the component words

fat and *cat*) meets the sequence for the first time, the figurative basis for its interpretation excludes it from the Grant and Bauer's category of idioms.

Irregularities and minor patterns

The adverbial *all of a sudden* exhibits irregularity. It does not fit ordinary syntactic patterns, because an adjective *sudden* appears in a slot normally occupied by nouns. The effort needed to learn the four-word sequence *all of a sudden*, with its meaning 'quickly and unexpectedly', can probably be justified for learners because it recurs fairly often in English conversation (more than 20 times per million words, an average of about one occurrence in every 6 or 7 hours of conversation; Biber et al 1999: 39, 1013). Despite the irregularity, this sequence is somewhat less idiomatic than *hot dog* because the constituent word *sudden* is a pointer to 'suddenness' as part of its meaning.

Rule-governed parts of a language, in principle, greatly reduce the burden on memory. The rules can be used in production and comprehension, instead of memorising each sequence as an item. Clear cases of this can be seen within the structure of words. (Just for this illustration, I shall be dealing with PREFIX + STEM sequences, not sequences of words.) There are at least a couple of hundred English adjectives — some of them frequently used — that consist of the negative prefix *un-* on an adjective or past participle, for example *uncomfortable, unchanged, unusual* (see Griffiths 2004: 15–16, for a list of about 190 past participles prefixed with this *un-*). Negative *un-* accords with a productive rule, which is to say it is freely used to make up new negative adjectives. On the other hand the negative prefix *in-* applies to considerably fewer adjectives. Three words that fit this minor pattern are *inadequate* (23/million), *informal* (24/million) and *invisible* (13/million), with average frequencies of occurrence per million words, from Leech, Rayson and Wilson (2001),

given in brackets. It is only through having the words stored in memory that one can know that *in-* is the prefix conventional employed for these forms. Someone who does not know could use the default pattern to produce *unadequate*, *unformal* or *unvisible*, none of which occur even once in the 100-million-word sample of British English used by Leech et al (see the exhaustive listing at <www.comp.lancs.ac.uk/ucrel/bncfreq>). Many native users of English who heard or read a form such as *unvisible* would judge that the learner had made an error, but – importantly – they would nonetheless understand the word.

Returning to word sequences, I have understood EFL students who used the expressions *play judo*, *play karate*, *play kendo* or *play archery*, even though these sequences do not sound right to me; in my English one *does judo/karate/kendo/archery*. *Play* is not a bad guess to have made for martial arts that are nowadays leisure pursuits, but *play* is the verb for games such as baseball, football, basketball, tennis and chess. *Do + activity* is a minor pattern that, in native-speaker English, is used for talking about martial arts and other activities that have a more 'serious' reputation than mere games.

Phrasal verbs and compounds

Phrasal verbs such as *carry* (SOMETHING) *out* and *find* (SOMETHING) *out*, both of which are reported (Biber et al 1999: 410) as occurring with a frequency of more than 100/million in some registers of English, often have meanings that are to a degree idiomatic. The two just cited can have the meanings 'do' and 'discover', respectively, which are not simply built up from the meanings of the parts. At the beginning of the paper I mentioned the high frequency 'intransitive phrasal verb' *come on*, and three different idiomatic meanings that it commonly can carry. Different phrasal verbs of all kinds occur a lot in native-speaker English: more than 1,800/million words in conversation and more than 1,900/million in

written fiction (Biber et al 1999: 409).

Compound words – for example *supermarket, eye-witness, central heating* and *internet service provider* – are sequences of words that conventionally go together to express meanings that are to some extent idiomatic (for instance, a *supermarket* is not really a 'market that is super' and a paraffin heater placed on the floor in the middle of a room is not *central heating*). Furthermore, in pronunciation, they are usually stressed as if they were single words. As the examples show, they are spelt with and without hyphens, and with and without spaces. Biber et al (1999: 326) in their analysis of 40 million words of English usage found (a) that "There are few really frequent compounds …" and (b) they are particularly common in news reports. The small number of compounds that occur frequently are ones like *bathroom* and *gunfire*, which would tend to be taught as single words anyway. Less frequent compounds will be discussed in a section below concerned with sequences as data from which structural patterns can be induced.

Recurring sequences identified by computer

To linguists nowadays, a *corpus* (plural *corpora*) is a large collection of computer-readable text, usually with numerous annotations, called *tags*, that encode information such as the grammatical category of each word. Careful thought goes into the selection of material for a corpus. With a good corpus, such as the 100-million-word one set up and used by Leech and his colleagues "It is possible to extrapolate from corpus frequencies to inferences about the language as a whole, because the compilers have taken pains to sample different kinds of speech and writing (e.g. conversation, novels, news reporting) broadly in accordance with their representation in everyday language use."Leech et al (2001:1). Comparing the relative frequencies of a wide range of lexical bundles across substantial samples of naturally produced usage was not possible before

the advent of computer corpus studies.

The frequently-occurring word sequences that have been brought to the attention of English language specialists through computer analysis of large corpora are often different in character from most of the examples given in earlier sections of this paper. For one thing, they tend to be low in idiomaticity (Biber et al 1999: 1025): "expressions (such as *what on earth, ... a piece of cake, ... on the double, ...*) are used occasionally as idioms in fiction (generally less than five per million words), [but] they are rarely attested in the other registers". Language teachers seem always to have known about idioms, but apart from their surprise value for sparking the attention of some students who might otherwise have become bored, most such idioms are probably not worth much class time.

We also learn from Biber et al (1999: 995) that "In conversation, only 15% of the lexical bundles can be regarded as complete structural units." One of many examples is *I was going to,* which occurs at a rate of $\geq$ 100/million (Biber et al 1999: 1002). They are not saying that the majority of utterances starting with this sequence were incomplete, but rather that there were different completions in different conversations. Students who have learnt such sequences would be in a position to notice the different sorts of constituent that serve as completions, which amounts to intuitive learning of some patterns of grammaticality in English, a matter that will be taken up again soon.

What use are word sequences for language learners?

Language teachers are traditionally advised — see, for instance, Ur (1996: 60) — that idioms should be taught as part of vocabulary, for the obvious reason that their meanings cannot be (fully) calculated from knowledge of the parts. But we must remember that idioms are actually quite rare, as noted earlier. Frequently-occurring ones must be in the syllabus, but

there are not likely to be many of them.

How frequent is frequent enough to justify inclusion in a course, for words or sequences that have to be learnt as items? Nation (2001: 325) offers a reasoned answer, along the following lines. In English, the 2000 most frequently used single words – each a headword subsuming a cluster of closely related forms – work very hard, for instance they account for around 80% of the words found in average academic texts. Acquiring control of them must be a priority for language learners. Word sequences that occur as often as members of these top 2000 word families should equally be prioritised. The lower bracket of frequency for this is roughly 50 occurrences per million words of running text. *Thank you very much, take place* and *take part* are somewhat idiomatic sequences that come close to this lower boundary (Biber et al 1999: 1007, 1028).

Two important benefits that recurring sequences hold for learners will be examined in the coming two subsections of this paper. Nation (2001: 336) gives the following concise justification:

> The memorisation of unanalysed chunks is an important learning strategy, especially for a learner who wants to quickly gain a degree of fluency in limited areas. It has other learning benefits as well, particularly in that it quickly provides a fund of familiar items that can be later analysed to help support the development of rules.

Sequences promote fluency

A sequence of words that is stored in memory can be produced or recognised as an item without attending consciously to the parts (Nation 2001: 320). This gives to the processing of those stretches of speech or written material the speed and ease that we call fluency.

Biber et al, who documented the prevalence of structurally incomplete sequences in English usage, also report (1999: 995) that "in many cases, the last word of the bundle is the first element of [another phrase]". For instance: *what* in the conversational bundle *I don't know what* (≥ 100/million) is the first word of a *wh*-complement clause; *to* in *you want me to* (≥100/million) is the first word of an infinitival complement; *of* in the academic phrase *in the case of* (≥100/million) is the first word of a prepositional phrase (1999: 1002−3, 1017). Anyone who has spoken in circumstances outside of the most relaxed domestic situations knows that valuable time is gained if one has a reasonably appropriate start for an utterance (or for a major component of an utterance already underway) together with the beginning of its principal constituent. That time can be used in framing the important material that completes the embedded constituent. This is true when operating in one's first language; even more so in a language only part-way through being learnt.

To the extent that an expression is irregular or an idiom it has to be stored as a whole. However, as Ellis (2003: 66−7) points out, expressions that could be put together according to regular patterns can also be memorised. As already noted, Biber et al (1999: 1025) found relatively unidiomatic sequences recurring with high frequency. High frequency of occurrence would facilitate the learning of such sequences, and the contribution that they make to fluency is a sound motive for learning them.

In a survey of vocabulary learning, Carter (2001: 46−7) notes that it has been plausibly argued that reduction of processing effort reduces the stress of communication for learners. This would mean that memorising lexical bundles also contributes to learners' 'affective strategies', their ways of handling emotional issues, such as anxiety over having to communicate in a language not yet confidently controlled (Oxford 2001:

168).

Sequences are a pattern database

Biber et al (1999: 993) report that long bundles often consist of shorter bundles in sequence. This opens two possibilities for learners: they can analyse memorised longer sequences into shorter ones, and they can assemble short ones that they already know into different longer sequences. Assembly from 'prefabricated' parts, of course, contributes to fluency (as well as producing some learner errors when illegitimate combinations are made).

Many sequences that give the appearance of being fixed are flexible at their joints, and this offers learners the possibility of discovering structural patterns in them, for instance *leave a sinking ship* has flexibility over the determiner: besides *a*, it could be *the* or *this*, to mention the most obvious possibilities.

Ellis (2003: 72−4) hypothesises a learning path according to which learners begin with chunks, then extract what he calls 'low-scope patterns' from them and eventually arrive at control over constructions. I shall use two illustrations from the analyses of Biber et al (1999) to illustrate potentially useful low-scope patterns.

While Biber et al (1999) reported that only a few individual compound words appear with notably high frequency, one particular compound structure, a noun modifying a following noun, is a characteristic feature of news reports and fairly common in academic prose as well. There are four very popular nouns in the manifestations of this construction: *government, police, home* and *world*. The frequency of each of these as the initial noun is $\geq$100/million, and each of the four occurs as a modifier on more than 100 different head nouns (1999: 589, 592). Another twelve

nouns, including *city, family, party* and *TV*, similarly act as modifiers to more than 100 different heads, but their separate frequencies of occurrence are a bit lower: ≥50/million, still over the threshold that Nation (2001), as mentioned earlier, sets for items being worthy of attention in class. Learning that these 16 words are high frequency first members of NOUN + NOUN sequences is something that could significantly help learners towards understanding news reports. And once they have experienced quite a lot of different instantiations, they will effectively know the N+N compound construction of English.

Finite complement clauses of the kind that could be marked by *that* are very common in conversation, around 7,000/million (Biber et al 1999: 674, 1010), the vast majority appearing as complement to an immediately preceding verb, such as *think* or *know*. Here we have a pattern useful for understanding or producing conversational speech: PRONOUN + *think/know (that)...* where the three dots can be replaced by a free-standing finite clause of the kind that EFL learners are usually taught at beginner level.

Koprowski (2005), who is persuaded of the need to teach high frequency sequences, apparently does not accept the usefulness of learners being helped to discover the patterns implicit in sequences that they know. Thus he criticises the authors of one textbook for explicitly introducing 25 different completions for sequences starting with *take* and *put* (2005: 329−30), because some of the sequences have low frequencies and/or narrow ranges of occurrence. Apparently he would have included only widely-used high frequency sequences, such as *take part in* (which is one that the book happened to omit). He disapproves of another textbook for introducing 26 sport nouns as objects for the verbs *play, do* or *go*, since his corpus research shows only a few of them (*play football* is one) occur frequently and widely enough to merit inclusion. He would have excluded sequences like *do judo* and *do weightlifting* because their

frequency and range suggests that they are not encountered often enough to be worth teaching.

Certainly it is important to try to ensure that frequently occurring forms are taught, but it is also important for the students' development beyond mere memorisation that they should be encouraged to explore patterning, and I do not see that text frequency has as much relevance for noticing patterns. What matters then is that the joints in the sequences should show up, revealing the construction; and for that the number and variety of substitutes is more likely to be important.

A final point about learnt sequences as a library of patterns is that the predominance of different kinds of sequence helps learners distinguish between usage styles that characterise different domains. For instance Biber et al (1999: 409) supply text frequency counts to show that phrasal verbs are colloquial in tone, validating their statement that "Overall, conversation and fiction show much greater use of the most common phrasal verbs than news and academic prose."

A considerable amount of detailed cataloguing of differences between academic writing and conversation is summarised as follows by Biber et al: "most lexical bundles in conversation are building blocks for verbal and clausal structural units, while most lexical bundles in academic prose are building blocks for extended noun phrases or prepositional phrases." (1999: 992). Examples of the relevant academic prose style sequences are *the nature of the, one of the most, on the basis of* (all three $\geq$40/million, 1999: 1015, 1017).

Which patterns for EFL learners and when?

Swan (2005: 394), in a thoughtful critique of excessively 'communicative' approaches to TEFL, argues for

planned approaches involving, among other elements, careful selection and prioritizing, proactive syllabus design, and concentrated engagement with a limited range of high-priority language elements, so as to establish a core linguistic repertoire which can be deployed easily and confidently.

The present focus is the selection and prioritising of word sequences. Studies of large computer corpora provide a sound empirical base for selecting sequences to receive priority in teaching, and Nation's (2001: 325) approximate cut-off frequency of 50/million has already been mentioned. However, selectiveness and concentration on high priority items comes at the expense of naturalness. Authentic videos, recordings and written texts may not have the desired concentration of the selected high-frequency sequences. Real usage almost invariably includes lots of dilution from low frequency items – individually they occur with low frequency, but there are many different ones (all requiring teacher- or textbook-explanation or student research). Texts will definitely be unnatural if they concentrate on the high frequency sequences and follow Koprowski's (2005: 328) recommendation against the inclusion of low frequency material: "Coursebook writers, as educators, have a professional and pedagogic responsibility to minimize or eradicate the inclusion of questionably useful lexical phrases and, at least, exclude the extremely rare ones."

Restricting myself to word sequences, I agree with Swan – and to an extent with Koprowski – that beginner syllabuses should be directed at establishing basic skill with a core linguistic repertoire. This will entail editing of materials, to increase the concentration of high frequency sequences and sweep aside many real but rare expressions. From intermediate level onwards it seems to me that something different is needed, in two respects:

1) Students must have practice in which they grapple with real or realistic situations and usage – that is with authentic text – to enable them to develop communicative strategies for coping with sequences that they have not met before. (Frequency counts from corpora are still relevant for the teacher or course designer, as a guide on where glossing or explanation may be needed.)
2) From time to time they should also be instructed, with lots of examples (not chosen solely on grounds of frequency and range), in the patterns found in sequences that appear in the communicative tasks that they are working on for (1), to develop structural knowledge and a feeling for textual possibilities that will help them face the endless variety thrown up in real usage.

REFERENCES

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

Ellis, Nick C. (2003) 'Constructions, chunking, and connectionism: the emergence of second language structure', in C.J. Doughty and M.H. Long (eds) *The Handbook of Second Language Acquisition*, Oxford: Blackwell, 63–103.

Carter, Ronald (2001) 'Vocabulary', in R. Carter and D. Nunan (eds) *The Cambridge Guide to Teaching English to Speakers of Other Languages*, Cambridge: Cambridge University Press, 42–7.

Grant, Lynn and Bauer, Laurie (2004), 'Criteria for re-defining idioms; are we barking up the wrong tree?' *Applied Linguistics*, 25: 38–61.

Griffiths, Patrick (2004) The verb and adjective *un-* prefixes of English, 英語英米文学論叢 (Beppu University English studies), 36: 1–18.

Hunston, Susan (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.

Koprowski, Mark (2005) Investigating the usefulness of lexical phrases in

contemporary coursebooks, *ELT Journal*, 59: 322−32.
Leech, Geoffrey, Rayson, Paul and Wilson, Andrew (2001) *Word Frequencies in Written and Spoken English*, Harlow: Pearson.
Nation, Paul (2001) *Learning Vocabulary in another Language*, Cambridge: Cambridge University Press.
Oxford, Rebecca L. (2001) 'Language learning strategies', in R. Carter and D. Nunan (eds) The Cambridge Guide to Teaching English to Speakers of Other Languages, Cambridge: Cambridge University Press, 166−72.
Swan, Michael (2005) Legislation by hypothesis: the case of task-based instruction, *Applied Linguistics*, 26: 376-401.
Ur, Penny (1996) *A Course in Language Teaching: practice and theory*, Cambridge: Cambridge University Press.