CAMBRIDGE

Cambridge Papers in English Language Education

Learning language in chunks



Scott Thornbury



Contents

	Page
Introduction	4
Key terms	5
Key issues	6
Research and literature review: significant findings	7
Key considerations	20
Summary: What are the implications for teachers?	21
Conclusion	23
References	24
Recommendations for further reading plus useful websites	28
Author biography	29

Introduction

It was over 30 years ago that Michael Lewis published The *Lexical Approach* (Lewis, 1993), prompting a radical re-think of the way that we view language – and, by extension, of the way that we teach it.

In contrast to the then prevailing structural account, in which language was viewed as comprising grammatical structures into which single words are slotted, Lewis argued that 'language consists of chunks which, when combined, produce continuous coherent text' (Lewis, 1997: 7).

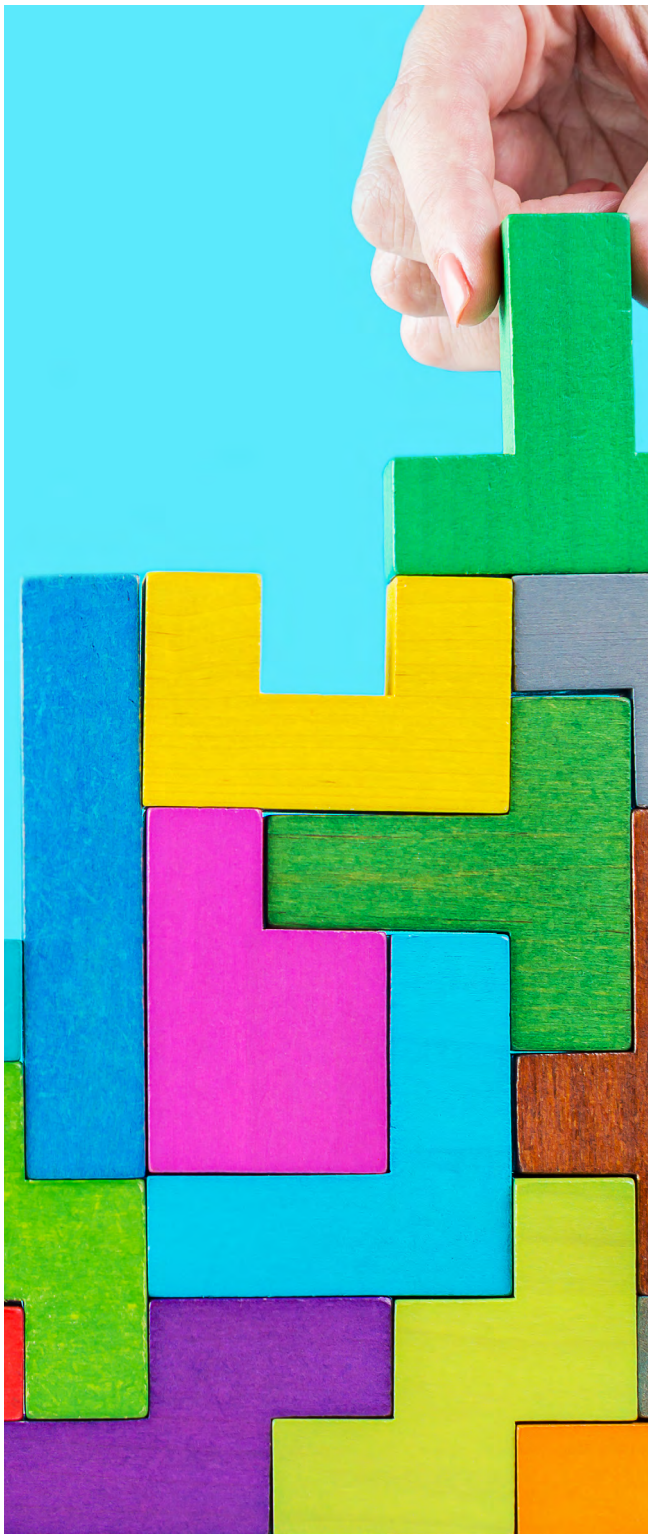
By 'chunks', Lewis was including

- collocations (*wrong way, give way, the way forward*)
- fixed expressions (*by the way, in the way*)
- formulaic utterances (*I'm on my way; no way!*)
- sentence starters (*I like the way...*)
- verb patterns (*to make/fight/elbow one's way to/towards...*)
- idioms and catchphrases (*where there's a will, there's a way; way to go!*)

...everything in fact that doesn't neatly fit into the categories of either grammar (as traditionally conceived) or single-word vocabulary.

Lewis was by no means the first to describe language in these terms: his singular contribution was to argue that, in order to accommodate this alternative description, it was language *teaching* that needed to be reformed – or, indeed, revolutionized. (The strapline for his book was *The state of ELT and a way forward*).

This paper charts the extent to which a Lexical Approach, or 'learning language as chunks', as Lewis and subsequent scholars conceived it, is being applied three decades on, and the research that underpins such an approach.



Key terms

chunk: an all-purpose term for any formulaic sequence (FS), lexical/phrasal expression or multiword item (MWI).

cluster (or ***bundle***): any commonly occurring sequence of words, as identified by statistical measures, irrespective of meaning or structural completeness, e.g. *at the end of the, you know what I*.

collocation: two or more words that frequently occur together, e.g. *false eyelashes, densely populated, file a tax return*.

concordance: short extracts from a corpus generated by a search tool and displayed as lines which position the search word (or words) at the centre of each line, allowing typical patterns, e.g. of collocation, to be observed.

construction: In 'construction grammar' (see, for example, Hilpert 2019), a construction is any conventionalised pairing of a form and a meaning, whether the form is a single word (*stock*), the plural morpheme s (as in *stocks*), a collocation (*stocks and shares*), a multiword idiom (*take stock of; lock, stock and barrel*) or an abstract grammatical pattern, such as the transitive verb + object construction; constructions collapse the traditional distinction between lexis and grammar, and proponents of 'usage-based' learning theories consider them to be the building blocks of language use and language acquisition.

corpus: a database of texts, typically authentic, which is digitally accessible for the purposes of calculating frequency, identifying collocations, etc.

formulaic language: According to Wray (2000: 465), 'a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.'

idiom: an expression whose meaning is not the sum of its individual words, i.e. it is 'non-compositional', e.g. *a wild goose chase, run out of steam, plain sailing*.

idiom principle: the principle whereby 'a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments' (Sinclair, 1991: 110). This contrasts with the 'open choice principle' where the only restraint on word choice is grammaticalness.

idiomaticity: the degree to which a particular wording (which may be one of several possibilities) is conventionalized in the speech community. For example, the standard way of asking a person's age in English is 'How old are you?' and not (the grammatically acceptable but non-idiomatic) 'How many years do you have?' (which is how the Spanish equivalent would be translated).

lexical approach: an approach to language teaching that foregrounds the contribution of vocabulary, including lexical chunks, to language use and acquisition.

lexical phrase: one of many alternative terms to describe multiword items.

MI (Mutual Information): a statistical measure of the strength of a collocation based on the likelihood of its individual elements occurring together more frequently than would be expected by chance: *short + straw*, for example, has a higher MI score than *long + straw*, while *draw + the short straw* has a higher MI score than *choose + the short straw*.

n-gram: a cluster defined in terms of its length, e.g. common 3-word n-grams are *I don't know, a lot of, I mean I* (from O'Keeffe *et al*, 2007: 66).

phraseology: a general term to describe the recurring features of language that are neither individual words nor grammatical structures.

usage-based learning: the theory that argues that language is acquired through exposure to language in use, and internalized as 'a huge warehouse of constructions that vary in complexity and abstraction' (Wulff 2019: 21).

Key issues

Lewis's Lexical Approach was strongly, even fiercely, argued, but was only sketchily supported by evidence. In the intervening years, researchers – directly or indirectly – have been investigating his claims, with a view to answering these key questions (among many others):

- To what extent does language consist of chunks?
- How might the learning of chunks benefit language learning overall?
- What challenges do chunks present to learners?
- How can chunks be integrated into the second language curriculum?
- How are chunks best learned and taught? What materials and activities might support their acquisition?

This paper addresses each of these questions in turn.



Research and literature review: significant findings

1. To what extent does language consist of chunks?

In order to estimate the proportion of spoken or written text that is 'chunk-like', we first need to define 'chunk'. This is easier said than done: the number of terms that are used to capture the phenomenon is bewildering.

One writer (Wray, 2002) listed over 50, but current practice seems to favour, as an (uncountable) umbrella term, *formulaic language*, embracing different types of *multiword items* (MWIs), or what most non-academic texts for teachers refer to simply as *(lexical) chunks*. (Krishnamurthy [2002: 289] prefers the term 'chunk' since, being relatively recent, it has 'less baggage associated with it.') These, in turn, can be subdivided into such overlapping categories as *collocations, lexical phrases, phrasal verbs, functional expressions, idioms*, and so on (see the **Key terms** on page 5 for definitions).

What the items in these categories have in common is that

- they consist of more than one word
- they are conventionalised
- they exhibit varying degrees of fixedness
- they exhibit varying degrees of idiomaticity
- they are probably learned and processed as single items (or 'holophrases')

Word combinations are *conventionalized* if they occur together with more than chance frequency. Corpus linguistics has exponentially enhanced our knowledge of what combinations of words are significantly frequent. Thus, the sequence *no way of [+ -ing]* can be completed with virtually any verb, but only one – *knowing* – is significantly frequent.
(It is more than ten times as common as the next most frequent combination: *no way of telling*). Moreover, according to the *Corpus of Contemporary American English* (Davies, 2008-) it is relatively common across a variety of registers: spoken language, fiction, news, and academic writing.

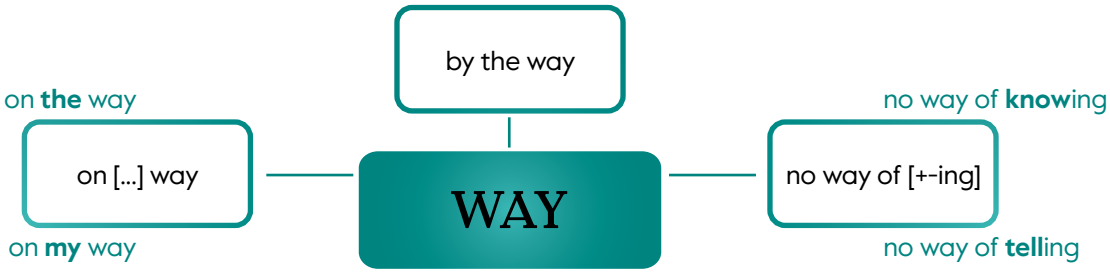


Diagram showing some examples of the fixedness of key phrases used with way.

With regard to *fixedness*, an example of a fixed chunk is *by the way*, which, as a discourse marker, allows no variation: **by a way; *by the ways* . *On the way,* however, allows some variation, as in *on my way. By and large* is an example of a chunk that is not only fixed but also idiomatic, i.e. it is 'non-compositional': its composite meaning cannot be inferred from its individual words. *By the way,* on the other hand, is less idiomatic since – even in its sense of marking a new direction in the discourse – its meaning is relatively easily derived from its parts.

In terms of their psycholinguistic status, that is, the way that they are mentally stored and accessed, there is growing evidence, e.g. from eye-tracking, read-aloud and brain scanning studies (see, for example, Siyanova-Chanturia & Van Lanker Sidtis 2019; Conklin 2020), that, at least in the first language (L1), chunks are processed holistically, rather than as a sequence of individual words. Indeed, Haselow (2017) cites research on neurologically impaired speakers that suggests that 'novel speech' and 'formulaic speech' are processed in different areas of the brain. This may explain the fact that word sequences are articulated differently in terms of their prosody, depending on whether their meaning is literal or figurative, e.g. 'they showed me the ropes', and that their stress and intonation is relatively fixed, e.g. 'my ears were BURNing', but never (when used idiomatically), 'my EARS were burning' (Lin 2019).

The fact that chunks are processed holistically is attributed to frequency effects: the more often a sequence (of morphemes or words) is encountered, the more likely it is that it is represented and retrieved as a single unit (Siyanova-Chanturia & Martinez, 2014). Frequency also accounts for the observation that second-language (L2) speakers process chunks less rapidly than do L1 speakers: 'This has led to the claim that non-native speakers process multiword items more compositionally', i.e. as sequences of individual words, rather than as single units (Conklin, 2020, 178).

Nevertheless, the pervasiveness of formulaic language is a fact of life: one frequently cited estimate is that nearly 60% of spoken language (slightly less of written) is formulaic to some degree. Biber *et al*, (1999: 990) using slightly different criteria, found that 45 per cent of the words in their extensive corpus of conversational English (but only 21% in academic prose) occurred in what they called 'bundles', i.e. 'recurrent expressions, regardless of their idiomaticity, and regardless of their structural status.' Some high frequency four-word bundles in spoken English include *I don't know what, I don't think so, I was going to, do you want to*. On the other hand, they found many fewer instances of idiomatic phrases, of the type *in a nutshell, a piece of cake, fall in love or a slap in the face:* these tend to occur more often in fiction than in spoken language. Idiomatic expressions that do occur with high frequency in both spoken and written language are verb + noun phrase combinations with *have, make* and *take* (as in *have a look, make sense, take time*) and phrasal verbs, especially (in spoken language) those with *come, go* and *put*.

A number of studies have attempted to compute the frequency of chunks compared to that of single words, and have established that there are many chunks that are as frequent as, or more frequent than, the most frequent individual words. In one study (Shin & Nation, 2008), using the 10 million word spoken section of the British National Corpus (BNC), the researchers identified the most frequently occurring collocations, and found that 84 collocations qualified for inclusion in the top 1,000 word band (examples being: *you know, I think, a bit, as well, in fact*...) with increasing numbers of collocations eligible for inclusion in successive bands. In similar fashion, Martinez & Schmitt (2012), using slightly different criteria, identified over 500 'phrasal expressions' that would qualify for inclusion in a list of the top 5000 word families in the BNC, both written and spoken. Examples include: *after all, as soon as, make sure, once again*.[1]

1 The complete list is available by clicking on PHRASE list at https://www.norbertschmitt.co.uk/vocabulary-resources

Nevertheless, these high frequency chunks are the exception, rather than the rule. Thereafter, frequency of occurrence - and hence opportunities for repeated encounters - trails off precipitously. Boers and Lindstromberg (2009), for example, found only one instance of the collocation *tell + truth* in a hundred-page police story. As noted above, such infrequent exposure accounts for the fact that non-native speakers display a limited repertoire of chunk-like language. As Cobb (2019: 196) notes, 'while formulaicity is pervasive in language, particular formulae sadly are not'.

Other researchers have investigated – not just the frequency – but the distribution of lexical chunks (specifically n-grams) in different registers of both spoken and written text. Biber et al (2004) for example, conclude that their patterns of use are not accidental, but that they form the 'building blocks' of discourse and often correlate with specific textual functions, such as expressing the speaker/writer's attitude, or foregrounding new information. As such they represent an important indicator of a text's register as well as a measure of the speaker or writer's command of that register.

2. How might the learning of chunks benefit language learning overall?

There are at least three reasons that have been put forward for prioritizing the learning of lexical chunks, each of which we shall look at in turn:

- they facilitate fluent processing
- they confer idiomaticity
- they provide the raw material for subsequent language development

Fluency: As long ago as 1925, Harold Palmer anticipated the Lexical Approach by posing the question 'What is... the most fundamental guiding principle [to conversational proficiency]?' and he answered it thus: *'Memorize perfectly the largest number of common and useful word-groups.'* (Palmer 1925: 187, emphasis in original). It took another seven decades for Pawley & Syder (1983), in a seminal paper that sought to solve two of the 'puzzles' of native-like proficiency, to reach a similar conclusion: 'Lexicalised sentence stems and other memorized strings form the main building blocks of fluent connected speech' (1983: 214). In other words, the possession of a memorized store of chunks allows more rapid processing, not only for production but also for reception, since 'it is easier to look up something from long-term memory than compute it' (Ellis *et al*, 2008: 376). Or, as another scholar neatly put it, 'speakers do as much remembering as they do putting together' (Bolinger, 1976: 2). In much the same way that artificial intelligence (AI) software is able to predict, on the basis of an enormous data base, what word or words

come next when generating or interpreting a text, we humans are similarly endowed with a capacious memory for frequently recurring sequences, whether sounds, letters, morphemes or words – a capacity that facilitates speed of processing and, ultimately, fluency.

Research that supports this view includes Towell *et al* (1996), who compared the spoken fluency of advanced speakers of French before and after an extended stay in France, and found that the more fluent speakers were able to speak faster and with less hesitation due to the effective use of chunks. And Boers *et al* (2006) conducted a study where two groups of learners were taught the same curriculum but only one group was given 'lexical-phrase oriented pedagogy' as well. Both groups were assessed on a speaking task: the experimental group was generally judged more fluent *and* their fluency correlated with the number of chunks they used. The researchers also noted that the more confidently chunks were used, the more they contributed to the perception of fluency (Boers & Lindstromberg, 2009: 36). In another study, Götz (2013: 171) also found that her subjects' command of formulaic expressions positively impacted perceptions of their fluency, such that she recommends 'a paradigm shift in syllabus design … to make teaching and learning processes lean more towards processes of first language acquisition, namely to teach language in chunks and formulaic sequences rather than stick to the strict divide of grammar and vocabulary exercises.'

Idiomaticity: The possession of a store of formulaic language helps resolve another of the 'puzzles' of native-like proficiency alluded to by Pawley and Syder (1983): that of native-like selection, or idiomaticity, i.e. the capacity a language user has to distinguish 'those usages that are normal or unmarked from those that are unnatural or highly marked' (p. 194). For example, given all the possible ways of performing a particular speech act, such as expressing regret, speakers of English typically choose to say *I'm [very/so] sorry*, rather than, say, *It causes me pain* (as in German *Es tut mir leid*) or *It tastes bad* (as in Spanish *Me sabe mal*), both of which would be grammatically well-formed in English. Wray makes the useful distinction between chunks that are *speaker-oriented*, i.e. that enable fluent production, and those that are *hearer-oriented*, i.e. that achieve social and interactional purposes, such as polite formulae, e.g. *I wonder if you'd mind…?* or expressions that assert group identity, as in 'generation Z talk': *it's giving; secure the bag; touch grass.*[2]

For language learners (and assuming native-like fluency is their goal), knowing how things are done in the target language is clearly an advantage. As Boers and Lindstromberg (2009: 37) note, 'The use of chunks can help students to be perceived as idiomatic language users, disposing of a relatively impressive lexical richness and syntactic complexity.' (Arguably, the emphasis placed on learning and testing phrasal verbs in many ELT courses reflects their iconic status as markers of idiomaticity.)

Nevertheless, the aim of (native-like) idiomaticity may not necessarily be an appropriate one for learners whose goal is to achieve communicative competence in English as an international language or English as a lingua franca (ELF), or in any variety of English that is considered non-standard. As Bokhorst-Heng (2012: 223) notes, the Lexical Approach 'fell short of fully appreciating and incorporating the diversity of the English language into its pedagogy and curriculum,' a weakness that she attributes to 'native-speakerism'.

It may also be the case, as some researchers suggest, e.g. Mauranen (2009), that lingua franca English is evolving its own standards of idiomaticity, in which chunks that might be considered deviant by native speaker standards are becoming conventionalised in ELF: *in my point of view* and *as the matter of fact*, to name just two. Idiomaticity, in the end, may be a moving target.

2 https://en.wikipedia.org/wiki/List_of_Generation_Z_slang

Language development: Another argument in favour of a lexical approach is that 'lexical phrases may also provide the raw material itself for language acquisition' (Nattinger & DeCarrico, 1989: 133). That is to say, the phrases are first learned as unanalyzed wholes. 'Later, on analogy with many similar phrases, [the learners] break these chunks down into sentence frames that contain slots for various fillers' (*ibid.*)

Lewis (1997: 211) himself made a similar point: 'The Lexical Approach claims that, far from language being the product of the application of rules, most language is acquired lexically, then "broken down"... after which it becomes available for re-assembly in potentially new combinations.' Proponents of 'usage-based' theories of language acquisition, such as Ellis (1997: 126), concur: 'Learning grammar involves abstracting regularities from the stock of known lexical phrases.' Wulff (2019: 26) elaborates on this usage-based view, arguing that (first language) learners are sensitive to the frequency and distribution of chunks, and that 'this sensitivity to sequence information in language processing demonstrates learners' implicit knowledge of memorised sequences of language and forms the basis for linguistic systematicity and creativity.'

However, attempts to research these claims have met with mixed results, one problem being the difficulty of deciding if a learner's utterance is the result of holistic learning or of (re-)analysis and subsequent creativity – or of a combination of both.

Wong Fillmore (1979), in a study of five Spanish-speaking learners of English, found evidence that formulaic sequences were gradually analysed, so that their constituent elements were available to be re-combined creatively: 'The formulas the children learned and used constituted the linguistic material on which a large part of the analytical activities involved in language learning could be carried out' (212).

On the basis of this and similar studies (mainly with young learners), Ellis and Shintani (2014: 71) accept that 'the prevailing view today is that learners unpack the parts that comprise a sequence and, in this way, discover the L2 grammar. In other words, formulaic sequences serve as a kind of starter pack from which grammar is generated.'

Other researchers are less convinced. Granger (1998: 157-8), for example, analysed a corpus of adult L2 learner language and concluded that 'there does not seem to be a direct line from prefabs to creative language... It would thus be a foolhardy gamble to believe that it is enough to expose learners to prefabs and the grammar will take care of itself.' Wray (2002: 201) suspects that what might work for young learners does not necessarily work for literate adults, whose inclination is to unpack formulaic expressions – not for their syntax – but for their words, and concludes that 'there is very mixed evidence regarding the effectiveness of teaching formulaic sequences.' Even proponents of usage-based learning theory, such as Wulff (2019: 28) are cautious, citing evidence to suggest that '[second language] learners vary considerably in their ability to decompose prefabricated routines in search of the underlying grammatical rules they exemplify'. Finally, Scheffler (2015) argues that – even if these 'unpacking' processes apply in first language acquisition – the sheer enormity of the input exposure required in order to 'extract' a workable grammar is simply unfeasible in most L2 learning contexts.

Further, it has also been argued that over-reliance on formulaic language may have adverse effects, contributing to the premature stabilization of the learner's developing language system. As Swan (2006: 5-6), for example, puts it, 'Much of the language we produce is formulaic, certainly; but the rest has to be assembled in accordance with the grammatical patterns of language [...]. If these patterns are not known, communication beyond the phrase-book level is not possible.' And Wray (2000: 472) goes so far as to suggest that 'formulaic sequences may be used by some adult learners as a means of actually avoiding engaging with language learning' (emphasis added). That is, they are used as communication strategies at the expense of the development of a productive grammar. However, in a later book on the subject (2002: 188), she is more conciliatory: 'Despite such reservations, it remains highly plausible that formulaic sequences are supporting the acquisition process, whether this be simply by maintaining in the learner a sense of being able to say something, even when there is only a small database to draw on, or by providing a wealth of stored nativelike data for later analysis.'

In short, the jury is still out on the role that chunks play in overall language development, and an exclusive focus on chunks may even be counter-productive, given the risk of over-reliance and consequent fossilization. Nevertheless, their key role in facilitating fluency and, to a lesser extent, idiomaticity, is uncontroversial.

3. What challenges do chunks present to learners?

'For second language (L2) learners, conventionalized recurrent phrases and constructions have almost always presented an area of difficulty' (Hinkel 2019a: 3). Why is this?

For a start, vocabulary learning – whether of single words or chunks – is an enormous challenge for second language learners, especially those who are not fully or partly immersed in the target language. For example, Webb and Nation (2017: 47) report a study in Taiwan that showed that, after nine years of instruction, fewer than half of the learners had mastered the 1,000 most frequent (single) words in English. The problem is exacerbated with the learning of multiword items, which are – on the whole – both less frequent but more numerous ('several hundreds of thousands' according to Pawley & Syder [1983: 213]), hence the chances of incidental learning, through reading and/or listening, are low. Also, because many of them are idiomatic, they are not easily understood and are therefore overlooked or ignored.

In terms of production, learners seem to prefer the less risky strategy of producing utterances 'compositionally', that is, word by word, even if this compromises fluency. Collocations – such as *make a journey* versus *do some sightseeing* – often seem arbitrary, and, for this reason, phrasal verbs are best avoided. Studies of learner talk show relatively few instances of formulaic language, except at the most advanced levels. Keckses (2007: 198), for example, found that less than eight per cent of the total words in recorded conversations on non-native speakers

living in the USA comprised formulaic language: compare this to the 60 percent figure for native speakers, cited earlier. And Jones et al (2018), in a corpus-based study of learners at three different CEFR bands (B1, B2 and C1) found no significant increase in the amount of formulaic language across the levels. They add that 'chunks which were favoured by learners tended to be those which could be used in a variety of speech and with a variety of functions such *as a lot of* and *I agree with you*' (75).

Idiomaticity, as we have seen, presents a real challenge to learners, since there are many ways a chunk can 'go wrong'. Prodromou (2008: 221) notes that in his own 100,000 word corpus of advanced L2 speech 'there are virtually no grammatical errors. There are, however, 40 non-canonical versions of idiomatic phraseology.' His list includes: *from my heart* (for *by heart*), *lift an eyebrow*, (for *raise an eyebrow*), *dining and wining* (for *wining and dining*), and a *streak of good luck* (for *a stroke of good luck*). Moreover, Prodromou found that his native speaker informants were less prepared to regard non-canonical forms produced by learners as being evidence of creativity than had they been produced by native speakers. He concludes, 'as far as idiomaticity is concerned, L1-users are playing at home, with rules that they can bend according to need; L2-users are playing away and if they break the rules they may be penalised' (238).

Little wonder that learners avoid chunks!

4. How can chunks be integrated into the second language curriculum?

(And, specifically, which chunks should be learned when?)

Since the advent of the Lexical Approach, there has been a perceptible shift in ELT coursebook design in the direction of including a more prominent focus on formulaic language. Nevertheless, there is a common perception amongst corpus linguists, in particular, that ELT materials have not kept pace with developments in the field. Hunston (2002: 38), for example, observes that 'unfortunately, phrases tend to be seen as tangential to the main descriptive systems of English, which consist of grammar and lexis.' Wolter (2020: 493) attributes this neglect as being due to inertia: 'In some parts of the world the preferred L2 pedagogy still places a strong emphasis on grammar with little attention paid to vocabulary beyond single words.' In similar vein, Granger & Meunier (2008: 251) complain that 'vocabulary teaching today is still too often exclusively word-based.' Where chunks, such as collocations, are included in course materials their selection appears somewhat arbitrary, 'largely grounded on the personal discretion and intuition of the writers' (Koprowski, 2005: 330). The problem is, perhaps, that, as Durrant (2019: 224) ruefully admits, 'we do not yet have a satisfying means of identifying formulas for teaching.'

In an attempt to rectify this situation, a number of researchers have proposed criteria for selecting lexical chunks for inclusion in language teaching curricula. These criteria include

- utility
- frequency
- fixedness
- idiomaticity (vs transparency)
- teachability

Utility: With regard to selecting chunks on the basis of their utility, or usefulness, one legacy of the functional-notional syllabuses of the 1970s has been the inclusion of functional language in the form of formulaic expressions associated with different speech acts, such as asking directions, making requests, etc. This was the direction advocated by Nattinger (1980: 342), an early proponent of a focus on chunks: 'Since patterned phrases are more functionally than structurally defined, so also should be the syllabus... In that way, the items we select to teach would not be chosen on the basis of grammar but on the basis of their usefulness and relevance to the learners' purpose in learning.'

While functions are no longer the primary organizing feature of mainstream courses, course designers are now better equipped in terms of identifying the most common exponents of different functions. *The Functional Language Phrase Bank by Cambridge University Press & Assessment,* for example, is compiled on the basis of corpus data, and typical exponents of common functions, such as giving advice, offering help or expressing agreement are tagged for frequency. Likewise, the *Academic Phrase Bank*[3] (2023) is organized into typical functions used in academic discourse, both spoken and written, such as *Describing trends, Explaining causality, Giving examples*. And the practice of teaching useful classroom language in the form of fixed expressions (*What does X mean? How do you say Y? Can you spell that?* etc.) is consistent with this functional imperative.

Frequency: When Palmer (1925, quoted earlier) recommended memorizing *'the largest number of common and useful word-groups'*, there were little means – apart from intuition – of establishing which 'word groups' were more common than others. With the advent of corpus linguistics, this problem has largely been solved, although the focus of corpus-based frequency research was initially on individual words. Significant breakthroughs in terms of ranking multiword items in terms of frequency include Biber et al. (1999) sampling of the most frequent lexical 'bundles' in the British National Corpus; Liu's (2003) list of most frequently used chunks in spoken American English; Shin and Nation's (2007) list of the most frequent collocations; Simpson-Vlach & Ellis's (2010) frequency list for academic English; Martinez and Schmitt's (2012) 'phrasal expression' (or PHRASE) list, and Garnier and Schmitt's (2015) Phrasal Verb Pedagogical (PHaVE) list. Both this and the PHRASE list are freely accessible online[4]. As Schmitt describes the latter (on his website), 'The list consists of the 505 most frequent non-transparent [i.e. idiomatic] multiword expressions in English, intended especially for receptive use.' The PHaVE list catalogues the 150 most frequent phrasal verbs, reducing the total number of meanings to under 300. More recently still, the Academic Phrasebank (mentioned above) is also available online[5].

Here, for example, are the top 20 three-word chunks (technically 'n-grams') extracted from the five-million word CANCODE corpus of spoken English (from O'Keeffe et al 2007):

	item	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

3 https://www.phrasebank.manchester.ac.uk/

4 https://www.norbertschmitt.co.uk/vocabulary-resources

5 https://www.phrasebank.manchester.ac.uk/

While all this frequency information is useful, its applications to syllabus design are less obvious. For a start, the compilers of these frequency lists are using different definitions of 'chunk' as a criterion for selection, as well as different measures of frequency. Moreover, many chunks – particularly phrasal verbs – are polysemous, i.e. they have more than one meaning: e.g. *make up, take off*. And, of course, lists lack contextual information, as well as information about the variability or fixedness (see below) of each chunk. But at least such lists allow both teachers and materials writers to check their intuitions in terms of which chunks are best candidates for inclusion in the curriculum.

Fixedness and idiomaticity: As a basis for selection, frequency, however, is problematic: as Boers & Lindstromberg (2009: 14) point out: 'Below the small group of highly frequent chunks, frequency distribution rapidly levels off, confronting the learner [and the teacher and the coursebook author] with a great many chunks of medium frequency.' In order to choose between these, they recommend enlisting criteria of fixedness and of idiomaticity. They argue that chunks which are relatively fixed in their form (such as *first and foremost, by leaps and bounds*) are, once learned, easier to deploy as a single item, and therefore more facilitative of productive fluency. Although, as Hinkel (2019a: 6) points out, 'to date, dozens of corpus analyses have largely established that, in fact, most idiomatic expressions and multiword units are variable in their forms and, to some extent, in their meanings.'

With regard to meaning, expressions which are idiomatic – hence semantically opaque – such as *every so often* and *by and large* – are likely to cause problems in comprehension, and therefore should be prioritized over those chunks that are relatively transparent.

In similar fashion, Martinez (2013) advocates selecting chunks on the basis, not just of frequency, but also of transparency (or lack thereof). Thus, an expression such as take time is both frequent and transparent – and, as such, probably doesn't merit detailed attention. On the other hand, an expression such as *take place* (meaning 'occur') is frequent but not transparent, and is therefore likely both to impede comprehension, and to be under-used in production. Hence it merits an instructional focus. Similar principles inform the PHRASE list of Martinez and Schmitt, mentioned above.

In compiling their 'Academic Formulas List', Simpson-Vlach & Ellis (2010) used both quantitative and qualitative data in order to select those chunks that, in an academic register (both spoken and written), have 'high currency and functional utility.' Items were initially derived from a corpus based on raw frequency of occurrence plus their MI score (a statistical measures of the probability of co-occurrence). Twenty experienced EAP teachers were then asked to a rate a sub-set of these items according to their 'formula teaching worth': i.e. how formulaic, how functional and how teachable they were. The results were extrapolated to the larger set of items to produce a definitive list, which was then sub-divided into functional categories, such as quantity specification (*a list of, all sorts of*), cause and effect (*as a consequence; the reason why*), and vagueness markers (*and so on, blah blah blah*).

Teachability: While 'teachability' is a somewhat elusive criterion, Boers and Lindstromberg (2009) argue that idiomatic expressions can be made more memorable, and hence more teachable, once their 'mnemonic potential' has been unlocked through teacher 'elaboration'. For example, when learners understand the sporting references that are encoded in expressions like *jump the gun, neck and neck,* or *on the ball,* they are more likely to remember them.

Likewise, drawing attention to the commonly used phonological repetition in expressions like *make-or-break, short and sweet, fair and square, time will tell*, etc. can enhance their memorability – 'for although these chunks have considerable mnemonic potential, relatively few learners will unlock it without prompting or guidance' (Boers & Lindstromberg 2009: 123). Other approaches to helping make chunks more

'learnable' include asking learners to make drawings of phrasal verbs (White 2012) or to act them out (Littlemore 2011) so as to make their idiomatic or figurative meanings more accessible. All things being equal then, it makes sense to select those chunks that are not only relatively frequent but are also teachable, in the sense that their mnemonic potential can be unlocked.

Other criteria for selection of lexical chunks that have been proposed include 'prototypicality' (Lewis 1997: 190) and 'generalizability.' On the assumption that memorized chunks provide the 'raw material' for the development of the second language grammar, then there is a case for choosing to teach chunks that embed prototypical examples of the target grammar. For example, the classroom questions (mentioned earlier) *What does X mean? How do you spell it?* etc. are potentially analysable into instances of *do*-auxiliary use which can be generalizable to the creation of new questions. This is also an argument for *not* teaching idiomatic expressions that are 'non-canonical', i.e. that do not reflect current usage, such as *come what may, long time no see, once upon a time*, etc. This may also be the thinking behind Nattinger & DeCarrico's (1992: 117) argument that 'one should teach lexical phrases that contain several slots, instead of those phrases which are relatively invariant.'

To summarise: given that there are many more combinations of words than individual words, and given the learning load that it is thereby implicated, the inclusion of lexical chunks into the curriculum requires rigorous and principled decisions with regard to selection and sequencing – decisions that are not always at the forefront when it comes to course planning and design.

5. How are chunks best learned and taught?

Having selected chunks for a pedagogical focus, how should that focus be implemented? Responses to this question have tended to fall into five groups, which, while not mutually exclusive, may be summarised as:

1. The phrasebook approach
2. The awareness-raising approach
3. The data-driven-learning approach
4. The analytic approach
5. The communicative approach

1. The phrasebook approach

Phrasebooks for travellers, of course, have always accepted the usefulness of memorizing set phrases (with or without unfilled slots) for specific situations, without the need for any explicit analysis. A phrase book approach to the learning of chunks makes similar assumptions. Organizing chunks according to their specific functions and contexts of use (e.g. greetings, service encounters, classroom language etc.) may help make the learning load manageable.

Moreover, if chunks are going to serve to lubricate fluent speech in real time, they need to be readily and accurately retrievable from long-term memory. Hence, as with all vocabulary learning for production, an element of deliberate memorization is essential. Nattinger himself was not averse to the idea that techniques associated with (by then discredited) audiolingualism, such as pattern practice drills, might be rehabilitated for the purposes not only of committing lexical phrases to memory but of modelling their generative potential: 'Pattern practice drills could first provide a way of gaining fluency with certain fixed basic routines...The next step would be to introduce the students to controlled variation in these

basic phrases with the help of simple substitution drills, which would demonstrate that the chunks learnt previously were not invariable routines, but were instead patterns with open slots.' (Nattinger & DeCarrico, 1992: 116-17).

More recently, proponents of skill-acquisition theory (e.g. Suzuki 2019) are not averse to activities that promote automatization, such as imitation and memorization. After all, any classroom activity that both increases the frequency of exposure and helps draw learners' attention to a language item must be beneficial.

Less behaviouristic, perhaps, is the technique of 'shadowing', whereby the learner listens to extracts of authentic talk, and 'subvocalises' at the same time. For younger learners, preselected chunks can be embedded in chants and songs – see, for example, Carolyn Graham's (1979) work on 'jazz chants'.

To summarise: the practical applications of the phrasebook approach might be:

- organizing chunks functionally and/or situationally
- rote learning of formulaic expressions
- drilling
- shadowing
- jazz chants

2. The awareness-raising approach

In formulating his Lexical Approach, Lewis showed no particular allegiance to any existing theory of second language learning, although he often makes reference to Krashen's (1985) Input Hypothesis and the necessity for high quantities of roughly-tuned input as a source for learning. Hence, in place of the dominant PPP [present-practise-produce] methodology of the time, he offers OHE (observe-hypothesise-experiment), an inductive, awareness-raising approach, predicated on the learners noticing common sequences in the input. Lewis calls this process 'pedagogical chunking' (1997: 54) but, like many instructional approaches that promote discovery learning, OHE leaves quite a lot up to chance. As Boers (2021: 102) points out:

> A practical problem with asking learners to identify phrases in texts is how learners can be expected to recognise phrases unless they are already (somewhat) familiar with them. After all, any word sequence could theoretically be a conventional multiword expression, and it may be naïve to expect L2 learners to have reliable intuitions about which sequences are more likely candidates than others.

Accordingly, Hinkel (2019b: 112) adopts a more proactive approach to awareness-raising, in which the teacher's role is:

1. To bring learners' attention to word combinations as they occur in context.
2. To focus on more valuable, frequent, and useful collocations and to avoid spending time on those that are low frequency.
3. To provide learning strategies and techniques to help learners remember and practise collocations.

Characteristics of both the learner-centred and the teacher-fronted approach to awareness-raising include:

- extensive reading and listening tasks, preferably using authentic material
- 'chunking' texts, i.e. identifying possible chunks, and checking these against a collocation dictionary, or by using online corpora (e.g. COCA: Davies, 2008) to check the relative frequency of word sequences[6]
- using online tools, such as the Academic Collocation List Highlighter (Smith 2022),[7] that identifies known collocations in texts
- listening to extracts of authentic speech and marking a transcript into tone units in order to identify likely chunks
- record-keeping and frequent review
- recycling chunks in learners' own texts, either spoken or written

6 For ways learners can be helped to identify chunks see Thornbury 2019: https://www.cambridge.org/elt/blog/2019/11/01/chunk-spotting-users-guide/

7 https://www.eapfoundation.com/vocab/academic/acl/highlighter/

3. The data-driven learning approach

More recently, the processes of 'observe and hypothesise' have been facilitated by learner access to online corpora and corpus analysis tools, and what is called *data-driven learning* (DDL). This is really an extension of the previous, awareness-raising, approach but exploits digital tools that were not readily available in the 1990s, when the Lexical Approach was first mooted. As Green (2019, p. 135) describes it, 'learners use authentic language data, presented in the form of concordance lines, to act as "linguistic detectives" as they find answers to their language questions'. Cobb (2019) argues persuasively that such use of corpus data makes up for the fact that many lexical chunks are infrequently repeated and/or widely dispersed in extended texts and therefore not easily acquired by reading or listening alone: 'only computationally assembled data can provide this exposure' (p. 197). Boers (2021: 104) warns, however, that, in using a DDL approach, 'it is … important to make a distinction between effectiveness (i.e, how much learning occurs) and efficiency (i.e., the time invested to obtain good learning outcomes).'

Characteristics of this approach include:

- guided study of concordance lines, using worksheets, in order to identify the collocations or other contingent features of a word or chunk
- using software that detects and marks up formulaic language in texts
- using artificial intelligence (AI) tools to proofread learner-produced texts
- using digital flashcard tools such as Anki or Quizlet to store and review lexical chunks

4. The analytic approach

While Boers and Lindstromberg (2009) agree with Lewis that class time should be devoted to raising awareness about the role of chunks, they are sceptical as to whether learners will be able to identify chunks unaided. Moreover, their own research supports the view that directing learners' attention to the compositional features of chunks ¬ – e.g. their metaphorical origin or their phonological repetition – can 'unlock' their mnemonic potential and hence optimise their memorability (see above). They summarise their more analytic approach in these terms:

> (1) *teach chunks* instead of relying on learner-autonomous, incidental chunk-uptake owing to awareness-raising alone; (2) *select* chunks for targeting not just on the basis of frequency but also on the basis of evidence of collocational strength and 'teachability'; (3) in order to improve the chances of retention, *complement noticing* by also encouraging elaboration of meaning and form; and (4) *reveal non-arbitrary properties* of chunks to make them more memorable (p. 126).

To the above list might be added the use of contrastive analysis and translation, i.e., comparing chunks in the target language with their direct translations in the mother tongue (see Laufer & Girsai 2008), and, as mentioned earlier, the use of visualization, using drawing or gesture, to unpack 'hidden' meanings. The practical application of these principles is spelled out in more detail by Lindstromberg and Boers, in their activity book, *Teaching chunks of language: from noticing to remembering* (Lindstromberg & Boers, 2008).

Typical activities include:

- targeted teaching of selected chunks, e.g. those that are derived metaphorically from a particular 'source domain', such as seafaring: *give someone a wide berth; take the helm,* etc.
- noticing patterns of sound repetition, as in *short and sweet, take a break,* etc,
- use of memory-training techniques, such as mnemonics and visualization
- frequent recycling and review

5. The communicative approach

Finally, coming from a background in communicative language teaching, with a focus on fluency in particular, Gatbonton & Segalowitz (2005) build on their earlier work on 'creative automatization' by proposing an approach (called ACCESS) that incorporates stages of controlled practice of formulaic utterances embedded within communicative tasks. Put simply, this involves an initial stage in which short chunks of functional language are presented and practised, before learners take part in an interactive task that requires the repeated use of these chunks in order to fulfil a communicative purpose.

A well-known task-type that fits this format is the 'Find someone who' activity: learners have to survey one another according to prompts that require the use of a lexical phrase with an open slot, e.g. *Have you ever...?* which might first be highlighted and practised in advance of task performance. Gatbonton & Segalowitz (2005: 338) sum up the rationale for their version of the Lexical Approach in these terms:

> The ultimate goal of ACCESS is to promote fluency and accuracy while retaining the benefits of the communicative approach. In ACCESS, this is accomplished by promoting the automatization of essential speech segments in genuine communicative contexts.

A similar approach was explored by Wray & Fitzpatrick (2008), in which intermediate/advanced learners predicted conversational situations they were likely to encounter, and then worked with native speakers to script and rehearse these conversations, including the incorporation of formulaic language. The researchers report that the subjects found that the use of memorized sentences in anticipated conversations [was] a liberating experience, because it gave them exposure to an opportunity to sound nativelike, promoted their fluency, reduced the panic of on-line production in stressful encounters, gave them a sense of confidence about being understood, and provided material that could be used in other contexts too (p. 143).

To summarize, the practical applications of a communicative approach to teaching chunks might include:

- the use of survey-type activities and guessing games that involve the repetition of formulaic expressions
- repeating speaking tasks, e.g. with different partners and/or to a time-limit, in order to encourage 'chunking' - see Bygate (2018) for more on the value of task repetition
- scripting, rehearsing, memorizing and performing dialogues that include formulaic expressions

Key considerations

Whatever procedures we adopt, it is worth bearing in mind that chunks are really just 'big words', and hence most – if not all – of the principles of effective vocabulary teaching apply equally to the teaching of chunks as they do to the teaching of individual words. These include taking into account

- the distinction between teaching for production (i.e. speaking and writing) as well as for recognition (listening and reading);
- the necessity of focusing on meaning as well as on form;
- the importance of teaching vocabulary in context, rather than as isolated items;
- the need for the deliberate (or intentional) teaching of vocabulary, rather than relying solely on incidental learning (Webb & Nation 2017);
- the value of having learners make decisions about the items they are learning, according to the principle that 'the more one engages with a word (deeper processing), the more likely the word will be remembered for later use' (Schmitt 2000: 121);
- the importance of learners forming associations with vocabulary items in order to establish mutually-reinforcing semantic networks as an aid to memory;
- the necessity of regular review, including 'spaced repetition,' i.e. reviewing previously learned material at increasingly larger intervals of time;
- the value of having learners make their own decisions as to what and how they learn vocabulary, including setting their own targets and measuring their success at meeting these.



Summary: What are the implications for teachers?

The current state of play with regard to the four questions posed above suggests a number of implications with regard to classroom teaching. In brief these are:

- 'One of the future challenges for teachers will be to help learners become aware of the pervasiveness of phraseology and its potential in promoting fluency' (Granger & Meunier, 2008: 248). Fortunately, there are now a number of relatively recent titles that provide practical ideas for doing so: Lindstromberg & Boers (2009), Dellar & Walkley (2016), and Selivan (2018).
- At the level of selecting and sequencing of chunks, teachers may need to be more systematic and more rigorous, taking into account not only frequency, but such factors as utility, idiomaticity, fixedness, generalizability, and teachability.
- At beginners and elementary levels, chunk learning should take the form of the formulaic ways that certain common speech acts – such as making requests, apologizing etc – are realised.
- Teachers should be encouraged to exploit the texts they use not only for their grammatical content, or their single-word vocabulary items, but also for any lexical chunks and patterns that fulfil the selection criteria listed above.
- Familiarity with online corpus tools, particularly those that provide collocational data, should be encouraged (see below for a list of recommended sites).
- Teachers should train learners in strategies for identifying possible chunks in the input that they are exposed to, as well as strategies for recording and reviewing them, and for re-integrating them into their output.
- For teachers of younger learners, in particular, the design and use of activities such as songs and chants so as to maximise their chunk-learning potential, should be promoted.
- Teachers of specialized course, such as English for academic purposes, should foreground and highlight those formulaic features of different registers and genres, including commonly used chunks, that are typical.
- Chunks that have been identified as useful or worth learning should be 'put to work'. As Ur (2019: 33) advises, 'it is useful to do activities that not only consolidate basic knowledge of the items, but also require students to probe further into meaning and use, to analyze, compare, contrast, and so on.'
- Testing of both spoken and written language should include criteria that address the candidates' command of formulaic and/or idiomatic usage.

At the same time, a balance needs to be maintained with other components of the curriculum. As Granger & Meunier (2008: 251) summarise it:

> Phraseology is a key factor in improving learners' reading and listening comprehension, alongside fluency and accuracy in production. However, its role in language learning largely remains to be explored and substantiated and it should therefore not be presented as the be-all and end-all of language teaching. Teachers have to do a "delicate balancing act" which consists in exposing learners to a wide range of lexical strings while at the same time ensuring that they are not overloaded with them and are also able to abstract key concepts and useful rules of grammar.

A further caveat relates to learners' communicative needs. As mentioned earlier, for learners of English as an international language, the time spent trying to achieve native-like idiomaticity through the acquisition of a 'warehouse' of memorised chunks, might be better spent on other aspects of the language. Moreover, as Wray (2019: 256) somewhat drily notes, 'learners experiencing little genuine practical interaction in the L2 will not need many communicative FSs [formulaic sequences].'



Conclusion

The ubiquity and usefulness of lexical chunks is generally well accepted, but how chunks should best be integrated into existing teaching approaches is less clear: we have seen five different 'schools of thought', and it may be the case that the judicious selection of techniques from each approach may be the wisest course. And that implementing the principles of effective vocabulary teaching applies equally well to the teaching of chunks as it does to the teaching of individual words.

References

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

Biber, D., Conrad, S., & Cortes, V. (2004). "If you look at ...": Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3): 371-405.

Boers, F. (2021) *Evaluating second language vocabulary and grammar instruction*. Routledge.

Boers, F., Eyckmans, J., Kappel, J., Stengers, J. & Demecheleer, M. (2006). 'Formulaic sequences and perceived oral proficiency: putting a Lexical Approach to the test'. *Language Teaching Research* 10(3): 245-61.

Boers, F. and Lindstromberg, S. (2009). *Optimizing a lexical approach to instructed second language acquisition*. Houndmills: Palgrave Macmillan.

Bokhorst-Heng, W.D. (2012) Lexical innovation in English as an International Language: Implications for English teaching. In Alsagoff, L. McKay, S.L. Hu, G., & Renandya, W.A. (eds.) *Principles and practices for teaching English as an international language*. Routledge.

Bolinger, D. (1976). Meaning and memory. *Forum linguisticum*. 1, 1-14.

Bygate, M. (ed.) (2018) *Language learning through task repetition*. John Benjamins.

Cobb, T. (2019) From corpus to CALL: The use of technology in teaching and learning formulaic language. In Siyanova-Chanturia, A. & Pellicer-Sánchez, A. (eds) *Understanding formulaic language: a second language acquisition perspective*. Routledge.

Conklin, K. (2020) Processing single-word and multiword items. In Webb, S. (ed.) *The Routledge Handbook of Vocabulary Studies*. Routledge.

Davies, M. (2008-) *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Available online at https://www.english-corpora.org/coca/

Dellar, H. & Walkley, A. (2016). *Teaching lexically: Principles and practice*. Peaslake: Delta.

Durrant, P. (2019) Formulaic language in English for academic purposes. In Siyanova-Chanturia, A. & Pellicer-Sánchez, A. (eds) *Understanding formulaic language: a second language acquisition perspective*. Routledge.

Ellis, N. C. (1997). Vocabulary acquisition: word structure, collocation, word-class, and meaning. In Schmitt, N. & McCarthy, M. (eds) *Vocabulary: Description, Acquisition and Pedagogy*, pp. 122-39. Cambridge: Cambridge University Press.

Ellis, N.C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42(3): 375-396.

Ellis, R. & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research*. London: Routledge, p. 71.

Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20, 29-62.

Garnier, H. & Schmitt, N. (2015) The PHaVE List: a pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, 19, 645-666.

Gatbonton, E. & Segalowitz, N. (2005). Rethinking communicative language teaching: a focus on access to fluency. *Canadian Modern Language Review* 61(3): 325-53.

Götz, S. (2013) *Fluency in native and nonnative speech*. John Benjamins.

Graham, C. (1979). *Jazz chants for children*. Oxford: Oxford University Press.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie (ed.) *Phraseology: Theory, analysis and applications*, pp. 145-160. Oxford: Oxford University Press.

Granger, S. & Meunier, F. (2008). Phraseology in language and teaching: Where to from here? In Meunier, F. & Granger, S. (eds) *Phraseology in foreign language learning and teaching*, pp. 247-52. Amsterdam: John Benjamin.

Green, B. (2019) Teaching and assessing multiword expressions using an open educational resources academic English corpus. In Hinkel, E. (ed.) *Teaching essential units of language: Beyond single-word vocabulary*. Routledge.

Haselow, A. (2017) *Spontaneous spoken language: An integrated approach to the emergent grammar of speech*. Cambridge University Press.

Hilpert, M. (2019) *Construction grammar and its application to English (2nd edn)*. Edinburgh University Press.

Hinkel, E. (2019a) A short preamble: a bit of history in the life of phrases and multiword units. In Hinkel, E. (ed.) *Teaching essential units of language: Beyond single-word vocabulary*. Routledge.

Hinkel, E. (2019b) Teaching strategies and techniques: Collocations and multiword units. In Hinkel, E. (ed.) *Teaching essential units of language: Beyond single-word vocabulary*. Routledge.

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

Jones, C., Byrne, S. & Halenko, N. (2018) *Successful spoken English: Findings from learner corpora*. Routledge.

Keksces, I. (2007) Formulaic language in English lingua franca. In Keksces, I. & Horn, L. (eds.) *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects*. Mouton de Gruyter.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal*, 59(4), 322-332.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. Harlow: Longman.

Krishnamurthy, R. (2002). Language as chunks, not words. *JALT Conference Proceedings*. Tokyo: JALT.

Laufer, B. & Girsai, N. (2008) Form-focused instruction in second language vocabulary learning: A case for contrastive analysis. *Applied Linguistics, 29*, 694-716.

Lewis, M. (1993). *The Lexical Approach*. Hove: Language Teaching Publications.

Lewis, M. (1997). *Implementing the Lexical Approach*. Hove: Language Teaching Publications.

Lin, P. (2019) Formulaic language and speech prosody. In Siyanova-Chanturia, A. & Pellicer-Sánchez, A. (eds) *Understanding formulaic language: a second language acquisition perspective*. Routledge.

Liu, D. (2003) The most frequently used spoken American English idioms: a corpus analysis and its implications. *TESOL Quarterly,* 37, 671-700.

Lindstromberg, S. & Boers, F. (2008). *Teaching chunks of language: from noticing to remembering*. Helbling Languages.

Martinez, R. (2013). A framework for the inclusion of multi-word expressions in ELT. *ELT Journal*, 67(2): 184-198.

Martinez, R. & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics* 33(3): 299-320.

Mauranen, A. (2009) 'Chunking in ELF: expressions for managing interaction.' *Journal of Intercultural Pragmatics* 6(2): 217-33.

Nattinger. J. (1980). A lexical phrase grammar of ESL. *TESOL Quarterly*, 14(3): 337-44.

Nattinger, J. & DeCarrico, J. (1989). Lexical phrases, speech acts and teaching conversation. In Nation, P. & Carter, R. (eds) *AILA Review 6: Vocabulary Acquisition*. Amsterdam: AILA.

Nattinger, J. & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Palmer, H. (1925) Conversation. Re-printed in Smith, R. (1999) *The Writings of Harold E. Palmer: An Overview*. Tokyo: Hon-no-Tomosha.

Pawley, A. & Syder, F. (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Richards, J., & Schmidt, R. (eds) *Language and communication*, pp. 191-227. Harlow: Longman.

Prodromou, L. (2008) *English as a lingua franca: A corpus-based analysis*. Continuum.

Scheffler, P. (2015). Lexical priming and explicit grammar in foreign language instruction. *ELT Journal*, 69(1): 93-6.

Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Selivan, L. (2018). *Lexical grammar: Activities for teaching chunks and exploring patterns*. Cambridge: Cambridge University Press.

Shin, D. & Nation, P. (2008). Beyond single words: the most frequent collocations in spoken English. *ELT Journal*, 62(4): 339-348.

Simpson-Vlach, R. & Ellis, N.C. (2010). An academic formula list: New methods in phraseology research. *Applied Linguistics*, 31(4): 487-512.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, A. & Martinez, R. (2014). The idiom principle revisited. *Applied Linguistics* 36(5): 249-69.

Siyanova-Chanturia, A. & Van Lanker Sidtis, D. (2019) What online processing tells us about formulaic language. In Siyanova-Chanturia, A. & Pellicer-Sánchez, A. (eds) *Understanding formulaic language: a second language acquisition perspective*. Routledge.

Suzuki, Y. (2019) Skill acquisition theory: Learning-to-use and usage-for-learning SLA. In McManus, K. (ed.) *Usage in second language acquisition: Critical reflections and future directions*. Routledge.

Swan, M. (2006). Chunks in the classroom: let's not go overboard. *The Teacher Trainer*, 20(3): 5-6. Reprinted in Swan, M. (2012). *Thinking about language teaching: Selected articles 1982-2011*, pp. 114-18. Oxford: Oxford University Press.

Thornbury, S. (2019) Chunk spotting: a user's guide. At https://www.cambridge.org/elt/blog/2019/11/01/chunk-spotting-users-guide/

Towell, R., Hawkins, R. & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics* 17 (1), 84 – 119.

Ur, P. (2019) Teaching multiword units: A teacher's perspective. In Hinkel, E. (ed.) *Teaching essential units of language: Beyond single-word vocabulary*. Routledge.

Webb, S. & Nation. P. (2017). *How vocabulary is learned*. Oxford: Oxford University Press.

White, B. (2012) A conceptual approach to the instruction of phrasal verbs. *Modern Language Journal*, 96, 419-439.

Wolter, B. (2020) Key issues in teaching multiword items. In Webb, S. (ed.) *The Routledge Handbook of Vocabulary Studies*. Routledge.

Wong Fillmore, L. (1979). Individual differences in second language acquisition. In C. Fillmore, D. Kempler & W. Wang (eds), *Individual Differences in Language Ability and Language Behavior*, (pp. 203-228. New York: Academic Press.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics* 21(4): 463 – 489.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2019) Concluding question: Why don't second language learners more proactively target formulaic sequences? In Siyanova-Chanturia, A. & Pellicer-Sánchez, A. (eds) *Understanding formulaic language: a second language acquisition perspective*. Routledge.

Wray, A. & Fitzpatrick, T. (2008). Why can't you just leave it alone?: Deviations form memorized language as a gauge of nativelike competence. In Meunier, F. & Granger, S. (eds) *Phraseology in foreign language learning and teaching*, pp. 123-48. Amsterdam: John Benjamin.

Wulff, S. (2019) Acquisition of formulaic language from a usage based perspective. In Siyanova-Chanturia, A. & Pellicer-Sánchez, A. (eds) *Understanding formulaic language: a second language acquisition perspective*. Routledge.

Recommendations for further reading plus useful website

There are three books written expressly for teachers that develop Michael Lewis's original ideas in different, but practical ways:

Dellar, H. & Walkley, A. (2016). Teaching lexically: Principles and practice. Peaslake: Delta.

Lindstromberg, S. & Boers, F. (2008). Teaching chunks of language: from noticing to remembering. Helbling Languages.

Selivan, L. (2018). Lexical grammar: Activities for teaching chunks and exploring patterns. Cambridge: Cambridge University Press.

And the following collection of papers represents a state-of-the-art and evidence-based overview of key issues on the subject of formulaic language by key researchers in the field:

Siyanova-Chanturia, A. & Pellicer-Sánchez, A. (eds) (2019) Understanding formulaic language: a second language acquisition perspective. Routledge.

Websites:

BNCLab: https://bnclab.lancs.ac.uk/query

A user-friendly corpus site, based on the British National Corpus, and which focuses on a variety of sociolinguistic variables, including gender, age, and region.

Corpus of Contemporary American English: https://www.english-corpora.org/coca/

The 'flagship' site of a suite of corpora allowing free searches for words and phrases, including frequency and collocational information.

Norbert Schmitt's website: https://www.norbertschmitt.co.uk/vocabulary-resources

An applied linguist's website, which houses – among other things – the PHRASE list of the 505 most frequent 'non-transparent' multiword expressions in English.

SkELL: hhttps://skell.sketchengine.eu/

An open-access version of 'Sketch Engine' which easily allows you to check how words and phrases are used by real speakers of English.

Author biography

Scott Thornbury is a teacher educator and methodology writer. Until recently, he taught on an MA TESOL program run from The New School in New York. Two of his latest books (for Cambridge) are *30 Language Teaching Methods* and *66 Essentials of Lesson Design*. For the last twenty years, he has been the series editor of the Cambridge Handbooks for Language Teachers.

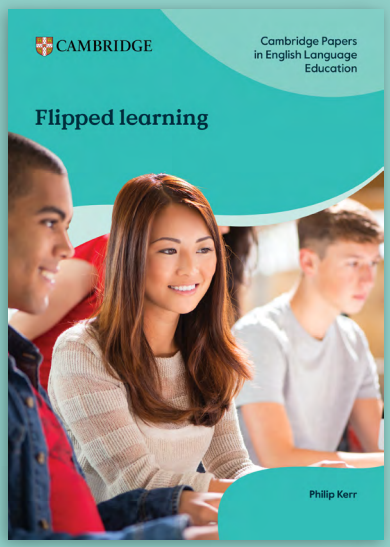


Explore more

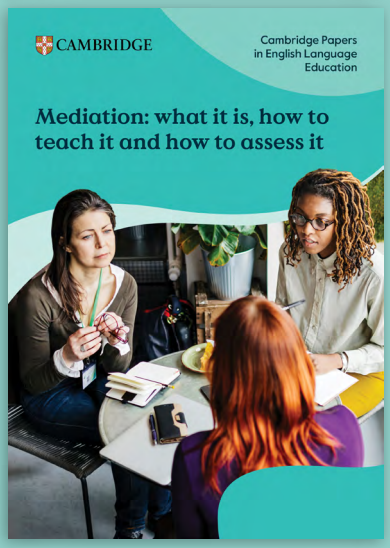
Cambridge Papers

in English Language Education









cambridge.org/english-insights



Find out more at
cambridge.org

We believe that English can unlock a lifetime of experiences and, together with teachers and our partners, we help people to learn and confidently prove their skills to the world.

Where your world grows

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

All details are correct at the time of going to print in July 2026.

© Cambridge University Press & Assessment 2026