Estudios de
lingüística inglesa aplicada

elia

FORMULAIC LANGUAGE: FIXED AND VARIED

Norbert Schmitt
University of Nottingham

Formulaic language has been shown to be an important component of language usage. This paper summarizes evidence of this importance and then goes on to focus upon two key characteristics of formulaic language: fixedness and variability. Formulaic language is usually conceptualized as being basically fixed, but examples are given to illustrate that in many cases formulaic language contains a considerable amount of variation. The degree and type of variation depends on which kind of formulaic language is being addressed: idiom, variable expression, or lexical bundle. Idioms, which are supposedly fixed, show the greatest amount of variation, while variable expressions and lexical bundles seem to contain much more stable fixed cores. It is suggested that variable expressions (and perhaps lexical bundles) may be stored in the mind as individual units, because there are relatively few instances to store. Conversely, idioms may involve so many variants that only the canonical form may be stored as a template, from which truncated and novel forms can be recognized. Teaching implications of these different forms of storage are considered.

Key words: formulaic language, multiword units, language processing, lexical storage

1. The Importance of Formulaic Language

One of the current ‘hot’ areas in applied linguistics is the study of formulaic language. It is becoming increasingly clear that it is an important element of language learning and use, in ways outlined over the years by Pawley and Syder (1983), Nattinger and Decarrico (1992), Moon, (1997), Wray (2002), and Schmitt and Carter (2004), among others. The main reasons why we should be interested in formulaic language are summarized as follows:

● Normal discourse, both written and spoken, contains large percentages of formulaic language. Erman and Warren (2000) calculated that 52-58% of the language they analyzed was formulaic, and Foster (2001) came up with a figure of 32% using different procedures and criteria.

● If much discourse is made up of formulaic language, then this implies that proficient language users know a large number of formulaic expressions. Pawley and Syder (1983: 213) suggest that the number of "sentence-length expressions familiar to the ordinary, mature English speaker probably amounts, at least, to several hundreds of thousands.” Jackendoff (1995) concludes from a small corpus study of spoken language in a TV quiz show that people may know at least as many formulaic sequences as single words. It must be said however, that there is little hard research yet to either support or refute these assertions.

● Formulaic language is not a homogeneous phenomena, but rather quite varied. Schmitt and Carter (2004) illustrate this diversity:

> … formulaic sequences can be long (*You can lead a horse to water, but you can't make him drink*) or short (*Oh no!*), or anything in between. They are commonly used for different purposes. They can be used to express a message or idea (*The early bird gets the worm* = do not procrastinate), functions ([*I'm*] *just looking* [*thanks*] = declining an offer of assistance from a shopkeeper), social solidarity (*I know what you mean* = agreeing with an interlocutor), and to transact specific information in a precise and understandable way (*Wind 28 at 7* = in aviation language this formula is used to state that the wind is 7 knots per hour from 280 degrees). They realize many other purposes as well, as formulaic sequences can be used for

> most things society requires of communication through language. These sequences can be totally fixed (*Ladies and Gentlemen*) or have a number of 'slots' which can be filled with appropriate words or strings of words (_[someone/thing, usually with authority]_ *made it plain that* _[something as yet unrealized was intended or desired]_). (p. 3)

● Moreover, formulaic language is used to realize a number of different functions in language use, including:

Functional use. There are recurring situations in the social world that require certain responses from people. These are often described as *functions,* and include such (speech) acts as apologizing, making requests, giving directions, and complaining. These functions typically have conventionalized language attached to them, such as *I'm (very) sorry to hear about* ____ to express sympathy and *I'd be happy/glad to* _______ to comply with a request (Nattinger and DeCarrico, 1992). Because members of a speech community know these expressions, they serve a quick and reliable way to achieve the related speech act.

Social interaction (phatic communion). People commonly engage in 'light' conversation for pleasure or to pass the time of day, where the purpose is not really information exchange or to get someone to do something. Rather, the purpose is social solidarity, and people rely on non-threatening phrases to keep the conversation flowing, including comments about the weather (*Nice weather today*; *Cold isn't it*), agreeing with your interlocutor (*Oh, I see what you mean*; *OK, I've got it*), providing backchannels and positive feedback to another speaker (*Did you really?*; *How interesting*). Research has shown that such phrases are a key element of informal spoken discourse.

Discourse organization. Formulaic phrases are a common way to signpost the organization of both written (*in other words, in conclusion*) and spoken discourse (*on the other hand, as I was saying*).

Precise information transfer. *Technical vocabulary* are words which have a single and precise meaning in a particular field (*scalpel* is a specific type of knife used in medicine). But this phenomena is not restricted to individual words. Indeed, fields often have extended phraseology to transact information in a way which minimizes any possible misunderstanding. For example, in aviation language, the phrase *Taxi into position and hold* clearly and concisely conveys the instructions to move onto the runway and prepare for departure, but to wait for final clearance for takeoff.

● The use of formulaic language helps proficient speakers be fluent. Pawley and Syder (1983) suggest native-speakers have cognitive limitations in how quickly they can process language, but they are also able to produce language seemingly beyond these limitations. They look at the psycholinguisitc literature and conclude that native speakers are unable to process a clause of more than 8-10 words at a time. When speaking, they will speed up and become fluent during these clauses, but will then slow down or even pause at the end of these clauses. Presumably these pauses permit the speaker to formulate the next clause. Speakers seldom pause in the middle of a clause. Together, this evidence suggests that speakers are unable to compose more than about 8-10 words at a time.

On the other hand, native-speakers can fluently say multi-clause utterances. Consider the following examples:

1) You don’t want to believe everything you hear.

2) It just goes to show, you can’t be too careful.

3) You can lead a horse to water, but you can’t make him drink.

They have increasingly more words, and Example 3 is clearly beyond the limit of 8-10 words. Yet native speakers can say them all without hesitation. Pawley and Syder suggest that these examples can be fluently produced because they are actually already memorized, that is, as prefabricated phrases which are stored as single wholes and are, as such, instantly available for use without the cognitive load of having to assemble them on-line as one speaks. Pawley and Syder suggest that the mind uses its vast memory to store these prefabricated phrases in order to compensate for a limited working memory (and the capacity to compose novel language on-line).

Overall, these points illustrate that formulaic language is intrinsically connected with functional, fluent, communicative language use.

2. Multi-Word Units: How Fixed?

I believe the above summary shows that formulaic sequences are an important element of language. If this is accepted, then as language specialists we need to know how these sequences behave. A considerable amount of work has been done on formulaic language (the best survey is Wray, 2002), but this research has tended to be widely scattered across a number of fields (child L1 acquisition, psychology, corpus linguistics), and often been of peripheral interest to the researchers (e.g., in child L1 acquisition, the focus is the newly uttered word strings themselves, not the fact that they are formulaic in nature). This diffusion is illustrated by the wide variety of terminology Wray (2002, p. 9) found for the various sorts of formulaic language:

chunks
collocations
conventionalized forms
formulaic speech
formulas
holophrases
multiword units
prefabricated routines
ready-made utterances.

To further our knowledge of formulaic language, we can either look at their *psycholinguistic* aspects, for example, how formulaic language is processed and acquired (see Schmitt 2004 for this perspective), or we can look at what might be called the *linguistic* aspects, e.g., the forms which are used and their characteristics. It is this second approach I would like to pursue in this paper.

Moon (1997) is one of the scholars who has looked at the forms of *multiword units (MWU)* (her preferred term) in detail.[1] She focuses on three features of MWUs:

1. Institutionalization the degree to which a multiword item is conventionalized in the language

2. Fixedness — the degree to which a multiword item is frozen as a sequence of words

3. Non-compositionality — the degree to which a multiword item cannot be interpreted on a word-by-word basis, but has a specialized unitary meaning.

(Moon, 1997, p. 44)

Of these features, perhaps fixedness is the most important, as the whole idea behind MWUs is that at least some of their components are fixed, allowing them to be memorized and used as wholes, rather than being newly created for each use. On first inspection, one might assume that all MWUs are completely fixed, but this is not the case. Of course some are, and idioms are usually cited as examples. For instance, corpus evidence shows that *once in a blue moon* occurs almost exclusively as that exact phrase, and not as variations such as **twice in a blue moon, *once in a yellow moon,* or **once in a blue time*. In other words, if we want to use an idiom to express the notion ‘something which occurs very infrequently’, we can only use the intact idiom *once in a blue moon*, not some variation. Another example of a totally fixed MWU is the warning phrase *Watch Out!* It is instantly recognizable, precisely because it is fixed, and little processing should be required to understand it. We could shout something like *Watch the car coming behind you!*, but if milliseconds count, then a shorter, more conventionalized warning is likely to be most effective.

However, many MWUs are not completely fixed, and in fact allow for a surprising amount of flexibility. Moon (1997, p. 53) illustrates this:

not touch someone/something with a bargepole (British vs. American
not touch someone/something with a ten foot pole English)

burn your boats (varying a lexical component)
burn your bridges

cost an arm and a leg (verb variation)
pay an arm and a leg
spend an arm and a leg
charge an arm and a leg

every cloud has a silver lining (truncation)
silver lining

break the ice (transformation)
ice-breaker
ice-breaking

In fact, it seems that once a MWU becomes well-known in a speech community, it can be creatively adapted and still be comprehensible. It is worth expanding on Moon's example of truncation. The well-known idiom *Every cloud has a silver lining* occurs in the 100-million word corpus 7 times. But the phrase *silver lining* occurs in the corpus 75 times representing the same meaning. Clearly people prefer to use a shortened version of the idiom in practice, often in highly creative ways:

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

Beyond idioms, it is not surprising that many other types of MWUs also contain variation, for it is an advantage in much of language use to allow more flexibility of meaning. For example, if we wish to express the notion that some activity or achievement is unusual, unexpected, or exceptional, then we can use phrases like *Diane thinks nothing of running 5 miles before breakfast* or *He thinks nothing of driving 100 miles per hour on the freeway.*

The underlying structure to these sentences is '_____ *thinks nothing of* _____', which allows the flexibility to express the 'unexpected' notion in many different situations.

In this paper I will expand upon this idea of variation in formulaic language, partially because the more I work with formulaic language, the more variation I find. But more importantly, I think this variation may have important implications for how we theorize that formulaic language is stored and processed. This in turn has implications for how we may best go about teaching MWUs. I will attempt to draw out these implications in the final two sections of the paper.

3. Variation in Formulaic Language

3.1. Grammatical and Lexical Variation

I will start the exploration of variation in formulaic sequences with the notion of grammatical variation. I took an idiom which on the face of it appears fixed: *stand shoulder to shoulder.* In all of the corpus analyses in this paper, I referred to the Longman Corpus, a 100-million word corpus based primarily on the British National Corpus. I found that the expected 'canonical'[1] form *stand shoulder to shoulder* occurred 11 times, such as the following example:

-[…] *where the grizzled heroes finally stand shoulder to shoulder.*

As one might expect, anytime a phrase has a verb in it, that verb is likely to change its inflection according to the tense. Therefore, we find cases of simple present, past, and continuous forms (number of instances in the corpus are in parentheses):

[1] The canonical form is the most standard form, and thus the one someone is most likely to know. With idioms, the canonical form is likely to be the full idiom, rather than a truncated variant.

-*While France stands shoulder to shoulder with Germany* […] (1)
-*Now the trees were fenced with armed men standing shoulder to shoulder.* (9)
-[…] *for we stood shoulder to shoulder with the Omanis in their struggle.* (15)

In addition to this kind of grammatical variation, there are also cases of different word choices which do not change the meaning to any great degree:

-*He and I fought shoulder to shoulder against appeasement.* (3)
-[…] *as they worked shoulder to shoulder in a school bus-size laboratory.* (2)

These are only a few examples, but the point is that corpus evidence clearly shows that

a) formulaic sequences often have variation in tense
b) formulaic sequences often have variation in lexical choice of one or more constituents.

This variation is ubiquitous, but it would be a mistake to think that it affects all types of formulaic language equally. One of the problems with most research and discussion into formulaic language is that it is treated as one homogenous set. In reality, there seem to be a number of different kinds of MWU, and each category is likely to used and even processed in somewhat different ways. It would not be surprising that variation would affect each type differently. To illustrate this, I will examine three different types of formulaic language which have varying degrees of fixedness: idioms, variable expressions, and lexical bundles.[2]

[2] In this paper, I shall use the terms formulaic language, multiword unit (MWU), and chunk interchangeably as cover terms for any kind of language that is formulaic in nature. However,

3.2. Idioms and Variation

To explore how idioms can vary and to what degree, I first chose the idiom *scrape the bottom of the barrel*. This phrase has three content words: *scrape*, *bottom*, and *barrel*. In order to explore the variation in the phrase, I broke the phrase into various components to see how each element would vary. First I searched for *scrape the* to see how the later parts of the phrase would vary. As expected there were instances of the full canonical form:

-*The company evidently had to scrape the bottom of the barrel for material*. (5)

There were also cases were the normal order of words was transposed:

-*I began to scrape the theoretical barrel-bottom.*

But more common were instances were some of the key elements were elided:

-[…] *the poor buggers scrape the barrel; the whole of their midfield couldn't* […]

-*Even to produce that list, he'd had to scrape the barrel a bit.*

-*He thought also of his own daughter who was making him scrape the bottom of his pocket.*

I next explored the variation at the beginning of the idiom by searching for *bottom of the barrel*. I found 16 cases in the corpus, of which 13 carried the meaning of 'low quality'.

the terms idiom, variable expression, and lexical bundle are used to describe particular categories of formulaic language.

-*This is really scraping the bottom of the barrel.*

-*But now we're down to the bottom of the barrel.*

-*Being a grunt, you were like the bottom of the barrel.*

Of the 13, only 5 had a version of *scrape*, so once again we see that component largely elided. If we take another idiom, *packed like sardines*, we find similar results. There are 5 instances of this idiom in the Longman Corpus:

-[…] *turtles are packed like sardines into more tanks.*

-*Packed like sardines, the motley crowd of tourists* […]

-[…] *literally packed like sardines in a box.*

There are 3 cases of the variant *packed in like sardines*:

-*Everyone was packed in like sardines and she was quite unable to move.*

-*They offer us a form of transport where people are packed in like sardines* […]

Beyond the above two variations, there are a number of different ways in which the idea of 'sardines = no space' is realized:

-[…] *because refugees were crammed like sardines on mattresses in their offices* […]

-*We are all squashed in here like sardines*

-We were downstairs, laid out like sardines under the Morrison shelter.

Interestingly, *packed together like sardines,* one of the forms cited in the *Oxford Learner's Dictionary of English Idioms* (1994) does not occur in the corpus.

So what are we to make of the analysis of *scrape the bottom of the barrel* and *packed like sardines*? Rather than being idioms with fixed forms, we find an amazing amount of variation. With all this variation, it seems more reasonable to ask which components are fixed, rather than which are not. For *scrape the bottom of the barrel*, it seems that only any two of the three content words *scrape, bottom,* and *barrel* are necessary, and they can be in any order. For *packed like sardines,* it seems *sardines* is the key word, together with one of a number of verbs realizing the meaning of 'tightly packed', such as *packed* or *crammed.* In sum, it seems the English speech community uses many variants of these idioms, with little of either one being frozen in terms of being absolutely required, or even existing in a certain position in the idiom.

3.3. Variable Expressions and Variation

We now turn to another kind of formulaic language, variable expressions. This is Sinclair's (2004) term for a phrase which has some fixed elements and some semantically-constrained 'slots'. One common example which occurs 125 times in the corpus is:

_______ *think nothing of* _______

I searched for *think nothing* and found that in 125 out of the 170 cases of this expression (74%), the preposition *of* was included as part of the string, indicating that it should indeed be considered as part of the canonical form, although the last two examples below show other variants.

-*She thinks nothing of going out at ten o'clock at night.*

-*He thought nothing of playing in 10 or 11 consecutive events.*

-[…] *adolescents capable of subduing the earth around them and thinking nothing of it* […]

-*Your average person in the States thinks nothing about going to Bali.*

-*Alan Beith thinks nothing to striding round five villages.*

As these examples show, other than the grammatical inflection of *think* and alternative prepositions in 25% of the instances, the 'fixed' element of this expression is actually quite stable. This makes sense, because there are two slots in this expression, and if the fixed components were not there as a reliable anchor to the expression, it would lose its holistic nature and become uninterpretable. One can reasonably think of this fixed string as the *core* of the expression, around which the flexible elements are added. In essence, the variation in these expressions resides in the slots, not in the core.

Another frequent variable expression (1344 instances) is

_____ *made it clear that* _____

-The United States has made it clear that the country can expect no further help.

-That means making it clear that it will not allow anyone to steal the election.

-Melville makes it clear that he is a "rugged individual"

Again we find the core component relatively fixed, but here it is frequently modified with specifying adjectives. In 200 out of the 1344 cases, adjectives like *quite, very, absolutely, perfectly, abundantly, pretty, fairly, painfully, crystal,* and *explicitly* boost the intensity of the expression:

-The meeting made it crystal clear that Carter was determined to go ahead.

-Bush made it abundantly clear that he thought the US's economic might […]

This analysis suggests that the fixed elements of variable expressions are actually more fixed than the fixed elements of idioms.

3.4. Lexical Bundles and Variation

Another type of formulaic language is the recurring strings of words identified by corpus analysis. These strings have been given various names (e.g., *sentence stems*), but are best known as *lexical bundles* following Biber *et al.*'s (1999) extensive discussion in Chapter 13 of the *Longman Grammar of Spoken and Written English.* Lexical bundles are extended collocations—bundles of words with a tendency to occur together. They are identified by using a concordancer to isolate the words which occur in multi-word sequences a minimum number of times. For example, four-word sequences needed to occur at least ten times per million words in order to be considered a lexical bundle by Biber and his colleagues. I arbitrarily chose a couple of lexical bundles from the *Longman Grammar of Spoken and Written English* to analyze for variation. The first lexical bundle I examined was *have a look at*, which occurred 756 times in the corpus.

-Let's have a look at your discovery.

-Let's have a look at what happens when we [...]

Biber *et al.*'s methodology of asking the computer to look only for word sequences which occurred in exactly the same form in the corpus means that they did not capture any of the potential variation of the bundles. To explore whether *have a look at* allows variation, I searched the corpus for this string, but used a wildcard in place of the content word *look*. I found 1297 cases of *have a X at*, but several 100 line samples produced no substitute word which means "look". It seems that this bundle has no common variant which substitutes for content word *look*. Doing a search with *X a look at* produced 510 cases of *take a look at*. This could either be interpreted as a very common variant of *have a look at,* or *take a look at* could be considered as a separate bundle in its own right. Either way, there are very few variations besides these two main forms, indicating that this lexical bundle seems to be relatively fixed.

I did the same kind of analysis with *it should be noted that*, which occurred in its canonical form 546 times in the corpus.

-*It should be noted that a few of them have reversed the process.*

When I searched for the string with a wildcard in the place of *noted*, there were 1091 cases, which means that in almost exactly half of the cases there was a substitute word for *noted*. Some of the most common substitutes are:

-*It should be remembered that it is impossible to anticipate every minor detail.* (137)

-*It should be emphasized that there was no criticism of the other volunteers.* (68)

-*It should be stressed that income is only one factor in determining consumption.* (56)

-*It should be clear that the attitude of workers is determined by many forces.* (39)

-*It should be recognized that many of the problems that face us* […] (38)

There were many other words occurring in the wildcard slot, including *said, added, obvious, recalled,* and *apparent*. In this lexical bundle, there is a great deal of variation in this 'content word' slot, although in about half the cases, *noted* is used, making it far and away the most typical form for expressing the notion of 'highlighting' inherent in this bundle.

I next used a wildcard to check the variation in the modal of this bundle *it X be noted that. Should* took this slot in 546 cases; a number of different modals made up the other 160 cases.

-it may be noted that	(54)
-it must be noted that	(49)
-it will be noted that	(37)
-it might be noted that	(12)
-it can be noted that	(5)

However the form of *be* seems stable; when searching for *it should X noted that,* in the 546 instances, only the form *be* was used.

In sum, it seems that lexical bundles are not always fixed in the sense that they are the only form which can impart a certain meaning. Even here there is variation, although this probably depends a great deal on the individual bundle. If a bundle has a modal verb, it is likely to allow other modals; also, some bundles seem to allow variation in content words while others do not. For example, in the 3-word lexical bundle *I want to, want* can easily be replaced by *wish* or *like*, but in the bundle *the number of,* it is difficult to think of any content word that could replace *number* and mean the same thing (*amount* and *degree* would change the meaning somewhat). However, the bottom line is that lexical bundles do contain variation.

4. How Do Proficient Speakers Store and Process the Variation in Formulaic Language?

4.1. Recognizing Idioms

From the above discussion we have seen that there is variation in multi-word units. Although only a few examples were given, that fact that variation was so easy to find suggests that many MWUs contain such variation. More research is needed to establish this, but I suspect that variation is the norm, with most or perhaps even all MWUs containing variation of some kind. If this is the case, then it raises interesting questions about how MWUs with variation are stored and processed by the mind. The 'holistic storage' theory, where each MWU is stored as an individual memorized chunk, is the commonly espoused view. This approach seems to make good sense as long as the MWUs are intact, unchangeable wholes, but runs into problems if variation is inserted. For example, how are novel variations of the canonical form recognized? We know that proficient users are creative with MWUs, and once a MWU is established in a speech community, it is often truncated or the order of the main components switched around. To illustrate this, I have contrived some new versions of well-known idioms, which do not occur at all in the Longman corpus, and as far as I know, are completely unique. Nevertheless, although you almost certainly have not seen these forms before, I am fairly certain that you will be able to understand them well enough to answer the questions.

-I hated taking the subways in Japan. The sardine-like train cars always made me sick.

What exactly is the complaint about the Japanese trains?

-Wasting time in meetings drives me crazy. The worst are the bush-beaters.
-It would be better if they were expelled immediately.

How do the bush beaters waste time in a way that drives the speaker crazy?

Sardine-like refers to the idiom *packed in like sardines*, and thus the complaint is that the subways are far too crowded. *Bush-beaters* is related to *beat around the bush*, which means that the speaker dislikes the way that such people talk in circles and never get to the point. If you were able to make these connections and catch the meaning, then you were able to interpret these completely novel MWUs, even though you have never seen these forms before, and even though the forms themselves are quite dissimilar to the underlying canonical form. The problem this presents for holistic storage is obvious; these forms were not previously stored in your mind, yet you were in all likelihood able to interpret them. It follows that some other process is necessary to help the mind make this type of connection.

In considering how this process might operate, it is useful to know just how minimal the formal connections need to be between a novel form and the canonical form, or to state the issue another way, just how little of a MWU needs to be given in order to interpret it. We can explore this in the following contrived example involving an idiom.

My friends and I went out dancing last Saturday night. We first went to the XENON club, and were having a great time. But one of my friends almost got into a fight with a group of very big guys and we just managed to escape __ ___ ____ __ ___ _____. After that we just felt like going home and having a quiet pizza.

Can you fill in the blanks and understand the complete meaning? It is highly unlikely. But let us see what happens when we fill in parts of the idiom. First let us insert some of the function words. Can you make out the idiom now?

-But one of my friends almost got into a fight with a group of very big guys and we just managed to escape __ the ____ of our _____. …

When I have tried this informally with students and conference attendees, they were largely unable to recognize the idiom. However, many more recognized it when the first function word was added.

-But one of my friends almost got into a fight with a group of very big guys and we just managed to escape by the ____ of our _____. …

So it seems that knowledge of idioms can be strong enough that it can be recognized without any content words actually being given. But if content words are given, then the idiom appears to be much easier to recognize. Would you be able to recognize it with only the following content word present?

-But one of my friends almost got into a fight with a group of very big guys and we just managed to escape __ ___ ____ __ ___ teeth. […]

How recognizable would the MWU be with the first content word given?

-But one of my friends almost got into a fight with a group of very big guys and we just managed to escape __ ___ skin __ ___ _____.

In this example, it seems that content words are the key prompts to recognition, as one might expect. Content words carry referential information, and are much less frequent than function words. However, if an idiom starts with a number of function words, an idiom may well be recognized before a content word appears. This suggests the intuitive idea that words at the beginning of the idiom are relatively more important for recognition, in the same way that the cohort model posits that the first letters of a word are important for its recognition (Marslen-Wilson and Tyler, 1980). I didn't look at the possibility here, but recognition may well be best prompted by a certain combination of content words, in this case *skin* and *teeth*. There is no way to judge the relative importance of these different prompt factors at the moment, but once we have a clearer idea of exactly what prompts the recognition of MWUs, we will be much closer to understanding how they are processed.

4.2. Recognizing Variable Expressions

As we have seen before, various kinds of MWU behave differently when it comes to variation. This suggests that it would be a mistake to treat all kinds of MWU the same. Let us try the same exercise as above with variable expressions, to explore what the minimal elements of recognition are for this type of MWU. In this variable expression, the variable slot is filled with *my skills*.

I took my pilot's flying examination yesterday. The test pilot was very rigorous and really ___ my skills __ ___ ____. But luckily my instructor prepared me well and I managed to pass.

Nobody I have tried this with has been able to figure out the expression from only from context, and very few can recognize the expression with only function words inserted.

-The test pilot was very rigorous and really ___ my skills to the ____.

Likewise, giving the first word, the delexicalized verb *put*, helped few people.

-The test pilot was very rigorous and really <u>*put*</u> *my skills __ ___ ____.*

Providing the content word in the expression, *test*, did not seem to help most people either.

- The test pilot was very rigorous and really ___ my skills __ ___ <u>*test*</u>*.*

Overall, it seems that there is no single element of the variable expression *put _____ to the test* which reliably allows recognition by itself. Rather, for most people, all of the fixed elements of the expression need to be present for recognition. This is congruent with earlier discussion that the fixed elements are necessary to provide the scaffold for the slot, with the implication that these fixed elements must be present for recognition. This is highlighted even further by frequency figures. The idiom *by the skin of [my, his, her* etc.*] teeth* only occurs 32 times in the Longman Corpus, but could be recognized in context with only certain elements being present. Conversely, *put _____ to the test* occurred around 270 times, yet is virtually unrecognizable without all of its fixed elements being present.

4.3. Recognizing Lexical Bundles

Now we will try the same thing with a third type of MWU: lexical bundles.

So, if the Consumer Price Index climbs 5 percent, their bosses may be obliged to give them 5 percent raises to maintain their standard of living. Though many people focus on ___ _______ __ ___ CPI survey, the truth is that it's a sloppy system.

You almost certainly cannot guess this bundle from the context, so let us fill in the function words.

-Though many people focus on <u>*the*</u> *_______* <u>*of*</u> <u>*the*</u> *CPI survey, the truth is that it's a sloppy system. …*

-You are probably still having trouble recognizing the bundle, so let us insert the content word.

-Though many people focus on ___ results __ ___ CPI survey, the truth is that it's a sloppy system.

With this content word, you can probably get the gist of the meaning in this context, but may well not be able recognize the bundle *the results of the*. In this case, although the whole bundle is basically fixed, it is difficult to recognize the bundle with elements missing. This is similar to the behavior of variable expressions, which also needed its fixed elements present to be recognized.

5. Implications of Variation for the Processing of MWUs

We have seen that various kinds of MWU have different degrees and types of variation. But what does this tell us about how they are stored and processed? It seems to me that the difference in variation implies that there are differences in storage and processing as well. First, let us discuss idioms. They possess very considerable variation, and yet people are able to process novel variants. If each variant was stored individually in its own right, these facts could be explained. This is a possible explanation, but with so many variants possible, it would be unparsimonious for speakers to have each variant stored individually. Besides, speakers are able to recognize novel variants, which would not yet be stored. If each variant is not stored individually, then how are they processed?

A possibility I would like to suggest is that the complete canonical sequence is stored and is used as a kind of template or exemplar. In this template, the key components would be the content words, particularly lower frequency ones which are less likely to combine widely. Judging from my very exploratory examples, the template can be accessed via one or more of the content words, but it is not usually necessary for all of the them to be recognized for access to occur. In contrast to content words, function words are less useful in accessing the template. My sense is that most templates have a 'core collocation' (usually made up of content words) which reliably leads to access of the template. Drawing on the examples I have used above, the core collocations would be:

- packed / sardines	(*literally packed like sardines in a box*)
- scrape / bottom / barrel	*(he'd had to scrape the barrel a bit.)* *(we're down to the bottom of the barrel)*

Variable expressions have different characteristics. They contain fixed elements that are relatively stable, and so there are relatively few variants. Of course the slot(s) can be filled with a wide range of possibilities, but if we assume that MWUs with some fixed elements and some open elements can be stored as an individual chunk, then the limited number of variants may well be stored as separate forms. In other words, if we assume that the following expressions

a minute ago
a hour ago
a day ago
a week ago
a month ago
a year ago

are all actually stored as the single variable expression

a (time period) ago,

then there is only one form to store, as the fixed elements *a* and *ago* cannot be changed. Even in variable expressions which do contain some variation, it is usually not great, and the few additional forms would not pose an onerous burden to memory. Variation which does exist is often in the tense/modal constituents (*stand shoulder to shoulder / stood shoulder to shoulder*). Perhaps each variation in these constituents entails a separate form to acquire, or it might be that these constituents behave more like a grammatical slot in the sequence, in which case there would be only one form to be stored ([inflection of *stand*] *shoulder to shoulder*).

Lexical bundles are similar to variable expressions in that they are relatively fixed, and so by analogy may also be stored individually. However, there is at least one key difference. Variable expressions have a close connection with meaning and functional language use. Lexical bundles on the other hand have been identified by corpus statistics and often have less obvious relationships with any particular meaning or language function. For example, *you know what* and *the fact that* occur frequently as part of language, but do not seem to realize any unique meaning or function in their own right. Rather, they are building blocks which come to gain meaning once combined with other words or lexical bundles. Given this lack of dedicated meaning, it is questionable whether they are actually stored in a formulaic manner at all. In fact, there is some preliminary evidence that at least some of them are not stored holistically (Schmitt, Grandage, and Adolphs, 2004).

It should be noted that some of the support for the above discussion derives from the fill-in-the-blank tasks illustrated in this paper. Of course, in the real world, people see or hear the entire MWU, that is, words are not normally blanked out. So just because people can't recognize the MWUs with blanks here, the processes may differ when they can see or hear the complete phrase. However, I would argue that the fill-in-the-blank tasks can be useful in illuminating the underlying processes, even if they do not exactly mirror them. In addition, some components of a MWU may be misheard or misread because of degradation of the speech stream or page. In these cases, the experience would not be far off of the one replicated in the tasks.

In sum, some types of formulaic language may well be stored holistically as individual chunks because there is little variation, and thus there are few different forms to hold in memory. Variable expressions and perhaps lexical bundles may fall into this category. On the other hand, other types of formulaic language, e.g., idioms, contain a great deal of variation, and it seems less obvious that a large number of variants would be individually stored. In this case, people who know the canonical form may use it as a template or exemplar; at least this is a theoretical explanation worth pursuing. The main point is that different kinds of formulaic language may involve different kinds of storage and processing.

6. Teaching Implications of Variation in Formulaic Language

If different types of formulaic language are stored and processed differently in the mind, then this means we might have to take different approaches in the teaching and learning of the various types. If variable expressions are stored individually because of their relatively high fixedness, and if there are few variants, then these forms may be worth teaching. This is because variable expressions are closely connected with the expression of particular meanings or functions, and so tend to be recurrent and useful. If only one or a few variants need to be addressed, this is a manageable learning burden considering the communicative benefits received.

Lexical bundles may or may not be stored holistically in the mind, but their use as 'building blocks' of language (i.e., often not realizing meaning content in their own right) suggests that they may be less amenable to explicit teaching, simply because the form-meaning relationship is less transparent. However, some of the lexical bundles identified by Biber *et al.* (1999), entail a clear meaning (e.g., *I don't know, I want to, on the other hand*), and these cases may well be worth teaching.

Idioms are highly conventionalized MWUs, so much so that they have the characteristic of non-compositionality. Furthermore, the more conventionalized a MWU is, the greater the chance that it may be creatively manipulated in language use and still be interpretable. This is because speakers know it so well that they can 'play' with it in creative ways, and their interlocutors know it so well that they can interpret the creative forms. We have seen just this with idioms. I have suggested that well-known, but variable, MWUs may be stored and processed as templates which can be recognized from only some constituents. If this is indeed so, then for this type of formulaic sequence, we might be best teaching the complete 'canonical' form, and then giving examples of how it can be truncated and manipulated.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

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

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 75-93). Harlow: Longman.

Jackendoff, R. (1995). The boundaries of the lexicon. In M. Everaert, E. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 133-166). Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W.D., & Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition, 8,* 1-71.

Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 40-63). Cambridge: Cambridge University Press.

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

Oxford learner's dictionary of English idioms. (1994). Oxford: Oxford University Press.

Pawley, A. & Syder, F.H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J.C. Richards & R.W. Schmidt (Eds.), *Language and communication* (pp. 191-225). London: Longman.

Schmitt, N. (ed.). (2004). *Formulaic sequences: Acquisition, processing, and use.* Amsterdam: Benjamins.

Schmitt, N. & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences:Acquisition, processing, and use* (pp. 1-22). Amsterdam: Benjamins.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 127-151). Amsterdam: Benjamins.

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