9

TESTING FORMULAIC LANGUAGE

Henrik Gyllstad and Norbert Schmitt

Introduction

As has been shown in this volume, formulaic language (FL) has been found to be a pervasive phenomenon in English, and is essential for using it effectively and appropriately. For example, it has been found that FL realizes key functions, e.g., requesting (*Would you please X?*) (Nattinger & DeCarrico, 1992) and meanings (*I'm not sure if X* = expressing uncertainty) (Biber, Johansson, Leech, Conrad, & Finegan, 1999). As the importance of FL has become increasingly apparent, moves have been made to identify Formulaic Sequences (FSs), compile lists of these items, and design tests. In comparison to single-word vocabulary, however, the identification and testing of FL is still in an embryonic stage. Single-word vocabulary testing has benefitted from the development of word lists, which have allowed the testing of the wider vocabulary of a language. For example, the *General Service List* (West, 1953) was influential in materials writing, and thus also the testing of the words appearing in those materials. Towards the end of the century, standardized tests of English single-word vocabulary began appearing, notably the *Vocabulary Levels Test* (VLT) (Nation, 1990), the *Eurocentres Vocabulary Test* (Meara & Jones, 1988), and the *EFL Vocabulary Tests* (Meara, 1992). But around the same time, as there was a growing awareness of the fact that that vocabulary consisted of much more than just individual words, attempts were made to measure knowledge of FSs with newly designed tests (e.g., the *Word Associates Test*, Read, 1998; *DISCO*, Eyckmans, 2009). However, the testing of FL has proved more difficult than that of individual words. Consequently, there is still no consensus on the best ways to measure FL and no test which has been recognized as a standard measurement. This chapter will review tests of FL to date, identify key issues, and suggest ways in which the field can move forward in developing the next generation of formulaic measurement

Tests of Formulaic Language

Despite a growing interest in FL generally over the last couple of decades, the field is far from having developed anything close to a standardized test. This is very different from single-word vocabulary where several tests are accepted and widely used (e.g., those previously mentioned). Several factors make the testing of FL a particularly challenging endeavour. First, FL is made up of numerous disparate categories, which all have their own particular characteristics: e.g., idioms (focus on non-compositionality), collocations (focus on word partnerships), lexical bundles (focus on recurring exact word strings), and phrasal verbs (focus on verb-based multiword units which typically are non-compositional). Therefore, creating a test format which can adequately measure every different category equally well is practically impossible. Second, there are a very large number of FSs, with Pawley and Syder (1983) suggesting that they amount at least to "several hundreds of thousands" (p. 213), while Jackendoff (1995) concludes that the phrasal lexicon may be the same or larger than the lexicon of single words. These two factors work against both the identification of the target population and the representative sampling of items from that population. This leaves researchers in a very challenging position, and it is probably next to impossible to develop a definite list of all the existing FSs in a language, and then to develop a test for these sequences.

Another factor involves the definition of FL (also connected with Point 1), obviously a prerequisite for testing. Definitions containing statistical criteria can be precise and measurable, e.g., MI scores in the frequency approach to collocations. But many criteria are much more subjective, e.g., the degree of compositionality for idioms (Grant & Nation, 2006), as they tend to rely on researcher intuition to some extent. An even fuzzier criterion is that of "holistic storage", which is mentioned in some definitions of FL (such as Wray, 2002), but is virtually impossible to operationalize. Also, as has been argued by Read and Nation (2004), what is "holistic" varies from person to person, and even varies from time to time within a person:

> the means of storage and retrieval of the same sequence can differ from one individual to another, and can differ from one time to another for the same individual depending on a range of factors such as changes in proficiency, changes in processing demands, and changes in communicative purpose.
>
> *(p. 25)*

Clearly, the rather complex and heterogeneous nature of FL presents challenges when it comes to testing and assessment, and there is a shortage of research to date on how it can and should be measured. A telling sign of this is the absence of assessment and testing as one of the identified main strands of activity in a thematic issue on FL in the *Annual Review of Applied Linguistics* in 2012 (Wray, 2012). The present situation is, thus, that there is no established best practice for how to test FL, let alone a standardized test of FL ability

Research Including Assessment of Formulaic Language

Despite the absence of tests of overall FL, a number of tests target knowledge of particular categories (e.g., collocations, idioms, word associations and phrasal verbs). Whereas idioms have received much attention in research on processing, from a testing point of view, the most frequent type of FS targeted is seemingly collocation. A number of studies have involved analyses of corpora of L2 essays written in English (e.g., Howarth, 1996; Laufer & Waldman, 2011; Nesselhauf, 2005; Siyanova-Chanturia, 2015) (see Chapter 12 in this volume for a discussion of these studies).[1] Furthermore, there are a number of studies in which some sort of elicitation technique has been used (e.g., Bahns & Eldaw, 1993; Farghal & Obiedat, 1995; Garnier & Schmitt, 2016). Finally, there are a handful of published studies in which the overarching aim has been to develop tests of collocation knowledge: Eyckmans (2009), Gyllstad (2009), and Revier (2009). There are also tests of word associations that are relevant in this regard: Read (1993), Vives Boix (1995), and Wolter (2005). Due to length restrictions, we will focus here on seven tests which exemplify a range of formats and that illustrate key issues for the development of FL tests.

The Word Associates Test (Read, 1993, 1998)

The *Word Associates Test* (WAT) is one of the oldest tests, which measures collocation knowledge as one type of FL as part of the test format (it also measures meaning knowledge based on synonyms[2]). It is probably also the best-known of the tests reviewed here. It was originally developed by Read (1993), and in its initial conception it was intended to measure knowledge of academic English vocabulary, as represented by the words in the University Word List (UWL) (Xue & Nation, 1984), an 800-word compilation based on various frequency counts of academic texts. In its revised version, it is aimed at measuring "the extent to which learners were familiar with the meanings and uses of a target word" (Read, 1998, p. 43), with the "uses" part measured by a matching collocation format. The test presents 40 adjectives like the example shown in Figure 9.1. The task is to circle the four words which associate with the target item, for example, in the figure, *quick* and *surprising* (synonyms) and *change* and *noise* (collocations).

The test features relatively few words selected for having strong and recognizable collocates. (This is one reason for using only adjectives for target words.)

Sudden

Beautiful	quick	surprising	thirsty		change	doctor	noise	school

FIGURE 9.1 An example task item from the Word Associates Test (new version).

Notably, the WAT was never designed to generalize to inferences about wider collocation knowledge, but rather a test which uses collocation knowledge as a proxy for "depth of knowledge". This concept is extremely vague (see Read, 2004), and it is interesting that Read used a type of FL to represent this more advanced quality of lexical knowledge. Perhaps this should not be surprising, as most research spanning from Bahns and Eldaw (1993) to Siyanova-Chanturia (2015) points to mastery of FL being one of the later aspects of lexical knowledge to be acquired.

The WAT uses a recognition format, partly due to practical constraints. Research shows that productive mastery of collocation is much more difficult than receptive knowledge, with Laufer and Waldman (2011) finding only about half the number of collocations in non-native essays compared to native ones. Read used a receptive format to ensure he would elicit collocation responses, even though they would be at the relatively easier receptive level of mastery. However, use of what is essentially a multiple-choice format leads almost inevitably to problems with examinees using test-taking strategies to answer the items (Gyllstad, Vilkaitė, & Schmitt, 2015). Schmitt, Ng, and Garras (2011) found that the method of scoring was crucial, and they suggested giving credit for items only if all correct options were selected, in order to compensate for guessing. The fact that the test is still in development 23 years after its inception (i.e., Read, 2016) illustrates the difficulty of measuring depth of knowledge, and for our current purposes, collocation knowledge in particular, as one type of FS.

COLLEX and COLLMATCH (Gyllstad, 2009) /DISCO (Eyckmans, 2009)

In an edited volume on researching L2 collocations (Barfield & Gyllstad, 2009), no fewer than four tests of collocation knowledge are presented: COLLEX and COLLMATCH (Gyllstad, 2009), DISCO (Eyckmans, 2009), and CONTRIX (Revier, 2009). The first three tests use versions of a recognition format and will be discussed in this section.

The COLLEX and COLLMATCH tests were designed to measure advanced Swedish learners' (upper secondary school and university) receptive recognition knowledge of English verb+noun word combinations. The tests were developed and evaluated through a series of test administrations, aimed at creating test versions yielding reliable and valid scores (see Gyllstad, 2007 for details). The COLLEX test is a 50-item test with a decontextualized format, as shown in Figure 9.2. The test taker must choose the alternative that is a frequent and natural word combination in English (b).

The most recent iteration of the COLLMATCH test format is essentially a yes/no test which targets collocations. It presents decontextualized items, which the examinees judge as being frequent and natural word combinations in English

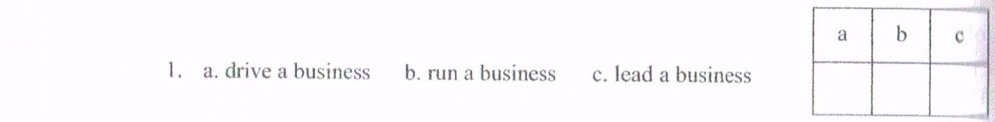

FIGURE 9.2 An example item from COLLEX.
(from Gyllstad, 2009, p. 157)

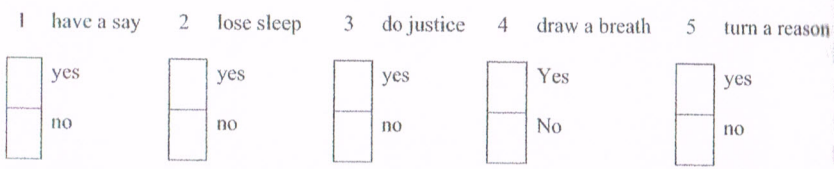

FIGURE 9.3 Five example items from COLLMATCH.
(from Gyllstad, 2007, p. 309)

Gyllstad administered both the COLLEX and COLLMATCH, along with the VLT (Schmitt, Schmitt, & Clapham, 2001), to 307 participants (mainly Swedish university students), and found that both collocation tests produced very similar scores in terms of percentage correct. They also correlated strongly with vocabulary size (VLT) (.83–.88) and with each other (.86). Thus, we find that the COLLEX (using a three-option multiple-choice format) and the COLLMATCH (using a yes/no format) provide very similar information, despite the differing formats.

This makes the yes/no format of the COLLMATCH interesting because of its advantage of speed; yes/no tests are typically quicker to take, and so more items can be tested than for other formats, e.g., COLLMATCH – 100 items; COLLEX – 50 items. Given the large number of FSs in language, this format allows a far larger sampling. A possible downside of the format is that there is no demonstration of knowledge, and cynical test takers could in theory simply guess and have a 50–50 chance of answering correctly. Yes/no tests of single words typically have non-words added (which the test takers obviously cannot know), and if these words are checked as known, then learners' scores can be adjusted downwards accordingly. (However, there is no consensus on the best adjustment formula, e.g., Pellicer-Sánchez & Schmitt, 2012.) In COLLMATCH, 70 of the 100 items are target collocations and 30 are pseudo-collocations. For all items, irrespective of category, z-scores were retrieved from the British National Corpus (BNC) to ensure significance for the target collocations and conversely lack of significance for the pseudo-collocations. Thus, it might be possible to adjust scores for guessing if a suitable adjustment formula can be found.

Another example which illustrates how the receptive format can be adapted

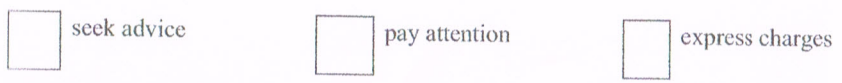

FIGURE 9.4 An example item from DISCO.
(from Eyckmans, 2009, p. 146).

in Figure 9.4, the test taker is asked to tick the *two-word* combinations that are idiomatic in English (*seek advice, pay attention*).

Eyckmans found that the DISCO was sensitive enough to indicate improvement in collocation knowledge after a 60-hour period of instruction, although it had limited power in indicating production of FSs in oral output.

It is interesting that all of these tests are intended to measure collocation knowledge, and although the authors provide a number of different kinds of validity and reliability evidence (see the respective studies for details), none have sufficient validation evidence which would indicate how their scores are to be interpreted in terms of overall knowledge of collocations. This is because the item selection approach used was largely a word-centred approach, whereby collocates are identified for high-frequency node words. A more holistic approach would entail selecting whole collocations from a frequency list. This criticism, essentially pointing to a lack of a good model of collocation knowledge and use, is not specific to these tests, but could be made of virtually every collocation measure, and is a weakness we think test developers need to address in the future.

CONTRIX (Revier, 2009)

The CONTRIX format is different from the previous formats in that, although employing a receptive format, it is claimed to assess L2 learners' productive knowledge of verb-object/noun collocations (e.g., *make a complaint*) (Revier, 2009). CONTRIX items consist of a sentence prompt containing a gap to be filled by selecting words from each of the three columns to the right. Test takers are asked to select (circle) the combination of verb, article, and noun that best completes the sentence. In the example in Figure 9.5, that would be *keep + a + secret*.

Using Schmitt's matrix for what type of knowledge is tested (2010, p. 86) (see Table 9.1), the format targets "meaning recognition" (by providing the components of forms to choose from). However, Revier (2009) argues somewhat unconventionally that it could also be said to tap into "productive knowledge for test takers must not only create (i.e., produce) meaning by combining lexical constituents, but they must also grammatically encode the noun constituent for determination" (p. 129). This is an interesting claim, but unfortunately, the initial pilot only investigated differences in scores between learners of different proficiency levels, and differences in scores between transparent, semi-transparent, and non-transparent collocations. Thus, there was no evidence that the test provides

The quickest way to win a friend's trust is to show that you are able to _____ .	tell	a/an	joke
	take	the	secret
	keep	--	truth

FIGURE 9.5 An example item from CONTRIX.
(Revier, 2009, p. 129)

A Productive Collocation Test (Schmitt, Dörnyei, Adolphs & Durow, 2004)

In the literature, a large number of studies have made use of tests of FL as part of traditional experimental and quasi-experimental learning designs (e.g., Henriksen, 2013; Schmitt, 2004). One example that illustrates a more conventional way of measuring productive knowledge of FSs (compared to the CONTRIX) comes from a study by Schmitt et al. (2004). The researchers created a productive test that was a type of cloze test. A range of academically based FSs were embedded in multi-paragraph contexts, with all or most of the content words in the target FS deleted, but leaving the initial letter(s) of each word. The meaning of the targeted sequence was provided next to the item in parentheses to ensure that the ability to produce the "form" of the formulaic sequence was measured, not comprehension of its meaning. One paragraph is extracted, which contains two items (*first of all, it is clear that*) (Figure 9.6).

This format is reminiscent of Laufer and Nation's (1999) format used for a single-word productive VLT. It has the advantage of being difficult to guess if one does not know the target item, while seemingly relatively easy to complete if the target sequence is known. The format would be classified as "form recall" according to Schmitt's (2010, p. 86) terminology, and the test would seem to provide evidence that a learner can spell the sequences in question. However, this is far from demonstrating that learners can think of the sequences unprompted on their own and independently use them in their writing and speaking. This highlights another problem common to almost all tests of FL: the uncertainty of how to interpret the scores in terms of how much FL learners can employ in their everyday use of the four skills.

The PHRASE Test (Martinez, 2011)

One of the few tests where the scores can be related to a fixed set of FSs is

Learning English as a second language is a difficult challenge,

but we do know several ways to make learning more efficient.

Fi_____ of a_____, almost every research study (the initial one)

shows that you need to use English as much as possible.

I_____ is cl_____ that the more you use English, (this is obvious)

the better you will learn it. There is not disagreement about

this.

FIGURE 9.6 Example items for testing production of FS.
(Schmitt et al. (2004, pp. 58–59)

At once: I did it **at once**.

a. one time

b. many times

c. early

d. immediately

FIGURE 9.7 An item for the PHRASE Test.
(Martinez, 2011, Slide 54)

expressions on the *PHRASE List* (Martinez & Schmitt, 2012). It uses a fairly standard four-option multiple-choice format, with the target item and a short, non-defining sentence as context (d) (Figure 9.7).

The key thing to note about this experimental test is that it was sampled from a finite list of phrasal expressions, and so the percentage correct on the test can be interpreted as the percentage known on the whole PHRASE List. This is in stark contrast to the other tests discussed, where there is no way to know how to interpret the scores in terms of overall size. This suggests that future tests of FL may need to focus on much more constrained, and thus identifiable, subsets of FL, in order to make the resulting scores more meaningful.

Principles for Developing New Tests of Formulaic Language

The previous review shows that, although measurement of FL subtypes has pro-

writers will need to enhance their development procedures to write tests that provide valid and reliable scores that are useful for teachers and students. Length constraints prohibit us from outlining every issue that needs to be considered when developing valid tests of FL, but we feel the following key issues need to be addressed.

Defining Constructs

In any language testing endeavour, there is a need to link an individual's test performance to a specific ability in reference to a construct. In terms of procedures for this, Bachman (1990, p. 40) suggests a sequence of three steps: (1) defining the construct theoretically, (2) defining the construct operationally, and (3) establishing procedures for quantifying observations. Once such definitions are in place, we can start working with the population of items that supposedly belong to the construct. Over the last couple of decades, a number of influential definitions of FL have been presented in the literature, most notably Wray's oft-cited 2002 version:

> a sequence, continuous or discontinuous, of words or other 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.
>
> *(p. 9)*

Wray's definition was designed to capture as wide a range of FL as possible for discussion in her seminal book, but this open-ended definition is simply too broad to operationalize in the development of tests (e.g., if something *is* or *appears to be* prefabricated, then it does not have to be prefabricated, thus making this criterion unworkable). In her 2008 book-length follow-up, Wray provides an excellent discussion on the way in which different definitions lead to wider or narrower scope when it comes to identification of formulaic exemplars, and how different research purposes require different approaches (Wray, 2008: Chapter 8). This applies to testing too, and the last point creates a natural link to the next issue, that of having a particular purpose in mind when designing a test.

Tests Need to be Developed for Particular Purposes

Tests of individual words have tended to be generic, with no indication given about which particular purposes, contexts, or learners the tests were suitable for. For example, developers of the well-known VLT only indicated that it provided an estimate of the vocabulary size at different frequency levels, but never specified what kind of learners it would be appropriately used with. This is true for both

et al., 2001). Likewise, users of the *Vocabulary Size Test* (VST) (Nation & Beglar, 2007) were initially given no guidance of whom to use it with or how. Nevertheless, Nation and Coxhead (2014) later found that some New Zealand participants performed quite differently on the test when a personal test administrator talked them through it and kept them engaged, compared to doing it themselves. Furthermore, this difference was the greatest for people in the lowest quartile of test takers. This discrepancy is not surprising, because no test works for every person in every situation. This means that the current situation where "one-size-fits-all" tests dominate is no longer tenable. Testing of FL must follow the lead of mainstream testing where tests are developed and validated for specific contexts and learners (Bachman & Palmer, 2010; Read & Chapelle, 2001) (see Voss (2012) for an example of this), and where appropriate statistical methods are used (see Bachman, 2004). Developers of formulaic measurement need to be clear about what their tests are trying to achieve, what the resulting scores mean, and whom to use the test with and in which situations. The reason for this is straightforward: if no context is specified for a test, then how can validation evidence be collected that it works? This calls for the creation of some type of manual accompanying the test which outlines (in plain speak) the essential requirements for choosing, administrating, and marking the test, and then for interpreting the resulting scores.

Selecting the Formulaic Sequences to Test

FL is ubiquitous in language. With so many FSs in language, it becomes essential to narrow them down in some way to have a reasonable chance of obtaining a viable measurement.

As explained in this volume, FL is not a homogeneous phenomenon, but is on the contrary, quite varied, made up of a range of different categories (e.g., idioms, phrasal verbs, lexical phrases, collocations, phrasal expressions, and discourse organizers). Each category has its own particular characteristics. Idioms and some phrasal verbs have idiomatic meanings. Other categories, like lexical bundles and discourse markers, have meanings that are typically transparent from the individual words in the sequence but bound in conventionalized strings expected by the speech community. Some categories, like collocations, can have both idiomatic and literal meanings (*top drawer* = highest drawer in a cabinet, and best example of something). The various categories are used for different functions, e.g., discourse markers are used to signpost discourse organization, while idioms usually express meaning (*silver lining* = there is hopefully some good in a bad situation).

With such a range and variety of FL, it is not surprising that no single, comprehensive compendium exists. Even if it did, it would almost certainly not be measurable. This makes it important to understand the reason for testing, in order to define which category or categories of FL to measure. For example, if the purpose is to measure knowledge of the most common FS, then testing phrasal expressions

frequency (among the most frequent 5,000 lexical items in English). Another purpose might be to measure writing ability, and testing discourse organizers might be a sensible part of this approach.

Once the category (or categories) of FL has been determined, the test developer needs to identify the population of FSs in that category, from which to sample in order to build the test. Most developers will rely on existing descriptions of the category, in the form of either a dictionary or list. There are numerous dictionaries focusing on various categories of FL. Idioms, phrasal verbs, and collocations are well supported with dictionaries from most of the major publishers. However, it should be noted that these resources vary greatly in how they were compiled, but because they are written for learners and not researchers, the rationale/procedure for inclusion (or omission) of items is either vague or left completely unstated.

The existing dictionaries have large numbers of items: e.g., the *Cambridge Idioms Dictionary* (2006, 2nd ed.) has approximately 7,000 items, *Collins COBUILD Phrasal Verbs Dictionary* (2012) has more than 4,000 items, and the *Oxford Collocations Dictionary for Students of English* (2009, 2nd ed.) has approximately 250,000 items. These large numbers of items can leave the test developer overwhelmed, as any sample small enough to be testable will only include a tiny fraction of the total items in the dictionaries (see sample rate identified later). To address this issue, a number of lists have been developed to identify a smaller number of the most useful items to teach and test, usually as indicated by high frequency. For example, Liu (2011) narrowed the nearly 9,000 phrasal verbs he analyzed down to 150, which made up nearly two-thirds of the phrasal verb occurrences in the BNC corpus. Likewise, Liu (2003) used frequency and range criteria to identify 302 idioms which occurred in the spoken discourse of professional, academic, and media language. Other lists which provide frequent and pedagogically relevant items include the Pearson *Academic Collocation List* (2,469 items; Ackerman & Chen, 2013) and the *PHRASE List* (505 items; Martinez & Schmitt, 2012).

But lists are not limited to indicating the most useful items in an FS category; they can also provide beneficial information about the items. For example, many lexical items are polysemous, and lists can potentially give information about the frequency of various meaning senses. For single-word vocabulary, this was most notably done by the *General Service List* (West, 1953), which gave percentages of the various meaning senses of the key 2,000 words in English. The same format was provided by the *PHaVE List* (Garnier & Schmitt, 2015), which provides percentage information about the meaning senses of the most frequent 150 English phrasal verbs. Another type of information is how FSs are used. The *Academic Formulas List* (Simpson-Vlach & Ellis, 2010) categorizes the identified formulas according to their function (e.g., Quantity specification – *both of these*, Contrast and comparison – *as opposed to*).

As useful as dictionaries and lists may be, it is important for test developers to

If word lists are not used, then the test developer must rely on other means to select the FSs to include on the test. Particularly for collocations and lexical bundles, statistical frequency-based criteria are often used (e.g., t-score, Mutual Information (MI), DeltaP, a certain number of occurrences in a corpus). Some researchers favour semantically based criteria based on the "phraseological school" approach (e.g., degree of compositionality, amount of variation allowed). It is beyond the remit of this chapter to go into these in detail (see Barfield & Gyllstad, 2009 and Schmitt, 2010, for overviews), but a logical requirement is that test developers need to carefully consider their purposes and which criteria are best suited to achieving those purposes.

Sampling

Once a source of appropriate FS has been selected, the next step is to sample a suitable number of items from it to fix on the test. This brings up the issue of sampling rate. With some language constructs, a limited number of items can give a good indication of the knowledge of the construct. For example, if a learner can answer several test items correctly demonstrating the past form of regular English verbs (*-ed*), this probably gives a good indication they can use this grammatical "rule" across the range of regular verbs. But FL is not a rule-based, but rather an item-based construct. Just because a learner knows one collocation or idiom, for example, does not imply they know a different one. Therefore, each lexical item, whether individual word or FS, needs to be tested separately. Given the large number of lexical items in any language, this causes problems for the test developer, as it is simply impossible to test every item.

The standard solution is to draw a representative sample from the overall population of items, and then use these to extrapolate to the complete population. For example, at the 3,000 level of the revised VLT (Schmitt et al., 2001), 30 items represented the 1,000 words in the level. If learners answered 50% of the items correctly (15), then the interpretation was that they also knew 50% of all the words in that level (500). Sampling always involves a tension between validity (more items give better test information) and practicality (fewer items lead to shorter and more practical tests). This leads to the obvious question of what is the lowest sampling rate which can produce valid test scores.

The answer to this question partly stems from the purpose of the test. If the intention is to obtain a very rough idea of the number of words/sequences a person knows, then a lower sampling rate might suffice. But if the test is supposed to produce a relatively accurate estimate, then a higher sampling rate is required. Unfortunately, there has been little research into how sampling rate affects the validity of vocabulary tests. Use of sophisticated statistics from *Item Response Theory* like Rasch analysis can allow the measurement of homogenous constructs with relatively few items (see McNamara, 1996), but it is very debatable whether

frequency bands (e.g., *VST* – 10 items; *X-Lex Test* (Meara & Milton, 2003) – 20; *VLT* – 30), with the unverified assumption that this was enough. Unfortunately, there has been little research into minimum sample rates for vocabulary tests, and none to our knowledge for tests of FL.

One study which does shed light on sampling is Gyllstad, Vilkaitė, and Schmitt (2015). They compared test scores from the 10 items on a four-option multiple-choice test (*VST*) with scores from a much more comprehensive 100-item test (which was assumed to be a better estimate of the 1,000-word frequency level). The 10-item test correlated at .50–.86 (r^2 = .25–.74). Unsurprisingly, more items led to increasingly higher correlations, with 30 items producing correlations of .85–.95 (r^2 = .73–.90). The researchers concluded that 10 items per 1,000 were sufficient to give a ballpark indication of vocabulary size, but that more items led to more accurate estimates, while any more than 30 items may well lead to practicality issues due to excessive test length. However, while this study is informative, the *VST* measures individual words, and it is unclear whether tests of FL would behave in a similar manner. Given the lack of research in this area, test developers will need to run their own validation studies to determine what sampling rate is appropriate for their particular purposes and needs. However, unless the set of FSs is quite constrained (e.g., the 207 core formulas on the *Academic Formulas List*, Simpson-Vlach & Ellis, 2010), the number of items on any FL test will likely need to be substantial.

However, the tests reviewed so far have not attempted to map onto a defined set of FS, but rather have attempted to measure more general collocation knowledge. The WAT has 40 items, the CONTRIX 45 items, the DISCO and COLLEX 50, and the COLLMATCH 100, but the number of items necessary to indicate general collocation knowledge as a construct is still undetermined. Future research will need to look at all of these formats in order to determine the number of items needed to provide useful information.

Choosing Appropriate Item Formats

We have looked at a number of different formats, but there is no way of saying that any format is better than the others. It all comes down to the test purposes and the type of learner taking the test. In this sense, tests of FL are no different from any kind of vocabulary test. As a way to avoid using terms like *receptive* and *productive*, Schmitt (2010), based on work by Laufer and Goldstein (2004), proposed the use of a two-by-two matrix for what particular type of form-meaning knowledge is targeted in a test. The matrix is aimed at single-word knowledge but can also be useful in guiding thought about tests of FL (Table 9.1).

Once a test developer is clear what degree of mastery should be tested for the specified purpose of the test, then Table 9.1 can help the developer think about what kind of item format is required to tap into that level.

TABLE 9.1 *Matrix for deciding what aspect and level of mastery an item is tapping into*

		Formulaic sequence knowledge tested	
		RECALL	RECOGNITION
Formulaic sequence knowledge given	MEANING	Form recall (supply all or part of the L2 sequence)	Form recognition (select the L2 sequence)
	FORM	Meaning recall (supply definition/ L1 translation etc.)	Meaning recognition (select definition/ L1 translation etc.)

Source: (adapted from Schmitt, 2010, p. 86)

Taking Advantage of Technology

Most tests of single-word vocabulary and FL to date have traditionally been of the paper-and-pencil variety. Some of these tests have been moved to computerized or Internet-based platforms, but for the most part, they are simply electronic versions of the paper-and-pencil formats. For example, the *Lextutor* website (www.lextutor.ca/tests/) provides web-based versions of a number of existing tests (e.g., the *VLT*, the *VST*, the *EFL Vocabulary Tests*, the *Phrasal Vocabulary Size Test*, BNC Version, Martinez, 2011). But the electronic age offers more possibilities than just reworking existing tests. One opportunity is to use adaptive tests to achieve a better and more focused sampling. With paper-and-pencil tests, the number of items at each level is fixed on the page, and learners must go through all of the items, regardless of whether they are too easy (e.g., high-frequency items which are very well-known) or too difficult (e.g., low-frequency items which may not be known at all). Computer-adaptive tests can use a few items at each frequency level to gauge the particular learner's general level, and then give many items at the frequency point where some, but not all, of the words/sequences are known. This allows many more items to be given in the "window", which is most informative of the learner's vocabulary size. However, the most sensible adaptive formula has not yet been established from the many options available, and research has only begun on the advantages and disadvantages of various algorithms (Kremmel, in preparation).

Adaptive tests also have the potential to give information on the quality (depth) of lexical knowledge. This has been demonstrated by the *Computer Adaptive Test of Size and Strength* (*CATTS*) test (Laufer & Goldstein, 2004), which gives examinees tests of form recall, meaning recall, form recognition, or meaning recognition, depending on the learner's responses. The result is an indication of the strength of knowledge of the form-meaning link, and the idea of using a computer-adaptive format should also work with FL.

Conclusions and Future Directions

In this chapter, we have concluded that there is a lack of standardized tests of FL, and instead that a number of tests exist that target subtypes of FS, such as idioms, collocations, phrasal verbs, and lexical bundles. Many are still experimental, and most of them lack the type and scope of validation research necessary to truly know how they work, and what their scores mean. For the majority, it is still difficult to determine what their scores say about knowledge of wider-ranging formulaic knowledge. Our suggestions for future tests largely revolve around a call for a much more rigorous specification of test purpose, and with it, a tighter description of the category(s) and scope of FL being measured. Ideally, this should be done through the issuing of a "user manual" accompanying the test. With this, desirable formats and the selection of target items will be much more obvious to achieve the stated purpose(s).

Beyond our suggestions, what might the future bring for the testing and researching of FL? In a conference colloquium dedicated to FL, Vilkaitė and Gyllstad (2014) discussed several possibilities. For example, in terms of identification of FS, they foresaw that intuition will continue to play a role, but with an increased use of several raters/judges to improve judgements (inter-rater reliability). Also, whereas offline/paper-and-pencil tests will continue to have appeal in traditional classroom settings, more sophisticated psycholinguistic and neurolinguistic approaches (e.g., eye-tracking, EEG, ERP, and fMRI) will be used in researching FL, especially when it comes to the question of holistic storage and researching differences (cross-sectionally) between groups of native speakers and learners, and within individuals over time (longitudinally).

It stands to reason that no one can truly foresee the future, but if future test developers follow up on our suggestions, and take on-board the other wealth of information available in this volume, the next generation of FL tests cannot help but be much improved.

Notes

1 Text analysis studies that have examined the use of FL in learner production could be seen as measurements/tests of productive FL ability, as are rating criteria and scoring rubrics.
2 This makes it difficult to interpret the scores as FL knowledge, as half the test measures a non-FL construct (i.e., single-word meaning).

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the academic collocation list (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12,* 235–247.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford: Oxford University Press.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System, 21,* 101–114.

Barfield, A., & Gyllstad, H. (Eds.). (2009). *Researching collocations in another language: Multiple interpretations.* Basingstoke: Palgrave Macmillan.

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

Cambridge Idioms Dictionary (2006). Cambridge: Cambridge University Press.

Collins COBUILD Phrasal Verbs Dictionary (2012). Glasgow: HarperCollins.

Eyckmans, J. (2009). Toward an assessment of learners' receptive and productive syntagmatic knowledge. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 139–152). New York, NY: Palgrave Macmillan.

Farghal, M., & Obiedat, H. (1995). Collocations: A neglected variable in EFL. *International Journal of Applied Linguistics, 28*(4), 313–331.

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research, 19*(6), 645–666.

Garnier, M., & Schmitt, N. (2016). Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System, 59,* 29–44.

Grant, L., & Nation, P. (2006). How many idioms are there in English? *ITL International Journal of Applied Linguistics, 15*(1), 1–14.

Gyllstad, H. (2007). *Testing English collocations* (Unpublished PhD thesis). Lund: Lund University.

Gyllstad, H. (2009). Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 153–170). New York, NY: Palgrave Macmillan.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics, 166,* 276–303.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development – A progress report. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 29–56). Eurosla Monograph Series 2.

Howarth, P. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making.* Lexicographica Series Maior 75. Tübingen: Max Niemeyer.

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: Lawrence Erlbaum.

Kremmel, B. (in preparation). Algorithms for vocabulary size tests.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3). 399–436.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33–51.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672.

Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly, 37*(4), 671–700.

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly, 45*, 661–688.

Martinez, R. (2011). *Putting a test of multiword expressions to a test.* Paper presented at the IATEFL Testing, Evaluation and Assessment SIG. University of Innsbruck: September 16, 2011. Retrieved from https://ufpr.academia.edu/RonMartinez/Talks.

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics, 33*(3), 299–320.

McNamara, T. (1996). *Measuring second language performance.* Harlow: Longman.

Meara, P. (1992). *EFL Vocabulary Tests.* University College, Swansea: Centre for Applied Language Studies.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. Retrieved from www.lognostics.co.uk/vlibrary/meara&jones1988.pdf.

Meara, P., & Milton, J. (2003). *X_Lex, the Swansea levels test.* Newbury: Express.

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* New York, NY: Newbury House.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nation, P., & Coxhead, A. (2014). Vocabulary size research at Victoria University of Wellington, New Zealand. *Language Teaching, 47*(3), 398–403.

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

Nesselhauf, N. (2005). *Collocations in a learner corpus.* Amsterdam: John Benjamins.

Oxford Collocations Dictionary for Students of English (2009). 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.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes – No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing, 29*(4), 489–509.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*, 355–371.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Lawrence Erlbaum.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 209–227). Amsterdam: John Benjamins.

Read, J. (2016). *A fresh look at measuring depth of vocabulary knowledge.* Paper presented at the Vocab@Tokyo conference. September 12–14, 2016, Tokyo.

Read, J., & Chapelle, C. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 3–32.

Read, J., & Nation, P. (2004). Measurement of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 23–35). Amsterdam: John Benjamins.

Revier, R. L. (2009). Evaluating a new test of whole English collocations. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 125–138). New York, NY: Palgrave Macmillan.

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

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Basingstoke: Palgrave

Schmitt, N., Dörnyei, Z., Adolphs, S., & Durow, V. (2004). Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 55–86). Amsterdam: John Benjamins.

Schmitt, N., Ng, J. W. C., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing, 28*(1), 105–126.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing, 18*(1), 55–88.

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

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System, 53*, 148–160.

Vilkaitė, L., & Gyllstad, H. (2014). *Formulaic language: How can it be assessed?* Paper presented at the AAAL conference. March 23, 2014, Portland, Oregon.

Vives Boix, G. (1995). *The development of a measure of lexical organisation: The association vocabulary test* (Unpublished PhD thesis). University of Wales, Swansea.

Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design* (Unpublished PhD thesis). Iowa State University.

West, M. (1953). *A general service list of English words.* London: Longman.

Wolter, B. (2005). *V_Links: A new approach to assessing depth of word knowledge* (Unpublished PhD thesis). University of Wales, Swansea.

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

Wray, A. (2008). *Formulaic language: Pushing the boundaries.* Oxford: Oxford University Press.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics, 32*, 231–254.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215–229.