Introduction

Besides single orthographic words (i.e., sequences of letters surrounded by spaces on each side), in most of the world’s languages there are multiword items (MWIs). For example, in the case of English, there are also structures including but not restricted to idioms (shoot the breeze), collocations (blissfully happy), proverbs (a watched pot never boils), binomials (wear and tear, part and parcel), compound nouns (with three spelling variants: concatenated: boyfriend, spaced: blood pressure, and hyphenated: word-play) (Kuperman & Bertram, 2013), and phrasal verbs (play up) (see Wood, this volume, for more examples and definitions). As a whole, these MWIs are very common (Erman & Warren, 2000; Pawley & Syder, 1983; Jackendoff, 1995), with Bolinger (1976, p. 1), among others, reminding us about the fact that “our language does not expect us to build everything starting with lumber, nails, and blueprint, but provides us with an incredibly large number of prefabs”.

In terms of terminology, Wray (2002) has illustrated how many labels have been used in the literature, among those “prefabs”, to capture these MWIs and to research them. On the assumption, then, that MWIs abound in language and that efficient language use (comprehension and production) relies on individuals having knowledge of and access to a very large number of these, it is important that knowledge and use can be measured in valid and reliable ways.

Some scholars, such as (McNamara, 2000, p. 55) suggest that the term measurement refers to “the theoretical and empirical analysis of scores and score meaning” (McNamara, 2000, p. 55). As such, measurement can be seen to be the rather more technical side of assessment and testing, where scores are assigned and subjected to various statistical calculations. Other scholars, like Bachman and Palmer (2010, p. 19/) pragmatically argue that “measurement”, “test”, and “assessment” are best seen as terms referring to the same thing: collecting information. However, they add that the important thing for test developers is to clearly specify the conditions under which test takers’ performance will be obtained and the procedures followed for recording this performance. Accordingly, in this chapter, “measurement”, “assessment”, and “testing” will be used interchangeably. The focus will be on measuring knowledge in a second (L2, L3, Ln) or foreign language.
Critical Issues and Topics

Prerequisites for Sound Measurement

Any measurement that lays claim to being a good measure of a specific type of knowledge needs to be valid and reliable. Measuring knowledge of MWIs is no different. Primarily, validity has to do with the notion that what we intend to measure or test is also what we in fact do measure (Fulcher, 2010). At the heart of any assessment enterprise is the “construct”, i.e., that knowledge, skill, or behavior purported to be tested. We will return to this concept later.

It is not uncommon for researchers and test developers to discuss different types of validity, such as construct, content, and criterion validity, but ever since Messick’s (1989) seminal work, validity is generally treated as a single concept but with a number of facets. It is important to note that validity is not technically a characteristic of a particular test but rather the inferences and interpretations made from the test scores and their use.

Reliability refers to the consistency of measures across different conditions in the measurement process (Bachman, 2004), and, moreover, whether there is cause for concern that additional construct-irrelevant factors also affect the measurement (Bachman, 1990). We always want to minimize the potential for error in our measurement, and if we can do that, then we will achieve a high level of reliability. A very commonly used analogy is that of a person weighing themselves. If I stand on a scale one day and weigh myself, and then do the same again ten minutes later, I expect the scale to show the same value again, of course on the assumption that I have not eaten or drunk a lot in between, and that I wear the same amount of clothes at both occasions. With respect to measuring knowledge of MWIs, if we for example designed a test to be a reliable measure of this construct, we would expect test takers to get the same or a very similar score if tested on two occasions, as long as no learning was assumed to have taken place in-between. Also, if many test takers who get a high total score on a test answer a question wrongly, and many test takers who get a low total score on the test answer that same question correctly, then it is likely that our measure of the construct will not be very reliable. Reliability can be measured in several ways, for example test-retest, parallel forms, split-half, and internal consistency (see Bachman, 2004, and Weir, 2005, for good accounts of these, and Fulcher, 2010, for calculation examples). There are also specific reliability measures to be used with criterion-referenced tests (see Bachman, 2004).

As was mentioned earlier, in any language measurement endeavor, a link needs to be made between an individual’s performance and a specific type of language knowledge or skill (a construct). Bachman (1990, p. 40) has suggested a procedure consisting of three steps to achieve this: (1) defining the construct theoretically, (2) defining the construct operationally, and (3) establishing procedures for quantifying observations. In the case of MWIs, we thus first need a theoretical definition. As an illustration, we could formulate a construct which is “receptive recognition knowledge of English proverbs” (e.g., proverbs like when in Rome, every little helps, and better the devil you know). Once such a definition is in place, we can start working with the population of items that supposedly can be linked to the construct and subsequently sample from this. This is when we operationalize the theoretical construct. This is a tall order, since we ideally want to base our population of items and sampling on some kind of model. Potentially, we could start by asking a sizeable group of native speakers for the proverbs they think are the most useful and frequently used in daily communication. Alternatively, we could specify that the proverbs should have a particular function in the language (in which case the theoretical construct may have to be adjusted). We could additionally ask them to rank them in bins by their relative frequency; for example, extremely frequent, highly frequent, moderately frequent, and frequent. We could then use a large native-speaker corpus to check things...
like raw frequency and see if we can validate the ranking. An additional check with a second native-speaker group is desirable. If in their judgment the items and their estimated relative frequency can be corroborated, we then have a sample of items. The final step is then to create procedures for measuring and quantifying our observation. For example, we could decide to use a receptive recognition task in which 30 frequent MWIs are presented together with ten distractors (nonexistent sequences of words for which we have evidence of their nonexistence) in a regular pencil-and-paper test format, in which test takers are asked to circle those that they know are frequent and conventional language sequences, or as an additional criterion those they know the meaning of. Our scoring procedures could then be based on a maximum score of 40 points, and we could award one point for each correctly circled proverb MWI, and also one point for each distractor that is left uncircled. Thus, an individual’s score is the number of items answered correctly. It would also be possible to employ a correction formula for guessing (see Eyckmans (2004) and Huibregtse, Admiraal, and Meara (2002) for accounts of such formulas).

**Different MWI Measurement Approaches**

**Traditional Test Items and Formats**

When it comes to measuring knowledge of MWIs, the literature abounds with various test item formats or elicitation formats. A useful approach for classifying item formats for single-word vocabulary knowledge has been suggested by Laufer and Goldstein (2004). They distinguish between active and passive types of knowledge and also recognition and recall. This makes it possible to create a $2 \times 2$ matrix whereby four kinds of knowledge can be defined. Based on this work, Schmitt (2010) has argued that the use of terms like “active” and “passive” is somewhat unclear and should be avoided, and that the use of the terms “form” and “meaning” are more transparent and should be used instead. Figure 25.1 shows what Schmitt’s take on Laufer and Goldstein’s original $2 \times 2$ matrix would look like for MWIs rather than single-word items.

<table>
<thead>
<tr>
<th>MWI knowledge aspect given as a prompt</th>
<th>MWI knowledge tested</th>
<th>MWI knowledge tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEANING:</td>
<td>FORM RECALL:</td>
<td>FORM RECOGNITION:</td>
</tr>
<tr>
<td>example: (Swe. L1) klen tröst</td>
<td>example: (Eng. L2) cold comfort</td>
<td>example: (Eng. L2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a. weak comfort</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. feeble comfort</td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. cold comfort</td>
</tr>
<tr>
<td>FORM:</td>
<td>MEANING RECALL:</td>
<td>MEANING RECOGNITION:</td>
</tr>
<tr>
<td>example: (Eng. L2) play with fire</td>
<td>example: (Swe. L1) leka med elden</td>
<td>example: (Swe. L1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a. spela med löggor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. skoja med elden</td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. leka med elden</td>
</tr>
</tbody>
</table>

*Figure 25.1* Matrix for deciding what type of MWI knowledge a test/elicitation item is tapping into

*Source: Adapted from Schmitt, 2010, p. 86*
In order to illustrate the use of the matrix in Figure 25.1, please consider the following item types taken from studies investigating MWI knowledge. In a study by Webb and Kagimoto (2011, p. 268), an L2 form recall item type is used to test L1 Japanese learners of English in terms of their collocation knowledge:

(1) 出世街道 __________

In (1) the L1 Japanese meaning for the English collocation “fast track” is provided and the test taker is asked provide the L2 form. In (2), we see an example of a use of form recognition items, taken from Gyllstad’s COLLEX 3 test format (2007, p. 108). In this item, prompted by the provided L1 Swedish meaning prompt “be en bön” to the left, a test taker is required to select one of the two decontextualized L2 MWI forms provided by ticking the corresponding box to the right. The target MWI in this case is the English collocation *say a prayer*.

(2) (be en bön) a. say a prayer b. tell a prayer

In (3), we find an example of a meaning recall format. The item type is taken from a study by Webb, Newton, and Chang (2013, p. 104) investigating incidental learning of collocation.

(3) meet demand __________

The format requires test takers to supply the meaning of the L2 MWI by writing an L1 translation, which in the case of this particular study was Mandarin Chinese. Finally, in (4), we see an example of a meaning recognition format. The item is taken from McGavigan’s (2009) study of Greek learners’ knowledge of idioms. The test taker is expected to select the most appropriate choice out of the four options provided that best matches the meaning of the idiom, in this case *the silver lining*. Interestingly, the multiple-choice item format is the same as that used in the Vocabulary Size Test (Nation & Beglar, 2007) and also the PHRASE Test (Martinez, 2011).

(4) *the silver lining*
   A. a wedding present
   B. there is always good in bad situations
   C. hidden treasure
   D. something which is mysterious

Yet another item type used to measure productive knowledge of MWIs is exemplified in a study by Schmitt, Dörnyei, Adolphs, and Durow (2004). The study investigated the acquisition of MWIs through two- to three-month-long treatments of exposure to target MWIs in a presessional EAP course. The researchers created and used a productive test that was a type of cloze test. In each item (see (5) below), a range of academically based MWIs were embedded in multi-paragraph contexts with all or most of the content words in the target MWIs deleted. However, the initial letter(s) of each word was left as a form prompt. The meaning of the targeted phrase was provided next to the item in parentheses to
ensure that the ability to produce the form of the MWI was measured, not comprehension of its meaning.

(5) A. The problem is that many people do not want the government to pay the banks. They feel that the banks caused their own problems by lending money too easily.

       B. I see what you mean. Many specialists told the banks that some countries had very weak economies and could not repay the loans. In spite of this, the banks loaned the money anyway. (I understand your argument)

An additional attempt to measure knowledge of MWIs can be found in Schmitt, Grandage, and Adolphs (2004). The study was aimed at investigating the psycholinguistic validity of corpus-derived recurrent clusters, such as as a matter of fact and in the same way as. The authors compiled a list of 25 such items based on corpora and textbooks, and devised a method intended to show whether these items were stored holistically or not in native and non-native-speakers’ mental lexicons (see Conklin, and Siyanova-Chanturia and Omidian, this volume). The measurement was done through a dictation task with oral responses, and the assumption was that if the stretches of dictation are long enough, short term working memory gets cognitively overloaded, and language users consequently have to rely on representations residing in long term memory to be able to repeat back the language stretches. Two examples of stretches used are given next in (6) and (7) (target items bold and underlined):

(6) I didn’t answer, letting his voice drift over me in the same way as the snow drifted over the hills in the distance.

(7) To make a long story short, I threw out his pack and drove off without him. I’ll never pick up a hitchhiker again!

For the target phrase in (6), 14 native speakers (3 correctly and 11 partially incorrectly) (out of 30) and 15 non-native speakers (1 correctly and 14 partially incorrectly) (out of 45) produced this item. For the phrase in (7), 26 native speakers (23 correctly and 3 partially incorrectly) (out of 30) produced this item compared to 39 of the non-native speakers (9 correctly and 30 partially incorrectly) (out of 45). In terms of measurement, the authors...
engage in discussions on how these scores can be interpreted. Furthermore, potential variables affecting the performance are discussed, and correlations between item corpus frequency, length, and transparency of meaning, and native-speaker performance are reported (only transparency of meaning was significant). However, no analysis employing inferential statistics on target item position in the stretches was carried out. The authors conclude that measuring holistic storage this way is fraught with problems, and this is of course the case. But the study showcases innovative attempts at measuring MWI knowledge outside of a controlled lab setting.

Extant Tests Targeting MWI Types

One type of measurement of MWIs involves what can be seen as traditional tests, either as paper-and-pencil tests, or computerized ones. In the literature, there are a number of tests that have been given a proper name and that in most cases have been subjected to more or less extensive validation procedures. Examples of such tests are the Word Associates Test (WAT) (Read, 1993, 1998), the Association Vocabulary Test (Vives Boix, 1995), V_Links (Wolter, 2005), CONTRIX (Revier, 2009, 2014), COLLEX and COLLMATCH (Gyllstad, 2007, 2009), DISCO (Eyckmans, 2009), and The Phrasal Vocabulary Size Test (Martinez, 2011). Despite these efforts, it is clear that there is a lack of truly standardized tests of MWIs, and instead a number of tests exist that target subtypes of MWIs, such as idioms, collocations, phrasal verbs, and lexical bundles.

Of those listed here, WAT is the oldest and probably also the most well-known test of knowledge that is at least partially referable to MWIs. Originally, it was designed to be a test of depth of vocabulary knowledge (Read, 2004). In its most recent configuration (Read, 2000), the test consists of 40 items, in which a prompt adjective word is supplied together with eight potential associated words, four of them adjectives and four nouns (see Figure 25.2).

The test format is a receptive recognition format designed to target two knowledge types: meaning knowledge and collocation knowledge. A test taker is asked to choose those words, in total four, from the left-most box that are synonyms or near-synonyms of the prompt word, and those words from the right-most box that are collocates. The test is capable of providing reliable scores, with Read (1993) reporting reliability coefficients of around .90 using an Item Response Theory (IRT) Rasch analysis (for accessible and nontechnical accounts, see McNamara, 1996 and Knoch & McNamara, 2015), where reliability measures are not dependent on the particular sample of participants tested. The test is in effect a multiple-choice test, with eight choices to be made for each of the 40 items (i.e., 320 choices). Recent validation of the test and its scores (Schmitt, Ng, & Garras, 2011) has shown that finding a suitable scoring system is indeed challenging.

In a volume on L2 collocation research (Barfield & Gyllstad, 2009b), four tests of MWIs are presented. Three of them involve decontextualized test items, and one includes items

![Figure 25.2](source: Adapted from Read, 2000, p. 184)
Measuring Knowledge of Multiword Items

with short sentence contexts. Gyllstad’s (2007, 2009) two contributions – COLLEX and COLLMATCH – are both receptive recognition formats and aimed at tapping into test takers’ knowledge of verb + NP collocations. Both tests are based on corpus analyses of frequent collocates of higher-frequency verbs, and Gyllstad used z-scores (similar to t-scores) from the BNC to inform selection of word combinations as being existing or non-attested in language use.

The COLLEX test (version 5) consists of 50 items of the kind shown in Figure 25.3. In each item, the test taker is asked to choose the option that is a conventional and frequently occurring word combination in the English language out of the three that are presented.

Based on a large-scale administration of the test (N = 269), Gyllstad reported on validity and reliability evidence by examining native-speaker group performance compared to non-native learner groups (L1 Swedish), and by comparing non-native-speaker groups of test takers of different overall proficiency levels. Using Classical Test Theory (CTT) approaches to item quality and test reliability (see Bachman, 2004), Gyllstad showed that a high level of internal consistency is attainable for COLLEX scores (Cronbach’s alpha = .89).

Eyckmans’s (2009) DISCO test is similar to COLLEX in that it is also a receptive recognition test comprising 50 items. In a DISCO item, test takers are asked to choose two options out of three that are deemed idiomatic verb + noun combinations (see Figure 25.4).

The items were constructed based on an initial selection of 40 high-frequency verbs, after which significant collocates were retrieved from the British National Corpus (BNC). Z-scores were used to identify suitable collocations, just like in Gyllstad’s (2007, 2009) studies. However one difference with Eyckmans’ items selection criteria is that she checked the phrasal frequency of her target items (in an additional corpus), and subsequently divided the items into three frequency bands. This is appealing as it implies modeling MWIs in a similar way to single words, in the sense that frequency is a strong predictor of phrase exposure and ease of acquisition on the part of learners. Using a computerized version of DISCO, Eyckmans tested a fairly small group of L1 Dutch learners of English (N = 25). A high level of reliability was observed for the scores (Cronbach’s alpha = .90). An ANOVA showed that “frequency band” was a significant factor, and the author argued

- a. tell a prayer
- b. say a prayer
- c. speak a prayer

**Figure 25.3** An item from COLLEX
*Source: From Gyllstad, 2007, p. 306*

- do time
- pay tribute
- run chores

**Figure 25.4** An example test item in the format used in DISCO
*Source: Adapted from Eyckmans, 2009, p. 146*
that DISCO was sensitive enough to pick up on differences in learners’ developing collocational knowledge.

Gyllstad’s (2007, 2009) COLLMATCH test (version 3) is a 100-item test also aimed at tapping into receptive recognition knowledge. In its earliest inception it started out as a grid format, but in its third version it is in effect a yes/no test. Test takers are asked to indicate for each item (see Figure 25.5 for five example items) whether the word combination exists in the English language. In order to counter overestimation of knowledge, 30 of the 100 items are pseudo-collocations, i.e., non-attested word combinations. For all items, irrespective of category, z-scores were retrieved from the BNC to ensure significance for the target collocations and conversely lack of significance for the pseudo-collocations. Thus, it would be possible to adjust scores for guessing if a suitable adjustment formula can be found (see Pellicer-Sánchez & Schmitt, 2012).

Gyllstad argued that COLLMATCH seems capable of yielding reliable scores (Cronbach’s alpha = .89), and that there is evidence of construct validity as well as concurrent validity (correlation between COLLEX and COLLMATCH at $r = .86$).

The COLLEX, COLLMATCH, and DISCO tests are all formats that require test takers to process decontextualized MWIs, and to match these sequences to potential representations in their mental lexicons. If the test taker has a salient representation that matches the MWI in the test, then they should be able to select this with relative ease. If no matching representation can be activated, then the test taker is likely to draw on other strategies, some of them construct-relevant, such as accessing partial knowledge of component words, and estimating the probabilities of certain words to be combinable with others. Other strategies are arguably more construct-irrelevant (Messick, 1995) to a greater or lesser extent, such as blind guessing and using processes of elimination. On a related note, although expressing some praise for high-quality and innovative test development research, in a critical commentary on the preceding tests, Shillaw (2009) argues that COLLEX, DISCO, and COLLMATCH may tap into test takers’ sensitivity to collocational potential rather than actual knowledge of collocations, and that all these tests are likely to require further validation.

Revier’s (2009, 2014) CONTRIX is different in a number of ways from the three tests described so far. Firstly, CONTRIX is a 45-item test of verb + noun collocations. An example of a CONTRIX item is shown in Figure 25.6. The item consists of a short-sentence context with a gap where one suitable collocation fits. The test taker must choose from the three-column matrix on the right by selecting a verb, a determiner (or none), and a noun that constitutes a conventional MWI in English. In the example, the expected answer is keep + a + diary. Secondly, following semantic classification models (Howarth, 1996; Nesselhauf,
Measuring Knowledge of Multiword Items

2005) the test is aimed at testing MWIs belonging to three categories: transparent, semi-transparent, and nontransparent.

The items for CONTRIX were developed by searching for collocates for 15 high-frequency verbs and creating three test parts of 15 verb + noun collocations, thus in total 45 items. In each part, the same verb appeared but with different noun collocates (e.g., play chess, play tricks, and play the field).

Revier argues perhaps somewhat unconventionally that CONTRIX could 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” (2009, p. 129). This is an interesting claim. If using the matrix adapted from Schmitt (2010) (see Figure 25.1), the format arguably taps into “form recognition”, on the basis that the short-sentence context implicitly hints at a meaning, and the test taker should then find the combination of forms that best fits that implied meaning. However, one could also argue that test takers who are not sure which meaning is targeted in the prompt could look at the provided word forms, and based on a specific combination “recall” the meaning of that MWI. Revier’s CONTRIX item format is thus an illustration of the fact that it is not always straightforward to clearly distinguish between form and meaning aspects of knowledge.

A final test format reviewed here, used to measure learners’ knowledge of MWIs, can be found in Nguyen and Webb’s (2017) receptive knowledge of collocation test. The test targets knowledge of 90 verb + noun and 90 adjective + noun items, and thus consists of a total of as many as 180 items. The items are pairings of words taken from the first thousand most frequent words of English (1K), the second thousand (2K), and the third thousand (3K). Thus, the selection of the target collocations was based on the frequency of the individual words that made up the collocations rather than their frequencies as whole collocations. This highlights the distinction between word-centered and holistic approaches to collocation (see Gyllstad, 2009, p. 169). The format is multiple-choice and two example items can be seen in Figure 25.7.

Much like the WAT, the format tests learners’ ability to recognize which of the verb alternatives in 1 goes with the noun friends. Similarly, they need to choose the adjective in 2 that goes with the noun car. As opposed to the WAT, though, in each item, only one alternative is correct. In terms of measurement reliability, in a study with 100 Vietnamese learners of English, a Cronbach’s Alpha of .77 was observed. With so many items, a higher alpha would be expected, but the authors note the rather homogeneous scores of the test takers, and we know that limited variance leads to lower reliability levels for internal consistency approaches to reliability. The learners’ scores decreased as a function of decreased word frequency level, something which is expected and observed also for single-word knowledge.

| Even as a child, John decided to _____ ___. As an adult, he really likes being able to read about his thoughts and other things that happened to him in his childhood. | push | a/an | secret |
| keep | the | idea |
| pull | – | diary |

**Figure 25.6** An example test item in the format used in CONTRIX

*Source: Adapted from Revier, 2014, p. 89*
Henrik Gyllstad

Analysis of Written/Spoken Texts (Manual and Automated Analyses of MWIs)

In addition to more traditional paper and pencil test formats for measuring knowledge of MWIs, one strand of research has focused on assessing the MWIs used in written and spoken production. This entails getting access to texts produced by learners (or/and native speakers) and using either manual procedures for identification of MWIs, or automated ones, or a combination of the two. Examples of such studies are Granger (1998), Howarth (1996, 1998), Erman and Warren (2000), Nesselhauf (2005), Siyanova and Schmitt (2008), Durrant and Schmitt (2009), Chen and Baker (2010), Forsberg (2010), and Laufer and Waldman (2011).

In studies that employ manual identification/measurement, it is of course essential that a clear set of criteria is used to guide this procedure. Erman and Warren (2000) is an oft-cited study worth taking a closer look at in this regard. The authors analyzed texts from a native-speaker spoken corpus, a native-speaker written corpus, and short extracts from a novel, and found that the average proportion of MWIs in texts (in their study called “prefabs”) was observed at between 52% (written texts) and 59% (spoken texts). There is no bullet point list of criteria for identification of MWIs provided in the paper, but the following main criteria can be discerned upon scrutiny (pp. 31–34):

1. A prefab is a word combination of at least two orthographic words favored by native speakers even though alternatives exist that are not conventionalized.
2. There is restricted exchangeability as to the word components of the prefab.
3. Completely compositional word combinations are not seen as prefabs.

Even though many easily identifiable prefabs appear in texts (e.g., idioms, compounds, non-compositional collocations, prepositional and phrasal verbs), Erman and Warren emphasize that the identification of prefabs is difficult, depending in part on the fact that what one language user sees or uses as a prefab may not be seen or used as such by another language user. This is very much echoed by Read and Nation (2004, p. 25) who when discussing formulaic sequences argue that what is “holistic” varies from person to person, and even varies from time to time within a person. On a critical note, measurement-wise, only single raters seem to have been used in Erman and Warren’s study, and thus no inter-rater reliability (IRR) was established. Best practice would dictate that at least a part of the data be subjected to ratings by at least two raters, using a coefficient for IRR. Commonly used IRR variants are Cohen’s Kappa, Cronbach’s Alpha, and Krippendorff’s Alpha. Hayes and Krippendorff (2007) provide a good discussion of suitability of different coefficients for different types of data. Furthermore, a validity issue resides in the fact that the theoretical definition mentions prefabs favored by native speakers, whereas the material, does not seem to have been checked by or with native speakers.
Another example of a study using a set of criteria for identifying/measuring MWIs can be found in Wray and Namba (2003). Compared to that of Erman and Warren (2000), their approach is much more comprehensive and avoids the pitfall of seeing language sequences as a binary phenomenon of MWIs or not MWIs (in their case formulaic sequences) by using a gradient Likert scale of five steps. No less than 11 criteria are used. Two of the criteria are provided here.

By my judgment, part or all of the wordstring lacks semantic transparency. [1 2 3 4 5]

By my judgment, the wordstring as a whole performs a function in communication or discourse other than, or in addition to, conveying the meaning of the words themselves. [1 2 3 4 5]

In many studies of written production, measurement of MWIs is aided by some kind of frequency criterion and word association measure. As an example, consider Durrant and Schmitt (2009), who analyzed a corpus of 96 NS and NNS texts, ranging in mean length between ~3,000 words and ~600 words. The MWI investigated was adjacent premodifier-noun word pairs. First, through manual extraction, 10,839 such MWIs were collected. Next, two frequency-based methods for determining collocation strength were applied. One of them entailed tallying the frequency of each MWI in the British National Corpus (BNC), and the other computing measures of collocational strength: t-score and Mutual Information (MI) score. Both are statistical measures based on taking the observed frequency of the MWI in a corpus and comparing it with the expected frequency by which it would be expected to appear by chance if looking at the frequencies of the component words. For a technical account of these and other types of word association measures, see Evert (2005), and for easy-to-follow calculation examples, see Schmitt (2010). Whereas the t-score gives us a measure of how confident we can be of there being an association, it has a bias in favor of high-frequency pairs (Evert, 2005, p. 84), e.g., hard work. MI scores, on the other hand, give us a measure of the strength of an association between two words. MI scores have been found to also have a bias but in this case towards word combinations consisting of relatively low-frequency words, but where the words appear rather exclusively with each other, e.g., ultimate arbiter. This applies even when the raw frequency of the word combination as a whole is low (Schmitt, 2010). By convention, t-scores of ≥ 2 and MI scores of ≥ 3 are seen as cutoffs for significant associations. Furthermore, because of the biases mentioned earlier, it is often recommended that MI scores be used in conjunction with a minimum raw frequency threshold (Stubbs, 1995), or in combination with t-scores.

Chen and Baker (2010) is an example of a study in which an automated frequency-driven approach to analyzing texts for MWIs was used. The authors retrieved lexical bundles from one L1 corpus of published academic texts and two L2 corpora of student academic writing, using the corpus tool WordSmith 4.0 (Scott, 2007). Frequency and distribution thresholds for determining four-word lexical bundles were set to ≥ 4 times (approximately 25 times per million words on average) and occurring in at least three texts, respectively. A distinction was also made between different types of bundles (types) and frequencies of bundles (tokens). The authors then relied on a structural classification of lexical bundles in the Longman Grammar of Spoken and Written English (Biber, Johansson, Leech, Conrad, and Finegan, 1999), by which 14 categories of lexical bundles are grouped in conversation and 12 categories in academic prose. The bundles from their search were classified and the results were then compared with the proportions of structural categories in the LSWE corpus.
Future Directions

Increasing Use of Psycholinguistic Techniques for MWI Measurement

Recently, we have seen an increase in the use of more psycholinguistically oriented techniques for MWI measurement. As this use is expected to increase in the future, a brief account is merited. Most of these techniques of measurement make use of processing time as an indirect indicator of how individuals cognitively master comprehension and production of MWIs, and how different types of MWIs are represented in their mental lexicons (see Conklin, Godfroid, this volume).

One type of measurement that has been used to investigate MWIs is the self-paced reading (SPR) task. The task requires participants to read language sequences (words or sentences) on a computer screen, and their reading times are measured in milliseconds (ms). The participant is asked to press a button once the displayed sentence or phrase has been read, and then a new sequence appears. The assumption for MWIs is that, at least for some types, e.g., idioms like *to flog a dead horse*, language users will read these more quickly on the basis that the idiom is represented as a holistic unit which allows for quicker processing compared to a similar language unit (control) like *to flog a dead cat*. An example of a study using the SPR technique is Conklin and Schmitt (2008). See also a study by Tremblay, Derwing, Libben, and Westbury (2011) investigating the effect of displaying target MWIs either word-by-word, phrase-by-phrase, or sentence-by-sentence.

Another useful technique for measuring MWIs is priming. In priming tasks, participants sitting in front of a computer screen are presented with a series of items, in which a word or wordstring is first shown (the prime) followed by another word (the target). If there is a relationship between the prime and the target (e.g., *say* → *prayer*) then the processing of the target is facilitated (processing measured in ms is quicker) compared to when there is no relationship between the prime and the target (e.g., *say* → *table*). It is widely believed that this priming of the target word happens thanks to a process called spreading activation (Collins & Loftus, 1975) by which words perceived to be related to the prime receive a residual level of activation prior to actual presentation. An example of a study using the priming technique is Wolter and Gyllstad (2011). These authors used a primed lexical decision task (LDT) to investigate participants’ collocation knowledge. The participant’s task was to identify whether or not the target string represented a real word in the specified language, by pressing a corresponding key on a keyboard. The time it took for them to respond was recorded as reaction time (RT). RT was faster for collocations (*say* + *prayer*) than for non MWIs (*do* + *prayer*), unrelated items functioning as baselines. Furthermore, RT was also faster for congruent items, i.e., collocations with word-for-word translations in terms of core meanings across Swedish and English, than for incongruent items, i.e., collocations for which word-for-word translations are not possible across the two languages. The sequence of presentation of an item is shown in Figure 25.8.

Eye tracking (ET) has presented itself as another useful technique for measuring knowledge, processing and representation of MWIs (see Conklin, Godfroid, this volume). ET is seen as a more ecologically valid technique compared to SPR. In ET experiments, participants’ eye gaze is measured as they read material on a computer screen. In MWI experiments, target items are embedded as regions of interest and presented in displayed sentences, and the assumption is that fixation, fixation count, first pass reading time, and total reading time (in milliseconds) are indications of how well the participant “knows” the target and the nature of the target representation. An example study using ET is Siyanova-Chanturia, Conklin, and Schmitt (2011). These authors investigated and compared native and
Measuring Knowledge of Multiword Items

non-native-speaker processing of idiomatic MWIs, either as literal items (at the end of the day = in the evening), or as figurative (at the end of the day = eventually). Another study investigating idiom processing in native and non-native speakers through ET is Carrol, Conklin, and Gyllstad (2016). This study was focused on measuring the influence of the L1 on the processing of idioms in an L2. See also Conklin, Pellicer-Sanchez, and Carroll (2018) for a recent and accessible account of the use of ET in applied linguistics.

A final type of measurement technique for MWI measurement is Event-Related Potentials (ERP). ERP is a noninvasive, but rather technical, neurophysiological technique measuring brain function through recordings of electrical activity in the brain. So far, the technique has been scarcely but increasingly used for studies into MWIs. See Siyanova-Chanturia (2013) for a good overview of methods and findings relevant for MWI research.

With a View to the Future: The Construction of MWI Tests

In this subsection, a number of principles for designing tests of MWIs will be suggested (see also Gyllstad & Schmitt, 2018). It stands to reason that although many attempts have been made to create tests measuring knowledge of MWIs, capable of yielding valid and reliable scores, they all have limitations and are more or less in need of further validation. In the following, some key issues will be discussed as to how future tests can be developed.

Test Development and Use for Particular Purposes

Just like many tests of single-word vocabulary, it is not uncommon for MWI tests to be developed for generic use, without indications of what purposes, contexts, or test takers are the specific targets. Thus, there is a need to move away from a situation where one-size-fits-all tests prevail. The testing of MWIs must to a greater extent follow the lead of mainstream testing where tests are developed and validated for specific contexts and learners (Bachman & Palmer, 2010). In doing this, test developers should use standard validation procedures and concomitant statistical methods (see Bachman, 2004). In order to be clear on what tests are designed to measure, developers of MWI knowledge tests should aim at compiling some type of test manual accompanying the test in which the purpose and rationale of the test is provided, observed reliability (ideally sample-independent reliability), evidence of validity for intended test use, and also how to score the test and – importantly – how to interpret these scores.
Selecting What MWIs to Test

At the beginning of this chapter, as well as in other chapters of this volume, it has been argued that MWIs are very common in language. There are also many subtypes, all with their own characteristics in terms of varying length (do sports vs. in a way vs. a watched pot never boils), frequency (blow the gaff [f 5,340 in a Google Books search], face the music [f 225,000 in a Google Books search]), semantic transparency and compositionality (by the end vs. by and large), and function (discourse markers used to signpost text organization, e.g., on the one hand; idioms used to express meaning, e.g., a green light = “getting permission for something to start”). Furthermore, we know that some items can have both a literal and figurative meaning (at the end of the day [= late in the day, time-wise] vs. at the end of the day [= to give one’s final opinion after considering all the possibilities]). Since this is the case, it is essential to have a precise definition of the MWI that is being measured in order to have a chance of obtaining a valid and representative sample.

Many specialized dictionaries exist that focus on various categories of MWIs and can be consulted when selecting suitable MWIs. These modern dictionaries are based on corpora, and they generally have a large number of items, e.g., the Oxford Dictionary of English Idioms (3rd edition) has ~6,000 items; the Oxford Idioms Dictionary for Learners of English (2nd edition, 2006) has ~10,000 items; the Cambridge Idioms Dictionary (2006, 2nd ed.) has ~7,000 items; Oxford Phrasal Verbs Dictionary (2006, 2nd edition) has ~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 ~250,000 items. However, these numbers can be daunting for test developers, when it comes to deciding which MWIs to use. Another approach is to use some kind of frequency list. As an example, Liu (2011) created a condensed list of 150 phrasal verbs based on an initial list of 9,000 phrasal verbs from the British National Corpus. Other lists which provide frequent and pedagogically relevant items include the Pearson Academic Collocation List (2,469 items) (Ackermann & Chen, 2017), the PHRASE List (505 items; Martinez & Schmitt, 2012), the PHaVE List (Garnier & Schmitt, 2015), which covers information about the meaning senses of the most frequent 150 English phrasal verbs, and the Academic Formulas List (Simpson-Vlach & Ellis, 2010), which categorizes the identified formulas according to their function (see Dang, this volume).

As yet another alternative, test developers can use corpora to cull their own selection of MWIs. Mark Davies has made many excellent corpora available to test developers, teachers, and students. They are available online at https://corpus.byu.edu/, and include The Corpus of Contemporary American English (COCA) (Davies, 2008), The British National Corpus (BYU-BNC) (Davies, 2004), and the Corpus of Global Web-Based English (GloWbE) (Davies, 2013). The interface used for these corpora enables test developers to control many aspects of the search criteria for MWIs, such as lemmatized searches, search span, and what specific language structures are targeted. In addition, association measures such as Mutual Information (MI) and raw frequencies are presented.

Sampling

After having chosen an appropriate list or selection of MWIs, test developers have to decide on a suitable number of items to sample for the test. Different language skills require different numbers of items. For example, if testing grammar skills like the application of the simple present third person verb inflection (–s), a fairly limited number of items provide a good indication of knowledge of the construct. This is because with the construct being
measured is generalizable knowledge about a rule or a system. However, MWIs are not rule-based in that sense, and even though frequency is a strong predictor of knowledge of MWIs (but not the only one, as variables like salience and semantic transparency can also have a strong effect (e.g., Wolter & Gyllstad, 2013; Gyllstad & Wolter, 2016)), just like for single words, we cannot strictly expect that learners’ knowledge of one phrasal verb or binomial implies knowledge of another. Ideally, if we know the underlying population of items, we can draw a (random, stratified) sample from that population, test it and then extrapolate to the complete population. This works relatively well if we use a frequency-based model of MWIs. The obvious question is of course what sampling rate can yield valid test scores (see Beglar, 2010; Gyllstad, Vilkaitė, and Schmitt, 2015 for discussions on this). We clearly need to strike a balance between a higher number of items, as this provides better and more valid test information, and a smaller number of items, as this leads to a shorter and more practical test. We have seen that MWI tests developed so far have featured anything from 40 items (WAT), 45 items (CONTRIX), 50 items (DISCO and COLLEX), to 100 items (COLLMATCH). However, for a majority of these, it is not at all clear how scores from these tests relate to the underlying population of items for the specified constructs. Evidently, future tests need to pay closer attention to this issue.

Further Reading


This monograph explains in detail not only how an assessment can be designed but also how an assessment use argument can be employed to justify the way in which the assessment is used.


This book is the second edition of one of the more influential handbooks on vocabulary to date. Chapter 12 (Finding and Learning Multiword Units) describes the nature of MWIs and their learning, and Chapter 13 (Testing Vocabulary Knowledge and Use), even though focusing primarily on single-word vocabulary, provides clear guidelines for important factors to consider for testing, also applicable to MWIs.


On the whole, this book provides a comprehensive account of various statistical concepts and analyses relevant for language assessment and testing. For some more advanced topics, it provides an overview which can serve as a point of departure towards more specialized accounts elsewhere. There is also a companion workbook for those wanting to learn and deepen their knowledge of the material in the main book.

Related Topics

Classifying and identifying formulaic language, factors affecting the learning of multiword items, measuring vocabulary learning progress, sensitive measures of vocabulary knowledge and processing, expanding Nation’s framework, key issues in researching multiword items

References


Martinez, R. (2011, September). Putting a test of multiword expressions to a test. Presentation given at the IATEFL Testing, Evaluation and Assessment SIG, University of Innsbruck, Austria.


