The Routledge Handbook of Vocabulary Studies

Stuart Webb

Key Issues in Researching Multiword Items

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780429291586-33
Anna Siyanova-Chanturia, Taha Omidian
Published online on: 13 Aug 2019

How to cite :- Anna Siyanova-Chanturia, Taha Omidian. 13 Aug 2019, Key Issues in Researching Multiword Items from: The Routledge Handbook of Vocabulary Studies Routledge
Accessed on: 05 Dec 2023
https://www.routledgehandbooks.com/doi/10.4324/9780429291586-33

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Introduction

Recent years have seen an unprecedented interest in multiword items as a linguistic, psycholinguistic, cognitive, and pedagogical phenomenon. They have firmly established themselves as a cornerstone of vocabulary research (see special issues dedicated to learning, processing, and using wordstrings in *Annual Review of Applied Linguistics*, 2012, Vol. 32; *Language Teaching Research*, 2017, Vol. 21:3; *The Mental Lexicon*, 2014, Vol. 9:3; *Topics in Cognitive Science*, Vol. 9:3). Yet some of the issues associated with researching multiword items (MWIs), their definition(s), approaches to their identification, and their properties and characteristics are still poorly understood, under-researched, or even disregarded. The present chapter will center on some of the key issues in researching strings above the word level. Specifically, it will focus on the need and the ways to define MWIs and approaches to studying them. It will also look at the properties of multiword speech that have figured prominently in the literature. In addition, directions and pathways for future research will be outlined. While we attempt to cover a variety of issues and topics that are central to MWI investigation, this review is, of course, selective.

Identifying the key issues in MWI inquiry is no trivial task. Not only do these ubiquitous strings of language vary in a multitude of ways, they have also been researched from a plethora of rather distinct perspectives. Some of the prominent strands of MWI research have included first (L1) and second language (L2) discourse/learner corpus research (Granger, 1998; Howarth, 1998; Nesselhauf, 2005); L1 acquisition (Clark, 1993; Peters, 1983); L2 acquisition (Schmidt, 1983; Vihman, 1982); online processing (Arnon & Snider, 2010; Bannard & Matthews, 2008; Siyanova-Chanturia, Conklin, & van Heuven, 2011; Siyanova-Chanturia & Janssen, 2018); L2 pedagogy (Boers & Lindstromberg, 2008; Meunier, 2012; Wood, 2002); academic discourse (Biber, Conrad, & Cortes, 2004; Hyland, 2008; Durrant & Mathews-Aydinli, 2011); prosody, fluency, and intonation (Lin, 2012, 2018; Van Lancker-Sidtis, 2003; Wood, 2004); pragmatics (Bardovi-Harlig, 2009; Bell, 2012); computational and corpus linguistics (Evert, 2009; Manning & Schutze, 1999), and others. Understandably, each of these strands will have its own key issues and topics in the examination of MWIs, which may be vastly different from those in another field. Despite this heterogeneity,
however, some common aspects that are likely to be of interest to researchers across the various lines of MWI inquiry can be identified.

**Critical Issues and Topics**

**Defining and Identifying Multiword Items**

While the issue of terminology is dealt with in detail elsewhere (Wood, this volume), an important distinction should be made between two commonly used umbrella terms: *formulaic language* and *multiword items/expressions*. The two are often assumed to refer to the same concept and used interchangeably. Yet, for research purposes, in particular, it is important to make an informed decision as to which umbrella term, and consequently the broad underlying phenomenon, is of interest. In this chapter, we use the term *multiword items* because our primary focus is on strings of language that are, by definition, longer than a single word. In contrast, formulaic language embraces multiword as well as single-word items, such as expletives and exclamations (*darn, wow*), conversational speech formulas (*yeah, hello*), and other formulaic items at the (single) word level (Van Lancker-Sidtis & Rallon, 2004; Wray, 2002). This distinction is important, for example, in clinical observations involving individuals with language disorders (aphasia, right hemisphere damage, Alzheimer’s disease and Parkinson’s disease) where one’s ability to produce and comprehend formulaic speech – of any size, length and complexity – is compared with the ability to deal with novel, propositional speech (Bridges & Van Lancker-Sidtis, 2013; Van Lancker-Sidtis, Choi, Alken, & Sidtis, 2015). The focus of the present chapter is thus on the key issues in researching multiword sequences, adjacent and not, but necessarily spanning two or more words.

A distinct but related issue pertains to choosing an approach to identifying MWIs. Although various approaches have figured in the literature, two in particular have been popular. As noted in Siyanova-Chanturia (2015a), some researchers have relied on language speakers’ intuitions (at times, their own) about “prefabricated” or “formulaic” strings of words vs. novel word combinations (Foster, 2001), or free vs. restricted combinations (Howarth, 1998; Nesselhauf, 2005). This approach can be referred to as a qualitative or “phraseological” approach to defining and identifying MWIs. Very loosely, the term phraseology can be defined as “the study of the structure, meaning and use of word combinations” (Cowie, 1994, p. 3168). The focus of this approach is mainly on distinguishing compositional phrases from non-compositional ones with an emphasis on certain semantic criteria, as well as intuitive judgments to determine the degree to which the meaning of an MWI can be inferred from its constituent parts. What lies at the heart of this approach is its reliance on rather impressionistic evidence for the definition and identification of MWIs.

A different approach altogether has been a quantitative, “frequency-based” one in which intuitions and native-speaker judgments play no role (Durrant & Schmitt, 2009). The frequency-based tradition defines an MWI as “the relationship a lexical item has with items that appear with greater than random probability in its (textual) context” (Hoey, 1991, p. 7). Two or more words can be considered an MWI if they are found together more often than the frequencies of their constituents would predict (Jones & Sinclair, 1974; Manning & Schutze, 1999; also see Hoey, 2005; Sinclair, 2004; Stubbs, 1995). By utilizing large text corpora and various statistical metrics, such as frequency and measures of association strengths (e.g., mutual information, t-score), proponents of the frequency-based tradition have developed different methodologies to study the use of MWIs in natural language. Two general approaches have emerged from this tradition: “corpus-based” and “corpus-driven”
Key Issues in Researching Multiword Items

(Tognini-Bonelli, 2001). In the corpus-based approach, corpus evidence is used to analyze different patterns of use for a set of pre-defined sequences. This approach relies heavily on theoretical presumptions regarding the semantic and syntactic properties of MWIs and uses corpus analysis as an evidential tool to validate these pre-established conceptions (Erman & Warren, 2000; Moon, 1998; Nattinger & DeCarrico, 1992). The corpus-driven approach, however, is more inductive in nature, in the sense that sequences emerge from the corpus analysis with little theoretical assumptions guiding their identification (Biber, 2009). In this approach, it is the emerged linguistic patterns that form the foundation of a linguistic theory, and not vice versa.

Neither the more qualitative (intuition-based), nor the more quantitative (frequency-based) approach is perfect; both have their strengths as well as their limitations. For example, while adopting a frequency-based approach can provide us with a systematic description of commonly used linguistic features in a corpus, the extent to which the identified features are the preferred way of expressing the intended meaning in a given context requires a qualitative investigation (Wray, 2002). On the contrary, relying solely on qualitative techniques and intuitive judgments runs the risk of overlooking rather non-salient MWIs (lexical bundles) whose characteristics are more amenable to empirical and frequency-based approaches than intuitive ones (Biber et al., 1999; Durrant & Schmitt, 2009). Because of the clearly complementary nature of these approaches, the two have recently started to be combined (Chen & Baker, 2010; Durrant, 2017; Garnier & Schmitt, 2015; Omidian, Shahriari, & Siyanova-Chanturia, 2018; Simpson-Vlach & Ellis, 2010). For instance, Durrant (2017) adopted a quantitative approach to explore disciplinary variations in the use of lexical bundles. He used the data elicited from this approach as the basis for a qualitative analysis of the identified bundles and their distinctive discourse functions which marked the differences between disciplines.

Whether one approach or a combination of two is adopted, what is critical is to explicitly define MWIs under investigation and, correspondingly, state the approach(es) to identifying them, as these factors will have implications for how the target MWIs will be treated and researched, and how the findings may (or may not) be comparable across relevant studies in the field. These considerations are important in the context of any study focusing on MWIs, irrespective of the methodology adopted, ranging from questionnaire-, test- and corpora-based studies to psycholinguistic explorations, employing reaction times and eye movements.

Properties of Multiword Items

MWIs are extremely heterogeneous and multifarious in nature. They come in various shapes and sizes, vary along all sorts of continua, and are anything but easy to define and classify (see Wood, this volume). In this section, we will cover some of the common aspects of MWIs that have figured in the literature. It is important to understand these properties as they will determine how certain MWIs may or may not be researched, and whether some kinds of MWIs might be more suitable than others for the research question at hand.

Frequency

Frequency is one of the most prominent statistical properties of multiword speech. Language users generally prioritize those sequences that are frequently used in their speech community. Scholars have argued that the frequency with which lexical phrases occur in language determines their degree of prominence among language users (Ellis, 2002;
also see Conklin, this volume). With this understanding, many studies have considered frequency of occurrence as the ultimate arbiter of MWI usefulness. Indeed, it seems reasonable to link frequency of a phrase to its usefulness. However, it is important to note that there are numerous instances of useful MWIs, which are extremely infrequent in even the largest corpora (in for a penny, in for a pound). As Wray (2002) argues, such sequences are commonly used only in the context in which their communicative function is warranted. In other words, the communicative function of a sequence plays an important role in its prominence and usefulness in different contexts. For example, the lexical bundle one of the most performs a basic function in the language, which is the reason why this sequence is highly frequent in virtually all types of language use (Biber et al., 2004).

On the contrary, the function of an expression such as dish the dirt is strongly associated with “gossiping”, which renders the sequence useful and frequent only in that particular context. As Simpson-Vlach and Ellis (2010) point out, many instances of high-frequency word combinations (that there is a) lack distinctive functions or meanings, which questions their value for pedagogical purposes. This issue was addressed by Durrant and Mathews-Aydınlı (2011), who adopted a “function-first” approach in which the corpus under analysis was annotated for various communicative functions, and formulas were then identified as the recurrent linguistic units associated with each function. The authors argued that frequency of occurrence is an insufficient guide to recognizing the communicative functions associated with an MWI.

In addition, frequency in itself does not provide much information regarding the formulaic status of an MWI. In essence, the argument pertains to the assumption that the frequency of a phrase is a function of its formulaicity (Herbst, 2011; Widdowson, 1990). Schmitt, Grandage, and Adolphs (2004) investigated the formulaic status of highly recurrent clusters (extracted from different corpora) by presenting them to participants in a psycholinguistic task. The authors found that the holistic storage of the target sequences varied among participants and frequency alone was insufficient for drawing conclusions regarding the formulaicity of sequences. Schmitt et al. (2004) concluded that “it is unwise to take recurrence of clusters in a corpus as evidence that those clusters are also stored as formulaic sequences in the mind” (p. 147). Similar concerns have been voiced by Ellis (2012) who also stressed the inherent inadequacy of frequency in providing an accurate account of formulaicity. It is thus important to point out that drawing any psycholinguistic conclusions regarding the degree of formulaicity of a phrase solely on the grounds of frequency is indirect and subject to scrutiny (Durrant, 2008; Durrant & Mathews-Aydınlı, 2011; Siyanova-Chanturia, 2015b). As argued by Durrant and Siyanova-Chanturia (2015, pp. 74–75), “it is especially important that we avoid the temptation of automatically ascribing all patterns found in the corpus to features of mind without further interrogation”.

**Familiarity and Predictability**

As noted earlier, frequency is often viewed as a key characteristic of MWIs, defining how these strings of language are acquired, processed, and used in an L1 and L2 (see Boers, this volume; Conklin, this volume). A related but distinct aspect of MWIs is familiarity. It seems sensible to propose that frequency should strongly correlate with familiarity. The more frequent a phrase is, the more likely it is to be familiar to a language user, and subsequently become encoded in their mental lexicon. However, familiarity and frequency are distinct and may be more or less important for various kinds of MWIs. There are sequences which are formulaic and familiar to the linguistic community but are, nevertheless, extremely infrequent.
Key Issues in Researching Multiword Items

For example, the expressions *never too late to mend* and *everything but the kitchen sink* are no doubt familiar to any native speaker of English; yet, they are rather infrequent, with the latter appearing just twice in the British National Corpus (BNC) and the former not being attested at all. Thus, very infrequent MWIs, like certain idioms and proverbs, may still be thought of and treated as highly conventional and formulaic (Hallin & Van Lancker-Sidtis, 2017; Moon, 1998; Reuterskiöld & Van Lancker-Sidtis, 2013). Recently, Hallin and Van Lancker-Sidtis (2017) investigated the prosodic characteristics of familiar infrequent Swedish proverbs spoken by a group of adults and children. In order to confirm the familiarity of the target proverbs, participants were asked to indicate, in a questionnaire, whether or not they had heard or used the target proverbs. The authors found that, although the included proverbs were extremely infrequent (0–0.3 occurrences per million words in a reference corpus), they had been either used or heard by all adult participants and the majority of children. The analyses of the prosodic characteristics revealed that the proverbs, despite their low frequency of occurrence, were spoken significantly faster and with less stressed tonal patterns than matched control sentences, a pattern of results similar to that reported in the studies with highly frequent MWIs (Bybee & Scheibman, 1999; Gregory, Raymond, Bell, Fosler-Lussier, & Jurafsky, 1999; Jurafsky, Bell, Gregory, & Raymond, 2001). This observation indicates that familiarity, irrespective of frequency, plays an important role in how MWIs are processed and used in speech.

One consequence of being a highly familiar string of language is that the sequence becomes uniquely recognizable, or predictable. That is, upon reading the initial constituent(s) of an idiom (*a piece . . .*), a saying (*early bird . . .*), or a collocation (*excruciating . . .*), a proficient speaker is very likely to complete the phrase correctly with *of cake*, *catches the worm*, and *pain*, respectively. Thus, the final constituent(s) of an MWI can be said to be almost redundant, in that they are uniquely anticipated prior to the reader reaching them. This feature of MWIs has, in particular, been interrogated in psycholinguistic (Siyanova-Chanturia et al., 2011; Underwood, Schmitt, & Galpin, 2004) and neurolinguistic studies (Molinaro & Carreiras, 2010; Vespignani, Canal, Molinaro, Fonda, & Cacciari, 2010). The latter suggest that distinct neural mechanisms may be involved in the processing of predictable phrasal information, where Word n is uniquely expected based on the reading of Word n-1. In a study looking at the comprehension of Italian idioms, Vespignani et al. (2010) put forward the notion of categorical template matching – a mechanism that operates specifically for multiword sequences and which is linked to the activation of a template believed to be stored in the semantic memory. This proposition is supported by probabilistic models of language, according to which information about the co-occurrence of single words (the fact that *bride* frequently appears next to *groom* to form a familiar binomial *bride and groom*) is represented in language users’ mind (Gregory et al., 1999; McDonald & Shillcock, 2003). In line with these models, the brain draws on large amounts of statistical information to estimate the probability of Word n following Word n-1 during language comprehension.

**Fixedness and Compositionality**

It is often assumed that multiword speech is fixed or semi-fixed (Moon, 1998; Schmitt, 2004; Wray, 2002). That is, while novel propositional speech is characterized by full syntactic flexibility, multiword speech is, by and large, rather fixed. Indeed, many MWIs are rigid, such that no changes are permitted without the phrase losing its original meaning (*kith and kin*, *by and large*, *first and foremost*). Other MWIs, however, allow some degree of variation, such as adjectival modification (*make a good impression*), quantification (*pull a few strings*),
passivization (*beans were spilled*), pluralization (*red herrings*), and so on. In such cases, the expression – when modified – does not lose its original meaning or the MWI status.

Idioms and their internal structure have, in particular, received attention in the literature. Researchers have proposed that, despite their fixedness, idiomatic expressions undergo analyses not unlike those of propositional speech, and have argued against the “wordlike” nature of idioms. For example, decompositional analyses have been reported in the processing of L1 idioms (Konopka & Bock, 2009; Peterson, Burgess, Dell, & Eberhard, 2001; Snider & Arnon, 2012; Sprenger, Levelt, & Kempen, 2006). In Sprenger et al. (2006), idioms and non-idioms were successfully primed by one of the constituent words, suggesting that during the planning of an idiom, individual components are accessed separately. It was further found that idioms’ literal word meanings were active during the production of idioms. These findings, as well as those of related studies by Konopka and Bock (2009), Peterson et al. (2001), Snider and Arnon (2012), and others in part support the compositional nature of idioms.

Thus, as with other properties, MWIs vary along a continuum of fixedness, with some items being rather fixed (*by and large*) and others being modifiable (*spill the beans → beans were spilled*). What existing evidence also alludes to is that even those MWIs that have traditionally been considered rather fixed – such as idioms – possess internal structure and undergo – at least to some extent – regular decompositional analyses.

The preceding considerations are important because they demonstrate that definitions depending solely on the linguistic aspects of MWIs may not provide a clear picture of their processing and status in the mind. Therefore, it is important for studies focusing on MWIs to avoid drawing psycholinguistic conclusions on the basis of linguistic evidence garnered from text (see Durrant & Siyanova-Chanturia, 2015 for a discussion).

**Adjacency**

Much of the research on MWIs in the field of applied linguistics has overlooked the effect of adjacency on the acquisition, processing and use of these units. Researchers in the neighboring fields such as cognitive science and psycholinguistics, however, have extensively studied the role of this phenomenon. This line of research has largely focused on the learnability and processing of adjacent (contiguous) and nonadjacent (separated by one or more intervening elements) linguistic units at both word and phrase levels (Gomez, 2002; Misyak, Christiansen, & Tomblin, 2009; Newport & Aslin, 2004). The findings reported in these studies point to an important role of adjacency in language. For example, Cleeremans and McClelland (1991) and Misyak et al. (2009) found that sensitivity to nonadjacent and longer-distance strings required more exposure than adjacent ones. Similar differences were reported by Gomez (2002) and Newport and Aslin (2004) who found a higher degree of learnability for adjacent sequences compared to their nonadjacent counterparts.

These studies, however, were largely conducted on lexical items in an artificial language, as the primary purpose of this line of research was to shed light on the capabilities of the human brain in acquiring and processing a language. Few studies have so far looked at adjacent vs. nonadjacent sequences in a natural language. In a corpus-based study, Cheng, Greaves, and Warren (2006) investigated the collocational patterns that may exist between two or more words, using the ConcGram software. The authors found that the majority of the collocational patterns identified by the software were noncontiguous, which was taken to argue that “searches which focus on contiguous collocations present an incomplete picture of the word associations that exist” (p. 431). This issue is acknowledged by Durrant (2009), who points out that the majority of the studies on collocation have exclusively focused on
adjacent MWIs, namely *lexical bundles*, and have overlooked those with nonadjacent constituents.

More recently, MWI adjacency has been explored in a series of psycholinguistic experiments (Vilkaitė, 2016; Vilkaitė & Schmitt, 2017). For example, Vilkaitė (2016) and Vilkaitė and Schmitt (2017) used eye movements to compare the reading of adjacent (*achieve status*) and nonadjacent collocations (*achieve a more secure status*) vs. their respective controls (*ignore status, ignore a more secure status*) in native and non-native speakers of English. Both adjacent and nonadjacent MWIs were read faster than their controls by native speakers (Vilkaitė, 2016). Although the evidence is still rather scarce, we can argue that frequent sequences, irrespective of adjacency, enjoy a processing advantage compared to lower frequency controls. Vilkaitė and Schmitt (2017), however, found that while non-native speakers read adjacent MWIs faster than their controls, they exhibited almost no processing advantage for nonadjacent MWIs relative to their controls. These findings suggest that adjacency may differently affect MWI processing in a more and less proficient language. It is therefore important to take into account the role of adjacency in the acquisition, processing, and use of MWIs when researching these units.

### Polysemy

Meaning variation has long been of interest to semanticists in the context of single words (Ravin & Leacock, 2000). More recently, the existence of polysemes has been documented in the context of strings longer than a word, such as MWIs (Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011; Macis & Schmitt, 2017). For example, Macis and Schmitt (2017) found that a large number of collocations in their study was polysemous, carrying both literal and figurative meanings. In an earlier study, Gardner and Davies (2007) reported that each of the phrasal verbs included in their list (a list of the most frequent English phrasal verbs) had 5.6 meaning senses on average. This led the authors to conclude that researchers “should continue to explore ways of identifying, tagging, and preserving the meaning senses of multiword items” (p. 355).

The polysemic nature of MWIs can complicate matters for researchers. For instance, Macis and Schmitt (2017) observed that the figurative meaning of a polysemous collocation such as *big nose* is exclusive to the world of wine tasting. This indicates that recognizing the multiplicity of meanings associated with certain MWIs relies heavily on understanding their intricate associations with the context in which they commonly occur. As argued by Sinclair (2004), studying linguistic items and their meanings in isolation would only add to their ambiguity. Another challenging aspect of researching polysemous sequences is that the meaning senses they impart may not remain the same across different varieties of English and registers. A case in point is Liu (2011) who found that a number of phrasal verbs on his list had certain meaning senses that were unique to American English and were completely absent in such sources as the BNC. Liu also noted that there is a significant difference between the semantic distribution of phrasal verbs such as *make up* in spoken and academic registers.

Such complications have led researchers to recognize the need to raise awareness of the importance of integrating the semantic analysis of MWIs with their identification (Durrant & Mathews-Aydınlı, 2011; Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011; Wray, 2002). However, while this proposition appears to enrich the research into MWIs, the greatest challenge that researchers face is how to account for the variant meanings of a polysemous sequence. Gardner and Davies (2007) used WordNet (Miller, 2003) to identify
Anna Siyanova-Chanturia and Taha Omidian

different meaning senses of the phrasal verbs on their list. While such electronic resources as WordNet (ibid.) and VIEW (Davies, 2005) have the potential to take account of various meaning senses of a sequence, relying solely on these databases has certain limitations. These limitations are acknowledged by Garnier and Schmitt (2015), who note the inadequacy of lexical databases in providing an accurate account of the meaning senses of highly polysemous phrasal verbs.

Another source of information that can be used for researching polysemous strings is dictionaries. On the surface, due to their relative inclusiveness, dictionaries may seem to provide reliable information about polysemous words or phrases. However, as pointed out by a number of researchers, the information included in dictionaries is limited in that it provides no insight regarding the typicality and usefulness of the listed meanings (Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011). Garnier and Schmitt (2015) demonstrated that the vast majority of occurrences of phrasal verbs can be covered by only a limited number of their senses. This indicates that, akin to words, MWIs have certain prototypical senses that are far more conventional and common than others. The absence of such information in dictionaries renders invisible the prominence of certain polysemes. This brings us to corpus analysis as the most systematic way of identifying frequent patterns of language use. The analysis of language corpora can be useful in capturing the commonality of linguistic items and informing the researcher about the patterns of their usage in the language. However, corpus data alone are not sensitive to possible semantic variations of a linguistic form, thereby requiring the researcher to manually examine different entries associated with each meaning sense. This can prove challenging since frequent MWIs may have several meaning senses dispersed over hundreds, if not thousands, of concordance lines. This discussion leads to the conclusion that, perhaps, the best treatment of polysemy would be to employ an amalgam of the methods available.

Semantic Prosody

Finally, we turn our attention to what is arguably one of the most under-researched aspects of MWIs: semantic prosody. Brought to light for the first time by Sinclair (1987), semantic prosody refers to the affective meaning of a word or phrase, established through its semantic association with a set of collocates (Louw, 1993). According to Louw (2000), the primary function of semantic prosody is to disclose the attitudinal orientations of the speaker/writer towards a certain pragmatic situation. Sinclair (2004) observed that MWIs such as \textit{naked eye} and \textit{true feelings} were semantically associated with notions of “difficulty” and “reluctance/ inability”, respectively. In Sinclair’s view, such prosodies are the pragmatic functions of the node item and its surrounding context. In fact, as opposed to some researchers who view semantic prosody as a property of the node item (Partington, 2004), Sinclair believes that such meanings are beyond the normal semantic values of the core unit and belong to a larger “unit of meaning” (the core and its collocates).

Semantic prosody has been a major finding of corpus analysis. It has been argued that, as a functional/pragmatic concept, semantic prosody is unlikely to be detected by intuition or unaided human observation, and so corpus analysis is the only promising way forward in researching semantic prosody (Louw, 1993; Partington, 2004; Sinclair, 2004; Stubbs, 1996). However, it is important to note that, as mentioned earlier, corpus data can only identify the most frequent patterns of language use, and thus evaluative judgments regarding the semantic associations between the identified patterns should be made by the researcher. This can prove difficult as recognizing the semantic prosody of certain words or word sequences is
not as straightforward as it may appear. For instance, Cheng, Greaves, Sinclair, and Warren (2008) investigated the semantic prosody of different collocational patterns generated by two frequently co-occurring words (referred to as **congrams**): **play/role**. The authors examined modifiers that commonly occurred between these two words (**important, key, major**, and so on) to determine the shared semantic profile of the whole unit of meaning. It was found that the collocational patterns containing the two words **play/role** were semantically associated with the concept of accomplishing “weighty/meaningful” tasks. However, the authors also noticed instances such as “who played only a cameo role” which were not, on the surface, in line with the identified affective meaning. This observation indicates that inferring the semantic prosody of linguistic units from corpus data is no trivial task and warrants a great deal of attention. As Dilt and Newman (2006) point out, this appears to be a common issue among studies on semantic prosody as “the researcher is required to make evaluative judgments in the absence of a set of principled criteria to guide the evaluation” (p. 233). Stewart (2010, p. 134) takes a step further to argue that the traditional link between corpus analysis and semantic prosody should be reconsidered, as the interpretive strategies adopted by analysts to infer semantic associations from corpus data may significantly differ from one researcher to another.

In addition, it has also been shown that the semantic prosody of a lexical item is sensitive to the genre or register in which it occurs (Hunston, 2007; Partington, 2004; Stewart, 2010). This issue was pointed out by Tribble (2000), who argued that “words in certain genres may establish local semantic prosodies which only occur in these genres, or analogues of these genres” (p. 86). For example, Stewart (2010) noted that while the sequence **rigor mortis set in** is generally considered to express a negative attitudinal meaning, it may not communicate the same prosody in scientific registers. This links up with the argument that the attitudinal meaning of a lexical item may not be transferable from one context to another (Whitsitt, 2005). Thus, any generalizations regarding the affective meaning of a word or phrase based solely on evidence garnered from a given genre or register can be tentative at best. These considerations are particularly important for areas of research focusing on the pedagogical aspects of MWIs. In fact, the implied corollary here is that teaching materials targeted at raising learners’ awareness of MWIs should take into account the association between context and meaning at more abstract levels such as semantic prosody.

**Future Directions**

The bulk of evidence reviewed in this chapter point to the many and varied challenges involved in the investigation of MWIs and the combinatory nature of words. As we argued, these challenges introduce key research issues that warrant further examination. This section will recapitulate some of these issues and will outline suggestions for future research.

One area that has received surprisingly little attention is multiword speech that is not contiguous. The vast majority of the studies to date have looked at adjacent MWIs. Yet, as Vilkaite (2016) notes, many noncontiguous strings are as frequent as adjacent combinations. We still know little about the use of nonadjacent MWIs in L1 and L2 speech and writing, as well as the mechanisms involved in the processing of such items. Although the studies by Vilkaite (2016) and Vilkaite and Schmitt (2017) have been valuable contributions to this poorly researched area, they centered on one kind of MWIs (verb + noun collocations) and have looked at native and highly proficient non-native speakers of English. Future research should look at other kinds of nonadjacent MWIs and explore the role of L2 proficiency in the acquisition, processing, and use of nonadjacent items. Further, while it is clear that such
items are ubiquitous in natural language, just how frequent they really are in various corpora (expert, L1 and L2, speech, and writing) remains to be addressed in future research. Finally, another – related – area ripe for future investigations is the way in which some MWIs, such as idioms and proverbs, might be modified (by means of insertion, passivization, pluralization, etc.) and the degree to which speakers of various proficiencies are tolerant to such modifications, as measured by offline (tests, questionnaires) and online measures (reaction times, eye movements).

As was noted earlier, semantic prosody is another aspect of multiword speech that has received little attention. Critically, the majority of the studies in this area have focused on the semantic prosody of single words. This is surprising as the unit of language that initially directed our attention to this phenomenon was the phrasal verb set in (see Sinclair, 1987). There are several reasons to consider the investigation of semantic prosody for MWIs. First, there are certain kinds of MWIs that appear to be commonly used to refer to unpleasant states of affairs, but they have also been attested in contexts with positive pragmatic effects (break out in peace had broken out). For example, Stewart (2010) points out that while the phrasal verb break out is conventionally used in unfavorable contexts (the violence broke out), it can impart positive prosodies when used ironically (peace finally broke out between the two kingdoms; also see Louw, 1993). Secondly, non-salient MWIs, such as lexical bundles, are intuitively considered to have neutral affective meanings in discourse. However, Cortes and Hardy (2013) found that, even in an impersonal genre such as the research article, certain lexical bundles were used mostly in positive contexts (one of the most). This indicates that the affective meaning of MWIs is an implicit aspect of these units that requires further investigation. Finally, as was noted earlier, the semantic prosodies of linguistic units appear to vary across different registers and genres. It remains to be seen whether the semantic prosodies of different kinds of MWIs are also sensitive to the register or genre in which they are commonly used.

A related issue that future research should attend to is polysemy. Even though several studies have consistently pointed out the importance of researching polysemes for pedagogical purposes (Gardner & Davies, 2007; Garnier & Schmitt, 2015, 2016; Liu, 2011; Macis & Schmitt, 2017), research in this area is still scarce. Garnier and Schmitt (2015) is the first study to systematically investigate various meaning senses of one type of MWI (phrasal verbs) and compile a list of the most frequent polysemes of these units in the language. Studies of this sort can further raise awareness about the multiplicity of meanings associated with certain kinds of MWIs and help both teachers and learners navigate the challenges posed by polysemous sequences.

MWIs are by definition highly familiar, conventional strings of language. However, not all MWIs are frequent. In fact, some are extremely infrequent such that they are not attested even in some of the biggest corpora. To the best of our knowledge, the relationship between frequency and familiarity and their respective influence on how MWIs are treated by a language user has not been investigated in detail. Researchers have traditionally looked at the role of frequency in MWI use and processing, where responses to higher frequency items are compared with response to lower frequency ones. However, some research suggests that high familiarity may have facilitative effects (in language processing) not unlike those ascribed to high frequency of occurrence (Hallin & Van Lancker-Sidtis, 2017). That is, highly familiar but infrequent items may be treated similar to frequent MWIs. It would thus be interesting to look at a range of MWIs varying in frequency and familiarity rankings to be able to tease these effects apart and look at their respective roles in language acquisition, processing, and use.
Finally, from a methodological standpoint, new ways of MWI identification, both qualitative and quantitative, can be fruitfully combined (for an example, see Ellis, Simpson-Vlach, & Maynard, 2008), and more powerful ways of data presentation and analysis explored. For example, as Arnon and Snider (2010) note, in the research looking at the role of phrasal frequency in MWI learning and processing, a continuous measure of frequency, rather than a binary one (high vs. low), is a better predictor of the dependent variable under investigation. It is further recommended that researchers use more advanced statistical approaches to data analysis, such as regression analyses and mixed-effects modeling (Baayen, Davidson, & Bates, 2008; Cunnings, 2012), rather than the more commonly used in applied linguistics ANOVAs and t-tests and their non-parametric equivalents. One of the advantages of mixed-effects modeling is the possibility to take into account the effects that unfold during the course of an experimental procedure (such as effects of learning) and to consider a range of relevant covariates (Baayen et al., 2008). Corpus studies can too benefit from more rigorous statistical approaches. For example, Gries and Deshors (2014) propose adopting regression analyses to explore variation and systematic deviations between L1 (baseline) and L2 (learner) corpora. It is argued that such analyses allow for “an unprecedented degree of granularity” and are important for the field of learner corpus research (and applied linguistics) to evolve and advance (Gries & Deshors, 2014, p. 109). A similar argument is put forward by Gries (2015) who suggests using multifactorial regression modeling, rather than the more conventional chi-square or log-likelihood, in which a dependent variable is predicted on the basis of independent variables using a regression equation (Gries, 2015, p. 161).

In conclusion, looking at MWIs and their multifarious nature, it seems virtually impossible to embark upon the study of these units without taking stock of the issues related to their definition, identification, and properties. This chapter has attempted to highlight some of these issues and describe the complexities involved in addressing them. It is hoped that by shedding light on the key aspects of MWIs, researchers will be able to make an informed decision as to what aspects of this linguistic phenomenon should be prioritized for research, and how to navigate the complexities that can potentially limit or change the course of the study at hand.

Further Reading


This corpus-based study investigates the use of MWIs and their communicative functions in essays produced by novice writers and published research articles in matched subject areas. This paper is interesting in that it explores the differences between the two corpora using a “top-down” approach in which function is prioritized over frequency. The implications of this study are important as they offer valuable insights into the connection between MWIs and the communicative function they commonly perform.


The chapter reviews the ways in which psycholinguistics and learner (as well as native-speaker) corpora have interacted in recent years. The evidence of such interdisciplinary interactions is critically evaluated, with a specific focus on corpus data as evidence for psycholinguistic hypotheses about language acquisition and use, and fruitful ways of integrating corpora and experimental research.

This paper investigates various meaning senses of highly frequent phrasal verbs. The authors use a combination of qualitative and quantitative analyses to compile a pedagogical list of phrasal verbs and their highly frequent meaning senses in the English language. This study is important as it attempts to address the pedagogical shortcomings of previous research on phrasal verbs by focusing on the polysemic nature of these units.


This review paper raises important questions with regard to the notion of semantic prosody and how this phenomenon has been defined in the literature. Hunston’s work is enlightening in that it pins down the areas of disagreement that exist among scholars regarding the identification of semantic prosody and how corpus evidence should be interpreted when researching this phenomenon.

**Related Topics**

Aspects of vocabulary knowledge, defining multiword items, the mental lexicon, factors affecting multiword items, learning multiword items, processing multiword items, knowledge of multiword items, issues in researching single-word items

**References**


Anna Siyanova-Chanturia and Taha Omidian


