The Cambridge Handbook of

English Corpus Linguistics

edited by **Douglas Biber**
and Randi Reppen

The Cambridge Handbook of English Corpus Linguistics

Edited by

Douglas Biber

and

Randi Reppen

Northern Arizona University

CAMBRIDGE
UNIVERSITY PRESS

6

Collocation

Richard Xiao

1 Introduction

While the study of recurrent co-occurrence of words dates back to as early as the mid eighteenth century, which saw the publication of Alexander Cruden's concordance of the Bible (Kennedy 1998: 14), serious linguistic research on collocation only started in the 1950s with the British linguist John Firth, who introduced the concept of collocation into modern linguistics. Collocation is one of the linguistic concepts "which have benefited most from advances in corpus linguistics" (Krishnamurthy 2000: 33–34). Indeed, corpus linguistics has not only redefined collocation but has also foregrounded collocation as a focus of research by neo-Firthian linguists as well as those of other traditions.

This chapter aims to provide a critical account of corpus-based collocation research. Following this brief introduction, Section 2 explores the state of the art in collocation research, on the basis of which Section 3 presents a cross-linguistic study of the collocational behavior and semantic prosodies of a group of near synonyms in English and Chinese. Section 4 concludes the chapter by summarizing the research.

2 State of the art in collocation research

This section starts with discussions of the definitional and methodological issues in collocation analysis (Sections 2.1 and 2.2), and then explores the meaning arising from collocation (Section 2.3) and collocational phenomena beyond lexical level (Section 2.4), which is followed by a discussion of the importance of collocation in language use (Section 2.5).

2.1 Definitional issues

The term "collocation" was first used by Firth (1957: 194) when he said "I propose to bring forward as a technical term, meaning by *collocation*, and apply the test of *collocability*." While collocation has been studied in linguistics for more than half a century, there is little consensus on the definition of the term. According to Firth (1968: 181–182), "collocations of a given word are statements of the habitual or customary places of that word." However, the meaning of "habitual or customary" is vague and has been interpreted and operationalized differently in different areas of linguistic research. For example, in traditional discourse analysis, Halliday and Hasan (1976: 287) use collocation as "a cover term for the cohesion that results from the co-occurrence of lexical items that are in some way or other typically associated with one another, because they tend to occur in similar environments," as exemplified by word pairs such as *letter*, *stamp*, and *post office*, or *hair*, *comb*, *curl*, and *wave*. This kind of conceptually based associations between lexical items have been called "coherence collocation," as opposed to "neighbourhood collocation" in corpus linguistics (Scott 2012), i.e. words that actually co-occur with a word in text (e.g. the co-occurrence of *my, this*, and *a* with *letter*). In spite of its name, coherence collocation, which takes a conceptual approach, differs clearly from Firth's (1957: 196) notion of collocation: "Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words." Coherence collocations are hard to measure using a statistical measure. In contrast neighborhood collocations can be retrieved using the computational method developed by Sinclair (1966: 415):

> We may use the term node to refer to an item whose collocations we are studying, and we may then define a span as the number of lexical items on each side of a node that we consider relevant to that node. Items in the environment set by the span we will call collocates.

In addition to coherence and neighborhood collocations, the term has also been used in computational linguistics to refer to a phrase that is semantically non-compositional and structurally non-modifiable and non-substitutable (Manning and Schütze 1999: 184), which, according to Evert (2008), has become better known now as "multiword expression." It is important to note, however, that a multiword expression as commonly understood is not necessarily non-compositional, non-modifiable, or non-substitutable as defined in Manning and Schütze (1999). This phraseological notion of collocation has been investigated under different names including, for example, "lexical bundle" and "word cluster" in corpus linguistics, "multiword unit" and "*n*-gram" in natural language processing, and "formulaic expression" in language education. Unfortunately, collocation has been used to refer to all three different types of recurrent co-occurrence which may or may not overlap with one other.

While collocation analysis has traditionally been concerned with contiguous word associations, recent developments in corpus linguistics have also made it possible to analyze the so-called "concgrams," i.e., sequences of associated words, whether consecutive or non-consecutive, which allow constituency variation (i.e. AB, ACB) and/or positional variation (i.e. AB, BA) (Cheng, Greaves, and Warren 2006: 413–414).

2.2 Methods used in collocation analysis

While some examples of collocation can be identified intuitively, "particularly for obvious cases of collocation" (Greenbaum 1974: 83), intuition is typically a poor guide to collocation. Greenbaum recognized this, and tried to address this problem by eliciting data on collocation from a number of informants "to provide access to the cumulative experience of large numbers of speakers" (ibid.). In those introspection-based elicitation experiments, he found it quite unsurprising that "people disagree on collocations" (ibid.). Intuition, as noted, may not be reliable, "because each of us has only a partial knowledge of the language, we have prejudices and preferences, our memory is weak, our imagination is powerful (so we can conceive of possible contexts for the most implausible utterances), and we tend to notice unusual words or structures but often overlook ordinary ones" (Krishnamurthy 2000: 32–33). Partington (1998: 18) also observes that "there is no total agreement among native speakers as to which collocations are acceptable and which are not." As Hunston (2002: 68) argues, whilst "collocation can be observed informally" using intuition, "it is more reliable to measure it statistically, and for this a corpus is essential." This is because a corpus can reveal such probabilistic semantic patterns across many speakers' intuitions and usage, to which individual speakers have no access (Stubbs 2001a: 153).

As noted earlier, the terms like "habitual" and "customary" as used by Firth (1957) are vague and impressionistic; they rely on the simple frequency counts of co-occurrence (Krishnamurthy 2002). This approach to collocation analysis, which is also adopted by neo-Firthian linguists, is labeled as "collocation-via-concordance" as opposed to "collocation-via-significance" in McEnery and Hardie (2012: 126–127). This latter approach depends on more rigorous inferential statistical tests than simple frequency counts and is now extensively used in collocation analysis. Indeed, the role of statistical tests in collocation analysis was well recognized decades ago, when Halliday (1966: 159) argued that "[the] occurrence of an item in a collocational environment can only be discussed in terms of probability," while Hoey (1991: 6–7) used the term collocation only if a lexical item co-occurs with other items "with greater than random probability in its (textual) context."

A number of statistical formulae are commonly used in corpus linguistics to identify statistically significant collocations, e.g. mutual information (MI), *t*-test, *z*-score test, and log-likelihood test. The MI score is computed by dividing the observed frequency of the co-occurring word in the defined span for the node word by the expected frequency of the co-occurring word in that span and then taking the logarithm of the result, as shown in equation (1):

$$\mathrm{MI} = \frac{\log(F_{n,cN}/F_{nF_c}S)}{\log 2} \qquad (1)$$

In the equation, N stands for the total number of words in a corpus (e.g. 98,313,429 words in the BNC via BNCweb), $\mathrm{F}_{(\mathrm{n})}$ for the frequency count of the node (e.g. *sweet* occurs 3,460 times in the BNC), $\mathrm{F}_{(\mathrm{c})}$ for the frequency of the collocate (e.g. *nothings* occurs 37 times in the BNC), $\mathrm{F}_{(\mathrm{n,c})}$ for the frequency of the node and collocate co-occurring within the defined span (e.g. within 4 words to the left and 4 words to the right of the node, with S=8), or 16 in the BNC example, while log2 is a constant roughly equivalent to 0.301. Based on equation (1), the MI score for the collocation of *sweet* with *nothings* within the 8-word span in the BNC is 10.58.

The MI score is a measure of collocational strength as well as the direction of association (i.e. attraction or repulsion between two lexical items). The higher the MI score, the stronger the association between two items; the closer to 0 the MI score gets, the more likely it is that the two items co-occur by chance. The MI score can also be negative if two items tend to shun each other. Conventionally an MI score of 3.0 or above is taken as evidence that two items are significant collocates (Hunston (2002: 71). In our example above, *nothings* can be said to be a statistically significant collocate of *sweet*.

However, as Hunston (2002: 72) suggests, collocational strength is not always reliable in identifying meaningful collocations. We also need to know the amount of evidence available for a collocation. This means that the corpus size is also important in identifying how certain a collocation is. In this regard, the *t*-test is useful as it takes corpus size into account. The *t*-test is based on the mean and variance of a sample in comparison with the expected mean when the null hypothesis holds. The *t*-score is calculated on the basis of the difference between the observed and expected means, scaled by the variance, to determine the probability of a particular sample of that mean and variance with the assumption of the normal distribution of the dataset, as expressed in equation (2).

$$t\text{-score} = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} \qquad (2)$$

In the equation, $\overline{x}$ and μ respectively represent the mean of the sample and the expected mean, S^2 is the sample variance while N refers to the sample size. In the BNC, for example, the frequency counts of *sweet* and *smell* are 3,460 and 3,508 respectively in N (98,313,429) tokens, and the two words co-occur 90 times within the 8-word span. Then the mean $\overline{x}$ can be expressed as $\frac{90}{98313429}$, i.e. 9.155 × 10^{-7}; the expected mean can be computed using the formula μ = P(*sweet*)P(*smell*), i.e. $(\frac{3460}{98313429})^{*}(\frac{3508}{98313429})$, roughly equivalent to 1.2547 × 10^{-9}. The sample variance S^2 = P(1–P), and for a very small P value, it is roughly equivalent to P, namely $\overline{x}$ in this case. Based on equation (2), the *t*-score for the co-occurrence of *sweet* and *smell* within the 8-word span can be obtained as 9.4.

Conventionally a *t*-score of 2.576 or above is considered to be statistically significant, which means that in our example above, *smell* is a significant collocate of *sweet*. While the MI test measures the strength of collocations, the *t*-test measures the confidence with which we can claim that there is some association (Church and Hanks 1990). As the *t*-test assumes the normal distribution of the population, which can rarely be guaranteed in language use, it is inappropriate to use the *t*-test if the data are known to be skewed. Collocations with high MI scores tend to include low-frequency words, whereas those with high *t*-scores tend to show high-frequency pairs as demonstrated by *nothings* and *smell* in the above examples, which are both among the most significant collocates of *sweet* identified on the basis of the two statistical measures.

The *z*-score test provides a measure of how far a sample is from the mean and in what direction. The test compares the observed frequency with the frequency expected if only chance is affecting the distribution. The *z*-score test is a measure which adjusts for the general frequencies of the words involved in a potential collocation and shows how much more frequent the collocation of a word with the node word is than one would expect from their general frequencies. It can be obtained using equation (3).

$$z\text{-score} = \frac{F_{n,c} - E}{\sqrt{E(1-P)}} \tag{3}$$

In the formula, $F_{n,c}$ and E are respectively the observed and expected frequency counts of co-occurrence while P refers to the probability of the collocate occurring where the node does not occur. P is expressed as $\frac{F_c}{N-F_n}$ and E as $P\,F_nS$, where F_n and F_c are the frequency counts of the node and collocate while N and S stand for the size of the corpus (i.e. token number) and the collocation span respectively. In the above example of *sweet* and *nothings* from the BNC, F_n, F_c and $F_{n,c}$ are 3,460, 37 and 16 respectively, while N and S remain the same. Based on equation (3), the *z*-score for the collocation between *sweet* and *nothings* is 156.78, which is considerably greater than 1.96, the critical value of a *z*-score (absolute value) that can be taken as evidence for significant collocations.

It can be seen from the above that in terms of the procedures of computation, the *z*-score is quite similar to the *t*-score, whereas in terms of output, the *z*-score is more similar to the MI score. A higher *z*-score indicates a greater degree of collocability of an item with the node word. As Dunning (1993) observes, the *z*-score assumes that data are normally distributed, an assumption which is not true in most cases of statistical text analysis unless either enormous corpora are used, or the analysis is restricted to only very common words (which are typically the ones least likely to be of interest). As a consequence, the *z*-score measure, like MI, can substantially overestimate the significance of infrequent words (see Dunning 1993). For lexicographical purposes, these are interesting (e.g. *sweet nothings* and *sweet marjoram*) and should be treated in a general-purpose dictionary. However, for pedagogical purposes, these expressions are of secondary importance compared with more basic collocations.

The solution Dunning proposes for this problem is the log-likelihood (LL) test. The LL measure does not assume the normal distribution of data. For text analysis and similar contexts, the use of LL scores leads to considerably improved statistical results. Using the LL test, textual analysis can be done effectively with much smaller amounts of text than is necessary for statistical measures which assume normal distributions. Furthermore, this measure allows comparisons to be made between the significance of the occurrences of both rare and common features (Dunning 1993: 67).

The log-likelihood test is probably the most complex of the four collocation statistics discussed in this chapter. The LL score is calculated on the basis of a contingency table, as shown in Table 6.1, by adding every cell in the table to the logarithm of that cell and applying the same to multiple combinations of table cells, with the final result multiplied by 2, as indicated in equation (4). In our example of *sweet* and *smell*, the calculated LL score is 688, which is much greater than 3.84, the critical value for statistical significance.

$$\begin{aligned}\mathrm{LL} &= 2 * (a * \log(a) + b * \log(b) + c * \log(c) + d * \log(d) \\ &\quad - (a+b) * \log(a+b) - (a+c) * \log(a+c) - (b+d) * \log(b+d) \\ &\quad - (c+d) * \log(c+d) \ + \ (a+b+c+d) * \log(a+b+c+d)\,) \qquad (4)\end{aligned}$$

Of the four association measures discussed above, the LL test produces consistently better results in collocation extraction by including both common and rare lexical items as collocates. While it is known that MI scores may unduly overvalue infrequent words, it is certainly used widely as an alternative to the LL and *z*-scores in corpus linguistics because of its cognitive relevance for collocations (see McEnery and Hardie 2012: 206, 224). While a standard statistics software package such as SPSS can be used to calculate the statistical test scores discussed above, many popular

Table 6.1 *Contingency table*

	Word A (e.g. *sweet*)	NOT Word A (e.g. *sweet*)
Word B (e.g. *smell*)	a	c
NOT word B (e.g. *smell*)	b	d

corpus analysis tools, both web-based and standalone, also include such statistical measures as built-in functions. For example, the Wordsmith Tools and BNCweb include all of these statistical tests in addition to other options, while AntConc includes MI and *t*-score, and Xaira includes MI and *z*-score.[1]

2.3 Collocational meaning

Shifting from form to meaning, Stubbs (2002: 225) hypothesizes that "there are always semantic relations between node and collocates, and among the collocates themselves." The collocational meaning arising from the interaction between a given node and its typical collocates is known as "semantic prosody," "a form of meaning which is established through the proximity of a consistent series of collocates" (Louw 2000: 57). Both individual words and phrases can have semantic prosodies (cf. Schmitt and Carter 2004: 7). The primary function of semantic prosody is to express speaker/writer attitude or evaluation (Louw 2000: 58). Semantic prosodies are typically negative, with relatively few of them bearing an affectively positive meaning. However, a speaker/writer can also violate a semantic prosody condition to achieve some effect in the hearer – for example irony, insincerity or humor can be explained by identifying violations of semantic prosody (see Louw 1993: 173).[2]

It would appear, from the literature published on semantic prosody, that it is at least as inaccessible to a speaker's conscious introspection as collocation is (see Louw 1993: 173; Partington 1998: 68; Hunston 2002: 142). Yet as the size of corpora has grown, and tools for extracting semantic prosodies have been developed, semantic prosodies have been addressed much more frequently by linguists. Table 6.2 gives some examples of semantic prosodies that have been investigated in the literature.[3]

Semantic prosody that belongs to an item is the result of the interplay between the item and its typical collocates. On the one hand, the item does not appear to have an affective meaning until it is in the context of its typical collocates. On the other hand, if a word has typical collocates with an affective meaning, it may take on that affective meaning even when used with atypical collocates. As the Chinese saying goes, "he who stays near vermilion gets stained red, and he who stays near ink gets stained black" – one takes on the color of one's company – the consequence of a word frequently keeping "bad company" is that the use of the word alone

[1] BNCweb can be accessed at http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php

[2] This view put forward by Louw (1993) is recently challenged by Wei and Li (2014), who distinguish between the Major Patterning and the Minor Patterning, thus calling into question the practice of treating counter-examples as exploiting so-called default prosody of a word for ironical purposes.

[3] A word in small capitals refers to the lemma of the word, including its morphological variations. For example, *CAUSE* refers to *cause, causes, caused*, and *causing*.

Table 6.2 *Examples of semantic prosodies*

Author	Negative prosody	Positive prosody
Sinclair (1991)	BREAK out	
	HAPPEN	
	SET in	
Louw (1993, 2000)	bent on	BUILD up a
	BUILD up of	
	END up *verb*ing	
	GET oneself *verb*ed	
	a recipe for	
Stubbs (1995, 1996, 2001a, 2001b)	ACCOST	career PROVIDE
	CAUSE	
	FAN the flame	
	signs of	
	underage	
	teenager(s)	
Partington (1998)	COMMIT	
	PEDDLE/peddler	
	dealings	
Hunston (2002)	SIT through	
Schmitt and Carter (2004)	bordering on	

may become enough to indicate something unfavorable (see Partington 1998: 67).

In Stubbs's (2002: 225) comment cited above, the meaning arising from the common semantic features of the collocates of a given node word can be referred to as "semantic preference," which is defined "by a lexical set of frequently occurring collocates [sharing] some semantic feature" (ibid.: 449). For example, Stubbs (2001b: 65) observes that *large* typically collocates with items from the same semantic set indicating "quantities and sizes" (e.g. *number(s), scale, part, quantities, amount(s)*). A more detailed study of *large* and its near synonyms such as *big* and *great* is undertaken by Biber, Conrad, and Reppen (1998: 43–53), who compare and contrast these words with respect to their semantic preferences shown by their collocates in a 5.7-million-word sample from the Longman–Lancaster Corpus. Partington (2004: 148) also notes that "absence/change of state" is a common feature of the collocates of maximizers such as *utterly, totally, completely*, and *entirely*.

Semantic preference and semantic prosody are two distinct yet interdependent collocational meanings. Partington (2004: 151) notes that semantic preference and semantic prosody have different operating scopes: the former relates the node word to another item from a particular semantic set whereas the latter can affect wider stretches of text. Semantic preference can be viewed as a feature of the collocates while semantic prosody is a feature of the node word. On the other hand, the two also interact. While semantic prosody "dictates the general environment which constrains the preferential choices of the node item," semantic preference "contributes powerfully" to building semantic prosody (ibid.).

2.4 Beyond lexical collocation

While collocation has commonly been viewed as a phenomenon of the association between individual words, whether in their orthographic or lemma forms, this does not need to be the case. Rather collocational phenomena can occur beyond word level to involve the characteristic co-occurrence between words and phrases with certain grammatical categories and syntactic contexts. Colligation is a type of this kind of higher-level abstraction, which refers to the relationship between words at grammatical level, i.e. the relations of "word and sentence classes or of similar categories" instead of "between words as such" as in the case of collocation (Firth 1968: 181). In other words, while collocation relates to the association between a lexical item and other lexical items (e.g. *very* collocates with *good*), colligation is concerned with the relationship between a lexical item and a grammatical category (e.g. *very* colligates with adjectives). According to Sinclair (1996, 1998) and Stubbs (2001b), the relationship between lexical units can be characterized at four different levels from the lowest to the highest: collocation (the relationship between a node and individual words), colligation (the relationship between a node and grammatical categories), semantic preference (semantic sets of collocates), and semantic prosody (affective meaning of a given node with its typical collocates).

Firth's (1968) definition of colligation has often been applied in a loose sense to refer not only to significant co-occurrence of a word with grammatical categories as noted above but also to significant co-occurrence of a word with grammatical words (e.g. Krishnamurthy 2000). On the other hand, colligation can also mean a word's syntactic preferences or its characteristic co-occurrence with syntactic contexts instead of particular grammatical categories. For example, the word *consequence* is typically used as the head rather than a modifier in a noun phrase (Hoey 2005: 48–49); the adjectives *alone* and *lonely* also display different colligational properties: the former functions as a complement whereas the latter is used as an attributive modifier. Further studies of such lexico-grammatical patterns can be found in Francis, Hunston, and Manning (1996, 1998) and Biber *et al.* (1999).

A specific approach that has been developed to study colligation in this latter sense, i.e. the interaction between lexical items and grammatical structures, is collostructional analysis (Stefanowitsch and Gries 2003). The term *collostruction* is a blend of *collocation* and *construction*, suggesting that this approach is largely concerned with the co-occurrence between lexical items and constructions. Starting with a particular construction, collostructional analysis uses statistical measures to determine the degree to which particular slots in a grammatical structure attract or shun particular lexical items (ibid.: 211). This new approach has certainly provided a useful tool in colligation analysis in its broad sense.

2.5 The importance of collocation in language use

Collocation is part of word meaning. Because of the “mutual expectancy” between two words, “You shall know a word by the company it keeps!” (Firth 1968: 179, 181), as illustrated by Firth’s (1957: 196) example: “One of the meanings of *night* is its collocability with *dark*.” In language comprehension and production, collocation plays an important role in disambiguating a polysemous word, and together with semantic prosody, helps the language user to choose near synonyms appropriately. Making use of collocation is also considered as a useful strategy that helps the language user to convey complex ideas efficiently (Bartsch 2004: 18–19). This is because, according to Hoey’s (2005: 13) lexical priming theory, words are “primed” to co-occur with other particular words (i.e. their collocates), together with their semantic prosody, in the same context.

While a good command of collocation is an important indicator of native-like proficiency and fluency (Bartsch 2004: 20), even a native speaker’s intuition is a poor guide to collocation and collocational meanings (see Section 2.2). Hence, not only has the development of corpus linguistics greatly facilitated collocation analysis but corpus-based collocation research over the past decades has also revealed a range of interesting findings that have hitherto been largely overlooked in previous non-corpus-based language descriptions.

First, while collocation analysis had been undertaken on specific texts such as the Bible before corpus linguistics was established (see Section 1), it is the growing number of large representative corpora and the development of increasingly sophisticated collocation statistics that have enabled large-scale collocation analysis and foregrounded collocation in linguistic research (e.g. Williams 2001; McEnery 2006a; Baker, Gabrielatos, and McEnery 2013), language teaching (e.g. Nattinger and DeCarrico 1992; Nesselhauf 2003), and natural language processing (NLP) tasks such as word sense disambiguation, parsing, computational lexicography, terminology extraction, language generation, and machine translation (see Manning and Schütze 1999; McKeown and Radev 2000). In discourse analysis, for example, collocation and collocational network (i.e. lexical items that are collocationally related), which provide a powerful way of visualizing collocation relationships, are a major form of analysis in McEnery’s (2006a) book-length exploration of discourses about bad language in English. Collocation has proven to be a useful tool in discourse analysis because it can not only reveal patterns of lexical association but also show how a word can acquire meaning in context, which may differ from or even contradict its literal meaning (e.g. semantic prosody). Collocation has also been recognized as playing a central role in language learning and in NLP because the collocational approach treats words not as words in isolation but as large lexical units.

Second, collocation information in dictionaries, if any, has traditionally been provided in the form of illustrative examples. Benson, Benson, and Ilson's (1986/2010) *BBI Dictionary* is probably the best-known dictionary of word combinations in the English language, and is considered a "monumental work" in the field of lexicography (Williams 2001: 63). Apart from specialized collocation dictionaries such as this, however, corpus-based collocation research has helped to produce better dictionaries in general and learner dictionaries in particular, in such a way that collocation information, together with frequency band and authentic examples, is routinely used in dictionaries published since the 1990s, particularly in dictionaries for language learners (Benson 1990; Walker 2009), including for example the *Longman Dictionary of Contemporary English* (3rd or later edition), *Oxford Advanced Learner's Dictionary* (5th or later edition), and *Cambridge International Dictionary of English*.

Third, corpus research has provided more reliable empirical evidence than intuition that facilitates the identification of collocational behavior and semantic prosody of an extensive range of lexical items that have until recently been hidden from intuition. Such knowledge is essential in improving language descriptions in general and in detecting subtle differences between near synonyms in particular (see Sections 2.2 and 3 for further discussion). Explicit teaching of collocation and semantic prosody also has an important role to play in language education (see Nesselhauf 2005).

On the other hand, corpus research on collocation needs to address two important gaps. The first relates to the collocation-via-significance approach. While a variety of statistical formulae are currently available to extract significant collocates, the lists of collocates extracted using different statistics differ sharply (see Section 2.2). Furthermore, existing statistical measures are based on the assumption of random distribution, which is not true in language use (Kilgarriff 2005); the distribution of words of different frequency is also skewed, as indicated by Zipf's law (Evert 2008). For these reasons there is a need to develop more scientific statistical models to address such issues in collocation research.

Second, corpus-based collocation studies have so far focused on, or indeed have largely been confined to, the English language. There has been little work done on collocation and semantic prosody in languages other than English. Still less work has been undertaken which contrasts collocation and semantic prosody in different languages (but see Sardinha 2000; Tognini-Bonelli 2001: 131–156; Xiao and McEnery 2006; Ebeling, Ebeling, and Hasselgård 2013); yet findings yielded by this kind of research can be particularly valuable in language typology, language education, contrastive linguistics, and translation studies.

As such, the case study to be presented in the section that follows will explore collocation and semantic prosody in two genetically distant languages, English and Chinese in this case, from a cross-linguistic perspective rather than in a monolingual context.

3 Collocation and semantic prosody of near synonyms in English and Chinese

By "near synonyms" we mean lexical pairs that have very similar cognitive or denotational meanings, but which may differ in collocational or prosodic behavior (Partington 1998: 77). As such, synonymous words are not collocationally interchangeable (see Conzett 1997: 70–87; Tognini-Bonelli 2001: 34). For example, Halliday (1976: 73) observes that tea is typically described as *strong* rather than *powerful* whereas a car is more likely to be described as *powerful* than *strong*, even though the two modifiers share similar denotational meanings. Likewise, while *weak* and *feeble* have similar cognitive meanings, native speakers of English prefer to say *weak tea* rather than *feeble tea* (Mackin 1978: 150). In addition to different collocational behavior, near synonyms can also differ in semantic prosodies, e.g. *fickle* is negative whereas *flexible* is positive (Tognini-Bonelli 2001: 18–24).

This section explores, from a cross-linguistic perspective, the collocational behavior and semantic prosodies of near synonyms, drawing upon data from two distinctly different languages, English and Mandarin Chinese. The near synonyms selected for cross-linguistic contrast in this study denote *consequence* because these words have been studied in English (Hoey 2005; Stewart 2009) and we want to move from the established patterns of English to investigate the patterns occurring in Chinese by examining the collocational behavior and semantic prosodies of the close equivalents of the words in question in Chinese, with the aim of addressing the following two research questions:

1. Does Chinese exhibit semantic prosody and semantic preference as English does?
2. How different (or similar) are the collocational behavior and semantic prosody of lexical items with similar denotational meanings (i.e. near synonyms) in genetically distant languages such as English and Chinese?

Before these questions are answered, it is appropriate to introduce the corpora and data analysis method used in this study (Section 3.1), which is followed by a discussion of the collocation and semantic prosodies of the chosen group of near synonyms in English (Section 3.2) and a contrastive analysis of the Chinese group (Section 3.3).

3.1 Corpora and data analysis method

As intuition is usually an unreliable guide to patterns of collocation and semantic prosody, this study takes a corpus-based approach to addressing these research questions. The principal corpus data used in this study are the FLOB corpus of British English (Hundt, Sand, and Siemund 1998), the Frown corpus of American English (Hundt, Sand, and Skandera 1999), and the *Lancaster Corpus of Mandarin Chinese* (LCMC; McEnery, Xiao, and Mo 2003). Each of these corpora follows the same corpus design, containing approximately one million words of samples collected from fifteen written text categories published around 1991. FLOB/Frown and LCMC are, as far as is practically possible, comparable corpora well suited for contrastive language research. However, there were points in our research when these corpora were not large enough to provide a reliable basis for quantification. On such occasions a supplementary group of data was also used to extract significant collocates, including the 100-million-word *British National Corpus* (BNC) for English and the *People's Daily* corpus (PDC2000) for Chinese, which covers one year's newspaper texts published in the *People's Daily* in 2000, totaling approximately 15 million words. As the supplementary corpora are not comparable either in sampling period or coverage, we clearly indicate where they are used in this study. These corpora were only used to add further weight to observations made in small comparable corpora.

In collocation analysis we chose to use the MI measure because it is built into the corpus tools we used, WordSmith and Xaira. Both tools allow users to set the minimum co-occurrence frequency of an item to be considered as a collocate of a given node word so that the drawback of the MI measure as noted in Section 2.2 can be partly offset. Given the size of the comparable corpora used, we set the minimum co-occurrence frequency to 3. Within a 4–4 window span, items which had a minimum co-occurrence frequency of 3 and a minimum MI score of 3.0 were accepted as the collocates of a node word. When using additional data from the BNC and PDC2000 corpora, the minimum co-occurrence frequency was set at 20. As we will see from the collocates extracted in Sections 3.2–3.3, these adjustments have allowed us to use the MI score safely.

In our analysis of semantic prosody the positive, neutral, and negative meaning categories correspond to Partington's (2004) favorable, neutral, and unfavorable prosodies. We evaluated each case in context. A pleasant or favorable affective meaning was labeled as positive while an unpleasant or unfavorable affective meaning was judged as negative. When what occurred was completely neutral, or the context provided no evidence of any semantic prosody, the instance was classified as neutral. Note that in the collocate lists presented in Sections 3.2–3.3, items with an unfavorable affective meaning are underlined and those with a favorable affective meaning are emboldened.

3.2 The *CONSEQUENCE* group of near synonyms in English

In English there are a number of words that mean "anything that is due to something already done," e.g. *result, outcome, consequence*, and *aftermath*. Table 6.3 shows the distribution of *CONSEQUENCE* across meaning categories in FLOB/Frown. It is clear from the table that while fixed expressions such

Table 6.3 *Distribution of* CONSEQUENCE *across meaning categories in FLOB/Frown*

Pattern	Negative	Neutral	Positive
as a consequence	6	7	4
in consequence (of)	8	3	1
consequence	27	7	6
consequences	85	20	1
consequent(ly)	15	73	5

as *as a consequence* and *in consequence (of)* can be negative, neutral, or positive, depending upon their contexts, *consequence* and *consequences* show a strong tendency towards a negative semantic prosody. The plural form *consequences* is even more likely to be used negatively. In the BNC, for example, collocates indicating the nature of consequences include, in the order of their co-occurring frequencies, *serious*, ***important***, *disastrous*, *adverse*, *dire*, *far-reaching*, *damaging*, *negative*, *profound*, *unintended*, *major*, *unfortunate*, *tragic*, *fatal*, *new*, *severe* and ***significant***. All of the underlined items express an unfavorable affective meaning.

In FLOB and Frown, significant collocates of *consequences* include (ranked by co-occurring frequency):

- Nature: **important**, adverse
- Affected target: social, financial, economic, ethical, moral, individual, public
- Action: HAVE, (there) BE, ACCEPT

Of the collocates indicating the nature of consequences, *important* is positive while *adverse* is negative. Interestingly, all instances of *important consequences* in FLOB/Frown collocate with *HAVE*/*there BE* to denote a positive pattern meaning. This observation is confirmed in a larger corpus. Of the 68 instances of *important consequences* in the BNC, 54 occurrences follow *HAVE* and one instance follows *there BE*. All 54 examples are positive while the remaining cases may be either positive or neutral.

When they are modified by collocates indicating an affected target, *consequences* are typically negative. As such, actions associated with them normally include *accept*, *alleviate*, *avoid*, *face*, *meet*, *minimize*, *offset*, *(be) responsible*, (*take*) *responsibility*, *suffer*, and *sustain*.[4] *Consequences* sometimes collocates with verbs such as *REAP*, as in (1):

(1) These officials generally attributed their problems to: [. . .] Some critics charged, though, that states were reaping the consequences of profligate spending during the growth years of 1984–1989. (Frown: H)

[4] In this list only *accept* is a significant collocate as defined in this study. In the BNC significant collocates indicating actions include *AVOID, ACCEPT, BE, CAUSE, CONSIDER, FACE, FOLLOW, HAVE*, and *SUFFER*.

REAP typically collocates in its literal meaning with names of crops and *harvest*, or metaphorically with words with a positive meaning such as *benefit(s)* and *rewards* (the three significant collocates are from the BNC). It seems that the apparently paradoxical combination of *REAP* and *consequences* in this example carries the implication that "you reap as you sow": the officials were suffering as a result of their own profligate spending.

In comparison with *consequence(s), aftermath* displays an even more pronounced tendency towards the negative pole of the semantic continuum. In FLOB and Frown, 14 occurrences of *aftermath* were found, mostly in the expression *in the aftermath of*. There is only one significant collocate indicating what causes the state of affairs referred to by *aftermath*. It is *war*. As the low frequency may result in unreliable quantification, we consulted the BNC, which provides 687 instances of *aftermath*. Significant collocates in the BNC typically include *war(s), world* (as in *World War I*) and *Gulf* (as in the *Gulf War*). Clearly these words are negative in their contexts.

Further away from the negative pole of the semantic continuum are *result* and *outcome. Result* is significantly more common than *outcome* (with 677 and 86 occurrences respectively in FLOB/Frown). It appears that both words are associated with a favorable affective meaning, e.g. *a good result*, *a great result*, *an excellent result*, *a brilliant result*, *a successful outcome* (see Hoey 2004a), as reflected by their significant collocates which indicate the nature of a result or outcome:

- Result: **better**, different, early, end, final, similar, direct, empirical, likely, experimental, **good**, negative, **desired**
- Outcome: likely, **positive, successful**

It is of interest to note that *negative* appears on the collocation list of *result*. A close examination of the concordances shows that in all of the three instances *negative* should be interpreted in a mathematical or medical sense, which has no impact upon affective meaning. The discussion above shows that the four near synonyms can be arranged, from positive to negative, on a semantic continuum as follows: *outcome/result, consequence*, and *aftermath*.

3.3 A contrastive analysis of the Chinese group of near synonyms

Shifting to consider these words in contrast, the Chinese translation equivalent of *result/outcome* commonly found in a bilingual dictionary is *jiéguǒ* "result" while the translation equivalent for *consequence/aftermath* is *hòuguǒ* "consequence." In addition, there are a number of obviously positive synonyms such as *chéngguǒ* "achievement" and *shuòguǒ* "great achievement," and negative synonyms including *kǔguǒ* "a bitter pill to swallow" and *èguǒ* "evil consequence."

There are 240 instances of *jiéguǒ* in LCMC, which are distributed across different meaning categories as follows: positive 33, neutral 129, and negative 78. Significant collocates of *jiéguǒ* include:

- Modifiers: *dàxuǎn* "general election," *bìrán* "inevitable," *shìyàn* "experiment," *diàochá* "investigation," *kěnéng* "possible," *jīngjì* "economic," ***hǎo*** "good"
- Actions: *biǎomíng* "show," *zàochéng* "cause," *zēngjiā* "increase," *chǎnshēng* "give rise to; arise," *yǒu* "have"

There are both similarities and differences in the distribution of *result* and its Chinese translation equivalent *jiéguǒ* across meaning categories. On the one hand, like its English counterparts *result* and *outcome, jiéguǒ* typically does not express a negative evaluative meaning. The semantic prosody of *jiéguǒ* is dependent upon its collocates. For example, when it collocates with *zàochéng* "cause," it indicates an unfavorable result; conversely, when it collocates with *chǎnshēng* "bring about," the result is evaluated favorably. The neutral use of *jiéguǒ* was mainly found in academic prose. As there are inherently positive synonyms in Chinese (e.g. *shuòguǒ* and *chéngguǒ*), as noted above, *jiéguǒ* is less frequently used than English *result* to indicate a positive semantic prosody.

In relation to *jiéguǒ, hòuguǒ* is typically negative, though it can be used neutrally, because in some instances there is no evidence of semantic prosody in its context. Of the 22 occurrences of *hòuguǒ* in LCMC, 19 are used negatively, with the remaining three being neutral. The only significant collocate of *hòuguǒ* in LCMC is *yánzhòng* "serious, grave." When the consequences are specified, they typically refer to undesirable situations such as "increasingly intensifying contradictions," "unsold goods piling up in stock," and "inflation." When the consequences are not made clear, there are usually modifiers expressing value judgments or indicating the nature of the consequences. These modifiers normally share a negative semantic preference including, for example, *yánzhòng* "serious, grave," *bùkānshèxiǎng* "too ghastly to contemplate," *bùkěwǎnhuí* "irrecoverable," *bùliáng* "adverse," *xiāojí* "negative," *náncè* "dubious," and *bùyánzìyù* "self-evident." In fact, *hòuguǒ* keeps bad company so frequently that simply using this word alone is usually sufficient to indicate some unfavorable result, as exemplified in (2):[5]

(2) a. nǐ xiǎngxiang nà huì yǒu zěnyàng de hòuguǒ (LCMC: G)
you think-think that will have what GEN consequence
'Just imagine what consequences will result.'

b. hng-hng, nà hòuguǒ, qǐng xiānsheng zìjǐ hǎoshēng
Humph then consequence please sir self carefully

[5] In the grammatical glosses of Chinese examples, ASP stands for *aspect marker*, BA for *ba*-construction, CL for *classifier*, GEN for *genitive*, and PRT for *particle*.

xiǎngxiang ba (LCMC: N)
think-think PRT
"Humph! Then you must think carefully of the consequences."

A more marked contrast was observed in the supplementary Chinese newspaper corpus PDC2000, where 472 instances of *hòuguǒ* were found. Of these, 470 instances show a negative affective meaning, with the remaining two being neutral. The PDC2000 corpus shows two collocates with a minimum co-occurring frequency of 20, *yánzhòng* "serious, grave" and *zàochéng* "cause."

Like English *consequences*, all of the neutral occurrences of *hòuguǒ* in LCMC were found in academic prose (e.g. 3a). *Hòuguǒ* was also found to occur in contexts where an option between a desirable effect (e.g. "peace") and an unpleasant consequence (e.g. "disaster") is available, as shown in (3b). Note, however, that whilst *hòuguǒ* can be used neutrally, the pattern meaning is quite unambiguous – a negative evaluation – when it collocates with verbs like *zàochéng/dǎozhì/zhìshǐ* "cause" (see Xiao and McEnery 2006).

(3) a. shēn céngcì rènshi de hòuguǒ biāozhì-zhe
deep level knowledge GEN consequence mark-ASP
gètǐ yìngfù yìngjī de nénglì (LCMC: J)
individual cope-with emergency GEN ability
"Deep level knowledge allows an individual to cope with emergencies."

b. qí yǐnqǐ de hòuguǒ jiāng bù shì hépíng,
they cause GEN consequence will not be peace
ér shì zāinàn (PDC2000)
but be disaster
"The consequences caused by any of such words and deeds will not be peace but a disaster."

In contrast to the typically positive *jiéguǒ* and the typically negative *hòuguǒ, shuòguǒ*, and *chéngguǒ* are inherently positive whereas *kǔguǒ* and *èguǒ* are inherently negative, regardless of genre. There are 4,572 instances of *chéngguǒ* and 109 instances of *shuòguǒ* in LCMC and PDC2000. The typical collocates of *chéngguǒ* include ***fēngshuò*** "rich and great," ***jiǎng*** "award," *zhuǎnhuà* "transform, turn into," *kējì* "sci-tech," *yánjiù* "research," *qǔdé* "gain," ***yōuxiù*** "excellent," ***gòngxiàn*** "contribution," and *shēngchǎnlì* "productivity." The significant collocates of *shuòguǒ* include *léiléi* "heaps of" and *jiéchū* "yield." *Chéngguǒ* is significantly more frequent than *shuòguǒ*, reflecting the fact that in the real world, results that can be labeled as *shuòguǒ* are considerably fewer than those labeled as *chéngguǒ. Kǔguǒ* occurs 32 times and *èguǒ* 42 times in the two Chinese corpora. All of these are negative, but no significant collocate was found for the two node words.

Like the synonyms of *result* in English, the six near synonyms of *jiéguǒ* in Chinese can be arranged on a semantic continuum, from positive to negative, as follows: *shuòguǒ, chéngguǒ, jiéguǒ, hòuguǒ*, and *kǔguǒ/èguǒ*. Our contrastive analysis of the collocational behavior and semantic prosodies of the two sets of near synonyms in English and Chinese suggests that *result/ outcome* in English and *jiéguǒ* in Chinese can be considered cross-linguistic near synonyms; likewise *consequence/aftermath* in English versus *hòuguǒ* in Chinese are cross-linguistic near synonyms. In relation to English, it appears that Chinese is more sharply divided between the clearly negative and positive ends of the continuum so that the Chinese words *shuòguǒ* and *chéngguǒ* (both highly positive) and *kǔguǒ* and *èguǒ* (both highly negative) can hardly find their cross-linguistic near synonyms in English at lexical level. It is also important to note that unlike English, in which different forms of a lemma may have different collocates and semantic prosodies (e.g. *consequence* vs. *consequences*), Chinese does not have a rich morphology which can affect collocation and semantic prosody in this way.

Our contrastive analysis shows that semantic prosody and semantic preference are as observable in Chinese as they are in English. As the semantic prosodies of near synonyms and the semantic preferences of their collocates are different, near synonyms are normally not interchangeable in either language. It can also be seen from the case study that the semantic prosody observed in general domains may not apply to technical texts. While English and Chinese are genetically distant and distinctly unrelated, the collocational behavior and semantic prosodies of near synonyms are quite similar in the two languages. This observation echoes the findings which have so far been reported for related language pairs, e.g. English vs. Portuguese (Sardinha 2000), English vs. Italian (Tognini-Bonelli 2001: 131–156), and English vs. German (Dodd 2000).

While the corpus-based approach can only reveal but not explain such cross-linguistic similarity, at least part of the explanation, in our view, can be found in the common basis of natural language semantics – "the conceptual system that emerges from everyday human experience" (Sweetser 1990: 1). However, as different languages can have different ranges of near synonyms, near synonyms and their close translation equivalents in different languages may also demonstrate, to some extent, different collocational behavior and semantic prosody. A more general difference between English and Chinese is that collocation and semantic prosody may be affected by morphological variations in English but not in Chinese, which lacks such variation.

4 Conclusion

This chapter has sought to provide a critical account of the current debates in corpus-based collocation research. The first main section (Section 2) explores the state of the art in collocation analysis, covering definitional and methodological issues, meaning arising from collocation, collocational phenomena beyond lexical level, as well as the importance of collocation in language use. The review in this section demonstrates that corpus linguistics has enabled large-scale collocation analysis and foregrounded collocation in linguistic research while corpus-based collocation studies over the past decades have uncovered a range of interesting collocational behavior and semantic prosody which have been hidden from intuition and can only be revealed by examining a large amount of attested data simultaneously. Such findings have not only helped to achieve improved language descriptions, but they also have an important role to play in practical applications such as language teaching, translation, and natural language processing. This review section concludes with a brief discussion of two major gaps to be addressed in future research, namely development of improved statistical measures and cross-linguistic research. To demonstrate the kind of research called for, the second main section in this chapter (Section 3) presents a contrastive study of collocation and semantic prosody in English and Chinese, via a case study of a group of near synonyms denoting *consequence* in the two languages, which suggests that, in spite of some language-specific peculiarities, even genetically distant languages such as English and Chinese display similar collocational behavior and semantic prosody in their use of near synonyms.