Collocation and colligation

Tomas Lehecka

Table of contents

Handbook of Pragmatics 2015. DOI: 10.1075/hop.19.col2

1. Introduction

Collocation and *colligation* are two closely related concepts associated with the distributional properties of linguistic items in actual language use. Specifically, collocation and colligation refer to the likelihood of co-occurrence of (two or more) lexical items and grammatical categories, respectively. Both terms have been attributed to J. R. Firth (1957: 194–195; 1968: 181–183; see Östman and Simon-Vandenbergen 2005 and Shore 2010 for a

summary of Firth's work). Since the terms were introduced, collocation in particular has become a fundamental concept in usage-based studies in many linguistic fields, most notably lexical syntax and semantics. Typically, collocations and colligations are studied in large electronic corpora which allows for statistical analyses of the co-occurrence patterns of linguistic items.

2. Collocation

Collocation refers to the syntagmatic attraction between two (or more) lexical items: morphemes, words, phrases or utterances. Most often, however, collocation analyses have been conducted on the word-level (see discussion in Hoey 2005: 158–159). The concept of collocation is based on the notion that each word in a language prefers certain lexical contexts over others, i.e. that any given word tends to co-occur with certain words more often than it does with others. For example, the word *grass* is often used together with *green*, and the lexeme LETTER is often used together with the lexemes WRITE AND READ (see e.g. Kjellmer 1996: 83). The strength of this kind of attraction between words can be measured through the statistical analysis of corpus data. The purpose of these statistical calculations is to find word pairs with significantly more co-occurrences than what would be expected by chance, given the words' total frequencies in the data. Thus, we can establish the most significant *collocates* of any given word in the language variety that the data represents (Sinclair 1966: 418, Berry-Roghe 1973: 103, Hoey 1991: 6–7).

The syntagmatic attraction, or *collocation strength*, between two words W1 and W2 (a *node* and its *collocate*) is calculated based on four observed absolute frequencies in the data: (i) the total number of word tokens in the corpus, (ii) the number of tokens of W1 in the corpus, (iii) the number of tokens of W2 in the corpus, and (iv) the number of tokens where W1 and W2 co-occur within a specified distance from each other (see *collocation window* below). The *observed* number of co-occurrences in the corpus is compared to the *expected* number of co-occurrences, i.e. the number expected by chance given (i), (ii) and (iii). If the observed number of co-occurrences of W1 and W2 is larger than what can be ascribed to chance, then W2 is a statistically significant collocate of W1.

Since the words in a language have very different frequencies (cf. Zipf's law, Zipf 1935) the collocation strength between different word pairs cannot be compared by considering absolute frequencies alone. Rather, the comparisons are done by using a statistical association measure which takes into consideration the uneven distribution of words in the data. There are currently over 50 different association measures used in statistics (Evert 2009: 1243). The ones that have been used the most in collocation analysis are the z-score, the t-score, MI (Mutual Information), the log likelihood ratio and Fisher's exact test. The choice of an association measure has a great impact on the results of a collocation analysis and therefore requires careful consideration. Unfortunately, many early studies on collocations employed the z-score and MI largely due to them being included in existing corpus software (e.g. AntConc, MonoConc Pro, WordSmith Tools) and not because of any methodological advantages these association measures might have over others (see also Gries 2015). A thorough comparison of the benefits and the downsides of different association measures can be found in Evert 2005; for a shorter discussion see Wiechmann 2008 and Evert 2004 and 2009.

What counts as a co-occurrence depends on how one has defined the *collocation window*, i.e. how far apart the node and the collocate can be in order to be considered as co-occurring. The collocation window is specified as the number of words from the node, for instance 5 words to the left and 5 words to the right of the node (L5–R5). The preferred collocation window size varies between different studies. According to Stubbs (1995a: 32–33), most of the early studies on collocations used L2–R2 or L3–R3 (i.e. 2 or 3 words to each side of the node word, respectively). According to Berry-Roghe (1973) and Jones and Sinclair (1974), L4–R4 is sufficient for uncovering the most significant collocates of a word. However, some scholars have argued that lexical attractions and dependencies manifest themselves even across considerably larger distances (see Siepmann 2005 on long-distance collocation; see also Fried and Östman 2003 and Östman 2005). Most studies on collocations do not take clause or sentence boundaries into consideration when specifying the collocation window (Halliday 1966: 151, Jones and Sinclair 1974: 21, Stubbs 1995b: 246), but some scholars have restricted the definition of collocation to only include clause-internal word combinations (Kjellmer 1987: 133, Smadja 1993).

Traditionally, collocation has been considered as a relationship between only two lexical elements (Jones

and Sinclair 1974, Sinclair 1987). Therefore, the frequent co-occurrence of more than two words has often been treated as a separate phenomenon (however, see Kjellmer 1984 and Smadja 1993). Frequently occurring multiword combinations have been labeled in different ways in the linguistic literature, e.g. *lexical bundles* (Biber and Conrad 1999), *clusters* (Kenny 2000: 99) and *multi-word strings* (Mauranen 2000: 120). In some cases, it has been argued that these multiword combinations are reducible to binary collocations. However, there is evidence to suggest that not all three-item-collocations can be split up into two basic constituents and that multiword collocations might, in fact, be quite common (see the discussion in Siepmann 2005: 417). For a thorough discussion of research on multiword combinations see Biber 2009 and Greaves and Warren 2010.

The term collocation has been used with slightly varying meanings in the linguistic literature. For the most part, that variation concerns the semantic status of collocations. Following the Firthian tradition, many scholars assume a purely statistical definition of collocation. According to this view, collocations are statistically significant co-occurrences of two or more words regardless of the meaning of these word combinations (e.g. Jones and Sinclair 1974, Sinclair 1991, 2004, Stubbs 1996, Hoey 2005). In order to differentiate between the different definitions of the term collocation, Evert (2009) calls this view of the concept *empirical collocation*. Within the tradition of phraseologically oriented research, however, collocation is often defined as a word combination that has been lexicalized to at least some extent, e.g. GIVE – SPEECH, instead of HAVE/MAKE/HOLD – SPEECH (Benson, Benson and Ilson 1986, Mel´čuk 1998, Hausmann 2003). Evert (2009) calls this view of the collocation concept *lexical collation*. Finally, in computational linguistics collocation is traditionally defined as a word combination with idiosyncratic semantic or syntactic properties, i.e. as a non-compositional word combination (Choueka 1988, Manning and Schütze 1999: 184). Sag et al. (2002) and Evert (2009) call this type of collocation *multiword expressions*. A thorough discussion of the different definitions of the term collocation can be found in Bartsch (2004: 27–64).

Collocation analysis is one of the most extensively used methods in corpus linguistics today. It has a fundamental place in the research on contextual semantics, i.e. the description of meaning of lexical items based on their contextual distribution in naturally occurring language. Contextual meaning has sometimes also been re-

ferred to as *collocational meaning* (Leech 1974, Partington 1998). In particular, collocation analysis has been used extensively to compare the meaning of near-synonyms. It has been demonstrated that near-synonymous words often differ as to the lexical contexts they prefer; for example, the adjectives *powerful* and *strong* have a differing set of significant collocates: *a powerful car*, *?a strong car*, *?powerful tea*, *strong tea* (Halliday 1966: 150, Church et al. 1991, Church et al. 1994). Similar comparisons have been made for, among others, *handsome – pretty* (Palmer 1976: 96, Leech 1974: 20), *between – through* (Kennedy 1991), *absolutely – completely – entirely* (Partington 1998), *big – large – great* (Biber, Conrad and Reppen 1998), *almost – nearly* (Kjellmer 2003) and *high – tall* (Taylor 2003). Gries (2001, 2003) has compared the collocates for the English adjective pairs ending in *-ic/-ical* (e.g. *alphabetic – alphabetical*), i.e. for adjectives with differing derivational suffixes.

For a long time now, collocation analysis has been an invaluable tool for lexicographers (Sinclair 1987: 319, Clear 1993: 271, Hoey 1997: 4). Lexicographic research has produced dictionaries which are based on the words' collocation patterns in large corpora, e.g. *Collins COBUILD English Dictionary* (Sinclair 1995), as well as dictionaries *of* collocations (Kjellmer 1994, Lea 2002). Furthermore, collocation analysis has been used widely within the theoretical approach of *frame semantics* and the FrameNet project (e.g. Ruppenhoffer, Fillmore and Baker 2002, Fillmore, Johnson and Petruck 2003). During the last two decades collocation analysis has also been applied in an ever increasing degree in computational linguistics for the purposes of machine translation, natural language processing, as well as vector-space modeling in the field of distributional semantics (Evert 2009: 1217–1218; for an overview and references see Evert 2005: 23–27).

3. Colligation

The term *colligation* has been used in a large number of different senses. In fact, the use of *colligation* has been even more varied than that of *collocation* (see above). Firth (1968: 181) used the term to refer to the syntagmatic attraction between grammatical categories, e.g. parts of speech or syntactic functions (whereas collocation, for him, was the syntagmatic attraction between lexical items). The most common use of the term colligation today, however, is to designate the attraction between a lexical item and a grammatical category (Sinclair 1998: 15,

Stubbs 2001a: 449, Tognini-Bonelli 2001: 163). For example, the English verb BUDGE is attracted to the construction [modal auxiliary verb + BUDGE], e.g. *will/won't budge* (Sinclair 1998: 13). In addition to studying the colligation patterns of single words, the concept of colligation has been applied also to multi-word phrases. For example, the English phrase *naked eye* is often preceded by a preposition and a definite article (e.g. *to the naked eye*, *for the naked eye*; Sinclair 1998: 15).

To date, the most extensive treatment of the concept colligation has been presented by Hoey (2005). In Hoey's theory of lexical priming (i.e. a statistically based theory of linguistic competence), colligations play a crucial part in what it means to know a language. According to Hoey (2005: 43) colligation actually encompasses three distinct aspects of distributional attraction between linguistic items: (i) the relationship between a lexical item and a grammatical context (e.g. [*consequence* + BE + subordinate clause]; Hoey 2005: 57–58), (ii) the relationship between a lexical item and a particular syntactic function in which the item can be used (e.g. *consequence* is often used as part of a complement; Hoey 2005: 44–48), and (iii) the relationship between a lexical item and the position in a phrase, clause, sentence, text or discourse where the item can be used (e.g. *consequence* is often used as part of the theme in a sentence; Hoey 2005: 49–52). Thus, Hoey uses colligation as a cover term which encompasses both grammatical patterns and patterns of information structure associated with a lexical item. It is important to note that all of the relationships above can be positive as well as negative, i.e. lexical items are primed to co-occur with some grammatical features while they are also primed to avoid others.

The main area of study in which colligation analyses have been employed is, in accordance with collocation analyses, the comparative study of near-synonyms. It has been demonstrated that near-synonyms often differ substantially with regard to the grammatical contexts in which they typically occur. For example, the English adjectives *little* and *small*, although similar in meaning, occur in significantly different grammatical contexts: *little* is considerably less likely to be used in the subject complement position (Biber, Conrad and Reppen 1998: 93). Similar comparisons of colligational preferences have been conducted for the English verbs QUAKE and QUIVER (Atkins and Levin 1995), BEGIN and START (Biber, Conrad and Reppen 1998: 95–100) and for Finnish verbs referring to the act of thinking (Arppe and Järvikivi 2007, Arppe 2008). Near-synonyms have also been shown to

prefer different morphosyntactic contexts, e.g. the Finnish adjectives TÄRKEÄ 'important' and KESKEINEN 'central' are primed to occur in different cases in the Finnish case system (Jantunen 2001: 183–185).

At a methodological-theoretical level, the findings of colligation analyses, as well as collocation analyses, suggest that different word forms of the same lexeme have often noticeably different distributional patterns. This effect has been demonstrated in a large variety of languages. For example, the English lexeme SEEK has different significant collocates depending on the inflectional form of the verb. While the forms *seek* and *seeking* show similar preferences regarding their lexical context, i.e. they share many significant collocates (e.g. *advice*, *government*, *political*), the form *seeks* attracts a distinctly different type of words (e.g. *attractive*, *black*, *caring*; Stubbs 2001b: 27–28; see also Sinclair 2004: 31). In similar fashion, Tognini-Bonelli (2001: 96–100) showed that the different inflectional forms of the Italian verb SAPERE 'to know' are primed to occur in different grammatical constructions. Interestingly, the preferred grammatical constructions differ even for the two alternative infinitive forms of the verb, *saper* and *sapere*. *Saper* is used significantly more often in contexts with other infinitives, whereas *sapere* is used more often in subordinate clauses. In Swedish, different inflectional and orthographical forms of adjectives have been shown to prefer different syntactic positions, e.g. the attributive function is preferred significantly more strongly for *stajlad* ('styled') than for the alternative orthographical form *stylad*. In general, the plural forms of Swedish adjectives are used relatively more often in the subject complement position than the singular forms of the same lexemes (Lehecka 2012a, 2012b). In Finnish, a language with a rich case system, the different inflectional forms of the adjectives KAUNIS 'beautiful' and HYVÄ 'good' collocate with different verbs (Jantunen 2004: 27f.).

The findings described in the previous paragraph have considerable implications for how we conceive of the structure of the mental lexicon. The differences in the distributional preferences of different forms of the same lexeme can, at least in certain cases, be seen as an indication of the individual forms having a relatively independent position in the speakers' mental representation of language. In other words, these individual word forms can be seen as having distinctive characteristics in the mental lexicon and might, therefore, be argued to be partially separate items from a cognitive perspective (cf. Sinclair 1991: 8, Stubbs 2001b: 27). Furthermore, Aston

and Burnard (1998: 8) point out that even the different senses of a polysemous word might have different collocation and colligation patterns. From a methodological perspective then, the collocation and colligation studies have demonstrated that distributional preferences of lexical items should, at the very least, be analyzed separately for the individual word forms of given lexemes (Firth 1968: 181, Stubbs 2001b: 27–28).

Finally, it is worth stressing that neither collocational nor colligational preferences of a word form are constant across all domains and registers of language but vary significantly between different types of text and different subcorpora (Biber, Conrad and Reppen 1998: 43–54, Butler 2004: 157, Newman and Rice 2006). For example, according to Firth (1957: 195), the English word *time* collocates normally with such word forms as *saved*, *spent*, *wasted*, *no* and *flies*. In a sub-corpus of sport journalism, however, these collocates are less significant in comparison to e.g. *half*, *full*, *extra* and *injury* (Partington 1998: 17). Correspondingly, Swedish adjectives colligate strongly with copula verbs and tend to occur in the subject complement position in informal socio-pragmatic contexts, but in more formal contexts they are relatively more frequent in the attributive function within a noun phrase (Lehecka 2012a, 2013).

4. Combining collocation and colligation analyses

Many scholars have stressed that the two types of distributional patterns discussed here, collocation (i.e. lexical attraction) and colligation (i.e. grammatical attraction), are closely interrelated. The relationship between these two phenomena has been described and studied in different ways. In what follows, I will present some of the most influential approaches in which collocation and colligation analyses have been combined. The presentation here does not attempt to be exhaustive but, rather, aims at giving a selective overview of the main theoretical and methodological frameworks in which collocation and colligation analyses have been applied together.

Much of the ground-laying work on collocations was done by John Sinclair. Sinclair's classic book *Corpus, Concordance, Collocation (1991)*constitutes a powerful argument for placing the interrelations between different levels of linguistic analysis (morphology, syntax, semantics etc.) at the center of (corpus)linguistic research. Sinclair focuses in particular on the nature of the relationship between the meaning of an utterance and its contex-

tual preferences. For example, the English letter string *second* can be claimed to have two primary meanings based on results of a collocation analysis: (i) 'next to the first' when it is used together with words such as *the*, *world*, *war*, *year*, *child* and *wife*, and (ii) 'a unit of time' when it is preceded by words such as *per*, *radians* and *cycles*. In addition to being attracted to different lexical contexts, the two primary meanings of *second* also prefer different grammatical contexts: *second* as 'next to the first' is often a part of a definite noun phrase, while *second* as 'a unit of time' can be usually found in an indefinite noun phrase (Sinclair 1991: 107). Hence, the collocational preferences of the two senses correlate with the colligational preferences.

According to Sinclair, the meaning of a word is heavily context-dependent and is to be analyzed not in terms of word-inherent semantic characteristics but rather through the lexical and grammatical elements with which the word co-occurs (Sinclair 1991: 108). However, the relationship between a word and its context is reciprocal. The lexical and grammatical context plays a crucial part when a speaker chooses which word he or she will use, but this word then instantly becomes a part of the context for all the other words, thus affecting their meaning (Sinclair 1998: 8). Therefore, when constructing multiword sequences, we co-select words which have compatible lexical and grammatical preferences (Tognini-Bonelli 2001: 101). If follows from this then, that information about the distributional preferences of the individual words or word forms is partly what it means to know a language. The knowledge of the collocational and colligational preferences of individual word forms is thus a part of the speaker's mental representation of the grammar of his or her language. Consequently, one simply cannot make any clear distinction between lexicon and syntax (Sinclair 1966: 411, 1991: 102–105). The grammar of a language, conceived as a set of abstract patterns in a language, is based on generalizations from the usage patterns of individual words and utterances (Sinclair 1991: 100). Thus, lexical items consist not only of meaning and form, but, rather, of a large amount of interrelated properties concerning their preferred patterns of use at several levels of linguistic structure.

Sinclair developed these ideas further in his later work, proposing a more specific account of the kind of distributional preferences included in speakers' mental representation of lexical items. He referred to these mental representations as *extended lexical units* or *extended units of meaning*. According to him, these units en-

compass information about four types of contextual parameters: (i) collocation, (ii) colligation, (iii) semantic preference, and (iv) semantic prosody (Sinclair 1996, 1998, 2004; for a summary of Sinclair's model, see Stubbs 2009: 123–126). By *semantic preference*, Sinclair means the tendency of a lexical unit to co-occur with words from a particular semantic field (in contrast to collocation which signifies the attraction between specific word forms). For example, the adjective *large* collocates generally with words from the semantic field of *Quantity and size* (e.g. with word forms such as *number*, *scale*, *part*, *quantities*, *amount(s)*; Stubbs 2001b: 65). Observations and judgments regarding the semantic preference of lexical items have often been based mainly on intuitive generalizations made from collocate lists. However, determining which semantic field a word represents is often far from straightforward. (For an attempt at a more empirically based method, see Dilts and Newman 2006 and Dilts 2009).

By *semantic prosody* Sinclair means, in turn, the communicative function of a lexical item. Semantic prosody incorporates the reasons for choosing a given lexical item in a specific context (Sinclair 1996: 87–88, 1998: 20, Stubbs 2001b: 65, 2001a: 449). However, it is worth noting that the term semantic prosody has been used in a variety of senses in linguistics. Formerly, the term was used (slightly confusingly also by Sinclair himself) primarily to refer to the feeling or the attitude which a lexical item contributes to an expression (mostly described on the axis from positive to negative; Sinclair 1991: 70–75, Louw 1993: 157, Stubbs 1995a: 42–48, 1995b: 247–250; for more references, see Xiao and McEnery 2006: 106). Hunston (2007) uses the terms *discourse function* and *semantic association* in order to distinguish between the two main senses of semantic prosody. As with semantic preference, it is difficult to determine an objective way of judging the semantic prosody of an item.

The four distributional parameters in Sinclair's model of extended lexical units are closely interrelated. The collocational preferences of a unit have a direct effect on which grammatical features will be found in the typical contexts of the unit (e.g. *budge* colligating with modal auxiliary verbs). In turn, the colligational preferences influence which words or word forms will be represented among the most significant collocates (e.g. *budge* collocates with *will/won't/can/could*). The collocational preferences are reflected at a more abstract level by the se-

mantic preference of a unit (e.g. *budge* is often used in utterances which express refusal). Concurrently, the semantic preference of a unit contributes to shaping the units' semantic prosody which guides the choice of the situational contexts in which the unit is used (e.g. *budge* expresses implicitly a frustration over an object not moving). Naturally, the preferred situational contexts influence in what specific lexical and grammatical contexts the unit will tend to be used, thus having an impact on the collocational and colligational preferences (Sinclair 1998).

Another comprehensive theoretical model that is founded on the concepts of collocation and colligation, is Michael Hoey's (1997, 2005) theory of *lexical priming*. In that framework, Hoey combines corpus linguistic methods with cognitive psychological theory. According to Hoey (2005: 2), statistical (or, in Evert's (2009) words, empirical) collocations explain why certain word combinations sound more natural or fluent than others. For example, *in winter* sounds more natural than *through winter*, which can be explained by the fact that *in winter* is by far the more frequent combination of the two in authentic language data (Hoey 2005: 5–7).

In Hoey's view, collocation is primarily a psychological association between two word forms which is evidenced by their statistically significant tendency to co-occur in corpus data. Collocations, as a psychological phenomenon, are formed by the cognitive process of priming. A speaker's mental representation of language is affected by his or her (implicit) memory of all previously experienced instances of language use. Every item in a speaker's mental lexicon is cumulatively loaded with all the linguistic and extralinguistic contexts in which it has been encountered and, thus, it becomes primed for use in certain contexts over others. Since every speaker has at least a partially different experience of language use, words can be differently primed for each person. Hence, the collocational and colligational preferences for individual word forms vary depending on the speaker (Hoey 2005: 8).

According to Hoey, the priming of lexical items with collocations in this psychological sense is the foundation of language structure in general (2005: 9). The recurrent use of word combinations gives rise to more abstract generalizations concerning language patterns in speakers' minds. These abstract patterns are what we call the grammar of a language. Thus, in Hoey's theory, grammar is an emergent system (for a comprehensive dis-

cussion of emergent grammar, see Hopper 1988, 1998; for psycholinguistic evidence for this theory, see Bybee and Hopper 2001). Every new encounter with speech or writing has an effect on how lexical items are loaded from a collocational perspective. Thus, our mental representation of language is constantly (even if ever so slightly) changing (Hoey 2005: 9).

Hoey (2005: 13) identifies ten different types of linguistic priming, i.e. distributional parameters which are encoded in the mental representation of each lexical unit. The elementary types are the same as Sinclair's: collocation, colligation, semantic preference (although Hoey uses the term *semantic association* for this concept) and semantic prosody (Hoey uses the term *pragmatic association*). Hoey's definition of colligation is, however, substantially wider than Sinclair's. According to Hoey, colligation encompasses not one, but three particular aspects of statistical attraction between lexical items and grammatical categories (see chapter 3 above).

Three of these elementary types of priming (collocation, colligation and semantic preference/association) operate over relatively short distances in a spoken or written co-text (according to Hoey: L4–R4). In addition, Hoey introduces three types of priming that refer to statistical associations at the level of discourse (Hoey 2005: 13). *Textual collocation* refers to the preference of items for having specific cohesive functions within a discourse (e.g. the phrase *Mrs. X* is primed to be followed by *she* later on in the discourse; Hoey and O'Donnell 2008: 193). *Textual semantic association* refers, in turn, to the preference of lexical items to be used in a certain semantic relation to the other items within the discourse (e.g. the phrase *to the right* is often used in a contrasting sense). Finally, *textual colligation* refers to the preference of items to be used in a certain position within a discourse (e.g. the phrase *sixty years ago* is often used as a text beginning in news stories; Hoey 2004: 32). The parameters of priming at all levels of abstraction affect each other (Hoey uses the term *nesting* to describe this phenomenon) and are thus interrelated in a highly intricate manner.

Both Sinclair's and Hoey's work, as well as the majority of corpus linguistic research in general, has focused on the distributional patterns of single word forms or lexemes in a corpus, thus taking the lexical item as a starting point and investigating the contexts and structures in which it occurs. Over the last few decades, however, it has become increasingly popular to study distributional preferences from the opposite perspective as well, i.e. by

starting out from particular grammatical structures or constructions and analyzing which lexical items are statistically most likely to occur in them. For example, Renouf and Sinclair (1991) studied English three-word sequences of the type [function word + content word + function word], e.g. [*a* + N + *of*]. By inspecting the content words most significantly attracted to the respective sequence, they concluded that each such sequence seems to be associated with a particular semantic field. In a similar fashion, Hunston and Francis (2000) studied which words are attracted to different *grammar patterns*, i.e. recurrent grammatical structures which contribute a specific meaning to an utterance. They found, for example, that verbs which occur in the ditransitive grammar pattern fall largely into five identifiable meaning groups, e.g. transfer of commodity (buy, give), communicating (ask, tell) and expressing a feeling (envy, forgive; for a comparison between Hunston and Francis's theory of pattern grammar and Hoey's theory of lexical priming, see Hoey 2009).

Currently, the most widely used approach dealing with the association of lexical items to specific grammatical constructions is probably *collostruction analysis* (for an overview, see Stefanowitsch and Gries 2009). This approach is an extension of both collocation and colligation analyses by combining corpus linguistic methods with the theory of construction grammar (e.g. Fillmore 1988, Goldberg 2006). In fact, collostruction analysis comprises three separate methods. Firstly, collexeme analysis measures the degree of attraction or repulsion of a lexeme to a slot in one particular construction, e.g. the construction [Head N [Modifier*waiting to happen*]] attracts accident and disaster (Stefanowitsch and Gries 2003). Secondly, *distinctive collexeme analysis* measures the preference of a lemma to one particular construction over other functionally similar constructions, e.g. the verbs give, tell and show favor strongly the ditransitive construction over the *to*-dative construction (Gries and Stefanowitsch 2004b). Thirdly, *co-varying collexeme analysis* measures the degree of attraction of lemmas in one slot of a construction to lemmas in another slot of the same construction, e.g. in the *into*-causative construction, [Subj V Obj *into* V-*ing*], the strongest attraction between lexeme pairs in the V and V-*ing* slots is demonstrated by bounce – accept and torture – confess (Gries and Stefanowitsch 2004a).

To date, perhaps the most detailed method of analysis of the interrelations between the distributional properties and preferences of lexical items is the so called *behavioral profiles analysis* (Divjak and Gries 2006, 2008

, Gries 2010a, Gries and Otani 2010). This method has been primarily used for cognitive semantic studies of near-synonyms and antonyms. Behavioral profiles analysis involves manual annotation of the investigated lexical items in the studied corpus with dozens of variables (called ID-tags in accordance with Atkins 1987), including morphological properties (e.g. number, case), syntactic properties (e.g. syntactic function, clause type), semantic properties (e.g. abstract/concrete, animate/inanimate), and collocations. This data is thereafter converted into a contingency table which indicates the relative frequency of co-occurrence between each investigated lexical item and the respective ID-tag. The data in the contingency table is subsequently evaluated using multifactorial statistical analysis, e.g. hierarchic agglomerative cluster analysis, illustrating how similar or dissimilar the investigated items are from a probabilistic perspective. For example, Gries and Otani (2010) showed that the lexemes big and little have a similar usage pattern in English (i.e. the variables have similar values for these lexemes). Interestingly, the usage pattern shared by big and little differs significantly from the one typically employed for another antonym pair, large and small.

5. Conclusion

The concepts of collocation and colligation have had a major impact on the development of corpus linguistic research. As has been demonstrated by the discussion above, these concepts have been refined by different scholars for a variety of purposes depending on which contextual factors are taken into account (e.g. morphosyntactic and pragmatic factors in addition to the lexical and syntactic factors). Furthermore, the different approaches to collocation and colligation have varied regarding the level of abstraction at which the contextual factors are studied (e.g. word forms, lexemes, semantic fields).

In a wider perspective, the numerous approaches combining collocational and colligational data (as well as data on other types of lexical properties), have had notable influence on the development of linguistic methodology in general. The studies within this line of research have demonstrated that the traditional levels of linguistic analysis (e.g. syntax and semantics) should not be treated as independent systems, but that they are closely interconnected. The *correlations* and *dependencies* between for example the collocational and the colligational

preferences should, therefore, constitute one of the main objects of study in linguistic research.

References

Arppe, A.

2008 *Univariate, Bivariate and Multivariate Methods in Corpus-based Lexicography: A Study of Synonymy*. University of Helsinki. http://urn.fi/URN:ISBN:978-952-10-5175-3 [10.10.2014]

Arppe, A. and J. Järvikivi

2007 "Every method counts: Combining corpus-based and experimental evidence in the study of synonymy." *Corpus Linguistics and Linguistic Theory* 3(2): 131–159.

Aston, G. and L. Burnard

1998 *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkins, B.T.S.

1987 "Semantic ID tags: Corpus evidence for dictionary senses." *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary* , 17–36. Waterloo: University of Waterloo.

Atkins, B.T.S. and B. Levin

1995 "Building on a corpus: A linguistic and lexicographical look at some near-synonyms." *International Journal of Lexicography* 8(2): 85–114.

Bartsch, S.

2004 *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence*. Tübingen: Narr.

Benson, M. , E. Benson and R.F. Ilson

1986 *Lexicographic Description of English*. Amsterdam: John Benjamins. doi: 10.1075/slcs.14

Berry-Rogghe, G.

1973 "The computation of collocations and their relevance to lexical studies." In *The Computer and Literary Studies*, ed. by A.J. Aitken , R.W. Bailey and N. Hamilton-Smith , 103–112. Edinburgh: Edinburgh University Press.

Biber, D.

2009 "A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing." *International Journal of Corpus Linguistics* 14(3): 275–311. doi: 10.1075/ijcl.14.3.08bib BoP

Biber, D. and S. Conrad

1999 "Lexical bundles in conversation and academic prose." In *Out of corpora. Studies in honour of Stig Johansson*, ed. by H. Hasselgård and S. Oksefjell , 182–190. Amsterdam: Rodopi.

Biber, D. , S. Conrad and R. Reppen

1998 *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511804489 BoP

Butler, C.S.

2004 "Corpus studies and functional linguistic theories." *Functions of Language* 11(2): 147–186. doi: 10.1075/fol.11.2.02but

Bybee, J. and P. Hopper

(eds.) **2001** *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins. doi: 10.1075/tsl.45 BoP

Choueka, Y.

1988 "Looking for needles in a haystack." In *Proceedings, RIAO Conference on User-oriented Context Based Text and Image Handling* , 609–623. Cambridge, MA.

Church, K. , W. Gale , P. Hanks and D. Hindle

1991 "Using statistics in lexical analysis." In *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, ed. by U. Zernik , 115–164. Hillsdale: Lawrence Erlbaum.

Church, K. , W. Gale , P. Hanks , D. Hindle and R. Moon

1994 "Lexical substitutability." In *Computational Approaches to the Lexicon*, ed. by B.T.S. Atkins and A. Zampolli , 153–177. Oxford: Oxford University Press.

Clear, J.

1993 "From Firth principles: Computational tools for the study of collocation." In *Text and Technology: In Honour of John Sinclair*, ed. by M. Baker , F. Gill and E. Tognini-Bonelli , 271–292. Philadelphia: John Benjamins. doi: 10.1075/z.64.18cle

Dilts, P.

2009 "Good nouns, bad nouns: What the corpus says and what native speakers think." *Language and Computers*

71(1): 103–117.

Dilts, P. and J. Newman

2006 "A note on quantifying 'good' and 'bad' prosodies." *Corpus Linguistics and Linguistic Theory* 2(2): 233–242.

Divjak, D. and S.T. Gries

2008 "Clusters in the mind? Converging evidence from near synonymy in Russian." *The Mental Lexicon* 3(2): 188–213. doi: 10.1075/ml.3.2.03div

2006 "Ways of trying in Russian: Clustering behavioral profiles." *Corpus Linguistics and Lingustic Theory* 2(1): 23–60.

Evert, S.

2009 "Corpora and collocations." In *Corpus Linguistics: An International Handbook, Vol. 2*, ed. by A. Lüdeling and M. Kytö , 1212–1248. Berlin, New York: Mouton de Gruyter. doi: 10.1515/9783110213881.2.1212

2005 *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

2004 *Association Measures*. www.collocations.de/AM/ [10.11.2014].

Fillmore, C.J.

1988 "The mechanisms of 'Construction Grammar'." In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, ed. by S. Axmaker , A. Jaisser and H. Singmaster , 35–55. Berkeley: Berkeley Linguistics Society.

Fillmore, C.J. , C.R. Johnson and M.R.L. Petruck

2003 "Background to FrameNet." *International Journal of Lexicography* 16(3): 235–250. doi: 10.1093/ijl/16.3.235

Firth, J.R.

1968 *Selected Papers of J. R. Firth 1952–59*. London: Longmans.

1957 *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Fried, M. and J. Östman

2003 "The explicit and the implicit in the Suasion Frame." In *Proceedings of CIL17*, ed. by E. Hajičová , A. Kotěšovcová and J. Mírovský , 1–22. Prague: Matfyzpress.

Goldberg, A.E.

2006 *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press. BoP

Greaves, C. and M. Warren

2010 "What can a corpus tell us about multi-word units." In *The Routledge Handbook of Corpus Linguistics*, ed. by M. McCarthy and A. O'Keeffe , 212–226. Abingdon: Routledge.

Gries, S.T.

2015 "Some current quantitative problems in corpus linguistics and a sketch of some solutions." *Language and Linguistics* 16 (1): 93–117. doi: 10.1177/1606822X14556606

2010a "Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics." *The Mental lexicon* 5(3): 323–346. doi: 10.1075/ml.5.3.04gri

2003 "Testing the sub-test: An analysis of English -ic and -ical adjectives." *International Journal of Corpus Linguistics* 8(1): 31–61. doi: 10.1075/ijcl.8.1.02gri

2001 "A corpus-linguistic analysis of English -ic vs. -ical adjectives." *ICAME Journal* 25. 65–108.

Gries, S.T. and A. Stefanowitsch

2004a "Co-varying collexemes in the into-causative." In *Language, Culture, and Mind*, ed. by M. Achard and S. Kemmer , 225–236. Stanford, CA: CSLI.

2004b "Extending collostructional analysis: A corpus-based perspective on 'alternations'." *International Journal of Corpus Linguistics* 9(1): 97–129. doi: 10.1075/ijcl.9.1.06gri

Gries, S.T. and N. Otani

2010 "Behavioral profiles: A corpus-based perspective on synonymy and antonymy." *ICAME Journal* 34. 121–150.

Halliday, M.A.K.

1966 "Lexis as a linguistic level." In *In Memory of J. R. Firth*, ed. by C.E. Bazell , J.C. Catford , M.A.K. Halliday and R.H. Robins , 148–163. London: Longman.

Hausmann, F.J.

2003 "Was sind eigentlich Kollokationen?" In *Wortverbindungen – mehr oder weniger fest*, ed. by K. Steyer , 309–334. Berlin: Walter de Gruyter.

Hoey, M.

2009 "Corpus-driven approaches to grammar." In *Exploring the Lexis-Grammar Interface*, ed. by U. Römer and R. Schulze , 33–48. Amsterdam: John Benjamins. doi: 10.1075/scl.35.04hoe

2005 *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

2004 "The textual priming of lexis." In *Corpora and Language Learners*, ed. by G. Aston , S. Bernardini and D. Stewart , 21.42. Amsterdam: John Benjamins. doi: 10.1075/scl.17.03hoe

1997 "From concordance to text structure: New uses for computer corpora." In *PALC '97. Proceedings of Practical Applications of Linguistic Corpora conference*, ed. by B. Lewandowska-Tomaszczyk and J. Melia . Lodz: Lodz University Press. 2–23.

1991 *Patterns of Lexis in Text*. Oxford: Oxford University Press. BoP

Hoey, M. and M.B. O'Donnell

2008 "The beginning of something important? Corpus evidence on the text beginnings of hard news stories." In *Corpus Linguistics*, Computer Tools and Applications: State of the Art, ed. by B. Lewandowska-Tomaszczyk. PALC 2007: 189–212.

Hopper, P.

1998 "Emergent grammar." In *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, ed. by M. Tomasello , 155–176. Hillsdale, NJ: Erlbaum. MetBib

1988 "Emergent grammar and the a priori grammar postulate." In *Linguistics in Context: Connecting Observation and Understanding. Lectures from the 1985 LSA/TESOL and NEH Institutes*, ed. by D. Tannen , 117–134. Norwood, NJ: Ablex.

Hunston, S.

2007 "Semantic prosody revisited." *International Journal of Corpus Linguistics* 12(2): 249–268. doi: 10.1075/ijcl.12.2.09hun

Jantunen, J.H.

2004 *Synonymia ja käännössuomi. Korpusnäkökulma samamerkityksisyyden kontekstuaalisuuteen ja käännöskielen leksikaalisiin erityispiirteisiin*. (University of Joensuu publications in the humanities 35.) Joensuu: Joensuun Yliopisto. http://epublications.uef.fi/pub/urn_isbn_952-458-479-4/urn_isbn_952-458-479-4.pdf [20.8.2014]

Jones, S. and J. Sinclair

1974 English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie* 24(1): 15–61.

Kennedy, G.

1991 "Between and through: The company they keep and the functions they serve." In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, ed. by K. Aijmer and B. Altenberg , 95–110. London: Longman.

Kenny, D.

2000 "Lexical hide-and-seek: Looking for creativity in a parallel corpus." In *Intercultural faultlines. Research models in translation studies I. Textual and cognitive aspects*, ed. by M. Olohan , 93–104. Manchester: St.

Jerome Publishing. TSB

Kjellmer, G.

2003 "Synonymy and corpus work: On almost and nearly." *ICAME Journal* 27: 19–27.

1996 "Idiomen, kollokationerna och lexikonet." *Lexico-Nordica* 3: 79–90.

1987 "Aspects of English collocations." In *Corpus Linguistics and Beyond: Proceedings of the Seventh International Conference on English Language Research on Computerised Corpora*, ed. by W. Meijs . Amsterdam: Rodopi.

1984 "Some thoughts on collocational distinctiveness." In *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*, ed. by J. Aarts and W. Meijs , 163–171. Amsterdam: Rodopi.

(ed.) **1994** *A Dictionary of English Collocations, Vol.* 1–3. Oxford: Clarendon Press.

Lea, D.

(ed.) **2002** *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Leech, G.

1974 *Semantics: The Study of Meaning*. Harmondsworth: Penguin Books.

Lehecka, T.

2012a *Interrelaterade lexikala egenskaper. Engelska adjektivimporter i en svensk tidningskorpus.* (Nordica Helsingiensia 32.) Helsingfors: Helsingfors universitet. http://urn.fi/URN:ISBN:978-952-10-8530-7. [20.8.2014]

2012b "Probabilistisk syntaktisk analys av engelska adjektiv i svensk tidningstext." *Maal og Minne* 104(1): 72–109.

2013 "Kollokationer och kolligationer: Om förhållandet mellan adjektivens semantiska och syntaktiska preferenser." *Folkmålsstudier* 51: 49–85.

Louw, B.

1993 "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies." In *Text and technology: In honour of John Sinclair*, ed. by M. Baker , G. Francis and E. Tognini-Bonelli , 157–176. Philadelphia/Amsterdam: John Benjamins. doi: 10.1075/z.64.11lou

Manning, C.D. and H. Schütze

1999 *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mauranen, A.

2000 "Strange strings in translated language: A study on corpora." In *Intercultural Faultlines Research Models in Translation Studies*, ed. by M. Olohan , 119–141. Manchester: St. Jerome Publishing. TSB

McEnery, T. , R. Xiao and Y. Tono

2006 *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Mel'čuk, I.

1998 "Collocations and lexical functions." In *Phraseology: Theory, Analysis and Applications*, ed. by A. Cowie , 23–53. Oxford: Clarendon Press.

Newman, J. and S. Rice

2006 "Transitivity schemas of English EAT and DRINK in the BNC." In *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, ed. by S.T. Gries and A. Stefanowitsch , 225–260. Berlin/New York: Mouton de Gruyter.

Östman, J.

2005 "Persuasion as implicit anchoring: The case of collocations." In *Persuasion across Genres*, ed. by H. Halmari and T. Virtanen , 183–212. Amsterdam: John Benjamins. doi: 10.1075/pbns.130.12ost

Östman, J. and Simon-Vandenbergen, A.

2005 "Firthian linguistics." In *Handbook of Pragmatics Online*, ed. by J. Östman and J. Verschueren . Amsterdam: John Benjamins. DOI:10.1075/hop.m.fir1 [10.10.2014] BoP

Palmer, F.R.

1976 *Semantics: A New Outline*. Cambridge: Cambridge University Press. BoP

Partington, A.

1998 *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins. doi: 10.1075/scl.2 MetBib

Renouf, A. and J. Sinclair

1991 "Collocational frameworks in English." In *English Corpus Linguistics*, ed. by K. Aijmer and B. Altenberg , 128–143. New York: English corpus linguistics.

Ruppenhofer, J. , C.J. Fillmore and C.F. Baker

2002 "Collocational information in the FrameNet database." In *Proceedings of the Tenth Euralex International Congress, Vol. 1*, ed. by A. Braasch and C. Povlsen , 359–369. Copenhagen.

Sag, I.A. , T. Baldwin , F. Bond , A. Copestake and D. Flickinger

2002 "Multi-word expressions: A pain in the neck for NLP." In *Proceedings of the 3th International Conference on Intelligent Text Processing and Computational Linguistics* , 1–15. Stanford, CA: Stanford University.

Shore, S.

2010 "J. R. Firth." In *Handbook of Pragmatics Online*, ed. by J. Östman and J. Verschueren . Amsterdam: John Benjamins. doi: 10.1075/hop.14.fir2 [10.10.2014]

Siepmann, D.

2005 "Collocation, colligation and encoding dictionaries. Part I: Lexicological aspects." *International Journal of Lexicography* 18(4): 409–443. doi: 10.1093/ijl/eci042

Sinclair, J.

2004 *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

1998 "The lexical item." In *Contrastive Lexical Semantics*, ed. by E. Weigand , 1–24. Amsterdam: John Benjamins. doi: 10.1075/cilt.171.02sin

1996 "The Search for Units of Meaning." *Textus* IX(1): 75–106.

1991 *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

1987 "Collocation: A progress report." In *Language Topics: Essays in Honour of Michael Halliday*, ed. by R. Steele and T. Threadgold , 319–331. Amsterdam: John Benjamins. doi: 10.1075/z.lt1.66sin

1966 "Beginning the study of lexis." In *In Memory of J.R. Firth*, ed. by C.E. Bazell , J.C. Catford , M.A.K. Halliday and R.H. Robins , 410–430. London: Longman.

(ed.) **1995** *Collins COBUILD English Dictionary*. London: Harper Collins.

Smadja, F.

1993 "Retrieving collocations from text: Xtract." *Computational Linguistics* 19(1): 143–177.

Stefanowitsch, A. and S.T. Gries

2009 "Corpora and grammar." In *Corpus Linguistics: An International Handbook, Vol. 2*. ed. by A. Lüdeling and M. Kytö , 933–951. Berlin/New York: Mouton de Gruyter. doi: 10.1515/9783110213881.2.933

2003 "Collostructions: Investigating the interaction of words and constructions." *International Journal of Corpus Linguistics* 8(2): 209–243. doi: 10.1075/ijcl.8.1.02gri

Stubbs, M.

2001a "On inference theories and code theories: Corpus evidence for semantic schemas." *Text* 21(3): 437–465.

BoP

2001b *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell. BoP

1996 *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Blackwell. BoP

1995a "Collocations and semantic profiles: On the cause of the trouble with quantitative studies." *Functions of Language* 2(1): 23–55. doi: 10.1075/fol.2.1.03stu BoP

1995b "Corpus evidence for norms of lexical collocation." In *Principle and Practice in Applied Linguistics*, ed. by G. Cook and B. Seidlhoper , 245–256. Oxford: Oxford University Press.

Taylor, J.R.

2003 "Near synonyms as co-extensive categories: 'high' and 'tall' revisited." *Language Sciences* 25(3): 263–284. doi: 10.1016/S0388-0001(02)00018-9

Tognini-Bonelli, E.

2001 *Corpus Linguistics at Work*. Amsterdam: John Benjamins. doi: 10.1075/scl.6

Wiechmann, D.

2008 "On the computation of collostruction strength: Testing measures of association as expressions of lexical bias." *Corpus Linguistics and Linguistic Theory* 4(2): 253–290. doi: 10.1515/CLLT.2008.011

Zipf, G.K.

1935 *The Psycho-biology of Language*. Boston: Houghton Mifflin.