Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 231 – 240

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Collocation, Co-collocation, Constellation… Any Advances in Distributional Semantics?

Moisés Almela*, Pascual Cantos, Aquilino Sánchez

University of Murcia, English Department, Murcia 30071, Spain

Abstract

This paper takes issue with the traditional notion that collocation is a binary combination. Theoretical and methodologically, the study is informed by the Lexical Constellation Model (LCM). This model has shown that the strength of syntagmatic attraction between two words in a collocation can be determined by associative relations involving more than two words. Therefore, the LCM proposes to complement the node-collocate distinction by incorporating the notion of co-collocate. In this study we examine evidence suggesting that the distinction between collocates and co-collocates can contribute to gaining a better understanding of the relations between meaning and context. The data analyzed consist of co-occurrences of *face* and *decision/choice* in the *enTenTen12* corpus.

Keywords: corpus linguistics, collocation, co-collocation, lexical constellations, semantics

1. Introduction

A widespread assumption in collocational studies is that collocation is a bipartite structure. This is in fact one of the few characteristics that allow us to establish a common ground among conflicting conceptions of collocation in the literature. As several authors have observed (Almela, 2007, 2006; Alonso Ramos, 1994-1995; Castillo Carballo, 1998; Corpas Pastor, 2001), there is little agreement among experts as to how the term *collocation* should be defined. Substantial discrepancies have arisen between those who regard collocation as primarily a statistical pattern, those who regard it as a special type of phraseological expression, and those who regard it as a selectional restriction.

* Corresponding author. Tel.: +34-868-88-3212
E-mail address: moisesal@um.es

doi:10.1016/j.sbspro.2013.10.643

However, despite these controversies, most models of collocation converge in considering that collocation consists of a relation between two lexical expressions.

The adequacy of the standard conception of collocation as a bipartite structure has been called into question by research into *lexical constellations* (Cantos & Sánchez, 2001; Almela, 2011a; Almela, Cantos & Sánchez, 2011a; Almela, Cantos & Sánchez, 2011b). The original contribution of the Lexical Constellation Model (henceforth: LCM) lies in emphasizing that different collocates of a node are not combined freely, a fact that had been neglected by the mainstream models. This suggests that the collocational pair is not always the appropriate framework for capturing the intricacies of the phenomenon of syntagmatic attraction, because it is often the case that the strength of syntagmatic attraction between two words *a* and *b* (e.g. *face* and *decision*) is sensitive to the presence or absence of another word (e.g. *difficult*). To deal with these cases, the LCM proposes an enhanced descriptive model of collocation which incorporates the notion of *co-collocate*, i.e. a collocate whose selection favors the co-selection of other collocates of the same node.

However, the LCM is still in its infancy. It has made interesting and innovative proposals, as well as provided empirical descriptions of several lexical items, but it still has a long way to go before constituting a fully-fledged model of collocational description. In sum, the LCM is a promising but still developing model. The goal of this paper is to contribute to this process of development. In particular, the focus of this study will be laid on discussing advantages of the LCM as a framework for describing the relationship between contextual and distributional similarity. In sections 2 to 4 we will discuss shortcomings of the standard approaches to collocation, compared to the LCM. Then, section 5 will show the LCM at work. The object of analysis will consists of complex co-occurrence patterns formed with different categories of collocates (verbs and modifiers) of the nouns *decision* and *choice*.

2. Binary approaches to collocation

As anticipated in the Introduction, the vast majority of approaches to collocation consider it as a binary word combination, and this is reflected in a terminology that analyses the internal structure of a collocation into two elements. In the literature produced in the Continent, the terms most usually employed to denote the two parts of a collocation are *base* and *collocator/collocate* (Spanish *base-colocativo/colocado*; German *Basis-Kollokator*; French *base-collocatif*). This distinction was introduced by Hausmann (1979) and has been later popularized in the lexicographical and phraseological literature on collocation (Castillo Carballo, 2001; Concepción García, 2001; Corpas Pastor, 1998, 2001; Hausmann, 1991, 1998; Írsula Peña, 1992, 1994; Liang, 1991; Martin, 2008; Wotjak, 1994; among others). In this perspective, each member of a collocation necessarily falls within one of two categories: it must be either the dominant element in the relation (the base) or the lexically dependent item (the collocator). Crucially, there is no intermediate or borderline category between base and collocator; there is no third type of collocational constituent within which one part of a collocation could be classified. In this sense the model can be described as dualistic: it is rooted in a binary (either-or) typology of collocational constituents.

Similarly, in the Meaning-Text Theory (MTT) framework, a word that forms part of a collocation is categorized as either the *keyword* or the *value* of the lexical-syntagmatic operation (the *lexical function*) that produces the collocation (Mel'čuk, 1998). More recently, MTT scholars like Alonso Ramos (2001, 2007) and Barrios Rodríguez (2010) have also adopted the terminology proposed by Hausmann, but this leaves the dualistic nature of the model unaltered—in fact, as Mel'čuk, Clas & Polguère (1995) conceded, the distinction between keyword and value is analogous to that between base and collocator.

A third approach to collocation that has been highly prolific and influential, especially in the field of English linguistics, is the so-called Sinclairian approach, which is equivalent to what Tognini-Bonelli (2001) labeled *corpus-driven* linguistics. In this framework, a word that forms part of a collocational pattern is classified as either a *node* or a *collocate* (Jones & Sinclair, 1974; Mason, 2000; Sinclair, 1991; among others). These two concepts are not tantamount to the foregoing terminological pairs (*base-collocator, keyword-value*) because the distinction of node and collocate is purely methodological: it does not distinguish between lexical items with different properties; it simply refers to different steps in the method of collocation extraction (input and output). Thus, each member of a collocation can be treated successively both as node and as collocate in different empirical studies. The node is simply the search term. However, despite this difference, there is something in the node-collocate distinction that is

shared by the aforementioned approaches, and again, this area of common ground is the adoption of a dualistic typology of collocational elements.

At this point, it is essential to underline that the binarity of the mainstream models does not refer to the number of words but to the number of constituents of the relation. Thus, the collocation *violation of the law* consists of four words and two collocational constituents. This is because the collocational link connects *violation* with *of the law* taken as a whole. It is not the case that *violation* relates separately to *of*, *the*, and *law*. Obviously, there is not a separate collocational link for each pair of words found in this sequence. Recognizing the internal complexity of one of the constituents does not imply per se a departure from the binarity of the traditional model.

For the same reason, the concept of *complex collocation* (García-Page, 2005; Koike, 2005) can be seen as an extension rather than a revision of the standard binary model of collocation. A complex collocation is defined as a collocation consisting of a simple lexeme and an idiom. Therefore, the difference between a complex collocation and a simple collocation lies not in the number of constituents of the relation (i.e. the number of elements connected by a collocational link) but rather in the internal complexity of one of the constituents.

Finally, a second type of extension of the binary model of collocation is based on *iteration*, i.e. successive applications of binary combinatorial operations. To our knowledge, in the literature there are two theoretical constructs that describe a process of iterative binary combination. The first one is the notion of *collocational network* (Williams, 1998; Alonso, Millon & Williams, 2011). A collocational network is formed through the successive application of the corpus-driven methodology for extracting node-collocate pairs. The network grows as each of the collocates obtained in a prior stage are subsequently entered as nodes in the next stage. For instance, if we input the verb *treat* as a node, we obtain collocates such as *patient* and *control*, among others, and when, in turn, these words are entered as nodes, we obtain further collocates, such as *undergo*, *admit*, or *ventilate*, for *patient*, and *use*, *serve*, or *match*, for *control*. The second theoretical construct devised to account for iterative binary combinations is the notion of *collocation chains*, put forward by Alonso Ramos & Wanner (2007). Collocation chains are defined as pairs of collocations in which one element is common to both collocations.

The main difference between collocational networks and collocation chains lies in the broader theoretical-methodological framewok. While collocational networks have been developed in the research framework of corpus-driven terminography, the description of collocation chains draws on the theoretical apparatus of Meaning-Text Theory. Thus, the expression *to commit a violation of the law* is analyzed as an overlap of two collocations (*commit a violation* + *violation of the law*) in which one and the same element functions as the base in collocation with *commit*, and as the collocator in collocation with *law*. The formal notation is given as follows: *to commit*$_{C}$ *a violation*$_{[B/C]}$ *of the law*$_{B}$. Despite these differences, the two constructs share a fundamental aspect, namely, the iterative application of binary structures. The step from the individual collocation to the network or chain is taken by combining two or more collocational pairs, and each of these pairs preserves the dualistic typology of node and collocate, or base and collocator. To put it another way, the categories used to describe each collocational pair (*node* and *collocate*, *base* and *collocator*) remain intact in the passage from the individual collocation to the network or chain.

3. Beyond the binary

Research into lexical constellations has suggested that the mere identification of overlaps among binary combinations is not sufficient to account for the relations among collocates in complex co-occurrence patterns. The main reason for this is that the node does not exert an unlimited influence on its environment (Cantos & Sánchez, 2001). The likelihood that two words, say *a* and *b*, occur in the same context is often determined not only by the influence exerted by *a* on *b*, or by *b* on *a*, but also by the influence from other neighboring words. That is, the probability for *a* and *b* to co-occur can be increased or decreased by the co-presence of *c*. When this obtains, *c* should be considered as integral to the collocation of *a* and *b*, but in the standard models of collocation there is no category that defines this role. In this context, *c* is neither a node nor a collocate, let alone a base or collocator.

In the LCM, the term used to denote the function of *c* in this context is *co-collocate*. A co-collocate is a collocate that exerts an influence on the likelihood of another collocation with the same node. This influence can be *positive* or *negative*. A positive co-collocate is one that makes a contribution to the activation of another collocation of the same

node. For example, the probability that *reject goods* converges with *faulty goods* is higher than the probability of *reject* co-occurring with *goods*. This can be interpreted as an indication that the selection of one of these collocations favours the selection of the other. By contrast, the relation of negative co-collocation obtains when the collocabiliy of a node and a collocate is weakened by the presence of other collocates of the same node. For example, the probability of *ship goods* converging with *faulty goods* is considerably lower than the probability of *ship* co-occurring with *goods* (Almela, Cantos & Sánchez, 2011b). This is an indication that the selection of one of these collocations repels the selection of the other.

The technique for detecting cases of positive and negative co-collocation is based on comparisons of conditional probabilities (Almela, Cantos & Sánchez, 2011b). The values compared are two: (1) conditional probability at the inter-collocational level and (2) conditional probability at the intra-collocational level (see formulae below). (1) is the probability that a collocate of the node is selected given as a fact the co-occurrence of the node and another collocate. In the formal notation, *n* stands for the node, and *c1* and *c2* represent two different collocates. (2) is the probability that the same collocate (c1) is selected given as a fact the occurrence of the node.

$$P(c1 \mid c2, n) \tag{1}$$

$$P(c1 \mid n) \tag{2}$$

If the value of (1) is higher than the value of (2), we say that *c2* is a positive co-collocate of *c1* relative to *n*. By contrast, negative co-collocation obtains when (1) is lower than (2). In this case we can say that *c2* is a negative co-collocate of *c1* relative to *n*. This formulation expresses the fact that the collocation of the node with *c2* diminishes the probability of finding the collocation (*n,c1*).

In the present paper our attention will remain focused on positive rather than negative co-collocation. The reason for this is that positive co-collocates are more directly relevant for semantic description. Although the motivation for negative co-collocation can be semantic, in the sense that it is often indicative of a lack of affinity between the respective word meanings, the negative co-collocates per se are not representative of meaning components of the node. Positive co-collocates, in contrast, are more likely to represent semantic features of the node. This question is addressed in the next section.

4. Constellations and meaning

The existence of a close correlation of meaning and distribution has been widely attested in the corpus linguistics literature. Many studies have shown that distributional classes tend to be aligned with semantic classes, i.e. that words which tend to occur in similar contexts will also tend to share aspects of meaning (Church, Gale, Hanks, Hindle & Moon, 1994; Renouf, 1996), and conversely, that differences in meaning will tend to be correlated with differences in syntagmatic behaviour (Partington, 1998). Today, corpus research into meaning-context correlations is advancing in several directions. The Behavioral Profile (BP) approach relies on fine-grained semantic and grammatical annotation for detecting subtle differences in the distribution of quasi-synonyms (Gries, 2010; Liu, 2010). The LCM explores the principle of meaning-context correlation in a different way. It provides a methodological and anaytical framework for extending this principle to apply not only to the relationship of nodes and collocates, but also to the effects of co-collocates. The underlying assumptions are two:

- The correlation of meaning and context applies in patterns with varying levels of complexity, including not only collocational pairs but also more complex collocational patterns, such as lexical constellations.
- Contextual specificity is proportional to semantic specificity. The more specific the context that defines a distributional class, the more restricted will be the meaning potential of the corresponding semantic class.
- Logically, one of the factors that determine the degree of contextual specificity of a distributional class is the complexity of the pattern in which it occurs. A distributional class defined by its co-occurrence with a word (e.g. with the node) will be less specific, and consequently less homogeneous, than a distributional class defined by its co-occurrence with another collocation (for instance, with the node and another collocate).

It follows as a corollary from this that a distributional class which occupies a slot in a lexical constellation will be less heterogeneous than a distributional class which occupies a slot in a collocational pair. A set of collocates defined by their occurrence in the same constellation (i.e. the same combination of node and co-collocate) will form a coherent subset –or a limited number of subsets– of the collocates occurring in a simple collocational pattern (i.e. collocates co-occurring with the same node). This prediction will be put to the test in the next section.

5. Lexical and conceptual constellations of *decision* and *choice*

5.1. Method

This section undertakes the construction and analysis of some constellations formed around the node *decision* and its verbal collocate *face*. The corpus used is *enTenTen12*, a large web corpus of English available at Sketch Engine. This corpus contains more than 12 billion tokens. The quantitative filters included a frequency threshold of 3 and the use of *logDice* as a measure of statistical significance—see Rychlý (2008) for advantages of logDice over other measures of collocational association. To this we added a qualitative filter: all collocates and potential co-collocates were extracted from a specific syntactic slot, namely, pre-modifiers of *decision*, both nominal and adjectival. All the collocational patterns occurred in the same type of construction, with *decision* functioning as a direct object of *face* in an active construction or as the subject in a passive construction. In short, we only considered constructions in which *decision* fulfils the semantic role of THEME (following the standard notation system employed in many linguistic schools, capital letters are used here to indicate semantic content). The reason for selecting this specific syntactic context is that in previous LCM research it has proven to be a highly productive source of positive co-collocation patterns. The data were extracted by using the Word Sketch application followed by manual supervision.

The first step was to extract the top logDice collocates of *decision* in the modifier slot (Table 1). The number of collocates in the list was limited to 50 for practical reasons (manageability, limitations of time and space, etc.). This was followed by the identification of modifiers standing in a relation of positive co-collocation with *decision*. The list was obtained by applying the formula described above, that is, the comparison of inter-collocational conditional probabilities and intra-collocational conditional probabilities (see section 3). The link of co-collocation can be sensitive to directionality (Almela, Cantos & Sánchez, 2011b), and in this case we decided to analyze the dependency of the modifier+noun collocation on the verb+noun collocation. Therefore, the variable *c2* in formula (1) was occupied by the verb *face*, and the variable *c1* in (1) and (2) represented the modifier. The value of this formula represents the proportion of verb+*decision* collocations that are accompanied by each modifier. Part of the data necessary for obtaining these values is shown in the Appendix. The rest of the data used in these calculations are the frequency of the node (*decision*) in the whole corpus, which is 2256081, and the frequency of the collocation *face* + *decision*, which is 3215. These data are not mentioned in the Appendix because they are constant (the same value would be repeated in all the rows).

The next step consists in abstracting from the surface lexical (formal) aspects of co-collocation and arriving at the underlying conceptual structures. The expected output is a semantic set or limited group of semantic sets of collocates. As with every study dealing with semantic description, recourse to intuition was inevitable. Here, intuition was used to detect shared meaning components among collocates.

Finally, the last step involves a comparison with constellations formed around other nodes. The principle of meanig-context correlation, well attested in the corpus linguistics literature, establishes that words which share semantic features will also tend to share contexts of use. Our contention is that this principle can be extended to include patterns of lexical constellation. This means that words that are lexically related in the vocabulary will tend not only to have overlapping collocational profiles but also to activate similar combinations of collocates and co-collocates. This can be illustrated with reference to constellations formed around *choice* and *face*. Since *choice* is a quasi-synonym of *decision*, our expectation is that, in principle, the positive co-collocations formed with *face* will be similar to those found in the context of *decision*. Following exactly the same methodology that we applied in the description of *face* + *decision*, we extracted the top 50 modifiers of *choice* and identified the relations of positive co-collocation with *face*. The data used in these calculations are displayed in the Appendix—the frequency of the noun

choice (2494194) and the frequency of the collocation *face+choice* (3817) have not been included in the Appendix because these data are constant in all the rows.

5.2. Results

The output of the first step –collocate extraction– consists of all the words listed in the left-most column in the Appendix. The list has been arranged in order of decreasing logDice score. Of all those 50 modifiers, the following ones qualified as entering a relation of positive co-collocation with *face*: *difficult, tough, important, hard, big, major, critical, financial, crucial, strategic, ethical, key, ultimate*. The list is arranged in order of decreasing difference between the two values of conditional probability (Table 1).

Table 1. Positive co-collocation between *face* and modifiers of *decision*

Modifiers	P(c1|n)	P(c1|c2,n)
difficult	0.60%	17.93%
tough	0.54%	13.93%
important	0.97%	3.73%
hard	0.29%	2.76%
big	0.40%	2.55%
major	0.39%	2.18%
critical	0.20%	2.15%
financial	0.47%	1.49%
crucial	0.12%	1.18%
strategic	0.31%	0.75%
ethical	0.08%	0.56%
key	0.21%	0.53%
ultimate	0.14%	0.22%

The data in the table reflects the capacity of *face+decision* for predicting co-collocations with specific modifiers of *decision*. In all these cases, the potential of the verb+noun collocation for triggering co-selection of a specific modifier is higher than the potential of the noun alone for predicting the selection of the same modifier. That is, the value of the inter-collocational conditional probability is higher than the value of the intra-collocational conditional probability, as explained in section 3. The highest increase is observed for *difficulty*, whose probability of being selected in the context of *decision* is increased by 30 times if we add the presence of the verb *face*. The second highest increase is with *tough*. In this case, the probability is increased by 26 times.

In the next stage, which leads to the semantic classification of co-collocational patterns, it was observed that nine of the thirteen modifiers in the list can be organized coherently into two semantic sets. First, there is a group of three adjectives, all quasi-synonyms, describing DIFFICULTY (*difficult, tough, hard*), and second, there is a group of six adjectives describing IMPORTANCE/ MAGNITUDE (*important, big, critical, major, crucial, key*). The presence of several near-synonyms in the list reinforces the initial intuition that co-collocation is rooted in semantic patterning. This issue is addressed in the Discussion section.

Finally, the same sequence of steps was applied in the description of constellations with *face+choice*. The list of top 50 logDice collocates, arranged in order of decreasing score, includes the items in the second column (from left to right) in the Appendix. The list of modifiers entering a relation of positive co-collocation with *face* is shown in Table 2.

Table 2. Positive co-collocations between *face* and modifiers of *choice*.

Modifiers	P(c1|n)	P(c1|c2,n)
difficult	0.21%	14.40%

tough	0.22%	9.92%
moral	0.08%	0.78%
ultimate	0.08%	0.22%
numerous	0.16%	0.22%
career	0.47%	0.53%

There are several similarities between Table 1 and Table 2. Firstly, the adjectives *difficult* and *tough* rank as the top two elements in both tables, and both belong to the same semantic set. Secondly, *ultimate* is shared by the two lists. Thirdly, one of the words in Table 1, *ethical*, is a near-synonym of one of the words in Table 2, namely, *moral*. To this we must add the fact that many other positive co-collocations of *face* and *choice/decision* have not been included in the table, because here we have take into account only the top 50 modifiers. For instance, *face+hard* is a co-collocation shared by *decision* and *choice*, but it has not been included in Table 2 because *hard* is not contained within the top 50 modifiers of *choice*. Thus, the list of shared co-collocations that we offer here is not exhaustive. It is reasonable to predict that a more comprehensive study describing a larger number of collocates and co-collocates would yield further commonalities between constellations of these two nouns.

5.3. Discussion

The results obtained above can be interpreted as an indication that the co-collocate (*face*) imposes a subcategorization on the modifiers that collocate with the nouns *decision* and *choice*. The main argument for this is that the list of words which occur in the modifier slot of the constellation is semantically more specific and less variable than the list of words which occur in the modifier slot of the simple collocational pair.

The list of modifiers collocating with *decision* (see section 4.1) is highly heterogeneous. It encompasses words as disparate as *informed, final, tough, purchasing, court, strategic, unanimous, recent, critical, policy,* or *sensible,* among others. It is difficult to subsume the different concepts expressed by these collocations under a limited set of headings. Far from representing an exception, this reflects a general fact. As a rule, the alignment of distributional classes and semantic classes is less systematic in the predicate slot than in the argument slot. In the literature there is ample evidence that the range of predicative lexemes (for instance, pre-modifiers) that collocate with an argument (for instance, a common noun) is less homogeneous than the range of arguments that collocate with a predicate (Apresjan, 2009; Bosque, 2001, 2004; Cruse, 2004).

Interestingly, the behaviour of distributional classes in lexical constellations is different from their behaviour in simple collocational environments. As stated, a distributional class of predicates exhibits little homogeneity when it is defined in relation to an argument, but if it is defined in relation to a constellation, i.e. to the combination of a node and a co-collocate, the resulting distributional class will gain consistency.

This remark does not contradict the traditional notion that semantic selection is oriented from predicates to arguments, because the co-collocate itself (the verb *face*) is a predicate, but it suggests that selection restrictions can extend over more than one layer of phrase structure. At one level, the valency carrier selects the semantic type of the head of the valency filler (*face* also collocates with nouns that are lexically related to *decision*, such as *choice* and *dilemma*). At a second level, the valency carrier can also impose restrictions on the semantic type of other elements occurring in the argument phrase, for instance, the modifier (Almela, 2011b). This phenomenon was termed *valency stratification* in Almela, Cantos & Sánchez (2011b). The concept refers to complex co-occurrence patterns in which a predicate restricts the semantic type of more than one slot-filler in the argument phrase. The standard corpus methodology of collocational description is not suited to capture the phenomenon of stratification in valency patterning, because it does not take into account the patterns of mutual preferences among different predicates/collocators of the same argument/base.

The nominal collocates of *face* provide further evidence for treating the pattern *face+difficult/hard/tough+ decision* as an example of valency stratification, i.e. as a pattern in which the verb imposes different selectional restrictions on different slots of the argument phrase. *Face* collocates prominently with object nouns denoting DIFFICULTY/OBSTACLE. Among top logDice collocates of *face* in the 'object-of' grammatical relation we find nouns such as *challenge, difficulty, problem, crisis, threat, obstacle, hardship, opposition, hurdle, danger*, etc., all of

which can be clearly subsumed under a common heading. Since many patterns of semantic valency have alternate realizations in the form of verb+noun and verb+adjective collocations (Almela, 2011), there are reasons to suggest that the semantic set formed by the modifiers *difficult, hard*, and *tough* is subjected to the influence exerted by the valency of the verb (the positive co-collocate *face*). In this perspective, expressions such as *face difficulties, face hardships, face a hard* + NN, *face difficult* + NN, etc., can be regarded as variable grammatical realizations of the same underlying semantic valency pattern. The potential of *face* for binding the concept DIFFICULTY is not circumscribed to a single syntactic slot.

6. Conclusion

The findings obtained from this study suggest that the methodology of lexical-constellation analysis, and the concept of co-collocation associated with it, can shed new light into the connection of syntagmatic behaviour and semantic properties of words. In this sense, the LCM can contribute to rendering corpus lexical studies more sophisticated and accurate. The case study analyzed here points to an interesting conclusion: the alignment of distributional classes and semantic classes is stronger in lexical constellations than in simple collocational pairs. The combinatorial preferences of the co-collocate (in this case, the verb *face*) are not unsystematic. They impose certain features of semantic subcategorization on the collocates of the node (i.e. on modifiers of *decision* and *choice*).

Obviously, more case studies are needed before this conclusion can receive substantial empirical support and be extended to apply to various areas of the lexicon. However, on the face of the evidence available, it is difficult to deny that there are some aspects of the relationship between meaning and context that cannot be adequately captured if the analysis of meaning by collocation remains focused on binary combinations. There is evidence that the selection of modifiers of a noun is not made independently of the selection of the verb. This indicates that the strength of syntagmatic attraction between two words can be determined by interactions and associative relations among more than two words, in which case the categories of node and collocate, base and collocator, or keyword and value are insufficient to account for the structure of a collocational pattern.

Acknowledgements

The research presented in this paper is generously funded by two grants: one from the Seneca Foundation, Region of Murcia, Spain (*Fundación Séneca, Agencia de Ciencia y Tecnología de la Región de Murcia*, Ref. 08594/ PHCS/08), and another one from the Spanish Ministry for Science and Research (*Ministerio de Ciencia e Investigación*, FF2012-38724). We are most grateful to these institutions for their financial support.

Appendix A. Data used for obtaining conditional probabilities

Modifiers of *decision*	Modifiers of *choice*	F(modifier,*decision*)	F(*face*,modifier,*decision*)	F(modifier,*choice*)	F(*face*,modifier,*choice*)
informed	*right*	35755	0	3494	0
final	*informed*	42055	2	10217	0
wise	*excellent*	15669	0	21118	0
right	*popular*	27112	0	24135	2
tough	*good*	12173	447	156074	10
purchasing	*ideal*	6799	0	14228	0
conscious	*healthy*	7191	0	17839	0
buying	*wise*	6884	1	10119	0
investment	*perfect*	17646	9	18779	0
court	*smart*	15061	5	9797	0
strategic	*obvious*	7022	24	9964	2
unanimous	*lifestyle*	4888	0	10482	0
smart	*preferred*	6900	0	6643	0
difficult	*poor*	13489	576	9123	3
sound	*personal*	5307	0	15456	5
recent	*great*	8668	0	38625	6
poor	*sensible*	6856	1	4593	0

bad	*career*	14284	1	11838	17
important	*top*	21986	120	9364	0
financial	*first*	10710	48	28468	1
judicial	*conscious*	3002	0	4159	1
major	*tough*	8775	70	5553	319
wrong	*final*	6062	0	6827	2
purchase	*logical*	5748	1	3621	0
rational	*food*	2669	0	18114	3
ultimate	*alternative*	3060	7	4143	0
intelligent	*correct*	2802	0	4555	0
correct	*suitable*	3915	0	4689	2
critical	*wrong*	4621	69	6115	1
good	*bad*	46265	1	10207	11
policy	*color*	11367	7	8341	1
quick	*multiple*	4355	2	4142	5
political	*fantastic*	5451	3	3582	0
key	*appropriate*	4816	17	4506	1
split	*second*	2037	2	6142	2
admission	*numerous*	2245	0	4075	7
controversial	*wonderful*	1949	2	3689	0
hiring	*possible*	1719	0	7466	2
crucial	*difficult*	2664	38	5217	463
hasty	*draft*	1560	0	2293	0
big	*natural*	8983	82	4993	0
personal	*moral*	6335	7	2110	25
hard	*viable*	6483	89	1846	0
ethical	*superb*	1777	18	1892	0
sensible	*different*	1687	0	10486	10
initial	*ultimate*	2514	1	1876	7
instant	*safe*	1789	0	3438	0
administrative	*free*	1722	0	6957	0
business	*limited*	18050	16	2043	20
management	*odd*	5994	3	1644	0

References

Almela, M. (2007). Concepto de 'colocación' en los estructuralismos inglés, alemán y español. *Revista Portuguesa de Humanidades, 11/1*, 229–249.

Almela, M. (2011a). Improving corpus-driven methods of semantic analysis: a case study of the collocational profile of incidence. *English Studies, 92/1*, 84–99.

Almela, M. (2011b). The case for verb-adjective collocations: corpus-based analysis and lexicographical treatment. *Revista de Lingüística y Lenguas Aplicadas, 6*, 39–51.

Almela, M., Cantos, P. & Sánchez, A. (2011a). From collocation to meaning: revising corpus-based techniques of lexical semantic analysis. In I. Balteiro (ed.), *New Approaches to Specialized English Lexicology and Lexicography* (pp. 57–72). Newcastle upon Tyne: Cambridge Scholars Press.

Almela, M., Cantos, P. & Sánchez, A. (2011b). Towards a dynamic combinatorial dictionary: a proposal for introducing interactions between collocations in an electronic dictionary of English word combinations. In I. Kosem & K. Kosem (Eds.), *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011, Bled, 10-12 November 2011* (pp. 1–11). Ljubljana: Trojina.

Alonso Ramos, M. (1994-1995). Hacia una definición del concepto de colocación: de J. R. Firth a I. A. Mel'čuk. *Revista de Lexicografía, 1*, 9–28.

Alonso Ramos, M. (2001). Construction d'une base de données des collocations bilingüe français-espagnol. *Langages, 143*, 5–27.

Alonso Ramos, M. (2007). Actantes y colocaciones. *Nueva Revista de Filología Hispánica, 2,* 435–458.

Alonso Ramos, M. & Wanner, L. (2007). Collocation chains: how to deal with them? In K. Gerdes, T. Reuther, L. Wanner (Eds.), *Proceedings of the Third International Conference on Meaning-Text Theory* (pp. 11–20). Wiener Slawistischer Almanach, Sonderband 69. Munich.

Alonso A., Millon A., Williams G. (2011). Collocational networks and their application to an E-Advanced Learner's Dictionary of Verbs in Science (DicSci). In I. Kosem & K. Kosem (Eds.), *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011, Bled, 10-12 November 2011* (pp. 12–22). Ljubljana: Trojina.

Apresjan, J. (2009). The theory of Lexical Functions: an update. In D. Beck, K. Gerdes, J. Milićević & A. Polguère (Eds.), *Proceedings of the Fourth International Conference on Meaning-Text Theory* (pp. 1–13). Montréal: OLST.

Barrios, M. A. (2010). El dominio de las funciones léxicas en el marco de la Teoría Sentido-Texto. *ELiEs, 30* (in http://www.elies.rediris.es/elies30/) [Date of access: 30.11.2010]

Bosque, I. (2001). Sobre el concepto de 'colocación' y sus límites. *Lingüística Española Actual, 23/1*, 9–40.

Bosque, I. (2004). La direccionalidad en los diccionarios combinatorios y el problema de la selección léxica. In T. Cabré (Ed.), *Lingüística teórica: anàlisi i perspectives* (pp. 13–58). Bellaterra: Universidad Autónoma de Barcelona.

Cantos, P. & Sánchez, A. (2001). Lexical constellations: what collocates fail to tell. *International Journal of Corpus Linguistics, 6/2*, 199–228.

Castillo Carballo, M. A. (1998). El término 'colocación' en la lingüística actual. *Lingüística Española Actual, 20/1*, 41–54.

Castillo Carballo, M. A. (2001). Colocaciones léxicas y variación lingüística: implicaciones didácticas. *Lingüística Española Actual, 23/1*, 133–143.

Church, K. W., Gale, W., Hanks, P., Hindle, D., & Moon, R. (1994). Lexical substitutability. In B. T. S. Atkins & A. Zampolli (Eds.), *Computational Approaches to the Lexicon* (pp. 153-177). Oxford: Oxford University Press.

Concepción García, A. (2001). Consideraciones para el estudio de la colocación sustantivo-verbo. In A. Veiga, M. González Pereira & M. Santo Gómez (eds.), *De lenguas y lenguajes* (pp. 95–103). Noia, A Coruña: Toxosoutos.

Corpas Pastor, G. (1998). Expresións fraseolóxicas e colocacións: clasificación. In X. Ferro Ruibal (Ed.), *Actas do I Coloquio Galego de Fraseoloxía* (pp. 31–61). Santiago de Compostela: Centro Ramón Piñeiro, Xunta de Galicia.

Corpas Pastor, G. (2001). Apuntes para el estudio de la colocación. *Lingüística Española Actual 23/1*, 41–56.

Cruse, D. A. (2004). *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

García-Page, M. (2005). Colocaciones simples y complejas: diferencias estructurales. In R. Almela Pérez (ed.), *Fraseología contrastiva con ejemplos tomados del alemán, español, francés e italiano* (pp. 145–167). Universidad de Murcia / Universität Leipzig.

Gries, S. (2010). Behavioral profiles. A fine-grained quantative approach in corpus-based lexical semantics. *The Mental Lexicon*, *5/3*, 323–346.

Hausmann, F. (1979). Un dictionnaire des collocations est-il possible? *Travaux de linguistique et de litterature, 17/1*, 187–195.

Hausmann, F. (1991). Collocations in the bilingual dictionary. In F. Hausmann, O. Reichmann, H. E. Wiegand, & L. Zgusta (Eds.). *An International Encyclopedia of Lexicography, Vol. 2* (pp. 2775–2778). Berlin: Walter de Gruyter.

Hausmann, F. (1998). O diccionario de colocacións: criterios de organización. In X. Ferro Ruibal (ed.), *Actas do I Coloquio Galego de Fraseoloxía* (pp. 63–81). Santiago de Compostela: Centro Ramón Piñeiro, Xunta de Galicia.

Írsula Peña, J. (1992). Colocaciones sustantivo-verbo. In G. Wotjak (Ed.), *Estudios de lexicografía y metalexicografía del español actual* (pp. 159–167). Tübingen: Niemeyer.

Írsula Peña, J. (1994). *Substantiv-Verb-Kollokationen. Kontrastive Untersuchungen Deutsch-Spanisch*. Frankfurt a. M.: Peter Lang.

Koike, K. (2005). Colocaciones complejas en el español actual. In R. Almela Pérez (Ed.), *Fraseología contrastiva con ejemplos tomados del alemán, español, francés e italiano* (pp. 169–184). Universidad de Murcia / Universität Leipzig.

Jones, S., Sinclair, J. (1974). English lexical collocations. A study in computational linguistics. *Cahiers de Lexicologie, 24*, 15–61.

Liang, S. Q. (2001). À propos du dictionnaire français-chinois des collocations françaises. *Cahiers de Lexicologie, 59/2*, 151–167.

Liu, D. (2010). Is it a chief, main, major, primary, or principal concern? A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics, 15/1*, 56–87.

Martin, W. (2008). A unified approach to semantic frames and collocational patterns. In S. Granger & F. Meunier (Eds.), *Phraseology. An Interdisciplinary Perspective* (pp. 51–65). Amsterdam: John Benjamins.

Mason, O. (2000). Parameters of collocation: the word in the centre of gravity. In J. M. Kirk (ed.), *Corpora Galore. Analyses and Techniques in Describing English* (pp. 267–280). Amsterdam/Atlanta, GA: Rodopi.

Mel'čuk, I., Clas, A. & Polguère, A. (1995). *Introduction a la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

Mel'čuk, I. (1998). Collocations and lexical functions. In A. P. Cowie (Ed.), *Phraseology. Theory, Analysis and Applications* (pp. 23–53). Oxford: Clarendon Press.

Partington, A. (1998). *Patterns and Meanings. Using Corpora for English Language Research and Teaching*. Amsterdam/Philadelphia: John Benjamins.

Renouf, A. (1996). Les nyms: en quête du thésaurus des textes. *Lingvisticae Investigationes, 20/1*, 145–165.

Rychlý, P. (2008). A lexicographer-friendly association score. In P. Sojka & A. Horák (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008* (pp. 6–9) Brno: Masaryk University.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.

Williams, G. (1998). Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics, 3/1*, 151–171.

Wotjak, G. (1994). Nichtidiomatische Phraseologismen: Substantiv-Verb-Kollokationen – ein Fallbeispiel. In B. Sandig (ed.), *EUROPHRAS'92. Tendenzen der Phraseologieforschung* (pp. 651–677). Bochum: Universitätsverlag Dr. N. Brockmeyer.