Towards a Dynamic Combinatorial Dictionary: A Proposal for Introducing Interactions between Collocations in an Electronic Dictionary of English Word Combinations

Moisés Almela, Pascual Cantos, Aquilino Sánchez
Universidad de Murcia (Spain)
Depto. de Fil. Inglesa, Fac. de Letras, Campus de la Merced, 30071 Murcia (Spain)
E-mail: moisesal@um.es, pcantos@um.es, asanchez@um.es

Abstract

This paper presents an academic (non-commercial) lexicographic project called *Dynamic Combinatorial Dictionary*, which is currently being developed by members of the LACELL Research Group at the University of Murcia. The aim of this project is to bring e-Lexicography in closer alignment with lexical models that cannot be implemented in printed dictionaries. Theoretically, the project is informed by the Lexical Constellation Model. The main difference between this model and the mainstream approaches to collocation lies in its suitability for recognising more than one domain of lexical attraction within the same collocational window. We will distinguish two different manifestations of this multiplicity of domains. The first one is the phenomenon of *indirect collocation*, which has been investigated in previous Lexical Constellation research, and the second one is *inter-collocability*. This concept refers to positive or negative dependency relations between collocational pairs (not between words). It will be argued that incorporating inter-collocability features into lexical entries can lead to significant advances in the field of combinatorial lexicography.

Keywords: collocation; lexical constellations; corpus linguistics; e-Lexicography; combinatorial dictionaries.

1. Introduction

The potential of electronic formats for increasing the variety of contextual data offered to the user, as well as for facilitating an interactive management of the information contained in lexical entries, is underexploited in current combinatorial dictionaries. This is in part due to the fact that the design of electronic combinatorial dictionaries is to a large extent informed by the design of earlier printed dictionaries. At present, the difference between electronic and printed developments in combinatorial lexicography lies more in the material format (i.e. in the medium) than in the kind and amount of information provided.

In this study, we present a proposal for exploiting more effectively and thoroughly the opportunities created by the electronic format in combinatorial lexicography. Our research is motivated by the idea that in an electronic dictionary it is possible to incorporate collocational information of a qualitatively different kind from the one that is offered to the user of a conventional collocation dictionary. More specifically, we submit that collocational information in an electronic dictionary need not be restricted to dependencies between words, and that it can be extended to include dependencies between different collocations.

The paper is structured as follows. First, in the next section we shall explain the theoretical framework of the proposal, which is based on Cantos & Sánchez's (2001) Lexical Constellation Model. It will be argued that the analysis of collocation as a relationship between lexical items is incomplete and should be complemented with a description of interactions between different collocations of a lemma. In section 3 the workings of the model are illustrated with reference to the collocational profile of the noun *goods*. The lexicographic treatment of this information is illustrated in Section 4, where we present parts of a sample entry from the *Dynamic Combinatorial Dictionary* (DCD). The advantages of the DCD over conventional approaches to combinatorial lexicography are also explained in this section.

2. The Lexical Constellation Model

The Lexical Constellation Model (henceforth: LCM) originated from the observation that the node, i.e. the word under investigation in corpus collocational research, does not exert an unlimited influence on its environment (Cantos & Sánchez, 2001). This means that the node is not the only lexical item to restrict the range of lexical choices in its textual environment. In the syntagmatic context of the node there are other lexical items which can be endowed with a context-predictive potential. To express it in more formal terms, we can say that what differentiates the LCM approach from mainstream approaches to collocation is its determination to resolve difficulties caused by the phenomenon of *lexical gravity overlaps* (or *lexical gravity interference*).

The term *lexical gravity*, as is well known, was used by Mason (2000) to denote the context-predictive potential associated with the selection of a word in the discourse. To quote the author, lexical gravity can be defined as "the restriction a word imposes on the variability of its context" (Mason, 2000: 270). The lexical gravity of a word is the influence it exerts on restricting the choice of possible words in specific positions of its textual environment.

The problem brought to the fore by LCM research is that lexical gravity can be exerted by more than one item in the same textual window. The imposition of restrictions on lexical choices in the context of the node is not an exclusive function of the node. The lexical gravity exerted by collocates of the node can interfere with the gravity attributed to the node. This begs the need to distinguish which features of lexical gravity are a contribution of the node and which ones are contributions of other elements. In this respect, the LCM outperforms the conventional approaches to collocation.

The received models of collocation are not suitable for dealing with the problem of lexical gravity interference. The reason for this is that they are linear, in the sense that they fail to divide the collocational patterns of the node into different domains of lexical attraction. The LCM seeks to resolve this problem by comparing the influence of the node and the influence exerted by other items or structures that co-exist within the same textual window.

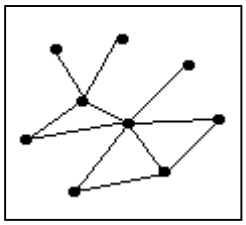


Figure 1: Structure of a plain collocational network

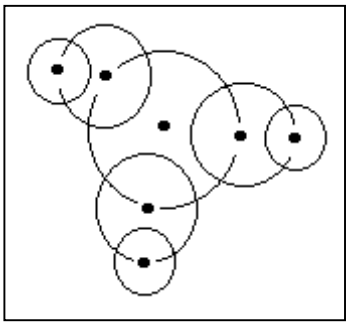


Figure 2: Structure of a lexical constellation (type 1)

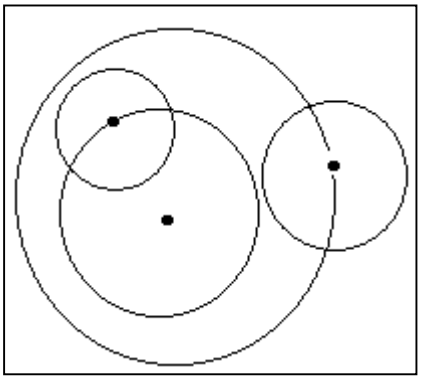


Figure 3: Structure of a lexical constellation (type 2)

The differences between plain (or linear) collocational analysis and constellational analysis are graphically represented in Figure 1, 2 and 3. In the three figures, a dot represents a lexical item, and a line represents a relationship of statistically significant co-occurrence.[1] Thus, a pair of dots connected by a line represents a collocational bi-gram. Additionally, in Figures 2 and 3 each circle symbolises a domain of lexical attraction.

Figures 2 and 3 represent different types of *lexical constellations*, the central category of description in the LCM. A lexical constellation is a collocational network hierarchically organised in two or more centres of lexical attraction. The first type of lexical constellation shown above (Figure 2) corresponds to the phenomenon of *indirect collocation*; the second type corresponds to patterns of *inter-collocability*[2] (Figure 3). These two classes of lexical constellations are described separately in the next subsections.

2.1 Indirect collocation

The phenomenon of indirect collocation was the first problem of lexical gravity interference to be investigated within the framework of the LCM (Cantos & Sánchez, 2001). This problem originates when a word so to say "intrudes" one of its collocates into the context of a another word. The phenomenon of indirect collocation is thus responsible for a large part of the "unwanted" items that are found in collocate lists.

The strategy adopted by the proponents of the LCM in order to detect cases of indirect lexical attraction among statistical collocates is based on comparisons of conditional probabilities. Once the statistically significant co-occurrences of a node have been extracted from the corpus using the conventional parameters of collocational analysis (defining a span, establishing a frequency threshold, selecting an association measure, etc.), the method proceeds to calculate the values of conditional probabilities between three different words: the node, the collocate and a candidate sub-collocate (i.e. a collocate which is suspected of being indirectly attracted because it shares more semantic features with other collocates than with the node).

[1] Following the Sinclairian line of thinking, *collocation* is defined in this study in statistical terms. Thus, it denotes a pair or group of words which co-occur with a probability greater than chance. However, we must be aware that this definition of collocation is controversial and has been criticised by notable experts in the field, especially by Bosque (2001). We will not tackle the debate here because the issue lies beyond the specific aims set for the present investigation.

[2] To avoid possible misunderstandings, a brief terminological note is in place here. The term *inter(-)collocation* is sometimes used in the literature to denote a reciprocal relationship of collocation. Thus, if a word *a* is a statistically significant co-occurrence of *b*, and *b* is a statistically significant co-occurrence of *a*, the two terms are said to form an inter-collocation. This notion of inter-collocation is not equivalent to the phenomenon that we call inter-collocability. The latter refers to a relationship between different collocational pairs.

Conditional probabilities are indicative of the strength of the dependency of one event on another event. For example, if we want to know how probable is the event *a* (say, the occurrence of a word *a*) given the occurrence of *b* as a fact, we can divide the total number of occurrences of *a* by the number of joint occurrences of *a* and *b* in a corpus. The value indicates the proportion of occurrences of *a* that take place in the company of *b*. This can be interpreted as an estimation of the dependency of the event *a* on the event *b*. The notation is P(b|a), which is read as follows: "the probability of *b* given the occurrence of *a*".

Thus, in previous research it was shown that *dental* collocates with *incidence* not because it is attracted towards *incidence* but because it has a strong dependency on another collocate of *incidence*, i.e. *caries* (Almela, 2011; Almela, Cantos & Sánchez, 2011). The probability of finding *dental* given *caries* in the Bank of English (55.2%) is more than a hundred times higher than that of finding *dental* given the occurrence of *incidence* in the same corpus (0.5%). This data is consistent with the observation that *dental* shares more semantic features with *caries* than with *incidence.* Thus, in Figure 2, the biggest circle can stand for *incidence*, the intermediate one for *caries*, and the smallest one for *dental*.

More generally, it was also found that collocates of *incidence* referring to body parts (*dental, heart, lung*, etc.) are more strongly attracted to other collocates of *incidence*, especially to those denoting a 'disease' or 'health problem' (e.g. *caries, attack, cancer*, etc.), than they are to the node. Hence, in expressions such as *incidence of dental caries*, *incidence of heart attack* or *incidence of lung cancer*, the modifier can be categorised as an indirect collocate of *incidence* (Almela, 2011; Almela, Cantos & Sánchez, 2011).

It is important to point out that grammar alone does not provide an explanation for the phenomenon of indirect collocation. Admittedly, in the above example there is a close correlation between phrase structure and the structure of the lexical constellation: the noun that modifies *incidence* is the direct collocate, and in turn, the modifier of the second noun is the indirect collocate of *incidence*. However, it should be added that in other cases the node has a closer syntactic connection with the indirect collocate than with the direct collocate. For instance, in expressions such as *caused by faulty design* or *caused by a defective gene*, the adjective is a direct collocate of the verb, and the noun is an indirect collocate (the data supporting this conclusion will be published in forthcoming research).

In sum, the analysis of indirect collocation in the LCM serves to uncover some discrepancies between statistical significance and lexical relevancy. From the fact that two or more words co-occur significantly in a corpus it does not necessarily follow that they are attracted to one another. One of the reasons for this is that there can be more than one centre of attraction within the same textual window.

The detection of cases of indirect collocation is useful in combinatorial lexicography, because it helps us to optimise the criteria for selecting the collocates of a headword. However, the implications are similar in printed and electronic dictionaries — the exclusion of irrelevant collocates is advisable in both types of dictionaries. Therefore, in what follows we will concentrate our analysis on the phenomenon of inter-collocability. As will be argued below, this second aspect of lexical constellations has important implications for the micro-structural design of collocation dictionaries, and consequently, it bears greater relevance for the discussion of issues that are specific to the field of electronic lexicography.

2.2 Inter-collocability

The concept of "inter-collocability" denotes the existence of dependency relations between different collocations of a word. The manifestations of inter-collocability are varied. In a positive sense, inter-collocability can be defined as the contribution which a collocation makes to the activation of another collocation of the same node. In a negative sense, inter-collocability can be defined in terms of restrictions on the combinational possibilities among different collocates of the same node.

As a method for identifying cases of inter-collocability we can use a variant of the technique employed for identifying cases of indirect collocation. Instead of calculating and comparing conditional probabilities between individual members of overlapping collocations, e.g. P(*a*|*b*), P(*a*|*c*), P(*b*|*c*), P(*b*|*a*), etc., we can calculate conditional probabilities between events of a larger size, for instance, the probability that a collocate of the node is selected given as a fact the co-occurrence of the node and another collocate: P(*c1*|*n,c2*), where *n* stands for the node, and *c1* and *c2* represent two different collocates. This value can then be compared with the corresponding conditional probability at the intra-collocational level, namely: P(*c1*|*n*).

Thus, given two collocates *c1* and *c2* of a node *n*, we can say that there is a relationship of positive inter-collocability between the pairs (*n,c1*) and (*n,c2*) if the probability of (*n,c1*) co-occurring with *c2* is higher than the probability of the node occurring with *c2* alone, or if the probability that (*n,c2*) co-occurs with *c1* is higher than the probability of the node co-occurring with *c1*. In the first case, we can say that *c2* is a "positive co-collocate" of *c1*, because the collocation (*n,c2*) is made more probable by the selection of *c1*; conversely, in the second case we say that *c1* is a *positive co-collocate* of *c2*, because the collocation (*n,c1*) is made

more probable by the selection of *c2*. The relationship of positive co-collocation can be mutual — that is, it can be observed in the two directions, from *c1* to *c2* and vice versa.

As for negative inter-collocability, we can say that *c2* is a *negative co-collocate* of *c1* if the capacity of the node for predicting the choice of *c1* is higher than the capacity of the collocation (*n,c2*) for predicting the choice of *c1*. This indicates that the collocation of the node with c2 diminishes the probability of finding the collocation (*n,c1*); conversely, we can say that *c1* is a negative co-collocate of *c2* if the selection of the collocation (*n,c2*) diminishes the probability of (*n,c1*). Like positive inter-collocability, negative inter-collocability can be mutual: *c1* and *c2* can be negative co-collocates of one another.

Inter-collocability is extremely frequent in patterns consisting of a verb and a noun phrase, especially when the noun phrase features a modifier-noun collocation. This reflects a characteristic of argument structure that we can describe as *valency stratification*. The capacity of a predicative lexeme, typically a verb, for restricting the lexical class of its arguments can extend over more than one layer of phrase structure.

At one level, the valency carrier restricts the class of the head of the valency filler (i.e. the noun heading the argument phrase). For instance, *return* selects nouns denoting 'data' (e.g. *value, string, integer, list*, *row, zero*, *tuple*, etc.) or 'goods' (*goods*, *vehicle, equipment, medicines*, etc.), among many others. This aspect of argument structure has been extensively described under different names. In valency theory it is described as a feature of *semantic valency*, along with semantic roles. In generative grammar, the terms employed are *selectional restrictions* and *s-selection*. Bosque opts for the term *lexical restrictions* (Bosque, 2001, 2004). This aspect of valency patterning has also been extensively described in valency dictionaries and similar reference works. In Herbst et al.'s (2004) *Valency Dictionary of English*, the arguments of verbs are assigned general semantic categories. For instance, the direct object of *translate* (in its primary meaning) is categorised as 'text'. Similarly, in P. Hanks' *Pattern Dictionary* the same argument of *translate* is categorised as 'document' — for more detailed information on the semantic categorisation of arguments in this dictionary, see Hanks & Pustejovsky (2005).

Less explored, however, is the second stratum of semantic valency. On top of restricting the lexical class of the head noun, the verb can also impose constraints on the collocability of different words within the argument phrase. Generally, these constraints exhibit a high level of semantic regularity — this justifies the treatment of valency stratification as a special feature of semantic valency patterning rather than as an idiosyncratic restriction.

One way of discovering patterns of valency stratification is to analyse adjectival co-collocates of verbs. The probability that the noun co-occurs with one or other adjectival collocate is often readjusted to the selection of a specific verbal collocate. Another way of approaching the phenomenon of valency stratification is by analysing inter-collocability relations in the reverse direction — that is, by analysing adjectival co-collocates of verbs. Because different modifier-noun collocations are associated with different verbs, the probability of finding a specific verb-noun collocation will be affected by the selection of different modifier-noun collocations. In principle, we can assume that these procedures are complementary. Both of them will be applied in the next section.

3. Lexical constellations at work

In this section the analytical framework sketched out above is applied to the description of lexical constellations formed with the noun *goods*. The analysis will be focused on capturing features of inter-collocability and valency stratification in verb-noun and modifier-noun collocations.

3.1 Method and results

The data and the examples have been extracted from the *ukWaC* corpus (1,565,274,190 tokens), accessible at the SketchEngine query system. All queries are syntactically restricted. We have taken into account only occurrences of the noun phrase (i.e. the adjective-noun collocation) as a direct object of the verb in an active construction, or as the subject in a passive construction (the connection between the two constructions is that in both cases the collocation ADJ+*goods* performs the semantic role THEME). The WordSketch function proved very useful in limiting our queries to the foregoing grammatical scheme. Nevertheless, manual supervision was required in order to detect possible parsing errors.

Following the remarks made at the end of section 4, we have approached the phenomenon of inter-collocability from two complementary perspectives. Tables 1-3 reflect the perspective provided by the analysis of adjectival co-collocates, and Tables 4-6 reflect the perspective provided by verbal collocates.

The criteria applied in the selection of the potential co-collocates were aimed at testing the initial hypothesis that the lexical constellations of *goods* follow highly systematic semantic patterns (at this point it should be remembered that in section 3 valency stratification was described as a special feature of semantic valency). The verbs *return, replace* and *reject* have been selected because they share important aspects of meaning. In collocation with *goods* they denote an action whereby the consumer does not accept the goods initially bought or received. As for the adjectives *faulty*, *defective* and *damaged*, they all describe a 'flaw' or 'imperfection'.

The results are shown in Tables 1-6. In each table the left-most column is a list of collocates of *goods*. In Tables 1-3 these collocates are modifiers (more specifically adjectives)[3], and in Tables 4-6 they are verbs. A frequency threshold and a statistical filter were applied to all the collocates. We made sure that all of them co-occur at least three times with *goods* (in the specified grammatical framework), and that they all are statistically significant co-occurrences of this noun. Statistical significance was defined in terms of *logDice* — for an explanation of the advantages of this measure the reader is referred to Rychlý (2008). Again, these data were obtained from the WordSketch function at SketchEngine.

The next two columns indicate raw frequency data. The first of them indicates the frequency of the whole 3-gram (verb, modifier, noun) in the corpus (for instance, the frequency of *return defective goods*). A minimum frequency threshold of 3 was also applied in this column. This was motivated by purely practical reasons that are independent of the research methodology: the list of 3-grams with a frequency lower than two would generate excessively long tables difficult to fit into the size of this paper.

The second frequency data column corresponds to the collocational pair formed by the noun (*goods*) and each of the collocates listed in the left-most column. Thus, in Tables 1-3 this column specifies the frequency of modifier-noun collocations (e.g. *faulty + goods, defective + goods*, etc.), while in Tables 4-6 the same column indicates the frequency of verb-noun collocations (e.g. *return + goods, replace + goods*, and so on). These data were obtained by checking the results in different SketchEngine tools (Concordance, WordSketch, Collocation, etc.).

As for the last two columns, they indicate values of conditional probabilities between collocations and between words, respectively. The first of these columns returns the value of P(*m*|*v,n*) in Tables 1-3, and of P(*v*|*m,n*) in Tables 4-6. The first formula can be read as "the probability that the modifier occurs given the occurrence of the verb+noun collocation" (where the noun is always *goods*). Correspondingly, the second formula can be read as "the probability that the verb occurs given the occurrence of the modifier+noun collocation". Finally, the right-most column returns the value of P(*m*|*n*) in Tables 1-3, and of P(*v*|*n*) in Tables 4-6. The first value reflects the probability that the modifier occurs given the occurrence of the noun; the second one specifies the probability that the verb occurs given the selection of the noun. In all the tables the order of the rows is determined by the difference between the values of these two columns. Thus, the word at the top of the list is the best candidate for positive co-collocate.[4]

	f(v,m,n)	f(m,n)	P(m|v,n)	P(m|n)
faulty	35	354	2.35%	0.36%
unwanted	21	149	1.41%	0.15%
defective	20	137	1.34%	0.14%
unused	7	20	0.47%	0.02%
undamaged	6	10	0.40%	0.01%
damaged	8	209	0.54%	0.21%
non-faulty	4	11	0.27%	0.01%
stolen	8	434	0.54%	0.44%

Table 1: Adjectival co-collocates of *return*.[5]

	f(v,m,n)	f(m,n)	P(m|v,n)	P(m|n)
faulty	30	354	19.11%	0.36%
defective	12	137	7.64%	0.14%
damaged	12	209	7.64%	0.21%
electrical	6	850	3.82%	0.86%

Table 2: Adjectival co-collocates of *replace*.[6]

	f(v,m,n)	f(m,n)	P(m|v,n)	P(m|n)
faulty	6	354	5.41%	0.36%
defective	3	137	2.70%	0.14%

Table 3: Adjectival co-collocates of *reject*.[7]

	f(v,m,n)	f(v,n)	P(v|m,n)	P(v|n)
return	35	1491	9.89%	1.50%
replace	30	157	8.47%	0.16%
receive	19	913	5.37%	0.92%
buy	17	1592	4.80%	1.60%
reject	6	111	1.69%	0.11%
supply	6	961	1.69%	0.97%
collect	3	270	0.85%	0.27%
sell	7	2237	1.98%	2.25%

Table 4: Verbal co-collocates of *faulty*.[8]

[3] A priori we did not decide to exclude noun modifiers from this list (e.g. *consumer goods*, *household goods*, etc.). However, for some reason, none of the modifiers that met the conditions set in the first three columns were nouns; all of them were adjectives.

[4] In Tables 5 and 6, the position of *deliver* at the bottom of the list might be misleading. The value of P(v|n) in this row is inflated by occurrences of the idiom *deliver the goods*. If we were able to exclude this idiom from the count of collocations of *deliver + goods*, the difference with P(v|m,n) would be greater in Table 5, and in Table 6 the value of P(v|m,n) would be higher than P(v|n). However, the occurrences of *deliver the goods* as an idiom cannot be separated automatically from those of *deliver the goods* as a collocation, and doing it manually is far too time-consuming a task to be considered a convenient method in lexicography.

[5] F(*return,goods*) = 1491

[6] F(*replace,goods*) = 157

[7] F(*reject,goods*) = 111

	f(v,m,n)	f(v,n)	P(v|m,n)	P(v|n)
return	20	1491	14.60%	1.50%
replace	12	157	8.76%	0.16%
reject	3	111	2.19%	0.11%
inspect	3	121	2.19%	0.12%
deliver	4	1930	2.92%	1.94%

Table 5: Verbal co-collocates of *defective*.[9]

	f(v,m,n)	f(v,n)	P(v|m,n)	P(v|n)
receive	15	813	7.18%	0.82%
replace	12	157	5.74%	0.16%
return	8	1491	3.83%	1.50%
inspect	3	121	1.44%	0.12%
deliver	4	1930	1.91%	1.94%

Table 6: Verbal co-collocates of *damaged*.[10]

The frequency of the noun remains constant in all the tables. The frequency of the noun *goods* in the corpus is 99393 (substantivisations of the adjective *good* were excluded from this count). Besides, the frequency of verb-noun collocations remains constant within each of the first three tables. Likewise, the frequency of modifier-noun collocations remains constant within each of the last three tables (4-6). Therefore, the figures are indicated in a footnote added to the caption.

3.2 Analysis and discussion

The results displayed in Tables 1-6 lend strength to the initial hunch that the lexical constellations of *goods* exhibit a high degree of semantic systematicity. The strongest positive co-collocates tend to be grouped together around a common core of meaning.

In Tables 1-3 the dominant group of adjectives is formed by words depicting a 'flaw': *faulty, defective, damaged*. Observe that *faulty* and *defective* occur in the three tables, and that in all of them *faulty* lies at the top. The fact that *unwanted* is a stronger co-collocate than *defective* in Table 1 does not run counter to the general pattern, because the meaning of *unwanted* is conceptually related to *faulty*, *defective* and *damaged* (as a rule, goods that are in a bad condition are not desired by the consumer).

Particularly significant are the values of conditional probabilities in Table 2. Observe that the capacity of the collocation *replace goods* for predicting the choice of *faulty* reaches 19.11 percent, a figure more than 50 times higher than the capacity of the noun *goods* for predicting the selection of *faulty*. This constellation is thus a very good example of the kind of dependency relation depicted in Figure 3 (see section 2). If we insert these lexical items in Figure 3 we obtain the picture below:

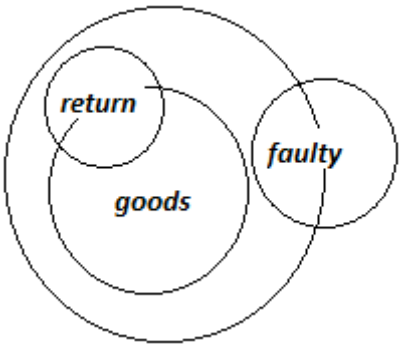


Figure 4: Positive inter-collocability

The results displayed in Tables 4-6 are equally coherent from the point of view of meaning. The dominant group is formed by verbs implying a decision of 'non-acceptance of the goods received': *return, replace, reject*. The verbs *return* and *replace* appear in the three tables, and in two of them, *return* is the strongest co-collocate.

Overall, the semantic regularities observed in these lexical constellations suggest that verb-noun collocations expressing 'non-acceptance of goods' are likely to converge with adjective-noun collocations describing goods as 'having a flaw'. This speaks strongly for the conception of lexical constellations as surface lexical realisations of underlying conceptual (cognitive) structures. In the same line of reasoning, it would be interesting to determine the extent to which lexical constellations are language-independent. Obviously, this objective cannot be pursued in the present article, because it requires more empirical research in English and in other languages.

Another interesting remark concerns the consistency of the findings obtained in the two groups of tables (1-3 and 4-6). The output of Tables 1-3 overlaps with the input of Tables 4-6, and vice versa. The dominant adjectives in Tables 1-3 coincide roughly with the elements analysed in Tables 4-6, and conversely, the dominant verbs in Tables 4-6 contain the elements analysed in Tables 1-3. This confirms the claim made in section 3.2 that co-collocation can be mutual. *Defective* is a co-collocate of *return*, and conversely, *return* is a co-collocate of *defective* (see Tables 1 and 4). The same holds true for other pairs: (*defective, replace*), (*defective, reject*), (*faulty, return*), (*faulty, replace*), (*faulty, reject*), (*damaged, return*), (*damaged, replace*). This reinforces the idea that the two perspectives on valency stratification (the one provided by verbal co-collocates and the one provided by modifiers) are complementary and lead to relatively similar results.

Finally, it should be noted that the prevalence of positive co-collocates over negative ones in Tables 1-6 results

[8] F(*faulty,goods*) = 354
[9] F(*defective,goods*) = 137
[10] F(*damaged,goods*) = 209

mainly from the decision to set a minimum frequency threshold for the 3-gram. If the analysis had been focused on verbs or adjectives occurring in low-frequency 3-grams, we would have obtained several prominent patterns of negative inter-collocability. Interestingly, these patterns can also be characterised by a high degree of semantic regularity.

A case in point is the relationship between verbs such as *ship* and *transport* and the adjectives analysed in tables 4-6. There is evidence that *ship* and *transport*, which are quasi-synonyms, are negative verbal co-collocates of *faulty*, *defective* and *damage*. The probability of these verbs given the occurrence of *goods* is 0.31 percent in the case of *ship* (309/99393), and 0.43 percent in the case of *transport* (426/99393). These figures, however low, are considerably greater than the probability of these verbs occurring in the context of modifier-noun collocations such as *faulty goods*, *defective goods*, or *damaged goods*. In almost all these cases the probability is zero. In the whole ukWaC corpus, which, it should be emphasised, contains more than one billion words, there is no single instance of 3-grams such as *ship defective goods*, *transport faulty goods*, *transport damaged goods*, etc. The sequence *ship faulty goods* yields one hit, but obviously the value of P(*ship*|*faulty goods*) is lower than P(*ship*|*goods*). Clearly, the collocations *ship/transport goods* tend to avoid the selection of modifiers describing a 'flaw' or 'imperfection'. This can be interpreted as an indication that semantic systematicity is a characteristic both of positive and of negative inter-collocability.

4. Lexical constellations in lexicography

From the previous sections we can draw the overall conclusion that the choice of a collocation influences the range of choice of other collocations in the same context. The choice of a collocation can contribute to activating or blocking other collocations of the same node. Once this fact has been established, the question that needs to be addressed is: should lexical constellations be recorded in combinatorial dictionaries, and if so, what are the appropriate lexicographic techniques for dealing with them? The first part of the question is answered in 4.1. The answer to the second part of the question is given in 4.2. In Section 4.3 we explain the guidelines for our lexicographic project and present some examples.

4.1 The relevance of constellational information

Lexical constellations provide a potentially useful type of information in a collocation dictionary. One of the main functions of this kind of dictionary is to assist the user –typically a foreign or second language speaker– in achieving native-like, fluent composition. Precisely, lexical constellations are one of the principal resources of fluency and cohesion in a text, because they make the word fit within a context broader than the simple collocational bi-gram. Compared to the simple collocation, a lexical constellation provides, so to say, an extended pattern of lexical cohesion.

Apart from this general consideration, there are two more specific arguments for introducing lexical constellations into collocation dictionaries. The first of these arguments concerns the strength of constellational patterns. In some respects, these patterns are stronger than most of the simple collocational bi-grams recorded in a conventional combinatorial dictionary. Observe, for example, that the dependency of the collocation *defective goods* on *return*, measured in terms of conditional probability, is ten times higher than the dependency of *goods* on *return* (see Table 5). In this light it is difficult to justify why the weaker pattern (the bi-gram) should be included in a dictionary while the stronger pattern (the constellation) is omitted.

A further argument for the incorporation of collocational data refers to the connection of form and meaning. The syntagmatic behaviour of words is closely associated with their semantic properties. Therefore, collocation is more than a surface co-occurrence pattern; it also provides a representation of word meaning (Renouf, 1996). Knowing the collocations of words is a contributing factor to the development of lexical semantic competence. This idea, which was formulated by Firth in his well-known definition of "meaning by collocation", has inspired much of the work conducted in corpus-driven lexicology, both theoretical and applied. Lexical constellations can help to provide a much more detailed and refined account of the connections between context and meaning. Notice, for example, that some semantic aspects of adjectives such as *faulty*, *defective*, or *damage* are better represented by their verbal co-collocates (*reject, return, replace*) than by the noun (*goods*). The discovery of a 'flaw' is causally connected with the decision of 'non-acceptance', and this decision is implied by the meaning of verbs such as *reject, return*, or *replace*, but not by the meaning of *goods*.

Considering these arguments, we can conclude that lexical constellations can improve the utility of collocation dictionaries. Having answered this question, the next problem to be resolved concerns the know-how. Clearly, the incorporation of lexical constellations requires the development of innovative practices, because current collocational dictionaries do not provide this kind of information. This gives rise to the question: what exactly are changes that have to be introduced in order to accommodate lexical constellations into collocation dictionaries? The issue is addressed below.

4.2 The treatment of constellational information

By definition, a lexical constellation always involves some form of interaction between different collocations of a node. However, in a standard collocation dictionary the different words in an entry are directly related to the headword and not to one another. Therefore, the main obstacle that has to be overcome in order to integrate constellations within collocation dictionaries is the lack

of explicit connection between different collocates of a headword.

The incorporation of lexical constellations requires us to take a step from an "intra-collocational" to an "inter-collocational" perspective. Thus far, the analysis of syntagmatic dependencies in collocation dictionaries, both printed and electronic, has been focused on relationships between the parts of a collocation. In this sense, we can say that combinatorial lexicography has done justice to Sinclair's (1991) remark that the choice of a word affects the choice of other words in its vicinity. What combinatorial lexicography has so far failed to reflect is the fact that the choice of a collocation can also affect the choice of other collocations in its vicinity.

In a conventional collocation dictionary the user is not provided with information concerning how different elements and sections in an entry can or tend to be combined in the discourse. There is, of course, information about the relationship between the headword and each of the collocates. However, this is not complemented by any specification of whether particular collocations or groups of collocations of the lemma tend to attract or repel each other.

For example, in *Macmillan Collocations Dictionary* (MCD), to quote the most recent major dictionary of English collocations, *faulty* and *return* are presented as different categories of collocates of the noun *goods*. *Faulty* is one of the three adjectives in this entry, along with *defective* and *damaged*, which are labelled as expressing the meaning 'not working properly'. *Return* is one of the four verbal collocates in the same entry which are ascribed the meaning 'send goods' (the others are *deliver*, *transport* and *ship*). From this information we can gather that *return faulty goods* and *transport faulty goods* are possible lexical combinations expressing the meaning 'send goods which do not work properly'. What we are not told, however, is that the selection of the modifier is adjusted to choice of a verbal collocate, and that the selection of *return goods* makes the selection of the adjective *faulty* highly probable, while the collocations *transport goods* and *faulty goods* tend to avoid each other. That is to say, the MCD does not inform us that some pairs of verbal and adjectival collocates of *goods* are more likely to converge in the same complex expression than others.

These facts are not reflected in the MCD or in any other major collocation dictionary, because the design does not contemplate any form of interaction between different collocations in an entry. The same remark applies to other important combinatorial dictionaries of English, notably the BBI and the OCD, or of other languages such as Spanish (e.g. REDES).

This criticism can also be made of electronic collocation dictionaries, such as the *Diccionario de Colocaciones del Español* (DiCE, an online dictionary of Spanish collocations), as well as of electronic versions of printed dictionaries (e.g. the OCD on CD-ROM). In none of these resources is the user provided with specifications of how the selection of a collocate influences the range of choice of further collocates of the same headword. Observe, for example, Figure 5, where we reproduce an entry from the OCD on CD-ROM. Here, we find some of the adjectival and verbal collocates of *goods* mentioned above (*faulty, defective, deliver, transport*, etc.), but again, no specification is given of their inter-collocability.

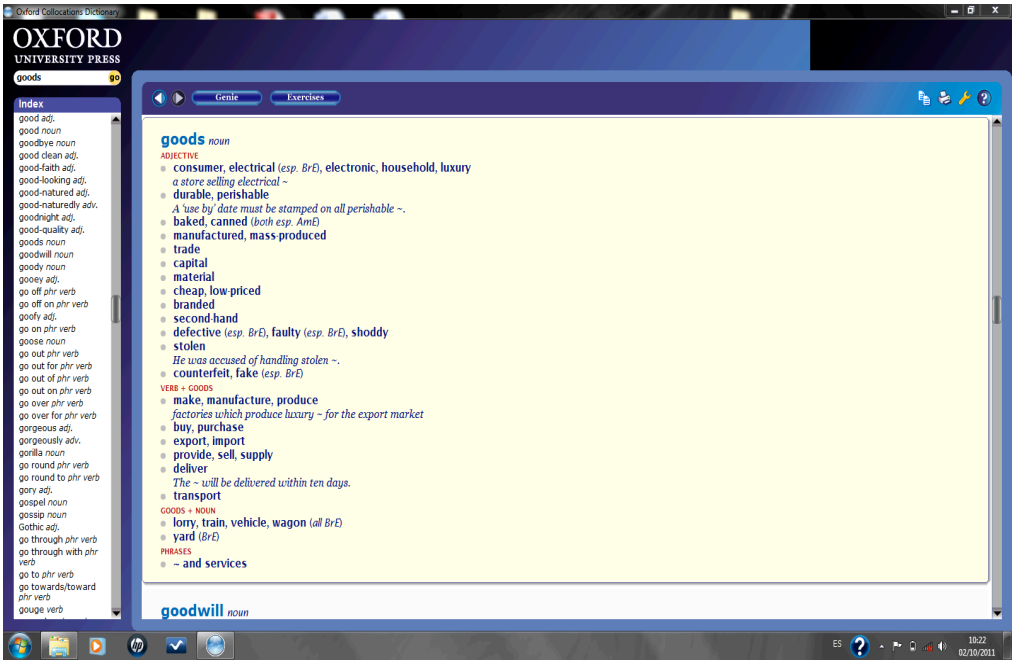


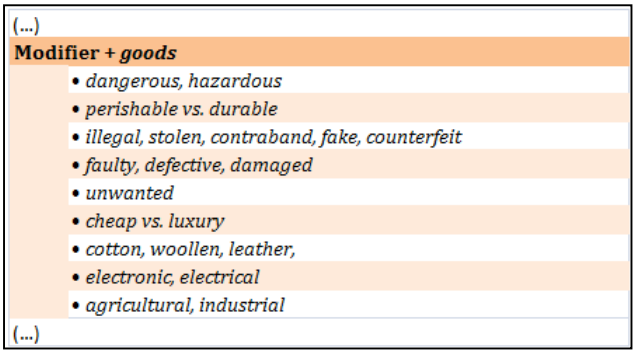
Figure 5: An entry from the OCD on CD-ROM

Logically, the possibilities of accommodating lexical constellations are not equal for printed dictionaries and electronic dictionaries. The printed format imposes a number of material conditions which render the incorporation of lexical constellations virtually impracticable. Supplying this kind of information in a printed dictionary would imply an excessive increase in size, probably beyond what is commercially viable. However, these practical difficulties can be resolved in an electronic dictionary. The user interface allows an interactive management of the information contained in lexical entries. With a simple click, the user can choose to expand the information on the collocations associated with a particular item, and precisely, one of the choices that can be made available in this menu is the generation of a list of collocates that are attracted to specific collocations of the lemma. For these reasons, we think that, in the present state of the art, the project of developing a collocation dictionary that includes lexical constellations is conceivable only in electronic format.

4.3 The *Dynamic Combinatorial Dictionary*

The treatment of lexical constellations in our lexicographic project, the DCD, follows four main guidelines: *dynamicity*, *progressiveness*, *compactness* and *systematicity*. Firstly, the micro-structural design is dynamic, because the information presented in a lexical entry is readjusted to the selections made by the dictionary user. This is why the project has been called a *Dynamic Combinatorial Dictionary*. This means, for example, that by clicking on the collocate *faulty* under the entry for *goods* the positive verbal co-collocates (e.g. *return, replace, reject*) are foregrounded, and the negative ones are omitted.

Secondly, the step from simple collocational bi-grams to lexical constellations is made in a progressive manner. As a default option, the entry offers only plain collocational information. The user is not provided with information on lexical constellations before s/he clicks on a specific collocate in search for more detailed information, and when this happens, the entry zooms in to show only the most relevant contextual data. That is, in the transition from purely collocational information to constellational information the dictionary leaves out all those elements which are not positive co-collocates of the items selected by the user. The principle behind this criterion is one of user-friendliness. It is not advisable to increase at the same time the level of detail and the amount of information. An increase in the depth of information should be compensated by a decrease in the width of information.


(...)
Modifier + *goods*
- *dangerous, hazardous*
- *perishable vs. durable*
- *illegal, stolen, contraband, fake, counterfeit*
- *faulty, defective, damaged*
- *unwanted*
- *cheap vs. luxury*
- *cotton, woollen, leather,*
- *electronic, electrical*
- *agricultural, industrial*

(...)

Figure 6: Extract from a DCD entry (first stage)


Verb + Modifier + *goods*
- *dangerous, hazardous*

⇨ TRANSPORT THINGS FROM ONE PLACE TO ANOTHER

Verb + Modifier + *goods*
- *illegal, stolen, contraband, fake, counterfeit*

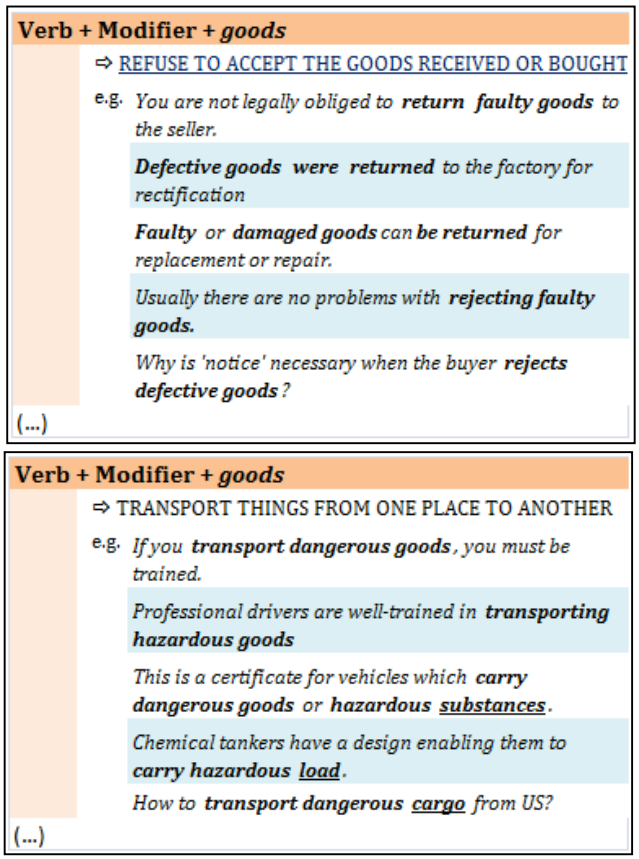
⇨ TAKE THINGS SECRETLY TO OR FROM A PLACE

Verb + Modifier + *goods*
- *faulty, defective, damaged*

⇨ REFUSE TO ACCEPT THE GOODS RECEIVED OR BOUGHT

⇨ PROVIDE NEW GOODS

Figure 7: Extracts from a DCD entry (second stage)


Verb + Modifier + *goods*

⇨ REFUSE TO ACCEPT THE GOODS RECEIVED OR BOUGHT

e.g. *You are not legally obliged to **return faulty goods** to the seller.*

***Defective goods were returned** to the factory for rectification*

***Faulty** or **damaged goods** can **be returned** for replacement or repair.*

*Usually there are no problems with **rejecting faulty goods.***

*Why is 'notice' necessary when the buyer **rejects defective goods**?*

(...)

Verb + Modifier + *goods*

⇨ TRANSPORT THINGS FROM ONE PLACE TO ANOTHER

e.g. *If you **transport dangerous goods**, you must be trained.*

*Professional drivers are well-trained in **transporting hazardous goods***

*This is a certificate for vehicles which **carry dangerous goods** or **hazardous substances**.*

*Chemical tankers have a design enabling them to **carry hazardous load**.*

*How to **transport dangerous cargo** from US?*

(...)

Figure 8: Extracts from a DCD entry (third stage)

The level of detail or granularity in DCD entries unfolds gradually through three different steps. In a first stage, the screen displays collocational pairs, similarly to conventional collocation dictionaries (Figure 6). In a second stage, the screen displays a semantic description of lexical constellations related to the collocate on which the user has clicked (see Figure 7). Finally, in a third stage, the user is provided with a series of examples representing different lexical realisations of the constellation (see Figure 8). This list is accessed by clicking on the semantic description of the constellation. Where relevant, the list includes references to other headwords sharing in the same lexical constellation pattern (e.g. *cargo, load, substance*, etc., in the lower part of Figure 8). In these cases, the words are underlined so that the user can follow the link to the corresponding noun entry.

Concerning the third guideline, i.e. compactness, information about lexical constellations is presented in a format as succinct as possible. One implication is that labels such as "lexical constellation", "inter-collocability" or "positive co-collocate" are not explicitly mentioned by any means in the entry. This marks a difference with some collocation dictionaries, especially in the Meaning-Text Theory (MTT) framework (notably the DiCE), which make extensive use of specialised terms that are not known to the wider audience and the lay speaker. These terms include MTT jargon such as *gloss* and lexical function labels such as 'Magn', 'Anti Bon', etc. In the DCD project we try to make the dictionary accessible by keeping metalinguistic data to a minimum. Metalinguistic information is reduced to basic grammatical categories (Verb, Noun, Adjective, etc,) and to semantic labels. For similar reasons, probability and statistical data are not shown to the user. The structure of constellations is signalled only by means of symbols such as arrows, and by highlighting words in authentic examples (see Figures 7 and 8).

Finally, the fourth guiding principle is the maximisation of systematicity. This apparently trivial statement contains important implications for the design of dictionary entries. It entails, among other things, the attempt at subsuming as much lexical information as possible under general combination rules. This implies first and foremost that semantic labels will be used to show the interconnectedness of several collocational patterns.

This practice, i.e. the grouping of different collocations under meaning categories, has been adopted to a greater or lesser extent by previous collocation dictionaries such as MCD, REDES and the DiCE, but no by others such as the OCD or the BBI. The specific challenge faced now by the DCD is to extend this strategy to apply to the description of semantic regularities underlying lexical constellations. This problem is resolved by inserting semantic paraphrases of constellations at an intermediate stage between collocational information and real examples of constellations (see Figures 7 and 8).

The rationale behind this emphasis on the connection of combinatorial and semantic properties of words is our strive for abridging the distance between the collocation dictionary and the general-purpose dictionary. In the line of neo-Firthian thinking, it is our conviction that a well-organised, detailed description of the syntagmatic behaviour of a word has a definitional value. Collocation provides a representation of word meaning, as Firth suggested.

5. Conclusion

In this article we have argued that the mainstream approaches to collocation have missed an important aspect of collocational patterning, namely, the operation of dependency relations between different collocations. Crucially, this level of analysis should not be confused with observation of dependency relations between the parts of a collocation. Collocability must be analysed at a different level than inter-collocability.

It has also been argued that the LCM provides an adequate analytical framework for inter-collocability. After applying the methodology of constellational analysis to collocational patterns of the noun *goods*, we have confirmed that different collocations influence in different ways the selection of other collocations of the same noun.

Finally, we have explained that dealing with lexical constellations in a dictionary is only possible in an electronic format and requires us to introduce a number of substantial changes with respect to the conventional micro-structural design of collocation dictionaries (including electronic ones). Some of these changes have been illustrated with reference to sample parts from the DCD.

6. Acknowledgements

The project presented in this paper is generously funded by a grant from *Fundación Séneca, Agencia de Ciencia y Tecnología de la Región de Murcia* (Ref. 08594/ PHCS/08). We are most grateful for this financial support.

7. References

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

Almela, M., Cantos, P. & Sánchez, A. (2011). From collocation to meaning: revising corpus-based techniques of lexical semantic analysis. In I. Balteiro (ed.) *New Approaches to Specialized English Lexicology and Lexicography*. Newcastle u. T.: Cambridge Scholars Press, pp. 47-62.

The BBI Dictionary of English Word Combinations

(1997). Compiled by M. Benson, E. Benson & R. Ilson. Amsterdam: John Benjamins.

Bosque, I. (2001). Sobre el concepto de 'colocación' y sus límites. *Lingüística Española Actual*, 23(1), pp. 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*. Bellaterra: Universitat Autonoma de Barcelona, pp. 13-58.

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

DiCE: Diccionario de colocaciones del español. Accessed at: http://www.dicesp.com.

Hanks, P., Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. *Révue Française de Linguistique Appliquée*, 10, pp. 63-82.

Herbst, T., Heath, D., Roe, I.F. & Götz, D. (2004). *A Valency Dictionary of English. A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Berlin: Mouton de Gruyter.

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*. Amsterdam: Rodopi, pp. 267-280.

Macmillan Collocations Dictionary for Learners of English (2010). Compiled by M. Rundell. Oxford: Macmillan.

Oxford Collocations Dictionary for Students of English (2009). Compiled by C. McIntosh. Oxford: Oxford University Press.

A Pattern Dictionary of English Verbs. Accessed at: http://deb.fi.muni.cz/pdev/.

REDES: Diccionario combinatorio del español contemporáneo (2004). Compiled by I. Bosque. Madrid: SM.

Renouf, A. (1996). Les nyms: en quête du thésaurus des textes. *Lingvisticae Investigationes*, 20(1), pp. 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*. Brno: Masaryk University, pp. 6-9.

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