Investigating Lexis

Investigating Lexis

Vocabulary Teaching, ESP, Lexicography and Lexical Innovation

Edited by

José Ramón Calvo-Ferrer
and Miguel Ángel Campos Pardillos

Cambridge
Scholars
Publishing

Investigating Lexis:
Vocabulary Teaching, ESP, Lexicography and Lexical Innovation

Edited by José Ramón Calvo-Ferrer
and Miguel Ángel Campos Pardillos

This book first published 2014

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-6807-8
ISBN (13): 978-1-4438-6807-5

TABLE OF CONTENTS

Foreword ix

Part I. Lexical Theory and Acquisition

Chapter One 3
"You Shall Know a Collocation by the Company It Keeps":
Methodological Advances in Lexical-Constellation Analysis
Moisés Almela

Chapter Two 27
From EFL to CLIL: Teaching and Learning Vocabulary in the Primary Classroom
María Tabuenca Cuevas and Gema Alcaraz Mármol

Part II. Legal Terminology

Chapter Three 41
Pronunciation Skills in Legal English for Interpreters:
English Latinisms and Cognates
Miguel Ángel Campos Pardillos

Chapter Four 57
The U.S. Supreme Court Cognitive Metaphors: Law, Deep Roots
and the Right Soil
Laura Vegara Fabregat

Chapter Five 79
A Comparative Study of Latinisms in the European Legislation:
Degrees of Syntactic Integration in English, Spanish and Greek
Document Versions
Rita Winiarska

Part III. Dictionaries: Past and Present

Chapter Six 95
Online Arabic-English-Arabic Specialised Dictionaries
Reima Al-Jarf

Chapter Seven... 103
General Remarks on the Challenges of Integrating Scientific
and Technical Words in General Dictionaries
Isabel Balteiro and José Ramón Calvo-Ferrer

Chapter Eight... 123
Gentyll On-Line Glossaries: Professional Titles of Women and Men
in a Series of Fields of Activity
Mercedes Bengoechea, María Rosa Cabellos and José Simón

Chapter Nine.. 133
The Dictionary of Richard Percyvall's *Bibliotheca Hispanica* (1591):
Structure and Composition
David Carrascosa Cañego

Chapter Ten 145
First Anglicisms in the Spanish Press: Treatment Given by the Royal
Spanish Academy Dictionary
María Vázquez-Amador & M. Carmen Lario-de-Oñate

Part IV. New Challenges, New Approaches

Chapter Eleven 163
Swingvergüenzas A Contra Blues: A Study on Creative Code Mixing
in Spanish Music
Paula López Rúa

Chapter Twelve 181
New Challenges in the Translation of Terminology for Software
Applications
José Ramón Belda Medina

Chapter Thirteen 191
The Terminology of the Video Game Industry: A New Type
of Specialised Language
Ramón Méndez González

Chapter Fourteen 205
The Multiple Shades of Erotica: Translating Romantic
and Erotic Fiction into Spanish
Scheherezade Surià

FOREWORD

This book is a collection of essays representing various aspects of lexicography; in all cases, its authors have attempted to combine state-of-the-art research with a user-friendly approach which makes it attractive for readers worldwide. It is divided into four major sections: (i) Lexical theory and acquisition, (ii) Legal terminology, (iii) Dictionaries, (iv) New challenges.

The first section, which deals with basic theoretical issues regarding vocabulary, consists of two studies. The first one, "You Shall Know a Collocation by the Company It Keeps": Methodological Advances in Lexical-Constellation Analysis" studies collocations, offering a new development within standard collocational analysis, that of lexical constellations, which improves semantic description through the use of clusters. The other paper in this section, "Vocabulary in Primary Language Learning" revisits the first stages of vocabulary learning and acquisition by analysing lexical choices in primary school textbooks, and the implications such choices have upon primary school teaching.

The second section focuses on legal terminology, an area of specialized vocabulary which has received comparatively less attention than others (e.g. business), but the importance of which is progressively gaining ground with the weight of transnational political bodies. The first paper, "Pronunciation Skills in Legal English for Interpreters: English Latinisms and Cognates" analyses a gap in interpreter training regarding the use of adapted and unadapted words of Latin origin in legal English, and emphasizes the need for a more thorough approach to specialized vocabulary learning which also includes pronunciation. This is followed by an innovative study of another almost uncharted territory: legal metaphors, entitled "The US Supreme Court Cognitive Metaphors and their Translation into Spanish: Law, Deep Roots and the Right Soil". This paper examines the conceptual framework underlying some common metaphors in legal English, insofar as they shape legal reasoning and become a challenge for translators. The third paper in this section, "A Comparative Study of Latinisms in European Legislation. Degrees of Syntactic Integration in English, Spanish and Greek Document Versions", deals with Latin expressions and how they have made their way into

English, Spanish and Greek syntax, while in most cases remaining clearly recognizable as a sort of legal lingua franca.

The third part of the book offers new insights into dictionaries, a tool which has had to face new challenges over the last decades and with the advent of globalization and new technologies. The first chapter, “Online Arabic-English-Arabic Specialized Dictionaries”, points out some of the shortcomings of the dictionaries presently available for specialized translation between English and Arabic, and makes a few suggestions for improvement in terms of coverage, accuracy and inclusiveness. The second contribution, “General Remarks on the Challenges of Integrating Scientific and Technical Words in General Dictionaries” deals with a problem that lexicographers have traditionally faced, and has received scant scholarly attention: i.e. to which extent scientific terms should be included in a dictionaries intended for the general public, and the amount of information they should offer. The third paper, “Gentyll On-Line Glossaries: Professional Titles of Women and Men in a Series of Fields of Activity”, is a poignant revelation of how, in spite of the alleged gender neutrality and political correctness in lexicographic practice, terminological banks continue to base human denomination on the masculine term. For their part, the fourth and fifth papers look at past lexicographic practice. The first one, entitled “The Dictionary of Richard Percyvall’s Bibliotheca Hispanica (1591): Structure and Composition”, provide insight into a 16th century grammar-cum-dictionary for English learners of Spanish, analysing the criteria used, the English translation suggested and the role of Latin in its creation process. The second one, “First Anglicisms in the Spanish Press: Treatment Given in The Royal Spanish Academy Dictionary” describes the practices by domestic prescriptive authorities with regards to the admission of Anglicisms in Spanish –a rather controversial issue, especially when the self-declared role of such authority is to “cleanse” the language from impurities.

The final section, entitled “New challenges, new approaches”, explores areas of vocabulary absent so far from academic analysis, either because they correspond to recent technological developments, or because scholars might have feared to tread into controversial territory. In the first study, “Swingvergüenzas A Contra Blues: A Study On Creative Code Mixing In Spanish Music”, cases of linguistic creativity are analysed, with a number of word-formation processes half-way between code-mixing and true borrowing whose stylistic effectiveness makes them attractive to music audiences. The second paper, “New Challenges in the Translation of Language for Software Applications”, looks into the localisation of software applications and the lexical problems such translation process entails and provides numerous examples to illustrate the major challenges translators face when adapting these terms into Spanish by using different lexical resources. The third contribution, “The Terminology of The Video Games Market: A New Type of Specialized Language” deals with a completely new product, whose innovativeness can only be equalled by its financial weight as a thriving area, and whose terminology may be challenging for those translators not familiar with the whole concept. Finally, “The Multiple Shades of Erotica: Translating Romantic and Erotic Fiction into Spanish” is a daring study of the lexical choices made by translators faced with erotic content, focusing on one of the most recent literary best-sellers whose explicit language regarding sexual organs and practices forces the translator to open new paths in the literary vocabulary in the target language.

PART I

LEXICAL THEORY AND ACQUISITION

CHAPTER ONE

"YOU SHALL KNOW A COLLOCATION BY THE COMPANY IT KEEPS": METHODOLOGICAL ADVANCES IN LEXICAL-CONSTELLATION ANALYSIS

MOISÉS ALMELA

Introduction

Traditionally, collocation has been conceived of as a bipartite structure. In all the dominant approaches to collocation, the structure of this type of word combinations is analysed into two parts – though not necessarily into two words, because one of the two parts of a collocation can be a complex item (García-Page 2011). In fact, as Martin (2008) remarks, the notion that collocation is made up of two parts is one of the few points over which most experts in the field are generally agreed.

This is not to deny that there are fundamental discrepancies concerning the way the two parts are categorised. Thus, the distinction between *base* and *collocator* in the literature on phraseology establishes a hierarchical relation between an autonomous item and a dependent element (Hausmann 1979, 1990, 1998; Írsula Peña 1994; Liang 1991), while the distinction of node and collocate in corpus linguistics is purely methodological and distinguishes only between input and output in the process of collocation extraction (Jones and Sinclair 1974; Krishnamurthy 2004; Phillips 1985; Sinclair 1991). However, over and above these differences, there is a common ground shared by all these approaches. All of them describe collocation as a relation between two parts.

The idea of collocation as a bipartite structure has been called into question by research conducted in the framework of the Lexical Constellation model (hereinafter, LCM), developed by members of the LACELL research group at the University of Murcia (Cantos & Sánchez, 2001; Almela 2011a; Almela et al. 2011a, 2011b). Central to the LCM programme is an attempt to optimise the methods of semantic description used in corpus-based lexicology. In pursuing this goal, previous LCM research has concluded that the established dualism of node and collocate is not suitable for capturing the complexity of collocational relations. The main reason for this is that the strength of attraction between a node and a collocate cannot be established independently of the effects produced by other collocations of the same node. Besides, these effects have implications for the analysis of meaning.

Hence, where the Firthian motto reads: "You shall know a word by the company it keeps", the LCM adds: "Collocations, too, shall be known by the company they keep". In fact, much of the company attributed to individual words in collocational studies is in reality attributable to an interplay of different collocational patterns. While collocational research has mostly been concerned with investigating the effect of a word on neighbouring words, LCM research has directed the attention towards the effects that a collocation produces on neighbouring collocations. For an overview of theoretical foundations of the LCM and potential applications in the field of lexicography, the reader is referred to Almela et al. (2011b).

The main difference between the LCM and other approaches to the phenomenon of interlocking collocation lies in the account of interactions among bi-grams. Like lexical constellations, *collocational networks* (Williams 1998, 2001; Alonso et al. 2008) and *collocation chains* (Alonso Ramos and Wanner 2007) represent forms of co-occurrence patterning where two or more collocations share one of their elements. With respect to these approaches, the specific contribution of the LCM lies in the development of a method for describing dependencies among different collocates of a node. In addition to stating the fact that two or more observed collocations have an element in common, the LCM apparatus is suited to determine whether the presence of one of these collocations increases or diminishes the probability of the other.

Further differences include the fact that collocational networks are conceived to describe the vocabulary of specialised sublanguages, while the LCM has so far been applied to the description of general English. Also worth mentioning is the difference with respect to the theoretical background research into collocation chains. The study of collocation chains carried out by Alonso Ramos & Wanner (2007) is informed by Meaning-Text Theory, whereas the LCM adopts a usage-based approach to language.

The aim of this study is to present a methodological innovation in the LCM. The new version presented in this paper introduces a step consisting in the use of clustering techniques. This step is aimed at guiding the description of the semantic relations underlying the network of *co-collocations* (i.e. collocations that are strengthened or co-activated by the presence of other collocations of the same node).

The paper is formally structured as follows. The next section provides a summary of the LCM methodological framework and explains the new step introduced in this study. Then, section 3 will present the new methodology at work. The collocational patterns analysed include interactions of verbs and premodifiers in the lexical environment of the noun *decision*. Finally, section 4 will discuss the findings and make suggestions for further research.

The Method of Constellational Analysis

At present, the method applied in the LCM consists of three main steps: (*i*) extraction of collocates, (*ii*) identification of inter-collocability relations, (*iii*) semantic grouping and analysis of co-collocates. The first step involves simply an extraction of statistically significant co-occurrences (collocates) of a node word. At this point there is no difference with respect to standard practices of collocation extraction in corpus linguistics. The specificity of constellational analysis is introduced in the second step. It is at this point where the distinction between the categories of *collocate* and *co-collocate* is established.

LCM research has identified so far two main forms of co-collocation, that is, of inter-collocational dependency (Almela et al. 2011b). The first one is *positive inter-collocability*, and the second one is *negative inter-collocability*.

Positive inter-collocability obtains in cases where one collocation 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.

Negative inter-collocability obtains when the collocability of a node and a collocate is restricted 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*. This indicates 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 et al. 2011b). The values compared in this phase of constellational are the following ones:

- 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;
- The probability that the same collocate (*c1*) is selected given as a fact the occurrence of the node: P(*c1*|*n*).

For the sake of brevity, we will refer here to the first value as P1, and to the second one as P2. P1 can be described as a value of conditional probability at the inter-collocational level, and P2 as a value of conditional probability at the intra-collocational level. If P1 is higher than P2, we can say that *c1* is a *positive co-collocate* of *c2* relative to *n* (Almela et al. 2011b). To render the terminology more symmetrical, we can further add that *c2* is a *positive co-node* of *c1* relative to *n*. This formulation expresses the fact that the collocation of *c1* and *n* is made more probable (or strengthened) by the presence of *c2*. An additional requirement for positive inter-collocability is a frequency threshold. If the frequency of the 3-gram (*n,c1,c2*) is lower than 2, the combination is excluded from being a candidate for positive inter-collocability.

Negative inter-collocability obtains when P1 is lower than P2. In this case we can say that *c1* is a negative co-collocate of *c2* relative to *n*, and conversely, that *c2* is a negative co-node 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*).

Like the distinction between node and collocate (see Introduction), the distinction between co-node and co-collocate is purely methodological: it depends on which collocational pattern has been used as input and which one has been obtained as output. If we are investigating the contexts of the pattern *faulty goods*, *return* will be obtained as a positive co-collocate, and *faulty* will be the co-node of *return*, but if we decide to investigate the contextual effects of *return goods*, it may turn out that we also obtain *faulty* as a positive co-collocate of *return*. The same applies to negative inter-collocability—more on this issue in Almela et al. (2011b). Therefore, it is important to emphasise that inter-collocability can be mutual, although it does not have to. It may operate in the two directions, from *c1* to *c2* and vice versa, or only in one direction. From the fact that *c1* is a positive or negative co-collocate of *c2* it does not necessarily follow that *c2* must also be a positive or negative co-collocate of *c1*. The reason for this is that conditional probabilities are directional. The probability of finding a word or expression *a* in the context of another word or expression *b* can be considerably higher or lower than that of finding *b* in the context of *a*.

In earlier versions of the method, the third step in constellational analysis consists in the semantic grouping of co-collocates and the lexical description of the resulting structures. Unlike the previous steps, this task is not automatised and consequently depends more on intuition. Despite this weakness, the step is absolutely necessary to account for a fundamental aspect of constellations, namely, the underlying semantic motivation of inter-collocational dependencies. Previous studies have shown that words from the same conceptual domain will also tend to share a substantial amount of their co-collocates. This suggests that lexical constellations may well be interpreted as representing specific surface realisations of more abstract semantic structures (Almela et al. 2011a).

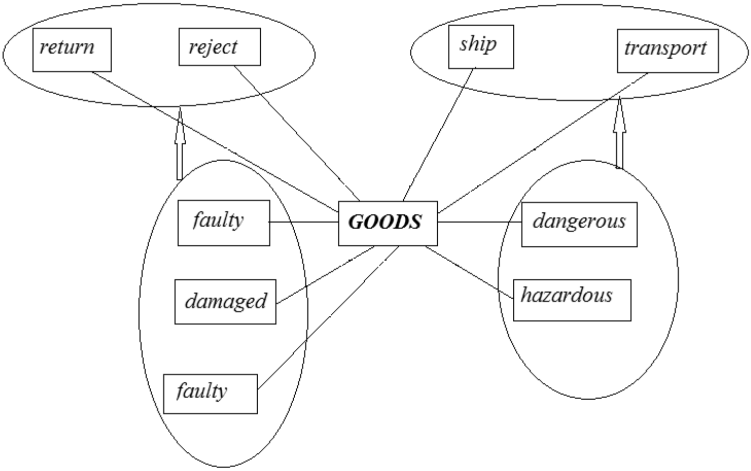

Figure 1. Some lexical constellations of *goods*

For example, verbal collocates of *goods* that express a process of TRANSPORTATION tend to function as positive co-collocates of adjectives expressing DANGER. Meanwhile, adjectival collocates of *goods* that describe some kind of FLAW in a product tend to function as co-nodes of verbs that describe a decision of NON-ACCEPTANCE of the goods. These two patterns of co-collocation are graphically represented in Figure 1. The node is emphasised in bold type and capital letters, and each line stands for a relationship of statistically significant co-occurrence between the node and a collocate. The arrows represent relations of positive inter-collocability, and the circles have been used to group together words that are related within the same conceptual domain. Thus, following the steps indicated in Almela et al. (2011a), the lexical constellations displayed in Figure 1 can be turned to lexico-conceptual constellations, as in Figure 2, where the link of positive inter-collocability holds between semantic groups, and not just between individual items.

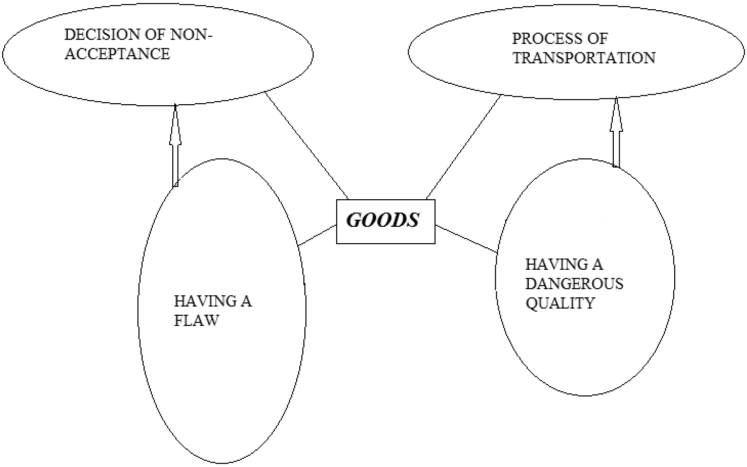
DECISION OF NON-ACCEPTANCE
PROCESS OF TRANSPORTATION
GOODS
HAVING A FLAW
HAVING A DANGEROUS QUALITY

Figure 2. Some lexico-conceptual constellations of *goods*

The semantic motivation of inter-collocability paves the way to interesting advances in the description of lexical structure. One of the reasons why collocational studies have proven so fruitful to lexicology is their capacity for identifying correlations of distributional classes and semantic classes. For example, we know that nouns occurring as direct objects of the verb *face* tend to share an aspect of their meaning. The list includes *difficulty, challenge, crisis, dilemma, threat, problem*, etc. This set of nouns constitutes both a distributional class and a semantic class. It is a distributional class because they share a distributional feature: they are statistically significant co-occurrences of the same item (*face*) and in the same slot (direct object). At the same time, they represent a semantic class, because they form a semantically coherent set. They all refer to concepts related to DIFFICULTY.

To a large extent, the potential of collocational data for providing semantically relevant information is owed to this alignment of distributional classes and meaning groups. However, the standard techniques of collocational description had allowed us to detect such correlations only at an elementary, coarse-grained level of analysis. It was possible to identify semantic sets of collocates, but not semantic sets of co-collocates. In this respect, constellational analysis supersedes the traditional methods of collocation-based analysis of meaning. The technique applied in the LCM makes it possible to identify correlations of distributional and semantic classes at more than one level (i.e. not only between nodes and collocates, but also between co-nodes and co-collocates).

For example, the traditional techniques of corpus collocational analysis determine that *tough* and *difficult* are collocates of *decision*. In the ukWaC corpus (at SketchEngine) these words form statistically significant pairs with *decision* within a 4:4 collocational span. The two pairs form statistically significant co-occurrences with three different measures (logDice, MI, T-score). What these techniques fail to tell us is that the presence of *tough* and *difficult* in the context of *decision* is motivated not by *decision* alone, but by interactions of the collocations *tough/difficult decision* with other collocational patterns of *decision*. While there are verbs that avoid co-collocations with *tough/difficult decision* –for example, in the ukWaC the frequency of occurrence of *review*, *reconsider* or *reverse* in this context is zero– there other verbal collocates of *decision* that are strongly associated with this pattern of inter-collocability. Thus, the verb *face* is a prominent co-node of *tough* and *difficult*. The probability that *tough* is selected as premodifier of *decision* is increased by almost 40 times if the collocation converges with *face*. A similar difference is obtained in the case of *difficult*.

P(*tough*|*face,decision*) = 38/304 = 0.125
P(*tough*|*decision*) = 962/302679 = 0.003
P(difficult|face,decision) = 80/304 = 0.263
P(*difficult*|*decision*) = 2034/302679 = 0.007

In the light of these data, it would be inaccurate to say that *tough* is a collocate of *decision* without specifying that it is also a co-collocate of *face*, because the probability that *tough* is selected as a premodifier of *decision* increases with the presence of the verb *face* but decreases with the presence of other verbal collocates of *decision*, such as *review*, *reconsider*, or *reverse*, inter alia.

Ultimately, the method of lexical-constellation analysis arises from a revision of the concept of *lexical gravity*. Mason (2000: 270) defined lexical gravity as “the restriction a word imposes on the variability of its context", but as Cantos & Sánchez (2000) remarked, the trouble with this concept is that, contrary to what collocational studies have often assumed, the node does not exert an unlimited influence on its environment. The restrictions imposed on the variability of the context of a word are shaped by an interplay of several factors, including the attraction between different collocational pairs (and not just between words). What is more important for lexical studies is that the patterns of interaction among collocates of a node are usually motivated by an underlying semantic structure.

To continue with the example above, the fact that *tough* and *difficult* are positive co-collocates of *face* is not unrelated to the fact that *face* (again as a verb) collocates with nouns that share a same semantic feature (DIFFICULTY) with the adjectives *tough* and *difficult* (e.g. *face* + *difficulty, challenge, crisis, dilemma, threat, problem*, etc.).

At present, a major weakness of the LCM method for grouping co-nodes and co-collocates into semantic sets is that it relies mainly on the analyst's intuition. The proposal we submit in this paper is intended to minimise the subjective component in this task and strengthen the commitment of the model to an objective methodology. One way to achieve this goal is to base the semantic grouping of co-nodes on classifications obtained from hierarchical cluster analysis. The amalgamation of co-nodes according to clustering techniques can be used as a guide to the manual amalgamation of co-collocates.

A Case Study: Lexical Constellations of Decision

In this section the analytical framework sketched out above is applied to the description of lexical constellations formed around the noun *decision*. The methodological decisions adopted in this case study are similar to those applied in the analysis of lexical constellations of *goods* in earlier research (Almela 2011a), except for the fact the present study introduces the use of cluster analysis as previous step to the semantic analysis of co-nodes and co-collocates.

The data and the examples have been extracted from the *ukWaC* corpus (1,565,274,190 tokens), accessible at the SketchEngine query system. The analysis is focused on capturing features of inter-collocability in verb + noun and premodifier + noun collocations. Previous research has proven successful in detecting cases of positive inter-collocability within this grammatical framework (Almela 2011b; Almela et al. 2011a). Therefore, 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 premodifier + *decision* performs the same semantic role). The WordSketch function has been useful in limiting our queries to the foregoing grammatical scheme. Nevertheless, manual supervision was required in order to detect possible parsing errors.

As instructed in the previous section, the method of constellational analysis applied in this study consists of four steps. Each of these steps is described in detail in the subsections below (3.1 to 3.4).

First Step: Collocation Extraction

The first step was to extract a list of verbal and modifier collocates of *decision*. Statistical significance was defined in terms of logDice − for an explanation of the advantages of this measure, cf. Rychlý (2008). The number of verbal collocates was limited to 25 due to limitations of space, given that a dendrogram containing more than 25 variables –in this case, the variables are verbs– would not fit in one page (see Figures 4 and 5 in the next subsection). Table 1 shows the list of 25 top verbal collocates of *decision* arranged in order of decreasing logDice score. The list of adjectives was allowed to be larger because a higher number of cases does not affect the size of the dendrogram but helps to make the analysis more accurate. Therefore, the list of premodifiers was extended to include 50 items (Table 2). Again, the collocates in the table have been arranged in order of decreasing logDice score.

1. *make*	6. *reverse*	11. *challenge*	16. *await*	21. *confirm*
2. *reach*	7. *appeal*	12. *announce*	17. *uphold*	22. *defend*
3. *influence*	8. *reconsider*	13. *review*	18. *regret*	23. *affect*
4. *inform*	9. *justify*	14. *welcome*	19. *implement*	24. *issue*
5. *take*	10. *overturn*	15. *defer*	20. *delay*	25. *explain*

Table 1. Top collocates of *goods* (grammatical relation: “object_of”)

1. *informed*	11. *recent*	21. *rational*	31. *quick*	41. *big*
2. *final*	12. *key*	22. *initial*	32. *correct*	42. *bad*
3. *tough*	13. *planning*	23. *wrong*	33. *buying*	43. *hard*
4. *conscious*	14. *wise*	24. *political*	34. *ultimate*	44. *tribunal*
5. *right*	15. *important*	25. *crucial*	35. *deliberate*	45. *subsequent*
6. *strategic*	16. *judicial*	26. *sensible*	36. *formal*	46. *tactical*
7. *unanimous*	17. *controversial*	27. *clinical*	37. *ethical*	47. *momentous*
8. *difficult*	18. *own*	28. *collective*	38. *future*	48. *early*
9. *purchasing*	19. *policy*	29. *court*	39. *brave*	49. *funding*
10. *investment*	20. *original*	30. *major*	40. *executive*	50. *majority*

Table 2. Top collocates of *goods* (grammatical relation: "modifies")

Second Step: Identification of Co-Collocates

In a second step, the values of conditional probabilities were calculated and compared following the procedure explained in section 2. Table 3 shows a sample of the data used for identifying co-collocates of one of the verbs, *overturn*. The first column is a list of potential co-collocates. The next column indicates the raw frequency of the whole combination (verb, premodifier, noun) in the corpus (for instance, the frequency of *overturn an earlier decision*). The third column provides the raw frequency data of the collocational pair formed by the noun (*decision*) and each of the collocates listed in the left-most column (for instance, the frequency of the collocation *original decision* in the corpus is 721).

	f(v,m,n)	f(m,n)	P(m|v,n)	P(m|n)
original	23	721	5.00%	0.24%
early	5	669	2.83%	0.22%
tribunal	13	189	1.09%	0.06%
court	2	653	1.09%	0.22%
initial	4	477	0.87%	0.16%
executive	5	207	0.43%	0.07%

Table 3. Frequency and probability data for calculating co-collocates of *overturn*

Along with the frequency of the noun, which is not shown in the table because it remains constant (302679) in all the rows, the frequency data in the second and in the third columns are necessary in order to calculate the values of conditional probabilities in the remaining columns. The fourth column returns the value of P(m|v,n), that is, the probability that the premodifier occurs given as a fact the occurrence of the verb + noun collocation; the second one specifies the probability that the premodifier occurs given the selection of the noun, which is formally expressed as P(m|n). Hence, these two columns return the values labelled as P1 and P2 in section 2.

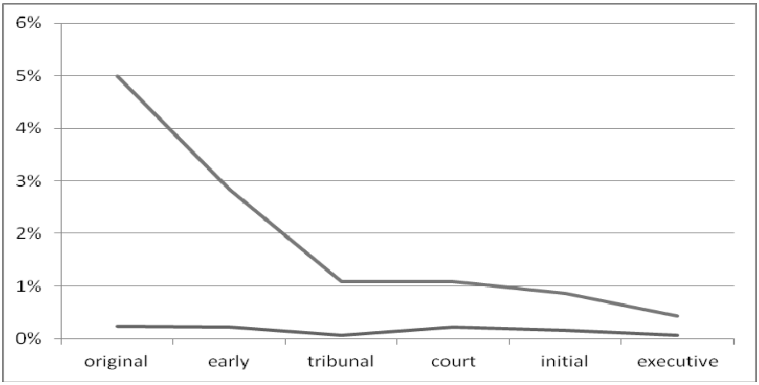

Figure 3. Comparing intra- and inter-collocational conditional probabilities from Table 3

The order of the rows is determined by the difference between the values of the last two columns. Thus, the premodifier at the top of the list is the best candidate for positive co-collocation. *Original* is the premodifier with the highest difference between the value of P1 (inter-collocational conditional probability) and of P2 (intra-collocational conditional probability). The difference between these two values is graphically represented in Figure 3. The top line represents P1, and the bottom line represents P2. The figure shows that all the premodifiers analysed in this table are positive co-collocates of *overturn* relative to the noun *decision*, and that *original* is the most prominent one.

Third step: cluster analysis

The third step in the new version of the model is cluster analysis. Using SPSS 19 and selecting square Euclidean distance as a measure, the verbs were grouped into clusters. In the settings, the verbs were entered as variables, and the premodifiers as cases. The reason for not doing the opposite can be explained with reference to figures 4 and 5. The dendrogram in Figure 4 shows the results from clustering the 25 top logDice premodifiers of *decision* according to their distribution in collocations of *decision* with its top 50 logDice verbs (always within the syntactic pattern specified at the beginning of this section). This clustering is much less discriminatory than the one displayed in the dendrogram of Figure 5, which shows the clustering of 25 top logDice verbal collocates of *decision* according to their distribution in collocations of *decision* with its top 50 logDice premodifiers. This is the main reason why in this study it was decided to treat the verbal collocates as potential co-nodes and the premodifiers as potential co-collocates, rather than the other way round.

The results displayed in the dendrogram of Figure 5 suggest the existence of three clusters of verbs according to their distribution in lexicogrammatical contexts of the type 'verb + premodifier + *decision*':

- First cluster: {overturn uphold confirm reverse reconsider review defend regret appeal challenge welcome}
- Second cluster: {influence inform affect justify implement}
- Third cluster: {defer delay announce await reach issue take explain make}

The first cluster is semantically more homogeneous than the other two. All the verbs in the first cluster express some aspect of a REACTION to a decision. In three verbs from the cluster, the reaction is described in terms of an axiological value, i.e. a judgment of APPROVAL or DISAPPROVAL: *defend, challenge, welcome*. In the remaining verbs from the cluster, the reaction described implies a process of REVISION. This subframe is organised around a temporal sequence. One of the verbs (*appeal*) refers to the initial stages, when the revision is demanded but still not undertaken, let alone completed. Two other verbs, *reconsider* and *review*, denote the process of undertaking the revision itself. Finally, the event denoted by four other verbs implies that the revision has already been completed. Different verbs denote different outcomes of the process of revision: *confirm* and *uphold* express an action by which the earlier decision is affirmed, while *reverse* and *overturn* express an action by which the earlier decision is altered or cancelled.

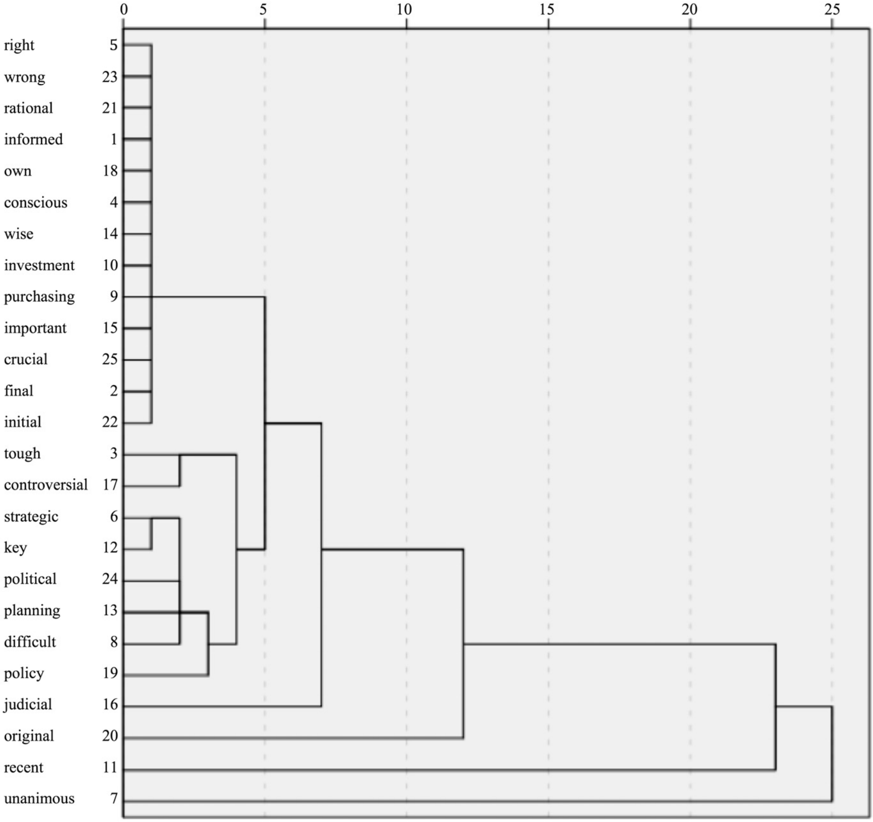

Figure 4. Cluster analysis of 25 premodifiers

More differences among the verbs from this cluster arise from their degree of contextual specialisation. *Uphold, overturn* and *appeal* are highly characteristic of the legal jargon, while the remaining verbs are more versatile in this respect. However, these are minor differences which do not alter the conclusion that the verbs from this cluster form a consistent semantic set structured around the central notion of REACTION to a prior decision. Therefore, there are good reasons to predict that the verbs from this cluster will tend to co-activate collocations describing decisions made in the past. The answer to this prediction will be given in 3.4.

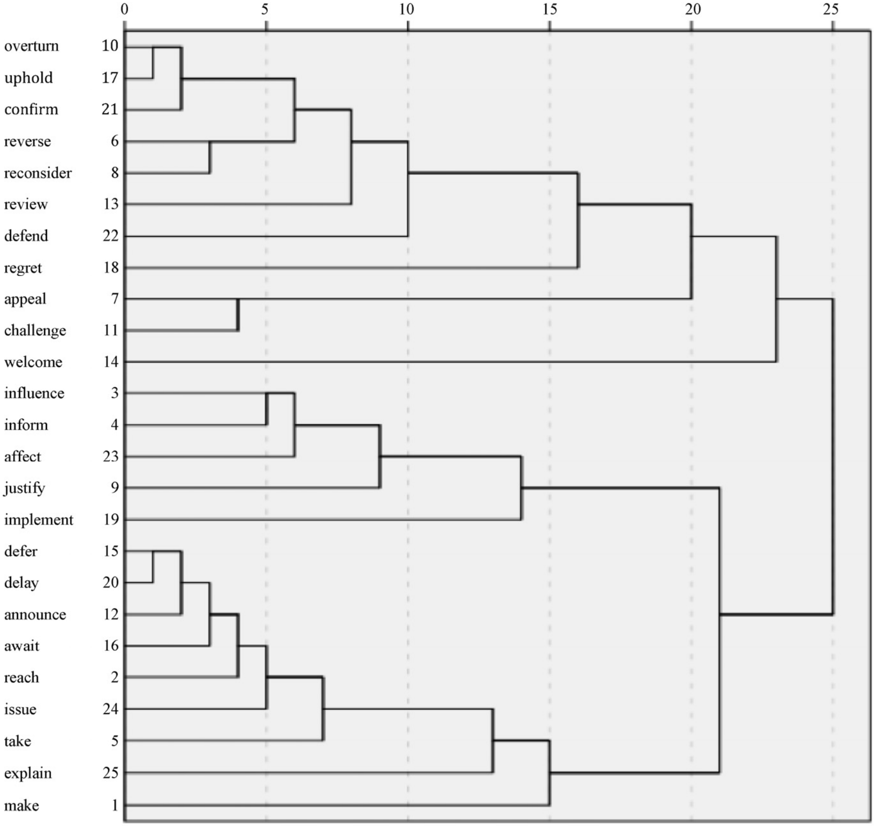

Figure 5. Cluster analysis of 25 verbal co-nodes

The second cluster is smaller, as it contains only five verbs. Three of these verbs have clearly related meanings: *influence, inform* and *affect* express ways of exerting some form of INFLUENCE on a decision. Thus, INFLUENCE can be deemed to constitute the central conceptual domain of the cluster. Nevertheless, it has to be conceded that the other two verbs, *justify* and *implement*, are less directly connected to this domain. Considering that the domain INFLUENCE conveys a semantic dimension of PROSPECT (i.e. projection of possible events or state of affairs in the future), it is reasonable to expect that verbs from this cluster will tend to function as co-nodes of premodifiers referring to FUTURE EVENTS. The answer will be given in section 3.4, when the whole process of constellational analysis is completed.

Lastly, the third cluster is less homogeneous than the other two. It contains two verbs of COMMUNICATION (*announce, issue*) and two verbs expressing POSTPONEMENT (*defer, delay*), plus three support verbs (*make, take, reach*) that as such are not associated with any specific conceptual domain. An indication of their status as support verbs in this context is the fact that the collocations *take/make/reach a decision* are equivalent to the simple verb *decide*, as far as denotation is concerned. The semantic relation expressed by these verbs corresponds to the lexical function Oper$_1$ in Meaning-Text Theory (see Melčuk 1998). One member of the cluster, *await*, cannot be pigeonholed into any of the aforementioned categories.

Fourth step: co-node and co-collocate grouping

In the fourth and last step the clusters obtained from the previous step are used as guidelines for the classification and description of inter-collocability relations. Put differently, the amalgamation of co-nodes according to clustering techniques in the third step is used now as a template for the manual amalgamation of co-collocates.

Following this procedure, the clusters of verbs obtained from the third step should be used now as a starting point for merging and dividing premodifiers into groups. The expectation was that co-nodes (in this case: verbs) that are grouped together in the same cluster would tend to select similar groups of co-collocates (in this case: premodifiers). If distributional classes are closely aligned with semantic classes, as is normally the case, we can predict that verbs from the same cluster will tend to activate premodifiers from related conceptual domains. Although the combination of verbs and premodifiers does not form part of any established grammatical typology of collocation, previous LCM research has shown that these two categories can interact and form semantic patterns (Almela 2011b; Almela et al. 2011b). Therefore, the classification of premodifiers according to clusters of verbs should lead, in principle, to coherent results.

Table 4 contains the lists of positive co-collocates of verbs amalgamated in the first cluster (negative inter-collocability has not been taken into consideration in this empirical study). Within each list the premodifiers have been arranged in order of decreasing prominence, that is, of decreasing difference between the values of inter-collocational and intra-collocational conditional probabilities (the same criterion has been applied in tables 5 and 6 for verbs from the second and the third cluster).

The results displayed in Table 4 are highly consistent and conform to the initial expectations that cluster analysis is an adequate starting point for grouping co-collocates under specific groups of co-nodes. Thus, the first pair of verbs to be joined together by a cluster, *overturn* and *uphold*, are also those that share exactly the same pattern of positive co-collocation. The lists of premodifiers associated with these verbs are identical. Even the order in which the premodifiers are arranged is exactly the same in both lists.

verbs (co-nodes)	premodifiers (co-collocates)
overturn	*original, early, tribunal, court, initial, executive, controversial*
uphold	*original, early, tribunal, court, initial, executive, controversial*
confirm	*original, early, controversial, funding, initial*
reverse	*early, original, court, initial, controversial, funding*
reconsider	*early, original*
review	*original, recent, own, early, initial, controversial, tribunal, formal, funding*
defend	*original, controversial, difficult*
regret	*early*
appeal	*court, tribunal, planning, original, bad, early*
challenge	*tribunal, court, planning, executive, bad, original*
welcome	*recent, tribunal, court, unanimous*

Table 4. Co-collocates associated with verbs from the first cluster

More generally, we can observe that the most repeated premodifiers in this cluster are related to two main conceptual domains. The first one includes premodifiers that are used to describe PAST EVENTS: *original, early, initial, recent*. All the verbs from this cluster are associated with at least one of the premodifiers from this list, and many verbs are associated with three of them. This is the case of *overturn, uphold, confirm,* and *reverse*, all of which exhibit positive co-collocation with *early, original* and *initial*. Notice also that *review* is associated with the four premodifiers from this group.

The second conceptual domain that is commonly found –though to a lesser extent– among co-collocates associated with the first cluster is structured around the notion LEGAL SYSTEM, more specifically in connection with the setting COURT OF LAW: *tribunal* and *court* (used as premodifiers) are positive co-collocates of five verbs from this cluster (*overturn, uphold, appeal, challenge, welcome*). Besides, *court* occurs in the list for *reverse*, and *tribunal* in the list for *review*.

verbs (co-nodes)	premodifiers (co-collocates)
influence	*policy, future, political, purchasing, investment, buying, subsequent, planning, strategic, key, major, clinical, judicial, funding, initial*
inform	*future, policy, investment, purchasing, funding, clinical, planning, buying, strategic, subsequent, ethical*
affect	*investment, purchasing, future, policy, planning, buying, funding, major*
justify	*political, funding, policy, investment, clinical*
implement	*policy, strategic, tribunal, key, court*

Table 5. Co-collocates associated with verbs from the second cluster

The discrimination between the first and the second cluster of verbs is convincing. By comparing tables 4 and 5 we can arrive at the following conclusions: firstly, the two clusters are highly coherent internally; and secondly, there is a noticeable semantic contrast between them, which speaks for the discriminatory power of the technique applied. While the prevailing conceptual domain in co-collocates associated with the first cluster is PAST EVENT, the dominant domain in the second cluster is exactly the opposite, that is, FUTURE EVENT. All the verbs from this cluster are co-nodes of premodifiers that in some or other way describe an action of planning or a projection of future events. This is obvious in the case of premodifiers such as *future, planning*, and *investment*. The same semantic dimension is less obvious though still present in premodifiers such as *strategic, policy, political*. These three words relate to an action of planning, and the notion PLANNING in turn contains a reference to FUTURE EVENTS. In a different sense, the word *subsequent* is also related to the notion of FUTURE EVENT, because it describes a temporal sequence.

Another conceptual domain that is repeated in the second cluster is the TRANSACTION frame. *Buying, purchasing* and *investment* (used as premodifiers) are perfect examples of this category – note that *investment* is related simultaneously to the two dominant conceptual domains in the cluster, since it combines the semantic features PLANNING and TRANSACTION. *Funding* is also related to this domain because, after all, it also makes reference to a transfer of capital.

verbs (co-nodes)	premodifiers (co-collocates)
defer	*final, funding, investment, hard, big*
delay	*final, planning, funding, purchasing, policy, key*
announce	*final, funding, majority, formal, unanimous, major*
await	*final, initial, court*
reach	*final, own, unanimous, informed, correct, collective, planning, majority, sensible, formal, initial, rational*
issue	*final, formal, recent*
take	*tough, final, strategic, difficult, major, brave, key, policy, hard, deliberate, important, political, formal, executive, conscious, momentous, crucial, sensible, early, collective, ultimate, right, controversial, tactical*
explain	(none)
make	*informed, right, own, final, conscious, important, wrong, purchasing, quick, correct, difficult, investment, rational, strategic, sensible, wise, bad, crucial, key, major, ethical, tough, buying, momentous, executive, big, clinical, initial, tactical, deliberate, collective, ultimate*

Table 6. Co-collocates associated with verbs from the third cluster

Unfortunately, the third cluster is slightly more heterogeneous than the other two (see Table 6). Besides, the discrimination between this cluster and the previous ones is less convincing than that between the first and the second one. Put briefly, the problem is that the third cluster contains many items that were found to characterise elements from the first cluster or from the second cluster (e.g. *early, initial, planning, funding, investment, policy, political*).

In part, this can be explained by the fact that the third cluster contains three ‘support’ verbs: *take*, *make* and *reach*. Support verbs are largely delexicalised and semantically underspecified. This also implies that their collocational behaviour is highly versatile – that is, they occur in a wide range of lexical contexts. This remark concerns their relations not only to collocates but also to co-collocates. Compared to the other verbs in the cluster (the purely ‘lexical’ verbs), the lists of positive co-collocates attributed to the support verbs –particularly to *make* and *take*, which are more frequent than *reach*– is significantly larger and more heterogeneous.