Collocation and textual cohesion:

A comparative corpus study between a genre of Written Sports Reports and a large reference corpus

By

Brett Edward Laybutt

A dissertation submitted to

The School of Humanities of the University of Birmingham

In part fulfilment of the requirements for the degree of

Master of Arts in TESL/TESL

Supervisor: Tilly Harrison

This dissertation consists of approximately 12,203 words

Centre for English Language Studies
Department of English
University of Birmingham
Edgbaston
Birmingham B15 2TT
England

20th September, 2009

To my wife

Abstract

One of the most problematic areas for foreign language learning is collocation. It is often seen as arbitrary and overwhelming, a seemingly insurmountable obstacle to the attainment of native-like fluency. The following study takes an approach suggested by recent corpus research by investigating the functional role of collocation for cohesion within a genre-specific corpus of written sports reports (WSR). Through a comparison with a large reference corpus, the study found that certain key collocations contributed to cohesion both within individual texts, or what will be termed `intratextual` cohesion, and also across texts within the genre, or `intertextual` cohesion. It was also found that many of these collocations are the result of underlying metaphors. The study suggests that, for foreign language learners, focusing on this functional role of collocation within genre may provide a more systematic and manageable technique for the study of collocation. It also goes on to suggest the need for a distinction between 'teaching genre', suitable for ESP or EAP classes, and 'genre teaching', which encourages learners to view language not as rule-generated but as a system of choice within differing stratum of text, context and genre.

Acknowledgements

First of all, I would to thank my supervisors over the course of this MA: Tilly Harrison, for all her insightful comments, suggestions and keeping me focused in writing this dissertation, and also Wendy Mah for all her advice and support during the coursework. I'd also like to thank Tony Cole for his statistical discussions. Lastly, I'd also like to thank my parents, without whom none of this would have been possible.

Table of Contents

List of Figures and Tables

Figures

Tables

Notation within the dissertation

As there is no generally agreed upon system of notation (Gledhill, 2000; Bartsch, 2004) for the study of collocation, this dissertation will use the following notation:

• ‘ball’	- a word-form
• ball	- a word-form as node
• <ball home>	- a single collocation in the GCL
• <ball + net>	- a collocation with up to two intervening words in the GCL
• *ball home*	- a single collocation in the BofE
• *ball* + *net*	- a collocation with up to two intervening words in the BofE
• span = 3:3	- a span of three words either side of the node
• span = 1:0	- a span of one word to the left of the node
• LL	- log-likelihood score
• N	- total number of occurrences of node
• n	- total number of occurrences of node + collocate
• [Event]	- a genre Move

Chapter 1 - Introduction

1.1 Background to the study

For learners of any language, the difficulty that arises from the fact that some word combinations are acceptable and some not, or 'collocational competence' (Hill, 1999), has long been recognised. As far back as the 1930s, Palmer (1933, cited in Nation, 2001, p. 317) discussed its importance and the fact that each must be "learnt as an integral whole". However, the concept of collocation has been largely neglected within mainstream linguistics (Bartsch, 2004) and it was not until the 1990s with the rise in computing power, and acceptance, of corpus linguistics that it received serious treatment. Recent corpus studies of collocation have focused in particular on the contribution of collocation for the creation and maintenance of representativity within genre (Williams, 2002) and the textual and discourse function of collocation for that genre (Gledhill, 2000). The role of collocation for the creation of cohesion within text and genre has, however, generally received scant treatment. Within English as a Foreign Language (EFL) research as well collocation is often dismissed as not playing any significant role in creating cohesion (McGee, 2009), or as too problematic for classroom treatment (Mahlberg, 2006).

1.2 Aims of the study

Most previous studies of collocation and genre have focused on restricted academic genres. The following study will instead conduct an investigation of collocation within a genre aimed at a more general audience, that of written sports reports (WSR). Formalised sport arose within a specifically British context to become a "recognisable social and cultural institution" (Rowe, 1999) and forms a readily recognisable 'discourse genre' (Ferguson, 1983; Ghadessy, 1988). Interest in British culture, of which sport constitutes a major aspect, is also an oft-stated reason for the study of English in an EFL context. Specifically within this author's own Japanese context, studies have shown that university students often prefer such integrative and personal motivations over instrumental ones (Benson, 1991). Also in a wider context, the British Council (http://www.britishcouncil.org/sport-premier-skills-home.htm), for example, is actively using football to promote English language learning. The investigation will be conducted from a systemic functional linguistics (hereafter SFL) perspective with two main objectives. The primary aim of the study will be to identify significant patterns of collocation within a specialised WSR corpus compared to a large corpus of general language. The secondary aim will then be to shed light on the possible textual role of this collocation for creating cohesion within the corpus.

1.3 Structure of the dissertation

The dissertation is divided into four central chapters. First, Chapter 2 will provide some background to the notion of collocation within systemic functional linguistics and its possible role in the creation of cohesion. This will be followed by an overview of corpus linguistics and its role in investigating collocation, cohesion and genre, followed by a brief overview of SFL notions of genre and a reading of a sample of sports reports establishing it as a possible genre. Next, Chapters 3 and 4 will go on to present the methodology for the analysis of collocation within a specialised corpus of Guardian newspaper sports reports (GCL) and the results of that study respectively. Finally, Chapter 5 will discuss the functional role for the creation of textual cohesion that the collocation identified in the study plays within the corpus and the possible implications of this for foreign language teaching.

Chapter 2 - Literature Review

2.1 Introduction

The following sections provide some background to the concept of collocation that will be used in this dissertation, with a SFL approach that seeks to investigate the role of collocation for cohesion within genre, in this case written sports reports (WSR). First, Section 2.2 provides a working definition of collocation and its relation to SFL theory, as well as its treatment within applied linguistics and importance for foreign language learning. Section 2.3 then gives an outline of corpus linguistics and corpus techniques for investigating collocation. Lastly, Section 2.4 outlines the SFL theory of genre, followed by a brief genre analysis of WSR.

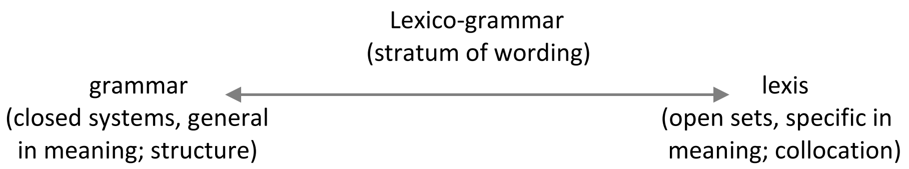
2.2 Collocation

The following section provides some background to the notion of collocation, its treatment within systemic functional linguistics and a working definition to be used within this paper. The term ‘collocation’ has been used in varied ways by different writers in different contexts. As such, there is no common, agreed upon, definition of the term (Bartsch, 2004). Rather than get lost in competing definitions, the following section outlines the approach to be taken here and presents a working definition that will be used for the later corpus analysis. This is followed by a brief discussion of the significance of collocation for users of a language and the particular problems faced by learners trying to acquire competence in a second or foreign language.

2.2.1 Background to collocation and text

Collocation was first used in a technical sense by the British linguist J.R. Firth (1957), who proposed that the meaning of a word is at least partly determined by its contextual environment, or “meaning by collocation” (Firth, 1957, p. 194). The main insight of this approach is that it questions the idea of separable word classes (Gledhill, 2000; Sinclair, 1991) and instead conceives of language as a cline from closed class grammatical words at one end to open class lexis at the other (Halliday, 2005; Halliday, 2004), as shown in Figure 1:

Figure 1: Lexico-grammar cline

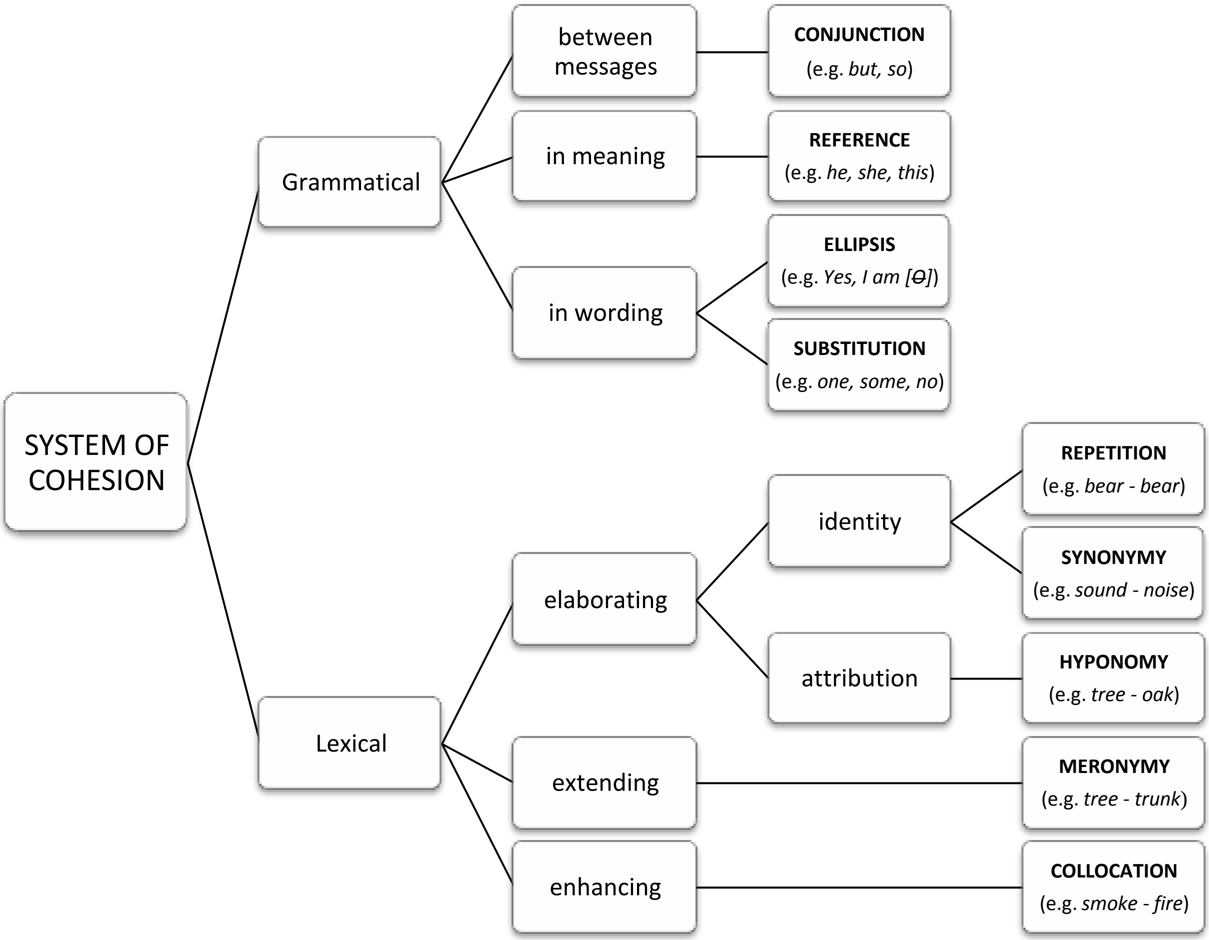

Lexico-grammar
(stratum of wording)
grammar
(closed systems, general in meaning; structure)
lexis
(open sets, specific in meaning; collocation)

(Halliday, 2004, p. 43)

Unlike the formalist approach of an internal rule-based system, language here is instead seen as a resource of choices to express meaning (de Beaugrande, 1996; Martin, 1992) and collocation in this system is at the lexical end of the cline. The notion of collocation also indicates that, rather being constrained by syntactic forces alone, there are paradigmatic and textual influences on lexical choice, which accounts for the fact that, for example, *strong tea* and *powerful car* are acceptable but **powerful tea* or **strong car* are not (Halliday, 2005). This concept led Sinclair (1991, p. 110) to propose a similar lexical cline, from the 'idiom principle' whereby "the language user has available…a large number of semi-preconstructed phrases that constitute single choices" to the 'open choice principle' where each word constitutes a "separate choice" (Sinclair, 1991, p. 175).

However, while collocation and its influence on linguistic choice may be readily observed, its precise role within text remains unclear. For SFL, a text can be defined as "a unit of language in use" (Halliday & Hasan, 1976, p. 1) and is distinguished from non-text by the two-fold concept of unity: unity of structure and unity of texture (Halliday & Hasan, 1985), also generally termed coherence and cohesion (Carter, 1998). The first of these will be discussed further in Section 2.4, but the second, cohesion, is concerned with how the text ties together internally in terms of its cohesive relations and is formed when one element of a text is dependent for its interpretation on another (Halliday & Hasan, 1976; Leckie-Tarry, 1995). Without it the surface features of a text may not relate to each other (Carter, 1998, p. 103) and it is thus central to the way in which text is produced and comprehended (Mahlberg, 2006). This is often divided into grammatical and lexical cohesion, outlined in Figure 2.

Figure 2: System of cohesion in English[1]


SYSTEM OF COHESION
Grammatical
Lexical
between messages
in meaning
in wording
elaborating
extending
enhancing
identity
attribution
CONJUNCTION (e.g. *but, so*)
REFERENCE (e.g. *he, she, this*)
ELLIPSIS (e.g. *Yes, I am [Ɵ]*)
SUBSTITUTION (e.g. *one, some, no*)
REPETITION (e.g. *bear - bear*)
SYNONYMY (e.g. *sound - noise*)
HYPONOMY (e.g. *tree - oak*)
MERONYMY (e.g. *tree - trunk*)
COLLOCATION (e.g. *smoke - fire*)

(Adapted from Halliday & Hasan, 1976; Halliday, 2004)

Collocation is thus seen as part of this system of cohesion to highlight semantic relations (Halliday, 2004). Yet its exact function has always remained "problematic" (Halliday & Hasan, 1976, p. 284) and, within SFL theory, has been largely "factored out" (Martin, 2002) as contributing little towards texture and the overall cohesiveness of text (Halliday, 2004).

[1] Although it has been further developed and refined (Mahlberg, 2006), this basic model is the one generally used within EFL and so will serve as the basis for discussion here.

Recently however, renewed attention has focused on collocation within the system of semiotic choice (Plum, 2006), described in more detail in Section 2.4, and its possible role in the description of recurrent patterns that may characterise particular genres (Williams, 2002; Gledhill, 2000). As Halliday (2004, p. 258) himself points out, the same items may "often appear in different collocations according to text variety". In particular, as will be outlined further in Section 2.3, corpus research has begun to highlight the "textual function" (Gledhill, 2000, p. 116) of collocation and how this is affected by, and varies between, differing genres. However, a precise definition of the term 'collocation' still remains "notoriously difficult" (Bartsch, 2004, p. 65). Even the term 'collocation' itself is not entirely fixed, with Wray (2000), for example, listing some 50 different, largely synonymous, terms used within the literature. The next section provides a working definition of collocation to be used here.

2.2.2 Definition of collocation

The definition of collocation still remains the subject of some debate. Nation (2001), for example, lists some ten different criteria for classifying collocation. In broad terms however, there are two main approaches, position and frequency (Nesselhauf, 2003). SFL theory generally takes the second approach (Plum, 2006) and, influenced by Firth's (1957) original, somewhat fuzzy, conception of collocation as "mutual expectancies", defines collocation in its most elemental form as "lexical items that regularly co-occur" (Halliday & Hasan, 1976, p. 284) or, in Sinclair's (1991, p. 71) terms, "a tendency for words to occur together" and identifiable by frequency of occurrence. More recently within broader research on collocation, however, it has become apparent that simple frequency of occurrence is not sufficient to fully account for the composition of collocation (Nation, 2001). For example, while two items may occur together frequently, for instance ‘of the’, this is not to say they form a significant collocation. As such, a more integrated approach is required. This study will thus define collocation, adapted from Bartsch (2004), as significantly frequent combinations of two words, one of which is lexical, in a direct syntactic relationship.

2.2.3 Collocation and EFL

As mentioned in Section 1.1, collocation, and its importance for learners of a language, has long been recognised within the area of applied linguistics. Lexical relations in general have also been viewed with increasing interest for their discourse function (Martin, 1992), especially within corpus linguistics, as will be discussed further in Section 2.3. According to Nation (2001), for example, vocabulary choices can reveal information about the the communicative messages of texts. For learners of a language, however, collocation can be the source of particular difficulties in attaining native-level competence due to the seemingly arbitrary nature of the word combinations and the sheer number of collocations present in the language (Nesselhauf, 2003).

It is unfortunately not the case, however, that collocation has found a consistent role within either EFL pedagogy or methodology. Although there have been some attempts, notably Nattinger & DeCarrico’s (1992) lexical phrases, Willis’ (1990) Lexical Syllabus, or more recently the continuing work on Lewis’ (2001) lexical approach and Data Driven Learning (Johns T. , 1991; Gavioli, 2005), these have not found widespread acceptance in the wider EFL context (Richards & Rodgers, 2001, p. 138). In terms of classroom application, there is largely still no consistent methodology beyond mere consciousness raising, and its treatment in published pedagogic materials remains patchy at best (Koprowski, 2005). The main problem for pedagogy is that collocation is generally viewed as somewhat overwhelming and lacking in any systematic treatment (Bartsch, 2004; Nesselhauf, 2003). Suggestions have included selection by frequencies (Shin & Nation, 2007) or range (Lewis, 2001) through to L1-L2 contrastive approaches (Bahns, 1993; Shirato & Stapleton, 2007) or focusing on verb forms (Nesselhauf, 2003). The next section will now describe the investigation of collocation through corpus linguistics.

2.3 Investigating collocation

This dissertation is a corpus study of collocation, as defined above. The following sections outline corpus linguistics and some of the key considerations in corpus design and construction that will later be used in this study. It will also describe recent corpus research on collocation and cohesion.

2.3.1 Corpus linguistics

A corpus can generally be defined as "a collection of texts in an electronic database" (Kennedy, 1998, p. 3) compiled for the purposes of linguistic analysis (Hunston, 2002). From small beginnings in the 1950s, when it was generally viewed as having "nothing to contribute" (Halliday, 1991), corpus linguistics has evolved to become an integral facet of the description of language (Rundell, 2008). McEnery and Wilson (2001), for example, discuss some fourteen different areas of language study in which corpora have been used, from the patterning of individual words to pragmatic and discourse analyses of text. In defining corpora, a broad distinction also is often made between 'general' corpora on the one hand and 'specialised' corpora on the other (Gavioli, 2005). This dissertation will be a comparative study from this dual perspective and the next section discusses the advantages and disadvantages of both.

2.3.2 General versus specialised corpora

From its outset, one of the main concerns of corpus linguistics was the construction of corpora that were representative as closely as possible of natural language (Sinclair, 1991; Kennedy, 1998). To this end, larger corpora offered not only ever increasing amounts of data but also, more importantly, different kinds (de Beaugrande, 1996). This allowed previously unnoticed patterns of language use to emerge, such as semantic prosodies (Sinclair, 1991), pattern grammars (Hunston & Francis, 2000) and collocation. It also suggested that language itself as a linguistic system, rather than being rule-based, is “inherently probabilistic” (Halliday, 1991, p. 31) and that choices made by users of the language are affected by more than just internal linguistic constraints (de Beaugrande, 1996). On the other hand, it is also the case that very large corpora tend to flatten or make invisible subtle but important facts about language (Hoey, 2004).

The notion of the influence of context on language choice has also informed the development of more recent smaller, specialised corpora. Interest in, and corpus studies of, genre has also grown in recent years (Hyland, 2002). The use of these specialised corpora can play a key role in the investigation of the linguistic characteristics of restricted academic disciplines (Hunston, 2002), specialised language (Bowker & Pearson, 2002) and the description of recurrent patterns that may characterise particular genres or registers within a sample of texts (Gledhill, 2000; Williams, 2002). It has found application in particular within such areas as English for Specific Purposes (ESP) and English for Academic Purposes (EAP) (Gavioli, 2005). From an SFL perspective, corpus studies have been conducted on such genres as newspaper editorials (Ansary & Babaii, 2005; Hasan & Babaii, 2005), press releases (Lassen, 2006), and application letters (Henry & Roseberry, 2001b). Studies of collocation within specialised corpora, however, have largely been confined to restricted academic genres and it still remains the case that the results from these collocation studies have largely failed to filter through to the wider world of general foreign language teaching (Sinclair, 2008). It may prove of more benefit, both to EFL students and teachers, therefore, to focus on broader, more general interest genres, such as WSR.

2.3.3 Corpus design for specialised corpora

Whereas most discussion of issues in corpus design has tended to focus in particular on representativeness in large corpora (Sinclair, 1991; Biber, 1993; Kennedy, 1998), the construction of specialised corpora, in particular for investigating collocation, also entails a number of design considerations and decisions. For the purposes of this dissertation, these can be grouped into four main categories:

- Selection of texts;
- Balance and representativeness;
- Size of corpus;
- Mark-up of texts

The first of these elements, selection, entails decisions regarding the number of texts to be included, the size of each text, and whether to use whole texts or samples (Bowker & Pearson, 2002). For a specialised corpus it is generally preferable to include whole texts as the position of an item in the text may also be significant (Bowker & Pearson, 2002). Most importantly, the criteria for selection must also be clear (Biber, Conrad, & Reppen, 1998). The second element, balance and representativeness, concerns how valid the claims of the corpus to be representative of the genre as a whole may be (Hunston, 2002), while the third has implications for the validity of any statistical claims for collocation that may be made.

The last point to be made concerns the mark-up of texts. There are two main approaches to the analysis of collocation in corpora, category-based or word-based. The first approach involves marking the text for parts of speech, or tagging, to investigate broad categories of language and, for specialised corpora, whether certain forms are more frequent within particular registers or genres (Hunston, 2002), such as the work of Biber (1995; 2003) using multi-variant analyses of, what he terms, register or text-type variation. The second approach, word-based, involves leaving the corpus un-tagged and investigating individual words for any patterns that may exist, such as Kennedy's (1991) study of the near-synonyms 'between' and 'through'. All the design decisions outlined here have certain advantages and disadvantages depending on the corpus involved and, importantly, the nature of the research questions that frame the corpus study (McEnery, Xiao, & Tono, 2006). The decisions taken for this particular study will be outlined in more detail in Chapter 3.

2.3.4 Corpus studies, collocation and cohesion

As described above, recent research on specialised corpora has highlighted the role of collocation within particular genres. This research has mainly focused on the contribution of collocation to the subject field and interpersonal tenor (defined in more detail below) that may characterise those genres (Hunston, 2002). Somewhat less attention, however, has focused on the contribution of collocation to textual cohesion, produced when one element of text is dependent upon another (Leckie-Tarry, 1995). Winter (1977, cited in Nation, 2001) introduced the notion of lexical items that function to signal the organisation of discourse, such as *compare, conclude, consequence* or *solution*, while McCarthy (1991) extended this notion to demonstrate how the lexical items produce cohesive ties to signal previous discourse. Hoey (1991), who preferred the term ‘links’ to ties, demonstrated the manner in which these links also produce cohesion across larger stretches of text. Corpus research has thus provided a somewhat more systematic view of the role of lexis within text and it is now recognised that these open-set lexical items in general perform a vital cohesive function. In a corpus study of a spoken genre (‘scheduling meetings’), Taboada (2004, p. 170), for example, found that cohesion was largely lexical at 70%.

Yet, even here, collocation is often overlooked and, within EFL in particular, the discussion of cohesion has largely focused instead on the closed-set grammatical items, especially conjunction. Hoey (1991), for example, specifically excluded collocation from his analysis. As described above, Halliday and Hasan (1976) identified collocation by frequency of occurrence and viewed it in paradigmatic terms, able to cut across sentence boundaries. Collocation can thus contribute to cohesion by providing semantic ties, which can be seen in the following example where the collocational tie between ‘smoking’ and ‘pipe’ contributes to the cohesion of the text:

> A little fat man of Bombay
> Was **smoking** one very hot day.
>   But a bird called a snipe
>   Flew away with his **pipe**,
> Which vexed the fat man of Bombay.
>
> (Halliday, 2004, p. 577)

As Moon (1998, p. 283) points out, however, this simple conception of collocation by frequency is “weak and insufficiently rigorous for formal analysis of the lexical organisation of texts”. More recent corpus linguistics research, questioning the division of lexical and grammatical cohesion (Mahlberg, 2006), suggests that open-set textual resources may, in fact, function to provide cohesion, yet also recognises that this varies with both register (Martin, 1992) and genre

(Gledhill, 1995). As such, corpus analysis through large corpora of the lexical items that contribute to cohesion is difficult and consequently undervalued. Yet this dissertation will investigate collocation from precisely this perspective of the textual role of certain key collocations within the genre of WSR for creating cohesion. The next section will go on to define genre further, leading to an analysis of WSR as a genre.

2.4 Analysing genre

Like collocation, genre is a “fuzzy concept” (Swales, 1990). Halliday (2004) himself even saw no need for the concept of genre, regarding register as adequate. The next sections provide an outline of the approach to genre and why the concept may be necessary. This will be followed by a framework of genre analysis that will then be applied in order to establish this particular WSR as a genre. Due to space limitations only the main theoretical points will be discussed.

2.4.1 Definition of genre

At its most basic level, genre is concerned with texts. As mentioned above, the underlying assumption of language here is that of a cline of lexico-grammar, which is a realisation of three meaning systems (Halliday & Hasan, 1985; Painter, 2001), or metafunctions: ideational meanings concerned with expressing content, interpersonal meanings concerned with the participants in discourse and stance, and textual meanings that create cohesion in text. These meanings are themselves a realisation of the context in which they occur. This ‘context of situation’, first suggested by Malinowski (1923, cited in Halliday & Hasan, 1985) and later expanded by Firth (Firth, 1957), sees meaning as being situated in, and influenced by, the surrounding context, comprised of three variables of discourse: field which refers to the nature of the activity in which the discourse is taking place, tenor which describes the relationship of the participants in that discourse, and mode which refers to the role language plays in the discourse (Halliday & Hasan, 1985). This is not to say, however, that the context of situation is itself linguistic (Martin, 2001), but rather it is realised by the three meanings systems above, as shown in Figure 3:

Figure 3: Realisation of register to text

SITUATION: Feature of the context	(realised by)	**TEXT:** Functional component of semantic system
Field of discourse (what is going on)	↘	Ideational meanings (transivity, naming, etc.)
Tenor of discourse (who are taking part)		Interpersonal meanings (mood, modality,person,etc.)
Mode of discourse (role assigned to language)		Textual meanings (theme, information, cohesive relations)

(from Halliday, 1985)

The semiotic systems of metafunction and register thus express how language is affected by the context in which it occurs. Cohesion helps tie this semiotic system together as text through surface textuality (Halliday & Hasan, 1985), or what could be termed ‘intratextual context’ (Leckie-Tarry, 1995). However, this in itself is not adequate to make sense of the text. As mentioned in Section 2.2.1, a text is distinguished from non-text by both cohesion and coherence.

As opposed to the surface ties of cohesion, coherence is concerned with the underlying structure of the text and how it ‘hangs together’ (Carter, 1998). A text may be cohesive, yet still not be coherent. This coherence of structure is formed through obligatory and optional recurring elements in sets of texts, the totality of which forms the Generic Structure Potential (GSP) (Halliday & Hasan, 1985) for that set. According to Halliday and Hasan (1985, p. 64), the GSP for ‘service encounters’, for example, is:

$$[(G)\bullet(SI)^\wedge]\ [(\overleftarrow{SE\bullet})\{SR^\wedge\overleftarrow{SC^\wedge}\}^\wedge S^\wedge]\ P^\wedge PC(^\wedge F)$$

where:

G = Greeting
SI = Sale Initiation
SE = Sale Enquiry
SR = Sale Request
SC = Sale Compliance
S = Sale
P = Purchase
PC = Purchase Closure
F = Finis

( ) = optional element
• = more than one option in sequence
[ ] = restraint in order
$\overleftarrow{\quad}$ = iteration
$\overleftarrow{\{\ \}}$ = iteration of elements is equal

In other words, there are certain obligatory elements that characterise the genre, in this case the ‘Sale’, ‘Purchase’ and ‘Purchase Closure’, and other optional ones that add elaboration but are not necessary. This was later expanded by Swales (1990), who proposed a two stage process of what he termed ‘moves’ within which are smaller ‘steps’, which will be used in the analysis in Section 2.4.2. There is still little agreement, however, as to the exact relationship between coherence and cohesion (McGee, 2009).

Genre is thus concerned with whole texts and the ways in which they relate to other texts (Halliday & Hasan, 1985), or ‘intertextual context’ (Leckie-Tarry, 1995), in order to engage in some kind of social activity. Importantly also, therefore, is that genre is a realisation beyond register of the ‘context of culture’, or the “complex knowledge system spread between various members of a particular culture” (Leckie-Tarry, 1995, p. 20). The context of culture is thus not ‘culture’ itself, per say, but the different ways in which members of the culture interact through language realised as genre to achieve certain goals (Martin & Rose, 2005). In summary, a genre can thus be defined as:

> a staged, goal-oriented, purposful activity in which speakers engage as members of our culture
>
> (Martin, 2001)

This study attempts to situate collocation within this system from a consideration of its functional role for textual cohesion. The next section will outline how the position described above may be applied to the analysis of a genre.

2.4.2 Analysing genre

Despite the definition of genre given above in Section 2.4.1, the precise nature of the process whereby members of the culture recognise particular genres is still not entirely clear. Genre can be considered an anticipatory process; that is, the recognition of genre affects how the text is subsequently processed by the listener/reader (Zwaan, 1994). If this is the case, then it would require both an external element of initial genre identification followed by an internal one that allows the recognition of the unfolding of stages within the text (Paltridge, 1994; Martin, 1992). These two concepts of external and internal genre recognition will form the framework for the genre analysis in Section 2.4.3 below.

2.4.3 Written sports reports (WSR) as a genre

The broad framework of external and internal genre analysis will now be applied to the Guardian corpus, exemplified by the Champions League game between Liverpool and PSV Eindhoven on 17th Nov. 2001 (Appendix 1). The Champions League is the premier Europe-wide club football competition, begun in 1955 and now comprising 32 teams in a group and knockout stage format (http://www.uefa.com/competitions/ucl/history/index.html - accessed 14th June, 2009). The competition is widely covered in the British media, through television and radio coverage, newspaper and internet reports and analysis, and discussed in blogs, radio phone-in and discussion programmes, not to mention its use in advertising. The Champions League thus operates through a number of differing genres within each of which are a variety of differing registers, or what could be termed ‘genre sets’ (Paltridge, 2007), which will be further discussed in Chapter 5.

Looking first at the external elements that characterise genre (Appendix 2), there are four main obligatory elements that characterise the genre: Result (R), Headline (H), Byline (B), Date (D) and the main report Text (T), although Result and Headline may be combined. There is also a seemingly obligatory element of Intertextual Links [IL] that possibly marks these texts as an internet genre separate from the print version. There are also optional elements of Picture (P) and Caption (C). An external GSP for the WSR could thus be represented as:

[H^R^D^IL^B^(P)^(C)^] T

where:
H = Headline
R = Result
D = Date
IL = Intertextual links

B = Byline
P = Picture
C = Caption
T = Text

The WSR genre is characterised internally by four Moves, shown in Table 1 below.

Table 1: Internal Moves and Steps for the WSR genre

MOVE	STEP	FUNCTION	TEXT FEATURES
Move 1 Providing context	EO	To establish evaluative frame for the text	Stakes…rewards…tonight
	MR	To provide the result	first place…secured…against PSV
	CO	To place the game within a wider (real world) context	the American owners; new four-and-a-half year salary; 13 months ago; long term
Move 2 General evaluation	ME:O	To provide general evaluation of the game overall	at the expense of; accommodating; triumphed convincingly; a point to prove
	ME:T	To evaluate the contribution of team players	Robbie Keane…tireless; impress…Ryan Babel; Lucas…jeered; All three impressed; opponents…weaker
	EQ	To provide a higher level of interpersonal evaluation	We…important…I'm pleased…managed to…positives…worked hard…the Liverpool manager said
Move 3 Events and evaluation	ME:M1/ E/ ME:G	To evaluate first half/ To describe Events/ To evaluate goal	casual; sterile; first half/ took the lead…36th minute; stretched…before the interval/ equalized…remarkable ease; ensured further invective; suspect guard
	ME:M2/ ME:G/ ME:P	To evaluate second half/ To evaluate goals/ To evaluate player	comfortable…second period/ goals…exquisite; stunning strike/ Isaksson…should have done better
	ME:P	To evaluate one particular player	Keane…selfless…gained the reward…deserved
Move 4 Reorientation	CR	To evaluate the match as a whole and place it within (competition) context	milestone…40th win…European games…continental triumphs
	FO	To orientate the game within the future context of the competitions	will

KEY: EO = evaluative orientation; CO = contextual orientation; ME:O = match evauation:overall; ME:P = match evaluation:team; EQ = evaluative quote; ME:M1 = match evaluation:first half; ME:M2 = match evaluation:second half; E = events; ME:G = match evaluation:goal; ME:P = match evaluation:player; CR = contextual reorientation; FO = future orientation

As opposed to Ghadessy (1988, p. 20) who saw the main purpose of the genre as “primarily a narration of events”, it will thus be suggested here that the overall communicative function of the written sports report is not merely to provide a factual ordering of events but to offer a level of evaluative interpretation of those events and act as an inter-discursive text that bridges the primary text of the game itself with secondary evaluative texts.

2.5 Summary

This chapter has outlined a definition of collocation and its place within genre theory from an SFL perspective, followed by a reading of Guardian Champions League reports as a genre. To summarise this position then, we can identify five elements of text and context:

- Text – the configuration of Ideational, Interpersonal and Textual meanings;
- Context of Situation – the configuration of field, tenor and mode features, realised as register;
- Context of Culture – the institutional and ideological background of the text, realised as genre;
- The 'intertextual' context – relations with other texts and their assumptions;
- The 'intratextual' context – coherence and cohesion within a particular text

(adapted from Halliday & Hasan, 1985; Leckie-Tarry, 1995)

Collocation, and its contribution to cohesion, will be investigated from these five perspectives. The chapter has also discussed the role of corpus linguistics in the investigation of collocation and genre. The use of specialised corpora has become established within the areas of ESP and EAP yet within the broader field of general EFL teaching the findings of corpus studies, and their classroom application, have still not been as widely applied or even accepted (Sinclair, 2008). As will be discussed further in Chapter 5, however, this dissertation hopes to demonstrate that the findings of corpora studies of genres such as sports reports may also have benefits and applications for general EFL teaching.

Chapter 3 – Methodology

3.1 Introduction

The following sections will outline the methodology to be used in this study. In Sections 3.2 and 3.3 the design and construction of the Guardian Champion's League corpus (GCL) and Bank of English (BoE) will be described. Next, Sections 3.4 and 3.5 will outline the procedure for the analysis and comparison of collocation between the two corpora. As mentioned in Chapter 2 above, there is no established methodology for the comparison of corpora. This study will follow two stages. The first step is the identification of collocation within the GCL for the purposes of analysis, in which significant keywords and function words are extracted from the corpus and then examined for collocation. The next step will be to compare this collocation with the larger BoE with the aim of investigating whether there exist any similarities or differences in the collocation patterning.

3.2 ‘Guardian Champions League’ corpus (GCL)

As mentioned in Section 2.3 above, there are several considerations in corpus design that must be taken into account, including selection, size and mark-up. The next two sections outline the decisions taken in constructing the GCL in light of these factors, and then give a detailed breakdown of the final composition of the corpus[2].

3.2.1 Collection of data

All texts to be used in the corpus were taken from the Guardian website (www.guardian.co.uk/football/championsleague). Only games from the group stages of the competition were selected and these were chosen according to the genre criteria outlined in Section 2.4.3. A total of 158 texts was collected and grouped according to season, spanning a period from the 2003/04 season to the 2008/09 season. These were then saved as *WordSmith 5.0*-readable Unicode text documents.

[2] A full copy of the corpus is available on request.

3.2.2 Construction of Corpus

Once the texts were collected all headlines and by-lines, etc. were removed, as these have their own particular linguistic characteristics (Perfetti, 1987; Moon, 1998), leaving only the main text report for the purposes of analysis. As this study is primarily a word-based study of collocation, the decision was made to leave the corpus un-tagged and un-lemmatised. There are a total of 6 sub-corpora, one for each season, giving a total corpus size of 111,136 words, which was felt to be adequate. A breakdown of the number of texts for each season, and the number of tokens, mean words and standard deviation, number of types and type-token ration, is given in Table 2.

Table 2: Breakdown of the Guardian Champions League Corpus

Year	Texts	Tokens	Mean words	Std. Dev. words	Types	Type-token ratio
2003/04	30	20,979	699.33	103.78	3,921	18.69
2004/05	33	21,320	646.55	125.15	4,136	19.39
2005/06	27	19,246	713.19	177.89	3,739	19.42
2006/07	27	20,012	742.04	110.14	3,881	19.37
2007/08	19	13,430	707.79	131.49	3,104	23.08
2008/09	22	16,149	734.45	115.50	3,403	21.06
OVERALL	158	111,136	703.88	131.50	10,162	9.14

3.3 ‘Bank of English’ corpus (BofE)

The ‘Bank of English’ corpus is a continually expanding monitor corpus aimed at providing a view of the “state of the language” (Sinclair, 1991, p. 26). It is designed to be as large as possible, reflecting Sinclair’s (1991) belief that large amounts of data are required for meaningful statistics, especially when it comes to collocation. The ‘Bank of English’ corpus is jointly owned by HarperCollins Publishers and the University of Birmingham (http://www.titania.bham.ac.uk/). In 2007 the corpus stands at 450 million words. A breakdown of the sub-corpora is shown in Table 3:

Table 3: Sub-corpora of the Bank of English (BoE) corpus

Sub-corpus	**Words**	**Country**	**Description**
usacad	6,341,888	US	Academic books
usephem	3,506,272	US	Ephemera (advertisements, leaflets, guides, etc.)
newsci	7,894,959	UK	New Scientist magazine
npr	22,232,422	US	Public radio
sunnow	44,756,902	UK	Sun/News of the World newspapers
brbooks	43,367,592	UK	General fiction and non-fiction books
brmags	44,150,323	UK	Magazines
guard	32,274,484	UK	Guardian newspaper
econ	15,716,140	UK	Economist magazine
bbc	18,604,882	UK	BBC radio
usspok	2,023,482	US	Spoken (conversations, telephone calls, service encounters, etc.)
wbe	9,648,371	UK	Business
strathy	15,920,137	Canada	Canadian mixed corpus
oznews	34,940,271	Australia	Newspapers
brephem	4,640,529	UK	Ephemera (advertisements, leaflets, guides, etc.)
usbooks	32,437,160	US	General fiction and non-fiction books
usnews	10,002,620	US	Newspapers
indy	28,075,280	UK	Independent newspaper
times	51,884,209	UK	Times newspaper
brspok	20,078,901	UK	Spoken (conversations, telephone calls, service encounters, etc.)
TOTAL	448,496,824		

This corpus will provide a base reference against which the GCL can be compared. The next section outlines the methods of analysis that will be used in the study.

3.4 Analysis of GCL corpus

The following section outlines the procedures that will be taken for the analysis of the ‘Guardian Champions League’ corpus in order to extract significant words which may then be analysed for collocation.

3.4.1 *WordSmith Tools 5.0*

Analyses of the corpus data will be conducted using *WordSmith Tools 5.0* (Scott, 2009), available from www.lexically.net/wordsmith/. This is a package of corpus analysis tools that was originally developed, and still used, for Oxford University Press in lexicography work (Scott, 2009). The package allows a range of analyses to be conducted and has been used for corpus studies in a wide variety of contexts (Berber Sardinha, 1999).

3.4.2 Keywords

A keyword is one which has an unusually high, or low, frequency in comparison to a base reference corpus (Berber Sardinha, 1999) and thus may characterise a text or a genre (Scott, 2009). These will form the basis for the first part of the corpus analysis. The first step in generating a keyword list is to produce a frequency list from the small corpus under study which can then be compared to a larger reference list to find those words which are statistically more significant (Scott, 2009). *Wordsmith Tools 5.0* provides a choice of either the chi-square test or log-likelihood test to produce the keywords. The chi-square and log-likelihood have been compared by Dunning (1993), Rayson & Garside (2000), Rayson, et al, (2004) and the log-likelihood was generally found to be the more accurate of the two. The decision was taken then to use the log-likelihood test for the keywords. From this keyword list the top five content words, excluding proper names, will be used as the basis for investigating collocation within the GCL, which can then be compared to the larger corpus.

3.4.3 Function words

Recent studies of collocation within genre have also increasingly focused on the role function, or grammatical, words play. While often overlooked (Bartsch, 2004) or seen as comparatively less significant than content words (Gledhill, 2000), function words and their collocation patterns may, in fact, provide significant information as to the characterisation of a genre (Gledhill, 2000; Groom, 2005). For this reason, significant function words as well content words will be investigated for their collocation properties within this study. These will both be identified by the keywords test, described in Section 3.3.2 above. The next section will outline the method of analysis for these keywords.

3.5 Analysis of collocation

The next section outlines the methodology for the analysis of collocation of the keywords and function words within the ‘Guardian Champion’s League’ corpus and how this collocation will be compared to the ‘Bank of English’ monitor corpus. First, the method for identifying collocation will be outlined followed by that for conducting a comparison with the BoE.

3.5.1 Identification of collocation

The key consideration in the identification of collocation is appropriate span, with suggestions generally ranging from one word either side of the node (Kennedy, 1991) to four (Sinclair, 1991), although suggestion have even ranged up to 50:50 (Stubbs, 1995). For this study, the first stage in identifying collocates will be through observation of word-based frequencies at an initial span of 3:3, followed by more detailed analysis of any patterns that emerge. To avoid the problem of a single text in the GCL containing a large number of the same collocate (Biber, Conrad, & Reppen, 1998), all collocates must also be present in more than one text.

It is also the case, however, that frequency alone may not be adequate and some measure of collocation strength is also required. There have been many suggestions as to the appropriate statistical methodology for the extraction of collocation from a corpus. Pecina (2005), for example, lists some 84 different association measures that have been proposed to calculate collocation strength (see Appendix 3). The decision here was taken to continue with the log-likelihood test to calculate both keywords and for the identification of collocation. For collocation, the log-likelihood test allows the comparison of both rare and frequent occurrences (Dunning, 1993) as it does not assume normal distribution (Oakes, 1998) and may thus distinguishing between significant occurrences and those resulting from syntactic forces alone or low frequencies of data. In terms of what value can be considered significant, Rayson, et al. (2004) found a critical value of 15.13 where p≤.000001. This value will be used for the study.

3.5.2 Comparison of collocation

Corpora have been compared in various ways yet it is also the case that most of these studies have been conducted on corpora of similar size. This study, however, is a comparison of two very different sized corpora and a reliable method of normalizing frequency counts has yet to be found (Baayen, 2005). The study here is concerned not so much with the strength of individual collocations but with the patterns of collocation surrounding keywords and their behaviour. As such, the approach within this study will thus be to compare this patterning between the GCL and BoE mainly in terms of their relative frequencies and, in this way, it can be seen whether certain collocations stand out in the specialised corpus relative to the larger one.

3.6 Summary

The main study of the ‘Guardian Champions League’ corpus will thus be composed of two parts. The first part will use the keywords function of *WordSmith Tools 5.0* to identify significant keywords and function words in the GCL. The second part will then compare the collocation patterns of these keywords and function words to a large reference corpus, the ‘Bank of English’, of 450 million words. The next chapter presents the results of this study.

Chapter 4 - Results

4.1 Introduction

The following sections present the results of the comparative corpus study of the GCL and BofE. While some similarities were found between the two corpora, it was also found that there were a number of key collocations with strikingly higher frequencies within the GCL as compared to the larger BofE. First, Section 4.2 identifies content and grammatical keywords within the GCL using *WordSmith Tools 5.0*, while Sections 4.3 and 4.4 analyse the collocation patterns of these content and grammatical keywords respectively in comparison to the BofE corpus.

4.2 Keywords

The GCL was compared to a BofE reference list taken from the BofE website at http://www.titania.bham.ac.uk/frequency%20lists/corpusrank.txt. Table 4 shows the top twenty content keywords (with proper names removed) within the GCL, as compared to the BofE list. From this the general domain of Champions League football can readily be identified.

Table 4: Top twenty content keywords by log-likelihood (ranked by keyness)

N	Key word	Freq. GCL	% GCL	Freq. BofE	% BofE	LL Keyness	P
2	CHAMPIONS	264	15.242	64		2376.890869	6.22E-20
3	GOAL	343	19.804	538		2336.740479	6.55E-20
7	LEAGUE	263	15.185	564		1662.112183	1.85E-19
8	TEAM	307	17.725	1534		1498.752441	2.55E-19
9	BALL	304	17.552	1700		1423.597778	2.98E-19
14	MINUTES	318	18.36	3793	0.0206787	1059.501465	7.42E-19
15	PLAYERS	177	10.219	654		957.1606445	1.02E-18
16	STRIKER	98	5.6582	8		945.4614868	1.06E-18
18	SHOT	216	12.471	1593		904.8990479	1.21E-18
19	HALF	331	19.111	6033	0.0328908	856.4360962	1.44E-18
22	GAME	233	13.453	2665	0.0145291	793.5245972	1.83E-18
23	KICK	130	7.5058	360		766.7202148	2.03E-18
24	WIN	171	9.873	1148		745.0040894	2.23E-18
27	GOALS	129	7.448	406		732.9768677	2.34E-18
28	MINUTE	183	10.566	1664		698.0903931	2.73E-18
30	HEADER	74	4.2725	14		679.7167969	2.97E-18
36	PENALTY	103	5.9469	282		609.4171753	4.2E-18
37	MIDFIELD	69	3.9838	35		573.1051025	5.11E-18
38	SCORED	92	5.3118	233		556.2512817	5.62E-18
39	MATCH	135	7.7945	1038		555.4172974	5.64E-18
42	FANS	82	4.7344	143		545.0278931	5.99E-18
46	KNOCKOUT	56	3.2333	8		524.4714355	6.78E-18
47	MIDFIELDER	51	2.9446	0		521.473999	6.91E-18

The following study will look at the collocation patterns of the top content words ‘goal’, ‘ball’, ‘minutes/minute’ and ‘shot’ in more detail. These were chosen as they were not as domain-specific as ‘champions’ and ‘league’ and may thus provide better collocation information in comparison with the BofE.

The top twenty function keywords, shown in Table 5, are perhaps notable for the large number of prepositions of location, movement and time, reflecting the action of a game of football.

Table 5: Top twenty function keywords by log-likelihood (ranked by keyness)

N	Key word	Freq. GCL	% GCL	Freq. BofE	% BofE	Keyness	P
55	AFTER	439	25.346	21220	0.116	453.7155762	1.08E-17
77	THEIR	710	40.993	52849	0.288	349.6662292	2.54E-17
99	FROM	880	50.808	76274	0.416	297.1157227	4.39E-17
120	HIS	1073	61.951	102929	0.561	265.6581116	6.43E-17
137	WIDE	94	5.4273	1776		237.2350616	9.53E-17
204	BUT	943	54.446	98522	0.537	170.1262207	3.17E-16
224	WAS	1649	95.208	198514	1.082	148.9938202	5.26E-16
227	OFF	223	12.875	14454	0.079	144.9200134	5.86E-16
246	AGAINST	178	10.277	10757	0.059	130.9644623	8.78E-16
252	WHEN	495	28.58	47185	0.257	124.7126312	1.07E-15
261	HAVE	801	46.247	87441	0.477	119.6321106	1.27E-15
269	BEFORE	240	13.857	18233	0.099	112.6014481	1.65E-15
398	BACK	252	14.55	22452	0.122	78.32037354	9.19E-15
417	BEYOND	64	3.6952	2846	0.016	73.61127472	1.29E-14
423	A	3050	176.1	430681	2.348	71.83927917	1.48E-14
448	CLOSE	73	4.2148	3692	0.02	70.75765991	1.61E-14
460	BEEN	454	26.212	49495	0.27	68.05059814	2.03E-14
482	WERE	558	32.217	64471	0.351	63.04746246	3.26E-14
491	HERE	169	9.7575	14224	0.078	61.56837463	3.82E-14

A combination of the top two preposition keywords, ‘after’ and ‘from’, and the top two auxiliary verbs, ‘was’ and ‘have’, will be analysed for this study. The next Sections will now look at these keywords and their collocation in more detail.

4.3 Content keywords

The following sections present a comparative description and analysis between the GCL and BofE of the keywords ‘goal’, ‘ball’, ‘minutes/minute’ and ‘shot’.

4.3.1 ‘goal’

Goal occurs in 134 texts (84.81%) with a total frequency of N = 343 within the GCL and N = 60683 within the BofE. At a right-side span of 0:1 in Table 6, there are only four significant collocations within the GCL. The top two of these, <goal difference> and <goal line>, may be considered compound nouns.

Table 6: Top GCL collocates of goal by log-likelihood (span = 0:1, LL≥15.13)

Guardian Champions League Corpus		
N = 343		
COLLOCATE	LL score	n
DIFFERENCE	44.04877853	5
LINE	20.57001495	5
FOR	18.66689491	14
WAS	15.26586056	16

A look at the BofE also reveals several collocates of goal that may be considered compound nouns, including *goal hero* (LL=2584.64), *goal scorer* (LL=1907.15) and *goal posts*

(LL=1292.18). The fact that some word-forms more readily combine into compound nouns would be useful information for learners.

The left-side collocates of goal within the GCL in Table 7 are notable for the use of ordinals, possessives and descriptors. These will be discussed further in Chapter 5.

Table 7: Top GCL collocates of goal by log-likelihood (span = 1:0, LL≥15.13)

Guardian Champions League Corpus		
N = 1649		
COLLOCATE	LL score	n
SECOND	99.3653717	20
A	98.91694641	52
THIRD	64.83300018	10
FIRST	63.45814133	17
OWN	47.10502625	9
OPENING	34.95968628	7
ACROSS	30.99990463	5
LATE	22.68929863	5
OPEN	21.3728447	4
ANOTHER	20.22229576	5
HOWARD'S	19.32189941	2
LEAGUE	18.04784203	7
ANELKA'S	17.60195351	2
LJUNGBERG'S	16.42316055	2
EQUALISING	15.52125359	2
LEHMANN'S	15.52125359	2
NISTELROOY'S	15.52125359	2

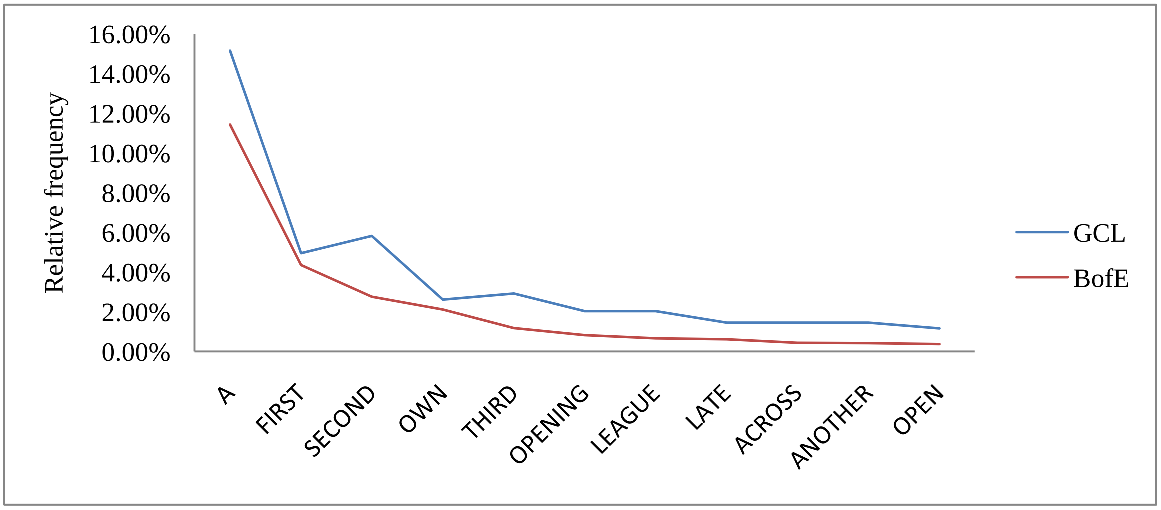
Comparing these left-side collocates by relative frequency (n/N*100%) between the GCL and the BofE in Table 8 also shows a general similarity between the two.

Table 8: Comparison of collocates of goal by relative frequency (span = 1:0, n≥2)

GCL Corpus			BofE corpus	
N = 343			N = 60683	
COLLOCATE	n	%	n	%
A	52	15.16%	6942	11.44%
SECOND	20	5.83%	1676	2.76%
FIRST	17	4.96%	2645	4.36%
THIRD	10	2.92%	719	1.18%
OWN	9	2.62%	1289	2.12%
OPENING	7	2.04%	497	0.82%
LEAGUE	7	2.04%	402	0.66%
ACROSS	5	1.46%	266	0.44%
LATE	5	1.46%	378	0.62%
ANOTHER	5	1.46%	259	0.43%
OPEN	4	1.17%	229	0.38%

The similarities between the GCL and BofE thus suggest that within the patterning of collocation surrounding ‘goal’, seen in Figure 4 below, no one collocate stands out as particularly significant for the GCL corpus.

Figure 4: Comparison of collocates of goal by relative frequency (span = 1:0, n≥2)



4.3.2 ‘ball’

As opposed to ‘goal’ above, ball displays both considerable variation and similarities between the two corpora. Ball occurs within the GCL in 120 texts (75.95%) with a total frequency of N = 304, and a frequency of N = 69119 within the BofE. Turning first to right-side collocates in Table 9, it can be seen that ball within the GCL is commonly followed by a prepositional phrase of movement or location.

Table 9: Top GCL collocates of ball by log-likelihood (span = 0:1, LL≥15.13)

Guardian Champions League Corpus		
N = 304		
COLLOCATE	LL score	n
INTO	154.7014313	29
AWAY	59.53800201	12
OVER	50.83135223	12
PAST	28.5637722	6
HOME	24.16307259	7
FROM	23.17192268	13
AROUND	19.29984093	3
ACROSS	16.08016777	3

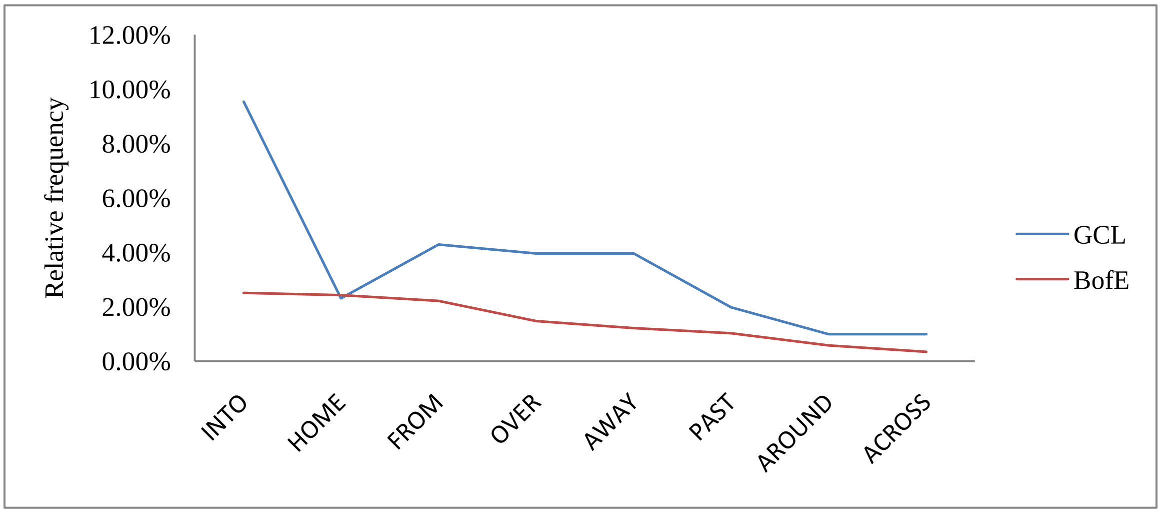
Comparing these collocates with the BofE by relative frequency (n/N*100%) in Table 10 below, it appears that there are some distinct differences. As opposed to the BofE, the GCL seems to most frequently use <ball into>. This indicates a goal scored, specifically with <ball into + net> (n=9) and <ball into + goal> (n=2). Other differences include the collocations <ball from>, <ball over> and <ball away>.

Table 10: Comparison of collocates of ball by relative frequency (span = 0:1, n≥2)

GCL Corpus			BofE corpus	
N = 304			N = 69119	
COLLOCATE	n	%	n	%
INTO	29	9.54%	1737	2.51%
AWAY	12	3.95%	836	1.21%
OVER	12	3.95%	1015	1.47%
PAST	6	1.97%	706	1.02%
HOME	7	2.30%	1676	2.42%
FROM	13	4.28%	1528	2.21%
AROUND	3	0.99%	391	0.57%
ACROSS	3	0.99%	232	0.34%

These differences can be clearly seen in Figure 5 below:

Figure 5: Comparison of collocates of ball by relative frequency (span = 0:1, n≥2)



On the other hand, it may also be seen that <ball home> is used with very similar frequencies in both corpora.

Moving to left-side collocates at a span of 3:0 in Table 11, it can be seen that there are noticeable differences between the two corpora. The BofE, as may be expected, demonstrates the use of ‘ball’ in a general sense, with *hit* + *ball*, *play* + *ball* and *kick* + *ball* (see Appendix 4). Interestingly, this last collocate is not among the significant collocates within the GCL. This corpus instead seems to employ material processes of movement that may, in fact, usually be associated parts of the body other than ‘foot’, especially ‘hand’ with collocates such as ‘rolled’, ‘gave’, ‘flicked’, ‘steered’ or ‘slotted’, shown below in Table C.

Table 11: Top GCL collocates of ball by log-likelihood (span = 3:0, LL≥15.13)

Guardian Champions League Corpus		
N = 304		
COLLOCATE	LL score	n
LOOSE	158.5554657	14
HAND	49.79420853	7
DEAD	48.52889252	5
PUT	46.47290421	9
LONG	42.60227203	8
GIVING	38.56528473	5
WITH	34.68873978	17
ROLLED	33.83707047	4
CHASED	28.71940994	3
DINKED	28.71940994	3
THROUGH	28.714468	7
GAVE	28.39023781	5
HE	25.02817726	13
CRASHING	23.61915207	2
CHESTED	23.61915207	2
SAW	23.56918716	4
FLICKED	23.25936127	3
CONTACT	20.06671333	3
STEERED	19.80550957	2
SLOTTED	19.80550957	2
KICKING	18.08485794	2
WEIGHTED	18.08485794	2
UNDERSTUDY	16.0027504	2

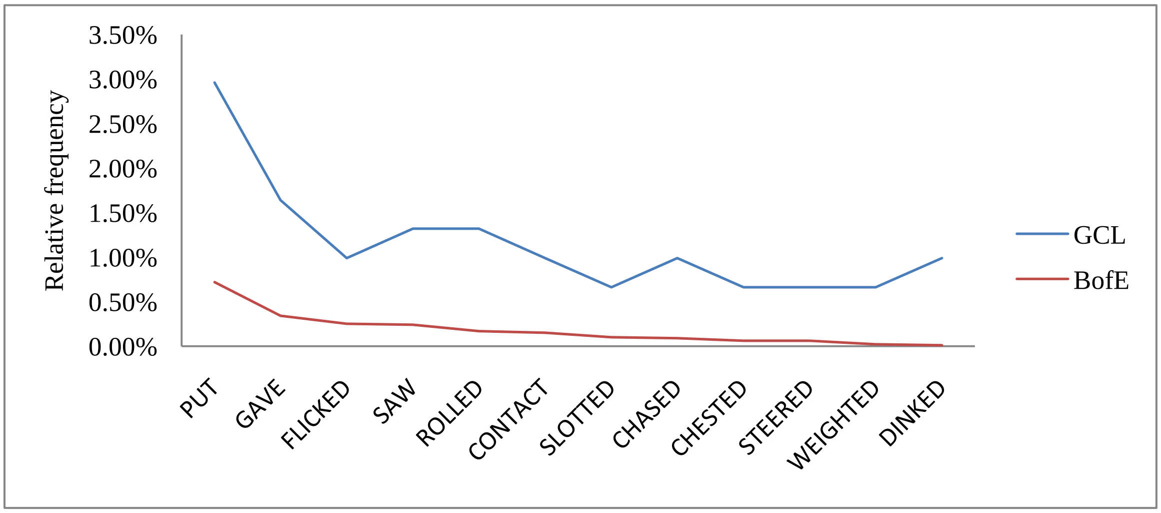
Compared to the BofE by relative frequency (n/N*100%) in Table 12 below, these processes have higher frequencies within the GCL. These collocates thus seem particularly associated with the GCL, although whether they are characteristic of just the GCL sub-register or the WSR genre as a whole would need more research.

Table 12: Comparison of material process collocates of ball by relative frequency (span = 3:0, n≥2)

GCL Corpus			BofE corpus	
N = 304			N = 69119	
COLLOCATE	n	%	n	%
PUT	9	2.96%	501	0.72%
ROLLED	4	1.32%	117	0.17%
CHASED	3	0.99%	60	0.09%
DINKED	3	0.99%	4	0.01%
GAVE	5	1.64%	236	0.34%
CHESTED	2	0.66%	44	0.06%
SAW	4	1.32%	163	0.24%
FLICKED	3	0.99%	171	0.25%
CONTACT	3	0.99%	107	0.15%
STEERED	2	0.66%	44	0.06%
SLOTTED	2	0.66%	66	0.10%
WEIGHTED	2	0.66%	14	0.02%

The higher frequencies of these material process choices can be clearly seen in Figure 6:

Figure 6: Comparison of material process collocates of ball by relative frequency (span = 3:0, n≥2)



4.3.3 ‘minutes/minute’

Minutes occurs within the GCL in 55 texts (34.81%) with a corpus frequency of N = 318, compared to N = 113568 within the BofE. For significant left-side collocates at a span of 1:0, the GCL is restricted to a single pattern of <n minutes>. Minutes in the BofE, on the other hand, takes a variety of premodification, shown in Table 13 below (cardinal numbers excluded):

Table 13: Top BofE collocates of minutes by log-likelihood (span = 1:0, LL≥15.13)

Bank of English Corpus		
N = 113568		
COLLOCATE	LL score	n
few	52022.0322	6654
within	8732.605377	1527
several	4022.67445	830
dying	1648.135357	249
half	1571.664768	519
closing	1338.623684	221
opening	923.058609	226
final	515.6320381	224
just	319.1516685	421
many	128.0200882	229
only	119.3697166	305
next	73.14442062	10

Expanding this to a span of 2:0 in Table 14 below, however, the GCL does contain both cardinals and other types of premodification. According to Butt, et al. (1985, p. 83), the most common ordering of premodification is Deictic Numerative Epithet Classifier (for example, ‘those two magnificent cedar trees’). The GCL, on the other hand, seems to be characterised by the patterns Classifier Numerative, <opening n minutes>, or Epithet Numerative <mere n

minutes>. The most significant pattern, however, appears to be a prepositional phrase, <after n minutes>, which is over four times more frequent by relative frequency in the GCL (28.08%) than the BofE (6.33%).

Table 14: Top GCL collocates of minutes by log-likelihood (span = 2:0, LL≥15.13)

Guardian Champions League Corpus		
N = 317		
COLLOCATE	LL score	n
AFTER	628.9321899	89
FIVE	227.4656372	27
TWO	123.4534836	22
SEVEN	118.6481628	14
SIX	96.68167114	14
EIGHT	93.88532257	11
OPENING	92.73947144	14
FOUR	79.52127075	13
THREE	73.66695404	14
WITHIN	60.51552582	8
NINE	40.07003403	6
FIRST	33.39347076	11
WITH	22.86668968	14
ONLY	19.29691315	7
ON	19.1470871	12
TEN	17.90468788	2
UNTIL	17.696558	4
FINAL	17.07387161	4
LAST	16.2387886	7
MERE	15.82308483	2
GOAL	15.6904459	7

The collocation surrounding minute also appears characteristic of the GCL. The top left-side collocates of minute within the BofE, shown in Table 15, reflect its general use within the language, with such phrases as *wait* + *minute*, *for* + *minute* or *cost (60p) per minute*, and general collocations such as *every* + *minute*. The GCL, on the other hand, is overwhelmingly focused on the single collocation <in the (nth) minute>. Of the 183 occurrences of minute, 126 take this form. It seems likely then that this collocation is a key distinguishing feature of the corpus.

Table 15: Top collocates of minute by log-likelihood (span = 3:0, LL≥15.13)

Guardian Champions League Corpus			Bank of English Corpus		
N = 183			N = 3246851		
COLLOCATE	LL score	n	COLLOCATE	LL score	n
IN	786.7926636	127	A	92632.95506	22384
THE	563.6112671	138	LAST	49866.79306	6912
THIRD	67.65892029	9	IN	47911.43821	13891
EIGHTH	58.82480621	5	THE	47659.44355	20919
FOURTH	53.45007706	6	PER	21480.03772	2929
SEVENTH	46.35961151	4	60p	16513.34248	1062
UNTIL	29.81506729	5	COST	14887.98502	1879
LAST	28.68949318	8	WAIT	10220.2627	1134
A	26.69762802	20	ONE	9961.134544	2661
NINTH	21.84224319	2	EVERY	8722.170269	1442
DECO'S	20.11941147	2	FOR	7400.029408	3510

Turning to right-side collocates in Table 16, another interesting difference between the two corpora becomes apparent. Minute within the BofE actually seems more strongly associated with sport than the GCL, through collocations such as *minute* + *goal*, which mostly take the pattern *ORDINAL* *minute*:

	goal
	penalty
nth minute	*winner*
	equaliser
	strike

These collocations are, however, notably absent from the GCL.

Table 16: Comparison of top collocates of minute by log-likelihood (span = 0:3, LL≥15.13)

Guardian Champions League Corpus			Bank of English Corpus		
N = 1649			N = 3246851		
COLLOCATE	LL score	n	COLLOCATE	LL score	n
WHEN	37.67674255	11	WHEN	10534.37167	2591
OPENER	30.16797066	4	GOAL	7912.519043	1026
EFFORT	24.3154335	4	PENALTY	7110.706944	789
WAS	24.15813446	14	1p	6557.994839	485
MATUZALEM	21.84224319	2	LATER	5193.538685	926
JOSÉ	19.51674461	3	FROM	4825.172329	2019
SAVED	19.51674461	3	HE	4684.810547	2362
AND	19.24719429	15	AND	4511.994095	4572
CARDOZO	18.03293991	2	AFTER	4483.123714	1297
FREE	17.69919586	4	SILENCE	4439.002686	502
CAME	17.1265316	4	LEAD	4424.815804	693
BEFORE	16.37539673	5	WINNER	4311.010033	579
VAN	15.92369938	4	EQUALISER	4295.64336	374
			STRIKE	4139.408772	537
			HEADER	4043.166634	404

This seems to show that it is the collocation patterns surrounding the keyword that most indicate this particular sub-register, an important feature that would need to be pointed out to learners of the language.

4.3.4 'shot'

Shot occurs in 109 texts (69.99%) with a total frequency of N = 216 within the GCL, and N = 62615 within the BofE. Looking first at right-side collocates in Table 17, shot in the GCL seems to mainly indicate the result of an unsuccessful attempt on goal, indicated by <shot deflected>

and <shot saved> and through the use of prepositional Circumstances, as well as which- and that-clauses.

Table 17: Top GCL collocates of shot by log-likelihood (span = 0:1, LL≥15.13)

Guardian Champions League Corpus		
N = 216		
COLLOCATE	LL score	n
WIDE	96.90955353	14
BEYOND	52.17559052	8
OVER	45.39262009	10
FROM	44.18400192	16
WHICH	33.12448502	8
ROUND	24.79850006	4
DEFLECTED	21.71107483	3
STOPPING	21.17630768	2
SAVED	18.52927399	3
PAST	18.45030785	4
THAT	17.50626183	11
INTO	17.16020393	6

The top collocates for the BofE in Table 18, on the other hand, are predominantly the sense of ‘shot with a gun’ (<shot dead>) or ‘to move or leave quickly’ (<shot past>, <shot through>).

Table 18: Top BofE collocates of shot by log-likelihood (span = 0:1, LL≥15.13)

Bank of English Corpus		
N = 62615		
COLLOCATE	LL score	n
DEAD	38830.8883	3725
DOWN	10684.7078	1982
AT	7996.89506	2978
FROM	4441.29742	1955
IN	3708.03667	3697
PAST	3337.09978	669
UP	3076.21761	1208
WIDE	2626.29278	419
THROUGH	2348.60705	675
INTO	2229.11257	845
BY	2112.82134	1405
HIMSELF	1960.91712	425

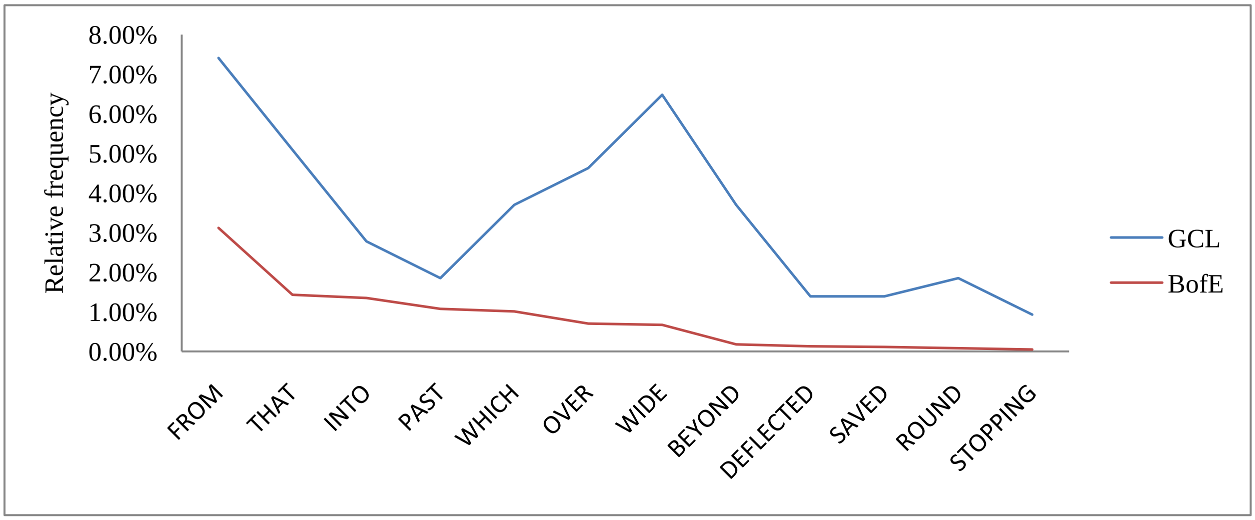
Turning to a comparison of the relative frequencies of these right-side GCL collocates in Table 19 however, clear differences between the corpora may be seen, with <shot from>, <shot wide>, <shot over> in particular having higher frequencies within the GCL.

Table 19: Comparison of collocates of shot by relative frequency (span = 0:1, n≥2)

Guardian Champions League Corpus			Bank of English Corpus	
N = 216			N = 62615	
COLLOCATE	n	%	n	%
FROM	16	7.41%	1955	3.12%
WIDE	14	6.48%	419	0.67%
THAT	11	5.09%	895	1.43%
OVER	10	4.63%	438	0.70%
BEYOND	8	3.70%	110	0.18%
WHICH	8	3.70%	634	1.01%
INTO	6	2.78%	845	1.35%
ROUND	4	1.85%	47	0.08%
PAST	4	1.85%	669	1.07%
DEFLECTED	3	1.39%	84	0.13%
SAVED	3	1.39%	71	0.11%
STOPPING	2	0.93%	31	0.05%

This can clearly be seen in Figure 7 below, although some similarity is also present in <shot past>.

Figure 7: Comparison of collocates of shot by relative frequency (span = 0:1, n≥2)



The top left-side collocates within the GCL in Table 20 below seem to describe the manner of the shot (<low shot>), as well as the field position (<yard shot>) and who it was taken by (<midfielder's shot>).

Table 20: Top GCL collocates of shot by log-likelihood (span = 1:0, LL≥15.13)

Guardian Champions League Corpus		
N = 216		
COLLOCATE	LL score	n
HIS	126.0599518	33
LOW	85.37536621	11
YARD	41.72498703	6
ANGLED	38.94545364	4
POWERFUL	38.94545364	4
RANGE	32.28086853	5
SWERVING	29.19507599	3
FOOT	25.75530434	4
DEFLECTED	21.71107483	3
CUM	21.17630768	2
MIDFIELDER'S	19.45407104	2

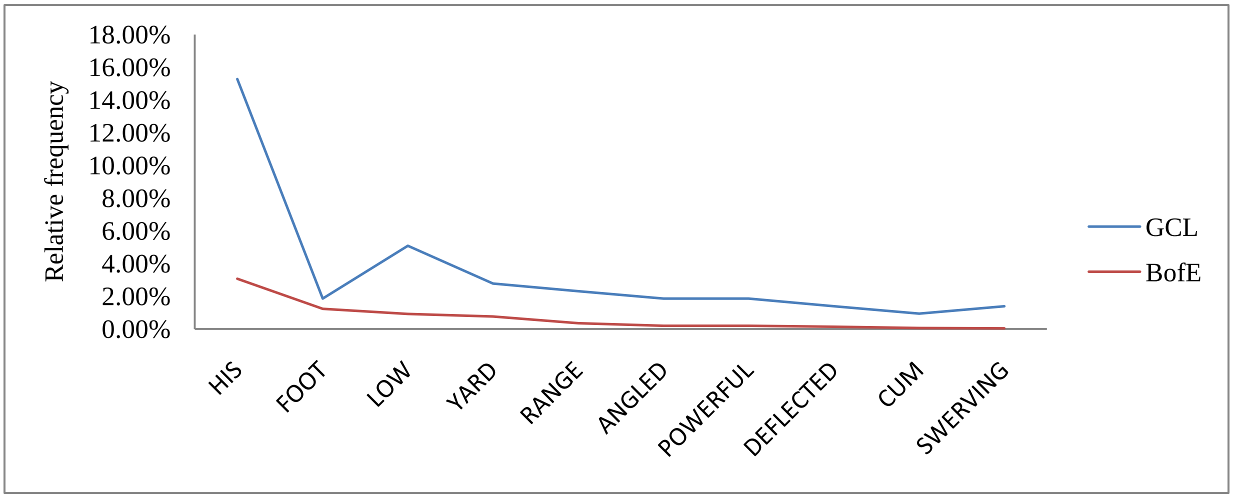
A comparison of the relative frequencies between the two corpora in Table 21 shows that this patterning of collocation appears to demonstrate some correlation with similar frequencies.

Table 21: Comparison of collocates of shot by relative frequency (span = 1:0, n≥2)

Guardian Champions League Corpus			Bank of English Corpus	
N = 216			N = 62615	
COLLOCATE	n	%	n	%
HIS	33	15.28%	1921	3.07%
LOW	11	5.09%	575	0.92%
YARD	6	2.78%	467	0.75%
RANGE	5	2.31%	213	0.34%
ANGLED	4	1.85%	126	0.20%
POWERFUL	4	1.85%	117	0.19%
FOOT	4	1.85%	764	1.22%
SWERVING	3	1.39%	20	0.03%
DEFLECTED	3	1.39%	84	0.13%
CUM	2	0.93%	29	0.05%

The two particular exceptions are <his shot> and <low shot>, which can be seen in Figure 8 below.

Figure 8: Comparison of collocates of shot by relative frequency (span = 1:0, n≥2)



In summary, it appears that, while there is some similarity between the two corpora, there are also distinct differences that highlight the GCL corpus. The next section will now present results for frequency keywords.

4.4 Frequency keywords

The following sections present an analysis of collocation for the function keywords ‘after’, ‘from’, ‘was’ and ‘when’.

4.4.1 ‘after’

After occurs in 142 texts (89.87%) within the GCL, with a total frequency of N = 440 and, in the BofE, N = 681330. The top left-side collocates of after within the two corpora, shown in Table 22, are both similarly concerned with time and the ordering of events, with the collocates <shortly after> and <soon after> both significant. It also possibly demonstrates the effect of Field with the collocation <minutes after> in the GCL reflecting the shorter time-span of the game of football, as opposed to the longer one of the BofE, with <days after> and <years after> significant.

Table 22: Top collocates of after by log-likelihood (span = 1:0, LL≥15.13)

Guardian Champions League Corpus			Bank of English Corpus		
N = 440			N = 681330		
COLLOCATE	LL score	n	COLLOCATE	LL score	n
SHORTLY	64.03852081	8	SHORTLY	61183.3577	6355
MINUTES	53.14352036	16	DAYS	42005.8642	8401
SOON	48.47557068	8	SOON	41740.1976	6888
WOUND	22.14520454	2	YEARS	38318.159	10922
GOALLINE	16.6157856	2	MONTHS	32420.1889	6628
AHEAD	16.48166656	4	LOOK	31575.2998	7017

The right-side GCL collocates in Table 23 indicate the action of the match through collocates such as <after + restart>, <after + deflection> or <after + corner>. It also seems to indicate a metaphor of FOOTBALL AS WAR through collocates such as ‘clash’, ‘defeats’ and ‘losing’. This concept will be returned to in Chapter 5.

Table 23: Top GCL collocates of after by log-likelihood (span = 0:3, LL≥15.13)

Guardian Champions League Corpus		
N = 440		
COLLOCATE	LL score	n
MINUTES	677.5009766	94
INTERVAL	143.0142822	21
RESTART	48.60065842	6
BEING	38.50834274	8
CLASH	34.77160263	4
HOUR	32.02558517	6
DEFLECTION	24.869627	4
BRIGHT	21.81460762	3
BREAK	21.23588562	4
DEFEATS	19.79940605	3
HALF	19.65776062	9
ONLY	19.36616325	8
LOSING	17.88441849	3
CORNER	16.25075912	5
TIME	16.1771431	7
RECENT	15.5701704	3

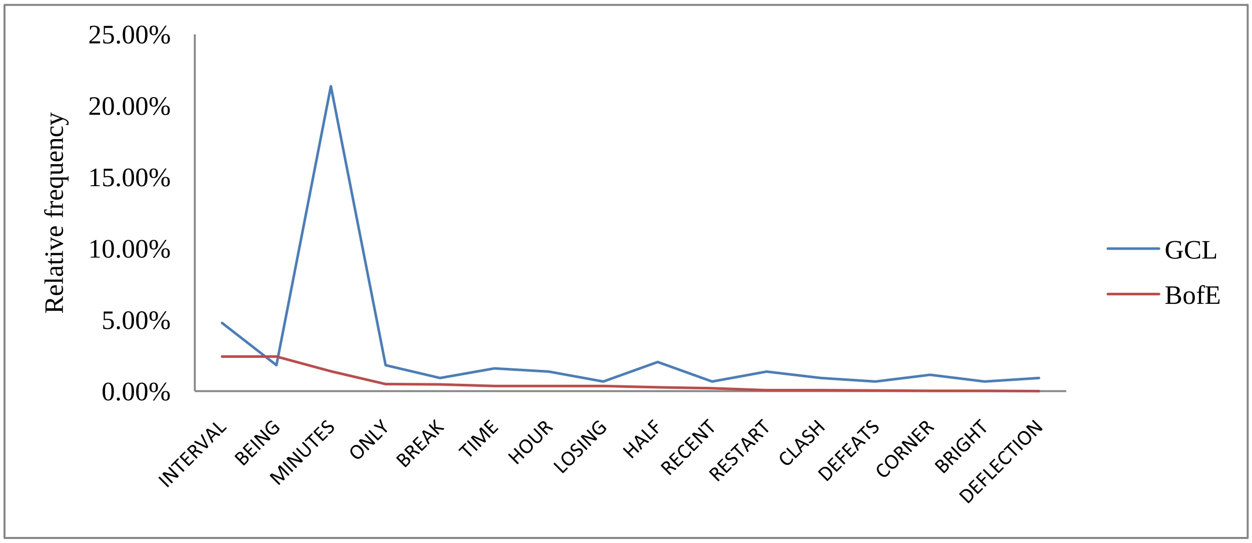
Table 24 below shows a comparison between the GCL and BofE of the relative frequencies (n/N*100%) of these right-side collocates.

Table 24: Comparison of collocates of <u>after</u> by relative frequency (span = 0:3, n≥2)

Guardian Champions League Corpus			Bank of English Corpus	
N = 440			N = 681330	
COLLOCATE	n	%	n	%
MINUTES	94	21.36%	9390	1.38%
INTERVAL	21	4.77%	16408	2.41%
HALF	9	2.05%	1851	0.27%
BEING	8	1.82%	16408	2.41%
ONLY	8	1.82%	3266	0.48%
TIME	7	1.59%	2940	0.36%
RESTART	6	1.36%	447	0.07%
HOUR	6	1.36%	2462	0.36%
CORNER	5	1.14%	104	0.02%
CLASH	4	0.91%	398	0.06%
DEFLECTION	4	0.91%	7	0.00%
BREAK	4	0.91%	3231	0.47%
BRIGHT	3	0.68%	80	0.01%
DEFEATS	3	0.68%	356	0.05%
LOSING	3	0.68%	2365	0.35%
RECENT	3	0.68%	1319	0.19%

The striking exception again is ‘minutes’ which is significantly more than twenty percentage points higher in the GCL, clearly seen in Figure 9:

Figure 9: Comparison of collocates of <u>after</u> by relative frequency (span = 0:3, n≥2)



4.4.2 ‘from’

From occurs within the GCL in 156 texts (98.73%) with a total frequency of N = 880 compared to N = 1920773 within the BofE. The top collocates in Table 25 reflect the action of the match, with <save from> and <header from>, for example, clearly indicating the game of football.

Table 25: Top GCL collocates of from by log-likelihood (span = 1:0, LL≥15.13)

Guardian Champions League Corpus		
N = 880		
COLLOCATE	LL score	n
SAVE	131.7410889	22
HEADER	78.76798248	16
MINUTES	62.64314651	23
APART	55.62653732	7
SHOT	44.18400192	16
GOALS	38.28239822	12
KICK	38.10361099	12
TACKLE	32.57513809	6
DRIVE	32.02467728	8
CROSS	29.46236229	10
YARDS	29.41236115	8
EFFORT	29.17728233	7
PRESSURE	28.2676754	7
PASS	26.16972733	9
DRAINED	24.55901146	3
RECOVERED	24.41725159	4
NEWS	23.57733727	4
BALL	23.17192268	13
CORNER	22.88850403	8
CHANGES	22.8212204	4
RUN	20.34476089	6
SCORED	19.62203979	7
SANCTION	19.35893631	2
REHABILITATION	19.35893631	2
ELIMINATION	19.35893631	2
UNRECOGNISABLE	19.35893631	2
DELIVERY	18.53606415	3
ASIDE	18.53606415	3
REBOUNDED	17.67966652	3
THEM	17.33353806	9
BAR	17.01677132	5
CAME	16.97137451	7
BACK	16.55997467	10
POINT	16.51265144	6
TUMBLE	15.55571556	2

The top BofE collocates (Appendix 5) seem to form a range of patterns, mostly combinations (often termed 'colligation') such as *ADJECTIVE PREPOSITION* (*suffering from* or *different from*) or *VERB PREPOSITION* (*come from* or *benefit from*). With the exception of <apart from>, however, from within the GCL seems instead to predominantly form Noun postmodification through Qualifiers, with some Circumstances. These also seem to be the result of grammatical metaphor, the significance of which will be discussed further in Chapter 5.

Table 26 below shows right-side collocates at a span of 0:3. This also indicates the particular field of the GCL. Whereas the BofE is mainly concerned with specific real-world locations, the collocation of from within the GCL is formed around several collocation patterns of <from + (field position)>, <from + (distance)>, <from + (set piece)>, or <from + (player/team)>.

Table 26: Top collocates of from by log-likelihood (span = 0:3, LL≥15.13)

Guardian Champions League Corpus			Bank of English Corpus		
N = 880			N = 1920773		
COLLOCATE	LL score	n	COLLOCATE	LL score	n
YARDS	288.5359497	39	OTHER	35262.5363	16758
RANGE	137.5177765	20	TIME	35252.1373	18150
ANGLE	95.77308655	12	1	16672.6765	10199
CLOSE	93.8724823	18	HOME	32893.5848	11874
INJURY	82.26631165	16	NATIONAL	23234.7858	8350
RIGHT	69.56195831	18	ITS	30050.6319	16932
LEFT	65.42538452	19	LONDON	26064.95	9305
EDGE	56.76322174	10	ALL	19413.5516	17242
CORNER	55.49393463	14	OUTSIDE	23474.4981	6143
CROSS	50.44950485	14	ONE	33404.9074	23043
HOME	43.95683289	15	POINT	23836.502	7652
SET	41.73886108	11	PUBLIC	25479.3733	8703

Table 26 (cont.): Top collocates of from by log-likelihood (span = 0:3, LL≥15.13)

Guardian Champions League Corpus			Bank of English Corpus		
N = 880			N = 1920773		
COLLOCATE	LL score	n	COLLOCATE	LL score	n
PIECE	40.27009201	6	KUWAIT	22030.6495	3925
STANDS	40.07041168	5	COUNTRIES	15546.7495	4711
SPOT	32.78650284	7	YARDS	22515.0647	3675
SIX	29.6529541	8	SOUTH	19947.9013	6891
TIME	28.5623436	13	RADIO	22561.0662	5703
DISTANCE	27.70001221	4	PERSPECTIVE	15164.6165	2559
FREE	27.27223015	9	SIDE	16436.618	6147
POSITION	25.13835716	6	BEING	19237.3384	9771
GROUP	23.60387421	11	SCHOOL	17375.3389	6481
PENALTY	22.74174309	8	WEST	16061.9028	5714
END	22.34193611	7	NEW	16744.8199	12712
PIECES	22.13378525	4	NORTH	15130.9108	5004
HALFWAY	20.77161789	3	POUNDS	15051.3209	5752
BYLINE	20.77161789	3	MR	14973.0855	10254
OUTSET	19.54451561	3	BEGINNING	14788.1904	3713
RESULTING	19.35893631	2	UNIVERSITY	14510.8215	4779
PORTO	19.01646614	6	WASHINGTON	13945.9858	4139
SUBSTITUTE	18.84011459	6	EAST	13891.092	4723
MIDFIELD	18.32886887	6	MOSCOW	13697.3271	3204
INSIDE	17.22341537	5	OFFICE	13624.6398	4853
LINE	16.2402935	6	SOURCES	12987.2466	2899
APPEARING	15.55571556	2	WORLD	12682.57	8369
ACUTE	15.55571556	2	CITY	11985.1226	5371
VISITING	15.15499783	3	SCRATCH	11825.1214	1649

4.4.3 ‘was’

Was occurs in 100% of texts with a total frequency of N = 1649 within the GCL and N = 3246851 within the BofE. The left-side collocates of the GCL in Table 27 reveal only two collocations as defined as a significant relationship containing at least one lexical word, <performance was> and <riposte was>.

Table 27: Top GCL collocates of was by log-likelihood (span = 1:0, LL≥15.13)

Guardian Champions League Corpus		
N = 1649		
COLLOCATE	LL score	n
IT	1379.925903	288
THERE	909.1323853	159
THIS	290.1841125	86
HE	286.6390686	105
THAT	55.49330902	56
IN	44.47477722	3
WHO	32.15494156	22
PERFORMANCE	20.28910637	9
RIPOSTE	18.59862518	3
I	17.81580544	13

It also does indicate, however, extensive use of pattern phrases, such as:

It was	only when (person) had (done) only in only a week earlier not until (n) minutes had elapsed as well for (team)	that

This also reveals the presence of a metaphor of FOOTBALL AS PERFORMANCE, through collocates such as:

night>
performance>
display>
<This was + game>
fixture>
match>
occasion>

This will be discussed further in Chapter 5.

Looking in more detail at right-side collocates, the BofE (Appendix 6) deals mainly with reporting factual events, such as <was born>, <was told>, <was arrested> or <was killed>, possibly reflecting the high use of passives in general newspaper registers (Kennedy, 1998). For the GCL in Table 28 below, however, the evaluative tenor of the GCL seems much more evident, through collocations such as <was terrific>, <was unimpressed>, <was easy> or <was obviously>. The field of the GCL is also indicated through such collocates as <was substituted>, <was replaced> or <was penalised>.

Table 28: Top GCL collocates of was by log-likelihood (span = 0:1, LL≥15.13)

Guardian Champions League Corpus		
N = 1649		
COLLOCATE	LL score	n
ENTITLED	41.87218475	8
BLOCKED	41.87218475	8
CAUGHT	34.3274765	7
SUBSTITUTED	33.80509567	5
TERRIFIC	33.80509567	5
CALLED	29.86663246	7
UNIMPRESSED	28.72034073	4
UNDERMINED	28.40734863	5
UNABLE	28.40734863	5
EASY	22.57437897	5
REPLACED	21.88183784	5
WITHDRAWN	20.80021858	3
FORTUNATE	20.41325951	4
DEEMED	20.41325951	4
DIFFICULT	19.34308624	6
LACED	18.59862518	3
TELLING	18.59862518	3
REWARDED	18.59862518	3
HEADED	18.47904968	7
PLENTY	17.9146862	4
ENOUGH	17.54946518	7
BOUND	17.04081917	3
AMPLY	16.84480476	2
RELUCTANT	16.84480476	2
ALARMING	16.84480476	2
SMILING	16.84480476	2
TESTIMONY	16.84480476	2
PENALISED	16.84480476	2
OBVIOUSLY	16.84480476	2
DISRUPTED	16.84480476	2
ANNOUNCED	16.84480476	2
FOILED	16.84480476	2
GIVEN	16.48926926	7
NOTHING	15.87089729	6
PROOF	15.82771587	3
DEFLECTED	15.5320673	4

Expanding the right-side collocates to a span of 0:3, the GCL seems to highlight another metaphorical set of collocates, in this case from LAW, with collocations such as <was proof>,

<was deemed>, <was adjudged> or <was testimony> and a related set concerned with FAIRNESS, such as <was entitled>, <was rewarded> and <was fair>, as shown in Table 29 below.

Table 29: Collocates of was related to LAW and FAIRNESS by log-likelihood (span = 0:3, LL≥15.13)

Guardian Champions League Corpus		
N =	1649	
COLLOCATE	LL score	n
ENTITLED	49.85089111	9
CAUGHT	41.87218475	8
INEVITABLE	33.80509567	5
ANONYMOUS	28.72034073	4
REWARDED	28.72034073	4
GIVEN	25.49587822	9
PROOF	24.22330093	4
FAULT	20.80021858	3
DEEMED	20.41325951	4
TELLING	18.59862518	3
FAIR	17.04081917	3
BOUND	17.04081917	3
PARTY	17.04081917	3
ADJUDGED	16.84480476	2
WRITTEN	16.84480476	2
DISTINCT	16.84480476	2
ANNOUNCED	16.84480476	2
INDICATION	16.84480476	2
TESTIMONY	16.84480476	2
DISRUPTED	16.84480476	2
PENALISED	16.84480476	2
FOILED	16.84480476	2

The significance of this will be discussed further in Chapter 5.

4.4.4 'have'

Although it may be argued that the collocation of was and have is predictively formed grammatically, the choice of process verb collocates, shown in Table 30, may nevertheless reveal a lot about the composition of the corpus. The field of discourse is indicated through the collocations <have won>, <have scored> and <have equalised>. It is also notable for the high number of mental process verbs, such as ‘imagined’, ‘expected’, ‘appreciated’, ‘noted’ and ‘felt’, which may reveal information about the tenor of discourse and the evaluative nature of the GCL as compared to general English.

Table 30: Top GCL collocates of have by log-likelihood (span = 0:1, LL≥15.13)

Guardian Champions League Corpus		
N = 801		
COLLOCATE	LL score	n
BEEN	1319.257202	191
DONE	111.4077377	16
WON	63.99913025	12
SCORED	61.00217438	14
EQUALISED	47.82299042	6
NOW	47.66205597	13
GONE	38.50254822	7
TO	38.37194443	55
MADE	37.83884048	12
TAKEN	37.04005814	8
GOT	34.66749191	8
LOST	32.901577	8
GIVEN	31.57877922	8
RAISED	23.56261444	4
IMAGINED	22.90593338	3
EXPECTED	22.68376732	5
ENJOYED	21.65936279	4
DECIDED	20.10416222	3
NOTCHED	19.73563004	2
DOUBLED	18.23644829	3
RECOVERED	16.83157158	3
APPRECIATED	15.93097591	2
SUFFICED	15.93097591	2
NOTED	15.70604992	3
FELT	15.5228529	4
PUT	15.19390869	6

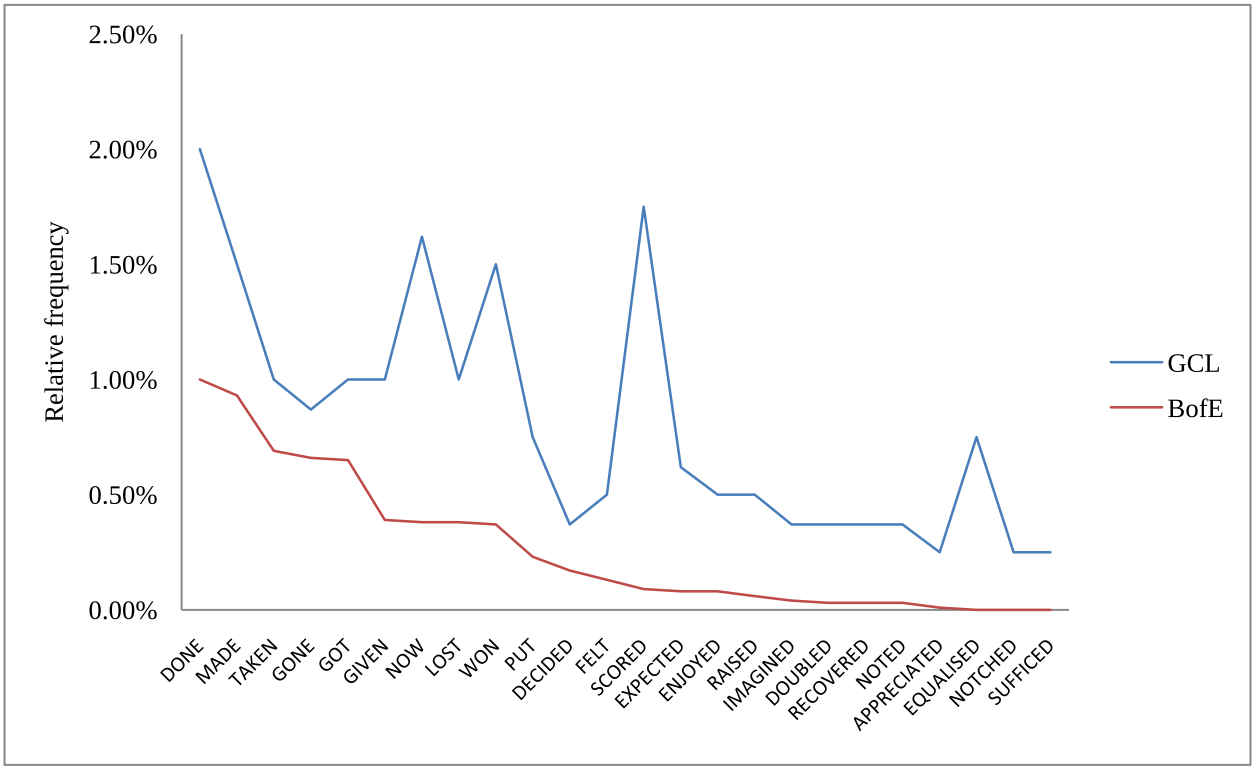
Table 31 shows a comparison with the BofE of this collocation by relative frequency.

Table 31: Comparison of collocation of have by relative frequency (span = 0:1, n≥2)

Guardian Champions League Corpus			Bank of English corpus	
N = 801			N = 2055640	
COLLOCATE	n	%	n	%
DONE	16	2.00%	20602	1.00%
SCORED	14	1.75%	1826	0.09%
WON	12	1.50%	7610	0.37%
MADE	12	1.50%	19031	0.93%
TAKEN	8	1.00%	14084	0.69%
GOT	8	1.00%	13270	0.65%
LOST	8	1.00%	7721	0.38%
GIVEN	8	1.00%	8091	0.39%
GONE	7	0.87%	13611	0.66%
EQUALISED	6	0.75%	77	0.00%
EXPECTED	5	0.62%	1664	0.08%
RAISED	4	0.50%	1290	0.06%
ENJOYED	4	0.50%	1591	0.08%
FELT	4	0.50%	2639	0.13%
IMAGINED	3	0.37%	817	0.04%
DECIDED	3	0.37%	3550	0.17%
DOUBLED	3	0.37%	554	0.03%
RECOVERED	3	0.37%	596	0.03%
NOTED	3	0.37%	660	0.03%
NOTCHED	2	0.25%	72	0.00%
APPRECIATED	2	0.25%	244	0.01%
SUFFICED	2	0.25%	53	0.00%
PUT	6	0.75%	4756	0.23%

There is some apparent correlation in the patterning between the two corpora. However, the collocations <have lost>, <have won>, <have scored> and <have equalised> mentioned above significantly stand out, seen more clearly in Figure 10:

Figure 10: Comparison of collocation of <u>have</u> by relative frequency (span = 0:1, n≥2)



Turning to left-side collocates in Table 32 below, the GCL is notable for the high use of modality (particularly those relatively rare, such as ‘ought’), whereas the BofE has higher scores for pronouns. This could again be a reflection of the more evaluative tenor of the GCL, which is also apparent through collocations such as <seemed + <u>have</u>>, <claimed + <u>have</u>> or <deemed + <u>have</u>>.

Table 32: Top collocates of <u>have</u> by log-likelihood (span = 3:0, LL≥15.13)

Guardian Champions League Corpus			Bank of English Corpus		
N = 801			N = 2055640		
COLLOCATE	LL score	n	COLLOCATE	LL score	n
WOULD	745.8592529	106	WE	861278.221	178973
MIGHT	719.9086304	87	WOULD	787104.418	148854
SHOULD	430.8342285	59	I	675444.034	198931
COULD	425.4238281	73	YOU	619288.07	163383
THEY	299.9582825	75	THEY	584808.522	146964
MUST	248.9338074	35	TO	302268.223	214420
WILL	241.5363312	55	AND	202631.368	174098
WE	234.5829163	50	THAT	220220.866	117742
MAY	174.8787384	29	MAY	276686.141	58836
OUGHT	144.5646057	17	T	251207.057	74967
YOU	113.5887146	21	COULD	257385.612	58499
NOT	108.2596359	40	WILL	235145.949	72000
THAT	94.19998932	49	MIGHT	247565.743	42838
HE	88.80767059	40	WHO	197199.788	61302
AND	72.01955414	61	SHOULD	201344.571	43139
I	62.89486694	19	MUST	162293.915	31632
IT	60.61651993	34	NOT	123869.385	54831
TO	60.22462082	65	PEOPLE	117611.74	36166
BUT	53.51572418	33	HE	78011.7013	55799
SIDE	50.81110764	17	DON	101286.092	27151
WHO	46.11138916	19	BUT	73242.39	46480
SURELY	35.00362396	6	WHICH	82167.9251	36724
TEAM	32.59581375	15	LL	81057.6731	19174
FEW	31.32825279	8	DO	66360.0171	26037
STILL	30.88377571	10	IF	51291.9451	26700
DID	30.2420311	12	SAID	49211.9291	28041
THIS	27.90800858	17	WHAT	47822.1711	24376
NOW	26.54616547	9	D	52338.897	16864
EASILY	25.53181458	5	DIDN	53257.9718	13449
SEEMED	24.45653534	6	NOW	44162.1105	20004
WHAT	21.40304184	6	MANY	43990.8108	16290
CAN	19.87962914	8	SEEMS	51171.5505	10871
NEUTRALS	19.73563004	2	THERE	33469.9394	23423
WHEN	17.77924728	14	DIDN	39787.2096	14580
CLAIMED	17.49082947	3	ALL	28530.7019	21276
DEEMED	17.49082947	3	DOES	38999.266	11478
GOALS	16.46161079	7	THESE	34868.0259	14442
THEN	16.16732025	8	WOULDN	43373.1083	8373
DON'T	16.02838135	4	CAN	27385.8903	18891
OCCASION	16.02838135	4	SHE	25448.3686	19409
AMERICANS	15.93097591	2	SOME	28728.5085	16125

4.5 Summary

The corpus study thus found several distinctive differences between the specialised GCL and the general BofE that may characterise, or be representative of, the particular sub-register of the GCL. The collocation surrounding the keywords contain several items that, in a sense, flag this sub-register for the (in this case) reader. As well as this, however, there are also several correlations between the two corpora. Comparisons of this kind between a specialised corpus and a large general one may thus prove useful, not just for identifying keywords as is generally the case but also for highlighting those collocations that signal the register. The next sections will go on to discuss the implications of these results further.

Chapter 5 – Discussion and Conclusion

5.1 Introduction

The following section presents a discussion of the results of the corpus comparison of the GCL and the BofE in Chapter 4 above. First, a summary of these results is presented in Section 5.2 followed by a brief discussion on some of the limitations of that study. This is then followed by a discussion of the results in Section 5.4, focusing on the functional role that collocation might play for creating cohesion within the GCL, and finally some of the implications of this for EFL teaching. Again, all examples are from the GCL unless otherwise indicated.

5.2 Statement of findings

There were a number of distinctive differences between the GCL and the BofE in the collocation patterns surrounding the keywords. Several individual collocations seemed to play a key role in the GCL, with considerably higher frequencies than within the BofE. Ideationally, certain distinctive collocations were seen to highlight the content of the GCL, which focused on the movement of the ball:

<gave		
<flicked		
<slotted	+	ball>
<rolled		
<flicked		
<dinked		

And on the field of play:

	(distance)	yards>
		range>
		close>
<from +	(position)	right>
		angle>
		edge>
	(player)	Cardoso>

Other collocates were seen to focus on the general play:

<goals	
<shot	
<ball	from>
<cross	
<save	

Or the result of that play and the game:

	won>
<have	scored>
	equalized>

Many distinctive collocations were also related to ‘time’, as in <after (n) minutes>, <with (n) minutes remaining>, <opening (n) minutes>, <after + restart> and also <after + (action)>. The most significant of these, however, was <in + minute>, which will be discussed further below.

In terms of interpersonal meanings the GCL was also characterised in particular through the use of modality. The otherwise relatively infrequent item ‘ought to have’, for example, is used to a significant degree within the GCL, and the 3rd conditional, as also pointed out by Ghadessy (1988), is a key feature of the genre. These interpersonal meanings were also seen, for example, in the more frequent usage of collocations surrounding the auxiliary was, such as <was terrific>, <was unimpressed>, <was easy> or <was obviously>, in the GCL than the BofE. The high frequency of mental processes collocating with ‘have’, such as <have imagined>, <have expected> or <have decided> was also notable within the GCL.

As mentioned in Chapter 4, a series of lexical metaphors runs throughout the corpus that also helps create interpersonal evaluation. A ‘goal’, for example, can be described as ‘deserved’. While the collocation <goal + deserved> may initially seem an unusual choice it is formed through lexical metaphor. Yet the agent of this is not the player who scored the goal but is, in fact, the ‘first half’ and the match itself is thus projected as a metaphorical agent that may feel and affect judgement of events. This is apparent with left-hand collocates of goal, Tables 33a, 33b and 33c, showing the goal being evaluated in terms of its frequency, manner and significance:

Table 33a: ‘Frequency'

COLLOCATE	Total
OPENING	7
LATE	5
RARER	1
RARE	1
CONCLUDING	1
EARLY	2
MINUTE	2
TIME	2
SEASON	1

Table 33b: ‘Manner'

COLLOCATE	Total
OWN	9
GOOD	3
WITTED	1
STUPID	1
SOFT	1
WONDER	1
BAD	1
FINE	1

Table 33c: ‘Significance'

COLLOCATE	Total
EQUALISING	2
DECISIVE	2
DISPUTED	1
BREAKAWAY	1
SIGNIFICANT	1
WINNING	1
IMPORTANT	1

Another way this seems to operate is within a FOOTBALL IS PERFORMANCE metaphor, in this case there is an *ideal* performance against which the match may be compared. This results in phrases such as those below where a player may be compared to what is expected of the ideal, or ‘more talented’ (L243), striker (L18,40,243), seen ‘showing improvement’ (L97), or the ‘audience’ for the performance ‘concerned’ about ‘boredom’ (L111):

	Concordance
18	Crespo did not look as well acquainted with the net as <might have been expected> for a striker
40	Cole <ought to have equalized> when spotted on the right of the area
97	there was an immediate improvement that <should have seen> Louis Saha equalise
111	a concerned audience who <must have imagined> that boredom was the main danger
243	a more talented forward would <surely have scored>

This also results in a metaphor of FAIRNESS where the matches are described in terms of ‘opportunities’ and ‘chances’, and goals are ‘awarded’ to which players are ‘entitled’.

Another possible interpersonal metaphor, GOAL IS A PRECIOUS OBJECT, for example, produces left hand collocates of goal in terms of number, possession and appearance, as seen in Tables 34a, 34b and 34c respectively:

Table 34a: ‘Number'

COLLOCATE	Total
SECOND	20
THIRD	10
FIRST	17
ANOTHER	5
NO	3
ONE	3
THREE	3
EQUALISING	2
FIFTH	2
TWO	2
NINTH	1
SEVENTH	1
SIXTH	1
FOURTH	1
FOUR	1
SOLITARY	1
12TH	1
10TH	1
100TH	1
14TH	1
50TH	1

Table 34b: ‘Possession'

COLLOCATE	Total
OWN	9
HIS	7
ANELKA'S	2
LJUNGBERG'S	2
NISTELROOY'S	2
LEHMANN'S	2
LAMPARD'S	2
HENRY'S	2
ROSICKY'S	1
DANI'S	1
KALOU'S	1
INZAGHI'S	1
MANDANDA'S	1
VIDIC'S	1
GAVRANCIC'S	1
FORLAN'S	1
ARUNA'S	1
DECO'S	1
MY	1
THEIR	1

Table 34c: ‘Appearance'

COLLOCATE	Total
PRETTY	1
BEAUTIFUL	1
STUNNING	1
FINE	1
TRADEMARK	1
GAPING	1

The goal is thus something that may be counted, owned or appreciated visually, but seemingly not, however, something to be done. The right hand collocate <goal for>, shown below, also implies that the goal is something that may be given (c.f. 'This is a present for you'), lost (L5) or fought over (L3, L10):

	Concordance
3	a famed and much-disputed <goal for> Liverpool against Chelsea
5	It was a bad <goal for> us to lose because we knew their danger was from set pieces
7	A <goal for> him then would not have been undeserved
8	a <goal for> Frank Lampard, who failed to connect,
9	an early <goal for> the Ukrainian
10	Robbie Keane's first <goal for> Liverpool but that was as close as Anfield came to witnessing a duel
11	Dutchman has not scored a Premier League <goal for> Liverpool
12	Michael Owen's first <goal for> Real Madrid
13	his first <goal for> the club
14	score his first <goal for> the Stamford Bridge club
15	he grabbed the <goal for> the Scots

The metaphors thus affect the choice of collocate, some of which are shown in Appendix 5. This use of metaphor thus functions as an evaluative device for the corpus that shifts the stance of the author away from a personal appraisal to that of a seemingly impersonal evaluation.

These results seem then to confirm the primary aim of this study that patterns of collocation vary between the specialised genre-specific corpus and a large corpus of general English, and also seem to add some empirical weight to the observation that different genres and registers choose to highlight different collocates (Lewis, 2001). From these differences it also seems apparent that it is the function keywords that most effectively highlight both the content of the GCL and how it differs from the general BofE, adding further evidence to the suggestion that focusing on function, as opposed to content, keywords may be preferable for investigating a particular genre (Gledhill, 1995; Groom, 2005). As such, a comparison with a large corpus such as this may be useful for revealing not just keywords but also their characteristic collocation patterns.

Finally, it must be also noted that a number of similarities between the two corpora were found. The frequencies of collocates surrounding goal, for example, showed a seemingly high correlation. Other similarities were found in the patterning surrounding the items shot and after. As Stubbs (1995) points out, similarities between different corpora may indicate general facts about the language, which would be of great value to any language learner. The next section will now go on to discuss the secondary aim and analyse the functional role of collocation for cohesion in the GCL.

5.3 Limitations of the current study

Before turning to the discussion, however, it is necessary to consider a number of limitations to the study that must be kept in mind. In terms of corpus design, what must be pointed out is that all the texts were taken from a single source. It cannot thus be said that the GCL corpus used here is representative of the WSR genre but is, in fact, a subset only of the Guardian register. As mentioned in Chapter 2, each newspaper reflects the demands and expectations of its readership, and each newspaper has its own style that reflects its ideological stance (Rowe, 1999). The genre, and register, choices are thus affected by the interaction of these factors. The GCL may merely reflect how the Guardian chooses to present sports reports. Care must be taken to view the results in light of this limitation.

It is obviously also the case that the BofE contains various word senses and parts of speech for each of the items analysed. According to the Oxford Dictionary (ODE, 2003), for example, 'shot' has six different word senses, not including phrasal verbs such as 'shot through'. While this is true, the fact that the GCL uses limited senses could also be considered significant, especially in light of the continuing debate over word senses, collocation and genre (Agirre & Edmonds, 2006). It could also be argued that, rather than a different sense, the use of 'shot' within the GCL is metaphorical. Other words related metaphorically to ‘shot’ within the corpus include ‘target’, ‘range’ and, of course, ‘goal’. The fact that certain words are overwhelming used metaphorically within certain domains would be important for any language learner. The study also revealed the presence of both similarities as well as differences between the two corpora. Rather than being limited to identifying keywords only, a comparison with the general BofE, such as this study, may thus be useful in revealing those collocates which are significant within the smaller corpus and those which are present in the language as a whole. Information such as this may also prove useful within EFL for classroom decisions as to which collocations to highlight in a particular course.

Due also to the limitations imposed by the length of this dissertation the following study must thus be rightly seen as a small exploratory investigation rather than an in depth study. The methodology taken here, however, may still reveal significant or interesting differences between the GCL and the larger reference corpus that may then point to or justify further study of the wider patterning within the corpus. As will be shown below collocations within the GCL seem to have specific functional uses that may contrast with their use in general language or spoken language. In terms of foreign language learning it may also demonstrate a useful starting point for learners studying genre through corpora. These points will be discussed further in Section 5.5.

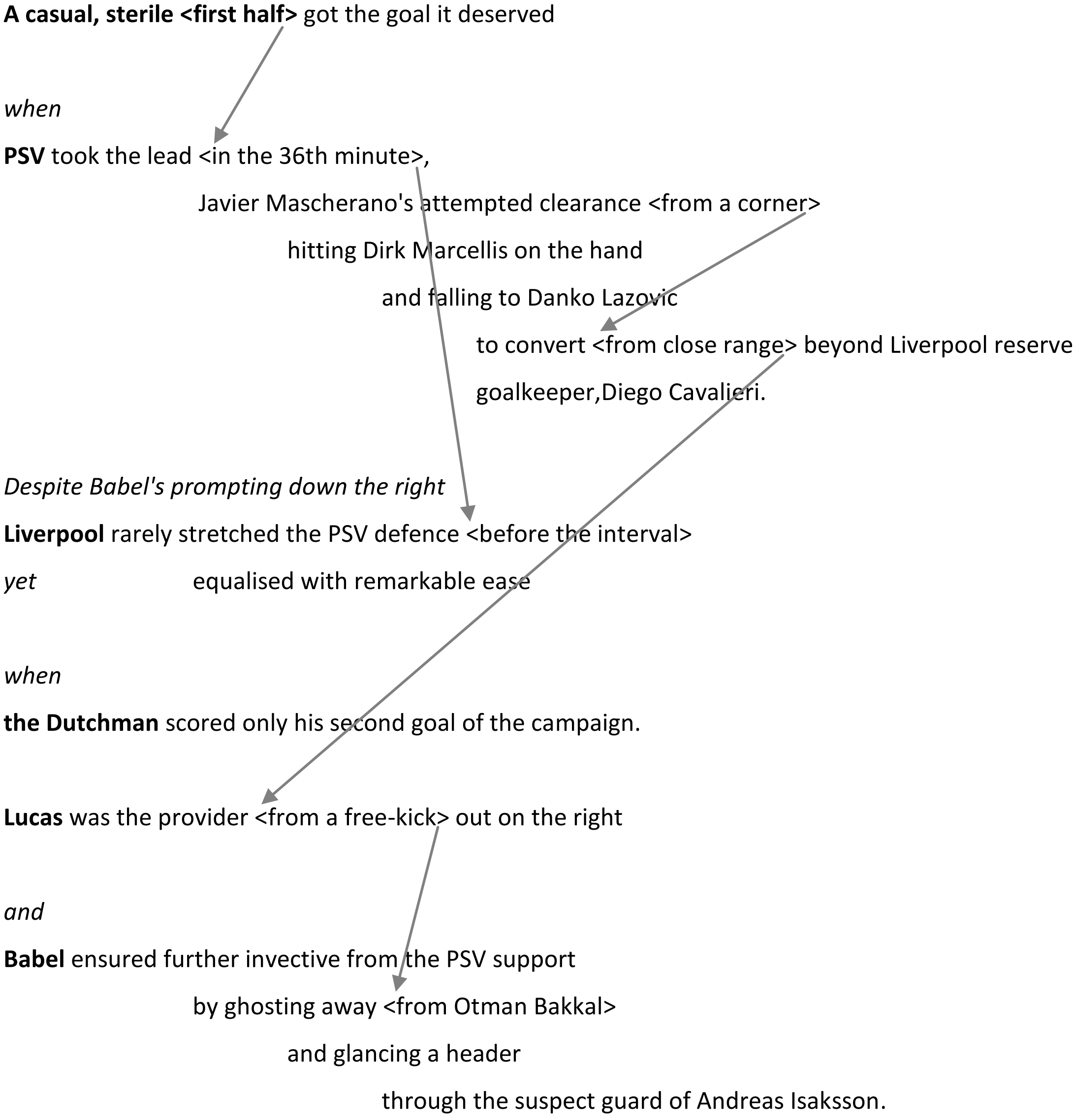
The next section will present a discussion of the results of the corpus study.

5.4 Discussion of results

As Conrad (2002) points out, the identification of significant items within a corpus must also be tied to their functional interpretation. Keywords may be more significant not just numerically but also functionally to organise text. The following section discusses the collocation identified by the quantitative study in terms of its functional properties and attempts to identify the textual function (Gledhill, 2000) for collocation within the corpus. In particular, the discussion will focus on the role of collocation in the creation of cohesion both between and within differing levels of text, register and genre. This will be demonstrated by focusing on the 7th paragraph of the text in Appendix 1.

While the GCL makes use of surface cohesion in the paragraph according to Halliday & Hasan (1976) and Halliday (2004)'s classification (see Appendix 8) this is not the only way the text is organised cohesively. Whereas Ghaddesy (1988) saw the basic sentence pattern of WSR as being Subject – Verb – Object – Adverb – Adverb, it may in fact be more useful to see the WSR as chains of hypotactic expansion. Collocation here serves to provide cohesion between the elements of these chains. This is shown in Figure 11 where cohesion is provided by the semantic link to collocations of match time and collocates of from (following Halliday, 2004, **topical Themes** are in bold and *textual Themes* in italics – hypotaxis is indented).

Figure 11: Theme and cohesion in the 7th paragraph



A casual, sterile <first half> got the goal it deserved

when

PSV took the lead <in the 36th minute>,

Javier Mascherano's attempted clearance <from a corner>

hitting Dirk Marcellis on the hand

and falling to Danko Lazovic

to convert <from close range> beyond Liverpool reserve goalkeeper,Diego Cavalieri.

Despite Babel's prompting down the right

Liverpool rarely stretched the PSV defence <before the interval>

yet equalised with remarkable ease

when

the Dutchman scored only his second goal of the campaign.

Lucas was the provider <from a free-kick> out on the right

and

Babel ensured further invective from the PSV support

by ghosting away <from Otman Bakkal>

and glancing a header

through the suspect guard of Andreas Isaksson.

Collocation thus seems to provide additional surface cohesion within the text through semantic reiteration and repetition. Collocation, however, does not only contribute to surface cohesion within a text. The suggestion will be made here that cohesion may be divided into cohesion within texts, or what will be termed 'intratextual' cohesion and an additional cohesion that functions between texts in the genre, or 'intertextual' cohesion. The cohesion provided by from above is dependent for its interpretation on the limited functional role of that item within the corpus as opposed to that within general language.

One way in which the corpus is organized textually is through choices of Theme. Choice of Theme, the "point of departure of the message" (Halliday, 2004, p. 64), can be an important indicator of register. Ghadessy (1998), for example, found that tagging a corpus for Theme more accurately characterised text type than Biber's (1995) multi-variant analysis. Within the GCL, collocation seems to play a key role in signalling these choices. Comparing <shot wide> and <shot from> in Table 35 below, it can be seen that in the former case the player is more clearly placed as the choice of Theme whereas in the former the shot itself is the main Theme.

Table 35: Theme/Rheme comparison of <shot wide> and <shot from>

Theme	Rheme
A deflected <shot from> Cardozo in the 73rd minute	squeezed past the far post
Shevchenko	skidded a <shot wide> of the post

This can also be seen with the collocation <n minutes later>. At a clause-final position <n minutes later> signals a goal being scored, especially one that equalises the scores, shown with double underline:

Concordance

34 Artmedia were back on level terms <five minutes later>.

35 Edu gave the ball away in midfield <eight minutes later>.

36 Kiev levelled though Sergei Fedorov <nine minutes later>.

25 Sutton's well-taken equaliser <five minutes later>.

26 A superb Steven Gerrard header restored hope for Liverpool <three minutes later>

31 Besiktas assumed the lead and then doubled it <four minutes later>

In the initial Theme position, however, <minutes later> mainly signals the action of general play with a ‘throw-in’, a ‘cross’ or a ‘save’ among others, shown with double underline:

Concordance

2 <Five minutes later>, with Ibrox rocking, substitute Peter Lovenkrands set off on a r

5 <five minutes later> they launched another move into the Lyon penalty area. This time

7 <Eight minutes later> Sutton took a throw-in to Larsson, who again threw in a trick an

8 <Two minutes later>, Kily Gonzalez, a former Valencia player under Cuper, showed why

9 <Two minutes later> Camara, moving swiftly on to a pass from McGeady, was chopped do

11 <Five minutes later> the German was called into an acrobatic save, tipping Carvalho's

12 <Five minutes later>, Lampard switched play to the right and Wright-Phillips, crossin

13 <Ten minutes later> the Czech reacted marvellously to push the ball over after a cor

14 <Two minutes later> a stunned Celtic side were two goals down when Cristiano Lucarel

19 <Seven minutes later> Frank de Bleeckere awarded a questionable free-kick against Fabi

20 <five minutes later>, Cihan Haspolatli sliced a clearance which was gathered by Jerma

22 <Four minutes later>, Paletta headed Marcelo Carrusca's corner high into the air and

27 <Seven minutes later> Jiri Jarosik set up Scott Brown at the edge of the area but his

28 <Five minutes later> Ronaldinho seized possession 40 yards out and drilled a diagonal

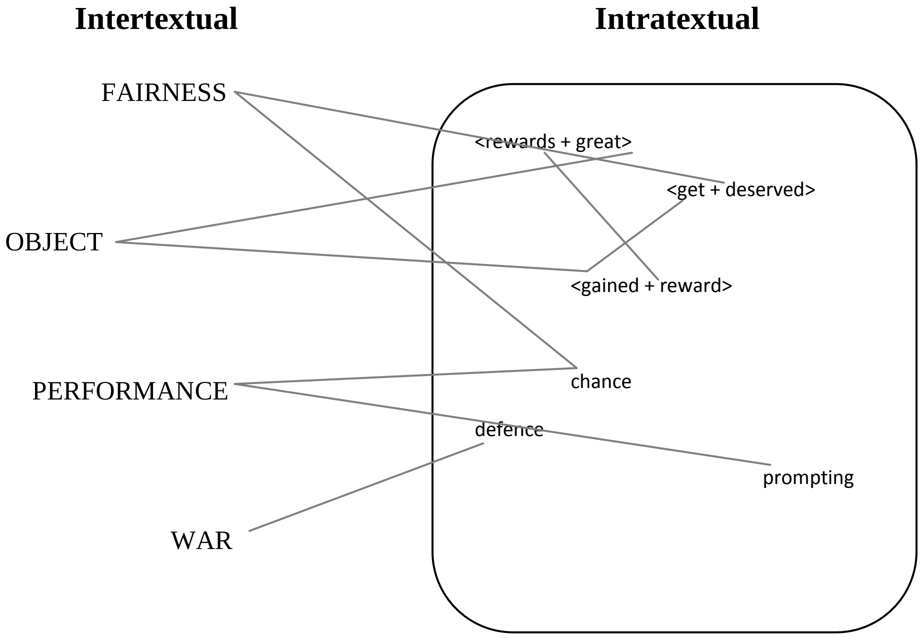
29 <two minutes later> a Daniel Jensen cross hit the bar with Valdes looking complacent

32 <three minutes later> Stanislav Varga's clearing header landed perfectly for Makaay to

The choice of collocation thus seems to be a marker of textual organisation within the GCL.

As described above, a series of metaphors runs throughout the corpus. While these collocations may not be individually significant, their interpretation is dependent upon this repeated functional role within the corpus and, taken collectively, they do seem to produce ties that contribute to intertextual cohesiveness, which we can see in an analysis of the 7th paragraph in Figure 12.

Figure 12: Intratextual and intertextual cohesion in the 7th paragraph



Collocation within the GCL appears to act as what McCarthy (1990) terms “lexical signals”, only instead of signalling textual discourse they seem to aid in producing cohesive ties across texts to signal the various generic stages. As shown in Chapter 4, for example, the collocation <in + minute> was highly significant within the GCL, yet it does not seem to add information to the text itself. As Martin & Rose (2005) point out, a text has its own internal logic and progression of time, which would seem to make rendering such overt expressions unnecessary. The collocation <in + minute> seems to function instead to signal generic Moves and other texts. Table 36, shows that it is seemingly associated with the [Event] move of the genre as, compared to the collocation surrounding other word-forms within this study, it is relatively free of evaluative lexis and metaphor that characterises the other Moves.

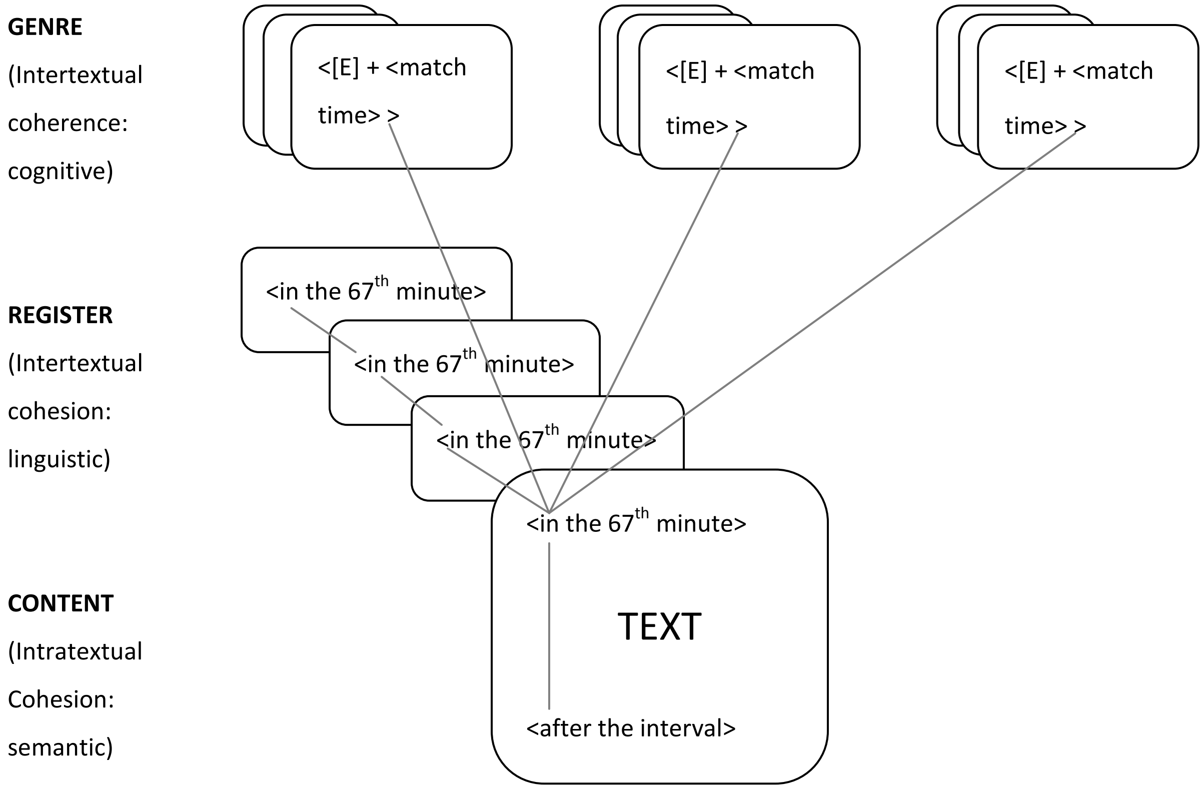
Table 36: Lexical collocates of 'in the nth minute' (span = 3:3, LL≥15.13)

Guardian Champions League Corpus		
N = 126		
COLLOCATE	LL score	n
LEAD	30.2848077	5
CAME	27.30320843	5
BALL	23.33693305	6
FREE	20.62075291	4
CROSS	19.43590218	4
KICK	18.93714346	4
SAVE	16.68152108	3

It seems possible then that, within the GCL at least, collocation also contributes to generic coherence of a text by providing certain recurring patterns that, in a sense, 'flag' the Moves for the reader and thus aid in following the text.

Out of the open-choice lexical choices provided within the lexico-grammar, collocation may thus provide cohesion by creating closed-set collocation items conditioned by generic constraints. The GCL is also, however, a sub-register of the WSR genre. As such, the collocation <in the (nth) minute>, for example, appears to function to produce cohesive ties on three different levels of context simultaneously: semantically within text, across texts within the sub-register and to signal the generic stage of [Event]. This may be represented as in Figure 13.

Figure 13: Representation of intratextual and intertextual cohesion



In summary, it appears that collocation within the GCL corpus contributes to the creation of cohesive ties for both intratextual and intertextual cohesion. The next section will now discuss the implications of this for general EFL classes.

5.5 Implications for EFL teaching

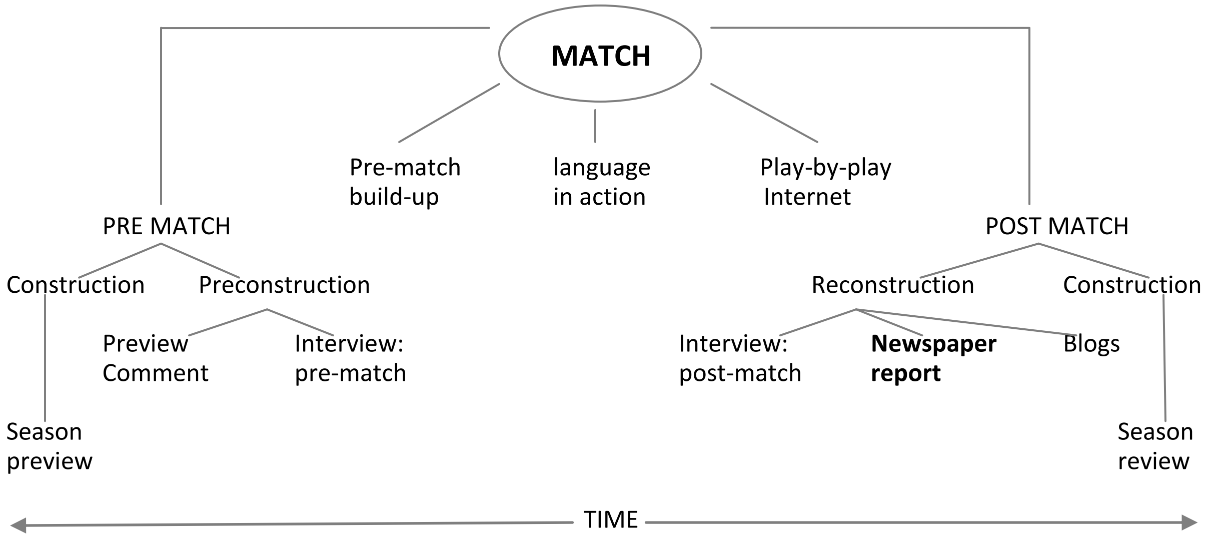
From the results of this study, there are a number of points to be made that may have implications for wider foreign language teaching. The first point relates to the teaching of collocation in general. As mentioned in Chapter 2, the two main difficulties for the teaching of collocation is the sheer number of combinations available and their seeming arbitrariness. However as hopefully shown in this study, while some of these collocations have general application throughout the language, others have more specific ideational, interpersonal and textual functions within particular genres and registers. As such, rather than treating 'collocation' as an overarching concept (informing language learners that a phrase is 'collocation' rather than 'because we say it like that' (Lewis, 2001) does not seem much more helpful), it may be more beneficial to demonstrate the function of collocation for cohesion at different levels of genre, register and text.

It also appeared from this study that the influence of metaphor is important for the choice of collocation. While individual collocates may not be statistically significant they may form part of a metaphorical or semantic set that creates cohesion. The collocation formed through the influence of metaphor is also a product of the culture in which it is situated. Wierzbicka (2006), for example, describes how the concept of ‘painess’ is, in fact, specifically a product of an English-speaking culture. The choice of metaphor for the register can thus only come from those made available from the culture, or system of genre. For learners of a language, especially those EFL students who may not be familiar with English metaphoric systems, awareness of this would prove valuable in demonstrating how collocation operates within an intertextual context across sets of texts.

The next point to be made is that collocation functions within texts, is a “textual phenomenon” (Hoey, 1991, p. 219), and texts do not exist in isolation. While this may seem an obvious point it is one that is often overlooked within EFL pedagogy and one that has a number of implications. The first is that a newspaper is also a social artefact and, originally, is a specific product of British society (Rowe, 1999). As such, the role of newspaper, and the communicative purposes of the genres within it, may differ from society to society and from register to register. Romaine (1994), for example, has traced the emergence of a genre of sports report within a Tok Pisin newspaper that has a possibly different communicative role to that of a British newspaper. As briefly outlined in Chapter 2, the communicative purpose of a text is also a product of the context of culture that allows certain responses to certain genres. Learning collocation thus involves not only learning its form but also its communicative function within the text and the place of that text within the culture.

Corollary with this point is that texts are found within the culture in sets (Hyland, 2007). A football match exists within a set of interconnected genres, including newspaper reports. This may be represented in Figure 14. The newspaper sports section also contains a number of social genres, including lists of results, reports, commentaries, previews, or news, among others, each of which may subsequently vary according to the sub-register of the particular newspaper.

Figure 14: Possible schema of genre set for footbal match



(Adapted from Martin, 2001)

The value in using more general texts, such as sports reports, for within EFL classrooms may lie not in being able to reproduce those texts, as is often the case within specialised ESP or EAP classes (Osman, 2004), but to demonstrate how they interact and fit into the system of language and culture, to help students "see the assumptions and values which are implicit in those genres and [help] them understand the relationships and interests in that context" (Hyland, 2007, p. 156). Knowledge of this textual role could be particularly valuable for EFL learners attempting to navigate between possibly unfamiliar genres (Mahlberg, 2006).

It may be useful then, in terms of EFL pedagogy, to suggest a distinction between 'teaching genres' on the one hand, which would be more appropriate for specialist ESP or EAP classes, and 'genre teaching' on the other, which utilises the insights gained from studies such this to teach language as a system of choice (Plum, 2006) rather than merely a sentence-level, rule-generated one. Collocation demonstrates how these choices vary through texts and genre, and the fact that local choices may also be the result of systemic constraints. It may also demonstrate the manner in which collocation creates both intratextual cohesion within texts and intertextual cohesion through ties to other texts. Taking these points into consideration within the EFL classroom may allow learners to more systematically study collocation and, more importantly perhaps, it may also provide a means of being able to *predictively* study collocation and allow for more autonomous learning. The next part will go on to suggest further research for collocation, corpus studies, genre and classroom practice.

5.6 Suggestions for further research

In terms of collocation research, this study seems to confirm Halliday’s (2004) assertion that the same words possibly collocate in different ways influenced by register and genre, as mentioned in Section 2.2.1. As also pointed out in Chapter 2, however, most research on collocation within this area has focused on the ideational and interpersonal factors which affect variation (Hunston, 2002). As hopefully demonstrated with this study, however, collocation also plays an important textual role for the creation of cohesion yet more research is needed into exactly how this operates. While the other textual resources for cohesion, outlined in Chapter 2, may be readily identified this contextual variation of collocation suggests that the cohesive ties produced by collocation are genre specific (Mahlberg, 2006) and that differing genres produce these ties through differing ways. This results in the difficulties in identifying these ties within large general language corpora and provides some explanation as to the reasons why they have so far been generally overlooked or excluded. The role of collocation for signalling textual relations and the way in which this relates to other lexical signalling devices, such as those described in Chapter 2, suggests an important role for genre-based corpus research. Collocation also seems to aid in the identification of Moves and generic coherence. Whether this is merely a feature of this corpus or may be found in other genres, and the effects of this in aiding reading comprehension, requires more research.

This generic variation may to a certain degree also help explain the difficulties in both defining and categorising ‘collocation’ itself. ‘Collocation’ was defined within Chapter 2 as significant combinations involving at least one content item and proceeded from this one underlying assumption, yet it may be that this definition only applies to some collocations and that others are formed through different criteria. Rather than employing one definition or another therefore, it may in fact prove of more use to consider ‘collocation’ as comprised of a configuration of several competing factors. As mentioned in Section 2.2.2, Nation (2001), for example, suggests the use of a set of ten scales to classify collocation while recently Frath & Gledhill (2005) suggested a referential conception of collocation. This study also suggests that functional and contextual factors must also be taken into consideration. More research is needed into whether there exists any relationship, or interaction, between these competing factors.

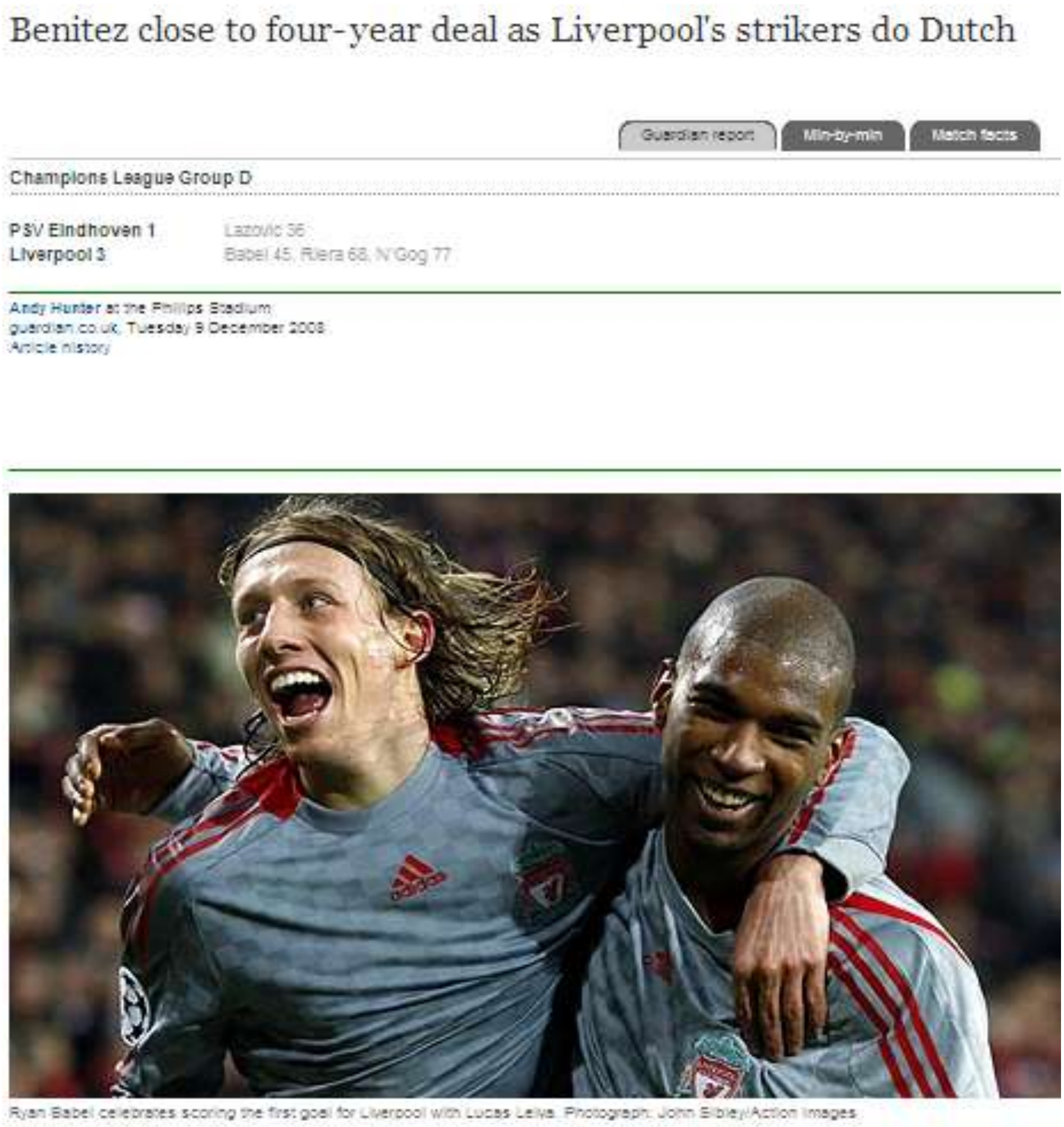
As described in Chapter 4, the use of possessives seems a characteristic feature of this corpus. This use of possessives was also a part of the system of metaphor used extensively throughout the GCL. For collocation and corpus studies, while the role of metaphor for the creation of interpersonal meanings has been extensively researched, more research is thus needed into how metaphor varies between and is influenced by genre. Skorczynska & Deignan (2006), for example, demonstrate how metaphors vary according to genre of economics texts, between general and specialist readerships, for the same topic. How this, in turn, creates cohesion across larger stretches of text also suggests further research opportunities.

As mentioned above in Chapter 2, discussion of the classroom application of genre has tended to focus on ESP and EAP classes, with a general, though understandable, pre-occupation with the writing process (Johns A. , 2002). As this study has hopefully shown, however, the genre approach of situating language choice within particular contexts and the role of collocation in signalling those choices may also have application within more general EFL classes. More research is needed, however, into specific classroom pedagogical practices and, importantly, the effects these practices may have on language acquisition. For learners of a language even broad information, such as the fact that a particular word collocates more strongly with the left-side than the right, may prove useful. Within the GCL for example, the right-side collocates of ‘goal’ do not seem significant yet for ‘ball’ they do. More research is required into the most effective ways of incorporating this information into classroom materials and practices, be it through explicit instruction, consciousness-raising, or extensive reading.

5.7 Conclusion

This dissertation conducted a comparative corpus study of collocation between a very large corpus and a small genre-specific one. The study found that, while there were similarities in the collocation patterns between the two corpora, several key collocations within the specialised corpus seemed to signal important differences between them. The study went on to investigate the functional role of these key collocations for the creation of textual cohesion within the genre, with the suggestion of a distinction between 'intratextual cohesion', that within a particular text, and 'intertextual cohesion', or that created between texts within the register and genre. The suggestion was also made that, for the study of collocation, language learners may be better served by taking a 'genre teaching' approach, focusing not only on the surface forms and patterns of individual collocations but also their functional role for cohesion within text, register and genre. In this way, a more systematic and manageable approach to the study of collocation may be achieved.

Appendix 1: Report: PSV-Liverpool, 9th December, 2008



Benitez close to four-year deal as Liverpool's strikers do Dutch

Champions League Group D

PSV Eindhoven 1	Lazovic 36
Liverpool 3	Babel 45, Riera 68, N'Gog 77

Andy Hunter at the Philips Stadium
guardian.co.uk, Tuesday 9 December 2008
Article history

Ryan Babel celebrates scoring the first goal for Liverpool with Lucas Leiva. Photograph: John Sibley/Action Images

The stakes were negligible yet the rewards proved great for Liverpool in Eindhoventonight : first place in Group D secured with ease against PSV and their manager, Rafael Benítez, on the cusp of sealing his long-term future at Anfield.

Sources close to Tom Hicks and George Gillett Jr have confirmed the American owners have agreed in principle to extend Benítez's contract until 2013. From the outset the Liverpool manager's priority has been a significant extension rather than a dramatic increase on his £3.5m-a-year salary, from a deal that has 18 months to run, and so the offer of a new four-and-a-half year contract is unlikely to meet any resistance. Financial bonuses in the package have still to be ironed out, but even without a considerable pay rise the deal will be worth at least £16m to Benítez.

From employers who 13 months ago were touting Jürgen Klinsmann as his successor, and initially wanted to follow the US sports-model of a one year extension for their manager, the offer represents a significant victory for the Spaniard. It also reflects the Liverpool hierarchy's determination to prevent off-field issues distracting from the team's pursuit of the Premier League title. Whether it signals a long-term commitment to Liverpool from Hicks and Gillett themselves, with interest from Dubai cooling, remains to be seen.

As for the task of negotiating first place at the expense of Atlético Madrid, PSV Eindhoven were as accommodating to Liverpool's needs as the Americans have been to Benítez. Already assured of a place in the knock-out phase, Liverpool reshuffled their pack, fell behind and did not exert themselves, but triumphed convincingly on the effort of those with a point to prove.

Robbie Keane returned to the side following his demotion at Blackburn Rovers and produced another tireless display that was short on chances but notably, playing off another striker, generated plenty of opportunity for those around him. Others with a mandate to impress included Ryan Babel, the former Ajax winger who has asked for — and been denied — a loan move back to Amsterdam given his lack of opportunity this term, the Brazilian midfielder Lucas, jeered by Liverpool supporters during his last start against Fulham, and the French striker David Ngog. All three impressed against PSV, although the calibre of their opponents was far weaker than they face in the Premier League.

"We spoke about how important it is to have the second leg of the knockout round at Anfield and I'm pleased we've managed to secure that," the Liverpool manager said. "There were also a lot of positives in the game. Ngog and Lucas are the two who deserve most credit, I think Ngog has a good future here. He is only young but he is clever and has good movement. Babel and Keane also worked hard. Of course strikers always want to score but Robbie worked extremely hard for us up front."

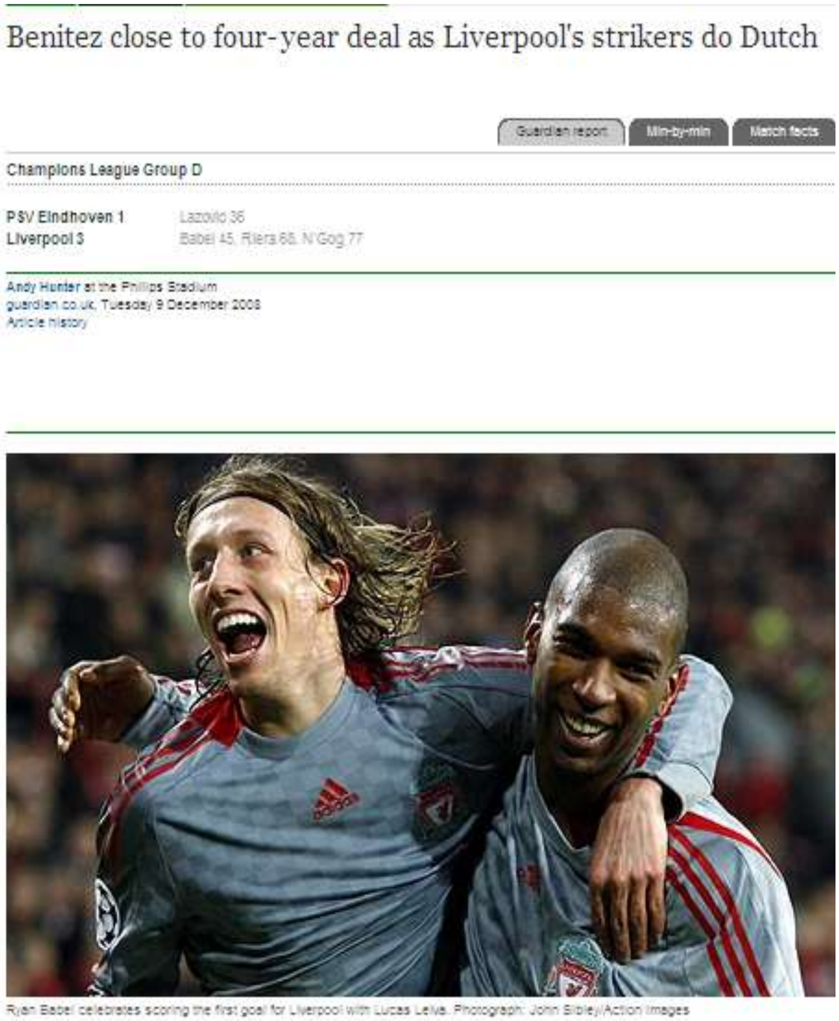
A casual, sterile first half got the goal it deserved when PSV took the lead in the 36th minute, Javier Mascherano's attempted clearance from a corner hitting Dirk Marcellis on the hand and falling to Danko Lazovic to convert from close range beyond Liverpool reserve goalkeeper, Diego Cavalieri. Despite Babel's prompting down the right, Liverpool rarely stretched the PSV defence before the interval yet equalised with remarkable ease when the Dutchman scored only his second goal of the campaign. Lucas was the provider from a free-kick out on the right, and Babel ensured further invective from the PSV support by ghosting away from Otman Bakkal and glancing a header through the suspect guard of Andreas Isaksson.

Liverpool were comfortable throughout the second period; the goals that secured their progress as group winners exquisite. Albert Riera drove the visitors ahead for the first time when invited to attack a retreating PSV defence and obliged with a stunning strike into the top corner from 25 yards. Again, Isaksson should have done better.

Keane had spent most of a frustrating evening shaking his head at errant passes from his team-mates, selfless play that brought no reward and some harsh treatment from the Russian referee, but finally gained the reward his display deserved with his part in Liverpool's third. The Republic of Ireland captain volleyed a superb pass through the PSV rearguard for Ngog to sprint clear and finish expertly under Isaksson, delivering another milestone for Benítez in the process.

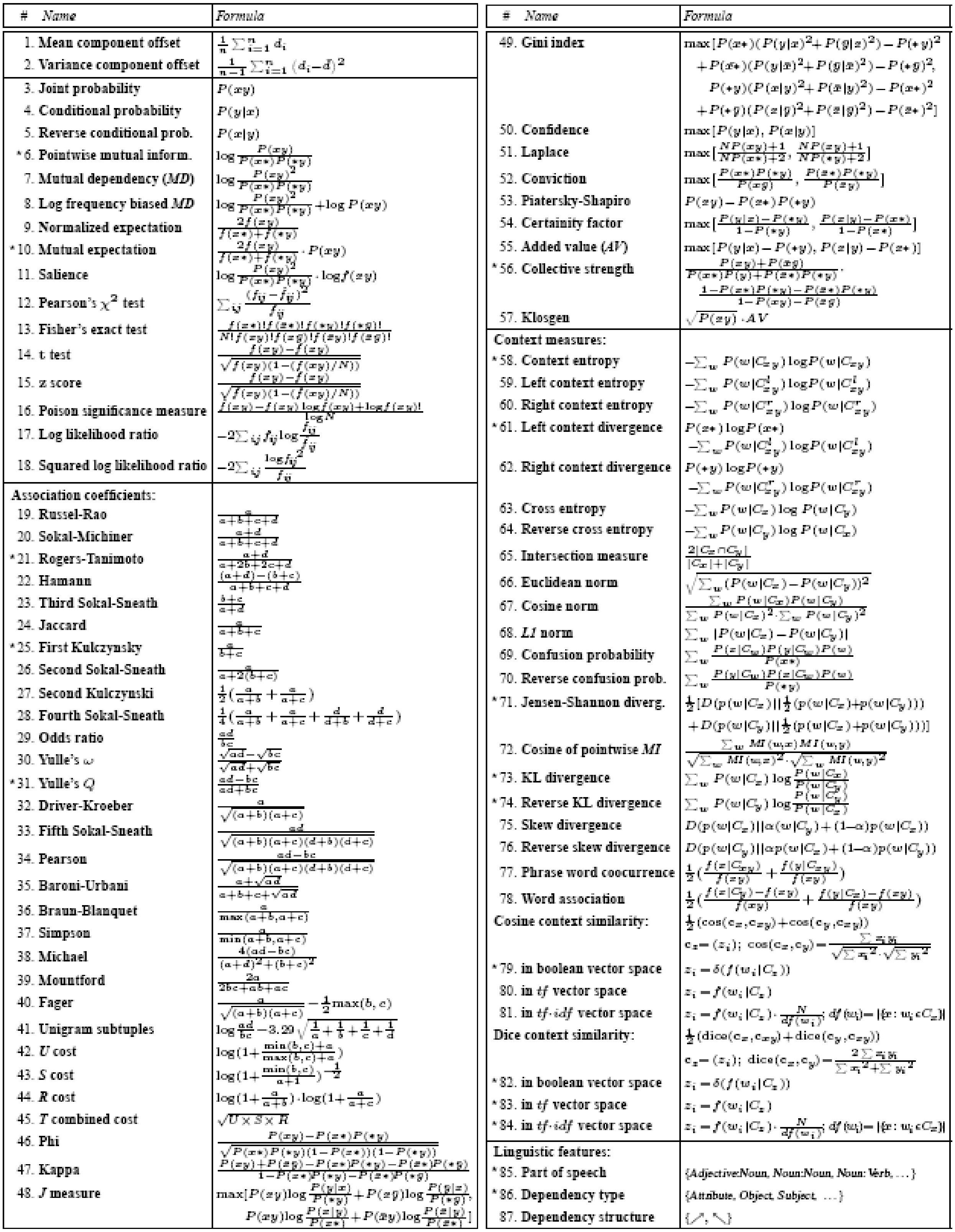
This was Benítez's 40th win in 67 European games as manager, taking him beyond Bob Paisley's record of 39 continental triumphs. The word from the US is that there will be many more.

Appendix 2: Macro-features of written sports report genre



Element	Text
Headline (H)	Benitez close to four-year deal as Liverpool's strikers do Dutch
Intertextual Links (IT)	Guardian report Min-by-min Match facts
Context (C)	Champions League Group D
Result (R)	PSV Eindhoven 1 Lazovic 36 Liverpool 3 Babel 45, Riera 68, N'Gog 77
By-line (B)/ Date (D)	Andy Hunter at the Phillips Stadium guardian.co.uk, Tuesday 9 December 2008 Article history
Picture (P)	
Caption (C)	Ryan Babel celebrates scoring the first goal for Liverpool with Lucas Leiva. Photograph: John Sibley/Action Images
Text (T)	

Appendix 3: Association measures for collocation extraction (Pecina, 2005, p. 15)



# Name	Formula
1. Mean component offset	$\frac{1}{n}\sum_{i=1}^{n} d_i$
2. Variance component offset	$\frac{1}{n-1}\sum_{i=1}^{n}(d_i-\bar{d})^2$
3. Joint probability	$P(xy)$
4. Conditional probability	$P(y\|x)$
5. Reverse conditional prob.	$P(x\|y)$
*6. Pointwise mutual inform.	$\log\frac{P(xy)}{P(x*)P(*y)}$
7. Mutual dependency (*MD*)	$\log\frac{P(xy)^2}{P(x*)P(*y)}$
8. Log frequency biased *MD*	$\log\frac{P(xy)^2}{P(x*)P(*y)}+\log P(xy)$
9. Normalized expectation	$\frac{2f(xy)}{f(x*)+f(*y)}$
*10. Mutual expectation	$\frac{2f(xy)}{f(x*)+f(*y)}\cdot P(xy)$
11. Salience	$\log\frac{P(xy)^2}{P(x*)P(*y)}\cdot\log f(xy)$
12. Pearson's χ^2 test	$\sum_{ij}\frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$
13. Fisher's exact test	$\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$
14. t test	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$
15. z score	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(\hat{f}(xy)/N))}}$
16. Poison significance measure	$\frac{\hat{f}(xy)-f(xy)\log\hat{f}(xy)+\log f(xy)!}{\log N}$
17. Log likelihood ratio	$-2\sum_{ij} f_{ij}\log\frac{f_{ij}}{\hat{f}_{ij}}$
18. Squared log likelihood ratio	$-2\sum_{ij}\frac{\log f_{ij}^2}{\hat{f}_{ij}}$
Association coefficients:	
19. Russel-Rao	$\frac{a}{a+b+c+d}$
20. Sokal-Michiner	$\frac{a+d}{a+b+c+d}$
*21. Rogers-Tanimoto	$\frac{a+d}{a+2b+2c+d}$
22. Hamann	$\frac{(a+d)-(b+c)}{a+b+c+d}$
23. Third Sokal-Sneath	$\frac{b+c}{a+d}$
24. Jaccard	$\frac{a}{a+b+c}$
*25. First Kulczynsky	$\frac{a}{b+c}$
26. Second Sokal-Sneath	$\frac{a}{a+2(b+c)}$
27. Second Kulczynski	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
28. Fourth Sokal-Sneath	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$
29. Odds ratio	$\frac{ad}{bc}$
30. Yulle's ω	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
*31. Yulle's Q	$\frac{ad-bc}{ad+bc}$
32. Driver-Kroeber	$\frac{a}{\sqrt{(a+b)(a+c)}}$
33. Fifth Sokal-Sneath	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
34. Pearson	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
35. Baroni-Urbani	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$
36. Braun-Blanquet	$\frac{a}{\max(a+b,a+c)}$
37. Simpson	$\frac{a}{\min(a+b,a+c)}$
38. Michael	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
39. Mountford	$\frac{2a}{2bc+ab+ac}$
40. Fager	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{1}{2}\max(b,c)$
41. Unigram subtuples	$\log\frac{ad}{bc}-3.29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$
42. *U* cost	$\log(1+\frac{\min(b,c)+a}{\max(b,c)+a})$
43. *S* cost	$\log(1+\frac{\min(b,c)}{a+1})^{-\frac{1}{2}}$
44. *R* cost	$\log(1+\frac{a}{a+b})\cdot\log(1+\frac{a}{a+c})$
45. *T* combined cost	$\sqrt{U\times S\times R}$
46. Phi	$\frac{P(xy)-P(x*)P(*y)}{\sqrt{P(x*)P(*y)(1-P(x*))(1-P(*y))}}$
47. Kappa	$\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$
48. *J* measure	$\max[P(xy)\log\frac{P(y\|x)}{P(*y)}+P(x\bar{y})\log\frac{P(\bar{y}\|x)}{P(*\bar{y})},$ $P(xy)\log\frac{P(x\|y)}{P(x*)}+P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$
49. Gini index	$\max[P(x*)(P(y\|x)^2+P(\bar{y}\|x)^2)-P(*y)^2$ $+P(\bar{x}*)(P(y\|\bar{x})^2+P(\bar{y}\|\bar{x})^2)-P(*\bar{y})^2,$ $P(*y)(P(x\|y)^2+P(\bar{x}\|y)^2)-P(x*)^2$ $+P(*\bar{y})(P(x\|\bar{y})^2+P(\bar{x}\|\bar{y})^2)-P(\bar{x}*)^2]$
50. Confidence	$\max[P(y\|x),\ P(x\|y)]$
51. Laplace	$\max[\frac{NP(xy)+1}{NP(x*)+2},\frac{NP(xy)+1}{NP(*y)+2}]$
52. Conviction	$\max[\frac{P(x*)P(*\bar{y})}{P(x\bar{y})},\frac{P(\bar{x}*)P(*y)}{P(\bar{x}y)}]$
53. Piatersky-Shapiro	$P(xy)-P(x*)P(*y)$
54. Certainity factor	$\max[\frac{P(y\|x)-P(*y)}{1-P(*y)},\frac{P(x\|y)-P(x*)}{1-P(x*)}]$
55. Added value (*AV*)	$\max[P(y\|x)-P(*y),\ P(x\|y)-P(x*)]$
*56. Collective strength	$\frac{P(xy)+P(\bar{x}\bar{y})}{P(x*)P(y)+P(\bar{x}*)P(*\bar{y})}\cdot$ $\frac{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(xy)-P(\bar{x}\bar{y})}$
57. Klosgen	$\sqrt{P(xy)}\cdot AV$
Context measures:	
*58. Context entropy	$-\sum_w P(w\|C_{xy})\log P(w\|C_{xy})$
59. Left context entropy	$-\sum_w P(w\|C^l_{xy})\log P(w\|C^l_{xy})$
60. Right context entropy	$-\sum_w P(w\|C^r_{xy})\log P(w\|C^r_{xy})$
*61. Left context divergence	$P(x*)\log P(x*)$ $-\sum_w P(w\|C^l_{xy})\log P(w\|C^l_{xy})$
62. Right context divergence	$P(*y)\log P(*y)$ $-\sum_w P(w\|C^r_{xy})\log P(w\|C^r_{xy})$
63. Cross entropy	$-\sum_w P(w\|C_x)\log P(w\|C_y)$
64. Reverse cross entropy	$-\sum_w P(w\|C_y)\log P(w\|C_x)$
65. Intersection measure	$\frac{2\|C_x\cap C_y\|}{\|C_x\|+\|C_y\|}$
66. Euclidean norm	$\sqrt{\sum_w(P(w\|C_x)-P(w\|C_y))^2}$
67. Cosine norm	$\frac{\sum_w P(w\|C_x)P(w\|C_y)}{\sum_w P(w\|C_x)^2\cdot\sum_w P(w\|C_y)^2}$
68. *L1* norm	$\sum_w\|P(w\|C_x)-P(w\|C_y)\|$
69. Confusion probability	$\sum_w\frac{P(x\|C_w)P(y\|C_w)P(w)}{P(x*)}$
70. Reverse confusion prob.	$\sum_w\frac{P(y\|C_w)P(x\|C_w)P(w)}{P(*y)}$
*71. Jensen-Shannon diverg.	$\frac{1}{2}[D(p(w\|C_x)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))$ $+D(p(w\|C_y)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))]$
72. Cosine of pointwise *MI*	$\frac{\sum_w MI(w,x)MI(w,y)}{\sqrt{\sum_w MI(w,x)^2}\cdot\sqrt{\sum_w MI(w,y)^2}}$
*73. KL divergence	$\sum_w P(w\|C_x)\log\frac{P(w\|C_x)}{P(w\|C_y)}$
*74. Reverse KL divergence	$\sum_w P(w\|C_y)\log\frac{P(w\|C_y)}{P(w\|C_x)}$
75. Skew divergence	$D(p(w\|C_x)\|\|\alpha(w\|C_y)+(1-\alpha)p(w\|C_x))$
76. Reverse skew divergence	$D(p(w\|C_y)\|\|\alpha p(w\|C_x)+(1-\alpha)p(w\|C_y))$
77. Phrase word coocurrence	$\frac{1}{2}(\frac{f(x\|C_{xy})}{f(xy)}+\frac{f(y\|C_{xy})}{f(xy)})$
78. Word association	$\frac{1}{2}(\frac{f(x\|C_y)-f(xy)}{f(xy)}+\frac{f(y\|C_x)-f(xy)}{f(xy)})$
Cosine context similarity:	$\frac{1}{2}(\cos(c_x,c_{xy})+\cos(c_y,c_{xy}))$ $c_x=(x_i);\ \cos(c_x,c_y)=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\cdot\sqrt{\sum y_i^2}}$
*79. in boolean vector space	$x_i=\delta(f(w_i\|C_x))$
80. in *tf* vector space	$x_i=f(w_i\|C_x)$
81. in *tf·idf* vector space	$x_i=f(w_i\|C_x)\cdot\frac{N}{df(w_i)};\ df(w_i)=\|\{x: w_i\in C_x\}\|$
Dice context similarity:	$\frac{1}{2}(\mathrm{dice}(c_x,c_{xy})+\mathrm{dice}(c_y,c_{xy}))$ $c_x=(x_i);\ \mathrm{dice}(c_x,c_y)=\frac{2\sum x_i y_i}{\sum x_i^2+\sum y_i^2}$
*82. in boolean vector space	$x_i=\delta(f(w_i\|C_x))$
*83. in *tf* vector space	$x_i=f(w_i\|C_x)$
*84. in *tf·idf* vector space	$x_i=f(w_i\|C_x)\cdot\frac{N}{df(w_i)};\ df(w_i)=\|\{x: w_i\in C_x\}\|$
Linguistic features:	
*85. Part of speech	{*Adjective:Noun, Noun:Noun, Noun:Verb*, . . . }
*86. Dependency type	{*Attribute, Object, Subject*, . . . }
87. Dependency structure	$\{\nearrow, \nwarrow\}$

Appendix 4: Top BofE collocates of ball by log-likelihood (span = 3:0)

Bank of English Corpus		
N = 69119		
COLLOCATE	LL score	n
TO	10977.4543	7499
HIT	9346.40785	1284
ZOE	6123.04942	502
LOOSE	5684.13166	600
WITH	5249.407	2696
HITTING	4900.60343	507
OFF	4257.86562	1080
ON	3716.97626	2346
KICKING	3701.22033	378
KICK	3401.3876	445
PLAY	3330.75046	693
GET	3220.0685	942
ALAN	3193.70848	468
GOLF	3079.33061	426
CRYSTAL	2991.51288	334
FIRST	2884.12671	997
THROUGH	2846.62283	792
HE	2804.4891	1944
NEW	2707.39848	1021
BAT	2509.68492	281
BONUS	2508.49702	301
LONG	2484.32202	687
KICKED	2377.18513	280

Appendix 5: Top BofE collocates of from (span = 0:1)

Bank of English Corpus		
N = 1920773		
COLLOCATE	LL score	n
APART	162022.203	19507
COME	77089.0372	19328
REPORTS	76989.8924	14099
FAR	73754.5493	16697
CAME	64728.0314	15672
COMES	61214.3018	11709
RANGING	54244.6944	6039
SUFFERING	53447.3484	7682
BENEFIT	40532.8969	6931
DIFFERENT	39942.8829	10567
EXCERPT	39361.9142	4354
DERIVED	38666.4277	4207
REMOVED	34629.7952	5273
COMING	29382.6039	7056
AVAILABLE	29125.5867	7183
MILES	28622.2585	5888
SUFFER	27411.2191	4264
BACK	25137.4335	12182
RECOVERED	23224.2541	3282
EMERGED	22423.2856	3792
REPORT	21162.5215	6664
PRESSURE	20559.984	5080
RECOVERING	19547.7798	2499
RANGE	19264.4904	4905
EVERYTHING	19248.5082	5390
NEWSCOPY	19179.9758	2001
BANNED	19035.5278	3141
EXEMPT	18259.5535	2317
EXCLUDED	18100.0522	2372
BENEFITED	17810.2165	2043
WITHDRAW	17493.0088	2580
RESULTING	17144.3836	2590
ARISING	17112.2291	1969
RETURNED	16909.5061	3839
SEPARATED	16274.0351	2360
TAKEN	15806.1927	5505

Appendix 6: Top BofE collocates of was (span = 0:1)

Bank of English Corpus		
N = 3246851		
COLLOCATE	LL score	n
BORN	97972.26767	16606
GOING	97276.68295	28704
GIVEN	39095.25821	12318
TOLD	34663.04014	12675
TAKEN	33482.24264	10630
FORCED	33332.15963	7475
ABLE	31474.41923	9813
BEING	29045.60118	15503
ARRESTED	28437.90825	5364
FOUND	27912.9746	11052
ONE	26515.86243	27821
KILLED	25897.93002	6650
SENT	24181.86345	6885
SUPPOSED	23614.84885	4825
DOING	22403.57846	8044
DUE	22363.43527	6294
MADE	22030.13768	13356
JAILED	20655.78805	3274
UNABLE	19911.05066	4279
NOTHING	18915.21154	7523
TRYING	17808.9554	6541
APPOINTED	17628.01803	3696
HAPPENING	15762.2362	3421
DELIGHTED	15510.48368	3164
SHOT	14811.19262	4909
LOOKING	14689.97706	6284
NAMED	14606.35701	3867
SURPRISED	14232.62961	3277
AWARDED	13606.25533	2629
WRONG	13428.96967	4816
RELEASED	13257.43047	3713
BROUGHT	12978.77851	4930
ASKED	12903.45793	5988
WEARING	12897.37505	3227
SOMETHING	12802.77084	7447
LAUNCHED	12535.7579	3314

Appendix 7: Examples of metaphor in the GCL

OBJECT	**METAPHOR**	**CORPUS EXAMPLE**
Football	FAIRNESS	deserved, chances, opportunity
Competition	JOURNEY WITH OBSTACLES; WAR	progress, advance, outset campaign
Match	PERFORMANCE; WAR	applause open game, close range
Team	PERFORMANCE; ARMY	display, support, prompt attacking thrusts
Players	PERFORMERS; SOLDIERS	display, effort threaten, defiant, massed ranks, defence
Goal	PRECIOUS OBJECT	convert, get, take
Result	WAR; JOURNEY STAGE	victory, defeat go through, next stage, bumpy
Ball	HAND POSSESSION	flicked, poked seized, grabbed
Shot	WEAPON	unleashed, powerful, send
Penalty/corner/free-kick	LAW	awarded, claim, dispute
Umpire	LAW	awarded, dismissed

Appendix 8: Examples of grammatical and lexical intratextual cohesion in the 7th paragraph

TEXT	GRAMMATICAL			LEXICAL		
	Between Messages	Meaning	Wording	Identity		Attribution
A casual, sterile first half						
got the goal it deserved		it		goal		
when PSV took the lead in the 36th minute,	when				PSV	
Javier Mascherano's attempted clearance from a corner hitting Dirk Marcellis on the hand			[~~attempted clearance~~] falling to…hitting			
and falling to Danko Lazovic to convert from close range beyond Liverpool reserve goalkeeper, Diego Cavalieri.	and			convert	Liverpool	
Despite Babel's prompting down the right		the right (exo.)			Babel	
Liverpool rarely stretched the PSV defence before the interval					Liverpool ; PSV	
yet equalised with remarkable ease	yet		[~~Liverpool~~] equalised	equalised		
when the Dutchman scored only his second goal of the campaign.	when	his		scored		the Dutch-man
Lucas was the provider from a free-kick out on the right		the right (exo.)				
and Babel ensured further invective from the PSV support by ghosting away from Otman Bakkal	and	the PSV support (exo.)			Babel	
and glancing a header through the suspect guard of Andreas Isaksson.	and					

Bibliography

Agirre, E., & Edmonds, P. (2006). *Word Sense Disambiguation: Algorithms and Applications.* New York: Springer.

Ansary, H., & Babaii, E. (2005). "The Generic Integrity of Newspaper Editorials: A Systemic Functional Perspective". *RELC Journal , 36* (3), 271-295.

Baayen, R. (2005). *Word Frequency Distributions.* New York: Springer.

Bahns, J. (1993). "Lexical collocations: a contrastive view". *ELT Journal , 47* (1), 56-63.

Bartsch, S. (2004). *Structural and Functional Properties of Collocations in English.* Tübingen: Gunter Narr Verlag.

Benson, M. J. (1991). "Attitudes and motivation towards English: A survey of Japanese freshmen". *RELC Journal , 22* (1), 34-48.

Berber Sardinha, T. (1999). "Using Key Words in Text Analysis: practical aspects". *DIRECT Papers 42* .

Biber, D. (1993). "Representativeness in corpus design". *Literary and Linguistic Computing , 8* (4), 243-257.

Biber, D. (2003). "Variation among University Spoken and Written Registers: A New Multi-Dimensional Analysis". In P. Leistyna, & C. Meyer (Eds.), *Corpus Analysis: Language Structure and Language Use.* Amsterdam/ New York: Rodopi.

Biber, D. (1995). *Dimensions of Register Variation: a cross-linguistic comparison.* Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Bowker, L., & Pearson, J. (2002). *Working with specialized language: a practical guide to using corpora.* New York: Routledge.

Butt, D., Fahey, R., Spinks, S., & Yallop, C. (1985). *Using Functional Grammar – An Explorer's Guide.* Sydney: National Centre for English Language Teaching and Research.

Carter, R. (1998). *Vocabulary – Applied Linguistic Perspectives* (2nd ed.). New York: Routledge.

de Beaugrande, R. (1996). "The 'pragmatics' of doing language science: The 'warrant' for for large-corpus linguistics". *Journal of Pragmatics , 25,* 503-535.

Dunning, T. (1993). "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics , 19* (1), 61-74.

Ferguson, C. (1983). "Sports Announcer Talk: Syntactic Aspects of Register Variation". *Language in Society* , 153-172.

Firth, J. (1957). *Papers in Linguistics, 1934-1951.* Oxford: Oxford University Press.

Frath, P., & Gledhill, C. (2005). "Free-Range Clusters or Frozen Chunks? Reference as a Defining Criterion for Linguistic Units,". *Recherches anglaises et Nord-américaines , 38*, 25-43.

Gavioli, L. (2005). *Exploring Corpora for ESP Learning.* Amsterdam: John Benjamins Publishing.

Ghadessy, M. (1998). "Textual features and contextual factors for register identification". In M. Ghadessy (Ed.), *Text and context in functional linguistics.* Amsterdam: John Benjamins Publishing Company.

Ghadessy, M. (1988). "The language of written sports commentary: soccer—a description". In M. Ghadessy (Ed.), *Registers of Written English: Situational factors and linguistic features.* London: Pinter Publishers.

Gledhill, C. (1995). "Collocation and genre analysis: the phraseology of grammatical items in cancer research abstracts and articles". *Zeitschrift für Anglistik und Amerikanistik , 43*, 11-29.

Gledhill, C. (2000). "The discourse function of collocation in research article introductions". *English for Special Purposes , 19* (2), 115-135.

Groom, N. (2005). "Pattern and meaning across genres and disciplines: An exploratory study". *Journal of English for Academic Purposes* , 257-277.

Halliday, M. (1991). "Corpus studies and probabilistic grammar". In K. Aijmer, & B. Altenberg (Eds.), *English Corpus Linguistics.* London and New York: Longman.

Halliday, M. (2004). *An Introduction to Functional Grammar* (3rd (Revised by Matthiessen, C.M.I.M) ed.). London: Hodder Arnold.

Halliday, M. (2005). *On Grammar.* (J. Webster, Ed.) London/New York: Continuum International Publishing Group.

Halliday, M., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Halliday, M., & Hasan, R. (1985). *Language, context, and text: Aspects of language in a social-semiotic perspective.* Geelong: Deakin University Press.

Hasan, A., & Babaii, E. (2005). “The Generic Integrity of Newspaper Editorials: A Systemic Functional Perspective”. *RELC Journal , 36*, 271-295.

Henry, A., & Roseberry, R. (2001b). "A narrow-angled corpus analysis of moves and strategies of the genre: 'Letter of Application'". *English for Specific Purposes , 20*, 153-167.

Hill, J. (1999). "Collocational competence". *English Teaching Professional , 11*, 3-8.

Hoey, M. (2004). "The Textual Priming of Lexis". In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and Language Learners.* Amsterdam: John Benjamins Publishing.

Hoey, M. (1991). *Patterns of Lexis in Text.* Oxford: Oxford University Press.

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern Grammar - A corpus-driven approach to the lexical grammar of English.* Amsterdam: John Benjamins Publishing.

Hyland, K. (2007). "Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing , 16*, 148-164.

Hyland, K. (2002). "Genre: Language, context, and literacy". *Annual Review of Applied Linguistics , 22*, 113-135.

Johns, A. (Ed.). (2002). *Genre in the classroom: multiple perspectives.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Johns, T. (1991). “Should you be Persuaded: Two Samples of Data-Driven Learning Materials”. *English language research journal , 4*, 1-16.

Kennedy, G. (1991). "Between and through: The company they keep and the functions they serve". In K. Aijmer, & B. Altenberg (Eds.), *English Corpus Linguistics.* London and New York: Longman.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* London: Longman.

Koprowski, M. (2005). “Investigating the usefulness of lexical phrases in contemporary coursebooks”. *ELT Journal , 59* (4), 322-332.

Lassen, I. (2006). “Is the press release a genre? A study of form and content”. *Discourse Studies , 8*, 503-528.

Leckie-Tarry, H. (1995). *Language and Context: A functional linguistic theory of register.* (D. Birch, Ed.) London & New York: Pinter.

Lewis, M. (Ed.). (2001). *Teaching Collocation: Further developments in the Lexical Approach.* Hove, England: Language Teaching Publications.

Mahlberg, M. (2006). "Lexical cohesion: Corpus linguistic theory and its application in English language teaching". *International Journal of Corpus Linguistics , 11* (3), 363-383.

Martin, J. (2002). "Meaning beyond the clause: SFL perspectives". *Annual Review of Applied Linguistics* , 52-74.

Martin, J. (2001). "Language, Register and Genre". In A. Burns, & C. Coffin (Eds.), *Analysing English in a Global Context.* New York: Routledge.

Martin, J. (1992). *English Text: System and Structure.* Amsterdam: John Benjamins Publishing.

Martin, J., & Rose, D. (2005). *Genre Relations: Mapping Culture.* London: Equinox Publishing.

McCarthy, M. (1991). *Discourse Analysis for Language Teachers.* Cambridge: Cambridge University Press.

McCarthy, M. (1990). *Vocabulary.* Oxford: Oxford University Press.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book.* New York: Routledge.

McGee, I. (2009). "Traversing the lexical cohesion minefield". *ELT Journal , 63* (3), 212-220.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A corpus-based approach.* Oxford: Clarendon Press.

Nation, I. (2001). *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Nattinger, J., & DeCarrico, J. (1992). *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Nesselhauf, N. (2003). "The Use of Collocations by Advanced Learners of English and Some Implications for Teaching". *Applied Linguistics , 24* (2), 223-242.

Oakes, M. (1998). *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.

ODE. (2003). *Oxford Dictionary of English (ODE), 2nd Edition (2003) :* (2nd ed.). Oxford: Oxford University Press.

Osman, H. (2004). “Genre-based instruction for ESP”. *The English Teacher , 33*, 13-29.

Painter, C. (2001). “Understanding Genre and Register: Implications for Language Teaching”. In A. Burns, & C. Coffin (Eds.), *Analysing English in a Global Context.* New York: Routledge.

Paltridge, B. (1994). “Genre Analysis and the Identification of Textual Boundaries”. *Applied Linguistics , 15* (3), 288-299.

Paltridge, B. (2007). *Discourse analysis: an introduction.* London/New York: Continuum International Publishing Group.

Pecina, P. (2005). "An extensive empirial study of collocation extraction methods". *Proceedings of the ACL Student Research Workshop* (pp. 13-18). Ann Arbor, Michigan: Association for Computational Linguistics.

Perfetti, C. (1987). "Comprehending newspaper headlines". *Journal of Memory and Language , 26*, 692-713.

Plum, G. (2006). *"Text and Contextual Conditioning in Spoken English: A genre approach".* Retrieved July 2nd, 2009, from Sydney eScholarship Repository: http://hdl.handle.net/2123/608

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, (pp. 1-6). Hong Kong.

Rayson, P., Berridge, D., & Francis, B. (2004). "Extending the Cochran rule forthe comparison of word frequencies between corpora". *Journées internationales d’Analyse statistique des Données Textuelles (7)*

Richards, J., & Rodgers, T. (2001). *Approaches and Methods in Language Teaching.* Cambridge: Cambridge University Press.

Romiane, S. (1994). "On the Creation and Expansion of Registers: Sports reporting in Tok Pisin". In D. Biber, & E. Finegan (Eds.), *Sociolinguistic perspectives on register.* New York: Oxford University Press, US.

Rowe, D. (1999). *Sport, Culture and the Media.* Buckingham, Philadelphia: Open University Press.

Rundell, M. (2008). "More Than One Way to Skin a Cat: Why Full-sentence Definitions Have Not Been Universally Adopted". In T. Fontenelle (Ed.), *Practical Lexicography: A Reader.* Oxford: Oxford University Press.

Scott, M. (2009). *WordSmith Tools 5.0.* Oxford: Oxford University Press.

Shirato, J., & Stapleton, P. (2007). "Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan". *Language Teaching Research , 11* (4), 393-412.

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

Sinclair, J. (2008). *How to use corpora in language teaching.* Phildelphia/Amsterdam: John Benjamins Publishing.

Skorczynska, H., & Deignan, A. (2006). "Readership and Purpose in the Choice of Economics Metaphors". *Metaphor and Symbol , 21* (2), 87-104.

Stubbs, M. (1995). "Collocations and Cultural Connotations of Common Words". *Linguistics and Education , 7,* 379-390.

Stubbs, M. (1995). "Collocations and semantic profiles: on the cause of the troubles with quantitative studies". *Functions of Language , 2* (1).

Swales, J. (1990). *Genre Analysis.* Cambridge: Cambridge University Press.

Taboada, M. (2004). *Building Coherence and Cohesion: Task-oriented Dialogue in English and Spanish.* Philadelphia/Amsterdam: John Benjamins Publishing.

Wierzbicka, A. (2006). *English: Meaning and Culture.* Oxford: Oxford University Press.

Williams, G. (2002). "In search of representativity in specialised corpora – Categorisation through collocation". *International Journal of Corpus Linguistics , 7* (1), 43-64.

Willis, D. (1990). *The lexical syllabus: a new approach to language teaching.* London and Glasgow: Collins ELT.

Wray, A. (2000). "Formulaic Sequences in Second Language Teaching: Principle and Practice". *Applied Linguistics , 21* (4), 463-489.

Zwaan, R. (1994). "Effect of genre expectation on text comprehension". *Journal of Experimental Psychology , 20* (4), 920-933.