E·S·P·Today
JOURNAL OF ENGLISH FOR SPECIFIC PURPOSES AT TERTIARY LEVEL

Ruth Breeze*

University of Navarra, Spain
rbreeze@unav.es

TEACHING THE VOCABULARY OF LEGAL DOCUMENTS: A CORPUS-DRIVEN APPROACH

Abstract

The globalization of business activity has been accompanied by an unprecedented need for lawyers to communicate internationally. One major aspect of their work centres on legal documents: lawyers have to draft and understand contracts and other essential documents, which are not only syntactically complex, but also contain highly technical vocabulary which may prove extremely challenging for non-native speakers. Even though non-native lawyers generally find the principles and concepts comprehensible, the lexis itself remains a considerable problem, unless specialized help is provided in the form of specific legal English instruction. This article describes a systematic approach to identifying the essential vocabulary of legal documents in English using WordSmith to find the most frequent words and word clusters (collocations and set phrases) in a 400,000 word corpus of authentic legal documents (DOCLEGAL), and then identify common verbs and prepositional phrases. The vocabulary of ten sub-corpora consisting of different document genres is then contrasted with that of the whole corpus, in order to identify keywords, keyword links and common clusters for these specific genres. The information from the corpus is then used to create exercises which help to familiarize students with these specialized words and expressions.

Key words

corpus-driven learning, materials design, legal English, English for specific purposes.

* Corresponding address: Ruth Breeze, Instituto de Idiomas/Instituto Cultura y Sociedad, Universidad de Navarra, C/Irunlarrea s/n, 31009 Pamplona, Spain.

Sažetak

Globalizaciju poslovnih aktivnosti prati izražena potreba za međunarodnom komunikacijom advokata. Jedan od najvažnijih aspekata njihovog rada tiče se pravnih dokumenata: advokati moraju da pišu i razumeju ugovore i ostala bitna dokumenta koja su ne samo sintaktički složena, već sadrže i usko specijalizovan vokabular koji može da predstavi priličnu prepreku za one advokate čiji maternji jezik nije engleski. Iako takvi advokati generalno razumeju pojmove i principe, sama leksika i dalje će predstavljati značajan problem ukoliko se ne pruži odgovarajuća pomoć u vidu specijalizovanih kurseva pravnog engleskog jezika. U radu se opisuje sistematski pristup identifikaciji ključnog vokabulara pravnih dokumenata na engleskom jeziku uz pomoć programa WordSmith, kako bi se pronašle najčešće reči i grupe reči (kolokacije i ustaljene fraze) u korpusu autentičnih pravnih dokumenata od 400.000 reči, a zatim izdvojili najčešći glagoli i predloške fraze. Vokabular deset podkorpusa koji se sastoje iz dokumenata različitih žanrova zatim je upoređen sa onim iz celog korpusa kako bi se izdvojile ključne reči, njihovo okruženje i najčešće grupe reči uobičajene za pojedine žanrove. Podaci dobijeni na osnovu korpusa zatim su upotrebljeni za izradu vežbanja pomoću kojih se studenti upoznaju sa specijalizovanim rečima i izrazima.

Ključne reči

učenje na osnovu korpusa, izrada nastavnih materijala, pravni engleski, engleski jezik struke.

1. INTRODUCTION

The last twenty years have seen an unprecedented rise in international business activity. One of the consequences of this is that lawyers, who previously tended to work within well-established national frameworks, have increasingly come into contact with international clients, handling agreements and disputes involving companies based in very different areas of the world, with widely differing legal systems and traditions. For practical reasons, a large proportion of international legal activity is conducted in English, and a huge number of legal documents (contracts, merger agreements, memoranda and articles of association, and so on) are drafted in English only, or in English plus other languages. This means that lawyers across the world, even those in non-common-law countries with no traditional link to the legal systems of the English-speaking world, now need to learn the basics of legal English. Moreover, they specifically need to learn to cope with the complex language of legal documents, which is opaque even to most

native speakers because of the complex syntax, specialized vocabulary and use of archaic conventional formulae (Charrow & Charrow, 1979; Cohen, 2008; Bhatia, 1993, 1998).

Perhaps because of the degree of difficulty it presents, vocabulary has often been the main focus of research into teaching legal English, and didactic material designed to initiate L2 students in this challenging area has often centred on a blend of terminology and content background (Russell & Locke, 1992; Chartrand, Millar, & Wiltshire, 1997; Ingels, 2006; Reinhart, 2007). As Bhatia, Candlin and Jensen (2002) conclude in their review of resources for teaching legal writing, much of the material available sits uneasily between English for Academic Purposes and content teaching, on the one hand, and L1 and L2 contexts, on the other. They call for "a more language and discourse-based approach" which is to be achieved "by grounding […] in research and evidence-based linguistic and discursive analysis of legal language", and specify that there should be a focus on "the discourses of the law rather than legal content" (2002: 316). The need to revisit the actual language of legal texts and to help students acquire appropriate strategies to deal with the lexical difficulties involved has been highlighted by Hafner and Candlin (2007), who suggest that L2 law students should be shown how to use corpora in order to improve their own legal writing in areas such as drafting opinions.

In general terms, recent research trends in the teaching of vocabulary have highlighted the need to study and teach words in their wider context, taking into consideration their possible collocations and combinations (O'Keefe, McCarthy, & Carter, 2007). Although some recent legal English textbooks have made a serious attempt to put these principles into practice (Krois Lindner, 2006; Brown & Rice, 2007), there is still a shortage of material in this area: researchers need to acquire deeper knowledge of the vocabulary of legal documents and the way words function in this context, and teachers need models for creating pedagogical activities that can help students develop a working knowledge of the language of legal documents.

This paper presents a systematic method for conducting applied research into the language of legal documents in order to develop practical exercises that will enable students to acquire proficiency in understanding and using these texts. The starting point for this was a 400,000 word corpus of legal documents, given the name DOCLEGAL, which was constructed using a range of representative authentic documents downloaded from online sources, including merger agreements, loan agreements, and contracts of sale. The research was carried out using various tools available in WordSmith (Scott, 1996). The research design was intended to provide a straightforward, systematic method for revealing some of the characteristic features encountered in these texts, and for exploring the way they function within these genres. This information is applied directly to the creation of materials for classroom use. In what follows, I explain the procedure for

identifying the specific features of these texts, and provide examples of learning activities that will help students understand and use them.

2. CORPUS AND METHOD

First, various tools provided by the WordSmith corpus analysis program were used to identify specific lexical aspects of legal documents that present special difficulties. The same program was then used to study these items in greater depth with a view to generating pedagogical exercises that can help students get to grips with the language of legal documents. The details of the corpus and research tools are explained below.

A 400,000 word corpus (DOCLEGAL) was compiled in 2011-12 using authentic legal documents from the area of commercial and corporate law, which had been obtained from the website http://www.onecle.com. Within the DOCLEGAL corpus there are ten sub-corpora consisting of: non-competition agreements, contracts of sale, license agreements, lease agreements, articles of association, debenture agreements, loan agreements, joint venture agreements, guaranty agreements and merger agreements. First, Wordsmith tools were used to investigate the lexis of the whole corpus. The functions used included classic wordcounts and frequency counts, keyword searches, and a concordancing function that enables researchers to detect clusters and patterns. Then the individual sub-corpora were researched in order to identify keywords, keyword links and common clusters for these specific genres. Finally, the data obtained from these researches were used to generate exercises designed to facilitate the acquisition of specialized words and expressions.

The following basic procedures were used to determine what vocabulary is important in legal documents, how we can get to know more about it, and how this knowledge can be operationalized in explanations and exercises for students.

1. Establishing what words are important in the main corpus: wordlists and frequency counts
2. Learning about collocations: researching common word clusters
3. Learning about collocations: researching prepositional phrases using concordance and clusters
4. Learning about verb use in context: researching frequent verbs using concordance and patterns
5. Establishing what is important in each sub-corpus: keywords
6. Learning about keywords in context: keywords, concordance and clusters

In each of the sections that follows, these procedures will be explained, and examples will be provided to show how the information obtained can be applied in practice.

3. THE PROCEDURES

3.1. Establishing what words are important in the main corpus: wordlists and frequency counts

As Scott and Tribble (2006: 31) comment, wordlists offer "an ideal starting point for the understanding of a text in terms of its lexis". Frequency wordlists throw the most frequent items to the top, and although in general corpora the most frequent tokens tend to be function words, the results in specialized corpora may bring out a very different pattern. This information is of paramount importance for teachers and course designers. As Flowerdew (1993) points out, frequency data provide a basis for establishing the relative importance of vocabulary items, which is essential information for course design and creation of didactic material. Table 1 shows the top 100 words in the corpus of legal documents.

1 THE	26 WHICH	51 BOARD	76 INFORMATION
2 OF	27 DATE	52 SET	77 PRIOR
3 TO	28 EACH	53 THAN	78 SUBJECT
4 OR	29 SECTION	54 PROVIDED	79 TERMS
5 AND	30 MAY	55 CLOSING	80 PURSUANT
6 IN	31 AT	56 IT	81 AFTER
7 ANY	32 HAVE	57 RIGHTS	82 EMPLOYEE
8 A	33 UNDER	58 PARTIES	83 LAW
9 SHALL	34 ARE	59 APPLICABLE	84 MADE
10 BY	35 IF	60 PERSON	85 REQUIRED
11 SUCH	36 AN	61 ARTICLE	86 WITHOUT
12 BE	37 STOCK	62 FORTH	87 ASSET
13 AS	38 PARTY	63 OBLIGATIONS	88 SECURITIES
14 COMPANY	39 SHARES	64 LICENSEE	89 RIGHT
15 WITH	40 BUSINESS	65 BEEN	90 WRITTEN
16 FOR	41 FROM	66 SHAREHOLDERS	91 PERIOD
17 OTHER	42 HAS	67 SCHEDULE	92 GENERAL
18 THIS	43 DIRECTORS	68 UPON	93 SERIES
19 AGREEMENT	44 NO	69 EXCEPT	94 PAYMENT
20 THAT	45 RESPECT	70 OTHERWISE	95 LICENSED
21 IS	46 TIME	71 SUBSIDIARIES	96 INTEREST
22 ALL	47 INCLUDING	72 PROPERTY	97 AMOUNT
23 NOT	48 MEETING	73 LICENSOR	98 BANK
24 ON	49 WILL	74 CORPORATION	99 EFFECT
25 ITS	50 NOTICE	75 MATERIAL	100 USE

Table 1. Top 100 words in DOCLEGAL corpus

It is enough to glance at this wordlist to perceive that the language of legal documents is quite unlike the type of English encountered in everyday spoken encounters or non-specialist texts. When compared with the top hundred words in the British National Corpus (Scott & Tribble, 2006; Aston & Burnard, 1998), which are nearly all grammatical, the list set out in Table 1 can be seen to be radically different. Whereas the first hundred words in the BNC include only three nouns,

namely *time*, *people* and *way*, the most frequent hundred words in the legal document corpus contains 43 potential nouns: *company*, *agreement*, *date**, *section*, *stock*, *party*, *shares*, *business*, *directors*, *respect**, *time*, *meeting*, *notice**, *board*, *rights*, *parties*, *person*, *article*, *obligations*, *licensee*, *shareholders*, *schedule*, *subsidiaries*, *property*, *licensor*, *corporation*, *material*, *information*, *subject**, *terms*, *employee*, *law*, *asset*, *securities*, *right**, *period*, *series*, *payment*, *interest*, *amount*, *bank*, *effect*, and *use**. Of these, several (those asterisked* above) may conceivably also be verbs (*date*, *notice*, *respect*, *subject*, *use*) or adjectives (*subject*, *right*). Without using tagging tools, it is not possible to establish which is which, but it is still noticeable that the number of "content words" (as opposed to "grammar words") is much higher in DOCLEGAL than in the BNC top hundred list.

Another striking feature is the prominence of *shall*, *under*, *set*, *provided*, *closing*, *applicable*, *forth*, *upon*, *except*, *otherwise*, *prior*, *pursuant*, *without*, *general*, and *licensed*, none of which is in the BNC top hundred. Aside from the highly predictable forms of *to be*, *to have*, and the modal *will*, the verbs listed here seem to point to the prevalence of specific legal formulae, as in the case of *shall* and *may*, and to the use of verbal linkers: *including*, *provided*. The heavy presence of *set* and *made* seems to indicate delexicalized uses of these verbs in set phrases.

Frequency lists such as these are essential as a starting point for researching and teaching vocabulary. In themselves, they provide pointers about which lexical items we ought to teach. On grounds of frequency alone, it can be seen that a range of non-core items like *shall*, *board*, *pursuant*, *prior*, and *otherwise* all need to form part of the repertoire taught in the legal English classroom. On the other hand, the frequency of prepositions such as *of*, *to*, *in*, *with* and *for* seemed interesting, in that it might point to an abundance of prepositional phrases or verbs that collocate with particular prepositions, which could provide useful practical insights into typical collocations and word clusters encountered in legal documents.

In themselves, frequency lists do not lend themselves particularly to the creation of learning activities. However, the following simple awareness-raising activity may be useful at the start of a lesson focusing on legal documents (Schmitt & Schmitt, 2005).

Exercise 1: Frequent words

How wide is your legal vocabulary? Look at each of the words below, and use the scale to give yourself a score for each word. After we have finished the unit, you will be tested on your knowledge of these words.

1. I have never seen this word before.
2. I have seen this word, but I'm not sure what it means.
3. I understand this word when I see or hear it, but I don't know how to use it myself.
4. I know this word, and I use it when I speak or write.

..... subsidiaries	..... licensor	..... employee	..... payment
..... shares	..... securities	..... asset	..... shareholders

3.2. Learning about collocations: researching common word clusters

Corpus tools are particularly useful for revealing the regular patterns that permeate written and spoken language. For linguists, this is one of the most exciting aspects of corpus research, since it is relatively simple to identify recurring groups of words, and to establish just how likely they are to recur again in a particular context. These recurring groups have been given many different names, from "multi-word units" to "prefabricated phrases" or "chunks" (O'Keefe et al., 2007: 63), to "clusters" (Scott & Tribble, 1996), but there is general agreement that they are fundamental to language use, since they enable people to bypass the analytical processes required to interpret or produce language from scratch (Wray, 2002). Moreover, it is now recognized that they are particularly important for L2 language learners, since mastery of ready-made clusters of language may speed up performance and facilitate fluent production (O'Keefe et al., 2007).

Although much research has focused on clusters in spoken language (Sinclair, 1991; Howarth, 1998; McCarthy, 1998, 2006; Biber, 2007), where they appear to be particularly prominent, it is known that written genres also contain formulaic sequences. Several studies have focused on the role of clusters in acquiring effective writing skills, particularly in academic contexts (e.g. Biber & Conrad, 1999; Hyland, 2008; Salazar, 2014), although it has been established that clusters are relatively rare in written academic registers when these are compared to spoken language or non-academic written registers (Biber, 2006, 2007). Less interest has centred on their role in reading comprehension, perhaps because clusters in areas such as academic prose pose few challenges to the reader. However, given the highly formulaic nature of legal documents, it may be supposed that clusters have a particular prominence in this genre. Moreover, the sheer complexity involved in understanding texts of this kind might suggest that better knowledge of recurring patterns in them might pay dividends.

Using the Wordlist function in WordSmith, I generated lists of the most frequent 2- to 8-word clusters (on the nature and frequency of clusters in this corpus see Breeze, 2013). Since these were extremely frequent, I decided to design some awareness-raising activities that would help students to become familiar with common clusters, and get used to the way that clusters link together to form phrases. With this aim in mind, exercise 2 was devised. It is based on the following 2, 3, 4, 5 and 6-word clusters which were all frequent in DOCLEGAL: "subject to" (490 occurrences), "other than" (410), "relating to" (386), "or otherwise" (338), "the date of this agreement" (103), "in effect" (86), "in the ordinary course of business" (76), "parties to" (48), "shall have the meaning set forth" (47), "relates to" (25), "the assets of" (25), "any interest in" (11). These clusters co-occur to a striking degree in particular types of contract clause.

Exercise 2. Common word clusters

1. The groups of words below include some common phrases that you will find in legal documents. Can you join the groups of words together to make a coherent sentence?

in Clause 3.1 of this Agreement shall have the meaning set forth

"Confidential Information"

(Answer: "Confidential Information" shall have the meaning set forth in Clause 3.1 of this Agreement.)

2. The groups of words below also need to be reordered to make a coherent sentence. Because this one is more difficult, the beginning and end have been done for you.

on the date of this Agreement
any Contract in effect
any business enterprise
relating to the disposition or acquisition of the assets of, or any interest in,
that relates to This Company's Business

This Company and its subsidiaries are not parties to or subject to……..
……..
…… other than in the ordinary course of business.

(Answer: This Company and its subsidiaries are not parties to or subject to any Contract in effect on the date of this Agreement relating to the disposition or acquisition of the assets of, or any interest in, any business enterprise that relates to This Company's Business other than in the ordinary course of business.)

3.3. Learning about collocations: researching prepositional phrases using concordance and clusters

A second, more focused approach to clusters was then devised, using WordSmith's Concord tool to identify prepositional phrases that were frequent in these documents. By running Concord with common prepositions as the search word, and then locating 2, 3 and 4-word clusters, we were able to obtain lists of frequent prepositional phrases such as those in Table 2:

IN	**FOR**	**AT**
in accordance with	for (the) purposes of	at any time
in connection with	for the benefit of	at the meeting
in respect of	for a period of	at the time of
in the event of	except for	at least

in all material respects	for the avoidance of	at all times
in the case of	for and on behalf of	present at
in the ordinary course of	responsible for	at law
in part	liable for	at the request of
in good faith	for the time being	at the address
in addition to	for any (other) purpose	at a rate
in excess of		
in the form of	**WITH**	
in order to	with respect to	**TO**
in relation to	in accordance with	with respect to
in full force	in connection with	to the extent (that/of)
in any way	together with	prior to
in lieu of	with respect to	from time to time
in any manner	in compliance with	subject to
in writing	consistent with	relating to
in good standing	with regard to	pursuant to
in no event		to the knowledge of
in compliance with	**BY**	in addition to
in witness whereof	by way of	equal to
in a manner	by reason of	in relation to
in any event	by virtue of	party to
in the absence of	by the terms of	
	by and between	

Table 2. Frequent prepositional phrases in DOCLEGAL

Given the high frequency of many of these prepositional phrases (e.g. "in accordance with" occurs 435 times in the main corpus), this information is useful for the classroom. Students need to become familiar with the most widely occurring instances of such phrases, and they need to understand the role that these phrases play in legal documents. The following example adapted from the corpus incorporates several common prepositional phrases in a typical clause:

Twig Marks. Twig Media hereby grants to Jones a non-exclusive, perpetual, irrevocable, non-transferable, fully paid up, worldwide right and license to use the Twig Marks *in connection with* the marketing and provision of goods and services, with the right to sublicense such rights *in the ordinary course of* business. Neither Jones nor any of its sublicensees will obtain any right, title or interest in the Twig Marks *by virtue of* their use of the Twig Marks. Any goodwill that is created through the use of the Twig Marks by Jones or any of its sublicensees will be solely *for the benefit of* Twig Media. All uses of the Twig Marks by Jones or any of its sublicensees will be: (i) *in accordance with* Twig Media's then-current trademark usage policies, and (ii) *subject to* inspection and monitoring by Twig Media to ensure that such uses are *in accordance with* such policies. At Twig Media's request, Jones and its sublicensees shall promptly make any changes *with regard to* usage of the Twig Marks as Twig Media deems appropriate. Any such changes shall be made *at the expense of* Jones and its sublicensees.

Example 1. Prepositional phrases in license agreement

As Example 1 illustrates, these prepositional phrases play a key role in the text. In many instances, a prepositional phrase appears to be used instead of a shorter, more mundane alternative ("by virtue of" in this context could be substituted in everyday language by "by", "for the benefit of" by "for", "in accordance with" by "according to"). However, such prepositional phrases are a prominent feature of legal documents, and have a technical function in spelling out the legal implications of what is being stated (thus "for the benefit of" is clearer than "for", which might be ambiguous in this context).

From the students' point of view, these phrases present a double layer of difficulty. On the one hand, although they make sense in context, their meaning is hard to paraphrase. On the other, non-native students find English prepositions confusing, and often have difficulty remembering which one is used in a particular expression, particularly as far as similar pairs (in/on, of/from) are concerned. In order to familiarize students with these phrases, a scaffolded two-stage activity was designed on the basis of the above excerpt. The first exercise is designed as an introduction to these prepositional phrases in familiar legal contexts. In this exercise, the prepositions are given, and the student only has to identify the core of the phrase from a list that is given (i.e. the student has "at the… of", and knows that the context involves money, so should be able to guess "at the expense of"). In the second exercise, the clause identified in the corpus is used as the main text, and the students carry out a gapfill exercise in which they have to use one of the prepositional phrases in each gap.

Exercise 3: Prepositional phrases 1

The following sentences contain some prepositional phrases that are common in legal documents. The main word in each phrase has been taken out, and only the prepositions have been left in the text. Choose the right word/s from the box to complete each prepositional phrase:

connection	ordinary course	virtue	benefit
subject	regard	expense	accordance

Which of these phrases could be used in the following extracts?

1. All damage shall be repaired at the …….. of the landlord. (Answer: at the expense of)
2. Claims are permitted by ……. of the parties' relationship with the company. (Answer: by virtue of)
3. The Lender has agreed to make Advances to the Borrower upon the terms and …….. to the conditions set forth therein. (Answer: subject to)
4. Licensee shall submit to Licensor, for Licensor's review, Licensee's marketing plans for the current Contract Year for the Licensed Territory with …….. to the Licensed Products. (Answer: with regard to)

5. The Lender may sign and endorse any invoices in …….. with the Guaranty Collateral. (Answer: in connection with)
6. Accounts may be created for the …….. of a minor. (Answer: for the benefit of)
7. The accounts are valid and genuine, have arisen out of bona fide sales, and have been billed or invoiced in the …… …….. of business. (Answer: in the ordinary course of)
8. The rights and licenses set forth in this Section include the right of Brown to disclose the Confidential Information included in such Enterprise, provided that such disclosure is in …….. with the confidentiality obligations set forth in this Agreement. (Answer: in accordance with)

3.4. Learning about verb use in context: researching frequent verbs using concordance and patterns

Another important feature of legal documents is the special use of certain common verbs, such as *set*, *arise* and *hold*. Taking the most frequent verbs obtained using WordSmith wordlists as a guide, I used Concord to obtain information about their collocates and the patterns in which they occurred. It was thus possible to build exercises to help students understand the way these verbs were used in the context of legal documents.

3.4.1. Example: ***set***

The verb "set" occurs 761 times in the DOCLEGAL corpus. The most common collocations were "set forth" (611), "set forth in" (424 occurrences) followed by "as set forth" (184), and there were also many longer clusters, such as "as set forth in" (118), "the meaning set forth in" (76), "except as set forth in" (67), and many variations on this theme. As Table 3 shows, the Patterns function in Concord (Scott & Tribble, 2006: 40-41) also made it possible to establish that "set forth" is frequently used in relation to conditions and meanings (usually expressed as the subject of a passive) which are set forth in the document (agreement, etc.) itself, or in a section or schedule.

N	L4	L3	L2	L1	R1	R2	R3	R4	
1	the	the	the	meaning		forth		the	agreement
2	shall	have	except	conditions		out		schedule	
3	and	has	and	expressly			for	this	
4	subject	and	paid	are			section	section	
5	any	that	shall	shall	aside		such	and	

Table 3. Patterns obtained with Concord using "set" as node

Exercise 4: Frequent verbs – *set*

The phrasal verb "to set forth" is found very frequently in legal documents, particularly when referring to terms, conditions and meanings. Look at the examples from commercial contracts:

1. "Expiration Date" shall have the meaning **set forth in** Section 2(c) herein.

2. NOW, THEREFORE, in consideration of the foregoing recitals and the conditions and other terms **set forth in** this Agreement, the Parties agree as follows:

What would you say in normal English to explain the same idea?

Now express the ideas below using the verb "set forth".

1. "Fundamental Change" has the meaning that is explained in the Warrant Certificate.
(Answer: has the meaning set forth in)

2. All notices shall be delivered as we have explained in section 3.
(Answer: as set forth in section 3).

3. Licensee shall pay Licensor in accordance with the manner that it explains in Section 2(b).
(Answer: in accordance with the manner set forth in)

3.4.2. Example: ***arise***

Forms of the verb "to arise" occur a total of 283 times in DOCLEGAL, the commonest forms being *arising* (243 occurrences) and *arise* (21). *Arising out of* (107) and *arising from* (49) are the commonest combinations obtained from the Wordsmith cluster function.

Concordancing tools can be further used to detect the kind of patterns and contexts in which "arising" is used. From this, we can establish that what *arises* is often an obligation or problem: claims, suits, legal proceedings or actions, damages, liabilities, taxes, disputes, omissions all form the subject of "to arise" in DOCLEGAL. On the positive side, rights and benefits may also form the subject of "to arise". All these things usually arise out of or from an agreement or a breach of an agreement. It is thus possible to identify typical occurrences of this verb and develop an exercise on this basis.

Exercise 5: Frequent verbs – *arise*

Read the following extract from a contract clause:

No party to this Agreement (or any of its Affiliates) shall, in any event, be liable or otherwise responsible to any other party (or any of its Affiliates) for any consequential or punitive damages of such other party (or any of its Affiliates) **arising out of this Agreement** or the performance or breach hereof.

Which of these phrases is closest in meaning to the words in bold in the example?
a. Appearing in this Agreement
b. Occurring as a consequence of this Agreement
c. Occurring outside this Agreement

Where, in the following sentences, should the phrase "arising out of" be inserted?

1. In the event of any controversy or claim between or among any of the Parties this Agreement, the Parties shall try to settle their differences amicably between or among themselves.
(Answer: In the event of any controversy or claim between or among any of the Parties **arising out of** this Agreement, the Parties shall try to settle their differences amicably between or among themselves.)

2. Any suit, action or proceeding based on any matter this Agreement shall be brought in the United States District Court for the District of Delaware.
(Answer: Any suit, action or proceeding based on any matter **arising out of** this Agreement shall be brought in the United States District Court for the District of Delaware.)

3. No Party shall be liable to any other Party for special, indirect, punitive or consequential damages, including lost profits and opportunity costs, a breach of this Agreement.
(Answer: No Party shall be liable to any other Party for special, indirect, punitive or consequential damages, including lost profits and opportunity costs, **arising out of** a breach of this Agreement.)

3.4.3. Example: ***hold***

The verb "to hold" occurs a total of 306 times in DOCLEGAL, the commonest forms being *held* (198 occurrences) and *hold* (91). This is a difficult verb for many non-natives, because its forms are irregular, and because its usual meaning in the context of documents (equivalent to "have") differs slightly from its main everyday meaning (which is generally physical). In order to explore the use of this verb in legal documents, the Concord tool was used, with its cluster and pattern functions. The pattern function, in particular, was useful in showing that shares or stock, assets, licenses, interests, etc. are usually what is held. The cluster function, on the

other hand, revealed that the combinations *held by* (89) and *held or used* (18) are frequent. The root form of the verb, *hold*, is also found in variations of the expression "hold harmless", a legal formula meaning not to hold liable for any loss or damage. This is encountered either directly as *hold harmless* (15 occurrences) or *hold someone/something harmless* (6). The other instances of *hold* are in the context of "holding a meeting" or "holding office", as well as in the cluster "have and hold premises". Exercise 6 below focuses on the most typical use of "to hold", meaning "to have, to possess", and is designed to make students focus on the form of the verb (present simple or past participle).

Exercise 6: Frequent verbs – *hold*

Sometimes irregular verbs can be confusing. For example, "to hold" is an irregular verb that is often found in legal documents, but you are more likely to find it in its past participle form, "held". Look at the following sentences and decide whether to use "hold" or "held".

1. Schedule 5.4(e) sets forth a true and correct list of all holders of Company Common Stock and Preferred Stock and the number of shares **hold/held** (Answer: held) by each such holder.
2. The Corporation shall effect such redemption pro rata according to the number of shares **hold/held** (Answer: held) by each Holder of Series A Preferred Stock.
3. In an AGM, the supervisory committee and shareholders who individually or jointly **hold/held** (Answer: hold) 5% or more of the Company's voting shares shall have the right to put forward provisional motions.

3.5. Establishing what is important in each sub-corpus: keywords

The keywords function offers the most efficient way to establish which words are particularly frequent in a particular set of texts. The main problem in the area of specialized language, such as the language of legal documents, is that it is not easy to define what an appropriate reference corpus might be (Scott & Tribble, 2006: 58), although there is a general rule that the reference corpus should be at least five times the size of the node text (Berber Sardinha, 2004: 102). Different reference corpora result in different keywords appearing. When a very large, general reference corpus is used, the keywords are those items that are more salient in the text than in general English: in this case, they would be the typical features encountered in legal English. However, when a comparison is made with a relatively specialized reference corpus, the keywords will be those words that are particularly salient in the text itself, that is, words related to what the particular text in question is about. In this case, the ten sub-corpora were compared with the whole legal documents corpus, in order to identify which words were particularly important in each sub-corpus. This procedure should eliminate "general" legal vocabulary

common to all of the document types, which is presumably best obtained by means of general wordcounts across the whole corpus, and bring out the "specific" vocabulary that is most characteristic of each type of document. The keywords (down to keyness of 45) were calculated for the sub-corpora as compared to the whole corpus. Tables 4 to 7 show the keywords for four of the sub-corpora.

1. executive	11. employment
2. firm	12. restricted
3. interests	13. class
4. area	14. covenants
5. buyer	15. activities
6. agreement	
7. damages	
8. exchangeable	
9. service	
10. liquidated	

Table 4. Keywords in sub-corpus of non-competition agreements (keyness >45)

1. company	13. option
2. merger	14. contemplated
3. stock	15. certificate
4. holding	16. sheet
5. subsidiaries	17. outstanding
6. election	18. transactions
7. options	19. balance
8. shares	20. surviving
9. holder	21. corporation
10. stockholder	22. series
11. common	23. effective
12. share	

Table 5. Keywords in sub-corpus of merger agreements (keyness >45)

1. bank	15. ratio
2. borrower	16. indebtedness
3. advance	17. debt
4. loan	18. margin
5. interest	19. fee
6. rate	20. monthly
7. committed	21. continuation
8. amendment	22. minus
9. accounts	23. modification
10. principal	24. default
11. credit	25. debtor
12. prime	26. deposit
13. extension	27. equal
14. maturity	28. base

Table 6. Keywords in sub-corpus of loan agreements (keyness >45)

1. tenant	10. completion
2. landlord	11. standard
3. premises	12. solicitor
4. lease	13. fire
5. lessee	14. insurance
6. superior	15. condition
7. rent	16. repairs
8. building	17. damage
9. lessor	18. possession

Table 7. Keywords in sub-corpus of lease agreements (keyness >45)

As above, in the case of the most frequent words in the whole DOCLEGAL corpus (Exercise 1), simple exercises can be devised to raise students' awareness of important lexical items in a given field. Exercise 7 is an example of a straightforward activity of this kind (see Exercise 1 above).

Exercise 7: Keywords

How much of the key vocabulary found in lease agreements do you already know? Look at each of the words below, and use the scale to give yourself a score for each word. After we have finished the unit, you will be tested on your knowledge of these words.

1. I have never seen this word before.
2. I have seen this word, but I'm not sure what it means.
3. I understand this word when I see or hear it, but I don't know how to use it myself.
4. I know this word, and I use it when I speak or write.

..... tenant	..... premises	..... insurance	..... solicitor
..... landlord	..... lessor	..... lessee	..... lease

3.6. Learning about keywords in context: keywords, concordance and clusters

These keywords provide a starting point for further analysis of vocabulary patterns in specialized texts, since each keyword can then be studied using the clusters and patterns functions in Concord, in order to establish how the word behaves in terms of collocations, and in what contexts it typically appears. Thus if we take the keyword "lessee" in the subcorpus of lease agreement as the node in the patterns function, we obtain the following results:

N	L4	L3	L2	L1	R1	R2	R3	R4	
1	the	lessor	that	the			the	the	
2	said	the	and			and		entitled	
3	and	premises		but				pay	
4	out	provided		deposit				its	and
5	any	lessee	premises				the		pay
6	shall	out	for		for	not	that		
7	lease	lease	provided			hereby	pay	and	lessee
8	for	and	lessee				lessor	said	

Table 8. Patterns obtained with Concord using "lessee" as node

As Table 8 shows, "lessee" is strongly linked with two other keywords "lessor" and "premises", since these both frequently occur close to "lessee", either to the left ("premises", "lessor") or to the right ("lessor"). If similar tables are generated for these two keywords, they are also found to be closely linked to each other, and to "lessee". This means that these three keywords are likely to be found close to each other in the lease agreements: these keywords belong to each other's "natural environment". On the basis of this knowledge, it is possible to return to the corpus and extract samples of language in which the three linked keywords appear in close proximity:

The Lessor has agreed to give the said Premises with the said amenities therein on Lease to the Lessee for a period of 36 months with an option for renewal as aforesaid subject to the Lessee observing and complying with the terms and conditions mentioned herein.

Example 2. Extract from lease agreement

As we have seen, "lessor", "lessee" and "premises" form what might be termed a keyword triangle, in that each of them is likely to be found close to the other two. However, the situation is actually more complex than this, because "premises" also forms part of a second, similar keyword triangle involving "landlord" and "lease". When some keywords are used as node in the patterns function, more complex linkage networks also emerge. For example, in the case of lease agreements, the word "premises" also occurs at the centre of a keyword "star" (Scott & Tribble, 2006: 47-50). This information was obtained by using the WordSmith "patterns" and "mutual information" applications to establish what words co-occurred within five words of each keyword, and what the strength of the association between particular keywords was. For example, "premises" is found close to a set of other keywords that are not mutually linked to each other: "premises" is found close to "possession", "condition", "damage" and "building", even though these items do not occur close to each other. Similarly, "lease" forms the centre of a star, around the

periphery of which we find the keywords "rent", "rental", "completion", "lessee", "premises", "landlord" and "tenant". Such knowledge arguably enables us to construct a diagram of the "lexical environment" of a particular word in a given genre.

Knowledge of this type enables us to return to the original corpus and extract samples of language that are reasonably "typical" of the genre, where certain keywords co-occur, in order to create exercises that encourage students to focus on these important words in their usual surroundings. Exercise 8 below exploits the keyword triangle lessor-lessee-premises, so that students can practise distinguishing between three unfamiliar words that will be encountered again and again in the context of lease agreements. Exercise 9, on the other hand, uses what I have established to be a wider network of interrelated keywords in lease agreements ("premises", "landlord", "tenant", "rent", "lease", "damage"), in a controlled exercise that helps students to build an understanding of the patterns in lease agreement clauses. Although the exercise initially appears to be taxing, once the student has realized that each sub-clause contains similar ideas and structures, it turns out to be relatively simple.

Exercise 8: Keywords in context – *lease agreements 1*

Choose the appropriate word (lessor, lessee or premises) to fill each gap in order to complete the text below.

The …… (Answer: lessor) has agreed to give the said …… (Answer: premises) with the said amenities therein on Lease to the …... (Answer: lessee) for a period of 36 months with an option for renewal as aforesaid subject to the …… (Answer: lessee) observing and complying with the terms and conditions mentioned herein.

Exercise 9: Keywords in context – *lease agreements 2*

Complete this extract from a lease agreement by using the words from the box.

premises x 2	tenant x 2	landlord	rent	lease	damage

(a) If the …… (Answer: premises) or any part thereof shall be damaged by fire or other casualty, …… (Answer: tenant) shall give immediate notice thereof to …… (Answer: landlord), and this …… (Answer: lease) shall continue in full force and effect except as hereinafter set forth.

(b) If the …… (Answer: premises) are partially damaged or rendered partially unusable by fire or other casualty, not caused by …… (Answer: tenant), the …… (Answer: damage) thereto shall be repaired by and at the expense of the Landlord and the …… (Answer: rent), until such repair shall be substantially completed, shall be apportioned from the day following the casualty according to the part of the premises which is usable.

4. CONCLUSIONS

This paper has applied a systematic data-driven approach to researching and teaching the language of a highly specific set of genres within English for specific purposes. For teachers of legal English, such a method makes it possible to focus students' attention selectively on different aspects of the lexis of legal documents, and to generate large numbers of practice exercises based on real examples. In principle, such a methodology could easily be transferred to other areas of language teaching, particularly those where lexical issues pose difficulties to the learner. The benefits of having a readily available specialized corpus and appropriate processing tools are beyond doubt. In Flowerdew's words, "high face-validity is given to an ESP course if the learning materials contain actual examples of use which are drawn from the content area and which the learner is likely to have come across" (Flowerdew, 1993: 239). ESP teachers have to meet the challenge of selecting the right kind of learning material, diagnosing potential difficulties, and preparing useful and varied activities that draw students' attention to important issues and facilitate the acquisition of coping skills. Against this background, this paper has provided an example of how practising teachers can bridge the gap between corpus and classroom in English for specific purposes.

[Paper submitted 10 Apr 2015]
[Revised version accepted for publication 12 May 2015]

References

Aston, G., & Burnard, L. (1998). *Exploring the British national corpus with SARA*. Edinburgh: Edinburgh University Press.

Breeze, R. (2013). Lexical bundles in four legal genres. *International Journal of Corpus Linguistics*, *18*(2), 229-253.

Berber Sardinha, A. (2004). *Lingüística de corpus* [Corpus linguistics]. Barueri SP, Brazil: Manole.

Bhatia, V. (1993). *Analysing genre: Language use in professional settings*. London: Longman.

Bhatia, V. (1998). Intertextuality in legal discourse. *The Language Teacher*, *22*(11), 13-39.

Bhatia, V., Candlin, C., & Jensen, C. (2002). Developing legal writing materials for English second language learners: Problems and perspectives. *English for Specific Purposes*, *21*, 299-320.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*, 263-286.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard, & S. Oksefjell (Eds.), *Out of corpora: Studies in honor of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Brown, G., & Rice, S. (2007). *Professional English in use: Law*. Cambridge: Cambridge University Press.
Charrow, R., & Charrow, V. (1979). Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Review*, *79*, 1306-1374.
Chartrand, M., Millar, C., & Wiltshire, E. (1997). *English for contract and company law.* London: Sweet and Maxwell.
Cohen, R. (2008). The law, the musician, his band and their partnership agreement: Comprehensibility, comprehension and compliance in a legal text. *Entertainment and Sports Law Review*, *6*(2). Retrieved from http://www2.warwick.ac.uk/fac/soc/law/elj/eslj/issues/volume6/number2/
Flowerdew, J. (1993). Concordancing as a tool in course design. *System*, *21*, 231-244.
Hafner, C., & Candlin, C. (2007). Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes*, *6*, 303-318.
Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24-44.
Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41-62.
Ingels, M. (2006). *Legal English communication skills*. Leuven: Acco.
Krois Lindner, A. (2006). *International legal English*. Cambridge: Cambridge University Press.
McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.
McCarthy, M. (2006). *Explorations in corpus linguistics*. Cambridge: Cambridge University Press.
O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom*. Cambridge: Cambridge University Press.
Reinhart, S. (2007). *Strategies for legal case reading and vocabulary development*. Ann Arbor: University of Michigan Press.
Russell, F., & Locke, C. (1992). *English law and language*. London: Cassell.
Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing*. Amsterdam: John Benjamins.
Schmitt, D., & Schmitt, N. (2005). *Focus on vocabulary: Mastering the academic word list*. White Plains, NY: Longman.
Scott, M. (1996). *WordSmith*. Oxford: Oxford University Press.
Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

RUTH BREEZE has a Ph.D. in Applied Linguistics and has published widely in the area of specialised language and media language. She is a member of the GradUN Research Group in the Instituto Cultura y Sociedad at the University of Navarra, Spain. Her most recent books are *Corporate Discourse* (Bloomsbury Academic, 2013) and the edited volume *Interpersonality in Legal Genres* (Peter Lang, 2014).