International Journal of
LANGUAGE & LAW

Marín, JLL 6 (2017): 18–45
www.languageandlaw.de

Legalese as Seen Through the Lens of Corpus Linguistics

— An Introduction to Software Tools for Terminological Analysis

María José Marín[*]

Abstract

In spite of the plethora of possibilities offered by Corpus Linguistics to the study of legal English, the research devoted to the study of this English variety based on this discipline is not as fruitful as that dedicated to other branches of ESP. The present research could be regarded as an introduction into major issues related to the design and compilation of a legal corpus such as the application of appropriate sampling strategies to ensure its representative value. This study also examines the implementation of Automatic Term Recognition (ATR) methods for the analysis of legal terminology and the automatic deployment of collocate networks. The first section explores such a controversial issue as establishing the ideal size for a specialised corpus applying the type/term ratio to a corpus of judicial decisions, the *BLaRC*, used as reference. In section 3, the assessment of different Automatic Term Recognition (ATR) methods is described. Out of five different methods, Drouin's (2003) *TermoStat* is found and recommended as the most efficient one in legal term mining. Finally, sections 4 and 5 demonstrate the practicality of collocate networks (Williams, 1998; 2001) in their capacity to reveal lexico-grammatical patterns which provide plenty of information for the study of legal text. A case study of the sub-technical legal term *party* using *Lancsbox* – designed by Brezina, McEnery & Wattam (2015) – is presented in section 5.2, where its general and specialised contexts are examined. Such scrutiny brings to the foreground interesting data such as the relevance of marriages of convenience in a collection of judicial decisions.

Keywords

Legal English, Corpus Linguistics, Terminology, Automatic Term Recognition, Collocate Networks, Lancsbox

Submitted: 27 October 2016, accepted: 31 July 2017, published online: 13 August 2017

[*] University of Murcia, Spain, mariajose.marin1@um.es.

DOI: 10.14762/jll.2017.018

1. Introduction

As commonly agreed by scholars, legal English (also known as *legalese*) is a peculiarly obscure and convoluted variety of English. David Mellinkoff, one of the first scholars devoted to the study of *legalese*, affirms that "the language of the law has a strong tendency to be: wordy; unclear; pompous [and] dull" (Mellinkoff, 1963: 63). The presence of Latin borrowings and Old French phrases, synonyms, archaisms and redundancy, as well as the widespread use of "common words with uncommon meanings" (Mellinkoff, 1963: 11) characterise its lexicon.

Traditionally, most of the work devoted to the description of legal English features (Mellinkoff, 1963; Alcaraz, 1994; Tiersma, 1999; Borja, 2000) has been either based on the authors' knowledge and intuitions on the subject or on relatively reduced language samples. These studies have often presented a top-down characterisation of the major traits of this ESP variety, following a deductive approach whereby the rule usually precedes the actual description of the examples provided. Nevertheless, there is a growing tendency towards corpus-based and corpus-driven[1] descriptions of *legalese* which provide a bottom-up characterisation of this ESP branch (Marín & Rea Rizzo, 2012; Biel & Engberg, 2013; Goźdź-Roszkowski & Pontrandolfo, 2014; Breeze, 2015).

Scholars have profusely discussed the advantages and disadvantages of employing language corpora as a source of information for linguistic analysis (Sinclair, 1991; McEnery & Wilson, 1996; Dudley-Evans & St. John, 1998; Kennedy, 1998; McEnery, Xiao & Tono, 2006; Tognini-Bonelli, 2001; Gries & Wulff, 2010). The Chomskyan distinction between competence and performance stands at the very basis of the earliest criticism against this discipline, which can be traced back to the 50s and 60s. Following Chomsky (1965), intuitive examples, as traditionally formulated by linguists, reflect linguistic competence as they arise from our tacit knowledge of the system and should serve as dependable references to base language theory upon. Conversely, those examples taken from corpora reflect performance, which usually mirrors competence poorly. As Chomsky puts it,

> "the problem for the linguist (...) is to determine from the data of performance the underlying system of rules that have been mastered by the speaker-hearer and that he puts to use in actual performance" (1965: 4).

Along these lines, some authors supporting this attitude have often deemed corpus samples skewed, frequently leading the linguist to erroneous generalisations on the language and offering "truncated concordance lines [which] are examined atomistically" (Flowerdew, 2009: 395). However, as Widdowson (2000) acknowledges, neither purely intuitive approaches to language description nor those based uniquely on

[1] In corpus-based linguistic studies a query is formulated in advance so as to find evidence in a corpus, whereas corpus-driven analyses base their conclusions solely on linguistic findings obtained from corpora and adopt an inductive approach to language description.

DOI: 10.14762/jll.2017.018

Corpus Linguistics are complete without each other. As a matter of fact, what the latter can do

> "is reveal the properties of text, and that is impressive enough. But it is necessarily only a partial account of real language. For there are certain aspects of linguistic reality that it cannot reveal at all. In this respect, the linguistics of the attested is just as partial as the linguistics of the possible" (Widdowson, 2000: 7).

In spite of earlier criticism and due to the fast growth of corpora and processing software nowadays, researchers can rapidly access and analyse large amounts of data that could have not even been thought of in the 50s and 60s. Tools like *Sketch Engine* (Kilgarriff et al., 2014) allow the user to search keywords, collocate patterns (sketches) and concordance lines employing as reference gigantic corpora like *enTenTen12*, of 12 billion words. Such plethora of data grants the reliability of the conclusions drawn from the observation of the language samples thus obtained, although the degree to which corpus data should be employed as the only source to base language description upon still remains an open question. In our view, intuition should go hand in hand with data collection, as remarked by Partington (1998), and aid the researcher, for instance, to discard ungrammatical examples. Similarly, the direct observation of the data can also contribute to the confirmation of hypotheses or *a priori* formulated theories and call our attention to new aspects of the language that could not be detected otherwise.

The applications offered by Corpus Linguistics to the study of general and specific languages are manifold, allowing for a descriptive approach to real language usage and also for the processing of large amounts of text. Nevertheless, the techniques and tools available may not always be well-known or easy to handle for non-specialists in the field such as law practitioners or linguists not accustomed to using corpora as part of their research methodology. This study was thus conceived as an introduction into this linguistic discipline for the analysis of legal English, especially aimed at those researchers unfamiliar with the wide array of corpus analysis tools available and the number of possibilities they offer.

Section 2 of this paper offers a general overview on such fundamental questions related to corpus design as how to determine the ideal size of a corpus or how to structure it. Additionally, section 3 presents a reflection on the usefulness of automatic term recognition tools by assessing their efficiency in legal term extraction. In section 4, the work by Williams (1998; 2001) and Brezina, McEnery & Wattam (2015) on collocational networks is presented. The article concludes with a case study of the term *party* in the general and the specialised fields using the software package *Lancsbox* (Brezina, McEnery & Wattam, 2015), which enables the user to obtain the lexical network of a given word/term and extend its context of usage up the seventh collocational level.

The three research questions (RQs) which motivated this study are the following:

RQ1: What key issues must be considered in the design and compilation of a legal corpus? How can they be tackled?

RQ2: How can automatic term recognition methods contribute to the study of legal texts? Can we trust these methods as dependable tools to rely on?

RQ3: How can collocation patterns add to the study of legal texts? Are there any automatic tools which facilitate such task?

2. Corpus description and justification

Answering the first research question on the most relevant issues to be considered in the design and compilation of a specialised corpus and how to tackle them is not an easy task. There seems to be general agreement on the importance of applying the appropriate sampling strategies in the selection of texts, since using a reliable method in corpus design is fundamental for the results obtained from its analysis to be representative of a given language variety. Biber (1993; 1998), McEnery & Wilson (2001), Sinclair (2005), McEnery, Xiao & Tono (2006), Tognini-Bonelli (2001) or Gries & Wulff (2010), to name but a few, provide a detailed insight into such and other issues, which are seminal in Corpus Linguistics. Following these authors, there are questions such as establishing the word targets or considering the communicative relevance of the text types included in a corpus that must be carefully tackled in its design and compilation.

This section presents a discussion on some of these issues[2] and the decision-making process in the design of the *British Law Report Corpus* (*BLaRC* henceforth), the legal text collection employed in this research.

2.1. Communicative relevance of law reports in common law legal systems

The *BLaRC*,[3] an 8.5 million word legal English corpus containing 1,228 legal texts, is a collection of British law reports issued by British courts between the years 2008 and 2010. Law reports are collections of judicial decisions or judgments which stand at the very core of common law systems and act as the main source of law followed by statutes and equity, hence their relevance within the British system. Following Sinclair, "the contents of the corpus should be selected […] according to their communicative function in the community in which they arise" (in Wynne, 2005: 5), a statement which insists on the aptness of this genre for the compilation of a legal corpus.

[2] See Marín & Rea Rizzo (2012) for further details.

[3] The corpus is freely available online at http://lextutor.ca/conc/eng and http://flax.nzdl.org/greenstone3/flax
.

The United Kingdom belongs to the realm of common law, where judicial decisions are based on previous cases always abiding by the doctrine of *stare decisis* (to stand by what has previously been decided) or principle of biding precedent. The decisions made by a higher court should act as binding precedent as long as they are related to the case in question in their essence. Determining what the essence of a given case is, that is, establishing the *ratio decidendi*, is part of the judge's role. "Cases must be decided the same way when their material facts are the same, [...] but the legally material facts may recur and it is with these that the doctrine is concerned", according to Williams (in Bhatia, 1993: 128). Nevertheless, judges are also subject to statutory principles, which must be interpreted whenever applicable and also act as a source of law. Statutory law has gained relevance as a major legal source in the UK in the last 150 years (Geary & Morrison, 2012; Orts, 2006), even so, law reports still stand at the very basis of the legal system and legal practitioners must know them well.

Actually, law reports must be cited and act as one of the essential elements which lawyers build their arguments upon and judges base their decisions on. This is why, in the UK, they are made public through different institutions, both public and private, i.e., the Incorporated Council of Law Reports of England and Wales (ICLR) or publishing houses like Butterworth or Lloyds. Due to the widespread use of information technologies, there is a tendency towards digitalising these texts and storing them in online databases. The British and Irish Legal Information Institute (bailii.org) offers an open-access online database where the judicial decisions made at British courts (as well as many other documents from various sources) can be consulted and downloaded.

As regards the generic classification of law reports, it varies depending on the perspective adopted for their analysis. Law reports may appear in generic classifications as part of the oral mode (Danet, 1980), within the category "recording and law making" (Maley, 1994) or as public unenacted law (Orts, 2009), amongst others.

Another relevant communicative function of law reports, as highlighted by Bhatia (1993) and Nesi & Gardner (2012), is the role they play within Higher Education. Becoming a solicitor or a barrister in the UK requires passing a hard process of accreditation which law faculties prepare students for. Amongst many other requirements, the suitors must be able to write case reports, thus having to apply and cite law reports as the major source to base their arguments on. Writing case reports is not only part of their training but also of their professional activity although only barristers can "be called to the bar", that is, argue a case in court on behalf of their clients.

Finally, law reports are rather comprehensive texts since they not only cover all the branches of law, but also present full sections of other legal texts such statutes, wills, contracts, deeds and the like. Nesi & Gardner (2012: 177) provide a description of the macrostructure of law reports which follow four principal stages:

i) case identification;
ii) case facts;

iii) arguing of the case (case history, presentation of arguments, *ratio decidendi*), and
iv) judgement.

Citing sections of statutes or the contents of some other private documents is a usual procedure when arguing a case, hence the relevance of this legal genre not only from a legal but also from a linguistic point of view if a terminological study (like the one presented below) is to be carried out.

2.2. Corpus size and representativeness: establishing the word target

Representativeness is vital in corpus design. Douglas Biber (1993) – a fundamental reference in this field – refers to the crucial role performed by corpus sampling strategies, which may be decisive to determine whether a corpus is representative of the variety of the language it aims at covering or simply an illustrative sample of it with no predictive value. Biber insists on the transcendence of this issue owing to the fact that "representativeness refers to the extent to which a sample includes the full range of variability in a population" (Biber, 1993: 246).

Therefore, the concept *representative*, as defined by Biber, points at two major questions, on the one hand, the capacity of a corpus to comprise the different textual types in a given variety or language and, secondly, its ability to account for variation within it. For the design of the *BLaRC*, which was created primarily to identify and analyse its legal terminology implementing different automatic methods, a decision was made to focus solely on law reports, given their relevance within the British legal system in comparison with other legal text types, as stated above. Furthermore, law reports touch upon all areas of law so the corpus was structured according to the field the texts pertained to so as to be able to account for terminological variation across legal areas.

Nevertheless, the question whether a specialised corpus is big enough to be representative of a given variety of the language, even if it is balanced and well sampled, still remains open to debate. There seems to be no clear agreement concerning the recommended size for a specialised corpus basically due to the fact that most approaches to this question are made on a theoretical basis. Whereas Pearson (1998) proposes a million words as a reasonable number (she poses that the limit should rather be established by the number of texts available and convertible into digital format), Sinclair (1991) believes that corpora must be as large as possible, establishing 10 to 20 million words as the recommendable target for a specialised one.

On the other hand, Kennedy (1998) does not consider that a big corpus necessarily represents the language better than a small one. In addition to this, Flowerdew underlines that the size of a specialised corpus necessarily depends on the aim the corpus has been designed for, given that "specialised corpora are constructed with an *a priori* purpose in mind" (Flowerdew, 2004: 25). Nevertheless, only a few authors draw their conclusions in this respect from actual data. Heaps (1978), Sánchez & Cantos Gómez (1997) or Corpas Pastor & Seghiri Domínguez (2010, citing Young-Mi 1995) propose measures to try and determine the most suitable size for a corpus.

Regarding the size of the *BLaRC*, an *a priori* decision had to be made for its compilation, since finding out about such data as type/token or term/type ratios to establish a word target based on actual data would require the existence of the corpus itself prior to its processing. Consequently, and following Biber's criteria on sampling and Sinclair's recommendations on specific corpus size, the initial target was set at 8.5 million words. As described in section 2.3, there were other external criteria which conditioned the structure and content of the corpus itself.

Following Sánchez & Cantos Gómez' (1997) study, which aims at formulating a method to try to determine the optimum size for a corpus to be representative of given language variety based on how the type/token ratio[4] progresses as the corpus grows bigger, type/term increase was measured in the *BLaRC*. Finding out the proportion of new terms appearing in a corpus as its size augments might be an objective way of determining whether the size of that corpus would suffice to study its terminology, as is the case with the *BLaRC*.

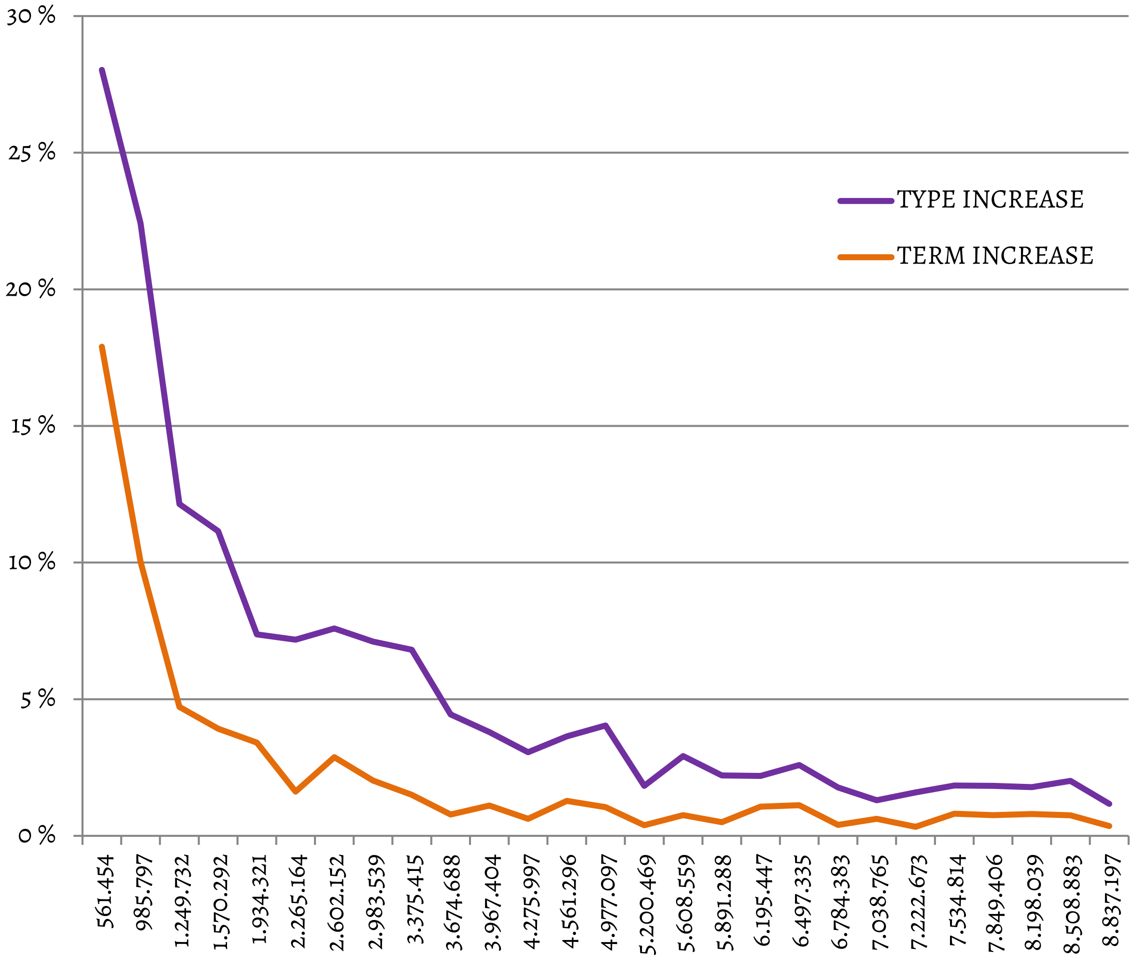
The terms in the *BLaRC* were first extracted automatically using Drouin's software *TermoStat* (2003) and then validated by comparison with a specialised legal English glossary of 10,088 terms.[5] Both the glossary and the lists generated by *TermoStat* (after progressively bringing together the 27 sub-corpora the main corpus was divided into) were compared using an excel spreadsheet so as to find out how many new terms appeared as new sub-corpora were added to the main corpus. The graph in Figure 1 illustrates the type/term ratio in the *BLaRC*, that is, how the percentage of terms and types, on the y-axis, relates to the total number of tokens in it. As can be observed, the former is inversely proportional to the latter, on the x-axis.

Figure 1 clearly illustrates how types and terms behave similarly, reducing their number as the corpus augments its size. Concerning the proportion of new terms appearing as the corpus grows bigger, they experiment a dramatic drop of 12.3 points from 17 % to 4.7 % as the corpus doubles its size from 500.00 words to 1.2 million approximately. Once the corpus reaches 1.2 million tokens, the decrease of new terms is less sharp falling from 10.03 % to 4.72 %. From that point on, although slightly recovering, this percentage drops to 1.62 % for sub-corpora 1 to 7 (2.26m tokens). It remains constant at 1.02 % on average until the corpus grows to 6.78 million words, decreasing to 0.4 % and not experimenting any significant changes from that point on.

[4] *Types* could be defined as the different words found in a corpus and the tokens associated to them through the type/token ratio coefficient are the repetitions of the same word within that corpus.

[5] Merged from four online legal glossaries available at www.legislation.gov.hk/eng/glossary/homeglos.htm, www.judiciary.gov.uk/glossary, sixthformlaw.info/03_dictionary/index.htm, and www.nolo.com/dictionary.

Figure 1: Type/term increase in the *BLaRC*.



Note: The x-axis represents the number of tokens.

Judging from the above, it appears that the initial target established for a corpus like the *BLaRC* may suffice to attain the objectives set for its compilation, that is, to analyse its terminology applying different automatic text analysis tools. As a matter of fact, 2.6 million words would have been enough due to the low increase in the percentage of new types and terms appearing as the corpus grew bigger. This is the reason why a pilot corpus of that size (*The United Kingdom Supreme Court Corpus*) was extracted from the *BLaRC* in order to facilitate the implementation of the methods described in section 3 and the analysis of the data.

2.3. Distributional criteria and word targets per category

The number of texts comprised by the *BLaRC* is not evenly distributed amongst its categories (which follow the geographic and hierarchical distribution of the courts and tribunals in the UK). Great variation was found depending on the text source (court or tribunal[6]). The reasons for the irregular distribution of the texts available are varied, in some cases, especially regarding tribunals, they may have started working recently or disappeared due to the *Tribunals, Courts and Enforcement Act, 2007*. In some others, the high figures coincide with a densely populated area (one of the criteria supporting text distribution within the corpus) or with a court whose decisions, due to its high status in the hierarchy (i.e. any of the chambers of the High Court of Justice of England and Wales), set binding precedent and may thus be more relevant for legal practitioners when it comes to arguing a case.

Nevertheless, the targets established for the sections and subsections of the corpus were kept proportional to the total number of texts available within the covered time span. Subsequently, the sub-targets were set according to this criterion: if the number of texts in a section was higher, they were assigned a larger word target, thus being more representative of the language variety as that is the proportion they keep in real life, or at least this was assumed to be so.

These decisions were made following Biber's (1993; 1998) recommendations so as to ensure the ability of a corpus to represent a variety of the language properly. When designing the corpus itself, researchers should bear in mind variability, which "can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness" (Biber, 1993: 247). The geographical and institutional criteria that influenced the structure of the corpus above might fall within the "situational" perspective, according to Biber, whereas thematic and terminological criteria could be classified as linguistic.

All the same, a corpus should not be intended to systematise reality in a mathematical way, in this case, we simply intended to be as coherent as possible in every step we took towards corpus design. As Sinclair puts it when dealing with the issue of sampling a corpus and the structural criteria to employ when designing it: "real life is rarely as tidy as this model suggests" (Sinclair, 2005: 3). Moreover,

> "We remain (…) aware that the corpus may not capture all the patterns of the language, not represent them in precisely the correct proportions. In fact, there are no such things as "correct proportions" of components of an unlimited population" (Sinclair, 2005: 4).

Having said so, the total number of texts available between 2008 and 2010 was 16,612. Therefore, the word targets were established with respect to it, as already stated. Table 1 shows how this distribution was organised for the section devoted to those texts coming from English and Welsh institutions, by showing the total number of texts available per sub-category, their percentage with respect to the total amount of texts and the corresponding word target achieved following this proportion.

[6] Note that "the essential difference between a tribunal and a court is that a tribunal does not administer any part of the 'judicial power of the state'. It has a specific jurisdiction as allocated by Parliament and does not enjoy a broad jurisdiction defined in general terms" (Geary, 2012: 51).

Table 1: England and Wales courts and tribunals.

Court / Tribunal	available texts	% of total	final word target
Court of Appeal (Civil Division)	2,640	15.89 %	956,398
Court of Appeal (Criminal Division)	1,136	6.84 %	414,683
High Court (Administrative Court)	2,039	12.27 %	731,693
High Court (Admiralty Division)	17	0.11 %	8,842
High Court (Chancery Division)	1,009	6.07 %	366,298
High Court (Commercial Court)	379	2.28 %	142,701
High Court (Court of Protection)	26	0.16 %	34,007
High Court (Senior Costs Off.)	70	0.43 %	29,302
High Court (Family Division)	199	1.20 %	84,557
High Court (Mercantile Court)	8	0.05 %	6,152
High Court (Patents Court)	105	0.64 %	40,420
High Court (Queen's Bench Division)	709	4.27 %	255,301
High Court (Technology and Construction Court)	284	1.71 %	101,066
Patents County Court	12	0.08 %	15,242
Magistrates' Court (Family)	98	0.59 %	33,680
County Court (Family)	56	0.34 %	20,702
Care Standards Tribunal	70	0.43 %	27,762
Lands Tribunal	115	0.70 %	44,004
Total	**8,972**	**54.06 %**	**3,322,810**

Note: The final word target was obtained by calculating the number of words which the percentage displayed in the third column represented with respect to the initial word target, 8.5 million words. In order not to include truncated versions of some of the decisions in each section, the final word target sometimes exceeded the expected size slightly, respecting the actual length of the decisions comprised in the corpus.

3. Applications of CL techniques to the study of *legalese*: Automatic Term Recognition methods

Once the corpus has been properly compiled and structured, the applications to the study of the language samples comprised in it are manifold. Amongst other, we find Automatic Term Recognition (ATR henceforth). Yet, as stated in research question number 2: How can ATR methods contribute to the study of legal texts? Can we trust these methods as dependable tools to rely on?

To begin with, ATR methods can become extremely useful tools for the researcher interested in handling large amounts of information that could not be processed manually. In fact, getting to know the most significant terms in a corpus of specialised texts can definitely contribute to a better understanding of the texts themselves, since terms could be defined as "linguistic representations of domain-specific key concepts in a subject field that crystallise our expert knowledge in that subject" (Kit & Liu, 2008: 204) and also lead to the identification of relevant topics that would otherwise remain unnoticed. In sum, specialised terms could be regarded as conceptual vehicles which can be employed to transmit specialised knowledge amongst scientists, researchers, or professionals in all specialised areas, hence their relevance and the need to identify them within a text collection. Actually, mining the specialised terms from a text collection might be the point of departure for further enquiry into the texts in a corpus by focusing, for instance, on collocate patterns (either as pairs of collocates of collocate networks), as shown in the last sections.

In order for ATR methods to be trusted as useful tools for term mining, and given the peculiar statistic behaviour of legal terminology, it becomes necessary to test them in order to select the most efficient ones in legal term extraction. It is commonly acknowledged that legal English is deeply intertwined with general language (Alcaraz, 1994; Borja, 2000; Mellinkoff, 1963; Tiersma, 1999), displaying specific features such as the abundance of sub-technical terminology, in other words, of "common words with uncommon meanings", (Mellinkoff, 1963) whose frequency and distribution might often be similar in the general and specific fields. ATR methods resorting to corpus comparison employ such parameters as frequency and distribution to perform their function. If a given term behaves similarly (in statistical terms) in both contexts, an ATR method implementing corpus comparison may be likely to fail or be less efficient and produce output lists of candidate terms that might contain a high percentage of noise (of false terms).

Consequently, ATR methods must be tested so as to identify the most effective ones in legal term recognition. In the past, the literature on ATR methods and software tools has been profusely reviewed (Maynard & Ananiadou, 2000; Cabré Castellví, Estopà Bagot & Vivaldi Palatresi, 2001; Drouin, 2003; Lemay, L'Homme & Drouin, 2005; Pazienza, Pennacchiotti & Zanzotto, 2005; Chung, 2003; Kit & Liu, 2008 or Vivaldi et al., 2012, to name but a few) often classifying these methods according to the type of information used to extract candidate terms (CT) automatically. One of the research foci of these works is the level of efficacy such methods can reach, concentrating on the amount of true terms (those terms confirmed as such after validation) they are capable of identifying automatically. In general, the most widespread procedure to determine the efficacy of ATR methods consists in comparing the list of CTs identified by each of them against a gold standard, that is, a glossary of specialised terms which ATR method designers employ as reference.

In Marín (2014; 2015) we find the evaluation of ten different ATR methods leading to the identification of the most efficient ones in the legal field. Table 2 displays the rate of efficiency reached by those ATR methods devoted solely to single-word term recognition. The figures show that it is Drouin's (2003) method which manages to successfully extract a greater rate of legal terms both on average (73 % of the terms identified were confirmed as true terms) and also for the top 200 candidate terms in the output lists (88 % of these were confirmed as legal terms).

Table 2: Average precision reached by SWT recognition methods (Marín, 2015: 11).

ATR Method	Avg. Precision 2,000 CTs	Precision Top 200 CTs
TermoStat (Drouin, 2003)	73.0 %	88.0 %
Kit and Liu (2008)	64.0 %	84.0 %
Keywords (Scott, 2008)	62.0 %	85.0 %
TF/IDF (Sparck Jones, 1972)	57.4 %	74.5 %
Chung (2003)	42.5 %	48.5 %

Note: ATR = Automatic Term Recognition; CT = Candidate Term.

The assessment process carried out by Marín (2014; 2015) consisted in the automatic validation of the candidate term lists produced by each method against a legal English glossary used as gold standard (see footnote 5 on the description of the glossary). The output lists were compared with the gold standard using an excel spreadsheet with the aim of determining the overlap percentage existing between both lists. Whenever a candidate term was found in the glossary, it was confirmed to be a true term. Therefore, the percentages found in the table above could be interpreted as the average level of precision achieved by each of the evaluated methods.

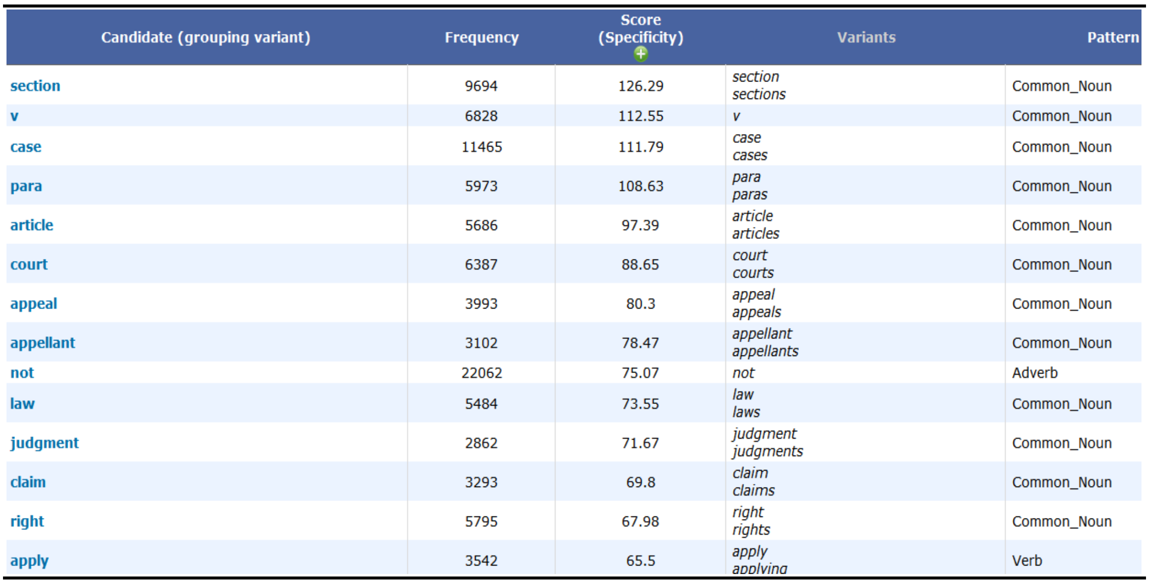
As regards Drouin's *Termostat* (2003), it is based on previous work on lexicon specificity such as Muller's, Lafon's, or Lebart & Salem's (in Drouin, 2003). Drouin claims that the frequency of technical terms in a specialised context differs, in one way or other, from the same value in a general environment and that "focusing on the context surrounding the lexical items that adopt a highly specific behaviour [...] can help us identify terms" (Drouin, 2003: 100). This author uses a corpus comparison approach which provides information on a candidate term's standard normal distribution giving

> "access to two criteria to quantify the specificity of the items in the set [...] because the probability values declined rapidly, we decided to use the test-value since it provides much more granularity in the results" (Drouin, 2003: 101).

Drouin applies human and automatic validation methods to evaluate the levels of precision and recall of his method. The author also resorts to three specialists who identify the true terms (TT) from the list generated by *TermoStat* noticing that subjectivity played a relevant role in this evaluation phase and that it might also be interesting to study human influence on validation processes. Regarding automatic validation, he compares the lists of CTs with a telecommunications terminology database. *TermoStat* reaches 86 % precision in the extraction of SWTs.

The ATR method designed by Drouin (2003) offers a user-friendly online interface,[7] which allows the researcher to upload their corpus (it accepts French, English, Spanish, Italian and Portuguese texts) and process it easily, obtaining the ranked list of candidate terms and other useful information for the analysis of the terminology comprised in it. Once the corpus is processed (it allows for the upload of files up to 30 megabytes), *TermoStat* produces a list of lemmatised[8] terms which are ranked according to their level of specialisation. Drouin's method resorts to corpus comparison for term extraction, using a reference corpus of newspaper articles as the general language corpus.

Figure 2: Output list of candidate terms extracted by *TermoStat*.



Candidate (grouping variant)	Frequency	Score (Specificity)	Variants	Pattern
section	9694	126.29	*section* *sections*	Common_Noun
v	6828	112.55	*v*	Common_Noun
case	11465	111.79	*case* *cases*	Common_Noun
para	5973	108.63	*para* *paras*	Common_Noun
article	5686	97.39	*article* *articles*	Common_Noun
court	6387	88.65	*court* *courts*	Common_Noun
appeal	3993	80.3	*appeal* *appeals*	Common_Noun
appellant	3102	78.47	*appellant* *appellants*	Common_Noun
not	22062	75.07	*not*	Adverb
law	5484	73.55	*law* *laws*	Common_Noun
judgment	2862	71.67	*judgment* *judgments*	Common_Noun
claim	3293	69.8	*claim* *claims*	Common_Noun
right	5795	67.98	*right* *rights*	Common_Noun
apply	3542	65.5	*apply* *applying*	Verb

As shown in Figure 2 the output list includes not only is the term's specificity value (spécificité) but also its frequency as lemma (fréquence), its variants (variants ortographiques), and its part-of-speech tag (matrice). The lexical categories identified by *TermoStat* are: nouns, adjectives, adverbs and verbs. It also detects multi-word terms having nouns and adjectives as phrase heads.

Table 3 displays the top 25 candidate terms (prior to the validation of the method) as ranked by *TermoStat* according to its level of specialisation, or specificity level, that is, after implementing the algorithm designed by the author. As it can be observed in the table below, not all the terms identified by the system could be regarded as legal terms proper. As already stated, this table includes all the candidate terms Drouin's method managed to extract before the whole list was validated against our legal glossary. We decided to offer this data for the reader to acknowledge the possibilities at hand using

[7] Online at http://termostat.ling.umontreal.ca.

[8] The term *lemma* refers to the root word without any inflectional suffixes (for instance, the infinitive of a verbal form). Lemma frequency includes all the occurrences of any of the possible realisations of the root word. Those methods which resort to lemmatisation tend to be more efficient than those which do not.

this term extraction method, which managed to identify 88 % legal terms out of the top 200 candidate terms extracted automatically from the *BLaRC*.

Table 3: Top 25 terms as identified by Drouin's *TermoStat*.

Rank	Term	Specificity level	Rank	Term	Specificity level
1	section	126.29	14	order	64.39
2	v (versus)	112.55	15	decision	63.53
3	case	111.79	16	person	62.83
4	para (paragraph)	108.63	17	proceeding	61.70
5	article	97.39	18	relevant	59.02
6	court	88.65	19	purpose	58.45
7	appeal	80.30	20	defendant	57.72
8	appellant	78.47	21	provision	57.55
9	law	73.55	22	principle	55.77
10	judgment	71.67	23	application	55.50
11	claim	69.80	24	jurisdiction	55.50
12	right	67.98	25	paragraph	54.69
13	apply	65.50			

4. Term collocates and lexical networks: Williams (2001) and Brezina, McEnery & Wattam (2015)

Closely linked to the automatic identification of specific terms is the relevance, not only of the terms themselves, but also of other words which tend to co-occur with them, that is, their collocates. Yet, going back to the research questions posed in the introduction, how can such patterns contribute to the study of legal text? Are there any automatic tools which facilitate such task?

Collocational patterns reveal the context in which a word occurs and provide plenty of information about the meanings and connotations associated with a word in context. When it comes to sub-technical or polysemous terms, their collocates can also help us distinguish between their specialised and general meaning but, most importantly, can point at other questions that may remain unnoticed on a superficial reading of legal texts. Nevertheless, for the identification of collocational patterns in a text collection, especially if it is a large corpus, it is necessary to employ automatic tools that facilitate the task. Let us first define and consider some theoretical questions related to the concept of collocation and then move onto the actual usage of collocation extraction software and its applications to the study of legalese.

Broadly speaking, in Firth's words, a collocate is "the company a word keeps" (1957: 6). The concept *collocation* has been revisited since then (Cruse, 1986; Gries, 2013; Sinclair, 1991; Stubbs, 2001) and more specific and accurate definitions have been provided, John Sinclair's being a classic reference in the field. Sinclair (1991; 2005) deems the statistical data associated with two co-occurring words as fundamental for their identification, as collocates can be mined automatically by applying measures of association like mutual information (Church & Hanks, 1990) or log-likelihood (Dunning, 1993), amongst others. Williams elaborates on this idea by delimiting the concept of collocation as

> "the habitual and statistically significant relationship between word forms within a predefined window and for a defined discourse community, expressed through an electronic corpus of texts" (2001: 5).

On a semantic level, based on the work by Stubbs (2001) on semantic preference and discourse prosody, Baker (2016: 2) insists on the mutual influence that collocates have on each other as regards their meaning, affirming that "collocates help to imbue words with meaning as words can begin to take on aspects of the meaning of the words that they collocate with".

However, as Baker (2016) acknowledges, the study of collocates has been limited to the analysis of word pairs until recently, often due to the limitations of tools like *AntConc* (Anthony, 2014) or *Wordsmith* (Scott, 2008), only capable of extracting pairs of collocates, disregarding the potentiality of collocational or lexical networks (Williams, 2001) in the study of the interaction amongst terms and their vicinity in a corpus.

Geoffrey Williams (2001) is one of the first authors to explore word associations beyond word pairs in specialised contexts based on the work by Phillips (cited in Williams, 2001). Williams proposes the lexical network model, which puts forward a quantitative approach to the study of word usage through the analysis of their collocates and co-collocates. The context is thus extended since lexical networks, which revolve around a central word or node, spread out progressively by also including the node's co-collocates and, in turn, the collocates of those co-collocates.

Williams' (1998) idea that collocational or lexical networks may enhance quantitatively and, above all, qualitatively our understanding of specialised vocabulary meant a step forward in the study of term usage and meaning and authors like Baker (2005; 2016), McEnery (2006) or Marín (2016) acknowledge this fact. However, in spite of the above, the process undergone in the production of lexical networks could be time consuming, as Baker (2016) and Marín (2016) affirm, requiring the manual arrangement of the networks (often populated by thousands of elements), since automatic corpus tools only allow for the study of one collocational level.

There is a plethora of tools capable of processing electronic text designed with different purposes (Sternfeld, 2012) although not many of them can obtain the lexical networks of a term automatically. This is the case of *Voyant Tools* (Sinclair et al., 2012) and *Lancsbox* (Brezina, McEnery & Wattam, 2015). Both offer plenty of possibilities to exploit corpora. The former is extremely powerful in loading large amounts of text online and offers very visual applications like *Cirrus, ScatterPlots* or *TermsRadio*, amongst other. Nevertheless, as regards collocate networks, the proposal by Brezina, McEnery & Wattam's (2015) proposal appears to be grounded and motivated by more solid linguistic criteria, allowing for a deeper analysis of the collocate networks of terms. It goes further than *Voyant Tools* into the contexts of usage not only of the central node of the networks but also of its collocates and co-collocates. Furthermore, *Lancsbox* implements the possibility of modifying the measures applied to obtain a word's collocates and thus test the efficacy of the tool in producing relevant collocate inventories, depending on the users' preferences.

One of the advantages of using *Lancsbox*[9] is that it not only manages to obtain a word's network very quickly, but also visually represents the network through a graph that displays the node's collocates, connecting them with vectors whose size varies according to the strength of the collocational bond calculated by the tool (the shorter the vector, the stronger the link between words) and indicating collocate directionality. *Lancsbox* also presents the possibility of adjusting association measures by testing which one produces the most interesting results. Amongst other, measures such as *MI3, delta-p* or *log-likelihood* can be implemented in the production of a word's lexical network, represented by a graph, as shown below.

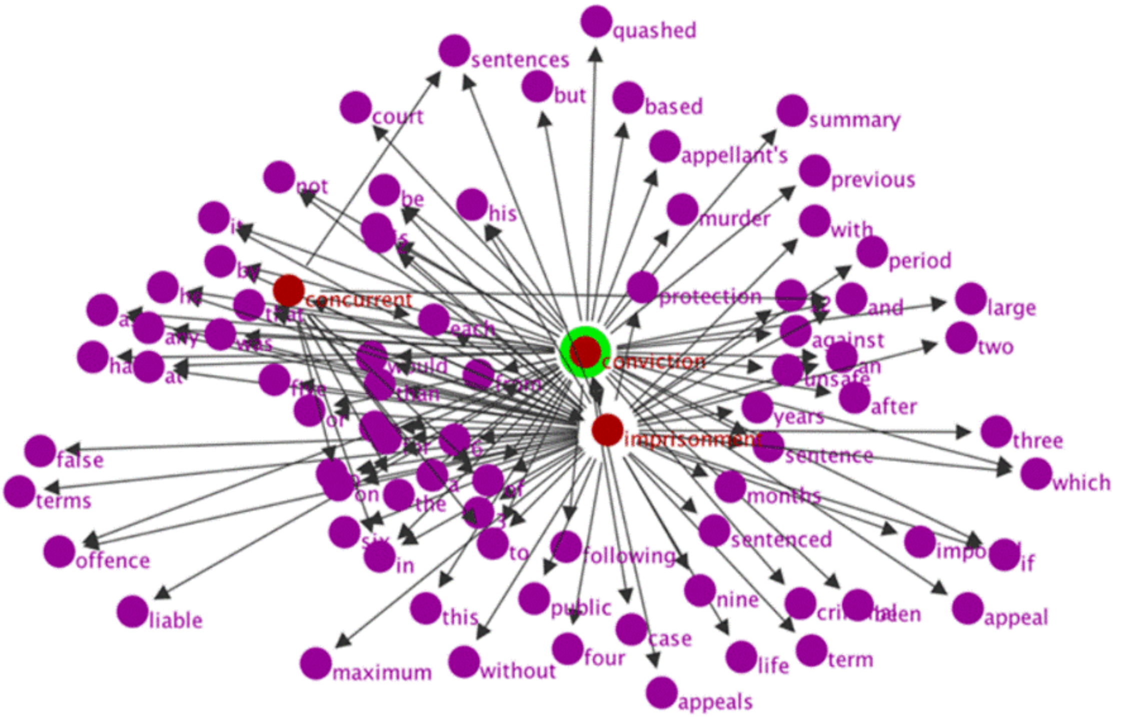
Once they are obtained, the graphs contain detachable tabs, which permit the user to generate embedded collocate networks, always displaying the relationship amongst all their constituents and the main node, as illustrated by Figure 3. If we click on any of the collocates (in purple), a new collocational level will be shown, which includes the collocate's collocates, that is, those words which tend to co-occur with each of the node's collocates. This can be done up to seven times, thus allowing for a subsequent development of the networks to the seventh collocational level.

As shown in Figure 3, which displays the collocational network of the term *conviction* (circled in green), it presents first level collocates such as *imprisonment, summary, appeal* or *sentence*. If we had not resorted to *Lancsbox,* the collocational network would have stopped at this point, however, this tool enlarges the context by displaying the collocates of *imprisonment* (in red), namely, *concurrent, conviction, sentence* or *protection* and of those words which also collocate with it, such as *concurrent* (the third sub-node, which constitutes the third collocational level in the network below). Whenever any of these share any collocates, they are linked with an arrow which indicates collocate directionality. Owing to the fact that the corpora employed in this study are considerably large (13.7 and 8.5 million words respectively), the networks might be excessively populated, as displayed in Figure 4. This is why the frequency thresholds must be adjusted to pre-

[9] Available at http://corpora.lancs.ac.uk/lancsbox/index.php.

vent this from happening. In any case, the tables appearing to the left of the graphs (as shown in Figure 4), once they are generated, allow the user to navigate through the whole collocate inventory easily.

Figure 3: Specialised collocational network of the word *conviction* (in *BLaRC*).



One of the advantages of *Lancsbox* is the possibility of adjusting the settings to limit the number of collocates in the networks or to change the association measures employed to mine them, as already stated. This is why Brezina, McEnery & Wattam (2015) perform a case study analysis where different measures are used in the replication of McEnery's (2006) examination of swearing language (the words *swearing* and *drunkenness* exemplify the study). In spite of all the multiple applications and advantages of *Wordsmith* (Scott, 2008), the software McEnery uses to extract the collocates in his study, it does not offer the possibility to implement MI3 (the cubed version of Church & Hanks' (1990) mutual information measure). In a nutshell, what mutual information does is basically to compare

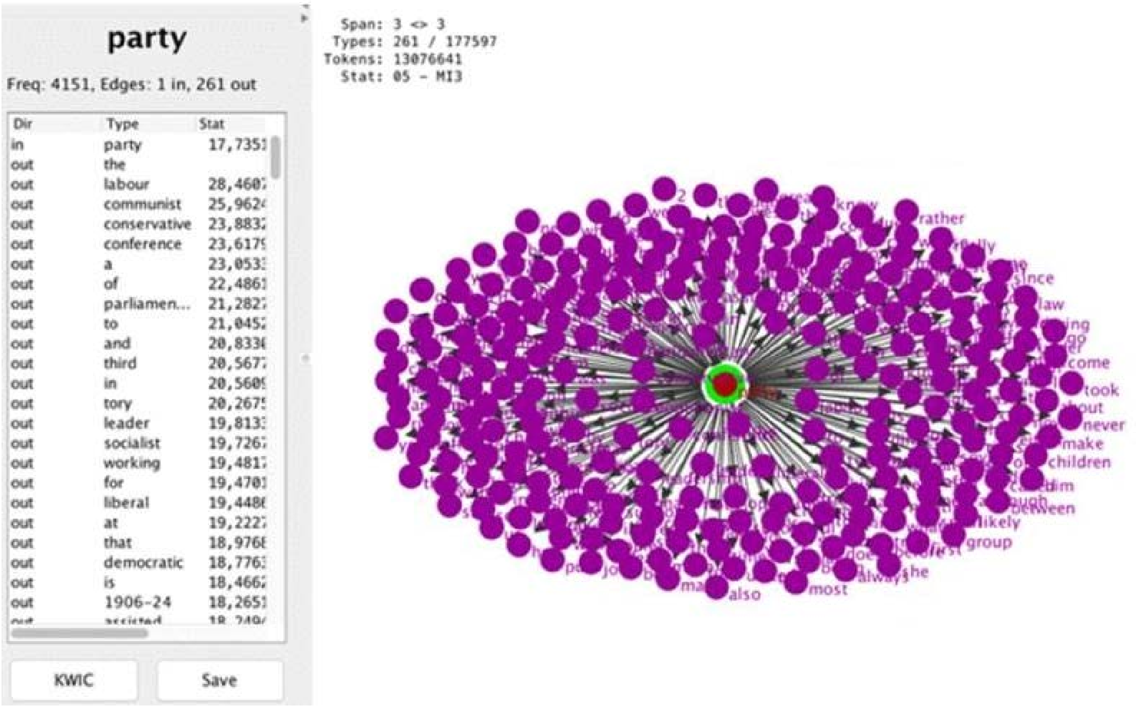
> "the probability of observing *x* and *y* together (the joint probability) with the probability of observing x and y independently (chance). If there is a genuine association between x and y, [...] then the joint probability will be much larger than chance" (1990: 77).

Therefore, if a collocate pattern was assigned a high MI score owing to its joint statistical behaviour, it would be identified as relevant within a given text collection.

As already stated, McEnery opts for mutual information (MI), highly precise, although it often shows a certain "propensity to highlight unusual combinations [...] that co-occur only once or twice in the corpus" (Brezina, McEnery & Wattam, 2015: 159). A collocate frequency threshold would thus become necessary for the networks not to become unmanageable and excessively populated if MI was to be applied. On the contrary, MI3 tends to push more frequent combinations to the top of the rank, leaving the most unusual patterns aside or either relegating them to the bottom of the collocate inventories, in other words, "the measure gives more weight to observed frequencies and thus gives high scores to collocations which occur relatively frequently in the corpus" (Brezina, McEnery & Wattam, 2015: 160).

The data associated with each of the constituents of the network can also be read in detail and saved in .csv format. The extension .csv stands for "comma separated values", which can be easily imported into an excel spreadsheet. As seen in Figure 4, a table displays the collocates of the selected item (highlighted in green in the graph) and also the value assigned to each pattern by the algorithm implemented through MI3 together with the raw and relative frequency of each pattern on the list.

Figure 4: *Lancsbox* table and graph as shown by the interface control panel.



party

Freq: 4151, Edges: 1 in, 261 out

Dir	Type	Stat
in	party	17,7351
out	the	
out	labour	28,4607
out	communist	25,9624
out	conservative	23,8832
out	conference	23,6179
out	a	23,0533
out	of	22,4861
out	parliamen...	21,2827
out	to	21,0452
out	and	20,8336
out	third	20,5677
out	in	20,5609
out	tory	20,2675
out	leader	19,8133
out	socialist	19,7267
out	working	19,4817
out	for	19,4701
out	liberal	19,4486
out	at	19,2227
out	that	18,9766
out	democratic	18,7763
out	is	18,4662
out	1906-24	18,2651
out	assisted	18,240

Having said this and leaving aside the fact that *Lancsbox* is capable of producing the lexical network of a term on the fly, which, on its own, is a major improvement, Brezina, McEnery & Wattam emphasise that the main potential of this software is its capability to unveil the semantic interaction amongst the words in a corpus by extending a word's context beyond the word itself and avoiding the painstaking and time-consuming process of doing it manually, as Baker (2016) and Marín (2016) also acknowledge.

5. Subtechnical legal terms and collocational networks: A case study

Following from the above, the applications of *Lancsbox* to the analysis of corpora and their lexicon are manifold. As Marín (2016) demonstrates in the proposal of an algorithm to study the level of specialisation of subtechnical vocabulary, the relevance and significance of this particular type of legal terminology in a corpus of judicial decisions was considerable. The comparison between the list of specialised legal terms extracted from the *British Law Report Corpus* and the list of the 3,000 most frequent words of English found in the *British National Corpus* (2007) yielded 45.41 % overlap, thus showing "that approximately half of the legal terminology identified in the *BLaRC* is shared with the general field, since almost 50 % of it matched the general vocabulary lists" (Marín, 2016: 81).

As shown in section 3, this is a common feature of the legal English lexicon, however, very little has been said about the meaning of these words in context. Words such as *trial*, *relief*, *battery* or *charge* (which are statistically profiled in Marín's analysis) present a specialised meaning in the legal context which very rarely occurs in the general one. Sections 5.1 and 5.2 present a case study illustrating the applications of Lancsbox to the study of subtechnical legal terms.

5.1. Methodology

Two corpora were employed in this analysis, one of them the *BLaRC* (8.5 million words), the other one *LACELL*, a 13.7 million word general English corpus containing texts from various British sources such as newspapers articles, book chapters (academic, fiction, etc.), magazine articles, brochures, letters and the like. Both corpora were processed using *Lancsbox* (Brezina, McEnery & Wattam, 2015). The thresholds established to limit the amount of collocates generated by the system were, firstly, >10 frequency, according to which, the pairs of collocates and co-collocates should co-occur at least 10 times in the corpus to be mined by the system. Secondly, the collocate window cut-off point was 3, that is, the collocates included in the network should fall within the three immediate words to the left and right of the node (the search word) or any constituent of the network. Following Brezina, McEnery & Wattam (2015) and Baker (2016), the association measure implemented for the calculation of the term's collocate network was MI3, whose capacity to leave irrelevant patterns aside by pushing them to the bottom of the collocate ranks has already been discussed.

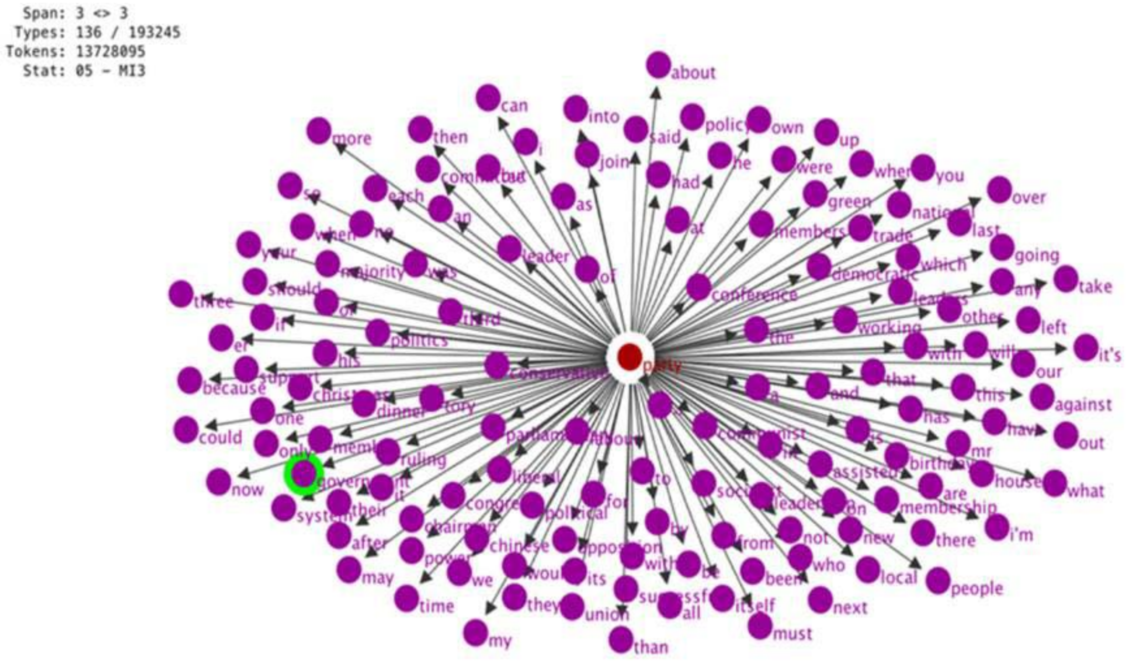
The word selected for this case study is *party*, a sub-technical word whose presence in both corpora is remarkable, hence its sub-technical character, displaying 4,808 raw frequency in the general corpus (3.5 relative frequency) and 40 % distribution (it appears in 1,712 out of 4,281 texts). In contrast, its frequency in the specialised corpus is 4 points higher than the same value in the general corpus (if we compare their relative frequencies), as it occurs on 10,351 occasions (7.5 relative frequency). In addition, it presents higher distribution values, covering 73 % of the texts in it.

Nonetheless, the major difference found between the use of *party* in the general context and the specific field, as might be expected, is related to its meaning in both areas. It is at this point that the software package *Lancsbox* can provide evidence of the context which surrounds the term, establishing which of its meanings in each corpus is the most representative one. The collocates associated to each of the senses of the word *party* (the main node of the lexical network obtained with *Lancsbox*), illustrate how the meaning of *party* can be understood as a "political group" or "celebration" or acquire its legal sense in the specialised corpus, meaning "person/s taking part in a legal proceeding".

5.2. Results and discussion

Figures 5 and 6 display the first level collocate networks of the term *party* in both the specialised and the general fields. In a first approach, and judging by the stronger lexical collocates of *party* in the general corpus (this is indicated by the shorter vector that joins them and by the coefficient displayed in the table attached to the graph, not in the figure), the primary meaning of the term is clearly "political group/association", in fact, the words *labour, communist, conservative, parliamentary, tory, leader or socialist* appear amongst the top 25 collocates identified by *Lancsbox.*

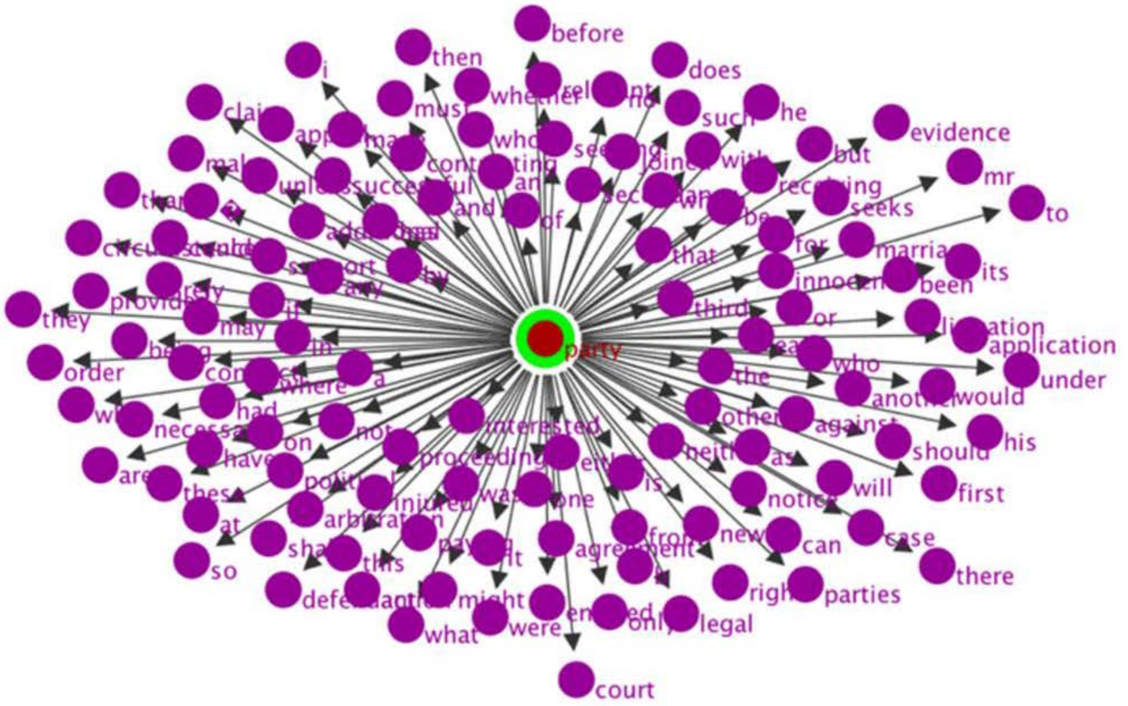
Figure 5: 1st level collocate network of *party* in the general corpus.



```
   Span: 3 <> 3
  Types: 136 / 193245
 Tokens: 13728095
   Stat: 05 - MI3
```

Evidence of its secondary general meaning, "celebration", much less frequent in *LACELL* than the former, was also found in the general corpus, as it was expected. Words like *dinner, cocktail, birthday, Christmas* or *tea* can be found within the top 125 collocates of *party*. On the other hand, its collocates in the legal corpus clearly signal its legal sense since we find words like *third, proceedings, innocent, agreement, contracting, marriage, aggrieved* or *arbitration* ranking amongst the top 50 collocates of the term.

Figure 6: 1st level collocate network of *party* in the legal corpus.



A detailed observation of the elements found in these networks can also help us identify certain topics related to the node (the main search word), which could be explored further by extending the network to a lower collocational level through the selection of any of the collocates in the set displayed above.

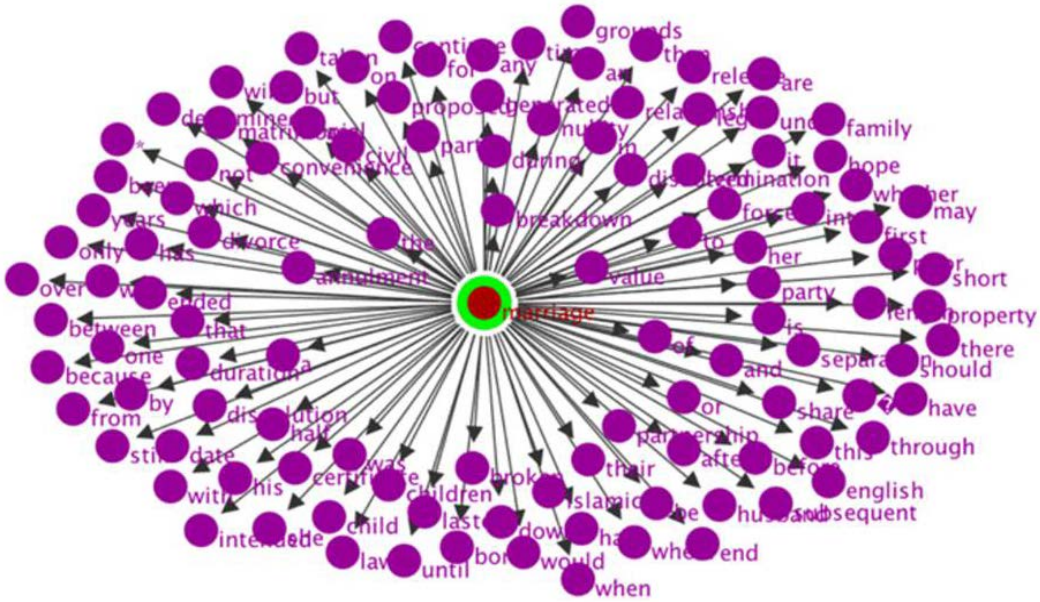
Government is one of them. If the collocates associated with it are examined on a second collocational level, still within the general English corpus *LACELL*, we observe a portrayal of this institution as reflected on texts coming from various sources such as the press, books, brochures, advertisements, written correspondence, etc. One of the possibilities of analysis could be grouping the network constituents according to semantic categories, as they refer to the different functions, organisation and features of this ruling body. The words *local, central, departments, federal* or *regional* belong in this area. On the other hand, the term *government* is associated with the ideology of the parties exercising that function, the collocates *tory, conservative* or *labour* are indicative of this fact. Another group of collocates which also contribute to the linguistic characterisation of this institution are those which refer to the power it exerts. Words like *reform, control* or *power* fall within this category. In addition, the word *government* occurs with words related to its *public service* role being envisaged as an institution which acts as *guidance* in public matters, takes care of *people's welfare* and is at their *service* (the words in italics are also collocates of *government)*. Likewise, its collocates display a general concern about economic issues since the words *funding, spending, expenditure* or *taxes* appear in its lexical network. The concordance lines below attest how all these examples can be analysed and interpreted in context. *KWIC* (key words in context), a software utility included in *Lancsbox,* offers this possibility:

(…) he said that the GOVERNMENT would REFORM taxation (…)
(…) agreement must be struck between the CENTRAL and LOCAL GOVERNMENTS both on the central-bank system (…)
(…) The problem for this TORY GOVERNMENT is that their ideology is (…)
(…) interpreting the true spirit of GOVERNMENT GUIDANCE in plan making, (…) is but one consideration.
(…) seeing these as the main GOVERNMENT contributions to WELFARE, or the general good (…)
Rifkin told the Commons health SPENDING meant a third of the annual budget (…)

Concerning the legal context, the collocate network of *party* clearly reveals the legal sense of the term in the field, as expected. Other legal terms such as *proceedings, litigation, convention, liability* or *appeal* collocate with it as well as other words which, although not being used exclusively in the legal area, are associated with its legal meaning, namely, *contracting, marriage, innocent, financial* or *witness.* These collocates provide plenty of data on the nature of some of the cases which were brought before British courts between 2008 and 2010.

One of the words that caught our attention amongst the constituents of this network was *marriage*. The fact that an issue such as marriage might be so relevant as to rank in 30th position within the collocate inventory of *party* was interesting enough to delve into its lexical deployment in the legal corpus.

Figure 7: 2nd level specialised network of *marriage*.



The list of constituents of the lexical network of *marriage* is noticeably long, as shown by Figure 7, being also connected to a large number of collocates of the first level network node, *party*. According to their meaning, the most relevant lexical collocates of *marriage* point at two major elements of this relationship as reflected on the texts in the corpus. On the one hand, its legal character, on the other hand, the economic terms which the legal concept *marriage* revolves around. Amongst the former group we find *annulment, divorce, separation, civil,* or *nullity*. The latter category comprises words like *contract, value, banking, property, valuation* or *acquire*.

Within the group of collocates of the term *marriage,* the words *convenience* and *genuineness* caught our attention. According to the Immigration Act 1999 (sections 24 and 24 A), amended in this respect by the Immigration Act 2014 (section 55), a marriage of convenience is defined as a civil relationship where

> "one or both of the parties is not a British citizen […] there is no genuine relationship between the parties; either or both of the parties enter into the marriage […] for the purpose of circumventing immigration controls […]"

But how do these different aspects reflect on those judicial decisions where the collocate pattern *marriage of convenience* is employed? Firstly, we find several collocates which refer to the definition of the term itself as found in the law, namely, *sham, bogus, circumventing* or *genuine.* If we analyse the concordances of the collocate pattern *sham marriage* (which the law identifies with *marriage of convenience),* in an appeal to the Supreme Court by the Secretary of State for the Home Department of the UK, we find that

> "persons seeking leave to enter or remain in this country may marry here, not for the reasons which ordinarily and legitimately lead people to marry, but in order to strengthen their claims for leave to enter or remain. Such marriages have been variously described as 'bogus' or 'sham' and as 'marriages of convenience'."

The texts in the legal corpus also gathered sociological information in relation to the topic that may have remained unnoticed on a superficial analysis of a smaller text sample, unless we went deeper into the interconnections amongst the constituents of lexical networks at different levels. Words such as *prevalence, incidence, recurrence* or *usual* can be found amongst the collocates of the term *convenience,* which may lead us to explore the issue further by reading the concordances associated to these terms and exploring other references (newspapers, legal texts, journal articles) to support our findings in this respect.

Lastly, the second level collocate network of *convenience* also contains words and terms which point at the legal reaction to this phenomenon on the part of the legislative or executive bodies. As proved by data, marriages of convenience appear to be a significant judicial problem in the UK and words such as *prevent, supress, measures, fighting, battle* or *policing* may also be pointing at that fact. Let us observe in greater detail what the texts have to say about this issue:

> (…) it operates to PREVENT MARRIAGES of CONVENIENCE (…)

(…) section makes no reference to MARRIAGES of CONVENIENCE or SHAM MARRIAGES (…)
(…) MEASURES to be adopted on the COMBATING of MARRIAGES of CONVENIENCE (…)

In response to research question 3 on the usefulness of collocational patters in the study of legal text, this analysis has attempted to illustrate the multiple possibilities that the exploration of collocational networks offers to the researcher interested not only in the linguistic dimension of these texts but also in their legal or sociological one. The fact that these networks can be obtained easily by simply uploading a corpus using automatic processing tools like *Lancsbox,* simplifies the process enormously, since obtaining them semi-automatically requires lots of effort and time prior to the actual analysis of their content.

6. Conclusion

The present research has been conceived as an introduction into the design and compilation of legal corpora and their processing using automatic corpus analysis tools. Such introduction has been carried out through the description and processing of two corpora, a general one of 13.7 million words, *LACELL* – used as reference whenever a general English corpus was required for comparison – and *BLaRC,* a legal one of 8.5 million words, made up entirely of judicial decisions.

Concerning the first research question posed in the introduction, an effort has been made to highlight the relevance of sampling criteria in corpus compilation, focusing, on the one hand, on the communicative relevance of the texts in the corpus and on the other hand, on the structure of the corpus itself.

Firstly, law reports have been presented as a fundamental legal genre all legal practitioners must know and cite, hence their importance within this ESP variety. Secondly, as regards the structure of the corpus, such a controversial issue as establishing the ideal word target has been tackled, concluding that, after calculating the type/term ratio in our legal corpus, a 2.5 to 3 million word target could suffice to study its lexicon, since the proportion of terms per word type dropped drastically at that point. The general structure of our legal corpus has also been presented in section 2.3., where a proportion in the word targets for each corpus category and subcategory was kept according to the number of texts available for each of them.

The second research question in the introduction enquired about the usefulness of Automatic Term Recognition (ATR) methods in the analysis of legal text. As shown in section 3, ATR methods can be of great help to the researcher when handling large amounts of data which could not be processed otherwise. Terms encapsulate specialised meaning, however, not all automatic term recognition methods are equally efficient in legal term identification. One of the reasons that could account for this phenomenon is the close relationship between legal terms and everyday vocabulary, where

DOI: 10.14762/jll.2017.018

large percentages of the former can be found. This is why different ATR methods were tested in order to select the most efficient ones in the legal field. The result of the assessment of five different ATR methods has been presented in section 3. After the validation process, it was found that Patrick Drouin's *TermoStat* (2003) managed to identify correctly 73 % legal terms in the *BLaRC*, ranking first in legal term mining. *TermoStat* is therefore recommended as the best method to extract legal terminology, which often poses difficulties in the accomplishment of this automatic task, as already stated.

Finally, the third research question posed in the introduction has been answered in sections 4 and 5, where one of the latest trends in Corpus Linguistics has been presented, that is, the use of software tools for the examination of collocate networks. A case study has been carried out in section 5 using one of these tools: *Lancsbox* (Brezina, McEnery & Wattam, 2015). One of the advantages of exploring the collocate patterns in a corpus is that they are capable of bringing to the foreground relevant aspects of its content and form that may otherwise remain unnoticed. Thanks to *Lancsbox* the task of producing collocate networks can be accomplished on the fly, allowing for the deployment not only of a word's collocate network but also of the networks associated with its collocates and the collocates of those collocates up to a seventh hierarchical level. The possibilities of enlarging the context of usage of a given word and analysing it through such connections are manifold.

To conclude, section 5 has demonstrated how the meaning of the sub-technical term *party* radically changes from one context to the other and how those meanings are organised in a hierarchical way in both contexts. Such change has been observed through the analysis of the constituents of the collocate networks extracted from both corpora, which have shown how the prevailing sense of the term *party* in the general corpus was that of "political group/association", followed by "celebration", whereas it meant "person/persons taking part in a legal proceeding" in the legal corpus, as was expected. Moreover, the collocate networks were explored in greater detail revealing interesting data such as the incidence of a topic like *marriage* in a corpus of judicial decisions, which, in principle, might not appear to be so relevant for a text collection comprising decisions from the criminal and civil fields. In fact, this analysis has gone beyond the merely linguistic level entering the sociological/legal dimension and allowing for a deeper understanding of such phenomena. In its creators' own words:

> "collocation networks as an analytical tool have a large potential in a number of areas of linguistic and social research such as discourse studies, psycholinguistics, historical linguistics, second language acquisition, semantics and pragmatics, lexicogrammar, and lexicology" (Brezina, McEnery & Wattam, 2015: 165).

Nevertheless, further research still remains to be carried out, particularly in the legal field, to test and exploit the potential of collocate networks, which this research has intended to suggest.

References

Alcaraz Varó, Enrique (1994). *El inglés jurídico: textos y documentos*. Madrid: Ariel Derecho.

Anthony, Laurence (2014). *AntConc* (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from www.laurenceanthony.net.

Baker, Paul (2005). *Public Discourses of Gay Men*. London: Routledge.

Baker, Paul (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, 21 (2), 139–164. DOI: 10.1075/ijcl.21.2.01bak.

Bhatia, Vijay (1993). *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Biber, Douglas (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8 (4), 243–57. DOI: 10.1007/978-0-585-35958-8_20.

Biber, Douglas, Conrad, Susan, & Reppen, Randy (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biel, Łucja & Engberg, Jan (2013). Research models and methods in legal translation. *Linguistica Antverpiensia*, 12, 1–11. Available at lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/316/225.

Borja Albí, Anabel (2000). *El texto jurídico en inglés y su traducción*. Barcelona: Ariel.

Breeze, Ruth (2015). Teaching the vocabulary of legal documents: a corpus-driven approach. *ESP Today*, 3 (1), 44–63. Available at www.esptodayjournal.org/esp_today_back_issues_vol4.html.

Brezina, Vaclav, McEnery, Tony & Wattam, Stephen (2015). A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20 (2), 139–173. DOI: 10.1075/ijcl.20.2.01bre.

British National Corpus (2007). BNC XML Edition version 3, distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at www.natcorp.ox.ac.uk.

Cabré Castellví, María Teresa, Estopà Bagot, Rosa & Vivaldi Palatresi, Jordi (2001). 'Automatic term detection: a review of current systems', in Bourigault, Jacquemin & L'Homme (Eds.), *Recent Advances in Computational Terminology* (53–87). Amsterdam: John Benjamins. DOI: 10.1075/nlp.2.04cab.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Boston: The Massachusetts Institute of Technology (MIT).

Chung, Teresa (2003). A corpus comparison approach for terminology extraction. *Terminology*, 9 (2): 221–246. DOI: 10.1075/term.9.2.05chu.

Church, Kenneth Ward & Hanks, Patrick (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16 (1), 22–29. Available at dl.acm.org/citation.cfm?id=89095.

Corpas Pastor, Gloria & Seghiri Dominguez, Míriam (2010). *El concepto de representatividad en lingüística de corpus: aproximaciones teóricas y consecuencias para la traducción*. Málaga: Servicio de Publicaciones de la Universidad de Málaga.

Cruse, David Alan (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Danet, Brenda (1980). Language in the Legal Process. *Law and Society Review*, *14* (3), 445–564. DOI: 10.2307/3053192.

Drouin, Patrick (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9 (1): 99–117. DOI: 10.1075/term.9.1.06dro.

Dudley-Evans, Tony & St John, Maggie Jo (1998). *Developments in English for Specific Purposes*. Cambridge: Cambridge University Press.

Dunning, Ted (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19 (1), 61–74. Available at dl.acm.org/citation.cfm?id=972454.

Firth, John Rupert (1957) *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Flowerdew, Lynne (2004). The argument for using English specialised corpora to understand academic and professional language. In Connor & Upton (Eds.), *Discourse In The Professions: Perspectives From Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/scl.16.02flo.

Flowerdew, Lynne (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics* 14 (3), 393–417. DOI: 10.1075/ijcl.14.3.05flo.

Geary, Adam & Morrison, Wayne (2012). *Common Law Reasoning and Institutions*. London: University of London.

Goźdź-Roszkowski, Stanisław & Pontrandolfo, Gianluca (2014). Legal phraseology today: corpus-based applications across legal languages and genre. *Fachsprache: International Journal of Specialized Communication*, 3–4, 130–138.

Gries, Stefan Thomas (2013). 50-something years of work on collocations: What is or should be next. *International Journal of Corpus Linguistics*, 18 (1), 137–166. DOI: 10.1075/ijcl.18.1.09gri.

Gries, Stefan Thomas & Wulff, Stephanie (Eds.) (2010). *Corpus-linguistics applications. Current studies, new directions*. Amsterdam/New York: Rodopi.

Heaps, Harold Stanley (1978). *Information Retrieval: Computational and Theoretical Aspects*. New York: Academic Press.

Kennedy, Graeme (1998). *An introduction to corpus linguistics*. New York: Longman.

Kilgarriff, Adam, Baisa, Vít, Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychlý, Pavel & Suchomel, Vít (2014). The Sketch Engine: Ten Years On. *Lexicography*, 1, 7–36. DOI: 10.1007/s40607-014-0009-9.

Kit, Chunyu & Liu, Xiaoyue (2008). Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14 (2), 204–229. DOI: 10.1075/term.14.2.05kit.

Lemay, Chantal, L'Homme, Marie-Claude & Drouin, Patrick (2005). Two Methods for Extracting 'Specific' Single-word Terms from Specialised Corpora: Experimentation and Evaluation. *International Journal of Corpus Linguistics*, 10 (2), 227–255. DOI: 10.1075/ijcl.10.2.05lem.

Maley, Yon (1994). The Language of the Law. In J. Gibbons (Ed.), *Language and the Law*. London: Longman.

Marín, María José (2014). Evaluation of five single-word term recognition methods on a legal corpus. *Corpora*, 9 (1), 83–107. DOI: 10.3366/cor.2014.0052.

Marín, María José (2015). Measuring precision in legal term mining: a corpus-based validation of single and multi-word term recognition methods. *ESP World*, 46, 1–23. Available at www.esp-world.info/Articles_46/MARIN_MEASURING%20PRECISION%20IN%20LTM-AN.pdf.

Marín, María José (2016). Measuring the degree of specialisation of sub-technical legal terms through corpus comparison: a domain-independent method. *Terminology*, 22 (1), 80–102. DOI: 10.1075/term.22.1.04mar.

Marín, María José & Rea Rizzo, Camino (2012). Structure and design of the BLRC: a legal corpus of judicial decisions from the UK. *Journal of English Studies*, 10, 131–145. DOI: 10.18172/jes.184.

Maynard, Diana & Ananiadou, Sophia (2000). TRUCKS: A model for automatic multi-word term recognition. *Journal of Natural Language Processing* 8 (1), 101–125. DOI: 10.5715/jnlp.8.101.

McEnery, Tony (2006*). Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. Abington, UK: Routledge.

McEnery, Tony & Wilson, Andrew (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, Tony, Xiao, Richard & Tono, Yukio (2006). *Corpus-based language studies: an advanced resource book*. Routledge Applied Linguistics: New York.

Mellinkoff, David (1963). *The Language of the Law*. Boston: Little, Brown & Co.

Nesi, Hillary & Gardner, Sheena (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.

Orts Llopis, María Ángeles (2006*). Aproximación al discurso jurídico en inglés: las pólizas de seguro marítimo de Lloyd's*. Madrid: Edisofer.

Orts Llopis, María Ángeles (2009). Legal genres in English and Spanish: some attempts of analysis. *Iberica*, 18, 109–130. Available at www.aelfe.org/documents/07_18_Orts.pdf.

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

Pazienza, Maria Teresa, Pennacchiotti, Marco & Zanzotto, Fabio Massimo (2005). Terminology extraction: An Analysis of Linguistic and Statistical Approaches. *Studies in Fuzziness and Soft Computing*, 185, 225–279. DOI: 10.1007/3-540-32394-5_20.

Pearson, Jennifer (1998). *Terms in Context*. Amsterdam: John Benjamins.

Sánchez Aquilino & Cantos Gómez, Pascual (1997). Predictability of Word Forms (Types), and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus. *International Journal of Corpus Linguistics*, 2 (2), 251–272. DOI: 10.1075/ijcl.2.2.06san.

Scott, Mike (2008). *WordSmith* Tools version 5. Liverpool: Lexical Analysis Software. Available at www.lexically.net/wordsmith.

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

Sinclair, John (2005). Corpus and Text: Basic Principles. In Wynne 2005 (see below). Available at ota.ox.ac.uk/documents/creating/dlc/chapter1.htm.

Sinclair, Stéfan, Rockwell, Geoffrey & the Voyant Tools team (2012). Voyant Tools [Computer software]. Retrieved from www.voyant-tools.org.

Sparck Jones, Karen (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 11–21. DOI: 10.1108/eb026526.

Sternfeld, Joshua (2012). Pedagogical Principles of Digital historiography. In Hirsch (Ed.), *Digital Humanities Pedagogy*. London: Open Book Publishers. Available at books.openedition.org/obp/1645.

Stubbs, Michael (2001). *Words and Phrases*. London: Blackwell.

Tiersma, Peter (1999). *Legal Language*. Chicago: The University of Chicago Press.

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

Vivaldi, Jorge, Cabrera-Diego, Luis Adrián, Sierra, Gerardo & Pozzi, María (2012). Using Wikipedia to Validate the Terminology Found in a Corpus of Basic Textbooks. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12)*. Istanbul, Turkey. Retrieved from www.lrec-conf.org/proceedings/lrec2012/index.html.

Widdowson, Henry (2000). The limitations of linguistics applied. *Applied Linguistics*, 21 (1), 3–25. DOI: 10.1093/applin/21.1.3.

Williams, Geoffrey (1998). Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics*, 3(1), 151–171. DOI: 10.1075/ijcl.3.1.07wil.

Williams, Geoffrey (2001). Mediating between lexis and texts: collocational networks in specialised corpora. *ASp, la revue du GERAS*, 31, 63–76. Available at asp.revues.org/1782.

Wynne, Michael (Ed.) (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow books. Retrieved from ota.ox.ac.uk/documents/creating/dlc.