HOW RELEVANT ARE LATIN WORDFORMS AND CLUSTERS IN LEGAL ENGLISH? A CORPUS-BASED STUDY ON THE REPRESENTATIVENESS AND SPECIFICITY OF SUCH ELEMENTS IN UKSCC: AN *AD HOC* LEGAL CORPUS

Mª José Marín Pérez
Camino Rea Rizzo
Universidad de Murcia

Abstract

The use of Latin wordforms and phrases in legal English is one of the features that contribute to its high degree of complexity. Most text and academic books on this ESP variety describe their use from a prescriptive perspective often based on a limited number of linguistic samples, however, to the best of our knowledge, no corpus-based studies on Latin words and phrases have been carried out to the date. Therefore, this article approaches the issue employing a corpus-based methodology aiming to establish the degree of representativeness of these elements as well as to classify them into different levels of specialization. Likewise, it also explores the different results obtained before and after discriminating single-word Latin units

Resumen

El uso de palabras y clusters latinos en el inglés jurídico es una de las características que contribuyen a su alto nivel de complejidad. La mayoría de los libros de texto y académicos sobre esta variedad del inglés describen su uso desde un enfoque prescriptivo basado en un número reducido de textos, sin embargo, no se han realizado estudios descriptivos sobre este tema hasta la fecha que se basen corpus de mayor tamaño. Por este motivo, este artículo estudia el tema desde la perspectiva de la lingüística del corpus para establecer el grado de representatividad de las palabras latinas además de clasificarlas según su nivel de especialización. Asimismo, también explora los diferentes resultados obtenidos antes y después de discriminar las unidades de un solo elemento de las de dos o tres. Para llevar a cabo tales tareas, se

Submission: 05/01/2012- Acceptance: 13/03/2012

from two/three-word clusters. In order to complete such tasks, an *ad hoc* 2.5 million-word legal corpus, UKSCC, has been compiled according to corpus standards and employed as the source to obtain the data for this study.

Keywords: Latin, legal English, corpus-based analysis, specialised corpora, representativeness, term identification.

ha compilado un corpus *ad hoc* de 2.5 millones de palabras con el fin de obtener los datos necesarios para este estudio. El corpus se ha denominado UKSCC.

Palabras clave: Latín, inglés jurídico, corpus especializados, representatividad, identificación de términos.

1. INTRODUCTION

Scholars agree to assert that legal English is a complex variety of the language for varied reasons (Mellinkoff 1963; Alcaraz 1994; Tiersma 1999; Borja 2000; Orts 2006). As Mellinkoff (24) claims: "The language of the law has a strong tendency to be: wordy; unclear; pompous [and] dull." Moreover, Maley (1994:11) believes that legal language has never been "in tune with common usage. It has always been considered a language apart."

The use of archaic language, redundancy or long and confusing syntactic structures full of parentheses and enumerations, amongst other features, contributes to its obscurity. This has resulted into citizens' initiatives such as the *Plain English Movement* in the UK, which fights "against the use of jargon and gobbledygook in public information from both private and public service organisations," as stated on their website.[1]

One of the characteristics that adds to the high degree of complexity of legal English is the use of Latin words and phrases, as highlighted by scholars (Mellinkoff 1963; Alcaraz 1994; Borja 2000; Orts 2006), which dates back to the 11th and 12th centuries when the Normans would draft legal documents in Latin, the language of science and culture.

However, the behaviour of such lexical elements in legal texts has always been approached from a prescriptive perspective after the specialists' intuition or through

[1] http://www.plainenglish.co.uk

the observation of a limited number of linguistic samples. To the best of our knowledge, no descriptive studies have been carried out employing a corpus-based methodology. Hence, this paper aims at answering some questions related to the representativeness and level of specialisation of Latin wordforms and clusters in law reports from a descriptive perspective, that of corpus linguistics. For this purpose, an *ad hoc* 2.5 million-word legal corpus of judicial decisions from the UK Supreme Court has been created and analysed.

In section 2, a justification and description of UKSCC, the United Kingdom Supreme Court Corpus, is presented followed by section 3, where the description of the methodology employed to analyse the data obtained and the results of such analysis are offered. To finish, the major conclusions drawn from this study and some further research questions are shown in section 4.

2. UKSCC: PILOT CORPUS DESCRIPTION AND JUSTIFICATION

UKSCC is a pilot legal corpus which has been compiled according to corpus linguistics standards as stated in Sánchez et al. (1995) and Wynne (2005) for general corpora and its adaptation to specific corpora (Pearson 1998; Rea 2010). It is a 2.5 million-word specialised corpus integrated into a larger one: BLaRC (The British Law Report Corpus), still in its compilation phase.

The reasons to single out this legal genre to study the linguistic behaviour of its lexicon are multifarious. To begin with, the UK belongs to the realm of common law, as opposed to civil or continental law, which is the judicial system working in most Western European countries. In purely common law systems, the acts passed at their parliaments have gained greater importance being most often cited in case decisions. However, case law stands at the very basis of common law systems which rely on the principle of binding precedent to work, that is to say, a case judged at a higher court must be cited and applied whenever it is similar to the one being heard in its essence (the *ratio dicendi*). Another fact that makes law reports an outstanding genre in common law legal systems is that they not only cover all the branches of law, but might also present full embedded sections of other public and private law genres displaying therefore great lexical richness and variety. Following Sinclair "the contents of the corpus should be selected […] according to their communicative function in the community in which they arise" (Sinclair 2005:5). Consequently, such texts as these have been selected to form the corpus due to the fundamental role they play in common law legal systems.

Regarding the main objective of compiling both BLaRC and UKSCC, they are aimed at providing a useful and reliable source of specific vocabulary to elaborate didactic materials owing to the scarce amount of existing legal corpora[2] and the methodological void derived from it. As McEnery and Wilson (1996:121) affirm, "such corpora can be used to provide many kinds of domain-specific material for language learning, including quantitative accounts of vocabulary and usage which address the specific needs of students in a particular domain more directly than those taken from more general language corpora." Furthermore, the use of corpus-based approaches in lexical studies allows us to "document such patters [of use], providing the information needed for more informative and helpful language instruction and materials development" (Biber et al. 1998:53).

As far as UKSCC (the pilot corpus) is concerned, it is a synchronic, monolingual and specialised collection of 193 judicial decisions from the UK Supreme Court and the House of Lords[3] issued between 2008 and 2010. The Supreme Court was selected as a text source for the pilot corpus due to its relevance within the British judicial system (all the decisions made at the Supreme Court set precedent and are cited whenever applicable), and the wide lexical variety of the texts coming from it. It is at the top of the UK judicial pyramid and deals with cases belonging to all branches of law therefore producing certainly rich and varied texts as far as the lexicon is concerned, which is the linguistic focus of this study.

The texts included in UKSCC are full authentic transcriptions of judgments as produced by the courts' official shorthand writers. In order to reach the Supreme Court, a case requires having obtained leave of appeal on several occasions, which enables it to follow a complex route to get to such high level in the institutional pyramid. This long path implies greater argumentation every time the case is heard at a different court, i.e. from a crown court to the High Court of Justice –criminal division– and from there to the Supreme Court, making the texts more and more complex and lexically richer as they go up in the judicial hierarchy.

---

[2] See Marín and Rea (2011) for a review on legal corpora.

[3] The *Constitutional Reform Act*, 2005, created the Supreme Court which started to work as the court of last resort of the UK in October 2009. Until then, it had been the so-called "Law Lords" of the House of Lords who carried out that function. This is the reason why the texts selected from 2008 to 2010 come from both sources.

3. METHODOLOGY, RESULTS AND ANALYSIS

3.1. FIRST RESULTS: SINGLE WORD UNITS

The first step taken towards the study of the presence of Latin words in legal English was to manually retrieve Latin words and phrases from the frequency list of 27,059 types from UKSCC, considering the term *Latin words* as those words which have been borrowed directly from Latin into English without changing or adapting to the phonetics of English regardless of them being lexicalised or not. The list of true terms was obtained from the literature consulted, both text and academic books,[4] which provides a wide inventory of these vocabulary items in legal English. It was used as the gold standard acting as reference for the manual extraction of Latin single and multi-word terms from UKSCC.

Table 1 illustrates the frequency of occurrence of Latin types, their text range or distribution (the number of corpus texts they appear in) and their keyness scores. A word is considered key "if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger word-lists" (Scott 2008:184). A bigger corpus is needed to calculate a word's keyness in a specific corpus. With this purpose, UKSCC has been compared with LACELL, a 21 million-word general English corpus compiled by the LACELL research team at the University of Murcia. It is a balanced synchronic corpus of general English including both written texts from diverse sources such as newspapers, books (academic, fiction, etc.), magazines, brochures, letters and so forth, and also oral language samples from conversation at different social levels and registers, debates and group discussions, TV and radio recordings, phone conversations, everyday life situations, classroom talk, etc. Its geographical scope ranges from USA, to Canada, UK and Ireland.

TYPES	FREQUENCY	TEXT RANGE	KEYNESS
In	66560	193	4373.02
V	7214	193	22486.09
Re	757	131	506.08
Non	712	135	84.12
Per	639	131	-284.48
Ex	550	95	681.11
De	510	67	13.71

[4] See Mellinkoff (1963), Alcaraz (1994), Borja (2000) and Orts (2006), for academic references on Latin vocabulary in legal English and Fernández (1994), Rice (2007), Krois-Linder and Firth (2008), Frost (2009), Callanan (2010) and Orts (2010) for textbook references.

Sub	314	66	208.08
Post	266	73	-37.17
Facto	230	22	734.78
Inter	170	93	143.18
Anti	155	33	-11.50
Interim	147	32	172.03
Turpi	134	3	588.69
Causa	129	3	497.18
Alia	124	77	422.65
Parte	115	26	400.30
Facie	114	50	378.96
Ante	112	5	230.19
Prima	109	49	286.53
Audit	108	6	71.35
Vires	98	21	375.69
Alter	91	44	25.09
Et	83	35	1.48
Media	78	17	-154.30
Memorandum	66	21	67.86
Affidavit	65	17	176.22
Dicta	57	35	227.29
Forum	56	21	-0.01
Pro	56	18	-14.73
Plus	54	28	-207.15
Ad	52	21	-4.26
Jure	52	5	205.863
Quantum	52	18	83.96
Onus	50	16	101.39
Sic	44	23	1.31
Consensus	42	11	-0.37
Doli	42	1	168.71
Pari	41	8	145.63
Se	39	23	-0.40
Via	39	21	-54.65
Incapax	36	1	158.15
Nil	36	13	-9.41
Nos	36	21	42.55
Obiter	36	23	158.15
Rea	35	12	55.64
Strata	35	2	64.97
Par	33	12	-0.70
Proviso	33	20	59.01
Fortiori	32	28	116.40
Medium	32	22	-81.26
Mens	32	11	64.27
Ab	30	17	23.94
Rata	30	5	61.89
Seq	29	17	108.19

Passu	28	4	104.00
Meruit	27	5	118.61
Actus	26	7	100.28
Est	25	9	0.03
Mutandis	25	19	91.46
Veto	25	3	0.00
Dictum	24	19	55.36
Ibid	24	15	-2.80
Mutatis	24	18	87.30
Vis	24	8	2.78
Ratio	23	16	-30.07
Annum	21	12	0.01
Ipso	18	11	51.08
Versa	17	13	-0.09
Bona	15	9	8.24
Conveniens	15	3	58.65
Nexus	15	6	19.22
Quasi	15	10	2.35
Qui	15	4	16.68
Camera	14	5	-95.00
Fide	14	8	7.59
Coelum	12	1	52.71
Intra	12	7	3.90
Corpus	11	6	0.94
Magna	11	1	11.12
Minimis	11	6	37.63
Quod	11	4	34.48
Absentia	10	2	26.01
Habeas	10	5	20.55
Litem	10	3	43.93
Reus	10	5	43.93

Table 1. *Latin types from the UKSCC frequency list*. (Due to the extension of the whole table, those items showing <10 occurrences have been left out).

UKSCC contains 187 Latin types in total (including those occurring <10 times) which represent 0.68% of the full corpus type list. By examining the first ten, it appears that, similarly to general English, most of them are function words: eight prepositions like *in, versus (v), per, ex, de, sub, post* and *inter,* one noun, *re,* and the negative adverb *non* (these morphological categories correspond to Latin, not to other possible uses in English). The case of *in* stands out, as it is, by far, one of the most frequently used words not only in Latin but also in general English corpora such as

the British National Corpus[5] (it is in the sixth frequency rank position) being a preposition in both languages. *In* appears in legal Latin phrases like *in camera* or *in personam,* as well as in an infinite number of English ones. Hence, its context must necessarily be consulted (by using a concordancer) in order to assign it an accurate frequency value for its Latin use. This is why *in* has been eliminated from the comparison since it could lead to misleading results if the frequency data were analysed without looking at the context. Furthermore, having manually checked the first 5,000 out of 66,560 concordances for *in,* no Latin use has been recorded in UKSCC.

The second of these items, *v* (the initial standing for the preposition *versus*), deserves special attention too as it appears in all 193 texts on 7214 occasions. This is directly related to the format of the texts themselves. All Supreme Court judgments start with an introduction to the case where several elements are always included so that, when lawyers and judges consult them, they can easily identify the source, date, names of the parties (it is here where the preposition versus appears −*Boss Holdings Limited v Grosvenor West End Properties and others*−) or issue of the case in question.

Except for *versus* and *in,* the rest of wordforms in the "top ten" list occur considerably less often. Actually, they are between the 400th and 1800th positions in the UKSCC frequency rank although well distributed throughout the corpus. On the other hand, 65 types (around 35.15% of the list) occur between 1 and 3 times in 1 or 2 texts at most leaving a relatively reduced amount of items ranging from 4 to around 700 occurrences, which represent 0.44% of UKSCC, that is, 120. On average, these 120 types appear 124.76 times, whereas UKSCC mean raw frequency is 193.88 after eliminating < 4 occurring items. Of these 120 wordforms, only 17 fall within the first most frequent 2000 words of UKSCC. This fact might be an indicator of their technicality although more parameters must be taken into account to make this claim.

For the sake of comparison, a list of 35 nouns referring to criminal offences[6] in the UK was employed showing that the values for those items within the same frequency range is 90.20, 34 points lower than Latin types. Therefore, as far as frequency is concerned, Latin wordforms are slightly below the average although

[5] BNC is one of the most widely used general English corpora. It has 100 million words being formed by written and spoken samples of British English from varied sources.

[6] The Supreme Court deals with all types of matters, not only criminal, so this list does not cover all kinds of issues that might be subject to be heard at this court. It is just a sample list used to establish whether the values obtained might be high or low in comparison with other relevant lexical items of legal English.

they seem to be used more often than other characteristic legal words such as crime nouns.

The notion of text coverage also appears as a very relevant parameter when attempting to establish the level of representativeness of a group of wordforms. It refers to the percentage of running words in a text covered by a given list of wordforms or families, that is, how helpful a set of words might be for a reader to understand a group of texts. Paul Nation and his associates have extensively worked on this area (Nation and Waring 1997; Nation 2001) and established that the first 2,000 word families of English present in the *General Service List* by West (1953) together with the 570 families from the *Academic Word List* by Coxhead (2000) cover 85-90% of the words in a text. Then, 5% would be left for technical words in the academic field (Coxhead and Nation 2001). As far as UKSCC is concerned, both the BNC and AWL lists of the most frequent 2,570 word families cover 90.21% of the total running words (tokens) in the corpus, while the remaining 9.79% are not found on those lists and are thus potential candidates to be technical terms.

In order to calculate text coverage, the frequency values of all the items in the list were added up and then divided by the number of tokens in UKSCC. The 187 Latin types of the pilot corpus cover 0.0059% of the tokens in it whereas crime nouns cover only 0.00095%, almost six times less than the former. Moreover, frequency level is higher for Latin types, 945.73 on average and 89.98 if we ignore the first two items *in* and *versus,* while it is 70.97 for the latter, that is to say, Latin types not only display six times as much text coverage as crime nouns but also they appear more frequently.

As regards their text distribution, Latin wordforms occur in 23.20 texts on average (again, the first two items of the list have been ignored for this calculation). Only 19 of them are above the mean value while 32 occur in less than 10 texts. However, if compared with crimes nouns, Latin types are much better distributed throughout the corpus as the latter are present in only 9.86 texts on average. Thus, could it be stated that Latin terms are well distributed within UKSCC? It could, especially when compared with the mean text distribution value of the 6674 keywords in UKSCC: 32.31, although the fact that 67 out of 187 stand below that figure might also be an indicator of their highly specialised character.

As far as keyness is concerned, when put in contrast with UKSCC keyness value (both data were calculated using the *Keywords* tool in Scott's (2008) software *Wordsmith 5),* Latin types display a reasonably high one. Whereas the pilot corpus keyness average is 116.08, Latin terms show 404.44, that is, they are four times as relevant as the keywords found in the study corpus, a datum that may point at their high level of representativeness, much higher than all the keywords in the pilot corpus. Nevertheless, if we leave the first two items out of this count, the mean drops

sharply to 94.32, 20 points below the corpus average and slightly lower than crime nouns (97.07). There is a group of wordforms (22) which fall within relatively high keyness values ranging from 118 to 734. Hence, they could be said to be highly representative of the genre. As for their morphological categories, 25% of them are function words: *inter, sub, ante, re, ex*, the rest are content words: nouns like *obiter, vires,* or *causa;* adjectives: *prima, turpi*; adverbs: *interim;* and one verb: *meruit*. On the other hand, 18 types present negative keyness values due to the fact that they are unusually infrequent in the study corpus as opposed to the general one.

A comparison with the legal section of BNC[7] of 2.2 million words has also been carried out so as to confirm the presence of Latin words in other legal corpora. Due to the size difference between both corpora, the counts for the most frequent 20 Latin types in UKSCC and the legal section of BNC were normed to a basis per 1,000 corpus tokens (Biber et al. 1998: 263-4), so the results show how many times a given type appears in each corpus per 1,000 words. The mean value for the 30 most frequent Latin types found in both UKSCC and BNC is 0.17 and 0.249 respectively. Half of the types display practically identical values which proves that the presence of these items in both corpora is quite similar, enabling us to draw certain generalizations about the role played by Latin words not only in UKSCC but also in other similar legal corpora. As a matter of fact, the Latin types in UKSCC and BNC present higher figures than the normalised average frequency value for the whole of UKSCC within the same frequency range (50-7200), namely, 0.142.

If contrasted with the same value for crime nouns, it appears that they occur 0.036 times every 1,000 tokens in the pilot corpus and 0.04 in the legal section of BNC, again six time less than Latin types. The elimination of the preposition *versus* from the list, due to its high value derived from the format of the texts which forces its appearance in all of them, does not alter the comparison much as the numbers are still much higher for Latin terms: 0.08 and 0.10 in the pilot corpus and BNC respectively.

Summing up, taking into consideration all the information obtained in this first approach to the presence of Latin wordforms in UKSCC, it can be stated that:

1- The mean frequency of occurrence for those items ranging between 4 and 700 (0.44% of the items on UKSCC frequency list), is 124.76, that is, about 70 points below the average for those within the same frequency range in UKSCC. However, if weighed against crime nouns, the latter display lower values (90.2) which may point at the greater relevance of Latin types within the corpus.

[7] These concordances are freely accessible at: http://www.lextutor.ca/concordancers/concord_e.html

This fact is confirmed by comparison with the legal section of BNC where the figures are similar for Latin wordforms, 0.1 occurrences every 1,000 words, against crime nouns: 0.04.

2- Latin types provide 0.0059% text coverage whereas the figure for crime nouns is far below: 0.00095%.

3- Latin wordforms are well distributed throughout UKSCC as they are just 9 points below the mean value of this parameter for the whole corpus appearing in 23.2 texts, while crime nouns show much lower figures as far as text range is concerned, only 9.36 on average.

4- Keyness, however, is similar for both types of wordforms. Actually, it is roughly higher for crime nouns: 97.07 against 94.3 for Latin types standing about 20 points below UKSCC average for this parameter, 116.08.

Nevertheless, these first results should be referred to with certain caution as it must not be forgotten that the software employed to obtain these lists, *Wordsmith 5*, does not distinguish the different senses of words and would thus compute, for instance, the sequence *re* as many times as it appears in the texts regardless of its use and meaning. As it happens with *in*, it will make no distinction between the prefix *re* as in *re-marry* (meaning *repetition of an action*) and the real Latin term RE as in *In RE Lo-atLine Electric Motors Ltd [1988] Ch 477,* a case citation formula where *re* means *case* or *issue.* This is why this list must be filtered so as to identify the really technical Latin terms of legal English present in the specialised corpus and the ones which have been lexicalised and belong to general English, or have acquired a specialised sense in a legal context.

For these reasons, it is necessary to study the context of occurrence of these types in order to decide whether a given token is being employed as a real Latin term. Besides, by observing their context and immediate collocates, Latin types can also be identified as a single wordforms or as part of a larger unit.

3.2. CLASSIFICATION OF LATIN TYPES ACCORDING TO THEIR LEVEL OF SPECIALIZATION

There exist different methods to analyse the information obtained from linguistic corpora. The literature on the subject shows how authors employ such methods to analyse and classify specific lexicon (Yang 1986; Farrell 1990; Coxhead 2000; Nation 2001; Rea 2008, amongst others) using diverse criteria to divide the

vocabulary of specialised texts into technical, sub-technical, academic or general. The use of stop lists to discriminate general from specific terms is present in most of them being West's *General Service List* (1953) one of the earliest methods in this respect which includes the most frequent 2,000 word families of English. Other more recent general vocabulary listings employed with this purpose are Averil Coxhead's *Academic Word List* (2000), or *The British National Corpus* lists (2007), amongst others.

The comparison between general and specific corpora is also a very useful method to characterise the lexical profile of ESP varieties due to the fact that it provides objective information on the frequency of occurrence of words in both corpora. The concepts of keyness and text range (or text distribution) are fundamental for the identification of technical vocabulary in a specialised corpus.

However, authors do not explain their methods in a practical way that can be actually applied to corpus analysis. According to Rea (2008:105), only Chung (2003) describes the method to employ accurately when trying to establish a cut-off point to discriminate terms (words with a specialised meaning) from non-terms (both general and also sub-technical wordforms which acquire a new meaning in a specialised context) after comparing a general and a specific corpus. Chung maintains that "a ratio of 1:50 provided the most effective cut-off point. To be classified as a technical term, a type had to occur at least 50 times more often in the technical text than in the comparison corpus, or only occur in the comparison corpus" (53).

Chung reaches this conclusion after validating her method by comparison with a qualitative one: the *rating scale approach*. She asks two experts to classify the vocabulary in a 5,500 word text from her anatomy corpus. The experts are trained in order to classify the words into four different categories depending on their level of specialization.

Meanwhile, she applies a quantitative procedure consisting in calculating the ratio of occurrence of the types in the anatomy text given to the experts. She normalises the frequencies of the text types in both her anatomy corpus and a general one and calculates the ratio by dividing the former by the latter. Then, basing her classification on these results and on the absolute frequency figures obtained, she also produces different groups and compares them to the ones by the specialists. The results of the comparison yield 86% coincidence on average, especially regarding highly specialised words and non-terms.

In the following subsection, these different methods will be applied with the purpose of obtaining a reliable classification of the Latin types identified.

3.2.1. GENERAL AND ACADEMIC VOCABULARY

To start with, UKSCC was processed with RANGE (Paul Nation's software) to obtain the list of the types present in the first 2000 of BNC. Then, the data were compared with the list of 187 Latin types using an excel spreadsheet in order to identify the ones falling within this inventory. As a matter of fact, due to the software not taking into account such phenomena as polysemy or homonymy, after studying their concordances in BNC,[8] words like *re, non,* ex, *alter, ad, se, nos, mens, est, camera,* and *ne,* were eliminated as they were not used as Latin words but as English ones. Hence, the list of Latin types used as general vocabulary are: *per, sub, post, pro, plus, nil, quid* whose frequency of occurrence is so low (just one occurrence each) that we can certainly affirm that, as a whole, the Latin words used in law reports do not belong to the domain of general vocabulary, as it may seem *a priori.*

Regarding academic vocabulary, it displays higher raw frequency values. Having processed UKSCC with RANGE using AWL as the only baseword list, it appears that only six Latin types: *plus, media, medium, via, ratio* (which occur 54, 78, 29, 39, and 23 times respectively) belong to the group of 570 families defined by Coxhead. They are no longer function words, as it happens with the general list, but rather content ones, although they are not numerous either.

All in all, if only 11 out of 187 Latin types (leaving *in* and *alter* aside for the reasons explained above) have been found in the general and academic lists of vocabulary –which represents 5.88% of the total–, it could be claimed that their level of specialisation is high, but, to what extent? To answer this question, the context of occurrence of these wordforms will be studied so as to differentiate those which are employed as single word terms from those which belong to two/three word clusters, then, both sorts of units will be studied and classified into levels of specialisation applying Chung's method complemented by other qualitative procedures.

3.2.2. TERMS AND NON-TERMS

To begin with, having manually obtained and read the concordances for the lists above, single word Latin terms were differentiated from multi-word units basing such discrimination on the statistical relevance/irrelevance of their collocates. In addition, a few of them whose characters coincided with the corresponding Latin word were

[8] BNC concordances checked at Mark Davies' website: http://corpus.byu.edu/bnc

excluded from the inventory as they are used as French words in the pilot corpus, namely, *qui*, *ne,* and *est* (*est* appears as a Latin word on just 5 occasions in one single text as part of a three word cluster: *cuius est solum,* in the rest of cases it is French). A frequency threshold was established so that those terms displaying ≤3 occurrences would be left out of the listings produced. As a result, 32 single word Latin units were extracted from the initial list, that is, 14.43% of 187 types. Let us then compare the information obtained previous to this filtering with the one illustrated in the table below.

Regarding their raw frequency in the pilot corpus, the results differ noticeably from the ones in section 4.1. Single word Latin terms display a mean value of 32.31, three times less than crime nouns and four time less than the unfiltered Latin types in table 1. Only two of them show > 100 frequency counts while 10 occur on less than 10 occasions. The remaining 20 show values between 13 and 66, much lower than the corpus average applying the ≤ 3 threshold cut: 193.88.

Text coverage is also affected by the filtering of the list as it changes from 0.0059% to 0.0041%. In this case, there is a slight drop not as sharp as the one experimented by frequency. Although the frequency scores have noticeably decreased, the number of running words in the corpus covered by these vocabulary items is still much higher than other specialised vocabulary like crime nouns.

The figures for text distribution or range sharply fall by 50% as single word Latin types appear in 12.81 texts on average as opposed to 23.2 in section 4.1. Fifteen of them (50%) only do so in < 10 texts, while only 8 occur in ≥ 20. In this case, the data are clear indicators of their high technical level as they are employed in relation with very specific topics, hence their low text distribution throughout the corpus.

Finally, as shown in the table below, keyness also changes to a certain extent. Ten wordforms show a negative keyness value which implies that they appear more often in LACELL, the general corpus, than in UKSCC and could thus be regarded as words belonging to general usage, they are: *plus, via, ratio, nil, ibid, vis, subpoena, consensus, persona, forum.* Besides, four of these types coincide with the ones already classified as general: *plus* and *nil,* and academic: *plus, via* and *ratio* by comparison with *BNC* and *AWL.* Only four (12.5%) of the 32 wordforms in the list show a higher keyness value than the corpus average: 116.08, as it happened with the unfiltered list (11.76% of 187 were above this figure in table 1). On the other hand, nine of them display <10 value in this section. Therefore, they could be considered relevant for the genre although considerably far from the average count obtained for crime nouns, showing values of 28.17 and 97.07 respectively.

TYPE	FREQ UKSCC	TEXT RANGE	KEYNESS	RATIO
Obiter	36	23	158.15	∞

Gravamen	4	4	17.93	∞
Dicta	57	35	227.29	159.6
Affidavit	65	17	176.22	18.82
Extempore	7	4	17.85	14.7
Dictum	24	19	55.36	12.6
Vide	4	1	10.74	11.2
Strata	35	2	64.97	8.4
Proviso	33	20	59.01	7.92
Nexus	15	6	19.22	4.84
Amicus	4	1	5.08	4.8
Interim	147	32	172.039	4.67
Alibi	7	5	8.4	4.52
Memorandum	66	21	67.86	4.13
Quantum	25	13	83.97	3.28
Audit	108	6	71.35	2.96
Lex	6	5	2.49	2.19
Caveat	5	5	1.3	1.82
Locus	9	6	1.94	1.71
Quasi	15	10	2.35	1.57
Sic	44	23	1.31	1.20
Veto	25	3	0.00	0.98
Forum	56	20	-0.01	0.98
Persona	8	4	-0.03	0.93
Consensus	42	11	-0.37	0.90
Subpoena	8	2	-0.2	0.85
Vis	13	8	-0.65	0.79
Ibid	24	15	-2.8	0.71
Nil	36	13	-9.41	0.61
Ratio	23	16	-30.07	0.38
Via	29	31	-54.65	0.27
Plus	54	28	-207.15	0.22

Table 2. Filtered data: single word Latin types.

Following the same procedure deployed to extract single-word terms, the list of two/three-word Latin clusters in table 3 was produced. The greatest difference between single and multi-word Latin terms lies basically in the fact that 19 out the 50 items in the table do not appear in LACELL, the reference corpus, therefore displaying higher keyness scores. This, coupled with the fact that only one of them shows a negative value for this parameter, are clear indicators of the greater representativeness of these units. Moreover, their average keyness value is 92.80, almost five times as much as single word terms.

On the contrary, raw frequency counts are very similar for both listings: 32.31 for single wordforms and 27.66 for two/three-word clusters. Likewise, both lists share text coverage values, slightly higher for two-word terms: 0.00055% against 0.00041% for single word units.

Finally, as regards text distribution, the figures are also quite similar to those in table 2 since single word terms are present in 12.81 texts while they do so in 10.6 in table 3 on average. However, 35 two-word terms occur in <10 texts as opposed to 15 in table 2, once more signaling the higher technical character of the former.

TYPE	FREQ UKSCC	TEXT RANGE	KEYNESS	RATIO
Ex turpi causa	129	3	578.10	∞
Doli incapax	36	1	161.33	∞
Quantum meruit	27	5	121.00	∞
Mutatis mutandis	24	18	107.55	∞
Alter ego	21	5	94.11	∞
Forum non conveniens	13	3	58.26	∞
Actus reus	10	5	44.81	∞
Ad litem	10	3	44.81	∞
Usque ad coelum	8	1	35.85	∞
Pari delicto	7	1	31.37	∞
Ratione personae	6	3	26.89	∞
Doli capax	5	1	22.41	∞
Debet ese	4	1	17.93	∞
Ad factum	4	1	17.93	∞
Res iudicata	4	2	17.93	∞
De novo	4	3	17.93	∞
Praesumptio juris	3	1	13.44	∞
Jus cogens	3	1	13.44	∞
In par materia	3	2	13.44	∞
De jure	52	5	210.42	145.6
Pari passu	28	4	111.23	117.6
Ex parte	115	26	447.92	96.6
Ultra vires	79	16	302.41	82.95
Et seq	29	17	110.72	81.2
A fortiori	32	28	119.19	67.2
Novus actus	16	1	59.59	67.2
Inter alia	124	77	436.87	47.34
De minimis	11	6	38.58	46.2
Inter partes	5	2	17.23	42
Prima facie	109	49	371.54	39.80
De facto	230	111	779.21	38.64
Ipso facto	17	10	53.73	28.56
Mens rea	32	11	96.98	24.43
In absentia	10	2	26.84	16.8
Amicus curiae	4	3	10.74	16.8
Magna carta	11	1	28.57	15.4
Sui generis	3	3	7.16	12.6
Pro rata	30	30	67.07	10.95
Habeas corpus	10	5	21.88	10.5
In camera	9	3	17.4	8.4

Ex officio	5	1	8.6	7
Sine qua non	5	4	8.6	7
In personam	6	2	8.72	5.6
Ab initio	6	4	7.97	5.04
Per se	20	10	18.04	3.5
Quid pro quo	6	3	5.14	3.36
Ex gratia	6	3	4.71	3.15
Bona fide	14	8	8.86	2.73
Per annum	21	12	0.01	1.02
Vice versa	17	13	-0.09	0.92

Table 3. Filtered data: two/three-word clusters.

To conclude, a classification into levels of specialisation of both sets of vocabulary items will be carried out following the method described by Chung (2003) based on the ratio of occurrence of the vocabulary items and their raw frequency counts. The calculation of a word's ratio consists in dividing its normed frequency in the study corpus (in this case UKSCC) by the same value in the reference corpus (LACELL). Since a ≤3 frequency threshold was established for practical reasons, the first and fourth groups defined by Chung have not been included in this study (both refer to items occurring only once in the specific corpus).

After contrasting her results with a qualitative classification method, as already explained, known as the *Rating Scale approach,* and obtaining 86% level of coincidence, Chung concludes that the wordforms falling in groups 1, 2 and 3 are terms (technical words) and those in groups 4 to 6 are non-terms. Nevertheless, as remarked by Rea (2008: 109), this method is an attempt to find a reasonably reliable way to classify vocabulary automatically. When in doubt, the last decision to include a wordform in a given category must be made by the researcher employing other criteria (often qualitative) to do so.

This is precisely why the contexts of occurrence of some of the items falling in groups 5 and 6 (non-terms in principle) were checked using the *Concordance* tool in *Wordsmith 5* to make sure that they actually appeared in a non-specific environment in the reference corpus. As a result, only 8 out of 83 Latin wordforms were re-classified as terms owing to the fact that they did occur either in judicial decisions or at least in a legal context in the general corpus, namely, *affidavit* (R= 18); *prima facie* (R=39); *mens rea* (R=24); *inter partes* (R=42); *amicus curiae* (R=16); *in camera* (R=8)*; lex* (R=2) and *amicus* (R=4). As shown by the figures, except for three of them, the rest display reasonably high ratio values in comparison with the total average: 9.28 for single-word units and 34.06 for two/three-word clusters (not including the infinity value in this calculation). Once more it appears that, if ratio be

regarded as an indicator of the level of specialization of a vocabulary item, multi-word ones are, by far, much more technical than single word units.

The fact that only 9.6% of the items were not accurately classified applying Chung's ratio method proves how reliable it can be as a point of departure to discriminate terms from non-terms in an automatic way, especially when managing large amounts of data. Nevertheless, the researcher must always intervene in the process to guarantee that the conclusions obtained are reliable and reflect the actual usages of the vocabulary items under examination.

Therefore, after filtering the initial list of 187 Latin wordforms, they will be classified into levels of specialization as follows:

1- General and academic: *plus* (used in everyday English and academic environments)*; ratio* and *via* (just academic).

2- Non-technical vocabulary (shared by different environments, not exclusive of judicial decisions): *Extempore, vide, nexus, alibi, caveat, locus, quasi, persona, subpoena, vis, de minimis, ipso facto, in absentia, magna carta, sui generis, habeas corpus, ex officio, sine qua non, in personam, ab initio, quid pro quo, ex gratia, bona fide, vice versa, affidavit, dictum, strata, proviso, interim, memorandum, quantum, audit, sic, veto, forum, consensus, ibid, nil, inter alia, de facto, pro rata, habeas corpus, per se, per annum*.

3- Technical vocabulary (only occurring in judicial decisions): *gravamen, lex, amicus, affidavit, actus reus, ad litem, usque ad coelum, pari delicto, ratione personae, doli capax, debet esse, ad factum, res iudicata, de novo, praesumptio juris, jus cogens, in par materia, obiter, dicta, ex turpi causa, doli incapax, quantum meruit, mutatis mutandis, alter ego, forum non conveniens, de jure, pari passu, ex parte, ultra vires, et seq, a fortiori, novus actus, inter alia, in camera, inter partes, amicus curiae, prima facie, mens rea.*

From the list provided, it can be inferred that only 3.61% of the Latin wordforms analysed are used as general or academic vocabulary while 54.21% of this inventory are non-terms, that is, they are not restricted to the environment of judicial decisions but shared by other varieties and usages of English. In contrast, 42.21% belong to the technical category formed by those vocabulary items which are exclusively employed in the specific environment of law reports, a considerably high value for such a restricted vocabulary inventory.

As for their structure, it is noticeably remarkable that only 4 out of 35 technical terms, 11.42% of the total, are single-word units while the percentage multiplies by five within the non-term group: 53.33%. In contrast, almost 88.5% of them are multi-word clusters in the specialised group. Considering all other indicators that had

already pointed at the greater technical character of multi-word Latin units as opposed to single wordforms, it has been demonstrated that they are much more specialised.

Furthermore, it also appears that single Latin wordforms are highly lexicalised, they do not only occur in the specialised environment of judicial decisions but also in other genres within legal English, other varieties of English, or even in general English contexts as shown by the linguistic evidence provided by corpora.

CONCLUSIONS AND FURTHER RESEARCH

This study has described the linguistic behaviour of Latin wordforms and clusters within judicial decisions from a corpus-based perspective. An *ad hoc* specialised corpus, UKSCC, has been designed and compiled in order to employ it as a reliable source of information. Different parameters related to the statistic relevance, significance and representativeness of the items selected have been studied in comparison with a general English corpus, LACELL, namely, frequency, text distribution and keyness reaching the following conclusions:

1- Given the striking differences between filtered and unfiltered data, it appears essential to apply different methods to filter type lists in order to obtain reliable results. Regarding the frequency of occurrence of Latin types, the counts are divided by four after filtering the list of single word units, 32.31, and by five in the case of two/three-word clusters, 27.66, standing far from the mean value for this parameter in the whole corpus. Similarly, text range is also affected by filtering as it sharply falls by 50% in both cases. However, text coverage does not change so dramatically: from 0.0059% to 0.0055% and 0.0041% for multi-word and single word units respectively. Regarding keyness, the greatest difference before and after filtering the list can be found amongst single word terms which display a mean value of 28.17 (five times less than the corpus average) against two/three-word items: 92.8.

2- Consequently, the use of reliable methods to filter and classify vocabulary such as Chung's (2003) to ensure the reliability of the results obtained becomes fundamental. In this case, the application of Chung's method yields 90.4% success in term identification.

However, the researcher must guarantee that his/her conclusions are based on real fact by also intervening manually through the study of the contexts of usage

of the most problematic vocabulary items, needless to say that corpus processing software only computes character sequencing and does not account for such phenomena as synonymy, homonymy or polysemy, amongst others. As Chung herself affirms, the most reliable method is the qualitative one although it is time-consuming and not applicable to manage large amounts of data for obvious reasons.

3- As regards the actual behaviour of Latin wordforms and clusters, once the list has been properly filtered, it appears that, in general, they are not noticeably frequent within the corpus as they stand, especially if compared with the corpus average, around 80 points below it, slightly less in the case of clusters, as shown above. The percentage of text coverage provided by these units is, however, considerably high in comparison with other typical vocabulary items such as crime nouns, 0.0048% against 0.00095% respectively.

Nonetheless, the most outstanding feature of these items is the difference existing between single-word units and clusters as regards their level of representativeness. While single-word items display much lower keyness counts than the corpus average (only 32 show higher ones), clusters score 3.2 times as much as single wordforms in spite of their lower frequency (5 points less than the latter) thus being more representative of the genre.

Likewise, clusters are more specialised than single wordforms: 88.5% of the items included in the group of technical terms, after applying Chung's vocabulary classification method, are multi-word units whereas the non-term and general vocabulary groups are composed of 57.44% single wordforms against 42.56% of clusters.

Finally, the low text distribution levels of Latin types reinforce their specialised character. They appear in a relatively small amount of texts on average (approximately 10 to 12 corpus texts). Moreover, 46.87% of single wordforms and 70% cluster occur in less than 10 texts confirming, once more, the greater technical character of the latter.

To conclude, due to the data obtained after this analysis and the fact that, to the best of our knowledge, there are no descriptive studies accounting for the behaviour of Latin wordforms and clusters in legal English, it appears interesting to explore in greater depth the contexts of usage of these words so as to provide data that could serve to delimit the boundaries between the specialised and shared usages of non-terms in order to come up with a reliable inventory of semi-technical Latin words.

Similarly, as further research, this information could be also contrasted with other characteristic items of the legal vocabulary to enrich this corpus-based study with more descriptive information on the lexical traits of legal English.

REFERENCES

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

Biber, Douglas. "Representativeness in Corpus Design." *Literary and Linguistic Computing* 8.4 (1993): 243-257.

Biber, Douglas, Susan Conrad and Randi Reppen. *Corpus Linguistics. Investigating Language Structure and Use.* New York: Cambridge University Press, 1998.

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

Callanan, Helen and Linda Edwards. *Absolute Legal English.* London: Delta, 2010.

Chung, Teresa Miwha. *Identifying Technical Vocabulary.* Doctoral thesis. Victoria University of Wellington, 2003.

Coxhead, Averil. "A New Academic Word List." *TESOL Quarterly* 34.2 (2000): 213-238.

Coxhead, Averil and Paul Nation. "The Specialised Vocabulary of English for Academic Purposes." *Research Perspectives on English for Academic Purposes.* Eds. Flowerdew John, and Matthew Peacock. Cambridge: Cambridge University Press, 2001: 252-267.

Fernández, Ramón Luis and Isidro Almendárez. *A Guide to Legal English/ Inglés para juristas.* Madrid: Síntesis, 1994.

Farrell, Paul. *Vocabulary in ESL: A Lexical Analysis of the English of Electronics and a Study of Semi-technical Vocabulary.* Dublin, Ireland: Centre for Language and Communication Studies, 1990.

Frost, Andrew. *English for Legal Professionals.* New York: Oxford University Press, 2009.

Krois-Linder, Amy and Matt Firth. *Introduction to International Legal English: A Course for Classroom or Self-study Use*. Cambridge: Cambridge University Press, 2008.

Maley, Yon. "The Language of the Law." *Language and the Law.* Ed. John Gibbons. London: Longman, 1994: 11-50.

Marín, María José and Camino Rea. "Design and Compilation of a Legal English Corpus Based on UK Law Reports: The Process of Making Decisions." *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de córpora*. Eds. María Luisa Carrió Pastor and María Ángeles Candel Mora. Actas del III Congreso Internacional de Lingüística de Corpus. Valencia: Universitat Politècnica de València, 2011: 101-110.

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

McEnery, Tony and Andrew Wilson. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

Nation, Paul. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press, 2001.

Nation, Paul and Robert Waring. "Vocabulary Size, Text Coverage and Word Lists." *Vocabulary: Description, Acquisition and Pedagogy*. Eds. Schmitt, Nobert, and Michael McCarthy.Cambridge: Cambridge University Press, 1997.

Orts, María Ángeles. *Aproximación al discurso jurídico en inglés. Las pólizas de seguro marítimo de Lloyd´s*. Madrid: Edisofer, 2006.

——. *Inglés jurídico y económico: manual para su aprendizaje y traducción*. Murcia: Diego Marín, 2010.

Pearson, Jennifer. *Terms in Context*. Amsterdam: John Benjamins Publishing Company, 1998.

Rea, Camino. *El inglés de las telecomunicaciones: estudio léxico basado en un corpus específico*. Tesis doctoral. Universidad de Murcia, 2008. URL: http://www.tdx.cat/handle/10803/10819

——. "Getting on with Corpus Compilation: From Theory to Practice." *ESP World* 9.1 (27) (2010): 1-23.

Rice, Sally and Gillian Brown. *Professional English in Use: Law*. Cambridge: Cambridge University Press, 2007.

Sánchez, Aquilino, Ramón Sarmiento, Pascual Cantos and José Simón. *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis*. Madrid: SGEL, 1995.

Scott, Mike. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software, 2008.

Sinclair, John. "Corpus and Text: Basic Principles." *Developing Linguistic Corpora: A Guide to Good Practice*. Ed. Martin Wynne. AHDS Literature, Languages and Linguistics: University of Oxford, 2005. URL http://ota.ahds.ac.uk/documents/creating/dlc/index.htm

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

West, Michael. *A General Service List of English Words*. London: Longman, 1953.

Wynne, Martin. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 2005.

Yang, Huizong. "A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts (An Interim Report)." *Literary and Linguistic Computing* 1.2 (1986): 93-103.

How to cite this article:

Marín Pérez, Mª José. "How Relevant are Latin Wordforms and Clusters in Legal English? A Corpus-based Study on the Representativeness and Specificity of such Elements in UKSCC: An *ad hoc* Legal Corpus." *ES. Revista de Filología Inglesa* 33 (2012): 161-182.

Author's contact: mariajose.marin1@um.es