Research Paper

Multi-Word Sequences in Legal Discourse

TORIKAI Shinichiro

Key words: n重字、連語、コロケーション、定型語句論、司法英語
n-gram, multi-word sequences, collocation, phraseology, legal discourse

Abstract

The objectives of this article are to select a sample entry word for the project of compiling a corpus-based production-oriented legal English dictionary for the Japanese students of law. The article first discusses collocation, the way words are actually used with other words in our communication. Next, the theory of n-grams is introduced to find frequent multi-word sequences used in the BNC and in the legal corpora which the author's project team compiled. Then, the characteristics of multi-word sequences are analyzed and compared between the general discourse and the legal discourse. It is found that in general discourse a wider variety of multi-word sequences are used at comparatively low frequencies while in legal discourse relatively limited types of multi-word sequences, and proper names of laws and legal institutes are used at high frequencies. Finally, a sample head word *application* is selected and illustrated with typical collocates in order to help Japanese students of law be able to use them in more productive ways.

1. Collocation and collocates

It is often said that a word is not used alone. It is commonly used with other words in our communication. For example, a word *dog* is used 11,795 times in *the British National Corpus* (hereafter *BNC* for short), but based on the analysis with a computer software, *Sketchengine*, *dog* is used 3,378 times with pre-modifiers which describe *dog* (e.g. *guide dog*, *stray dog*, *guard dog*, *pet dog*, etc.) and used 1,532 times with other nouns which *dog* pre-modifies (e.g. *dog owner*, *dog handler*, *dog warden*, *dog food*, etc.). *dog* is used 2,930 times as the subject of many types of verbs (e.g. *dog bark*, *dog eat*, *dog chase*, *dog run*, etc.) and 3,454 times as the object of various types of verbs (e.g. *walk dog*, *train dog*, *bark dog*, *keep dog*, etc.). This linguistic phenomenon that a word is used closely connected with other words is traditionally called collocation. *The Oxford English Dictionary* (1989) defines collocation as follows:

> c. *Linguistics*. Habitual juxtaposition or association, in the sentences of a language, of a particular word with other particular words; a group of words so associated.
>
> Introduced by J. R. Firth as a technical term in modern Linguistics,…

John Rupert Firth (1957), the founder of the London school, explained the nature of a word in terms of the relationship with other words.

With the advent of computer technology, the study of collocation became more active and widespread. This is mainly due to the fact that the computer can examine massive linguistic data and find frequent sequences of words almost instantaneously. John Sinclair, who started the COBUILD project, defined collocation as follows:

> Collocation is the occurrence of two or more words within a short space of each other in a text. (1991: 170)

Sinclair's definition of collocation is purely from the viewpoint of corpus linguistics which focuses on forms and sequences of words computers can recognize. Teubert (2004) emphasizes the importance of collocations and claims as follows:

> Not single words but collocations constitute the true vocabulary of a language.
>
> (p. 188)

His claim sounds rather too extreme as Leech (2011) criticizes by pointing out that Teubert does not consider *the idiom principle* claimed by Sinclair (1991). However, we can tell that collocations are the key to understanding how the words are actually used in real communication.

Yet, one may ask that if only some words co-occur side by side frequently, is it all right to call them collocations? If we look at natural language use, we can find hundreds of examples where some high frequency words co-occur with others extremely often. For example, *the* and *of* are ranked as the two most frequent words in the *BNC*; *the* occurring 5,415,707 times and *of* 3,027,441 times, and the combination of *of* and *the* occurs 753,195 times, the most frequent two-word combination in the *BNC*. But no one thinks of this combination as a collocation because the population of these two words is so large that it is quite natural that the chances of these two words co-occurring are also high.

Consequently, how can we distinguish real collocations from others? Corpus linguistics often tried to find the answer to this kind of question in statistics. Nowadays statistical formulae are built into most computer software. *Scketchengine*, for instance, has seven built-in statistical formulae, namely T-score, MI, MI3, log likelihood, min. sensitivity, logDice, MI.log_f, to help us find useful collocations.

The linguistic phenomena of collocation have received considerable attention from the dictionary writers and editors. Many English dictionaries published these days, i.e. *Longman Dictionary of Contemporary English* (2014) (hereafter LDOCE for short), *Collins COBUILD Advanced Learner's Dictionary* (2014) (hereafter COBUILD for short), *The Wisdom English-Japanese Dictionary* (2013), and *Genius English-Japanese Dictionary* (2014), and *Oxford Advanced Learner's Dictionary* (2015) have useful information on collocations, and many outstanding English collocation dictionaries and thesauruses are also published based on the achievements of corpus linguistics, i.e. *Oxford Collocations Dictionary for Students of English* (2009) and *Longman Collocations Dictionary and Thesaurus* (2013).

2. Multi-word sequences based on n-gram analysis

With the help of computer software, we can make an exhaustive collocation list of a particular word instantaneously if we specify the node word we want to survey. In this case, how can we find the word we think useful to survey its collocates in the discourse? One of the most effective ways is to make use of the n-gram model. The idea of n-grams was originally advocated by Claude Elwood Shannon (1964). He explains the n-gram model as follows:

> The zero-order approximation is obtained by choosing all letters with the same

> probability and independently. The first-order approximation is obtained by choosing successive letters independently… In the second-order approximation, diagram structure is introduced. After a letter is chosen, the next one is chosen in according with the frequencies with which the various letters follow the first one... In the third-order approximation, trigram structure is introduced. Each letter is chosen with probabilities which depend on the preceding two leters. (pp. 42-3)

Shannon (1964: 43) named second-order approximation digram structure, third-order approximation trigram structure, and tetragram structure a n-gram. An n-gram in linguistics is a sequence of a given number of particular linguistic elements, typically words or letters. The occurrence count of n-grams shows how frequently these n-grams appear in a particular piece of discourse. The idea of n-grams can be explained as follows. Suppose there is a piece of discourse as follows:

The boys played baseball in the park but the girls played baseball in the field.

When the value of n is 1, we divide the discourse by one word as follows:

The/boys/played/baseball/in/the/park/but/the/girls/played/baseball/in/the/field.

The consequences are: we have 15 subdivisions which are grouped by a word form; *the* (4), baseball (2), *in* (2), *played* (2), *boys* (1), *but* (1), *field* (1), *girls* (1), *park* (1). The word *the* is used four times; *baseball*, *in* and *played* are used twice.

When the value of n is 2, we divide the discourse by two word forms as follows:

The boys/played baseball/in the/park but/the girls/played baseball/in the/field.
The/boys played/baseball in/the park/but the/girls played/baseball in/the field.

Notice that there are two patterns of cutting the discourse: the one starts cutting after the second word *boys*, the other starts cutting the discourse after the first word *The*. The results are: we have 16 subdivisions, 14 of them are made up of two words and two of them with one word. The details are as follows: *played baseball* (2), *in the* (2), *baseball in* (2), *the boys* (1), *park but* (1), *the girls* (1), *boys played* (1), *baseball in* (2), *the park* (1), *but the* (1), *girls played* (1), *the field* (1), *field* (1), *the* (1).

When the value of n is 3, we divide the discourse by three words as follows:

The boys played/baseball in the/park but the/girls played baseball/in the field.
The boys/played baseball in/the park but/the girls played/baseball in the/field.
The/boys played baseball/in the park/but the girls/played baseball in/the field.

The results are: we have 13 three-word subdivisions, two two-word subdivisions and two one-word subdivisions. We have 11 different types of thee-word slots: *the boys played* (1), *baseball in the* (2), *park but the* (1), *girls played baseball* (1), *in the field* (1), *played baseball in* (2), *the park but* (1), *the girls played* (1), *boys played baseball* (1), *in the park* (1), *but the girls* (1), two different types of two-word subdivisions: *the boys* (1), *the field* (1), and two different types of one-word subdivisions: *field* (1), *the* (1).

The analysis of n-grams indicates the probability that a certain word is more likely to co-occur with particular words than others. For example, from the analysis of the above small experimental sample discourse, we found that *baseball* is more likely to occur after *play*, and so does *the* after *in* and *in* after *baseball*.

I did the same experimental analysis on the BNC using the n-gram formula built in *Sketchengine*. The findings are shown in Table 1 below:

Table 1. N-grams in the BNC

1-gram		2-grams		3-grams		4-grams		5-grams		6-grams	
the	54157.07	of the	7531.95	I do n't	371.47	I do n't know	119.04	at the end of the	37.9	on the other side of the	6.21
of	30274.41	in the	4801.92	one of the	297.8	the end of the	103.74	I do n't know what	18.01	at the end of the day	6.03
to	25669.78	to the	2869.59	the end of	207.19	at the end of	78.44	I do n't want to	17.97	ask the Secretary of State for	6.01
and	25121.71	on the	2075.57	as well as	168.46	I do n't think	69.85	in the middle of the	14.14	To ask the Secretary of State	5.95
a	20418.49	and the	1882.28	part of the	166.99	at the same time	48.12	as a result of the	13.72	mm mm mm mm mm mm	4.39
in	17877.23	to be	1878.62	do n't know	154.06	the rest of the	47.12	by the end of the	13.28	from the point of view of	4.03
that	10623.85	for the	1595.8	out of the	152.38	for the first time	47.12	the other side of the	12.08	by the end of the year	3.97
is	9729.21	at the	1380.48	a number of	137.9	per cent of the	45.22	the Secretary of State for	12.08	my hon. Friend the Member for	3.63
was	8778.6	that the	1271.84	a lot of	136.64	as a result of	44.68	at the time of the	10.23	in such a way as to	3.46
I	8618.25	by the	1253.65	end of the	133.97	one of the most	32.86	I do n't think I	9.96	in the middle of the night	2.64
for	8318.05	with the	1241.54	be able to	133.82	is one of the	32.67	the end of the year	9.35	the Department of Trade and Industry	2.57
it	8197.97	of a	1240.2	some of the	128.85	do n't want to	32.66	at the top of the	9.35	in the second half of the	2.47
on	6950.02	from the	1203.7	to be a	116.97	in the case of	32.45	for the first time in	8.66	at the other end of the	2.41
be	6485.75	in a	1064.86	the fact that	113.74	I do n't want	32.39	I do n't know how	8.48	Secretary of State for the Environment	2.37
with	6404.2	it is	909.29	per cent of	113.35	to be able to	31.67	the end of the day	8.45	I do n't know what you	2.35
The	6195.11	it was	864.02	there is a	104.77	the Secretary of State	30.59	I do n't think it	8.18	I do n't want to be	2.19
as	6035.79	as a	817.58	I did n't	103.53	On the other hand	28.36	on the part of the	8.14	The hundred shares index closed down	2.18
you	5747.82	do n't	815.16	in order to	102.22	in the form of	27.57	at the beginning of the	7.89	the Secretary of State for the	2.17
at	4872.49	is a	777.14	I ca n't	99.86	on the basis of	27.43	At the end of the	7.62	if he will make a statement	2.13
by	4867.25	with a	757.64	in terms of	93.56	the top of the	26.73	on the other side of	7.46	The hundred shares index closed up	2.11

As you see, 1-grams are the same as the occurrence counts of individual words in the *BNC*. The results of 2-grams are more like the combinations of the frequent 1-grams. All of these top 20 two-word sequences are the combinations of frequent grammatical words. In the 3-gram list some lexical words such as *end*, *part*, *know*, *number*, *able*, *fact*, and *terms* appear. Some of these three-word sequences are traditionally recognized as set phrases or phrasal expressions. All the 4-gram examples are cohesive sequences of four words we often recognized as a semantic unit in our daily communication. 5-grams and 6-grams are more like adding one or two more words to the 4-gram set phrase sequences.

These frequent multi-word sequences have received a lot of attention from many linguists. Sinclair (1991) argued the nature of these groups of words under the name of "idiom principle". Huston and Francis (1996) wrote "Pattern Grammar" based on these frequent word sequences. Biber (2006) named them "lexical bundles" and examined the use of these frequent multi-word sequences in the academic context. Yamada (2007) applied the n-gram theory to the analysis of Chinese classics. Stubbs (2007) argued phraseology from the viewpoint of n-grams. Koyama (2008, 2009) applied multi-word expressions to the ESP in science and technology. Simpson-Vlach and Ellis (2010) compared the academic speech and writing corpora of 2.1 million words each with the Switchboard corpus of 2.9 million words, the FLOB and Frown corpora. They extracted 607 written and spoken academic word sequences and categorized them under the functional categories. Martinez and Schmitt (2012) made the Phrasal Expressions List of 505 frequent non-transparent academic multi-word sequences. They name multi-word sequences formulaic language, and claim as follows:

> in essence, most definitions indicate that individual formulaic sequences behave much the same as individual words, matching a single meaning or function to a form, although that form consists of multiple orthographic or phonological words. (P.299)

3. Objectives, Data, and Methodology

The objectives of this paper are to explore the frequent multi-word sequences in legal discourse and try to apply the research findings to the project of compiling a corpus-based production-oriented legal English dictionary.

The data I am going to use in this article are the ones Professor Tamaruya, the College of Law and Politics, Rikkyo University, Associate Professor Takahashi, Miyagi University of Education, and I collected for the above project. This project is supported by the Japanese government funding for scientific research (# 16H03458). The legal corpora I am going to use are as follows:

Law Journals issued in the United Kingdom in 2015 (hereafter abbreviated as UK LJ): 5,911,156 words. The articles are downloaded from the following law journals:

> *Cambridge Law Journal*, *Dublin University Law Journal*, *Edinburgh Law Review*, *European Law Review*, *International & Comparative Law Quarterly*, *Journal of Business Law*, *Law Quarterly Review*, *Legal Studies, Modern Law Review*, *Oxford Journal of Legal Studies*, *Public Law, Edinburgh Law Review*, *UCL Journal of Law and Jurisprudence*

Law Journals issued in the United States in 2015 (hereafter abbreviated as US LJ): 5,952,782 words. The articles are downloaded from the law journals of the following universities:

> Yale University, Harvard University, Stanford University, Columbia University, University of Chicago, New York University, University of Pennsylvania, University of California – Berkeley, University of Michigan - Ann Arbor, University of Virginia

I am going to use the corpus software, *Sketch Engine*, and the built-in statistical formulae it contains.

4. Quantitative analysis of legal discourse based on n-grams

I analyzed how words are quantitatively used in legal discourse by using an n-gram research function. Appendix 1 shows how frequently the top 40 n-grams appear in legal discourse. I included in the appendix the top 40 n-grams of the BNC in order to compare legal discourse with general discourse. All the frequent counts are normalized per million words (hereafter abbreviated as cpm).

4. 1. Frequent 1-gram words

It is interesting that the top eight 1-gram words in the BNC, UK LJ and US LJ consist of exactly the same grammatical words, namely *the*, *of*, *to*, *and*, *are*, *in*, *that* and *is*, and the rest of the top 40 1-gram words are very alike. As many as 27 words are common in the top 40 1-gram lists of three corpora. They are:

> *a*, *an*, *and*, *are*, *as*, *at*, *be*, *but*, *by*, *for*, *from*, *has*, *have*, *in*, *is*, *it*, *not*, *of*, *on*, *or*, *that*, *the*, *this*, *to*, *was*, *which*, *with*

Although these words are common, their ranking orders are different depending on the corpus. For example, *it* is ranked 12th in the BNC, but is ranked 15th in the UK LJ and 18th in the US LJ. *At* is 19th in the BNC, 30th in the UK LJ and 34th in the US LJ. However, all these findings seem to indicate that even though the type of discourse is different, the fundamental grammatical structures constructed by these grammatical words are quantitatively similar. This becomes salient when we examine longer multi-word sequences later.

However, there are some significant differences between the general discourse and legal discourse. For example, personal pronouns are frequent in the BNC: *I* (ranked 10th), *you* (ranked 18th), *he* (ranked 21st), *his* (ranked 27th), *her* (ranked 36th), *we* (ranked 37th), *one* (ranked 39th), but they do not appear in the two legal corpora. Frequent use of personal pronouns often happens in our daily conversation as Biber (1988) claims that the frequent use of the first and the second pronouns are typical characteristics of "Involved Production" (p. 107), which Biber explains as follows:

> associated in one way or another with an involved, non-informational focus, due to a primarily interactive or affective purpose and/or to highly constrained production circumstances. (p. 105)

Telephone conversations and face-to-face conversations are, according to Biber, typical registers of the involved production where *you* and *I* are frequently used.

Another interesting linguistic phenomenon is the limited use of the past tense forms of the *be* verb in legal discourse compared with that in general discourse. As Table 2 shows, *was* is used 8,779 times (ranked 9th) and *is* is used 9,729 times (ranked 8th) in the BNC. We understand that *was* and *is* are used almost at the same frequencies in the BNC. On the other hand, the past tense *be* verbs are used much less frequently in legal discourse. *Was* is used 5,222 times (ranked 21th) in the UK LJ and is used 2,764 times (ranked 31st) in the US LJ. *Was* occurs about one third or one fourth as frequently as *is* in frequencies in legal discourse. The same phenomenon can be observed in the use of *were* and *are* in the BNC and in the legal corpora. The frequencies of *was* and *were* are only one third or one fourth of the frequencies of *is* and *are* in the legal corpora. This infrequent use of past tense in legal discourse is one of the salient characteristics of official documents (Biber 1988).

Table 2. Frequencies of *was*, *is*, *were* and *are*.

	was		is		were		are	
	cpm	rank	cpm	rank	cpm	rank	cpm	rank
BNC	8,779	9	9,729	8	3,120	35	4,551	22
UK LJ	5,222	20	14,870	8	1,641	54	4,776	24
US LJ	2,764	31	11,551	8	1,654	53	5,563	15

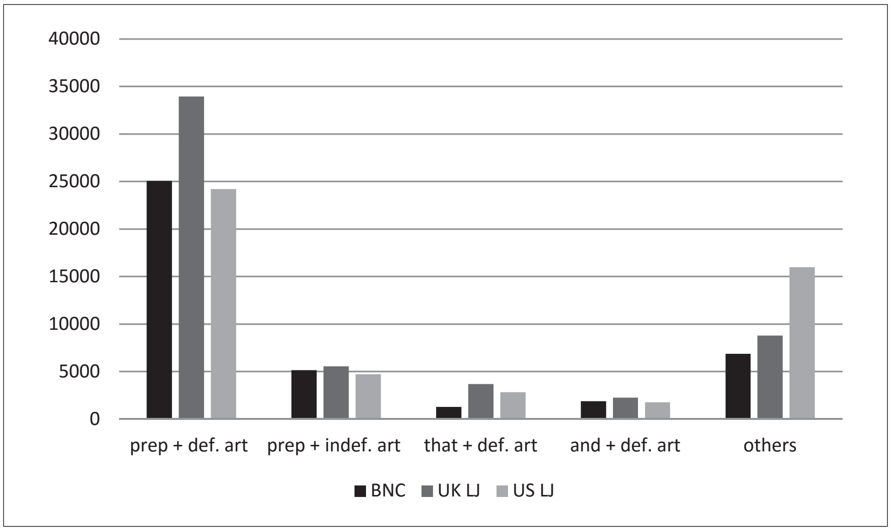
4. 2. 2-gram word sequences

Most of the top 40 2-gram word sequences in each corpus are basically the combinations of their top 40 1-grams. The preposition-definite article combination is outnumbered. In the BNC 11 out of 40 2-grams are this type of combinations which sum up to 25,046 counts per million words, accounting for 62% of the total number of the top 40 2-grams of 40,214 cpm. If we include the total number of preposition-indefinite article combinations, they will account for 75% of the top 40 2-grams in the BNC. This high percentage of preposition-article combinations is basically the same in the legal corpora. The total number of preposition-definite article combinations is 33,941 cpm, accounting for 63% of the top 40 2-grams in the UK LJ, and 22,937 cpm, accounting for 46% of the top 40 2-grams in the US LJ. If we include preposition-indefinite article combinations these numbers will increase to 39,493 cpm and 73% in the UK LJ, and 27,649 cpm and 56% in the US LJ. (See Table 3 and Figure 1)

Table 3. Frequencies of combinations of a preposition/*that/and* and the article

	BNC	UK LJ	US LJ
prep + def. art	25,046	33,941	24,204
prep + indef. art	5,152	5,552	4,711
that + def. art	1,272	3,696	2,821
and + def. art	1,882	2,245	1,774
others	6,861	8,785	15,974

Figure 1. Frequencies of preposition/*that/and* — article combinations



This high probability of the article occurring right after a preposition seems to be attributable to the structure of the prepositional phrase. It usually consists of a preposition and a noun phrase, and the noun phrase is more likely to start with the definite article than anything else. This very fundamental structure of the English prepositional phrase inevitably contributes to increasing the number of preposition-definite article combinations.

Other noteworthy examples are *don't* in the BNC and *the Court* in legal corpora. Biber (1988) pointed out that contractions are the third most salient characteristics of involved production. This explains why two more contractions, namely *didn't* (49th) and *can't* (96th), are found in the BNC while no contracted forms are found at all within the top 100 2-gram word sequences in the legal corpora. We can easily understand why a highly professional 2-gram like *the Court* is ranked high in legal discourse. Technical terms peculiar to legal English will become more common when the multi-word sequences become longer.

4. 3. 3-gram word sequences

The feature of 3-gram word sequences is completely different from that of 2-gram word sequences. Many of the 3-grams are often recognized as a set phrase or a multiple word expression. Unlike the 2-grams, the meanings of these 3-gram sequences are clear and understandable. This is mainly because most of these 3-gram word sequences contain a lexical word within them. The most common pattern of the three-word sequences is "a grammatical word + a lexical word + a grammatical word" (hereafter GLG for short). The following are all the 12 GLG structures found in the top 40 3-grams in the BNC (the 4-grams expanded from the equivalent are listed on the right side for further reference with a newly added word being underlined):

the end of (207 cpm; 3rd)	*the end of the*, *at the end of*, *by the end of*
as well as (168 cpm; 4th)	*as well as the*
a number of (138 cpm; 8th)	
a lot of (137 cpm; 9th)	
be able to (134 cpm; 11th)	*to be able to*
the fact that (114 cpm; 14th)	*the fact that the*
per cent of (113 cpm; 15th)	*per cent of the*
in order to (102 cpm; 18th)	
in terms of (94 cpm; 20th)	*in terms of the*
the number of (86 cpm; 23rd)	
the rest of (80 cpm; 26th)	*the rest of the*, *as a result of*

the use of (80 cpm; 28th)

In the BNC the total occurrence counts of the 12 GLG structures are 1,453 that account for 32% of the total occurrence of the top 40 3-grams. Seven of these GLGs expand to 10 different types of 4-grams by adding either a new preposition to the head of the GLG or the definite article to the end of the GLG.

The following are the 20 GLG structures found in the UK LJ:

the Court of (333 cpm; 1st)	*the Court of Appeal*, *the Court of Justice*, *of the Court of*
in order to (328 cpm; 2nd)	
the fact that (314 cpm; 3rd)	*the fact that the*
in relation to (298 cpm; 4th)	*in relation to the*
the right to (280 cpm; 5th)	*of the right to*
as well as (272 cpm; 6th)	*as well as the*
the context of (237 cpm; 11th)	*in the context of*, *the context of the*
the basis of (221 cpm; 12th)	*on the basis of*, *the basis of the*
a number of (211 cpm; 16th)	
the scope of (210 cpm; 17th)	*the scope of the*
the law of (209 cpm; 18th)	
the use of (202 cpm; 20th)	
the application of (199 cpm; 22nd)	*the application of the*
in terms of (177 cpm; 28th)	
in respect of (169 cpm; 30th)	
the nature of (162 cpm; 33rd)	*the nature of the*
the absence of (161 cpm; 34th)	*in the absence of*
in case of (160 cpm; 35th)	*in the case of*
the principle of (148 cpm; 38th)	
a matter of (144 cpm; 40th)	*as a matter of*

In the UK LJ the total occurrence counts of the 20 GLG structures are 4,445, accounting for 53% of the total occurrence of the top 40 3-grams. Seven of these GLGs expand to the 4-grams by adding a new preposition to either end of their 3-grams. Eight of them become the 4-grams by adding the definite article to the end of the GLGs, and one of them become two different types of 4-grams by adding a new technical proper noun.

The following are the 17 GLG structures in the US LJ:

as well as (237 cpm; 3rd)	as well as the
in order to (183 cpm; 5th)	
the use of (181 cpm; 6th)	
the fact that (168 cpm; 9th)	the fact that the
with respect to (166 cpm; 10th)	with respect to the
the context of (149 cpm; 12th)	in the context of
the value of (137 cpm; 17th)	the value of the
more likely to (132 cpm; 18th)	are more likely to, more likely to be
a number of (130 cpm; 20th)	
the scope of (129 cpm; 21st)	the scope of the
in favor of (125 cpm; 22nd)	
the number of (122 cpm; 26th)	
the absence of (119 cpm; 27th)	in the absence of
be able to (117 cpm; 30th)	
a matter of (112 cpm; 32nd)	as a matter of
the Court has (106 cpm; 34th)	
in terms of (105 cpm; 36th)	

In the US LJ the total occurrence counts of the 17 GLG structures are 2,418 accounting for 41% of the total occurrence of the top 40 3-grams. Five of these GLGs expand to five different types of 4-grams by adding the definite article, and three of these GLGs expand to three different types of 4-grams by adding a new preposition to the head of their 3-grams.

The unique feature of 3-grams found in the legal corpora is the frequent use of technical proper nouns. The following are the examples from the UK LJ:

the Court of (333 cpm; 1st),
Court of Appeal (266 cpm; 7th),
the Supreme Court (214 cpm; 15th)
House of Lords (144 cpm; 39th)

The same type of examples from the US LJ are:

the United States (442 cpm; 1st),
the Supreme Court (352 cpm; 2nd),
the federal government (180 cpm; 7th),
in the United (170 cpm; 8th)

the Court has (106 cpm; 34th)

N-grams with legal proper names appear more in the lists of 5-grams and 6-grams.

4. 4. 4-gram word sequences

The 4-gram word sequences are characterized by the expansion of some frequent 3-gram word sequences which we have seen. Typical examples of these 3-grams that are expanded to the 4-grams in the BNC are *I don't* (371 cpm; ranked 1st), *one of the* (298 cpm; ranked 2nd) and *the end of* (207 cpm; ranked 3rd). All these 3-grams are expanded to 4-grams by adding a frequent word as follows:

I don't (371 cpm; 1st)	→ *I don't know* (119 cpm; 1st)
	→ *I don't think* (70 cpm; 4th)
	→ *I don't want* (32 cpm; 14th)
cf. *don't know* (154 cpm; 6th)	→ *I don't know* (119 cpm; 1st)
	→ *don't know what* (25 cpm; 22nd)
one of the (298 cpm; 2nd)	→ *one of the most* (33 cpm; 10th)
	→ *is one of the* (33 cpm; 11th)
	→ *was one of the* (23 cpm; 28th)
the end of (207 cpm; 3rd)	→ *the end of the* (104 cpm; 2nd)
	→ *at the end of* (78 cpm; 3rd)
	→ *by the end of* (25 cpm; 23rd)

The way the frequent 4-grams are created from the 3-grams are similar in legal discourse. The following are some examples of them in the UK LJ:

the Court of (333 cpm; 1st)	→ *the Court of Appeal* (180 cpm; 2nd)
the fact that (314 cpm; 3rd)	→ *the fact that the* (104 cpm; 6th)
in relation to (298 cpm; 4th)	→ *in relation to the* (89 cpm; 11th)
the context of (237 cpm; 11th)	→ *in the context of* (186 cpm; 1st)
on the basis (221 cpm; 12th)	→ *on the basis of* (172 cpm; 3rd)
the case of (160 cpm; 35th)	→ *in the case of* (104 cpm; 5th)

The followings are the examples from the US LJ:

the United States (442 cpm; 1st)	→ *in the United States* (159 cpm; 1st)
	→ *of the United States* (75 cpm; 3rd)

the Supreme Court (352 cpm; 2nd) → *the Supreme Court has* (60 cpm; 19th)
as well as (237 cpm; 3rd) → *as well as the* (51 cpm; 23rd)
the same time (103 cpm; 37th) → *At the same time* (68 cpm; 14th)
a matter of (112 cpm; 32nd) → *as a matter of* (66 cpm; 15th)
the absence of (119 cpm; 27th) → *in the absence of* (63 cpm; 17th)

There is an interesting difference in the way 3-grams are expanded to 4-grams in the BNC and in legal discourse. In the BNC some frequent 3-grams such as *I don't* function as the common core part of some 4-grams, and become the basis of such frequent 4-grams as *I don't know*, *I don't think* and *I don't want*. Some 3-grams in the BNC are even interrelated. Very frequent 3-grams of *I don't* and *don't know* are combined and used as the most frequent 4-gram in the BNC, *I don't know*. In legal discourse, however, some 3-grams expand to 4-grams to become the complete name of a particular legal institute.

4. 5. 5-gram and 6-gram word sequences

5-grams and 6-grams are more like the consequences of the expansion of the related lesser n-grams. The same types of n-gram expansion I mentioned in 4.4 continue in 5-grams and 6-grams.

In the BNC the sentence initial type of 4-grams, *I don't know*, *I don't think*, *I don't want* expand rightward into 5-gram and 6-gram sentence initials by adding the first word of the following embedded clause.

I don't know (119 cpm; 1st) → *I don't know what* (18 cpm; 2nd)
→ *I don't know what you* (2 cpm;)
→ *I don't know how* (8 cpm; 14th)
→ *I don't know whether* (7 cpm; 21st)
→ *I don't know if* (7 cpm; 22nd)
→ *I don't know why* (7 cpm; 24th)
I don't think (70 cpm; 4th) → *I don't think I* (10 cpm; 10th)
→ *I don't think it* (8 cpm; 16th)
→ *I don't think so* (6 cpm; 29th)
I don't want (32 cpm; 14th) → *I don't want to* (18 cpm; 3rd)
→ *I don't want to be* (2 cpm; 16th)
→ *I don't want to go* (2 cpm; 30th)

Another type of expansion in the BNC is to add a new word, usually the article or a preposition at either end of the 4-gram sequences. The following group of examples shows how the 3-gram word sequence, *the end of*, develops into three 4-grams, five 5-grams and four 6-grams.

the end of (207 cpm; 3rd)
→ *the end of the* (104 cpm; 2nd)
→ *by the end of the* (13 cpm; 6th)
→ *by the end of the century* (2 cpm; 33rd)
→ *by the end of the year* (4 cpm; 7th)
→ *the end of the year* (9 cpm; 11th)
→ *the end of the day* (8 cpm; 15th)
→ *at the end of the day* (6 cpm; 2nd)
→ *At the end of the* (8 cpm; 19th)
→ *at the end of* (78 cpm; 3rd)
→ *at the end of the* (38 cpm; 1st)
→ *at the end of the year* (2 cpm; 23rd)
→ *by the end of* (25 cpm; 23rd)

Meanwhile, in the legal corpora more and more proper nouns come into the lists of the top 40 5-grams and 6-grams. In the UK LJ 18 out of 40 are proper nouns in the 5-gram list, but they increase to 24 in the 6-gram list. In the US LJ there are six 5-gram proper nouns but they increase to 10 in the 6-gram list. The following group of words show how three different types of 4-gram proper nouns expand to the 6-grams. The way they expand is not in single linear order. It is interesting that one 5-gram and two 6-gram sequences newly get a preposition before the full proper name of *the European Court of Human Rights* is completed.

the European Court of (53 cpm; 29th)
European Court of Human (47 cpm; 36th)
Court of Human Rights (47 cpm; 38th)
→ *European Court of Human Rights* (44 cpm; 2nd)
→ *the European Court of Human* (43 cpm; 3rd)
→ *of the European Court of* (cpm 14; 35th)
→ *the European Court of Human Rights* (40 cpm; 1st)
→ *of the European Court of Human* (12 cpm; 11th)
→ *by the European Court of Human* (6 cpm; 36th)

Another example shows how a multi-word phrase develops in legal discourse and disappears from the list. One of the most typical examples is *the context of*. It appears first in the top 40 3-gram list of the UK LJ. Then, it is ranked first both in the top 40 4-gram list and the top 40 5-gram lists, but it disappears from the top 40 6-gram list. The details are as follows:

the context of (237 cpm; 11th)
→ *in the context of* (186 cpm; 1st)
→ *in the context of the* (52 cpm; 1st)
→ (disappear)

Many of the multi-word phrases like *the context of* make their first appearance in the 3-gram lists. But they either disappear from the 6-gram list or drastically decrease their occurrence counts. This is because the nouns to be used after the last *the* are so diverse in kind that the frequency of each noun becomes too low to appear in the list.

4. 6. Overall observation on the extension of n-grams in general discourse and legal discourse.

Table 4 shows how the total cpm number of the top 40 n-grams in each corpus changes. The percentages indicate the ratio of the n-grams compared with the number of the n-grams in the left column. As you see, the total cpm numbers of 1-grams of the BNC and the US LJ are about the same, around 340 thousand, and the UK LJ around 390 thousand. As the value of n-grams increases from 1 to 6, the difference between the BNC and the US LJ becomes wider, and the total cpm numbers of 6-grams between these two corpora are about 1 to 7. The total cpm numbers of top 40 n-grams of the UK LJ follow basically the same progress. While the total number of 6-grams in the UK LJ reduces by half, the US LJ decreases only by 30%. This is because the frequencies of capitalized legal crèches remain about the same in the 6-gram list of the US LJ.

Table 4. Total number of the top 40 n-grams in the BNC, the UK LJ and the US LJ

	1-gram	2-grams	3-grams	4-grams	5-grams	6-grams
BNC	345,481	40,214 (11.6%)	4,587 (11.4%)	1,388 (30.2%)	368 (26.5%)	106 (28.8%)
UK LJ	390,407	54,219 (13.8%)	8,437 (15.5%)	3,092 (36.6%)	997 (32.2%)	425 (42.6%)
US LJ	347,268	49,484 (14.2%)	5,951 (12.0%)	2,381 (40.0%)	979 (41.1%)	704 (71.9%)

The above table suggests that the number of n-grams decreases more rapidly in general discourse than in legal discourse. The reason for that can be that the general discourse like the BNC is so diverse in register and genre that it contains so many different kinds of multi-word sequences that fit in different registers and genres. That is why the frequencies of the individual n-gram sequences disperse and become low in the BNC. Meanwhile, in the legal discourse which only consists of the same single register, the variety of multi-word sequences is more limited and particular ones peculiar to legal discourse are concentratedly used. This difference becomes salient when linguistic variety becomes particularly apparent in the 3-grams where lexical nouns start to appear. The following three statistical accounts strongly confirm the above postulation.

A. The total number of the top 40 3-grams in the BNC is 4587, while 8437 in the UK LJ and 5951 in the US LJ.
B. The 3-gram structure occurring 100 cpm i.e. *I can't* is ranked 19th in the BNC, while the 3-gram structure occurring 101 cpm i.e. *the risk of* is ranked 84th and the 3-gram structure occurring 99 cpm i.e. *to do so* is ranked 39th in the UK LJ.
C. The 3-gram structure ranked 100th in the BNC i.e. *you can't* occurs 47 cpm, while the 100th in the UK LJ i.e. *interpretation of the* occurs 90 cpm and the US LJ i.e. *to engage in* occurs 71 cpm.

All these facts indicate the general tendency that limited types of 3-grams are used more intensively in legal discourse than in general discourse.

5. How can we use n-grams to compile our legal English dictionary?

5. 1. Criteria to select sample n-grams from the legal discourse

There are some noteworthy arguments concerning how to select useful n-grams for pedagogical purposes. Biber (2006) chose n-gram sequences based solely on the frequency. His criterion is simple and straightforward but, as Simpson-Vlach & Ellis (2010) criticized, the problem is there are so many n-grams combining grammatical words such as *to do with the*. Martinez and Schmitt (2012) claim that frequency is not the only criterion, and introduce a new criterion, *compositionality*, which focuses on how much individual words in the sequence contribute to decoding the entire meaning of the multi-word sequences. They explain that *at all times* is more compositional than *at all* because we can guess the whole meaning of *at all times* more easily from the component words than *at all*. They claim:

> We therefore ended up with selection criteria that revolved around high

> frequency, meaningfulness, and relative non-compositionality. (p. 304)

Simpson-Vlach and Ellis (2010) compared the top 10 3-grams and the bottom 10 3-grams chosen by high frequency n-gram metric and high MI n-gram metric. They say:

> Ideally, though, we wanted to combine the information provided by *both* metrics to better approximate our intuitions and those of instructors, and thus to rank the academic formulas for use in pedagogical applications. (p495)

Frequency is definitely one significant factor, but there are some other factors such as usefulness and unitiveness. In the next section I will choose sample multi-word sequences for our project based on frequency, meaningfulness, and my own ESP instructor's intuition and experience.

5. 2. How to select the sample n-grams for our project

In order to select sample n-grams for our project, we first need to separate multi-word sequences for legal use from others. I grouped the top 40 3-grams listed in the UK LJ into four categories as shown below:

A. Proper noun 3-grams
the Court of, *Court of Appeal*, *the Supreme Court*, *House of Lords*

B. Legal use 3-grams
in relation to, *the right to*, *the context of*, *the scope of*, *the law of*, *of the law*, *the application of*, *the common law*, *in the context*, *in respect of*, *the absence of*, *the principle of*, *a matter of*

C. General use 3-grams
in order to, *the fact that*, *as well as*, *part of the*, *on the basis*, *the basis of*, *a number of*, *the use of*, *in terms of*, *nature of the*, *the nature of*, *in case of*, *in the case*

D. Grammatical use 3-grams
can not be, *there is a*, *that it is*, *one of the*, *there is no*, *in which the*, *it is not*, *such as the*, *is that the*, *that there is*

Proper noun 3-grams and grammatical use 3-grams are easy to identify. On the other hand, the distinction between legal use and general use 3-grams is somewhat problematic. I mechanically put in the group of general use 3-grams 1) the UK LJ 3-grams which are also found in the BNC 3-gram list, and 2) the UK LJ 3-grams which are expanded and used as the 4-grams in the BNC.

More problematic is the 3-grams in the legal use. Are they solely and exclusively used in legal discourse and not in other discourse? The frequent 3-grams of *in relation to* (298 cpm; ranked 4th), *the right to* (280 cpm; ranked 5th), (*in*) *the context* (*of*) (237+188 cpm; ranked 11th & 26th), *the scope of* (210 cpm; ranked 17th), and *the absence of* (161 cpm; ranked 34th) look typical of legal use. The following are the examples of *in relation to* from the legal corpus:

> Parliament had enacted legislation in relation to a banking activity,(UK LJ)
> The same problem was also addressed in relation to other estates in land before 1925, (UK LJ)

Although the occurrence count is low, this 3-gram sequence is used 45 cpm and ranked 110th in the BNC. The following are the examples:

> These are all examined in relation to the six elaborate mosaics listed above. (BNC)
> Couple of points that were made in relation to this particular report was the backing by the employee side, (BNC)

LDOCE and COBUILD define *in relation to* with an example use as follows:

relation **S2** **W1**
2 **in relation to sth** *formal*
b) *formal* concerning: *latest developments in relation to the disease*

relation ◆◆◇
7 PHRASE
If something is said or done in relation to a subject, it is said or done in connection with that subject. …*a question which has been asked many times in relation to Irish affairs.*

Another example of is (*in*) *the context* (*of*). This n-gram sequence is quite popular in the top 40 3-gram, 4-gram and 5-gram lists of the legal corpora as we have examined. However, no general and legal dictionaries I have consulted so far label (*in*) *the context* (*of*) as legal use or technical use. The example use of (*in*) *the context* (*of*) listed in the dictionaries all look formal but this does not mean these n-grams are exclusively for legal use. The following four example uses are from the legal corpus, the BNC and general English dictionaries, but they all look alike in terms of formality, style, lexical level and structural complexity, and give us an impression that they all belong to the same register or genre.

> Defining objectiveness in the context of the duty to act in good faith in the interests of the company is more complex (UK LJ)

> When considered in the context of levels of affinity obtaining between other mosaics in Britain, (BNC)

> These incidents are best understood in the broader context of developments in rural society. (LDOCE)

> We are doing this work in the context of reforms in the economic, social and cultural spheres. (COBUILD)

The following examples contain seemingly very legal terms, *law* and *right*. Thus, we may think they are typical of legal use.

> The focal point of this book is on the law of commercial contracts as constructed by the American and UK legal systems.(UK LJ)

> the bank had the right to have the account falsified, (UK LJ)

However, when we consult general English dictionaries, we find these two words are commonly used in our daily lives. LDOCE labels both *law* and *right* S1 and W1, and COBUILD gives three diamonds (◆◆◆) indicating they are most frequent words.

All those discussed so far seem to reveal the interesting nature of legal discourse. As the word “legalese” indicates, we often have an impression that legal discourse is full of jargon incomprehensive to lay persons. However, the above discussion strongly implies that legal discourse is not necessarily full of jargon. It is the level of formality, the conventional writing style and the preference to formal words and expressions that make the legal discourse look unfamiliar and unfriendly to us. In the next section, as the conclusion of the article, I will focus on one verb and its nominalized form and show how they are used as multi-word expressions in a conventional manner in legal discourse.

6. Conclusion—sample entry words: *apply* and *application*

The following passage is quoted from the UK LJ, typical legal discourse in our legal corpus. The thick underlined sequences are proper names of the law and the double underlined sequences are legal technical terms. The thin underlined parts are the n-grams of general or grammatical sequences. I shaded the proper name n-grams and the legal use n-grams.

In applying the law, special law has priority over more general law. Thus, the order of application of the law to a marine insurance contract is, first, the provisions of the Commercial Code on marine insurance, secondly, the Insurance Act 2008, and last, the general contract law contained in the Civil Code.

Even if we do not have a specialized legal knowledge we can understand the outline of the above passage. This is because although the formality is high and the style is professional the passage itself is written in the same ordinary English we encounter in our life.

In the passage below I erased all the general and grammatical n-grams. Those that remain are the technical terms and the names of the law which legal dictionaries such as Tanaka *et al*. (1991) and Garner *et al*. (1999) define and explain. If you are a legal professional, these are the terms your eyes immediately focus on when you read the passage. You can understand the main idea of what the passage is talking about only by picking up these terms.

, special law general law. ,
a marine insurance contract the
Commercial Code marine insurance, , the Insurance Act 2008, , the
general contract law the Civil Code.

The passage below shows the general and grammatical n-grams only. Unlike the above passage, the exact topic is not clear. However, you may be able to tell how the point of the argument is introduced and developed, and how the topic-related key words are presented in an orderly way in the paragraph. These n-grams organize the structure of the paragraph, on which the topic is discussed.

In applying the law, has priority over more . Thus,
the order of application of the law to is, first, the

provisions of the Commercial Code on , secondly, , and last, contained in .

When Japanese students of law produce legal discourse, the difficulties they have are not necessarily the proper names such as *the Commercial Code*, *the Insurance Act 2008*, or *the Civil Code*, or specialized legal terms such as *special law*, *general law*, *a marine insurance contract*, *marine insurance* or *the general contract law*. Much more problematic to them is how to construct their legal argument by using these proper names and legal terms. In other words, their realistic problem is that they do not know enough appropriate expressions conventionally used to introduce the legal proper names and technical terms in legal discourse. The purpose of our project is to provide the Japanese students of law with conventionally appropriate and frequently-used general and grammatical n-gram type sequences to help them express their thoughts and ideas in legal English.

In the above sample quotation, the topic is introduced into the paragraph by using the 2-gram sequence consisting of a preposition and the non-finite verb form, i.e. *in applying* (*the law*), then the verb *apply* is nominalized in the second sentence and used in the 2-gram word sequence, i.e. *application of* (*the law*). This sequential order of a verb → its nominalization is one of the common ways to introduce and develop the topic in the formal discourse (Biber, 2006).

I will use *apply* and *application* and illustrate how these words should be explained for the Japanese students of law in our dictionary. (All the numbers hereafter are crude occurrence counts per 6 million.)

APPLY (6099 times; ranked 96th in the UK LJ)
subjects of APPLY
court (151), rule (77), law (70), principle (48), provision (27), act (19), judge (18), test (18), convention (16), consideration (15), charter (14)

> First, the court **applied** its reasoning in BMO to the argument that provincial legislation did not apply.
>
> The grandchildren therefore, in such a situation, have an insurable interest in their grandparents. The rule **applies** vice versa. This interest can be extended to other members of the family such as siblings.
>
> The court held that Japanese law **applies** to the matters relating to the validity of the contract and legality of the voyage,

objects of APPLY

law (229), test (151), rule (146), principle (131), standard (62), provision (48), approach (36), doctrine (31), criterion (24), statute (21), convention (21), reasoning (19),

On this last issue, Lord Clarke **applied** English law to the deceit claim

The Adjudicator **applied** a similar balancing test to that in Best,

The simplest way to do this would be to **apply** the Chapter 58A rules to all applications for PEOs.

The Court straightforwardly **applied** ordinary accounting principles to require Hall to pay 37,054.69 to the plaintiff.

prepositions taken by APPLY

apply to (1,099), apply in (417), apply for (118), apply by (69), apply with (24), apply at (17), apply as (11)

The Insurance Act **applies** to all kinds of insurance contracts, whether the contract is called insurance, a co-operative agreement known as Kyosai, or others.

Admittedly, this would only **apply** in highly exceptional cases

the respondent might also be able to **apply** for an order restricting its liability

APPLICATION (3730 times; ranked 87th in the UK LJ)

propositions taken after *application*

application of (1,572), application for (125), application to (68), application in (33), application under (11)

APPLICATION of

law (256), rule (159), principle (152), art (56), test (54), article (45), provision (44), standard (35), doctrine (37), convention (34)

Professor Andrea Lista examines the **application** of EU competition law in the financial services industry,

A possible model paradigm for the **application** of competition law to the banking sector

The non-market aims include the **application** of the patent rules to safeguard human dignity and integrity

APPLICATION for

the Court upheld **applications** for judicial review brought by asylum seekers

he therefore dismissed the **application** for permission to appeal.

verbs with APPLICATION as the object 729

make (60), reject (30), consider (28), limit (25), bring (23) [application brought by], ensure (21), refuse (21), justify (16), dismiss (14), preclude (12), lodge (13), examine (11), trigger (11), permit (11),

fourteen mayors made an **application** to the European Court of Human Rights.

Despite this, an **application** was made to the Land Registry to close the leasehold titles on the basis of a letter from the bailiffs as to the date of re-entry.

A State may make an **application** for necessary measures to be taken in respect of the protection of its servants or agents

The Court rejected the **applications** for an anti-suit injunction and damages, on the basis that,

Bell considers the **application** of the principle to the situation of an unexpectedly re-appearing child:

It found that by limiting the **application** of the Article 4(1)(b) exception to situations

So held the CJEU in an **application** brought by various Dutch nationals concerning the refusal to issue them with a passport

References

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

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

Burchfield, R. *et al*. (Eds.) (1989). *The Oxford English dictionary*. Oxford: Oxford University Press.

Firth, J. R. (1957). Modes of meaning. *Papers of Linguistics 1934-51* (pp190-215). Oxford, Oxford University Press.

Garner, B. A. et al. (Eds). (1999). *Black's law dictionary*. St. Paul: West Group.

Hori, M. (2009). *Introduction to collocation studies in English*. Tokyo: Kenkyusha.

Hornby, A. S. *et al*. (Eds.) (2015) *Oxford advanced learner's dictionary*. Oxford: Oxford University Press.

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

Inoue, Y. & Akano, I. (Eds.) (2013). *The wisdom English-Japanese dictionary*. (3rd ed). Tokyo: Sanseido.

小山由紀江 (2008)「Multi-word Expressionに関する統計と教育への応用」『統計数理研究所共同研究リポート』NO.216 統計数理研究所 pp39-56

小山由紀江 (2009)「科学技術コーパスにおける特徴的 Multi-word Expressionの抽出とその評価」『統計数理研究所共同研究リポート』NO.233 統計数理研究所 pp51–68

Leech, G. (2011). Frequency, corpora and language learning. In F. Meunier, S. Cook, G. Gilquin & M. Paquot (Eds.), *A Taste for Corpora: In Honour of Sylviane Granger* (pp. 7-31). Amsterdam – Philadelphia, PA: John Benjamings.

Martinez, R. and Schmitt, N (2012). A Phrasal Expressions List. *Applied Linguistics, 33* (3), 299-320.

Mayor, M. et al. (Eds.) (2014). *Longman dictionary of contemporary English*. Harlow, Essex: Pearson.

Minamide, K. *et al*. (Eds.) (2014). *Genius English-Japanese Dictionary*. Tokyo: Kenkyusha.

Oxford University Press. (2009). *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Pearson. (2013). *Longman Collocations Dictionary and Thesaurus*. Harlow, Essex: Pearson

Shannon, C. E. (1964). The mathematical theory of communication. In C. E. Shannon & W. Weaver (Eds.), *The mathematical theory of communication* (pp. 29-125). Urbana: The University of Illiois.

Simpson-Vlach, R. & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4), 487-512.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. et al. (Eds.) (2014). *Collins COBUILD advanced learner's dictionary*. Glasgow: HarperCollins.

Stubbs, M. (2007). An example of frequent English phraseology: distributions, structures and functions. In R. Facchinetti (Ed.), *Corpus Linguistics 25 Years on* (pp. 89-107). Amsterdam-NY: Rodopi.

Tanaka, H. et al. (Eds.) (1991). *Dictionary of Anglo-American law*. Tokyo: Tokyo University Press.

Teubert, W. (2004). Units of meaning, parallel corpora, and their implications for language teaching. In U. Connor & T. A. Upton (Eds.), *Applied Linguistics: A Multidimensional Perspective* (171-189). Amsterdam: Rodopi.

山田崇仁 (2007)「Ngram方式を利用した漢字文献の分析」『立命館白川静記念東洋文字文化研究所紀要』第一号 立命館大学 pp.1-23

1-gram					
word	BNC	word	UK LJ	word	US LJ
the	54157.07	the	73908.83	the	61726.33
of	30274.41	of	45660.00	of	36665.67
to	25669.78	to	31504.33	to	29894.83
and	25121.71	in	21430.50	and	21316.50
a	20418.49	and	21400.67	a	19753.83
in	17877.23	a	20682.17	in	18155.67
that	10623.85	that	17150.67	that	17567.83
is	9729.21	is	14869.50	is	11550.83
was	8778.60	be	9220.33	for	8521.17
I	8618.25	as	8458.67	as	7360.00
for	8318.05	for	8192.50	not	6965.83
it	8197.97	not	7635.00	be	6697.50
on	6950.02	by	6878.33	on	6159.50
be	6485.75	on	6806.33	or	6071.50
with	6404.20	it	6715.83	are	5563.17
The	6195.11	The	6425.00	The	5505.67
as	6035.79	or	5552.00	by	5381.83
you	5747.82	law	5388.00	it	5267.83
at	4872.49	with	5232.33	with	5010.00
by	4867.25	was	5221.67	have	4337.67
he	4625.23	an	5048.17	an	4199.67
are	4550.86	this	4937.17	from	3809.83
have	4533.05	which	4892.17	this	3707.00
not	4325.22	are	4775.50	law	3612.33
had	4185.94	have	3876.33	would	3326.33
from	4102.06	from	3445.17	their	3067.17
his	3821.26	has	3275.17	more	2930.33
which	3612.03	would	2906.17	which	2884.17
or	3604.70	In	2819.67	can	2834.17
this	3435.63	at	2770.50	In	2834.17
they	3365.49	its	2635.67	was	2763.50
but	3215.13	can	2559.67	has	2757.00
an	3200.93	been	2425.17	they	2726.50
n’t	3164.73	such	2390.83	at	2635.167
were	3120.34	their	2364.67	may	2560.50
her	2895.03	legal	2342.67	its	2302.33
we	2651.80	but	2285.17	than	2236.83
been	2599.46	may	2151.00	other	2223.00
one	2585.03	This	2125.17	but	2220.00
has	2543.60	will	2047.83	will	2164.50
	345480.57		390406.50		347267.67

2-grams					
word	BNC	word	UK LJ	word	US LJ
of the	7531.95	of the	12561.17	of the	7719.50
in the	4801.92	in the	5361.67	in the	4394.67
to the	2869.59	to the	4810.00	to the	3441.17
on the	2075.57	that the	3695.67	that the	2821.17
and the	1882.28	on the	2666.17	on the	1928.17
to be	1878.62	to be	2347.00	and the	1773.67
for the	1595.80	by the	2254.83	of a	1573.00
at the	1380.48	and the	2245.17	for the	1325.00
that the	1271.84	of a	2054.50	to be	1308.00
by the	1253.65	for the	1887.67	as a	1283.00
with the	1241.54	it is	1756.33	by the	1195.33
of a	1240.20	with the	1643.00	with the	1185.50
from the	1203.70	as a	1476.50	it is	1152.67
in a	1064.86	is not	1165.33	in a	1078.83
it is	909.29	from the	1137.83	from the	1006.33
it was	864.02	in a	1086.33	is not	931.17
as a	817.58	is a	1055.50	the Court	894.00
do n't	815.16	to a	934.33	is a	797.83
is a	777.14	the Court	928.67	to a	776.50
with a	757.64	as the	890.83	as the	742.17
have been	701.37	does not	875.67	does not	732.00
will be	696.05	can be	834.83	the same	728.17
for a	688.58	the law	823.50	in which	709.67
was a	649.63	It is	797.00	would be	705.50
had been	641.76	has been	796.67	may be	689.67
is the	610.00	there is	775.83	at the	684.00
to a	584.05	that it	775.83	such as	648.83
has been	575.56	is the	774.50	can be	630.17
as the	563.94	have been	758.33	Supreme Court	625.33
the same	558.64	would be	753.67	do not	616.17
and a	551.88	not be	752.50	is the	598.17
one of	551.51	should be	748.83	about the	581.67
would be	546.49	at the	727.67	is that	564.17
can be	540.93	such as	726.17	there is	549.50
he was	533.11	may be	712.17	United States	544.00
into the	528.12	the same	657.67	should be	519.67
It is	517.01	the case	639.00	the law	518.17
the first	502.91	in which	634.00	rather than	505.17
It was	494.35	of this	628.83	the United	504.33
I do	476.94	is that	628.50	not be	501.50
	40213.71		54218.50		49483.50

3-grams			
word	BNC	word	UK LJ
I do n't	371.47	the Court of	333.00
one of the	297.80	in order to	327.67
the end of	207.19	the fact that	313.67
as well as	168.46	in relation to	297.50
part of the	166.99	the right to	280.33
do n't know	154.06	as well as	272.17
out of the	152.38	Court of Appeal	265.50
a number of	137.90	part of the	263.83
a lot of	136.64	on the basis	262.50
end of the	133.97	can not be	243.33
be able to	133.82	the context of	237.33
some of the	128.85	the basis of	221.00
to be a	116.97	there is a	215.17
the fact that	113.74	that it is	215.17
per cent of	113.35	the Supreme Court	214.33
there is a	104.77	a number of	211.33
I did n't	103.53	the scope of	210.33
in order to	102.22	the law of	209.00
I ca n't	99.86	there is no	206.00
in terms of	93.56	the use of	202.33
at the end	89.64	of the law	200.67
there was a	86.87	the application of	198.67
the number of	85.91	the common law	197.83
you do n't	82.22	in which the	194.67
that it is	81.44	one of the	193.17
the rest of	80.29	in the context	187.83
it would be	80.29	it is not	177.67
the use of	79.79	in terms of	177.00
do n't think	78.65	nature of the	173.50
that it was	77.48	in respect of	169.17
there is no	77.23	such as the	168.83
have to be	75.86	is that the	165.67
the same time	75.64	the nature of	161.50
the first time	74.19	the absence of	160.83
members of the	73.98	the case of	160.17
can not be	72.09	in the case	159.00
at the time	70.38	that there is	154.50
would have been	69.74	the principle of	147.83
to be the	68.99	House of Lords	143.67
it was a	68.41	a matter of	143.50
	4586.62		8437.17

word	US LJ
the United States	442.33
the Supreme Court	352.33
as well as	236.83
in which the	187.00
in order to	183.17
the use of	180.50
the federal government	179.67
in the United	170.00
the fact that	168.00
with respect to	165.67
one of the	157.83
the context of	149.17
there is no	144.17
part of the	144.17
can not be	142.33
it is not	140.83
the value of	137.17
more likely to	132.00
is that the	131.00
a number of	130.17
the scope of	129.33
in favor of	125.00
some of the	124.83
there is a	124.00
As a result	122.83
the number of	122.17
the absence of	118.50
such as the	118.17
that it is	117.33
be able to	117.33
likely to be	116.83
a matter of	111.50
in the context	109.83
the Court has	106.17
the common law	105.50
in terms of	104.83
the same time	102.83
of the law	102.17
to do so	99.00
the other hand	98.67
	5951.17

4-grams			
word	BNC	word	UK LJ
I do n't know	119.04	in the context of	185.83
the end of the	103.74	the Court of Appeal	180.33
at the end of	78.44	on the basis of	171.67
I do n't think	69.85	the House of Lords	111.50
at the same time	48.12	in the case of	104.33
the rest of the	47.12	the fact that the	103.50
for the first time	47.12	as a result of	102.50
per cent of the	45.22	the rule of law	93.33
as a result of	44.68	the Court of Justice	92.67
one of the most	32.86	the nature of the	92.33
is one of the	32.67	in relation to the	88.50
do n't want to	32.66	on the basis that	87.67
in the case of	32.45	the scope of the	86.67
I do n't want	32.39	the extent to which	77.67
to be able to	31.67	as a matter of	76.50
the Secretary of State	30.59	the application of the	75.17
On the other hand	28.36	in the absence of	74.50
in the form of	27.57	On the other hand	73.17
on the basis of	27.43	for the purposes of	72.83
the top of the	26.73	the context of the	70.67
in the middle of	26.41	as well as the	70.17
do n't know what	25.42	at the time of	70.00
by the end of	25.12	in the light of	64.50
as well as the	25.08	on the part of	64.00
on the other hand	24.60	the basis of the	58.50
the way in which	24.26	of the Court of	55.00
a member of the	24.15	the way in which	53.50
was one of the	23.29	of the right to	52.83
at the time of	22.88	the European Court of	52.67
the middle of the	22.19	the interests of the	52.33
a great deal of	22.05	for the purpose of	51.83
will be able to	21.81	that there is a	49.33
a wide range of	21.72	in accordance with the	48.83
the fact that the	21.48	in the form of	48.17
At the same time	21.08	the role of the	47.67
the back of the	20.73	European Court of Human	47.00
the nature of the	20.35	on the other hand	46.83
Secretary of State for	19.00	Court of Human Rights	46.50
in terms of the	18.84	in light of the	45.50
at the beginning of	18.39	of the common law	45.33
	1387.56		3091.83

word	US LJ
in the United States	158.50
in the context of	107.83
of the United States	74.83
THIS POINT IS NOT	74.17
TABULAR OR GRAPHIC MATERIAL	74.17
SET FORTH AT THIS	74.17
POINT IS NOT DISPLAYABLE	74.17
OR GRAPHIC MATERIAL SET	74.17
MATERIAL SET FORTH AT	74.17
GRAPHIC MATERIAL SET FORTH	74.17
FORTH AT THIS POINT	74.17
AT THIS POINT IS	74.17
on the basis of	68.50
At the same time	68.33
as a matter of	66.17
On the other hand	64.17
in the absence of	62.50
in a way that	60.00
the Supreme Court has	59.67
in the form of	55.83
the extent to which	55.50
as a result of	54.33
as well as the	50.67
in the first place	49.67
in the case of	48.00
are more likely to	46.67
at the time of	46.50
the nature of the	45.00
To the extent that	44.67
the value of the	44.17
in the face of	44.17
to the extent that	42.83
the fact that the	42.33
in light of the	40.83
with respect to the	37.67
the scope of the	37.50
the criminal justice system	36.17
at the expense of	34.50
on the other hand	34.33
more likely to be	31.17
	2380.5

5-grams			
word	BNC	word	UK LJ
at the end of the	37.90	in the context of the	51.50
I do n't know what	18.01	European Court of Human Rights	43.67
I do n't want to	17.97	the European Court of Human	42.50
in the middle of the	14.14	on the part of the	41.67
as a result of the	13.72	on the basis of the	38.17
by the end of the	13.28	as a result of the	37.00
the other side of the	12.08	in the light of the	34.00
the Secretary of State for	12.08	v Secretary of State for	33.17
at the time of the	10.23	at the time of the	32.50
I do n't think I	9.96	European Convention on Human Rights	32.50
the end of the year	9.35	the Court of Appeal in	31.17
at the top of the	9.35	the European Convention on Human	29.67
for the first time in	8.66	of the rule of law	27.17
I do n't know how	8.48	on the basis that the	26.50
the end of the day	8.45	of the House of Lords	26.50
I do n't think it	8.18	the House of Lords in	24.67
on the part of the	8.14	to negotiate in good faith	24.50
at the beginning of the	7.89	the extent to which the	24.50
At the end of the	7.62	of the Court of Appeal	23.50
on the other side of	7.46	by the Court of Appeal	23.17
I do n't know whether	7.39	of the Court of Justice	21.50
I do n't know if	7.27	Secretary of State for the	21.33
is one of the most	6.66	the best interests of the	20.67
I do n't know why	6.55	in the interests of the	20.33
in the same way as	6.49	within the scope of the	20.00
in the form of a	6.48	the interests of the company	20.00
at the bottom of the	6.44	of State for the Home	19.50
the way in which the	6.37	Court of Justice of the	19.33
I do n't think so	6.35	on the basis of a	18.17
hon. Friend the Member for	6.22	for the purposes of the	17.67
on the edge of the	6.11	the Court of Justice of	17.17
ask the Secretary of State	6.09	the way in which the	17.00
for the rest of the	6.08	State for the Home Department	17.00
at the back of the	6.03	for the benefit of the	15.50
in the case of the	5.99	of the European Court of	14.17
you do n't have to	5.96	in the same way as	14.00
To ask the Secretary of	5.95	in the case of a	14.00
in the light of the	5.64	in the Court of Appeal	14.00
the other end of the	5.49	in the context of a	13.83
at the same time as	5.38	in the case of the	13.67
	367.89		996.83

word	US LJ
THIS POINT IS NOT DISPLAYABLE	74.17
TABULAR OR GRAPHIC MATERIAL SET	74.17
SET FORTH AT THIS POINT	74.17
OR GRAPHIC MATERIAL SET FORTH	74.17
MATERIAL SET FORTH AT THIS	74.17
GRAPHIC MATERIAL SET FORTH AT	74.17
FORTH AT THIS POINT IS	74.17
AT THIS POINT IS NOT	74.17
at the time of the	18.50
is not to say that	18.00
the Necessary and Proper Clause	16.00
This is not to say	15.83
separation of funds and managers	15.17
the Bressman and Gluck study	14.83
as a result of the	14.83
the costs and benefits of	13.67
there is no reason to	13.50
the extent to which the	13.17
the separation of funds and	12.67
on the part of the	12.50
are more likely to be	12.00
in the absence of a	11.67
Court of Appeals for the	11.50
in the form of a	11.33
even in the absence of	11.33
in the context of the	11.17
To the extent that the	11.17
in the wake of the	11.00
of the law of nations	10.33
the scope of this Article	10.17
at the end of the	10.00
to the extent that the	9.83
at the expense of the	9.83
in the Bressman and Gluck	9.67
beyond the scope of this	9.67
it is not clear that	9.50
in the United States and	9.50
in a way that is	9.50
in the context of a	9.33
on the ground that the	8.83
	979.33

6-grams			
word	BNC	word	UK LJ
on the other side of the	6.21	the European Court of Human Rights	40.00
at the end of the day	6.03	the European Convention on Human Rights	28.83
ask the Secretary of State for	6.01	Secretary of State for the Home	19.50
To ask the Secretary of State	5.95	v Secretary of State for the	19.00
mm mm mm mm mm mm	4.39	of State for the Home Department	17.00
from the point of view of	4.03	the Court of Justice of the	16.83
by the end of the year	3.97	in the interests of the company	13.50
my hon. Friend the Member for	3.63	of the European Convention on Human	13.33
in such a way as to	3.46	Court of Justice of the European	12.83
in the middle of the night	2.64	of Justice of the European Union	12.00
the Department of Trade and Industry	2.57	of the European Court of Human	11.67
in the second half of the	2.47	of the House of Lords in	10.67
at the other end of the	2.41	the entry into force of the	10.33
Secretary of State for the Environment	2.37	the best interests of the child	10.33
I do n't know what you	2.35	the right to a fair trial	9.50
I do n't want to be	2.19	the object and purpose of the	9.33
The hundred shares index closed down	2.18	in such a way as to	9.17
the Secretary of State for the	2.17	the Court of Appeal held that	8.83
if he will make a statement	2.13	decision of the Court of Appeal	8.83
The hundred shares index closed up	2.11	by the Court of Appeal in	8.83
The pound is up at one	2.06	of the Court of Appeal in	8.50
pound is up at one dollar	2.05	in the best interests of the	8.33
at the end of the year	2.05	from the point of view of	8.00
This is not to say that	1.98	decision of the House of Lords	7.67
in the first half of the	1.92	it is difficult to see how	6.83
pound is down at one dollar	1.91	Vienna Convention on the Law of	6.83
The pound is down at one	1.91	to act in good faith in	6.67
ask the Prime Minister if he	1.80	in good faith in the interests	6.67
the point of view of the	1.79	good faith in the interests of	6.67
the Prime Minister if he will	1.79	faith in the interests of the	6.67
To ask the Prime Minister if	1.79	act in good faith in the	6.50
Still to come on Central News	1.77	Court of Justice of the EU	6.50
by the end of the century	1.76	Convention on the Law of Treaties	6.50
if he will list his official	1.75	the EU Charter of Fundamental Rights	6.33
Prime Minister if he will list	1.75	on the Rights of the Child	6.33
Minister if he will list his	1.75	by the European Court of Human	6.00
will list his official engagements for	1.74	is beyond the scope of this	5.83
he will list his official engagements	1.74	duty to act in good faith	5.83
at the turn of the century	1.73	Court of Appeal held that the	5.83
I do n't want to go	1.71	the decision of the Court of	5.67
	106.02		424.50

word	US LJ
TABULAR OR GRAPHIC MATERIAL SET FORTH	74.17
SET FORTH AT THIS POINT IS	74.17
OR GRAPHIC MATERIAL SET FORTH AT	74.17
MATERIAL SET FORTH AT THIS POINT	74.17
GRAPHIC MATERIAL SET FORTH AT THIS	74.17
FORTH AT THIS POINT IS NOT	74.17
AT THIS POINT IS NOT DISPLAYABLE	74.17
This is not to say that	12.83
the separation of funds and managers	12.67
in the Bressman and Gluck study	9.67
beyond the scope of this Article	8.17
U.S. Court of Appeals for the	7.83
the U.S. Court of Appeals for	7.17
is beyond the scope of this	6.67
the Federal Rules of Civil Procedure	6.50
shared presuppositions of speakers and listeners	6.50
THIS POINT IS NOT DISPLAYABLE The	6.50
the shared presuppositions of speakers and	6.17
by the shared presuppositions of speakers	6.00
framed by the shared presuppositions of	5.83
as framed by the shared presuppositions	5.83
meaning as framed by the shared	5.50
THIS POINT IS NOT DISPLAYABLE Figure	5.33
contextual meaning as framed by the	5.17
referents for claims of legal meaning	5.00
on the basis of sexual orientation	5.00
is no reason to believe that	4.33
THIS POINT IS NOT DISPLAYABLE Source	4.00
the congressional respondents in the Bressman	3.67
states and the District of Columbia	3.67
respondents in the Bressman and Gluck	3.67
of the Civil Rights Act of	3.67
congressional respondents in the Bressman and	3.67
of the Federal Rules of Civil	3.50
is not to say that the	3.50
among applications or classes of applications	3.50
violation of the law of nations	3.33
there is no reason to believe	3.33
the provision of Quality of Service	3.33
the causes of action available in	3.33
	704.00