You shall know a term by the company it keeps: Collocations of the term *evidence* in general and legal corpora

TOMASZ MICHTA

Abstract

Despite the progress made in the study of collocations, their use in specialised languages by and large continues to be underresearched. The present article attempts to go some way towards filling this gap by looking at variation in collocations of a single term (*evidence*) as extracted from a general corpus and a legal one, and by exploring the implications of such variation for the retrieval of legal collocations. In particular, the study looks at a) the overrepresentation of collocations in the legal corpus, b) the underrepresentation of collocations in the legal corpus, and c) the potential of both corpora for collocation retrieval. The findings suggest that there are striking differences between the use of collocations in each corpus and that such differences can radically affect the lists of collocations obtained from each corpus.

Keywords

collocation, corpus, general language, legal language, specialised language, terminology

Specyfika łączliwości wyrazowej terminu: Kolokacje terminu *evidence* w korpusie ogólnym oraz prawniczym

Abstrakt

Pomimo postępu, jaki dokonał się w badaniach nad kolokacjami, ich użycie w językach specjalistycznych jest w znacznej mierze nadal niewystarczająco zbadane. W niniejszym artykule podjęto próbę częściowego wypełnienia stwierdzonej luki w badaniach poprzez analizę zróżnicowania kolokacji pojedynczego terminu (*evidence*) pozyskanych z korpusu języka ogólnego oraz języka prawniczego. W szczególności w badaniu zanalizowano: a) zwiększony udział niektórych kolokacji w korpusie prawniczym, b) zmniejszony udział niektórych kolokacji w korpusie prawniczym, c) potencjał obu korpusów jako źródła do pozyskiwania kolokacji. Uzyskane wyniki wskazują na istnienie zdecydowanych różnic pomiędzy użyciem kolokacji w obu korpusach, co w sposób radykalny wpływa na listy kolokacji otrzymane na ich podstawie.

Słowa kluczowe

język ogólny, język prawa, język specjalistyczny, kolokacja, korpus, terminologia

1. Introduction

Much of today's research into collocations owes a debt to the work of Firth, who is often recognised as the father of collocation. Not only did he propose *collocation* as a linguistic term, but he also succeeded in drawing the attention of numerous scholars to the habitual company that words keep. His often-quoted sentence "You shall know a word by the company it keeps" (1968:

179) captures the essence of his approach: collocations are seen as being unique to individual words and as a key aspect of a word's meaning. Following in his footsteps, many linguists have investigated the phenomenon of collocations by adopting a variety of approaches, pursuing a range of different purposes, and producing a body of literature that is impressively vast.

The motivation behind exploring collocations is at least two-fold. First, they are of theoretical interest and can be invoked to explain the fact that certain words tend to habitually occur in the company of other words (see e.g. Firth 1957, Sinclair 1991, Hoey 2005). Seen from this perspective, collocations appear to be a perfect illustration of the non-random nature of language (Kilgarriff 2005). Consequently, they need to be studied as an important subject in and of themselves, but also one that contributes to our understanding of how language works in general. Second, and related to the first, the centrality of collocations in language and the concomitant challenges that they pose for speakers act as a driving force behind undertaking more practically-oriented investigations that aim to explore the nature of the problem and contribute to overcoming it (for an overview see Boers and Webb 2017).

The challenges related to the use of collocations are by no means limited to the context of general language. Saber et al. (2020: 106) have found that problems with collocations account for the largest proportion of errors made in scientific abstracts written by low-proficiency doctoral students. Their findings tally with opinions about the difficulty of using collocations expressed by a group of professional and non-professional translators (i.e. physicians) of medical texts in a study by Badziński (2019: 167-168) who, with the use of a questionnaire, established that it was collocations that his respondents considered to be the most challenging aspect of medical translation. Some scholars also point out that, outside general language tasks, native speaker competence may be of limited use as far as collocations are concerned (Benson 1989: 4, Frankenberg-Garcia 2018). Citing her own experience as a legal translator trainer, Giczela-Pastwa (2021: 191) opines that when "starting a course

in legal translation, students are usually not familiar with legal discourse in their native tongue". This lack of familiarity with legal discourse encompasses legal collocations.

Given the importance of collocations in specialised language and the problems related to their use, it is perhaps regrettable that there is a clear imbalance in the attention paid by scholars to collocations in general language (i.e. language for general purposes, or LGP) and those in specialised language (i.e. language for special purposes, or LSP). In contrast to the wealth of insights produced by studies into collocations in LGP, until relatively recently there had been few studies into collocations in LSP. Exceptions include: Picht 1987; Martin 1992; Meyer and Mackintosh 1994; Gledhill 2000; L'Homme 2000; Heid 2001; Michta 2007; and Ward 2007. In the particular case of legal English, research by Kjaer (1990a, 1990b), Goźdź-Roszkowski (2011) and Biel (2012, 2014) merits special attention. Commenting on the state of research into legal phraseology, Goźdź-Roszkowski and Pontranfoldo (2015: 130) note: "The legal domain and its phraseology has also received scant attention". However, although this statement was true when made, the situation has since started to improve.

Also regrettable is the fact that precious few specialised dictionaries actually contain word combinations (L'Homme and Leroyer 2009: 260), while those lexicographic resources that do deign to provide them often do so inconsistently (Montero-Martinez and Buendia-Castro 2012) and may prove wanting, especially as far as their treatment of verb + term combinations is concerned (Buendia-Castro Faber 2015). The problem of inadequate lexicographical treatment of specialised collocations is further compounded by the fact that there is only a handful of dictionaries and glossaries of specialised collocations.

2. Aims and scope

The present work is of an exploratory nature and seeks to make a contribution to the study of specialised (here: legal) collocations by looking at variation in collocations that feature a single

term between a general corpus and a legal one, and by exploring the implications of such variation for the retrieval of legal collocations. More specifically, the article sets itself a three-fold aim that consists in:

(a) establishing and exploring the most overrepresented collocations of the noun *evidence* in judgments handed down by the UK Supreme Court as opposed to general English;
(b) establishing and exploring the most underrepresented collocations of the noun *evidence* in judgments handed down by the UK Supreme Court as opposed to general English;
(c) investigating the potential of two different corpora (a legal one and a general one) for the retrieval of legal collocations, which is carried out with a view to including such collocations in lexicographical resources.

Thus this study represents another addition to the line of research presented in the recently published book chapter by L'Homme and Azoulay (2020), which compared sets of collocates for 15 lexical items retrieved from a) a specialised corpus on the theme of the environment and b) a general language corpus, and found marked differences between the sets retrieved from each corpus. Simultaneously, the article fits into what Biel (2010: 4) classifies as the first trajectory of research in corpus-based studies of legal language, a trajectory which focuses on external variation and investigates how "legal language differ[s] from general language and other languages for special purposes".

Several motivations have informed the choice of the term *evidence* for the present analysis. First, the word is reasonably frequent in both corpora, occurring in them a total of 8,224 (the legal corpus) and 20,336 (the general corpus) times. This means that even though the word can act as a legal term, it may be claimed to be part of general vocabulary. Second, its significance for the field of law is unquestionable and it was felt that priority should be given to a key legal term rather than one of marginal importance. Third, its sense structure is relatively straightforward. By way of example, the Oxford Learner's Dictionaries website only distinguishes between two senses of the

word: 1. "the facts, signs or objects that make you believe that something is true" and 2. "the information that is used in court to try to prove something" (OLD 2021). Clearly, both senses are related to each other. The second is the more legal one, although its connection to law is not indicated with a label, which in turn suggests that is not perceived as being particularly technical in general language contexts. In the legal corpus chosen for the purposes of this study, the technical meaning overwhelmingly dominates and it is captured by a definition that foregrounds its field-specific character: "[t]estimony and production of documents and things relating to the facts into which the court enquires and the methods and rules relating to the establishing of those facts before the court" (Richards and Curzon 2011: 181). Importantly, *evidence* does not acquire a range of new senses[1] in the legal corpus. The existence of a single dominant meaning thus prevented the analysis from being compounded by problems of polysemy, which legal English is notorious for (Matilla 2012: 30).

3. Data and methods

3.1. Data

For the purposes of this study, two language corpora have been selected: one general and one legal. The legal corpus comprises judgments handed down by the Supreme Court of the United Kingdom (UKSC), which acts as "the final court of appeal for all United Kingdom civil cases, and criminal cases from England, Wales and Northern Ireland" (UKSC 2021a). Originally compiled by the present author and Katarzyna Mroczyńska as part of a joint project aiming to provide a lexicographic description of collocations in the judgments of the UKSC (Mroczyńska 2020,

---

[1] One may of course split the meaning captured by the quoted definition into a few senses. *Black's Law Dictionary* (Garner 2004: 595), for example, lists 5 senses in its entry for *evidence*. The differences between them, albeit important whenever a high degree of precision is required, represent various focus points rather than wildly dissimilar concepts.

Michta and Mroczyńska 2022), the corpus (heireinafter: UKSCC) is intended as a reliable foundation for corpus studies of legal English. Featuring 636 actual judgments handed down by the UKSC, totalling 9.5 million tokens, and spanning a ten-year period from the Court's inception in 2009 to 2018, it represents a complete collection of UKSC judgments for the time frame mentioned. Arguably, its focus on a single text type and the rather small number of justices (12 at any one time (UKSC 2021b)), who contribute to the linguistic variety of the corpus, might be seen as a potential drawback. Yet, UKSC judgments are undoubtedly legal in nature and seem sufficiently suitable for the purpose of the study despite any linguistic peculiarities that they show compared to other legal English texts types.

As the general language corpus, the British National Corpus (BNC) was chosen to serve as a source of linguistic data. There are a number of arguments for making this particular choice. As its very name suggests, the corpus focuses on British English, thus matching the legal one in terms of the language variety it represents. Since collocations do vary between geographical language varieties (Mair 2007), ensuring this correspondence was seen as a priority. To match the legal corpus in terms of register (written), only the written subcorpus (hereinafter: WBNC) was selected for analysis. Another important consideration supporting the choice of the WBNC was its reasonably large size. Its number of tokens stands at 100.5 million (as calculated by Sketch Engine), which was considered sufficient given the study's objectives. The WBNC can also be described as balanced as it "contains texts from a wide range of different language genres and text domains" (Baker, Hardie, and McEnery 2006: 18), including a selection of academic and non-academic texts that are concerned with law. Last but not least, the WBNC is readily available online and can, for example, be downloaded free of charge or consulted via Sketch Engine. All the arguments cited so far weigh heavily in favour of using the WBNC, yet a certain drawback must also be mentioned. No texts were included in the WBNC after the release of the BNC in 1994. In fact, about 96 % of its content is comprised of texts from the period 1984-

1993.[2] This slightly diminishes the value of the WBNC as a reference point for the legal corpus, as the latter draws on more up-to-date texts. It was nevertheless decided that the benefits of relying on the WBNC outweigh this drawback, and any others, and it was therefore chosen as a reference corpus.

3.2. Methods

To achieve its aims, the study adopted an understanding of collocations that involves both frequency-based and syntactic criteria. In order for a word combination to potentially qualify as a collocation, the minimum frequency threshold in either of the corpora was set at 5. This rather liberal approach resulted in a rather long initial list of collocations, but all of the collocations that were subjected to closer inspection, and are discussed later in this paper, occurred more frequently in the corpus. As a safeguard against idiosyncratic collocations, another requirement was introduced so that a word combination had to appear in at least two texts in order to be considered for inclusion. Again, the approach adopted here as to what should count as a collocation contributed to the rather liberal character of the list of collocations. The final criterion was that constituents of a collocation, i.e. *evidence* (the node) and the collocate, had to be syntactically related. Only one such relation was analysed, i.e. modifier + noun. Several scholars (Bergenholtz and Tarp 1994; Michta et al. 2009: 93; L'Homme and Azoulay 2020: 154) have noted that focusing on this relation may also yield combinations that are terms, but distinguishing between collocations on the one hand, and terms on the other, was not the main focus of the study. Hence both terms and collocations were included in the analysis of collocates of *evidence*. Whenever the terminological status of certain combinations was deemed worthy of mention, the *Longman Dictionary of Law* (Richards and Curzon 2011) was consulted to establish whether a given combination typically acts

---

[2] The percentage was calculated using the BNC available via Sketch Engine.

as a term. This method, which involved taking recourse to legal resources, was felt to be a good heuristic vis-a-vis the fact that no universally agreed criteria exist that allow a clear-cut dividing line to be drawn between terms and collocations (see also Heid 2001: 791). To retrieve collocations, Sketch Engine was used since it offered all the necessary tools to ensure that all the requirements of collocations were met. In particular, the study relied heavily on Sketch Engine's word sketch functionality, which provides a condensed description of a word's grammatical and collocational behaviour (Kilgarriff et al. 2014: 9). The results produced by Sketch Engine were not accepted blindly. When a word was suggested as a potential collocate and it turned out not to act as a modifier in the corpus, the word was removed from further analysis.

In order to establish the most overrepresented collocations of the noun *evidence* in the UKSCC as opposed to WBNC, a list of key collocations was produced using the word sketch function, with the UKSCC acting as a focus corpus and the WBNC acting as a reference corpus. These key collocations were identified by Sketch Engine, which uses the simple maths method (Kilgarriff 2009). This approach calculates the keyness score according to the formula below (Lexical Computing 2015 : 3):

$$\frac{fmp(foc) + n}{fmp(ref) + n}$$

where fpm(foc) is the normalised (per million) frequency of the collocation in the focus corpus, fpm(ref) is the normalised (per million) frequency of the collocation in the reference corpus, and n is the smoothing parameter (n = 1 is the default value).

What follows from this formula is that, in general, a value of 1 indicates that a given collocation has the same normalized frequency in both the focus and the reference corpora. If it is higher than one, in general it indicates that a given collocation has a higher normalised frequency in the focus corpus (i.e. it is overrepresented there). By contrast, a value lower than 1 indicates that a given collocation has a lower normalised frequency

in the focus corpus (i.e. it is underrepresented there). It is worth noting at this point that a consequence of adopting the simple maths formula for keyness scores is that a score of 0.5 represents the same degree of underrepresentation as does a score of 2.0 regarding overrepresentation.

In order to establish the most underrepresented collocations of the noun *evidence* in the UKSCC as opposed to the WBNC, the same procedure was used. The only difference lay in the fact that the roles of the corpora (focus and reference) were reversed in Sketch Engine and the keyness score was calculated by the author using the formula mentioned earlier.[3] This modification enabled a list to be produced that also included cases where the UKSCC did not contain a single occurrence of a given collocation but the WBNC did. These cases would have been absent, had the reversal not taken place as Sketch Engine removes from a list of key collocations those candidates in the focus corpus that are not featured in it. However, collocations that are attested in the WBNC and not in the UKSCC also merit inclusion in the analysis as they constitute a prime example of variation in collocation use between corpora.

In order to investigate the potential of the WBNC and the UKSCC for the retrieval of legal collocations, two word sketches were generated: one for the UKSCC and one for the WBNC. The resulting lists of collocates were sorted according to the frequency[4] of the collocates and then compared. Unlike in the previous two stages, certain candidate collocates suggested by Sketch Engine were removed from the analysis. Such exclusion concerned (semi-)determiners (*first, other, own,* and *such*) as well as the adjective *only.* All of them exhibit low informative value and are typically omitted from dictionaries, terminological

---

[3] For the sake of consistency, when calculating the keyness score, the original roles of both corpora were left unchanged, i.e. the UKSCC acted as a focus corpus and the WBNC acted as a reference corpus.

[4] As pointed out by several authors (e.g. Łukasik 2017: 53, Michta 2018: 50, Rzepkowska 2021: 280), frequency may not be the ideal criterion for retrieving words and collocations from corpora for pedagogically-oriented publications. It is, however, an important one and provides a good starting point for analysing collocations.

glossaries or pedagogically-oriented lists of collocations. Their exclusion seemed all the more uncontroversial as, to all intents and purposes, they are unlikely to be what a user hopes to find in such resources.

4. Analysis and results

The structure of this section follows the order of aims presented earlier. First, the results of the analysis of the most overrepresented collocations of the noun *evidence* as opposed to the WBNC are presented, followed by the study's findings as to the most underrepresented collocations of the same term. Finally, the potential of both corpora for legal collocations retrieval is evaluated.

An initial list of potentially overrepresented collocations in the UKSCC featured 105 candidate collocates. The list was refined manually following the procedure described in the methods section. For reasons of space, a complete list of candidate collocates will not be presented here.[5] Instead, only those parts are provided that are key to the analysis.

The first observation that can be made is that the UKSCC shows a marked tendency to overrepresent certain collocates. As many as 41 received a keyness score of 2 or more. The data concerning them are included in Table 1.

Table 1

The most overrepresented collocates of *evidence*
in the UKSCC as opposed to the WBNC

No	Collocate	Raw frequency in the UKSCC	Raw frequency in the WBNC	Normalised frequency in the UKSCC	Normalised frequency in the WBNC	Keyness score
1	expert	188	56	19.79	0.56	13.70
2	oral	158	72	16.63	0.72	10.28
3	fresh	123	52	12.95	0.52	9.19

---

[5] The complete list is available from the author upon request.

4	relevant	78	24	8.21	0.24	7.44
5	hearsay	56	30	5.90	0.30	5.31
6	such	123	178	12.95	1.77	5.04
7	other	121	178	12.74	1.77	4.96
8	cogent	33	3	3.47	0.03	4.34
9	new	87	161	9.16	1.60	3.91
10	medical	66	116	6.95	1.15	3.69
11	live	23	1	2.42	0.01	3.39
12	credible	21	3	2.21	0.03	3.12
13	anony-mous	20	0	2.11	0.00	3.11
14	reliable	27	26	2.84	0.26	3.05
15	admissible	25	20	2.63	0.20	3.03
16	criminal	21	7	2.21	0.07	3.00
17	documen-tary	46	100	4.84	0.99	2.93
18	statistical	26	31	2.74	0.31	2.86
19	inadmissi-ble	19	6	2.00	0.06	2.83
20	opinion	19	6	2.00	0.06	2.83
21	prima facie	25	29	2.63	0.29	2.82
22	additional	25	30	2.63	0.30	2.80
23	unchallen-ged	17	2	1.79	0.02	2.74
24	sufficient	53	144	5.58	1.43	2.71
25	factual	19	14	2.00	0.14	2.63
26	objective	24	35	2.53	0.35	2.62
27	false	19	15	2.00	0.15	2.61
28	material	18	15	1.89	0.15	2.52
29	character	15	3	1.58	0.03	2.50
30	only	28	65	2.95	0.65	2.40
31	general	14	4	1.47	0.04	2.38
32	conclusive	34	93	3.58	0.92	2.38
33	DNA	13	0	1.37	0.00	2.37
34	own	17	20	1.79	0.20	2.33
35	compelling	16	21	1.68	0.21	2.22

36	post-pu-blished	11	0	1.16	0.00	2.16
37	written	18	35	1.89	0.35	2.15
38	insufficient	24	70	2.53	0.70	2.08
39	further	65	279	6.84	2.77	2.08
40	available	33	118	3.47	1.17	2.06
41	primary	12	13	1.26	0.13	2.00

In Table 1, it can be seen that the overwhelming majority of collocates which are overrepresented in the UKSCC are constituted by lexical words (also called content words). This is in line with the word sketch algorithm, which prevents grammatical words (also called function words) from being displayed as collocates. Yet some of the candidate collocates may also be classified as grammatical words. The word *such*, for instance, which is one of the most overrepresented collocates, may be more suitably described as a semi-determiner (Biber et al. 1999: 281). *Other* and *own* are also grammatical words and may be classified as determiners (Biber et al. 1999: 258, 271). Determiners and semi-determiners are unlikely to be counted as collocations in pedagogically-oriented resources and theoretically-oriented analyses (especially within the phraseological strand of collocation research, see e.g. Nesselhauf 2004: 11-18) as the fact they are used with a noun is more easily explained by referring to grammatical rules than to the phenomenon of collocations. While *such* could be dismissed as being largely irrelevant for the study of collocations, the fact it placed sixth in the list seems unlikely to be accidental. Rather, it appears to reflect a tendency observed in the UKSCC to employ *such* particularly often. This word can be used to mean "of a particular of similar type" (CD 2021), as it often does in general English as well. But it can also be used to mean "this specific person/thing", in which case it is typical of legalese (Garner 1995: 849) and is sometimes described as an anachronism (Tiersma 2000: 91) together with *aforementioned*, *same* and *said*, when used before a noun to serve deictic purposes. A concordance search revealed that it is clearly the first use that dominates in the UKSCC and con-

tributes most to its increased frequency.[6] A particularly important function of *such*, as used in the UKSCC, is that it enables discussions of legal matters to abstract away from aspects of particular evidence to the more general aspects of categories of evidence, thus also facilitating a link between legal norms and their application to particular cases.

As regards the lexical collocates included in the table, it seems clear that a substantial number of the resulting word combinations with *evidence* exhibit different degrees of terminological character. The following are given entry status in the *Longman Dictionary of Law* (Richards and Curzon 2011): *expert evidence*, *oral evidence*, *relevant evidence*, *hearsay evidence*, *documentary evidence*, *opinion evidence*, *prima facie evidence*, *character evidence*, *conclusive evidence*, *insufficient evidence*, and *primary evidence*. The list of terms could also be extended by adding *sufficient evidence*, which unlike *insufficient evidence* is not included in the dictionary, as well as *admissible evidence* and *inadmissible evidence*, as these are related to *admissibility of evidence*, which appears as an entry.

The presence of so many terms in the list can be explained by the nature of the two corpora used for comparison. It is only natural to expect a specialised corpus intended to represent a special language known for its rich terminology (legal English) to differ from a general corpus by containing a particularly high proportion of terms, which a general corpus by definition cannot exhibit. It may be, however, no easy task to pinpoint exactly which combinations function as terms in a particular LSP and what they actually mean since this is an area where general language competence may not suffice.

The legal terms included in Table 1 differ both with regard to their transparency and their recognisability as terms. The word

---

[6] The UKSCC contains sporadic instances of the construction *such evidence as (there) is/was*, some of which are found in quotes used in the judgments, e.g. "She did, however, observe that 'such evidence as was before the Judge' suggested that the expenditure would not have enhanced the value of the property, albeit without identifying what evidence she had in mind." [2017] UKSC 21. In cases like this the use of such comes close to the second function of *such* described in the text and can be seen as replacing "the".

oral, with the second highest keyness score, forms a term with the word *evidence*. The *Longman Dictionary of Law* (Richards and Curzon 2011: 184) states that this is evidence: "given in court by word of mouth. It may be testimony (i.e., what the witness perceived through his senses) or hearsay". The resulting meaning, as captured by the definition, is not so far removed from the meanings of its constituents used in general language so as to make the term incomprehensible. It could therefore be assumed to be relatively transparent. At the same time, it should be noted that while the meaning of a term might appear simple after it has been explained, it is often more challenging to guess the exact meaning by relying only on the knowledge of its constituents. A number of possible meanings could be suggested for the combination of *oral* and *evidence* based on what they mean in general language (or in special language for that matter). Such predictions may prove of limited use in specialised fields such as law, where there is a marked tendency and a considerable need to ensure a high level of precision by carefully delineating a concept from similar ones (Tiersma 2008: 21). In the specific case of *oral evidence*, one might be tempted to speculate as to what it means in law and since this is a term that is relatively transparent, the results of such speculation could even be close to the definition quoted earlier. However, it seems implausible that a person uninitiated with legal English terminology would be able to suggest that the term "includes evidence which, by reason of any disability, disorder or other impairment, a person called as a witness gives in writing or by signs or by way of any device" (Criminal Justice Act 2003, s. 140). In this regard, the example of *oral evidence* shows that terms might be deceptively simple.

Full transparency of meaning cannot be assumed for any of the terms in Table 2 (or any other terms), as it is in the nature of combinations of words acting together as terms to show at least a certain degree of opacity – something which, in fact, is part of what makes them terms. Certain terms, however, can at least more easily be recognised as such. A good example is provided by *prima facie evidence,* which ranks 21 in the table and

may be immediately suspected of being a term due to its Latin constituent, even if its meaning may be impossible to guess by a person unfamiliar with Latin. Rather confusingly, the term can be employed in two different senses. While it can be used to refer to "evidence adduced by a party [which] is so weighty that it could reasonably justify a finding in his favour" (Keane and McKeown 2012: 30), it can also mean "evidence adduced by a party which is, in the absence of contradictory evidence, so weighty that it does justify a finding in his favour" (Keane and McKeown 2012: 30). Even for a person familiar with the Latin phrase, it may be a challenge to predict that the term is used in both of these meanings.

The fact that quite a few terms appear in the list underscores the key role that they play in legal English as an inherent aspect of its lexicon. Underlying this fact is the need to categorise different varieties and types of evidence. Although a general definition for evidence can be given, a particular item of evidence is perceived differently depending on whether or not it can be qualified by *oral*, *hearsay*, *sufficient*, *documentary*, etc. Most importantly, by classifying an item of evidence using one of these modifiers important legal implications follow. While in general English the need to classify evidence may not be acute, in legal English it seems to be a major driving force behind term formation.

Table 1 also includes a number of adjectives that serve either evaluative or descriptive functions, e.g. *fresh*, *new*, *cogent*, *live*, *credible, reliable*, *general*, *further*, and *available*. With the exception of *cogent*, *live*, *credible* and *general,* they all appear fairly frequently in the WBNC. It is worth noting in passing that the words *fresh* and *new* can be considered as (near)-synonyms, for example here: "If the fresh evidence were always evidence of primary fact, or new expert evidence , the test might be satisfactory" [2011] UKSC 18. Additionally, each of the words can potentially be subject to terminologisation, e.g. by being defined for the purposes of a given legal act. For instance, the Criminal Justice Act 2003, s. 78(2) states that: "Evidence is new if it was not adduced in the proceedings in which the person was acquitted (nor, if

those were appeal proceedings, in earlier proceedings to which the appeal related)", thus severely restricting the meaning of *new* when used before *evidence*.

In the next step, the study also looked at the most underrepresented collocations of the noun *evidence* in the UKSCC as opposed to the WBNC. Given its restricted nature, it was natural to assume that a number of collocations would be overrepresented in the UKSCC. Yet, a number of collocations received a keyness score below 1, suggesting that they might constitute cases of underrepresentation. The data on the 39 collocates with the lowest keyness scores have been summarized in Table 2.

Table 2

The most underrepresented collocates of *evidence*
in the UKSCC as opposed to the WBNC

No	Collocate	Raw frequency in the WBNC	Raw frequency in the UKSCC	Normalised frequency in the WBNC	Normalised frequency in the UKSCC	Keyness score
1	Little	358	10	3.56	1.05	0.45
2	historical	84	0	0.84	0.00	0.54
3	archaeological	74	0	0.74	0.00	0.58
4	Research	70	0	0.70	0.00	0.59
5	considerable	81	1	0.81	0.11	0.61
6	first	54	0	0.54	0.00	0.65
7	early	50	1	0.50	0.11	0.74
8	more	118	6	1.17	0.63	0.75
9	clinical	33	0	0.33	0.00	0.75
10	contemporary	27	0	0.27	0.00	0.79
11	indirect	39	1	0.39	0.11	0.80
12	physical	22	0	0.22	0.00	0.82
13	much	112	7	1.11	0.74	0.82

14	less	21	0	0.21	0.00	0.83
15	fossil	19	0	0.19	0.00	0.84
16	recent	68	4	0.68	0.42	0.85
17	scant	16	0	0.16	0.00	0.86
18	definite	15	0	0.15	0.00	0.87
19	internal	14	0	0.14	0.00	0.88
20	overwhel-ming	37	2	0.37	0.21	0.88
21	visible	13	0	0.13	0.00	0.89
22	sound	13	0	0.13	0.00	0.89
23	abundant	36	2	0.36	0.21	0.89
24	vital	24	1	0.24	0.11	0.89
25	flimsy	12	0	0.12	0.00	0.89
26	molecular	11	0	0.11	0.00	0.90
27	current	22	1	0.22	0.11	0.91
28	wide-spread	10	0	0.10	0.00	0.91
29	reasona-ble	10	0	0.10	0.00	0.91
30	labora-tory	10	0	0.10	0.00	0.91
31	geological	10	0	0.10	0.00	0.91
32	dramatic	10	0	0.10	0.00	0.91
33	previous	9	0	0.09	0.00	0.92
34	irrefuta-ble	9	0	0.09	0.00	0.92
35	observa-tional	9	0	0.09	0.00	0.92
36	consi-stent	9	0	0.09	0.00	0.92
37	computer	9	0	0.09	0.00	0.92
38	firm	20	1	0.20	0.11	0.92
39	hard	100	8	0.99	0.84	0.92

The first observation that can be made about Table 2 is that most of the scores are not particularly low, especially in comparison with the previous table which contained the data for the

most overrepresented collocations. In fact, if the same keyness threshold had been applied to the analysis of both the most overrepresented and the most underrepresented collocations, the table would include a single record for *little*. The reason for this lies in the nature of the WBNC and its size. As a general corpus, the WBNC comprises a wide range of texts encompassing a multitude of topics. It can therefore be expected to include a large pool of collocations (including legal collocations) that reflects its diverse composition. By contrast, the legal corpus can give prominence to a smaller group of collocations which are frequently employed in it.[7] As a result and partly due to the size of the WBNC, for a collocation to meet the keyness threshold of 0.5, it would have to, for instance, appear at least 101 times in the WBNC and not a single time in the UKSCC. Such frequency requirements were considered too strict. Since the keyness score was employed only as a helpful tool for identifying cases of over- and underrepresentation, it was decided to discuss a similar number of collocates in both cases and, especially in the case of underrepresentation, place particular focus on general trends rather than individual examples.

Table 2 includes a small number of grammatical words that are common in general language, yet in the UKSCC are infrequent collocates of *evidence*. All of them are determiners: *much*, *more*, *little*, *less,* and *first*. Additionally, the first four can also be classified as quantifiers (Biber et al. 1999: 278). *Little* has the lowest keyness score (0.45) but the scores for the remaining four are also low. This is rather surprising. One might expect – given the high frequency of these words in general English – that the same pattern would also occur in the UKSCC. Yet, only sporadic instances of *much/more/little/less/first + evidence* are attested in the UKSCC. Closely related to *much* and *little*, albeit by no means synonymous, are the adjectives *sufficient* and *insufficient*, which by contrast occur in the UKSCC with a markedly higher

---

[7] A similar explanation is offered in the study by L'Homme and Azoulay (2020: 160-161), who used a specialised corpus on the theme of the environment.

frequency than in the WBNC. This fact could be interpreted as testament to the role that the concept of (in)sufficient evidence plays in the legal system, which also justifies the inclusion of *insufficient evidence* as an entry in the *Longman Dictionary of Law* (Richards and Curzon 2011). The high keyness scores for *sufficient/insufficient* evidence on the one hand, and low ones for *little/much* evidence on the other, suggests that sufficiency of evidence is accorded greater importance in law than its mere quantity.

As regards lexical collocates that are featured in Table 2, several of them are clearly linked to specific fields other than law, e.g. *historical, archaeological, research, clinical, fossil, molecular, laboratory,* and *geological. Historical* is the most underrepresented in the UKSCC, occurring 84 times in the WNBC while not being attested in the UKSCC at all. The low frequency of collocations of *evidence* with these words is to be expected in view of the fact that *historical, archaeological, fossil* etc. *evidence* is unlikely to be adduced in court. Undoubtedly important in the fields to which they share ties, these collocations in the WBNC convey a sense of *evidence* that is closer to sense 1 in the definition quoted earlier.

Another group of collocates that can be distinguished on the basis of Table 2, includes *early, contemporary, recent, internal, visible, sound, current, widespread, previous,* and *computer.* These are of a largely descriptive character and the attributes that they denote appear – on the basis of the corpus evidence – to be more relevant in contexts other than law. Take the word *recent,* it is typically used with *evidence* in the WNBC in medical texts, as illustrated by the following sentence: "Recent evidence has shown that the expression of enzymes participating in biotransformation may play a part in tumour" (BNC). It thus places evidence at some point in time, while such chronological ordering may be considered not to be key in law.

The table also includes examples of collocates that are evaluative in character: *considerable, scant, definite, overwhelming, abundant, vital, flimsy, reasonable, dramatic, irrefutable, consistent, firm,* and *hard.* Some of them may be felt to be rather

strong and emotional (e.g. *flimsy, dramatic, irrefutable, definite*), which may make them unlikely collocates of *evidence* in a UKSC judgment. More importantly, they may seem to be less closely linked to important characteristics of evidence as seen in law. A good case in point is provided by *abundant,* whose meaning could be expressed as "existing in large quantities; more than enough" (OLD 2021), pointing to the fact that it is akin to *sufficient* but stronger. Based on its meaning one could expect *abundant* to collocate with *evidence* in the UKSCC. Stylistically, the adjective is rather formal and for this reason too it would blend in well with the rather formal style employed in judgments. Nevertheless, it is *sufficient evidence* that occurs with a markedly higher frequency than *abundant evidence*. A reason for this could be the fact that sufficient evidence is related to what is known as the evidential burden. As noted by Keane and McKeow (2012: 37): "A party discharges an evidential burden borne by him by adducing sufficient evidence for the issue in question to be submitted to the jury (tribunal of fact)". Since the requirement mentioned in the quote involves only sufficient rather than abundant evidence, qualifying evidence as abundant might appear superfluous to a certain measure and also rather removed from the legal requirements that are in place. That is not to say that *abundant evidence* is an impossible combination in a UKSC judgment, but merely that it is unlikely as the corpus evidence from the UKSCC bears out. One can only agree with Sinclair, when he states that there "are virtually no impossible collocations, but some are much more likely than others" (1996: 411). It should be noted here that *abundant* is not the only adjective that has an analogue in Table 1. *Scant* expresses an idea roughly similar to *insufficient*, while *firm* and *hard* are similar to *compelling,* and last but not least *definite* and *irrefutable* are close to *conclusive*. The preference for certain collocates over others with a closely related meaning (as evidenced by *evidence* in the UKSCC) points to the conventional nature of collocations in legal English.

To achieve its third aim, a word sketch was generated for *evidence* in both the UKSCC and the WBNC and the candidate

collocates were sorted by frequency. The following discussion will be limited to the 50 most frequent collocates in each corpus. It is plausible to assume that dictionary compilers may want to focus on the frequency of a given combination as an important criterion for the selection of collocations. Since space is at a premium in many dictionaries and will prevent many dictionary compilers from including more than about 20 collocations, for the modifier + word category, a list of the top 50 candidate collocates provides a sound foundation for rank comparisons.

Table 3 displays the first fifty collocates of *evidence* in the UKSCC (listed in the second column) and the WBNC (listed in the third column). The first column indicates the rank of a collocate in the frequency lists obtained from each corpus. The third column, then, for each collocate obtained from the UKSCC, gives the rank assigned to it in the top 50 collocates obtained from the WBNC. To indicate that a given collocate was not among the top 50 collocates of evidence obtained from the WBNC, "n/a" has been used.

Table 3
Comparison of the ranks of collocates
in the UKSCC and the WBNC

Rank	Collocates in the UKSCC (sorted by frequency)	Rank of the collocate in the top 50 collocates in the WBNC	Collocates in the WBNC (sorted by frequency)
1	expert	27	further
2	oral	21	clear
3	fresh	29	new
4	new	3	empirical
5	relevant	49	sufficient
6	medical	10	good
7	further	1	direct
8	hearsay	44	available
9	sufficient	5	strong
10	documentary	12	medical

11	conclusive	14	scientific
12	available	8	documentary
13	cogent	n/a	hard
14	direct	7	conclusive
15	good	6	ample
16	clear	2	historical
17	reliable	48	considerable
18	statistical	42	circumstantial
19	additional	43	experimental
20	admissible	n/a	archaeological
21	prima facie	45	oral
22	insufficient	22	insufficient
23	objective	37	research
24	live	n/a	recent
25	credible	n/a	anecdotal
26	criminal	n/a	forensic
27	anonymous	n/a	expert
28	factual	n/a	convincing
29	false	n/a	fresh
30	inadmissible	n/a	real
31	opinion	n/a	early
32	material	n/a	independent
33	written	39	detailed
34	strong	9	indirect
35	unchallenged	n/a	overwhelming
36	ample	15	abundant
37	compelling	n/a	objective
38	empirical	4	substantial
39	character	n/a	written
40	detailed	n/a	conflicting
41	scientific	n/a	clinical
42	general	n/a	statistical
43	DNA	n/a	additional
44	primary	n/a	hearsay
45	conflicting	n/a	prima facie

46	convincing	28	positive
47	decisive	n/a	contemporary
48	experimental	19	reliable
49	post-published	n/a	relevant
50	specific	n/a	vital

Even a short glance at the table presented above indicates that the lists of collocations retrieved from the two corpora are strikingly different. This disparity was amplified in part as a result of removing some candidate collocates with the lowest informative value (e.g. *such*) from the list.

The overall picture that emerges from comparing the top 10 collocates retrieved from the UKSCC and the WBNC is one of vast discrepancies between them. Of the top 10 of the top 50 collocates of *evidence* in Table 3, only 3 are shared by both the legal and the general corpora (*new, further* and *sufficient*). In fact, it might be difficult to resist the impression that the lists of top 10 collocates in the UKSCC and the WBNC in fact correspond to two different nodes, rather than the same one but in different corpora. If the top 20 collocations are considered, the number of shared collocates grows to 9, achieving its highest overlap rate of 45 %. Expanding the list to the first 30, 40, and 50 collocates, results in a very slight increase in the number of shared collocates, amounting to 10, 14 and 16 respectively, and in a gradual decline in the percentage of overlapping collocates (33 %, 35 %, and 32 % respectively). As many as 23 collocates from the top 50 collocate list obtained from the UKSCC did not make the corresponding top 50 list of collocates based on the WBNC. An additional comparison between the ranks of the collocates in the two lists adds further weight to the claim regarding the wide discrepancies between them. The word *expert* ranks first on the UKSCC list but only 27th on the WBNC list. *Fresh* ranks 3rd and 29th respectively. *Relevant* is also an extreme case, ranking 5th and 49th. These are but a few examples. Many others can be found in the table. One conclusion that offers itself up from the comparison is that the potential of each corpus can be harnessed to produce lists of collocations with *evidence,*

yet the results will be considerably affected by the differences between the corpora and the language they are intended to represent.

5. Concluding remarks

The analysis reported in the previous section set out to establish and explore the most over- and underrepresented modifier + noun collocations of *evidence* in a legal corpus as opposed to a general corpus, and investigate the potential of the two corpora for the retrieval of such collocations. The main findings can be summarised as follows:

(1) The legal corpus was found to exhibit a strongly marked tendency to overrepresent certain collocations. As could be expected, a large subset of them was constituted by combinations that displayed various degrees of terminological character. However, a not insignificant number of combinations were identified whose overrepresentation could not be put down to their status as terms. Three of them were classified as (semi-)determiners. As for the remaining ones, it was suggested that what could better explain the prominence of such collocations in the corpus is the particular relevance of the attributes that the modifiers of *evidence* acquire when used in legal contexts.

(2) The legal corpus was found to reveal a slight-to-moderate tendency to underrepresent certain collocations, indicating a large pool of collocates of *evidence* that are shared by the two corpora and a somewhat limited systematic avoidance of a small group of collocates exhibited by the legal corpus. Apart from a significant proportion of collocates whose underrepresentation could easily be accounted for by pointing to the close link that they share to fields other than law, the study also identified 5 determiners and a large number of other collocates whose underrepresentation could be explained by their limited relevance in legal contexts or by the conventional nature of certain collocations, which results in a preference for particular collocates over others with a similar meaning.

(3) While both corpora were found to provide fertile material for col-
location retrieval, a marked divergence was noted between the lists
of collocations of *evidence* obtained from each of them. Many collo-
cates were assigned vastly different ranks in the lists and the over-
lap between the top 50 collocations from both corpora stood at just
32 %. This figure is strikingly close to the percentage reported by
L'Homme and Azoulay (2020: 166), despite the differences in scope
and methodology between their study and the current one.

Based on these findings, several tentative observations can be
made. Even for a word that is frequently used both in legal (spe-
cialised) and in general language and has a relatively simple
sense structure, its collocational profile can vary sharply be-
tween a general and a legal (specialised) corpus. This might ex-
plain why a high level of competence in general language does
not automatically translate into an equally high level of compe-
tence in legal (specialised) language and vice versa, which is a phe-
nomenon that has been observed, among others, by Franken-
berg-Garcia (2018) and Giczela-Pastwa (2021: 191). Due to the
extent of the differences in collocations retrieved from a general
corpus and a legal one, linguistic resources offering collocations
compiled from a general corpus may prove to be of limited use
when the context calls for legal (specialised) collocations and
vice versa. As a consequence, this may warrant the need to de-
velop such resources and language materials that target collo-
cations in specific LSPs rather than in general language. Despite
the insights produced into the nature of collocations, the study
is not without limitations. It has to be acknowledged that the
choice of corpora to represent general and legal language may
have exerted some influence on the number and diversity of the
collocations retrieved from them. The written component of the
general corpus covers texts that were produced earlier than
those in the legal one. Additionally, the UKSCC contains only
UKSC judgments spanning a limited time frame and cannot
thus capture the full complexity and wide diversity of legal Eng-
lish as used in the UK. Coupled with the study's focus on collo-
cations featuring a single term and, to boot, only on the modifier
+ noun type, all these facts limit the generalisability of the

findings, although their combined impact is difficult to estimate. It is hoped that future investigations will address some of these limitations since much further research into the nature of legal (specialised) collocations is needed.

References

[2011] UKSC 18 = R (on the application of Adams) v Secretary of State for Justice (2011). United Kingdom Supreme Court, case 18. Available at <http://www.bailii.org/uk/cases/UKSC/2011/18.html>. Accessed 15.06.2021.

[2017] UKSC 21 = BPE Solicitors and another v Hughes-Holland (in substitution for Gabriel) (2017). United Kingdom Supreme Court, case 21. Available at <http://www.bailii.org/uk/cases/UKSC/20 17/21.html>. Accessed 15.06.2021.

Criminal Justice Act 2003. Available at <https://www.legislation.gov. uk/ukpga/2003/44>. Accessed 15.06.2021.

Badziński, Arkadiusz (2019). "Problems in medical translation among professional and non-professional translators: Collocations as a key issue". *Beyond Philology* 16/4: 157–177.

Baker, Paul, Andrew Hardie, Tony McEnery (2006). *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Benson, Morton (1989). "The structure of the collocational dictionary". *International Journal of Lexicography* 2/1: 1–14.

Bergenholtz, Henning, Sven Tarp (1994). "Mehrworttermini und Kollokationen in Fachwörterbüchern". In: Burkhard Schaeder, Henning Bergenholtz (eds.). *Fachlexikographie: Fachwissen und seine Repräsentation in Wörterbüchern.* Tübingen: Gunter Narr Verlag, 385–419.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan (1999). *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education.

Biel, Łucja (2010). "Corpus-based studies of legal language for translation purposes: Methodological and practical potential". In: Carmen Heine, Jan Engberg (eds.). *Reconceptualizing LSP: Online Proceedings of the XVII European LSP Symposium 2009.* Aarhus: Arhus University Aarhus, 1–15.

Biel, Łucja (2012). "Areas of similarity and difference in legal phraseology: collocations of key terms in UK and Polish company law". In:

Antonio Pamies, José Manuel Pazos Bretaña, Lucía Luque Nadal (eds.). *Phraseology and Discourse: Cross Linguistic and Corpus-based Approaches*. Baltmannsweiler: Schneider Verlag, 225–233.

Biel, Łucja (2014). "Phraseology in Legal Translation: a Corpus-based Analysis of Textual Mapping in EU Law". In: Le Cheng, King Kui Sin, Anne Wagner (eds.). *Ashgate Handbook of Legal Translation*. Surrey: Ashgate Publishing.

BNC 2015 = What is the BNC? Available at <http://www.natcorp.ox.ac.uk/corpus/index.xml>. Accessed 15.06.2021.

Boers, Frank, Stuart Webb (2017). "Teaching and learning collocation in adult second and foreign language learning". *Language Teaching* 51/1: 77–89.

Buendía Castro, Miriam, Pamela Faber (2015). "Phraseological units in English-Spanish legal dictionaries: A comparative study". *Fachsprache* 37/3-4: 161–175.

CD 2021 = *Cambridge Dictionary*. Available at <https://dictionary.cambridge.org>. Accessed 15.06.2021.

Firth, John R. (1957). "Modes of meaning". In: J. R. Firth. *Papers in Linguistics, 1934-1951*. London: Oxford University Press, 190–215.

Firth, John R. (1968). "A synopsis of linguistic theory, 1930–1955". In: F. R. Palmer (ed.). *Selected Papers of J. R. Firth 1952–1959*. Bloomington: Indiana University Press, 168–205.

Frankenberg-Garcia, Ana (2018). "Investigating the collocations available to EAP writers". *Journal of English for Academic Purposes* 35: 93–104.

Garner, Bryan A. (1995). *A Dictionary of Modern legal Usage*. 2nd edition. Oxford: Oxford University Press.

Garner, Bryan A. (ed.) (2004). *Black's Law Dictionary*. 8th edition. St. Paul: Thomson West.

Giczela-Pastwa, Justyna (2021). "Developing phraseological competence in L2 legal translator trainees: A proposal of a data mining technique applied in translation from an LLD into ELF". *The Interpreter and Translator Trainer* 15/2: 187–204.

Gledhill, Christopher J. (2000). *Collocations in Science Writing*. Tübigen: Gunter Narr Verlag.

Goźdź-Roszkowski, Stanisław (2011). *Patterns of Linguistic Variation in American Legal English: A Corpus-Based Study*. Frankfurt am Main: Peter Lang.

Goźdź-Roszkowski, Stanisław, Gianluca Pontrandolfo (2015). "Legal phraseology today: Corpus-based applications across legal languages and genres". *Fachsprache* 37/3–4: 130–138.

Heid, Ulrich (2001). "Collocations in sublanguage texts: Extraction from corpora". In: Sue Ellen Wright (ed.). *Handbook of Terminology Management: Applications-Oriented Terminology Management*. Vol. 2. Amsterdam – Philadelphia: John Benjamins Publishing Company, 788–808.

Hoey, Michael (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Keane, Adrian, Paul McKeown (2012). *The Modern Law of Evidence*. 9th edition. Oxford: Oxford University Press.

Kilgarriff, Adam (2005). "Language is never ever ever random". *Corpus Linguistics and Linguistic Theory* 1/2: 263–276.

Kilgarriff, Adam (2009). "Simple maths for keywords". In: Michaela Mahlberg, Victorina González-Díaz, Catherine Smith (eds.). *Proceedings of the Corpus Linguistics Conference, CL2009, University of Liverpool, UK , 20–23 July 2009*. Liverpool: University of Liverpool. Available at <http://ucrel.lancs.ac.uk/publications/cl2009/171_FullPaper.doc>. Accessed 15.06.2021.

Kjaer, Anne Lise (1990a). "Context-conditioned word combinations in legal language". *Terminology Science & Research* 1/1-2: 21–32.

Kjaer, Anne Lise (1990b). "Phraseology research – State-of-the-art". *Terminology Science & Research* 1/1-2: 3–20.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel (2014). "The Sketch Engine: Ten years on". *Lexicography* 1: 7–36.

Lexical Computing (2015). *Statistics used in the Sketch Engine*. Available at <https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf>. Accessed 15.06.2021.

L'Homme, Marie-Claude (2000). "Understanding specialized lexical combinations". *Terminology* 6/1: 89–110.

L'Homme, Marie-Claude, Daphnée Azoulay (2020). "Collecting collocations from general and specialised corpora: A comparative analysis". In: Gloria Corpas Pastor, Jean-Pierre Colson (eds.). *Computational Phraseology*. Amsterdam – Philadelphia: John Benjamins Publishing Company, 151–175.

L'Homme, Marie-Claude, Patrick Leroyer (2009). "Combining the semantics of collocations with situation-driven search paths in specialized dictionaries". *Terminology* 15/2: 258–283.

Łukasik, Marek (2017). "Lexical minimum (re)defined." *Lingwistyka Stosowana* 23/3: 47–63.

Mair, Christian (2007). "Varieties of English around the world: Collocational and cultural profiles". In: Paul Skandera (ed.). *Phraseology and Culture in English*. Berlin: Mouton de Gruyter, 437–468.

Martin, Willy (1992). "Remarks on collocations in sublanguages". *Terminologie et Traduction* 2-3: 157-164.

Matilla, Heikki E. S. (2012). "Legal vocabulary". In: Peter M. Tiersma, Lawrence M. Solan (eds.). *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press, 27–38.

Meyer, Ingrid, Kristen Mackintosh (1994). "Phraseme analysis and concept analysis in exploring a symbiotic relationship in the specialized lexicon". In: Willy Martin, Willem Meijs, Margreet Moerland, Elsemiek ten Pas, Piet van Sterkenburg, Piek Vossen (eds.). *Euralex 1994 Proceedings*. Amsterdam, 339–348.

Michta, Tomasz (2007). "Zastosowanie narzędzi korpusowych do badania kolokacji w tekstach specjalistycznych na przykładzie angielskich tekstów chemicznych". In: Marek Łukasik (ed.). *Debiuty Naukowe* I. Warszawa: Katedra Języków Specjalistycznych UW, 55–61.

Michta, Tomasz, Maria Kloza, Jan Łompieś, Mariusz Mela, Wioletta Mela, Magdalena Miąc, Joanna Newska, Liliana Religa (2009). "Studencki Słownik Kolokacji Angielskiego Języka Medycyny". In: Marek Łukasik (ed.): *Debiuty Naukowe* III. Warszawa: Katedra Języków Specjalistycznych UW, 89–225.

Michta, Tomasz (2018). *A Model for an English-Polish Systematic Dictionary of Chemical Terminology*. Siedlce: Scientific Publishing House of Siedlce University of Natural Sciences and Humanities.

Michta, Tomasz, Katarzyna Mroczyńska (2022). *Towards a Dictionary of Legal English Collocations*. Siedlce: Wydawnictwo Naukowe Uniwersytetu Przyrodniczo-Humanistycznego w Siedlcach.

Montero-Martínez, Silvia, Miriam Buendía-Castro (2012). "La sistematización en el tratamiento de las construcciones fraseológicas: el caso del medio ambiente". In: Izaskun Elorza, Ovidi Carbonell i Cortés, Reyes Albarrán, Blanca García Riaza, Miriam Pérez-Veneros (eds.). *Empiricism and Analytical Tools for 21st Century Applied Linguistics*. Salamanca: Ediciones de la Universidad de Salamanca. 711–724.

Mroczyńska, Katarzyna (2020). "A dictionary of legal English collocations as an aid for mastering the legal English genre". *Linguistics Beyond and Within* 6: 130–141.

Nesselhauf, Nadia (2004). *Collocations in a Learner Corpus*. Amsterdam – Philadelphia: John Benjamins Publishing Company.

OLD 2021 = *Oxford Learner's Dictionaries*. Available at <https://www.oxfordlearnersdictionaries.com>. Accessed 15.06.2021.

Picht, Heribert (1987). "Term and their LSP Environment – LSP Phraseology". *Meta* 32/2: 149–155.

Rzepkowska, Agnieszka (2021). "Labour-law basic terminology based on the Polish Labour Code". *Półrocznik Językoznawczy Tertium* 6/1: 257–283.

Richards, P. H., L. B. Curzon (2011). *The Longman Dictionary of Law*. 8th edition. Harlow: Longman.

Saber, Anthony, Audrey Cartron, Claire Kloppmann-Lambert, Céline Louis (2020). "Towards a typology of linguistic and stylistic errors in scientific abstracts written by low-proficiency Geoscience and Mechanical Engineering doctoral students in France". *Fachsprache* 42/3-4: 90–114.

Sinclair, John (1966). "Beginning the study of lexis". In: C. E. Bazell, J. C. Catford, M. A. K. Halliday, R. H. Robins (eds.). *In Memory of J. R. Firth*. London: Longman, 410–430.

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

Tiersma, Peter M. (2002). *Legal Language*. Chicago: University of Chicago Press.

Tiersma, Peter M. (2008). "The nature of legal language". In: John Gibbons, M. Teresa Turell (eds.). *Dimensions of Forensic Linguistics*. Amsterdam – Philadelphia: John Benjamins Publishing Company: 7–25.

UKSC 2021a = Role of the Supreme Court. Available at <https://www.supremecourt.uk/about/role-of-the-supreme-court.html>. Accessed 15.06.2021.

UKSC 2021b = The Supreme Court. Available at <https://www.supremecourt.uk/about/the-supreme-court.html>. Accessed 15.06.2021.

Ward, Jeremy (2007). "Collocation and technicality in EAP engineering". *Journal of English for Academic Purposes* 6: 18–35.

Tomasz Michta
ORCID iD: 0000-0001-8997-6154
Uniwersytet w Białymstoku
Wydział Filologiczny
Plac Niezależnego Zrzeszenia Studentów 1
15-420 Białystok
Poland
t.michta@uwb.edu.pl