LEXICAL BUNDLES IN LEGAL TEXTS CORPORA – SELECTION, CLASSIFICATION AND PEDAGOGICAL IMPLICATIONS

Veronika Tománková

Abstract

This paper analyzes lexical bundles with the aim of determining specific features of the linguistic competence of the selected research population and drawing some pedagogical implications from the findings. The study focuses on lexical bundles retrieved from four legal genre-based corpora (Flowerdew 2005) compiled from texts submitted by respondents who are all legal professionals. In an effort to provide a more comprehensive view of lexical bundles, the phenomenon has been treated using both the conventional register analysis approach (Biber & Conrad 2009) and an approach adopted especially for legal texts by Breeze (2013). The study also attempts to submit proposals for the teaching of lexical bundles (O'Keeffe et al. 2007: 216) in the context of reading comprehension. Within the context of needs analysis, this study has been conducted as part of a larger study aiming to examine communicative competence of legal professionals from the four language skills perspective.

Keywords

legal English, legal English communicative competence (LECC), lexical bundle (LB), structural classification of a LB, functional classification of a LB, frequency analysis, content, LB, non-content LB

1 Introduction

Lexical bundles (LBs) are defined as recurrent, i.e. frequency-driven sequences of three and more words which function as *text building blocks* (Biber et al. 2002: 443) and are referred to by a variety of terms including *prefabricated patterns*, *routines*, *fixed expressions, formulas, lexical phrases* and *LBs* (Biber et al. 2004: 372). Based on a number of studies exploring the discourse of a professional or academic community (Cortes 2004, Biber et al. 2004, Jablonkai 2010, Dontcheva-Navratilova 2012, Cortes 2013, Grabowski 2013) and complying with the key requirement for the studied linguistic features to be *pervasive* in register analysis (Biber & Conrad 2009: 9, 53), the study looked at the LBs most frequently appearing in the sample legal texts representing four different genres.

Discourse and Interaction 9/2/2016, pp. 75-94
ISSN 1802-9930
DOI: 10.5817/DI2016-2-75

The texts from which the LBs were generated had been obtained in the course of semi-structured interviews where the respondents identified the significance of LBs for their job requirements, thus familiarizing the researcher with a wider socio-cultural context of their use and providing her with more insights into the respective legal genres under study. Through this procedure, an ethnographic element was imparted to the investigation. In agreement with Flowerdew (2005), I believe that corpus-based and genre-based approaches to text analysis in English for specific purposes (ESP) present a desirable combination of bottom-up and top-down processing respectively and establish a convenient starting point for effective classroom use.

Based on the assumption that the information on frequency does not inherently carry its explanation (Biber et al. 2004: 376), it is possible that this study will shed more light on the issue of lexical frequency in legal English texts and thus will help to determine its didactic value. In this respect and drawing on the view that a course of ESP in tertiary education contains a major general academic English component, I hope that the present investigation will offer partial insights into the proportion of legal and academic parts of the discourse. This study will also attempt to classify LBs in terms of their structure, function and connection to the text content, based on which I will try to draw pedagogical implications aiming at facilitating students' acquisition of these items and extending their communicative competence in the given ESP field.

The following research questions helped to focus this study:

1) What is the frequency of the most frequently occurring LBs determined both within the constraints of Biber and Conrad's (2009) approach and Breeze's (2013) approach?
2) What overlap is there between LBs generated separately within legal genre specific corpora (experimental corpora) and an academic LBs corpus (reference corpus) as yielded within Biber and Conrad's (2009) approach?
3) Which are the most common types of LBs structurally and functionally in the four experimental corpora as yielded within Biber and Conrad's (2009) approach?
4) What is the ratio of content vs non-content bundles within Breeze's (2013) approach?
5) What pedagogical implications can be drawn from findings generated under questions 1-4?

It is assumed that the ethnographic element as well as the employment of two distinct methods of analysis will enable the study to contribute to the current state of the art of LB investigation.

2 Lexical bundles

2.1 Definition

Unlike lexical chunks which may only be structurally complete units, being frequency-led, LBs may also be constituted by structurally incomplete units. Their "incompleteness" is given by their frequent position at the phrase or clause boundary meaning they tend to *bridge two structural units* (Biber et al. 2004: 377). The obvious importance of LBs for the attainment of a fluent text is further documented by the results of the study conducted by Biber et al. (1999: 995), according to which only 15 per cent of multi-word units present in the spoken word and five per cent of them present in the academic prose form structurally complete units.

As a linguistically unique construct, LBs represent the most frequent phrases in a register (Biber et al. 2004: 371). According to Breeze (2013: 230), the restricted lexical range and a definite set of genres which characterize professional language raise the likelihood of the occurrence of these recurring patterns in ESP texts. Indeed, a number of research studies show that these *conventionalized expressions* form a great part of coherent discourse (Hyland 2008: 8) and as such imply proficient use of professional language (Dontcheva-Navratilova 2012: 38).

The study of multi-word units such as LBs has experienced a dramatic growth recently. Its origins may be traced back to the neo-Firthian approach to word meaning according to which the meaning of a word is constituted as much from the actual use in combination with other words, i.e. collocations, as from the meaning it inherently carries in itself (O'Keeffe et al. 2007: 59). Collocations form an essential part of both written and spoken discourse and according to some authors, their active use is a condition for the attainment of proficient language fluency (ibid.: 60). According to Sinclair (1987a, 1987b, 1987c, 1991a, as quoted in O'Keeffe at al. 2007: 60), it is lexis, not syntax that is responsible for the *organization* and *patterning* of language whereas the role of syntax is diminished to that of a structure to which chunks are slotted.

Simply put, LBs *(on the basis of the, the fact that the)* or *extended collocations* (Biber et al. 1999: 989) reduce the text into more easily manageable chunks through which discourse processing is facilitated (Dontcheva-Navratilova 2012: 41). With the view of their three functional types (cf. Biber et al. 2004 below) it may be said that their existence both assists perception and contributes to the fluency of production (Dontcheva-Navratilova 2012: 41). From the didactic perspective, LBs may be viewed as an important component of discourse and the basic aspect of the knowledge shared by the professional community where their acquisition acts as a major factor of learners' socialization in a [professional] setting (ibid.: 38).

2.2 The frequency and cut-off point

The frequency of LBs may be viewed from a number of perspectives, primarily as a percentage of LBs in relation to the total text volume and the frequency of individual LBs in a given discourse per a set number of words, most commonly one million words (Conrad & Biber 2005: 61) which is the approach adopted in this investigation. However, given the fact that researchers at times process data from corpora of smaller volume, *normalization* or *standardization*, i.e. the conversion to the set amount of words, e.g. 100,000 or one million is necessary. Although Biber and Conrad (2009: 62) illustrate the standardization of the corpus data by conversion per one hundred words, the data in the present study were normalized per one million words to allow for comparison.

To qualify for the further classification process, the LB is required to meet the condition of the minimum frequency, the so-called cut-off point which is arbitrarily set by the researcher (Biber et al. 2004: 376). To illustrate this convention in the case of the authoritative list of four-word academic prose LBs (Biber et al. 1999: 1014-1024), the cut-off point was set at the level of ten occurrences per one million words while the subsequent study (Biber et al. 2004: 376) raised the threshold four times. However, not even this cut-off point may be necessarily perceived as extreme, as the presence of the most frequently occurring LBs in academic prose such as *in the case of* and *on the other hand* exceeds 100 hits per one million words (Biber et al. 1999: 994). In addition, when Biber and Conrad's (2009) approach to LBs is employed there is another requirement for the LB to satisfy which is the occurrence in the minimum number of texts across the given corpus in order to eliminate idiosyncratic features which reflect the unique author style rather than typifying the register in question (Biber et al. 1999: 993, Biber et al. 2004: 376, Biber & Conrad 2009: 7). Two authoritative studies (Biber et al. 1999: 992, Biber et al. 2004: 376) in this respect determine the minimum threshold of five texts without further specification of corpus size.

2.3 Length and classification of lexical bundles

According to a number of studies (Biber et al. 2004, Cortes 2004, Hyland 2008, Dontcheva-Navratilova 2012, Grabowski 2013 and others), four-word LBs lend themselves best to functional classification due to their easier specification when compared with three-word LBs and a lower degree of variability than five-word LBs. Moreover, the choice of four-word LBs is convenient due to its ready comparability with other studies or authoritative lists of LBs. In this investigation, I have used the analytical framework consisting of its structural and functional classifications (cf. below) as designed by Biber et al. (1999) and

Biber et al. (2004) where the concrete LBs listed also served as a reference corpus for the data generated within the four experimental corpora. Within Biber et al.'s (1999: 1014-1024) structural classification noun phrases with specific fragments *(the form of a, the way in which)*, prepositional phrases *(about the nature of, in England and Wales)* as well as verb phrases *(is based on the, should be noted that)* are included among others.

The aim of the functional classification is to find out how the given phrase *behaves* (Breeze 2013: 231) within discourse. According to the classification of Biber et al. (2004), there are three components which may be viewed as signals contributing to discourse fluency (Dontcheva-Navratilova 2012: 41). In this respect, the following categories are distinguished:

1) *stance expressions* which reinforce authorial presence (ibid.) and reflect the author's certainty towards presented information through personal *(I want you to)* and impersonal *(it is possible)* expressions (Biber et al. 2004: 389);
2) *discourse organizers (to look at the, on the other hand)* which reflect the relationship between preceding and following textual information (ibid.: 384-388) and thus perform a significant cohesive role (Dontcheva-Navratilova 2012: 41);
3) *referential expressions (and one of the, in terms of the)* which point directly to physical or abstract entities or the very context of the text aiming to identify such an entity (Biber et al. 2004: 384-388), by the means of which *topic continuity* is supported (Dontcheva-Navratilova 2012: 41).

2.4 Relevant research studies

In an effort to distinguish between the characteristic features of student and professional writing, Cortes (2004) identified and compared the main structural and functional types of LBs in published writing in history and biology to find and specify the potential differences between these two disciplines. The author (ibid.) found out that in professional history texts the two major types of structurally classified LBs are the noun phrase and the prepositional phrase, while in biology texts the range of structural types was significantly wider. In addition to the above-mentioned types, biology texts demonstrated the occurrence of the *it+be+adjective* or *verb+complement* phrases, which signals the use of pragmatic hedging in these disciplines (ibid.: 410). In an attempt to identify differences within the LB typology across various branches of science, Hyland (2008) observed the existence of *research-oriented bundles* in science and technical texts which focused on the communication of empirical methods and from the structural point of view demonstrated a strong inclination towards the noun phrase *(the performance of the)*. In his view, the prevalence of *text-*

oriented bundles in the humanities corpora gives evidence of the emphasis these disciplines place on fluent argumentation. These multiword sequences such as *in the case of* show a strong preference for explicit interpretation and ethical dimension over the presentation of empirical methods. Finally, the third type was represented by *participant-oriented* bundles which focus on the author of the text or its addressee *(it should be noted)* (ibid.: 13-14, 16, 18-19). Similarly, in the present study, I will try to identify the most frequent bundles in various legal genres including their structural and functional classification, indicate the significance of prevalence of certain types in the texts under study and if necessary point out to potential new classification categories.

3 Material and method

Using the texts submitted by a total number of 14 respondents representing three professional spheres (academic, state, and private), four distinct corpora determined by the text content (Academic and Study Legal Texts, Judicial Decisions, EU Legislation, and Contracts) were created which contained a total of 34 texts with a total volume of 421,760 words where the average text size was 12,405 words per text and the average size was nine texts per corpus. Using SketchEngine, a text corpus management and analysis software tool developed by Lexical Computing Ltd. in cooperation with the Faculty of Informatics, Masaryk University, Brno, the Czech Republic, the most frequent LBs were generated.

Corpus name	Number of texts	Total number of words
Corpus 1 Academic and Study Legal Texts (C1)	6 texts	129,552 words
Corpus 2 Judicial Decisions (C2)	9 texts	39,001 words
Corpus 3 EU Legislation (C3)	9 texts	77,387 words
Corpus 4 Contracts (C4)	10 texts	175,820 words

Table 1: Experimental corpora specifics

3.1 Analysis according to Biber and Conrad (2009)

As stated above, for analysis, primarily four-word bundles were chosen due to the easier specification of their structure and function as opposed to three-word bundles and due to a lower degree of variability as opposed to five-word and longer bundles. Apart from the minimum (standardized) frequency which was arranged at 30 hits per one million words, I set the minimum number of

corpus texts in which the LB was to appear at 50 per cent of the total number of texts of the given corpus.

Based on the assumption that ESP courses contain a strong general academic component, the LBs generated within each of the experimental corpora were first separately assessed with regard to the degree of overlap with the authoritative list of four-word academic LBs (Biber et al. 1999, Biber et al. 2004) which, as mentioned earlier, served as a reference corpus. Next, the LBs were classified by two independent raters both structurally and functionally where reliability was measured by direct percentage agreement yielding a result of 85.5 per cent (the raters were highly consistent) to determine the lexico-grammatical areas to take into account in curriculum creation.

3.2 Analysis according to Breeze (2013)

The aim of the above classification was to offer the view of the selected phenomenon from the perspective of apriori defined forms as represented by Biber's studies (Biber et al. 1999, Biber et al. 2004, Biber & Conrad 2009). Among other findings, this analysis was expected to provide information on the most frequent bundles as classified both structurally and functionally. In line with research studies dealing with legal English genres (Bhatia 1993: 107, Breeze 2013: 238, 242) the results confirmed the dominant position of noun and prepositional phrases (cf. below). In an attempt to provide a deeper insight into specialist legal English lexis and grammar, noun and prepositional phrases were subjected to a second analysis. To do so, the methodology of Breeze (2013) was employed which disregards the criterion of the minimum occurrence across corpus texts. This approach, which has led to the inclusion of LBs present within a single text only, also meant the generated LBs demonstrated a frequency as high as thousands per one million words. In an effort to make the resulting amount of data better manageable and to compensate for the non-existence of the criterion of pervasive occurrence across corpus texts, the cut-off point has been raised to 60 hits per one million words. For the same, i.e. economic reasons, only four-word bundles were included in the study. As implied above, this analysis only treated those structurally categorized LBs (noun phrases and prepositional phrases) which had been determined as most frequent by the the first analysis.

Using the methodology of Breeze (2013), the generated LBs were classified into *content* and *non-content phrases* where *the Court of Appeal* is an example of the former type while *the terms of the* represents the latter one. Content LBs are further subdivided into *abstract concepts, agents, documents, dates* (ibid.), which in the present study was changed into *time* to make it more inclusive, and *actions* while LBs failing to establish an immediate link to the text content

undergo no further subdivision. Raters' perception of the potential border-line cases was harmonized using the notion of specificity whereby the expression *the provisions of the* was classified as a non-content phrase whereas *the provisions of Charter* was labelled as a content phrase, i.e. demonstrating an immediate link to a unique document. This approach meant that the concept of non-content phrases adopted in this classification was fairly wide incorporating bundles containing words typically occurring in the academic register *(the basis of the*) as well as those commonly associated with the legal register *(the provisions of the)*. Also Breeze (2013: 242) labels the noun phrase *the parties to the* as non-content. Raters were again highly consistent in achieving the direct percentage agreement of 90 per cent.

I think that this method of legal text analysis, or rather its outcomes, may be used in language teaching as it is capable of distinguishing content-specific LBs which in my view demonstrate a certain degree of overlap with legal specialist lexis, and non-content LBs which offer an insight into discourse organization, particularly in the area of text coherence. With regard to the fact that these bundles always have to meet the condition of high frequency, their mastery may be perceived as an important, or even indispensable component of the communicative competence within a given professional orientation. Besides, this type of textual analysis may be conducted even for individual texts using the classification suitable for a student without a degree in philology.

4 Results and discussion

4.1 Analysis according to Biber and Conrad (2009)

The LBs generated within the first analysis amply fulfilled expectations of their high frequency in English legal texts. This is evidenced by the following figures showing the range of standardized frequency per 1,000,000 words calculated and provided by SketchEngine for the first ten most repeated LBs: C1: 225-88, C2: 616-147, C3: 651-250, C4: 212-37, where the relative frequency substantially exceeds that of LBs in academic texts (Biber et al. 1999: 994, cf. above). The figures below show the ten most frequent LBs for each corpus (following the application of the above-mentioned criteria, only nine LBs were generated in C4) where the bold typeface signals the presence of a given item in the reference corpus.

Corpus	Lexical bundles	Standardized frequency per 1 million words
Academic and Study Legal Texts (C1)	1 of the European Union 2 of the Member States 3 of the Court of 4 in accordance with the ***5 as well as the*** 6 for the protection of 7 in the field of 8 ***the fact that the*** 9 ***in relation to the*** 10 with regard to the	225.2 206.4 193.3 193.3 187.6 168.9 137.6 112.6 93.5 87.6
Judicial Decisions (C2)	1 the meaning of Article 2 for a preliminary ruling 3 on the ground that 4 ***on the basis of*** 5 ***the fact that the*** 6 ***on the other hand*** 7 in accordance with the 8 in accordance with Article 9 ***in the context of*** 10 the provisions of the	616.2 440.1 410.8 381.4 352.1 293.1 264.1 264.1 205.4 146.7
EU Legislation (C3)	***1 for the purpose of*** 2 in accordance with the 3 of the Member States 4 referred to in Article ***5 on the basis of*** ***6 for the purposes of*** 7 in a Member State 8 of the European Parliament 9 referred to in paragraph 10 Parliament and the Council	651.2 633.4 517.4 472.8 419.3 410.4 401.4 356.8 330.1 249.8
Contracts (C4)	1 in connection with the 2 in accordance with the 3 ***for the purposes of*** 4 ***in the case of*** 5 the provisions of this 6 ***on the basis of*** 7 ***for the purpose of*** 8 on behalf of the 9 in the course of	211.6 211.6 170.2 193.2 110.4 96.6 73.6 36.8 36.8

Table 2: Ten most frequent LBs across four experimental corpora

As noted above, the lists of LBs representing the domain of the written and spoken academic discourse (Biber et al. 1999, Biber et al. 2004) served as a reference corpus for the experimental corpora to acknowledge the presence of general academic vocabulary in the language course of any ESP variety in a higher education setting. Therefore, the first aspect to be examined was the level of occurrence of the individual LBs generated within each experimental corpus (legal texts) in the reference corpus (academic texts – spoken and written). The comparison yielded the following resuls showing that the majority of the obtained LBs tended to appear in legal corpora only, with C3 demonstrating the trend most powerfully (71%) and being followed perhaps surprisingly by the corpus containing academic and study texts (C1: 65%). The figures for C2 and C4 were 60 per cent and 55 per cent respectively. Overall then, only a minority of the LBs generated within each of the four experimental corpora was present in the reference corpus.

As mentioned above, the comparison of the general academic corpus and genre-based legal corpora, which was conducted regardless of further structural and functional classification, showed differences in terms of content overlap. Despite this, high agreement was expected in terms of LB typology, where prepositional and noun phrases did indeed present the most frequently occurring structural type, which corresponds with the findings of Biber (2006: 41) indicating the two types put together account for as much as 70 per cent of the total LBs. Within the experimental corpora, the noun phrases formed the following percentage values: 34 per cent in C1 *(the scope of the, the role of the)*, 30 per cent in C2 *(the meaning of Article, the provisions of the)*, 23 per cent in C3 *(the purposes of this, the purposes of the)* and eleven per cent in C4 *(the provisions of this)* while the presence of prepositional phrases was even higher reaching the levels of 54 per cent *(of the Court of, for the protection of)*, 70 per cent *(on the basis of, in the context of)*, 62 per cent *(for the purpose of, on the basis of)* and 89 per cent *(for the purposes of, in the case of)* in C1, C2, C3 and C4 respectively. The total sum of the two categories (C1: 88%, C2: 100%, C3: 85%, C4: 100%) amply confirmed the trend suggested by Biber (ibid.) and outlined pedagogical implications for the teaching of these lexico-grammatical units.

From the functional classification perspective which sheds some light on how the phrase behaves within the text, the most frequently used type was that of referential expressions (C1: 85%, C2: 80%, C3: 100%, C4: 100%) which identify an entity or its particular quality (Biber et al. 2004: 393). Within this group intangible framing attributes (C1: 74%, *in accordance with the, in relation to the*; C2: 50%, *the meaning of Article*, *on the ground that*; C3: 50%, *within the framework of*, *with regard to the*; C4: 56%, *for the purposes of*, *for the purpose*

of) were the most frequently occurring subtype rendering abstract characteristics of the given entity (ibid.: 395). It is possible that this phenomenon is caused by the need to express general and abstract concepts in legal texts across all genres, particularly in specialist texts and textbooks (C1). Given the fact that the legal text is both highly intertextual and intratextual (Vázquez Orta 2010), it also tends to develop a network of connections to other documents, laws and judgments as well as a network within the given text to effectively guide its reader. The results support this notion by a relatively high occurrence of the textual deixis LBs in all the corpora with the exception of C1, in which case the texts are perhaps used in more diverse contexts and the referencing conventions are thus more variable than in standardized documents, which means their presence was not frequent enough to qualify for the corpus inclusion under the register analysis constraints (Biber & Conrad 2009) (cf. above). Nevertheless, those concrete LBs which did meet the inclusion criteria are indicative of a relatively low degree of diversity in these expressions which is not totally unexpected given the fairly standardized structure of the C2, C3 and C4 texts (*in accordance with Article* in C2, *referred to in Article*, *referred to in paragraph* in C3, *in accordance with the* in C4 to name just a few). Finally, the generated data pointed to the necessity of establishing another reference-based classification category. The LBs performing the function of participant reference were identified across three experimental corpora (C1: 8%, C3: 27%, C4: 11%) with the trend being the most apparent in the EU legislation corpus (C3) (*of the Member States*, *the European Parliament and*, *be carried out by*, etc.). The fact that this type of LB was not present in the corpus of judicial decisions was rather surprising. However, the LBs such as *the parties to the*, *in which the defendant*, *raised by the plaintiff*, *which the court could*, etc. which did appear in C2 texts simply were not numerous enough to be added to the four corpora.

4.2 Analysis according to Breeze (2013)

As noted earlier, only noun phrases and prepositional phrases as the most numerous structural types identified in the first analysis were considered in the second analysis.

The data (cf. Tables 3-6 below) show a significant prevalence of content noun phrases over non-content ones across all the four experimental corpora with the tendency being the most pronounced in the case of the Academic and Study Legal Texts (C1) and the Judicial Decisions Corpus (C2). It may be said that this trend evidences the preference of repetition for the sake of accuracy over lexical variety in these legal genres.

Further subdivision within the content phrases domain brought information about the most frequently occurring subtypes where *abstract concepts* became the most prevalent group (C1: 53% of all the noun phrases) and where in a number of documents they even represented the very specialist subject matter under discussion *(general principles of law, law and legal language, the presumption of innocence)*. Content LBs related to *agents* demonstrated the highest presence within the Corpus of Judicial Decisions (C2) which resulted from the inclusion of text passages mentioning *the Zagreb Municipal Court* and *the Zagreb County Court* whose specificity naturally is not representative of the respective legal genre but which is a fitting illustration of the textual need to refer to the institutional authority whose decision forms the substantial part of the text content. Moreover, emergence of this type of highly specific LB neatly demonstrates the immediate accessibility of the text content via this analysis. Lastly, this category was able to catch those participant-oriented bundles which perhaps unexpectedly failed to surface in the corpus of judicial decisions (C2) in the previous analysis (cf. above). Documents-based content LBs show the highest distribution in the Corpus of Contracts (C4) and the Corpus of Academic and Study Legal Texts (C1) where in the former group the obtained LBs refer to other corporate documents *(our consolidated income statement)* or relevant acts which govern the given contract *(the US Securities Act)*. Finally, the category of time was represented only within the Corpus of Contracts (C4) as time is of essence for almost any contractual performance. In this corpus there were also three non-content phrases present with a very strong time component *(the date of this, the time of the, the date of the)*.

The data concerning prepositional LBs do not indicate an equally strong inclination towards the content sequences as was the case with the noun LBs, with C1, C3 and C4 even showing an opposite trend. This might have been caused by the above discussed criterion of specificity. Moreover, the fact that the prepositional sequence is likely to contain a definite or indefinite article besides a preposition means there is an insufficient space left in a four-word bundle to impart specificity to it. This, on the other hand, accentuates the existence of two-word phrases such *Member States* within e.g. the LB *in the Member States* which many times do not qualify to be included in a four-word noun LB.

In the course of both analyses the existence of five-word and longer bundles was observed. I believe these LBs play an important role in the LECC as well as offer additional research potential in the domain of LBs investigation in legal texts. A study of longer than four-word LBs was, however, beyond the scope of the present work.

Content – Abstract Concepts	**Content – Agents**
1 the financial interests of	1 Court of First Instance
2 the presumption of innocence	2 the European Anti-Fraud Office
3 a single set of proceedings	3 European Court of Justice
4 protection of fundamental rights	4 the European Public Prosecutor
5 general principles of law	5 the European Parliament and
6 the application of the	6 the Court of Justice
7 appeal and review mechanism	7 National Law and Institutions
8 the exchange of information	8 the competent authorities of
9 application of the Charter	**Content – Documents**
10 the protection of Fundamental	1 Charter of Fundamental Rights
11 Judgment of the Court	2 the arbitration clause in
12 Direct Effect of WTO	3 the 2000 Standard Terms
13 provisions of the Charter	4 the Treaty of Lisbon
14 protection of the euro	5 Round Evaluation Report on
15 protection of the EC	6 a valid arbitration clause
16 exclusion of the CISG	
17 Effect of WTO Obligations	**Non-content**
18 the national law of	1 the scope of the
19 the fight against fraud	2 the fact that the
20 law and legal language	3 the exclusion of the
21 denial of direct effect	4 the role of the
22 development of the software	5 the implementation of a
23 National law of Mediterraneo	6 the establishment of a

Table 3: Noun phrases generated in the Corpus of Academic and Study Legal Texts (C1)

Content – Abstract Concepts	**Content – Agents**
1 the meaning of Article	1 the Member State of
2 civil and commercial matters	2 the Zagreb Municipal Court
3 use of a language	3 the Member State in
4 a set of civil proceedings	4 Member State of the
5 enforcements of judgments in	5 Member State in which
6 private and family life	6 the Zagreb County Court
7 the recognition and enforcement	7 the Supreme Administrative Court
8 the proceedings concerning the	8 the European Union Legislature
9 the field of application	
Content – Documents	**Non-content**
1 a contract of transport	1 the place where the
2 the preamble to Regulation	2 the courts of the
	3 the length of the

Table 4: Noun phrases generated in the Corpus of Judicial Decisions (C2)

Content – Abstract Concepts	**Content – Agents**
1 the ordinary legislative procedure 2 the sole responsibility of 3 principle of equal treatment 4 interests of the Union 5 the place of arbitration 6 the recognition and enforcement 7 secondary and repeat victimisation 8 investigation and prosecution of 9 consent of the European 10 investigations and prosecutions of 11 proper administration of justice 12 performance of its functions 13 right to equal treatment 14 civil and commercial matters 15 the rules of procedure 16 rights of the defence 17 implementation of the principle 18 cooperation in criminal matters	1 European Public Prosecutor Office 2 The European Parliament and 3 The Committee of Ministers 4 The European Delegated Prosecutors 5 the court first seised 6 the Member States concerned 7 the competent national authorities 8 The Data Protection Officer **Content - Documents** 1 The Convention on the 2 Charter of Fundamental Rights 3 the Model Law on
Non-content 1 the request of the 2 the total number of 3 the competence of the 4 the basis of the 5 the place where the 6 the implementation of the 7 the substance of the 8 the rights of the 9 the territory of the 10 the person against whom 11 the party against who 12 the validity of the 13 the courts of the 14 the date on which	

Table 5: Noun phrases generated in the Corpus of EU Legislation (C3)

Content – Abstract Concepts	**Content – Agents**
1 material adverse effect on	1 our Board of Directors
2 the meaning specified in	2 European Federation of Energy
3 intersegment sales in the	3 the Supervisory Board of
4 investment in the Notes	4 Dukovany Nuclear Power Plant
	5 the Luxembourg Stock Exchange
Content - Time	6 the Board of Directors
1 year ended December 31	7 the Ministry of Finance
Content – Documents	**Non-content**
1 audited consolidated financial statements	1 the provisions of this
2 the U.S. Securities Act	2 the laws of the
3 financial statements for the	3 the purposes of this
4 the Fiscal Agency Agreement	4 the date of this
5 our consolidated income statement	5 the time of the
6 any Credit Support Document	6 the date of the
7 the Czech Nuclear Act	

Table 6: Noun phrases generated in the Corpus of Contracts (C4)

5 Pedagogical implications

The expression "chunk" free of the *lexical* attribute was first used by the leading cognitive psychology scholar Miller in his work on limited capacity of short-term memory where a chunk serves as a coping mechanism to compensate for one's limitations in immediate memory and to successfully deal with "informational bottleneck" (1956: 95). In linguistic processing this means that advanced language users do not deconstruct language to the minimum number of independent units, although they are naturally able to do so, but rather store longer chunks in memory to promptly use them in new target situations (Sinclair & Mauranen 2006: 33-34). In an effort to maximize the capacity of limited memory, the *effective chunking of incoming information* influenced by the schema theory (Rumelhart 1980, as quoted in Sinclair & Mauranen 2006: 37) even formed a substantial part of the instruction of reading in the 1970s and 80s (ibid.). Although the system suggested by Rumelhart (1980: 33) viewed schemata as representing the knowledge of concepts, it is possible that the theory can work for both complete and incomplete stretches of language, i.e. LBs where in addition to the above-mentioned processes, users are able to make associations between them. This network of stored chunks or LBs then forms a complete repertoire which can be retrieved and used innovatively in new communicative situations, which is thought to significantly contribute to fluency both in the domains of production and reception (Sinclair & Mauranen 2006: 33-34, 38-39).

In the area of second language acquisition this approach represents a digression from traditional grammars in that it sees language users as individuals seeking to achieve their communicative goals or purposes through an additive manner (*increments*, Brazil 1995, as quoted in Sinclair & Mauranen 2006: 28) where strict adherence to grammar rules is not of primary importance. Despite being originally formulated for spoken language, Sinclair and Mauranen (ibid.) extend the application of this theory to written language when they claim the increments theory can equally be employed in reading comprehension (ibid.).

Despite the fact that the research conducted by Sinclair & Mauranen (ibid.) differs from the present one substantially in that in the former research study LBs are not generated electronically but rather demarcated by their users based on the subjective perception of their frequency in texts, both research studies are data-driven and avoid predefined grammar categories (ibid.: 36). Although the degree of overlap between the two groups of lexical chunks or bundles yet needs to be determined, I believe the above-mentioned theories make a good case for teaching LBs within ESP classes. Moreover, electronically unassisted perception of the existence and importance of LBs on students' part seems to be a prerequisite of their successful learning and teaching process (Nation 2008: 122).

Even though many would argue for the inclusion of LBs in the ESP curriculum, the didactic value of a particular list of LBs needs to be determined by teachers (cf. Simpson-Vlach & Ellis 2010). There are a couple of studies available which deal with the teaching and learning of LBs while including challenges encountered in doing so. Byrd and Coxhead (2010) examine LBs in academic writing while simultaneously presenting practical tips for teachers on how to handle various difficulties arising in their teaching process such as the inclusion of shorter LBs within longer ones, the contradiction between the spontaneous use of LBs within authentic communicative situations and their analytical treatment within instruction, the lack of face validity in the case of well-known bundles such as *as a result of*, the necessity of students' exposure to LBs within actual academic texts as opposed to their exposure to discipline-driven lists of academic LBs, and finally the issue of insufficient information on the LB context (ibid.: 51-56).

In this respect, in agreement with Flowerdew (2005: 321, 329), I believe that the contextualized use of LBs coupled with genre-specific information and a situational analysis (Biber & Conrad 2009: 36) is the key to their efficient application in teaching. Better localized data provided by the analyst who in the case of the present study is also a corpus compiler equipped with specialist knowledge provided by informants, i.e. text users, makes the top-down processing of corpora much more effective (Flowerdew 2005: 329). In the design of possible

corpus-based learning activities, it is therefore advisable to try to follow this practical principle.

The following teaching activities which may facilitate language acquisition are to a certain extent determined by the method employed in the generation of LBs. The data obtained within the conditions of register analysis according to Conrad and Biber (2009) suggest that relevant classroom activities may include: (i) observation of the immediate or extended context depending on the type of LB in question and its position within a sentence or at a paragraph level, (ii) identification of the most frequent collocates for the most often occurring LBs, (iii) identification of the types of LBs (structural, functional) depending on the legal genre in question as well as (iv) recycling the most relevant LBs in student genre related writing, etc.

The data obtained from semi-structured interviews show lawyers demonstrate a high level of language awareness. This could be utilized within the instruction process where apart from the above-mentioned suggestions for classroom activities learners would be encouraged to recognize trends and patterns in their own data and invent their own categories in doing so.

In addition, the research data collected during interviews with respondents also reveal a dominant position of the reading skill (Tománková 2014) which is further confirmed by the fact that 75 per cent of the submitted representative documents are read by respondents as opposed to the remaining 25 per cent which the respondents authored or co-authored. I am therefore convinced that the method employed in the second analysis according to Breeze (2013) can be used in pre-reading activities of long texts to improve reading comprehension where the LBs extracted within one text can be used to create the text vocabulary profile and discuss the anticipated text content. During post-reading activities students may alternatively assess the level of assistance provided by the study of LBs prior to reading. I believe these two groups of exercises as well as the two methods: the conventional one presenting a more global view of the register in question which also includes genre-non-specific lexico-grammatical items, and the *exploratory* one offering a closer look at specialist lexis, complement each other and as such should be considered in curriculum creation.

Finally, the research results obtained in the first analysis clearly show that academic LBs (Biber et al. 1999, Biber et al. 2004) form an important part of legal texts. This indicates the necessity of incorporating into the curriculum aspects of general English for academic purposes while the extent of such inclusion may depend on the legal genres in question.

6 Conclusions

With the view of the dominant position of the reading skill within the respondents' professional target situations (Tománková 2014) and in line with Alderson (2007) who defines lexical frequency as a vital variable of text comprehension, the goal of this investigation was to identify and classify the most frequently occurring lexico-grammatical items present in texts typically consulted by lawyers within their job requirements and to indicate how these could be used in the classroom in order to facilitate the reading process and thus increase the communicative competence of learners.

The first of the two linguistic analyses sought to provide a global view on the most frequent lexico-grammatical items and contrast them with the most common multiword sequences in academic prose with the objective of establishing the proportion of the most widely occurring legal and academic LBs and specifying their structure and text function. On the other hand, the second analysis, freed from the minimum frequency across texts criterion, was designed to provide an alternative view of the phenomenon and thus draw attention to those LBs which go unnoticed in a linguistic analysis performed within the constraints of register analysis. As implied earlier, due to its supportive function, the second analysis targeted only the most frequent structural LBs as indicated by the primary analysis.

Given the standard requirements in the area of digital literacy, the methods used in data generation may be considered as imposing only basic demands on their users. I am thus convinced the exposure to LBs within the teaching process as well as self-study may become a dominant feature of ESP instruction and as such present further interesting research opportunities in the area of the measurement of the increase in the communicative competence conducted following intervention in the form of LB-driven instruction, where these multiword sequences are not only identified but also classified with regard to their structure and text function.

Overall, the inclusion of LBs in instruction imparts authenticity to teaching and learning, supports learner autonomy by taking the focus off the teacher who becomes a facilitator enabling students to access sources most likely to match their needs (O'Keeffe et al. 2007: 218), enhances digital literacy in students and most importantly actively engages and benefits both the student and teacher.

References

Alderson J. C. (2007) 'Judging the frequency of English words.' *Applied Linguistics 28* (3), 383-409.

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

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. Harlow: Pearson.
Biber, D., Conrad, S. and Leech, G. (2002) *Longman Student Grammar of Spoken and Written English*. Harlow: Pearson.
Biber, D., Conrad, S. and Cortes, V. (2004) 'If you look at...: Lexical bundles in university teaching and textbooks.' *Applied Linguistics 25*, 371-405.
Biber, D. (2006) *University Language: A Corpus-Based Study*. Amsterdam and New York: John Benjamins.
Biber, D. and Conrad, S. (2009) *Register, Genre and Style*. Cambridge: Cambridge University Press.
Breeze, R. (2013) 'Lexical bundles across four legal genres.' *International Journal of Corpus Linguistics 18* (2), 229-253.
Byrd, P. and Coxhead, A. (2010) 'On the other hand: Lexical bundles in academic writing and in the teaching of EAP.' *University of Sydney Papers in TESOL5*, 31-64.
Conrad, S. and Biber, D. (2005) 'The frequency and use of lexical bundles in conversation and academic prose.' *Lexicographica 20*, 56-71.
Cortes, V. (2004) 'Lexical bundles in published and student disciplinary writing: Examples from history and biology.' *English for Specific Purposes 23* (4), 397-423.
Cortes, V. (2013) 'The purpose of this study is to: Connecting lexical bundles and moves in research article introductions.' *Journal of English for Academic Purposes 12*, 33-43.
Dontcheva-Navratilova, O. (2012) 'Lexical bundles in academic texts by non-native speakers.' *Brno Studies in English 38* (2), 37-58.
Flowerdew, L. (2005) 'An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies.' *English for Specific Purposes 24,* 321-332.
Grabowski, L. (2013) 'Register variation across English pharmaceutical texts: A corpus-driven study of keywords, lexical bundles and phrase frames in patient information leaflets and summaries of product characteristics.' *Procedia – Social and Behavioral Sciences 95*, 391-401.
Hyland, K. (2008) 'As can be seen: Lexical bundles and disciplinary variation.' *English for Specific Purposes 27* (1), 4-21.
Jablonkai, R. (2010) 'English in the context of European integration: A corpus-driven analysis of lexical bundles in English EU documents.' *English for Specific Purposes 29* (4), 253-267.
Miller, G. (1956) 'The magical number seven, plus or minus two: Some limits on our capacity for processing information.' *Psychological Review 63*, 81-97.
Nation, P. (2008) *Teaching Vocabulary. Strategies and Techniques*. Boston: Heinle Cengage Learning.
O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From Corpus to Classroom*. Cambridge: Cambridge University Press.
Rumelhart, D. E. (1980) 'Schemata: The building blocks of cognition.' In: Spiro, R., Bruce, B. and Brewer, W. (eds) *Theoretical Isues in Reading Comprehension*.' Hillsdale, NJ: Lawrence Erlbaum. 33-58.
Simpson-Vlach, R. and Ellis, N. C. (2010) 'An academic formulas list: New methods in phraseology research.' *Applied Linguistics 31*, 487-512.
Sinclair J. McH. and Mauranen, A. (2006) *Linear Unit Grammar*. Amsterdam: John Benjamins.

Tománková, V. (2014) 'Analýza řečové dovednosti čtení v oblasti odborného právního anglického jazyka.' In: Janíková, V. and Píšová, M. and Hanušová, S. (eds) *Aktuální témata výzkumu učení a vyučování cizím jazykům III*. Brno: Masarykova univerzita. 281-303.

Vázquez Orta, I. (2010) 'A genre-based view of judgments of appelate courts in the common law system.' In: Gotti, M. and Williams, C. (eds) *Legal Discourse Across Languages and Cultures*. Bern: Peter Lang. 263-284.

Veronika Tománková is a Senior Lecturer of English for Specific Purposes and Academic English at Masaryk University, Brno, Czech Republic. Her professional interests include ESP course design, teaching and testing in higher education settings.

Address: Mgr. Veronika Tománková, Ph.D., Masaryk University Language Centre, Faculty of Education, Poříčí 7, Brno 603 00, Czech Republic. [e-mail: tomankova@ped.muni.cz]