

# ENGLISH LEXIS IN THE DOCUMENTS OF THE EUROPEAN UNION – A CORPUS-BASED EXPLORATORY STUDY

[doi.org/10.61425/wplp.2008.02.38.54](https://doi.org/10.61425/wplp.2008.02.38.54)

**Anna Trebits**

Language Pedagogy PhD Program, Eötvös Loránd University  
a.trebits@yahoo.com

**Abstract:** The present study aims to explore the characteristics of lexical elements in English language documents of the European Union (EU). As the EU is involved in a wide range of activities, the documents related to it exhibit features of texts written for legal, business, political and other specific purposes. The aim of this corpus-based research is to identify and describe the lexical elements whose mastery is indispensable for those who wish to work in cooperation with or inside a European institution. An EU English Corpus consisting of approximately 200,000 words was built using texts which are representative of the various fields of activities of the EU. The analysis uncovered the most frequent lexical elements and collocation patterns as well as their most common uses and senses in this particular variety of the English language. This article discusses and demonstrates the methods of analysis through the detailed presentation of three of the most frequently occurring lexical elements. The pedagogical relevance of the study is that its findings can be directly used for the development of special English language course books and supplementary materials for EU and International Relations courses.

**Keywords:** English for Specific Purposes, corpus-based research, collocations, course design, European Union

## 1 Introduction

The English language is incontestably becoming the *lingua franca* in Europe (as in most parts of the world) due, in part, to the ever growing importance of the European Union (EU) in the life of Europeans and European countries. Of much interest to teachers of English as a second or foreign language (ESL/EFL) is how to help students prepare for the linguistic challenges they may face if they wish to work in one of the several EU institutions, or if their everyday work in their home country requires them to work with EU documents in English. Among the first issues to consider when designing an English language course for prospective EU employees or students majoring in EU studies is how to compile the vocabulary component of the course. *EU English*<sup>1</sup> is becoming an especially important field within research on English for Specific Purposes (ESP) in countries which have recently joined the EU, as the new Member States are preparing for their turn of the EU Council Presidency. Hungary, for example, will preside over the Council of the European Union during the first half of 2011, which calls for the need to produce efficient EU English language development programs for EU experts and ministry officials, students in international relations, etc. Many reference materials and even some practice books have recently been published in Hungary for students, translators and speakers of English which aim to cover the most important terms and specialized vocabulary needed to understand and translate EU-related texts and documents (e.g., Bart & Klaudy, 2003; Jablonkai et al., 2006). The official website of the

---

<sup>1</sup> I use the term *EU English* to refer to English as it is used in the documents of the EU.

European Union ([www.europa.eu](http://www.europa.eu)) also contains a very useful and multilingual *EU Glossary* and even a *Euro-Jargon* section which both aim to clarify some of the essential terms and expressions.

Most publications approach this issue from a translator's point of view by providing equivalents and model texts in two languages or lists of specialized vocabulary in two or even more languages (e.g., Bart & Klaudy, 2003; EUABC multilingual online dictionary). The selection of lexical elements in most of these publications mainly reflects the assumption that knowing the special terms and concepts describing the history and the development of EU institutions (e.g., *Copenhagen criteria*, *enhanced cooperation*) enables intermediate or higher level (B1-B2 levels according to the Common European Framework of Reference) speakers of English to understand and produce EU-related texts efficiently in English. Yet, those who wish to be able to function effectively at an EU institution also need considerable practice in understanding and using both special and common words and collocations that they are most likely to encounter in EU documents in context. Learners of English who are experts in their own field, or secretaries and assistants who work with EU-related documents often speak English at intermediate (B1-B2) level (personal experience and personal communication from Human Resources employees at an EU institution). Therefore, it is very important to determine which lexical elements constitute the list of the most frequent elements in EU-related texts and to uncover patterns in which these are typically used in order to help students of English prepare for the real world linguistic features they are most likely to encounter. This paper reports on a corpus-based investigation of EU documents which is intended to identify the most frequent lexical elements in written texts of the EU and to provide examples as to how the usage patterns uncovered by the analysis could be described to serve as a basis for the compilation of specialized teaching materials designed to enhance the development of EU-related vocabulary.

## **2 Background to the study**

Three main fields of research constitute the background to this study: research on English for Specific Purposes (ESP) in general, research on English as it is used in the documents of the EU in particular and, thirdly, investigations concerning the use of corpora in second and foreign language teaching. The following sections outline the most important points in each of these areas insofar as they underpin the present study.

### **2.1 English for Specific Purposes (ESP)**

The importance of research on ESP in English language teaching (ELT) lies primarily in the realization that different speakers of English carry out various communicative tasks depending on their own situation, their own needs and goals and that language teaching has to cater for the specific needs of the students to make sure that they have the chance to learn the part of language that they really need. Language courses have to be tailored to prepare the students for the particular communicative events and tasks that they are most likely to face in real world situations in order to be effective. It is easy to imagine that a young Hungarian doctor who gets a job in the emergency ward of a US hospital needs to follow a different English language curriculum than a lawyer who is hired by the Court of Justice of the EU. Both of them have to learn a huge amount of General English, nonetheless, many of the situations in which they will use English are most likely to differ to such an extent that they

have to be taken into account when designing their particular courses and course materials. Faced with such specific situations as the ones outlined above, ESP is concerned with teaching English on the basis of what the immediate needs of the learners are and what their reason for learning is, therefore, it is essentially needs-based and task-oriented (Smoak, 2003).

The question of how learners' needs should be assessed has been widely researched and discussed in the literature (see Basturkmen, 1998; West, 1994) as it pertains to what parts of the language will be selected for study in a given language course. A comprehensive needs analysis considers the learner and the learning situation as well as the language variables in play in order to allow for an effective and appropriate course design both in terms of content and teaching method. The learner and learning situation variables include: learner beliefs, motivation, language learning strategies and abilities among others. For the purposes of this study, however, it is important to focus on two aspects that play a key role in determining the selection of the content of the course: the language level of the learners and language use in the real world ESP situations that the learners prepare for. As far as the language level of the learners is concerned, according to Dudley-Evans and St. John (1998), ESP courses in general assume a basic knowledge of the language and in most cases are designed for intermediate or higher level students. This assumption is probably easy to accept: in order for communication to take place regardless of its purpose, a basic knowledge of the language is necessary. Nevertheless, the specific aims and ESP needs of the learner can be incorporated in the language material from the beginning of language instruction. The General English and ESP components of the course probably overlap to a greater extent in the case of less advanced students than in the case of higher level learners of the language, and it can sometimes be difficult to draw the line between the two (Anthony, 1997). For example, learning how to describe the symptoms of an illness can be considered to be an ESP component in the case of the doctor's situation mentioned above, whereas it could also be part of a general English course.

The success of an ESP course hinges to a great extent on the selection of lexical and grammatical elements from the language appropriate to the learners' field of study. Researchers seem to agree that authentic, real world language has to be brought to the learners' attention and that teachers need to take time to study how language is used in real-life situations in a given ESP context in an effort to create a genuinely useful course and to design informative language development materials for their students (Robinson, 1991; Smoak, 2003; Wright, 1992).

## **2.2 Characteristic features of English used in EU texts**

At present the EU is involved in practically all walks of political, social and economic life of its member states. From agriculture and commerce to mobility programs for students and professionals, its decisions and recommendations have come to shape to a great extent the national decision-making processes over the past decades (Károly, 2007). From an ELT standpoint this has at least two important implications: first of all, studying the characteristics of EU English through authentic texts would benefit professionals from a wide range of fields. Second, EU English has indeed developed into a special language featuring characteristics of political, legal, business, and even academic discourse, thus mirroring the different functions and multiple activities of the EU as a cultural, political and historical reality.

A considerable amount of research has been done on the language of the EU particularly with the intention of pointing out issues in translation (see Trosborg, 1997; for an overview see Károly, 2007). EU documents including multilateral agreements, contracts, decisions, regulations, recommendations, general activity reports, etc. are created in one of the major languages (English, French or German) of the Union and most of them are subsequently translated into all the other official languages - a total of 23 languages since the last enlargement took place in 2007. In an effort to iron out problems arising from the text type requirements and norms of individual languages, EU texts are standardized in terms of terminology and structural organization which also makes the work of translators easier (Károly, 2007). This process results in what Trosborg (1997) calls the creation of “hybrid” texts. Hybrid texts are translations that are neither source-culture-bound nor target-culture-oriented but rather aim to be universal in terms of the use of lexical units and grammatical structures. Interestingly enough, the EU has already left its mark on the linguistic development of the languages of its member states by inventing and imposing its own linguistic culture, norms and conventions over its 50 years of existence.

Acknowledging the importance and the magnitude of its linguistic challenges, the EU's official website (<http://ec.europa.eu>) offers several ways for its citizens to familiarize themselves with EU terminology. The EU Glossary and Euro-Jargon pages are available in all EU languages and they provide an alphabetical list with explanations of the most important terms and expressions relating to EU institutions, EU history and the in-house language use of the organizations. A relatively recent initiative on the part of the European Commission is the launch of the IATE, a multilingual inter-institutional terminology database which has been accessible to European citizens since the end of 2006. As is stated on the IATE website, it “aims to provide a web-based infrastructure for all EU terminology resources, enhancing the availability and standardization of the information” with a view to facilitating the work of public administrations, professionals, businesses and citizens in general.

It is clear from the discussion above that EU English involves a number of lexical, grammatical and discourse features of ESP. Therefore, from a language teaching perspective, EU English needs to be analyzed in a similar way to other registers in order to create meaningful and useful learning materials.

### **2.3 Corpus linguistics and materials development in ELT**

The significance of corpus-based studies in language teaching is that they help to point out how English is actually used in context, thus bringing real world language use in naturally occurring texts closer to the learner and the teacher (Biber, Conrad, & Reppen, 1998). Rather than selecting the elements of a course on the basis of the teachers' and materials writers' intuition, corpus-based analyses make it possible to justify the choice of given lexical and grammatical items to be included in a course by empirical research findings (see e.g., Cullen & Kuo, 2007; Gardner & Davies, 2007; Gilmore, 2004; Liu, 2003). Using language data from a corpus – that is a carefully compiled collection of naturally occurring language (Biber et al., 1998) – may prove to be helpful in analyzing how frequently certain lexical elements are used in a given register and what characterizes their usage patterns (see Liu, 2003).

When interpreting the results of frequency studies, it is important to bear in mind that although they provide very useful information on what linguistic elements speakers of a language are most likely to use in a given register, frequency counts alone can also be

misleading. Some rarely occurring items may be crucial for language learners to master and thus their exclusion from the syllabus would actually hinder the learning process. In addition, great care needs to be taken in building the corpus itself so that it should be representative of the register under study so as to avoid that the data obtained from the analyses is distorted (Biber et al., 1998).

### 3 Research Aims

The present study intends to report findings that assist the teaching of English as it is used in written EU documents by identifying the most frequent content words and presenting a possible analysis of some of their usage patterns in a corpus of EU-related texts. The research aims are as follows:

- (1) Find the most frequent content and function words in the corpus.
- (2) Present the use of concordance lists in establishing collocation<sup>2</sup> and usage patterns<sup>3</sup> for some of the most frequent content words<sup>4</sup> and function words showing their pedagogical relevance.
- (3) Describe how the semantic analysis of a frequent word form can be useful in corpus-based materials design.

The subsequent sections will discuss how the corpus was built and present the analytical tools used in this study. Then the presentation and interpretation of the results will follow. The study concludes by relating the findings of the current study to language pedagogy and ESP research.

## 4 Method

### 4.1 The EU English Corpus

In view of my focus on EU English for learners of English who are experts in their own field but are not trained linguistically, it was important for the corpus to reflect the range of activities the EU is involved in but not to be biased towards any of the fields of activity in particular, so that the findings should remain balanced and pedagogically useful. It was assumed that, for example, a trained lawyer would understand the purely technical and specialized vocabulary of his/her own field (e.g., the English words *contract* or *obligation* are probably known to a lawyer who can speak English at B1-B2 (intermediate) level). Consequently, what learners actually need from an EU English course or practice book is the opportunity to learn (1) whether the specialized vocabulary of their own field may be used in a different sense in EU texts than in other texts, (2) EU-related vocabulary from other fields that they may often encounter in EU texts and (3) other non-specialized common words

---

<sup>2</sup> A *collocation* is an expression consisting of two or more words that correspond to some conventional way of saying things (Maning & Schütze, 1999). They include noun phrases, phrasal verbs, and other stock phrases. They are particularly interesting for language learners and teachers because often they show patterns of word usage that are not easy to explain yet native speakers all know and use them correctly.

<sup>3</sup> *Usage patterns* show the contexts in which a given word can be used and the functions that it can perform in a sentence.

<sup>4</sup> A *content word* is a word such as a noun, verb, or adjective that has a lexical meaning, rather than indicating a syntactic function, as a *function word* does.

frequently used in EU texts in context. Bearing in mind these assumptions, the following digitized texts and documents were used to compile the corpus:

- 19 information booklets of 20 pages on average on the different activities of the EU (e.g., competition, market, economy, travel, environment, justice, science, transport, etc.) intended for a general but informed audience. The booklets are downloadable from the Internet at: <http://ec.europa.int> .,
- the annual general report on the activities of the EU in 2006 (published in early 2007 and available on the Internet at: <http://ec.europa.int> .,
- sample test material from recruitment competitions in all subject areas (the sample tests are available from the European Personnel Selection Office (EPSO) website and are freely downloadable).

The last component of the corpus – the sample test – was included because those who seek a job at an EU institution have to pass the EPSO recruitment competition, which is not so much of a challenge in terms of the candidates' expertise as lawyers, economists or secretaries, but it definitely puts their knowledge of EU English to the test. The EPSO tests are also convenient for the purpose of exemplifying how English is used in EU documents because they include texts on the most important subject matters (economy, law, etc.) as well as on general knowledge and understanding of EU institutions and verbal reasoning. It is also important to note that all the texts included in the corpus are relatively recent and were published after 2000.

The final version of the EU English Corpus contains approximately 200,000 running words and it consists of a balanced selection of texts in terms of the Union's fields of activity. As this study focuses on one single register and within that only written texts, the size of the corpus seems to be adequate for its purposes. Table 1 summarizes the details of the EU Corpus.

Origin	Topics covered	Years of publication	Total number of words
<b>19 information booklets on the activities of the EU</b>	audiovisual arts, budget, consumer interests, customs policy, economy, enlargement, environment, foreign policy, globalization & trade, information society, justice, market, science, transport, travel, the story of the EU and its institutions, general information on the EU	2002-2007	121,149
<b>Annual general report on the activities of the EU</b>	agriculture, budget, customs, economy, education, enlargement, environment, globalization & trade, humanitarian aid, information society, justice, market, mobility, transport, overview of the activities of the institutions, police	2006	81,979
<b>EPSO sample tests</b>	general understanding of EU institutions, verbal & numerical reasoning, audit, law, economics	2006	3,629
<b>TOTAL</b>			<b>206,757</b>

Table 1. Summary of the contents of the EU Corpus

## 4.2 Tools of analysis

I used the lexical frequency and range computer programs by Heatley, Nation and Coxhead (2002) to establish the frequency lists of the EU Corpus and to compare them to the list of the first 3,000 most frequent words in the British National Corpus (BNC). The lists obtained from these programs allowed me to (1) draw up the list of the most frequent lexical elements and within those the most frequent content words in the EU Corpus, (2) examine how the most frequent content words are distributed in the frequency lists of the BNC and which of the most frequent content words in the EU Corpus are not on among the first 3,000 most frequent words of the BNC, and (3) check the EU Corpus for the vocabulary items posted on the Euro-Jargon and EU Glossary websites of the EU.

In order to examine the most frequently occurring content words in context and to reveal their most important collocations and usage patterns, I used the concordance function of the WordSmith Version 2.0 computer program (Scott, 1996). When analyzing the results generated by the concordance and collocation functions, I looked for patterns that were interesting and important from a language learning and teaching point of view. Finally, as one of the main goals of this research was to reveal how words with multiple meanings are used in the EU context, WordNet Version 2.0 (Miller, 2003) was also run on the most frequently occurring content words in order to see which of the different senses of the same word form can be found in the EU Corpus. An example of the semantic analysis is presented in 5.3.

## 5 Results and Discussion

The primary purpose of this corpus-based study is to generate findings that can be directly used in teaching English for 'EU purposes' to non-linguist learners and thereby contribute to English teaching and materials development in this very popular area of language learning. In the discussion that follows, I will point out the most important characteristics of English lexis revealed by the EU English Corpus and I will present examples of the most interesting findings shown by the analyses.

### 5.1 Most Frequent Lexical Elements in the EU English Corpus

The lexical frequency and range programs uncovered that almost half (46.5%) of the lexical elements in the EU Corpus were **not** among the first 3,000 most frequent words of the BNC, and about one third of the types came from the most frequent first 1,000 words in the BNC (see Table 2 for details). A closer examination of the types that were **not** in any of the BNC lists shows that the 30 most frequent are among the first 300 most frequent lexical units in the whole corpus. Therefore, as could be expected, many words in the corpus may be considered as EU-specific and as such are not likely to form part of the general English vocabulary of an average learner of English at intermediate (B1-B2) level.

BNC WORD LISTS	TOKENS (RUNNING WORDS) %	TYPES (INDIVIDUAL WORDS) %
The most frequent first 1000 headwords	74.19%	27.74%
The most frequent second 1000 headwords	10.79%	17.97%
The most frequent third 1000 headwords	2.41%	7.8%
<b>Not in the lists</b>	12.61%	<b>46.5%</b>

Table 2. Summary of the lexical range output

It is important to note that a close comparison between the lexical units that were not found in the BNC lists and the vocabulary lists posted on the *Euro-Jargon and EU Glossary* websites shows little overlap. This is probably due to the fact that the websites designed to disseminate the lexical novelties of the EU concentrate on general elements in the history of the EU and on explaining how the individual institutions work. For example, they contain the expression “*pillars of the EU*” to discuss in which policy areas the EU takes decisions but not the adjective “*sustainable*” which nonetheless occurs 121 times and ranks as the 207<sup>th</sup> most frequent lexical unit in the EU Corpus. Many of the lexical elements not found in any of the three BNC lists above are abbreviations, some of which are enumerated on the websites, while many of them are not (for example: *OJ* which stands for Official Journal and is used very frequently in the corpus ranking as the 55<sup>th</sup> most frequent lexical item). These findings lend empirical support to the relevance of corpus-based investigations in teaching EU English. A combined list of the 50 most frequent lexical occurrences in the four wordlists is tabulated in Appendix A.

## 5.2 Most Frequent Content Words in Context

In the following analyses, I will concentrate on some of the most frequently occurring content words in the corpus. The choice of the particular content words to be included in the present discussion to illustrate the methods of analyses was made 1) on the basis of the frequency of these words in the corpus and 2) on the basis of how many usage patterns they revealed that could be particularly useful for learners of English at intermediate (B1-B2 level). In each of the following sections of the discussion, I will attempt to illustrate how the corpus-based analysis of the EU Corpus can be exploited for pedagogical purposes.

### 5.2.1 Collocations

The *Longman Dictionary of Contemporary English* (1995) contains more than 50 entries of the word *have* while the *Longman Idioms* (2000) and *Phrasal Verbs* (1998) dictionaries contain 17 and 13 entries of the word respectively. Therefore, it was expected that a large number of occurrences of this word would be found in the EU Corpus as well in various collocation patterns.

The primary verb *have* is – not surprisingly – the 10<sup>th</sup> most frequent word in the EU Corpus and the second most frequent verb preceded only by another primary verb *be*. The reason for its frequent use is obviously due to the many grammatical functions that *have* can perform as an auxiliary verb, for example when forming the *perfect* tenses (e.g., *it has been decided ...*), in causative sentences (e.g., *have something done*) or in other constructions (e.g., *have to do something*, *had better do something*, *have someone doing something*, etc.). For the purpose of the present study, however, I focused on *have* as an ordinary verb and examined its collocates with a view to finding out which particular phrases and expressions can be found in this register.

The concordance search proved to be extremely productive and a total of 568 entries were found for the verb *have*. Table 3a shows how the different functions of the verb (auxiliary verb and ordinary verb) are distributed. Table 3b displays the top section of the list of the most common right collocates of *have* by order of frequency based on the concordance listing provided by the WordSmith program.

	<b>ORDINARY TRANSITIVE VERB</b>	<b>AUXILIARY VERB used to form perfect tenses</b>	<b>AUXILIARY VERB HAVE + to + infinitive verb</b>	<b>AUXILIARY VERB CAUSATIVE</b>	<b>TOTAL</b>
<b>TOTAL</b>	329	202	34	3	568

Table 3a. Distribution of the entries of *have* in the EU Corpus

<b>RIGHT COLLOCATES</b>	<b>RAW COUNTS</b>
been	93
a	58
to	52
the	37
an	13
access	10
right	10
impact	2

Table 3b. Most common immediate right collocates of *have* in the EU Corpus

The list of collocates shows that *been* is by far the most common right collocate of the word *have* which confirms the high number of its occurrences as an auxiliary verb. The combined raw counts of the definite and indefinite articles (*the*, *a*, *an*) as right collocates of *have*, however, outnumber the count for *been* which underpins the importance of the use of *have* as an ordinary transitive verb in EU-related texts. A close examination of the concordance lists reveals the extremely productive nature of this verb as regards the collocations in which it is used (e.g., *have access to*, *have an impact on*, *have the right to*, *have the chance to*, *have a better understanding of*, *have recourse to*, etc.). A full list of these collocations can be found in Appendix B. It is important to note that some of these collocations could be easier for learners than others depending on the extent to which they are language specific (e.g., *have access to* is likely to pose more problems for Hungarian learners of English than *have the right to*). This aspect of collocations needs to be taken into consideration when designing course materials.

The results of the concordance search also help to point out how they can be used in context, for example which adjectives appear typically with the collocates of *have*. A cluster analysis of one collocate *access*, shows about 10 adjectives which appear at least once next to it in the corpus (e.g., *free*, *easy*, *broadband*, *public*, *preferential*, etc.). The same analysis also shows that *access* occurs together with the preposition *to* 117 times out of 175, its total number of occurrences. Another important detail that can be brought to learners' attention is the use of the definite and indefinite articles in phrases with *have* which is a common problem for language learners. The concordance search results provide ample examples of use which can easily be incorporated into course materials and give learners the opportunity to learn and practice through real and meaningful examples the points that present the most difficulties for them.

### 5.2.2 Usage Patterns

Concordance listings proved to be extremely helpful in revealing interesting word usage patterns. The abbreviation *EU* is the 10<sup>th</sup> most frequent word in the whole EU Corpus

(see Appendix 1), the only content word preceding it is *European*. Therefore, it is important to examine its use and how this word can be taught to learners of English in detail. Table 4 shows the most frequent left and right collocates of this abbreviation from the concordance list produced by the WordSmith program.

LEFT COLLOCATES	COUNTS IN CORPUS	EU	RIGHT COLLOCATES	COUNTS IN CORPUS
<b>the</b>	<b>1067</b>	+	<b>countries, country</b>	<b>259</b>
all	37	+	's	196
each	31	+	has	80
another	30	+	is	79
other	27	+	and	61
for	20	+	citizens	44
non	19	+	institutions	35
any, new	11	+	law	26
with	9	+	rules	23
many	6	+	leaders	21
under	4	+	budget, governments	20

Table 4. The 10 most frequent left and right collocates of *EU* in the EU Corpus

As is obvious from the list of collocates in Table 4, the abbreviation *EU* can function as a noun (and it is most frequently followed by the verbs *has* and *is* showing that it is the subject of the sentence) and as the first part of a compound noun (its most frequent immediate right collocates are nouns such as *countries*, *citizens*, *institutions*, etc.). The full concordance list also displays a useful set of nouns and verbs that are used together with the abbreviation *EU* and that are worth studying (e.g., nouns: *economy*, *treaties*, *standards*, *membership*, etc., verbs: *funds*, *helps*, *works*, *supports*, etc.).

Selected lines of the results of the cluster analysis are shown in Table 5 below. Included in the table are clusters of words that would be among the most useful from a language teaching point of view. A number of prepositions that can occur with the collocation *the EU* and its use in the possessive case are apparent from the list in Table 5.

within the EU
between the EU ..... (e.g., countries)
throughout the EU ..... (e.g., countries)
and the EU ..... (e.g., countries)
by the EU ..... (e.g., countries)
across the EU ..... (e.g., countries)
from the EU
outside the EU
EU's single market / external borders / neighbours
EU-funded research

Table 5. Cluster analysis on the basis of the concordance listing for *EU* in order of frequency

Based on concordance and cluster searches that reveal the linguistic patterns in which a single word, an abbreviation or collocations may be used, it is easier for the language teacher to make decisions regarding how to present them to the learners and how to compile sample sentences and test materials for given items.

### 5.3 Semantic Analysis

Corpus-based studies for language education purposes often draw the reader's attention to the fact that many corpus-based vocabulary studies tend to oversimplify the results obtained by frequency counts by interpreting them as though frequency of particular word forms was equivalent to frequency of meaning (see e.g., Gardner & Davies, 2007). In fact, the different meanings of the most frequent words in a given corpus are just as important to characterize the corpus itself and to provide useful insights for the language learner and the language teacher into the register under study. The word *trade* is the 17<sup>th</sup> most frequent content word in the EU Corpus occurring a total of 441 times (see Appendix A). *Trade* can have many context-sensitive meanings attested by the 39 different uses of the word and its compounds in the *Longman Dictionary of Contemporary English* (1995). In order to gain a thorough understanding of the different uses of the word in this particular register, I examined the concordance listings by WordSmith, and compared them to the output produced by WordNet (Miller, 2003), an electronic lexical database that distinguishes between the different senses of the same word forms. Table 6 summarizes the outcome of a meaning-based analysis of the word *trade*.

WORD CLASS AND MEANING	NUMBER OF OCCURENCE
noun – the commercial exchange (buying and selling on domestic or international markets) of goods and services (WordNet)	<b>426</b>
noun – people who perform a particular kind of skilled work (WordNet) “trade union(s)”	<b>8</b>
noun – “trade-off”: acceptable balance between two opposing things that you want (Longman Dictionary)	<b>1</b>
noun – “trademark”	<b>3</b>
verb – to buy and sell goods and services, engage in the trade of (WordNet)	<b>3</b>
<b>TOTAL: 441</b>	

Table 6. Different meanings of *trade* in the EU Corpus

The analysis shows that the overwhelming majority of the occurrences of *trade* refer to its most common meaning in general English – commercial exchange of goods and services – in various noun phrases as shown in Table 7 where the most frequent collocates of *trade* are displayed. This shows that some of the most frequent collocations are: *World Trade Organization*, *free trade agreements*, *international trade relations*, *trade policy*, *barriers to trade*, *trade-related*, etc. Learning how these collocations are used in context in a sentence and not only their exact translation into the learners' first language is essential for language learners who need to deal with EU-related documents.

LEFT COLLOCATES	COUNTS	TRADE	RIGHT COLLOCATES	COUNTS
world	58	+	organization	28
free	38	+	policy	19
international	15	+	agreements	18
barriers to	9	+	relations	12
fair, community	5	+	area	9
external	4	+	-related	9
global	3	+	negotiations	7
internal	3	+	unions	8
arms	3	+	rules	6
national	3	+	areas	2
preferential	3	+	links	2

Table 7. The 10 most frequent left and right collocates of *trade* as a noun

Similar analyses are important when compiling course materials for learners of English in the case of words that have multiple meanings and word forms that belong to different word classes.

## 5 Conclusion

The pedagogical implications of corpus-based analyses of language use, and vocabulary in particular, are enormous and there is a considerable body of literature discussing the relevance of corpus studies in language teaching (Frankenberg-Garcia, 2005; Gardner & Davies, 2007). Rising to the linguistic challenges Hungary now faces after having joined the EU in 2004, and while preparing for the upcoming EU Presidency in 2011, is crucial to the country's future success. Corpus-based English language teaching materials would be very useful in designing effective language courses to enhance EU English language skills and proficiency in Hungarian language learners.

In the present study I attempted to show how the lexical aspects of this particular register – English lexis in EU documents – can be researched with the use of concordancing and semantic analysis computer programs to:

- (1) identify the most frequent content and function elements in the corpus,
- (2) present the use of concordance lists in establishing collocation and usage patterns for some of the most frequent content words showing their pedagogical relevance, and to
- (3) describe how the semantic analysis of a frequent word form can be useful in corpus-based materials design.

An examination of the most frequent lexical elements in the corpus revealed that almost half of the words in the corpus were not among the first 3,000 words of the BNC, consequently, the vocabulary of EU documents is likely to pose challenges for average language learners at B1-B2 level. Since the primary purpose of this research project is to contribute to producing learning materials for students of EU English at this proficiency level, this finding warrants further investigations, namely the description of usage patterns, collocations and semantic analyses of word forms with multiple meanings. To illustrate how these types of analyses can contribute to materials design, three lexical items from among the first 50 lexical elements in the corpus (*have*, *EU*, *trade*) were chosen. These items served as examples to demonstrate the use of the various functions of the WordSmith concordancing software (patterns, collocates, clusters, etc) and WordNet in analyzing vocabulary use in the current corpus of EU English.

The findings are very practical in nature and are directly applicable to course design for EU-studies and International Relations English language development programs. Each of the types of analyses presented in this paper can be applied to the lexical elements that the teacher or materials developer deems to be important for their learners. Language teaching practice, the results of a proper needs analysis and, last but not least, common sense can guide teachers and materials developers in choosing the items to be further analyzed and brought to learners' attention, for example by including data-driven learning activities using concordance samples in the syllabus. It is obvious that in order to be able to establish a more elaborate picture of the characteristics of the English language in the documents of the EU, more aspects of language use (e.g., grammatical structures, discourse features) need to be analyzed

and a larger corpus would need to be built which can be the object of further research in this area. Hopefully, this study will make a useful contribution to the growing body of literature on the relevance of corpus-based investigations in language teaching in general and to the teaching of EU English vocabulary in particular.

*Proofread for the use of English by: Gordon Dobson, Department of English Applied Linguistics, School of English and American Studies, Eötvös Loránd University, Budapest*

## References:

- Anthony, L. (1997). *English for Specific Purposes: What does it mean? Why is it different?* Retrieved from <http://www.antlab.sci.waseda.ac.jp/abstracts/ESParticle.html>
- Bart, I., & Klaudy, K. (2003). *EU-fordítóiskola. Európai uniós szövegek fordítása angolról magyarra*. Budapest: Corvina.
- Basturkmen, H. (1998). Refining procedures: A needs analysis project at Kuwait University. *English Teaching Forum*, 36(4).
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics*. Cambridge: Cambridge University Press.
- British National Corpus Online*. Retrieved from [www.natcorp.ox.ac.uk](http://www.natcorp.ox.ac.uk)
- Chilton, P., & Schäffner, C. (1997). Discourse and Politics. In T. A. Van Dijk (Ed.), *Discourse as social interaction* (pp. 207-226). Thousand Oaks, USA: Sage Publications.
- Cullen, R., & Kuo, I-C. V. (2007). Spoken grammar and ELT course materials: A missing link? *TESOL Quarterly*, 41(2), 361-385. <https://doi.org/10.1002/j.1545-7249.2007.tb00063.x>
- Dudley-Evans, T., & St John, M. (1998). *Developments in ESP: A multi-disciplinary approach*. Cambridge: Cambridge University Press.
- EUABC Online Dictionary*. Retrieved from <http://en.euabc.com/>
- European Personnel Selection Office Webpage*. Retrieved from [http://europa.eu/epso/index\\_en.htm](http://europa.eu/epso/index_en.htm)
- Fischer, M. (2007). Fordítás(politika) és terminológia az Európai Unióban [Translation(policy) and terminology in the European Union.]. In *A XVI. Magyar Alkalmazott Nyelvészeti Kongresszus előadásai*. [Proceedings of the 16th Applied Linguistics Congress] Vol. 3. (pp. 806-811). Pécs-Gödöllő: MANYE-Szent István Egyetem.
- Frankenberg-Garcia, A. (2005). Pedagogical uses of monolingual and parallel concordances. *ELT Journal* 59(3), 189-198. <https://doi.org/10.1093/elt/cci038>
- Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, 41(2), 339-359. <https://doi.org/10.1002/j.1545-7249.2007.tb00062.x>
- Gilmore, A. (2004). A comparison of textbook and authentic interactions. *ELT Journal*, 58(4), 363-374.
- Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range and frequency programs software. Retrieved from [http://www.vuw.ac.nz/lals/staff/Paul\\_Nation](http://www.vuw.ac.nz/lals/staff/Paul_Nation)
- Jablonkai, R., Erdei, J., Fekete, É., Homolya, K., & Nagy, É. (2006). *Angol-Magyar*

- Tematikus Gazdasági Szókinestár* [English-Hungarian thematic dictionary in economics]. Budapest: Aula Kiadó.
- Károly, K. (2007). *Szövegten és fordítás* [Discourse analysis and translation]. Budapest: Akadémiai kiadó.
- Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly*, 37(4), 671-700.  
<https://doi.org/10.2307/3588217>
- Longman Dictionary of Contemporary English*. (1995). Harlow: Longman.
- Longman Idioms Dictionary*. (1998). Harlow: Longman.
- Longman Phrasal Verbs Dictionary*. (2000). Harlow: Pearson Education Limited
- Manning, C. D., & Schütze, H. (1999). *The Foundations of statistical natural language processing*. Cambridge, Mass.:The MIT Press. Retrieved from  
<http://nlp.stanford.edu/fsnlp/promo/colloc.pdf>
- Miller, G. (2003). WordNet (Version 2.0) software. Retrieved from  
<http://www.cogsci.princeton.edu/wn/obtain.shtml>
- Robinson, P. (1991). *ESP today: A practitioner's guide*. Hemel Hempstead: Prentice Hall.
- Scott, M. (1996). *WordSmith tools*. Oxford University Press, Oxford.
- Smoak, R. (2003). What is English for Specific Purposes? *English Teaching Forum*, 41(2). Retrieved from [https://americanenglish.state.gov/files/ae/resource\\_files/03-41-2-g\\_0.pdf](https://americanenglish.state.gov/files/ae/resource_files/03-41-2-g_0.pdf)
- Trosborg, A. (1997). Translating hybrid political texts. In A. Trosborg (Ed.), *Text typology and translation* (pp. 145-159). Amsterdam, Philadelphia: John Benjamins.
- West, R. (1994). Needs analysis in language teaching. *Language Teaching*, 27(1), 1-19.  
<https://doi.org/10.1017/s0261444800007527>
- WordSmith lexical analysis tools software. Oxford: OUP. Retrieved from  
<http://www.lexically.net/wordsmith/>
- Wright, C. (1992). *The Benefits of ESP*. Published online by the Cambridge Centre for Languages. Retrieved from <http://www.camlang.com/articles.cfm>

## APPENDIX A

## List of the 50 most frequent words in the EU Corpus

Word Type	Rank	Frequency	Cumulative Percent
THE	1	16648	8.43
AND	2	7558	12.25
OF	3	7492	16.05
TO	4	5586	18.87
IN	5	4883	21.35
A	6	3713	23.22
ON	7	2847	24.67
FOR	8	2661	26.01
EUROPEAN	9	2256	27.16
<b><i>EU</i></b>	<b><i>10</i></b>	<b><i>1935</i></b>	28.13
IS	11	1822	29.06
BY	12	1271	29.70
WITH	13	1268	30.34
IT	14	1226	30.96
AS	15	1144	31.54
UNION	16	1111	32.10
COMMISSION	17	1110	32.67
ARE	18	1081	33.21
THAT	19	1046	33.74
ITS	20	877	34.19
S	21	869	34.63
THIS	22	835	35.05
WHICH	23	816	35.46
COUNTRIES	24	805	35.87
BE	25	787	36.27
FROM	26	778	36.66
AT	27	739	37.04
OR	28	725	37.40
COUNCIL	29	714	37.76
HAS	30	642	38.09
MEMBER	31	637	38.41
AN	32	630	38.73
STATES	33	592	39.03
ALSO	34	591	39.33
THEIR	35	570	39.62
NEW	36	568	39.91
<b><i>HAVE</i></b>	<b><i>37</i></b>	<b><i>567</i></b>	40.19
WAS	38	557	40.48
WILL	39	522	40.74
MORE	40	517	41.00
ALL	41	510	41.26
EUROPE	42	508	41.52
ADOPTED	43	474	41.76
OTHER	44	470	41.99
POLICY	45	466	42.23
FI	46	461	42.46
NOT	47	460	42.70
CAN	48	459	42.93
THEY	49	448	43.16
<b><i>TRADE</i></b>	<b><i>50</i></b>	<b><i>426</i></b>	43.37

In **bold** and *italic* characters are the lexical units analyzed in the present study.

**APPENDIX B**List of collocations with *have***HAVE + INDEFINITE ARTICLE + NOUN**

have an accident  
 have an advantage  
 have an agreement  
 have an approach  
 have an arrangement  
 have a bearing on something  
 have a budget (of)  
 have a choice  
 have a complaint  
 have a dispute  
 have a duty  
 have a right (to)  
 have a role (to play)  
 have an understanding (of something)  
 have an idea (of something)  
 have an image (of something)  
 have an impact (on something)  
 have an influence (on something)  
 have an involvement (with something)  
 have a job  
 have a legal status  
 have a link (with something)  
 have a need  
 have an effect (on something)  
 have an objective  
 have a part  
 have a problem (with something)  
 have a responsibility  
 have a strategy  
 have a tool to  
 have a value  
 have a view (of something)

**HAVE + NOUN**

have access (to)  
 have confidence (in something)  
 have faith (in something)  
 have health insurance  
 have leeway (to)  
 have money (to)  
 have protection (against)  
 have (no) power (to)  
 have reason (to)  
 have recourse (to something)  
 have resources  
 have strength  
 have (high/low) standards

**HAVE + DEFINITE ARTICLE + NOUN**

have the ability (to)  
 have the advantage (of)  
 have the chance (to)  
 have the honour (to)  
 have the right (to)

have the opportunity (to)  
have the objective (to)  
have the potential (to)  
have the competence (to)  
have the qualifications (to)  
have the use (of something)

<b>OTHER CONSTRUCTIONS</b>
----------------------------

have your say  
have something in place  
have something with you