

DOCTORAL DISSERTATION

RÉKA RÁTKAINÉ JABLONKAI

A corpus-linguistic investigation into the lexis of written English EU discourse:

An ESP pedagogic perspective

2010

Language Pedagogy PhD Program  
Programme director: Péter Medgyes, DSc  
Director of research: Krisztina Károly, PhD, habil.  
Director of studies: Dorottya Holló, PhD, habil.

Faculty of Pedagogy and Psychology, Eötvös Loránd University, Budapest

RÉKA JABLONKAI

A corpus-linguistic investigation into the lexis of written English EU discourse:

An ESP pedagogic perspective

Supervisor: Károly Krisztina, PhD, habil.

Members of the Dissertation Examination Committee:

Kinga Klaudy, DSc, Professor  
Pál Heltai, CSc, Associate Professor  
Tamás Váradi, CSc, Senior Researcher  
Dorottya Holló, PhD, habil., Associate Professor  
Réka Eszenyi, PhD, Lecturer  
Ildikó Lázár, PhD, Lecturer  
Gyula Tankó, PhD, Lecturer

Date of submission: 2010

## Table of Contents

Acknowledgements	7
Abstract	8
List of abbreviations	9
Chapter 1: Introduction	10
Chapter 2: Theoretical framework	15
2.1. Analysis of language variation	19
2.2. Research and pedagogical perspectives within the study of ESP	23
2.2.1. What is ESP?	23
2.2.2. Classification of ESP	26
2.2.3. Historical overview of the study of ESP	28
2.2.4. Theoretical influences on the study of ESP	31
2.2.5. Approaches to teaching ESP	34
2.2.5.1. Focus on learning	35
2.2.5.2. Focus on needs	38
2.2.5.3. Focus on skills and strategies	39
2.2.5.4. Focus on the discipline	40
2.2.5.5. Focus on language	41
2.2.6. Lexis in ESP	43
2.2.6.1. Categories of lexis in specialised texts	45
2.2.6.2. Multi-word items in ESP	46
2.2.6.3. Teaching lexis in ESP	48
2.2.7. Course and materials design in ESP	49
2.3. Lexis in the centre of course design	50
2.4. Summary	52
Chapter 3: Earlier analyses of EU discourse	53
3.1. Language in the EU – eurojargon and hybridity	53
3.2. English EU discourse in an ESP context	54
3.3. Summary	56
Chapter 4: Methodological framework: corpus linguistics	58
4.1. A corpus linguistic approach to text analysis	58
4.1.1. Brief history	59
4.1.2. Key concepts of corpus linguistics	62

4.1.2.1.	What is a corpus?	63
4.1.2.2.	Kinds of corpora	66
4.1.2.3.	Concordance, collocation and annotation	70
4.1.3.	Units of analysis in corpus research	71
4.1.4.	Approaches to corpus analysis	72
4.1.5.	Reasons for using corpora in linguistic research	74
4.1.6.	Fields enjoying the benefits of corpus linguistics	76
4.1.7.	Criticism of corpus linguistic studies	77
4.1.8.	Summary	80
4.2.	Data-driven learning	81
4.3.	Corpus research and ESP	82
4.3.1.	Analysis of the language of specific disciplines	84
4.3.1.1.	Word lists in ESP	85
4.3.1.2.	Key word analysis	90
4.3.1.3.	Collocational analysis in ESP	91
4.3.2.	Analysis of specific contexts	98
4.3.3.	Analysis of specific discourses	101
4.3.3.1.	Academic discourse	102
4.3.3.2.	Functions and structures of lexical bundles	104
4.3.3.3.	Professional discourse	106
4.3.4.	Analysis for course and materials design	106
4.3.5.	Analysis across languages	109
4.3.6.	Summary	110
4.4.	Issues in corpus design for ESP	111
4.4.1.	Guiding principles for corpus design	112
4.4.2.	An overview of the steps of corpus design and creation	114
4.4.3.	A Model for Corpus Creation for ESP	116
4.4.4.	Purpose, the guiding thread	119
4.4.5.	Representativeness	120
4.4.5.1.	Sampling methods	121
4.4.5.2.	Structure of the corpus	124
4.4.6.	The bigger the better?	125
4.4.6.1.	Small can be beautiful too	129
4.4.6.2.	Sample size	130

4.4.7.	Data collection	132
4.4.8.	Data entry	132
4.4.9.	Ethical and legal issues	134
4.4.10.	Summary	135
Chapter 5: Aims and research questions		136
Chapter 6: Research design and procedures of analysis		138
6.1.	Stage 1: Needs analysis and corpus creation	138
6.2.	Stage 2: Corpus analysis	139
6.3.	Design and compilation of the English EU Discourse Corpus	141
6.3.1.	Defining written English EU discourse	143
6.3.2.	Needs analysis survey for corpus design	143
6.3.2.1.	Questionnaire construction	144
6.3.2.2.	Establishing the validity and reliability of the instrument	146
6.3.2.3.	Participants	154
6.3.3.	Conducting the needs analysis survey	158
6.3.4.	Methods of needs survey data analysis	158
6.4.	Procedures and tools of corpus analysis	159
6.4.1.	Step 1: Establishing the EU Word List	159
6.4.1.1.	The notion of the word family	160
6.4.1.2.	Compilation of the English EU Word List	162
6.4.2.	Step 2: Collocational analysis of selected lexical items	166
6.4.2.1.	Selection of lexical items for collocational analysis	166
6.4.2.2.	Analysis of collocational patterns in the EEUD Corpus	169
6.4.3.	Step 3: Analysis of lexical bundles in the EEUD Corpus	170
Chapter 7: Results and discussion		176
7.1.	Results of the needs analysis	176
7.1.1.	How EU documents are used by professionals	179
7.1.2.	Description of the EEUD Corpus	182
7.1.2.1.	Sources of texts in the EEUD Corpus	182
7.1.2.2.	Balance for EU subject fields	183
7.1.2.3.	The time period represented by the EEUD Corpus	185
7.1.2.4.	EU institutions represented in the EEUD Corpus	186
7.1.3.	Limitations of the needs analysis and the EEUD Corpus	187
7.1.4.	Conclusions concerning the EEUD Corpus	188

7.2.	Results of the corpus analysis	189
7.2.1.	The EU Word List	189
7.2.1.1.	Evaluation of the EUWL	192
7.2.1.2.	Limitations of the EUWL	196
7.2.1.3.	Conclusions concerning the EUWL	196
7.2.2.	Collocations in the EEUD Corpus	198
7.2.2.1.	Grammatical behaviour of selected lemmas	199
7.2.2.2.	Semantic preference and semantic prosody of selected lemmas	205
7.2.2.3.	Limitations of the collocational analysis	209
7.2.2.4.	Conclusions concerning the collocational analysis	211
7.2.3.	Lexical bundles in the EEUD Corpus	212
7.2.3.1.	Structural analysis of lexical bundles in the EEUD Corpus	215
7.2.3.2.	Functional analysis of lexical bundles in the EEUD Corpus	219
7.2.3.2.1.	Stance bundles in the EEUD Corpus	223
7.2.3.2.2.	Discourse organisers in the EEUD Corpus	224
7.2.3.2.3.	Referential bundles in the EEUD Corpus	225
7.2.3.2.4.	Subject-specific bundles in the EEUD Corpus	228
7.2.3.3.	Length of lexical bundles in the EEUD Corpus	230
7.2.3.4.	Limitations of the analysis of lexical bundles	232
7.2.3.5.	Conclusions concerning the analysis of lexical bundles	233
7.2.4.	Lexis in the EEUD Corpus	234
	Chapter 8: Implications for teaching English for EU purposes	238
	Chapter 9: Conclusion	253
	References	258
	Appendices	278

## **Acknowledgements**

I am especially grateful to my supervisor Dr. Krisztina Károly for her guidance throughout my PhD course and for her useful insights and invaluable comments on earlier drafts of the dissertation.

I would like to thank Anna Trebits for providing me with her EU English Corpus and valuable comments on the findings. In a similar vein, thanks also go to Márta Fischer for commenting on the results, and offering valuable insights from the perspective of terminology.

I would also like to express my gratitude to the anonymous respondents to the questionnaire survey, and the interviewees, without whom it would not have been possible to base my research on insights into the actual language use of EU professionals.

I am also grateful to the British Council Hungary for their financial support which made a visit to the library of the University of Birmingham possible, and where I had access to the literature on corpus linguistics.

Thanks to Éva Nagy for offering her time to help me with the data analysis.

I would also like to thank Bálint Sass for drawing my attention to corpora that can be used for English language teaching practice in a Hungarian context.

I also wish to express my thanks to Dr. Enikő Csomay for introducing the concept of lexical bundles to me in her corpus linguistics course.

I am also indebted to my family, my husband and two daughters, for their patience and encouragement throughout my doctoral studies.

## **Abstract**

The aim of this dissertation is to extend research into the use of English in the context of the European Union. As previous studies have mainly focused on language policy, translation and terminology issues, and as there is little research into the English language use within the European Union for ESP pedagogic purposes, the specific goal of this study is to explore the discourse of written English EU documents with language learners in mind. In order to gain a comprehensive picture of this particular variety of English, the approach and methods of corpus linguistics have been found appropriate, given its focus on real language use and tools that allow the analysis of a large number of texts. Therefore, the, so called, English EU Discourse Corpus (EEUD Corpus) was compiled based on a needs analysis survey among members of the EU discourse community, as a starting point for further investigation. The corpus analysis concentrated on the frequent lexical items, their collocational behaviour, and frequent multi-word items. The investigation of frequently used lexical items applied the notion of the word family, and resulted in the EU Word List, with 513 word families frequently used in English EU texts. The results of the collocational analysis of a few selected lemmas show marked differences in the behaviour of the analysed lexical items in a general corpus, the BNC Written, and the specialised EEUD Corpus. Finally, the analysis of frequently used multi-word items shows the tendency of written English EU discourse – as represented by the EEUD Corpus – to apply a large number of lexical bundles in high frequencies; this suggests that a fairly large proportion of EU texts is made up of formulaic patterns. These findings, on the one hand, provide a clearer understanding of the special characteristics of EU discourse or ‘eurojargon’; and, on the other hand, they can serve as the basis for sound course and materials design for EU English courses. The study also provides sample tasks, in order to demonstrate how the results can be utilised for the actual ESP teaching practice.



## List of abbreviations

adj	adjective
AWL	Academic Word List
BE	Business English
BEC	Business English Corpus
BNC	British National Corpus
CLIL	Content and Language Integrated Learning
DDL	data-driven learning
EAP	English for Academic Purposes
EEP	English for Educational Purposes
EEUD Corpus	English EU Discourse Corpus
EFL	English as a Foreign Language
EGAP	English for General Academic Purposes
EGBP	English for General Business Purposes
EOP	English for Occupational Purposes
ESP	English for Specific Purposes
ESPj	English for Specific Purposes journal
EST	English for Science and Technology
EUWL	EU Word List
L1	first language
L2	second language
LOB Corpus	London–Oslo/Bergen Corpus
LDCE	Longman Dictionary of Contemporary English
MAWL	Medical Academic Word List
MJ score	mean joint frequency and importance score
MWI	multi-word item
POS tag	part-of-speech tag
SEWL	Student Engineering Word List
SFL	Systemic Functional Linguistics

## Chapter 1: Introduction

The accession of Hungary to the European Union raised several questions in language-related issues. These are probably most obvious in relation to translation and interpreting (Dróth, 2000; Klaudy, 2001; Pym, 2000). Language use within the European Union, however, has been the topic of studies focusing on language policy, translation and terminology as well (Fischer, 2006, 2007; McArthur, 2003; Truchot, 2002). Several studies have been published on the language varieties found in EU documents, mainly from the point of view of the translator. These analyses revealed that EU texts in many official languages have their characteristic syntax, lexis, terminology, and particular style (Born & Schütte, 1995; Dróth, 2000; Fischer, 2006, 2007; Károly, 2007; Klaudy, 2001; Pym, 1993; Pym, 2000; Schäffner & Adab, 2001a, 2001b).

According to the language policy of the EU, all documents should be made available in the twenty-three official languages (*Council Regulation (EC) No 1791/2006*). In practice, however, due to time and financial constraints, documents are produced first in one or just some of the working languages of the EU, and very often EU documents are issued first in English (Truchot, 2002). Moreover, according to EU statistics, English as a source language accounted for 75% of all pages translated in 2009 (EU DG Translation Homepage [http://ec.europa.eu/dgs/translation/index\\_en.htm](http://ec.europa.eu/dgs/translation/index_en.htm)). In a broader context, English is the language used worldwide as the international *lingua franca*, especially in business contexts (Kachru, 1985; Louhiala-Salminen, Charles, & Kankaaranta, 2005; Nickerson, 2005a; St John, 1996), and, increasingly, among EU member states as well. Truchot's (2002) findings on the proportion of texts drafted initially in English within EU institutions clearly demonstrate the rise of the use of English, not only in communication between member states, but also in internal communication within EU institutions, especially in written

communication. English gaining more and more ground within the EU as the *lingua franca* necessitates preparing future Hungarian EU professionals for the use of English within the EU context. Therefore, issues such as the comprehensive analysis of the variety of the English language used within the institutions of the EU, and teaching materials for EU English courses, need to be addressed in research, especially in light of the preparation for Hungary's upcoming Presidency of the Council of the European Union.

Most studies on written English EU discourse have focused on legal documents (e.g., Favretti, Tamburi, & Martelli, 2001). This is not surprising, as most of the translation work involves translating the EU *acquis communautaire*, the body of EU legislation. However, the characteristics of EU legal language alone would not suffice as a basis for English language courses for the EU.

There have only been a handful of studies investigating English EU documents specifically for pedagogic purposes. Tribble (2000) analysed one specific EU genre, that is, proposals for EU funding, in order to draw conclusions regarding the writing skills development necessary for writing such difficult texts in English. Trebits (2008, 2009a, 2009b) investigated particular lexical items like *EU* and *trade*, conjunctions and phrasal verbs in English EU documents, in order to formulate conclusions on the importance of lexis in teaching EU English courses. Although both authors examined relevant genres and linguistic aspects of English EU discourse, in order to provide future EU professionals with appropriate training in English, a more comprehensive approach to English EU discourse is needed. Furthermore, a comparison of the registers of newspaper articles on EU-related issues and official EU documents, revealed that there are significant differences between the two written registers, in terms of their lexis and discourse (Jablonkai, 2009a). This also implies that, instead of focusing on one specific genre or a few specific lexical items, the analysis should

have a broader scope including many EU genres, and examining several linguistic features of English EU documents in general.

Consequently, one of the aims of the present study is to identify EU genres that may be regarded as representative of EU discourse in English, and to construct a corpus referred to as the English EU Discourse Corpus (EEUD Corpus), which is appropriate for such a comprehensive analysis of written English EU discourse. Secondly, based on the investigation of the EEUD Corpus, the study intends to identify the lexical and lexicogrammatical features that are typical of English EU documents, and can thus form the basis of EU English courses in programmes of EU studies, as well as for occupational purposes within an EU context. In order to achieve these aims, the present study takes a comprehensive view of written English EU discourse, covering several genres, and all EU subject fields ranging from monetary policy to foreign and security policy. The results of the analysis will be used to formulate theoretical implications for the study of professional registers in general, and the written English EU discourse, in particular. Furthermore, conclusions will be drawn on the pedagogical implications for course design and materials development in teaching English for EU purposes.

The broad research questions formulated to guide the study of the lexis of written English EU discourse are as follows:

1. What genres can be regarded as representative of written English EU discourse?
2. What lexical items are typically associated with written English EU discourse?
3. What implications do the findings have for course and materials design in teaching English for EU purposes?

The dissertation is divided into nine chapters. Following the introductory chapter, Chapter 2 provides the theoretical framework for the study, focusing on text analysis and the analysis of language variation, as well as, researching and teaching English for Specific

Purposes (ESP). Firstly, a brief theoretical overview of the analysis of language variation is given, outlining relevant aspects of register analysis, and defining the register under study. Then, as the present study draws heavily on earlier findings in research and teaching ESP, a detailed discussion of theoretical influences, approaches and practical elements, such as lexis and course and materials design in ESP is given.

Chapter 3 presents the findings of earlier analyses of EU discourse. The main characteristics of EU texts are discussed, and the notion of the 'hybrid text' (Schäffner & Adab, 2001a, 2001b) – the text type proposed for characterising EU texts – is defined. The chapter also summarises the main findings of analyses of English EU texts for pedagogic purposes.

The methodological approach used in the current analysis draws on corpus-linguistic research. Therefore, Chapter 4 focuses on the most important theoretical and practical considerations of corpus linguistics in text analysis. The chapter also highlights the benefits of the empirical stance that is characteristic of corpus research for several fields in general, and for ESP and language teaching, in particular. Finally, relevant issues of corpus design and corpus building for ESP are discussed, and a *Model for Corpus Creation for ESP* is proposed.

The aims and research questions guiding the present investigation are summarised in Chapter 5, and an overview of the research design is given in Chapter 6. The research procedures are discussed in detail in Sections 6.3 and 6.4, outlining the two main stages of the research, that is, the corpus design and corpus building stage, and the corpus analysis stage. Section 6.3 describes how the proposed *Model for Corpus Creation for ESP* was applied as the theoretical and practical foundation of the corpus design and compilation stage of the study. Section 6.4 describes the three main procedures of corpus analysis, namely, the selection of lexical items particularly associated with written English EU discourse, the investigation of collocations of selected lexical items, and the frequency-based analysis of

multi-word items (MWI). The results of the investigation are presented and discussed in Chapter 7 in several sections, each section focusing on both the results and the limitations of the different stages and procedures of the research. The chapter ends with a summary of the main findings of the investigation by highlighting the principle characteristics of the lexis of official English EU documents.

Chapter 8 discusses the pedagogical implications of the study by pinpointing aspects of written English EU discourse that are relevant for teaching, and proposing practical ways of applying the findings of the current study in the ESP teaching practice.

Finally, the main conclusions of the study are drawn in Chapter 9, outlining the contribution of the present research to corpus linguistics, to register analysis in ESP, to a genre-based approach to ESP, and to ESP pedagogy. Suggestions for further research are also discussed.

## Chapter 2: Theoretical framework

The general aim of ESP has always been to serve the specific needs of learners. Needs in ESP have been defined as a multi-faceted concept including (a) professional, personal and linguistic information about the learners, (b) information about the environment and objectives of the particular course, (c) effective ways of learning the particular skills and language, and (d) information about professional communication, that is, how language and skills are used in their specific professional contexts (Dudley-Evans & St John, 1998, p. 125; Hutchinson & Waters, 1987, p. 53-64). In this respect, learners' linguistic needs can be best described by the linguistic analysis of the language used for communication in their workplaces. As a consequence, at a higher level of abstraction, ESP research has always shown particular interest in varieties of language used in specific situations, and in specific contexts. The particular situations and contexts that the present study focuses on concerns the communication within the institutions of the EU, and in a wider sense the communication in the larger EU context, which includes communication between member states, and also between member states or citizens and the EU institutions.

Varieties of language used in specific situations and in specific contexts are referred to as 'registers' in linguistics (Atkinson & Biber, 1994; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Eggins, 2004; Halliday, 1978; Leckie-Tarry, 1993). Studies investigating particular registers take an empiricist approach to linguistic analysis (e.g., Atkinson & Biber, 1994; Biber et al., 1999; Halliday, 1978). Following the Firthian tradition, this means that "the actual language text duly recorded is in the focus of attention" (Firth, 1968, p. 173). Several fields within linguistics, such as text analysis, discourse analysis, register analysis, genre analysis and corpus analysis, focus on the text as their main object of study. The perspective and approach these fields take focus on different aspects of text in general. As a consequence,

the terms ‘discourse’, ‘text’, ‘register’ and ‘genre’ are interpreted in several different ways in the literature. Therefore, this chapter will start by making a theoretical distinction between these fundamental concepts, and it will also outline the way they are understood in the present study.

This study is based on linguistic theories with an empirical view of language that focuses on its social functions in different contexts. The new concepts of **text** and **discourse**, **register** and **genre** have been introduced and discussed in the literature in order to define and delimit the scope of such empirical analyses. However, as pointed out in the literature of text analysis and discourse analysis (de Beaugrande & Dressler, 1983; Károly, 2007; Trosborg, 1997a), the terms ‘text’ and ‘discourse’ are very often used interchangeably by researchers. Some scholars refer to written language as text and to spoken language as discourse (Coulthard, 1985); others, on the other hand, argue that discourse refers to both written and spoken language, whereas text refers exclusively to written language (Sanders & Sanders, 2006). Approaching the two concepts from a communication studies perspective, researchers like Cook (1989) and Widdowson (1996) consider text as a physical product of discourse, whereas discourse, in their view, is the process leading to the text. The approach taken in the present study follows de Beaugrande’s (1997) distinction. According to him, “if we define a text as a communicative event, a discourse would be a set of interconnected texts, the primary instance being the conversation” (p. 21). Following on from this, the term **EU discourse** is used here to refer to all written and spoken instances of language use within an EU context, and the term **EU text** is applied to individual documents issued by EU institutions, such as the Treaty on European Union, Press release IP/08/83, and Community guidelines on state aid to maritime transport, as the main focus in the present study is on written EU discourse.

The concepts of register and genre have been introduced to text and discourse analysis in order to identify types of discourse or text (de Beaugrande, 1993). The terms register and



genre are, however, used inconsistently in the literature. Several articles on text typology attempted to clarify the distinction between the two concepts (e.g., Eggins, 2004; Leckie-Tarry, 1993; Trosborg, 1997a). Trosborg (1997a) concluded that **registers** were situationally defined language varieties, whereas **genres** were distinguished by the situation and their respective communicative purposes. In a similar fashion, Leckie-Tarry (1993) summarised the difference between the two concepts as follows:

The term 'register' tends to be the more neutral, generalized and embracing term, having a wider currency in the language teaching area, and a stronger historical basis. It tends to suggest a focus on the linguistic side of the text-context paradigm, on patterns of lexis and syntax rather than on discourse structure or textual organization, and on sections of discourse smaller than the whole text. 'Genre', in contrast, had the force of suggesting the priority of the context as a 'conventionalized occasion' over linguistic forms and patterns, the text as a complete event, with formalized organizational schemata. (p. 40)

Furthermore, Trosborg (1997a) pointed out that "registers are divided into genres" (p. 6). She illustrated the relationship between the two concepts by the example of the legal register:

The legal register may comprise the language of the law in legal documents (legislative texts, contracts [...]), the language of courtroom (e.g. the judge declaring the law [...]), the language of legal textbooks, and various types of lawyers' communication with other lawyers and with laymen. Only in the case of restricted registers is there a close relationship between register and genre. (p. 7)

The definition of 'genre' most widely accepted and applied in ESP was given by Swales (1990). According to this:

A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choices of content and style. (p. 58)

Based on this definition, examples of written **EU genres** include, among others, treaties, press releases, reports, presidency conclusions and grant agreements.

The present study, however, defines its object of study at a higher level of generality. It aims to examine the language variety in the EU context, including several genres, rather than focusing on one particular genre. Therefore, the concept of ‘register’ was found appropriate to define the language variety under study. Furthermore, Atkinson and Biber (1994), in order to delimit the scope of register studies, identified four characteristics of studies focusing on a register. According to these characteristics, **register studies** are descriptive analyses of “actually occurring discourse”, which aim to “characterize language varieties”, and describe language varieties in a formal linguistic way, and to analyse the situational characteristics of these varieties (Atkinson & Biber, 1994, p. 352).

In light of the above, the approach taken in this study may be claimed to fall within the area of register analysis, as it aims to:

- describe the “actually occurring discourse” of EU institutions as reflected in written English EU documents,
- characterise the variety of English used in official EU documents,
- present a formal linguistic characterisation of written English EU discourse primarily at the lexical and lexicogrammatical level.

Moreover, from a language pedagogic point of view, this study relies on previous research in the field of Language for Specific Purposes (LSP) and, within that, especially ESP. Register analysis for language teaching purposes in ESP dates back to the 1960s. Although this approach was abandoned later on, it has seen a revival with the emergence of computer technology in linguistic research (Dudley-Evans & St John, 1998).

Thus, the theoretical framework of this study is a synthesis of relevant theories in the following fields: analysis of language varieties, lexis in professional and occupational discourse, and research in ESP, especially with a focus on lexis and course and materials design. In addition, since the current research applies a corpus methodology, it also makes use of theories underlying corpus linguistics. The following chapters will outline the relevant findings in these fields, which served as a framework and a starting point for the present study.

## 2.1. Analysis of language variation

Halliday, McIntosh and Strevens (1964) devised a framework for the study of language varieties. They divided language into user-related varieties such as geographical, temporal, social dialects, and use-related varieties known as **registers**. Furthermore, in his later works, Halliday (1978) identified three main descriptive categories of register that are **field** (the activity or topic), **tenor** (the social relations between the speakers), and **mode** of discourse (channel of communication, predominantly spoken versus written). As Eggins (2004) pointed out:

These three dimensions [...] are used to explain our intuitive understanding that we will not use language in the same way to write or to speak (mode variation), to talk to our boss as to talk to our lover (tenor variation), and to talk about linguistics as to talk about jogging (field variation). (p. 9)

Along these dimensions, registers can be identified at any level of generality (Biber, 1995; Biber et al., 1999). Very general registers, for example, written and spoken varieties of a language, are defined by a single dimension, in the case of the example by their mode. On the other hand, at a very low level of generality, registers are classified in terms of all three dimensions with very specific parameters; for example, preamble sections of EU treaties or press releases of the European Central Bank, etc.

In order to further delimit the register under scrutiny, in the present research the concept of **discourse community** was applied. The concept is used in the field of research into academic discourse to define the writers and readers of a particular discipline sometimes described as an “academic tribe” (Hyland, 2000, p. 8). Swales (1990) gave six defining criteria for a **discourse community**: (1) common public goals, (2) mechanisms of intercommunication, (3) ways to provide information and feedback, (4) special genres, (5) specific lexis, and (6) threshold level of content and discursual expertise (p. 26). According to these criteria, professionals of a particular discipline, subject field or occupation form discourse communities. As a consequence, professionals like experts, translators, etc., working in the EU context can be considered a discourse community as they have their common goals, specific ways of communication, special EU genres, specific EU lexis, and they need to be familiar with certain EU concepts.

On the basis of the above, the language in EU documents was considered a register reflecting the lexical and structural choices specific to the discourse community of professionals working in the EU context. Thus, the particular register under study can be defined in terms of Halliday’s categories (1978) as follows:

- a) field of discourse: European Union,
- b) tenor of discourse: formal,
- c) mode of discourse: written.

Halliday, McIntosh and Stevens (1964) claimed that “registers [...] differ primarily in form” (p. 87). More precisely, they argued that registers can be described in terms of their lexis and grammar. Moreover, they suggested that differences in grammar were “less striking” (p. 89) whereas lexical characteristics were “most obvious” (p. 88), as in some cases a few lexical items are sufficient to identify a register, for example, the lexical item “tablespoonful” is characteristic of recipes and prescriptions, or technical terms are signals of particular technical registers. They also pointed out that “often it is not the lexical item alone but the collocation of two or more lexical items that is specific to one register” (p. 88).

Besides the investigation of the grammatical and lexical level in register studies, Halliday (1966) also argued for a lexicogrammatical level for describing particular linguistic varieties. The existence of collocations, where lexical and grammatical restrictions intertwine, was the main argument in support of the lexicogrammatical level. With the help of computerised corpora and automatized corpus analysis it became possible to investigate collocations on a larger scale. Based on corpus analyses of huge corpora, Sinclair (1991) also found that lexis and grammar are interdependent, and that lexical and structural choices correlate. Therefore, the present study took a lexically oriented approach to the description of written English EU discourse, by focusing on the lexical and the lexicogrammatical levels.

Language varieties have been the subject of several fields of linguistics, for example, sociolinguistics, systemic linguistics and applied linguistics. In sociolinguistics researchers investigate the language variation of, for example, different social classes, gender or age (e.g., Atkinson & Biber, 1994); in systemic linguistics researchers analyse language use in different contexts described according to field, mode and tenor (e.g., Halliday, 1976, 1978); and in applied linguistics it has been used to help determine the content and methodology of language teaching and translation training programmes (e.g., Biber et al., 1999; Biber, Conrad, & Reppen, 1998; Conrad, 1996). Among the areas of interest for this study are

investigations into occupational and professional varieties of language, and studies with a language teaching focus, especially teaching ESP.

Atkinson and Biber (1994) mentioned several studies analysing different professional registers in their review of register analyses. The professional fields investigated from a language variation point of view include law, medicine, science, media, business, bureaucracy and schooling. These analyses involved lexical, syntactic and discourse characteristics of the language and communication of the fields. Most of these studies describe distinctive features of the discourse of their professional fields, for example, binominal expressions like *on behalf of* as a distinctive marker for legal language (Gustaffson as cited in Atkinson & Biber, 1994, p. 354) and sentence complexity in Nigerian “bureaucrats” (Longe as cited in Atkinson & Biber, 1994, p. 356).

In applied linguistics, register analysis with a language teaching focus resulted in, for example, reference books in grammar. *The Longman Grammar of Spoken and Written English* (Biber et al., 1999) is based on corpus-based research into four major registers, namely, conversation, fiction, news and academic prose. The book is based on the findings of the lexicogrammatical analysis of the Longman Spoken and Written English Corpus, which contains 40 million words of text. Instead of a traditional grammar-based approach, the approach taken in this book provided learners with “both grammatical associations of lexical words and lexical associations of grammatical structures” (Altenberg & Granger, 2001, p. 5). In addition, the *Longman Grammar* also discussed the relevance of certain lexical, structural and lexicogrammatical features in different registers.

Moreover, in her register study Conrad (1996) examined three types of academic texts, namely, research articles in ecology, extracts from composition textbooks and textbooks used in ecology courses. She found that research articles and textbook extracts used the impersonal style in a different way, but they both lacked narrative features. Based on her investigation she

suggested that research into registers should be conducted to help teachers decide which texts they should introduce their students to in order to provide them with the necessary linguistic patterns in academic writing and reading courses. In a similar fashion, analysis of the linguistic patterns of texts from different academic disciplines can also provide valuable insights into the language use of a particular discipline (Biber et al., 1998).

At a theoretical level, the main conclusion that can be drawn from earlier research into language variation is that the concepts of register and genre provide a useful framework for studies investigating the discourse of specific disciplines or professional fields. At a more practical level, studies into varieties of language can reveal aspects of lexis, grammar and discourse relevant for language teaching, and their findings can serve as solid foundations for designing language teaching programmes for specific disciplines or professional fields.

## **2.2. Research and pedagogical perspectives within the study of ESP**

The field focusing on the research and teaching of the language of specific disciplines in English is called English for Specific Purposes (ESP). The following chapter will review the definitions, classifications, theoretical influences and approaches to ESP over its nearly fifty years of history.

### **2.2.1. What is ESP?**

Hutchinson and Waters (1987), in defining ESP, laid stress on the common features of language and learning in ESP, and general language teaching. They concluded that “ESP must be seen as an *approach* not as a *product*” (p. 19). The central element in their definition was learners' **needs**. As they put it: “ESP is not a particular kind of language or methodology” [...] “it is an approach to language learning, which is based on learner need” (p. 19). In other words, according to Hutchinson and Waters, it is not the language or the methodology that are in some way specific to ESP, but the learner's reason for learning, and this reason is the

foundation of all ESP courses. They even claimed that in theory there was no difference between ESP and general English. The distinction becomes clear when the specific needs of learners are analysed. The tool used to define these purposes is needs analysis, which has its own traditions and literature (West, 1994).

Following a similar approach, Robinson (1991) noted that:

an ESP course need not include specialist language and content. What is more important is the activities that students engage in. These may be specialist and appropriate even when non-specialist language and content are involved. We should be guided by what the needs analysis suggests and what we are institutionally capable of, and cases certainly exist where apparently general language and content are best. (p. 4)

Robinson (1991) defined needs as “what they [learners] have to be able to do at the end of their language course” (p. 7). Accordingly, she viewed ESP as a goal oriented instructional operation based on needs analysis. She added three more characteristics: (a) an ESP course is designed for a clearly defined time period, (b) learners are more likely to be adults, and (c) they are likely to have the same kind of job. She also emphasised that the difference between ESP and general language teaching lay not so much in the specific language, but in the specified needs of people taking these courses. In two aspects her definition is similar to that of Hutchinson and Waters (1987). Firstly, in both definitions, the central element is learners’ needs, and secondly, both refute the existence of a special language.

Researchers since the late 1990s have, however, emphasised the specificity of language and content of ESP classes (e.g., Dudley-Evans & St John, 1998; Hyland, 2002b, 2008; Strevens, 1988). Strevens (1988), for example, in his definition of ESP lists **special content** related to particular disciplines, and **special language** appropriate to relevant activities, together with **learners’ needs**, among the absolute characteristics of ESP. Dudley-Evans and



St John (1998) also stressed that in addition to learners' needs, ESP focused on the language of specific disciplines. Moreover, they also added the variable characteristic of an **ESP-specific methodology** that was applied in some cases. Consequently, they formulated their definition as follows:

1. Absolute characteristics:

- ESP is designed to meet specific needs of the learner;
- ESP makes use of the underlying methodology and activities of the disciplines it serves;
- ESP is centred on the language (grammar, lexis, register) skills, discourse and genres appropriate to these activities.

2. Variable characteristics

- ESP may be related to or designed for specific disciplines;
- ESP may use, in specific teaching situations, a different methodology from that of general English;
- ESP is likely to be designed for adult learners, either at tertiary level institution or in a professional work situation. It could, however, be used for learners at secondary school level;
- ESP is generally designed for intermediate or advanced students. Most ESP courses assume basic knowledge of the language system, but it can be used with beginners. (Dudley-Evans & St John, 1998, pp. 4-5)

A common element of all the definitions proposed over the past decades of ESP history has been the central role of learners' needs, or their reasons for learning the language. Recently, however, the idea of a special language and methodology has come to the foreground in ESP research and teaching. This development has been supported by new approaches to linguistic analysis such as text linguistics, discourse analysis, register analysis,

genre analysis and corpus linguistics, which provided analytical frameworks that are suitable for highlighting relevant differences in the language varieties used in different situations and contexts of particular disciplines and professions. As regards the methods for teaching ESP, Hutchinson and Waters (1987) clearly refuted the existence of a separate methodology for ESP, whereas Strevens (1988) and Dudley-Evans and St John (1998) included the defining characteristic of a methodology that was different from the methodology for general language teaching. Furthermore, as regards specificity in language use, recent studies into the language use of specific disciplines revealed considerable specificity in the discourse patterns and language features of different disciplines (e.g., Flowerdew, 1994; Hyland, 2002b; 2008; Nelson, 2000; Wang, Liang, & Ge, 2008). As Hyland (2002b) claimed “by stressing students’ target goals and the need to prioritise competencies, specificity clearly distinguishes ESP and general English” (p. 386).

Similarly, the present study adopted an approach that follows the trend towards specificity in ESP research and teaching, as this study is very much motivated by an interest in the activities, discourse practices and linguistic features of the language varieties of particular discourse communities. Consequently, teaching ESP is viewed here as an activity that is guided by learners’ needs and that is centred on the language features characterising the discourse of the specific discipline or profession it serves.

### **2.2.2. Classification of ESP**

Based on the degree of specificity of learners’ needs, purpose and language, ESP is commonly divided into English for Occupational Purposes (EOP) and English for Academic Purposes (EAP). Starting from a linguistic approach, in the teaching of ESP we have subdivisions according to differences in language, for example, EOP for waiters, pilots, studies in physics, economics, etc. These can be refined further into even smaller sub-subcategories. To establish a theoretical basis for ESP, Widdowson (1984) suggested the use

of a continuum with training at the one end, and education at the other end of the scale, and the purposes of ESP arranged along this scale of specificity. Figure 1 shows this kind of continuum, as suggested by Dudley-Evans and St John (1998), with General Purpose English at one end, and specified ESP courses at the other.

GENERAL				SPECIFIC
Position 1	Position 2	Position 3	Position 4	Position 5
English for Beginners	Intermediate to advanced EGP courses with a focus on particular skills	EGAP/EGBP courses based on common-core language and skills not related to specific disciplines or professions	Courses for broad disciplinary or professional areas, e.g. Report Writing for Scientists and Engineers, Medical English, Legal English, Negotiation/ Meeting Skills for Business People	1) an academic support course related to a particular academic course 2) one-to-one work with business people

Figure 1. Continuum of ELT course types (taken from Dudley-Evans & St John, 1998, p. 9)

In a different classification (Robinson, 1991), specificity is defined by the work experience learners have in their specific discipline or professional field. Based on this approach, Robinson sub-divided EOP into pre-experience, in-service and post-experience courses and EAP or English for Educational Purposes (EEP) into English courses for the studies in a specific discipline and English courses as school subjects (see Figure 2).

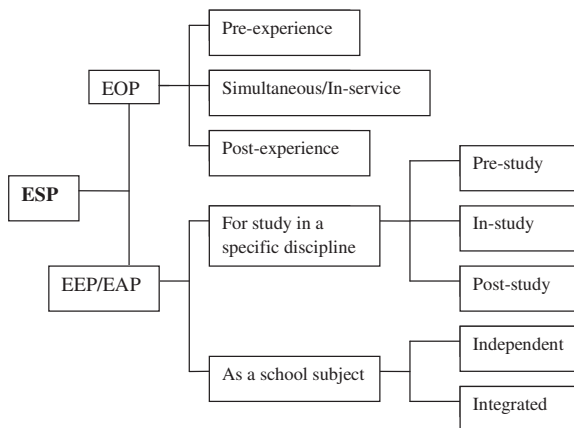


Figure 2. Classification of ESP based on experience (Robinson, 1991, pp. 3-4)

The present study was conducted with learners in mind who either learn English as a subject for their studies in International Relations or European Studies, or already work in some subject field related to the European Union. Therefore, findings can be used for course and materials design for pre-experience or in-service learners in English courses in Position 4 on the specificity continuum (Figure 1).

### 2.2.3. Historical overview of the study of ESP

In their historical overview, Hutchinson and Waters (1987) described the emergence of teaching and research on ESP as a development that evolved world-wide for three main reasons. Firstly, after the Second World War, in a technology and commerce driven world, there was great demand for an **international language**. Mainly due to the economic power and influence of the USA, English gradually became the *lingua franca* in business and economic contexts. In the 60s and 70s, this development created a need for courses with clearly defined aims. The second reason for the emergence of research on ESP was a development within linguistics which meant a shift from investigating the formal features of a

language, to exploring the language as actually used in **communication**. The assumption that the language used in different situations varies according to the **context** and **topic** could be applied to courses that were defined and guided by needs. The third reason was the development of educational psychology, which concentrated on the learner who has different **needs** and **interests**, which, in turn, have an impact on their motivation and therefore on the effectiveness of their learning.

Describing the present state of ESP, Dudley-Evans and St John (1998) suggested that the early history of ESP was essentially a history of English for Science and Technology (EST), whereas, in the last decade, the biggest area of growth was English for Business and EAP.

This view was reinforced by Hewings (2003) in his article as editor of the journal *English for Specific Purposes (ESPj)*, a well-established, internationally recognised journal publishing research in all branches of ESP. Hewings investigated the development of the field by examining, among other factors, the focus and topics of research articles published in the journal between 1980 and 2001. His findings show that a strong specialisation in ESP took place. He found that there had been a steady decline of articles on ESP in general and, simultaneously, an increase in studies of EOP, especially in the business context. Since 1997, however, an increasing interest in different aspects of EAP can be seen, as some 80% of the papers focused on issues in that area. Hewings suggested that this decline of interest in general ESP indicated that ESP is becoming more and more specialised. Teachers and researchers active in the field of ESP find considerable variations in the skills, approaches and materials relevant for their own particular fields like EAP or Business English (BE). Therefore, it is not relevant for the ESP community to investigate and describe general ESP programmes or textbooks any longer.

Hewings (2003) also described trends which will influence the development of ESP in the future. According to him, the five areas of growth are as follows: (1) **internationalisation**; that is, ESP research will spread geographically, especially in China and Eastern and Mediterranean Europe; (2) **specialisation**; that is, more and more specific contexts will be examined for the development of more specific ESP courses; (3) growth of **Business English**; although the growth of BE is indicated by the increasing number of books and materials in the field, the growth in research seems to lag behind. Research in BE might catch up, as BE courses are provided at universities; (4) the three theoretical fields, and the methodology for research within ESP that will continue to be particularly influential are **genre analysis**, **corpus analysis** and **systemic functional linguistics (SFL)**; (5) **English as an international language**; although this was not documented in *ESPj* at the time of Hewings's article, and this last trend was suggested as speculation, six years after the publication of Hewings's article there is evidence for this development. In addition to publishing several articles on issues of English as an international language, the *ESPj* has also produced a special issue on English as an international *lingua franca* in business contexts (Nickerson, 2005b).

The focuses and topics of articles published by *ESPj* between 2002 and 2008 show that these tendencies continue to be relevant. As demonstrated in Figure 3, most articles (91%) deal with the analysis of English language used in specific professions or jobs, such as Medical English, Legal English or English for Economics, and only 5% of the papers discuss issues of ESP in general. This neatly illustrates the specialisation within the field, too. The two most important areas continued to be BE and EAP. BE accounts for 17% of all articles between 2002 and 2008. Nearly a quarter of these articles deal with English as a *lingua franca* in a business context. The area that accounts for almost two-thirds of all the articles is EAP. Compared to earlier periods, there has been a slight decline, as the proportion of articles on

EAP was nearly 80% between 1987 and 2001 (Hewings, 2003). It still shows that many researchers working in the field of ESP are preoccupied with issues in EAP.

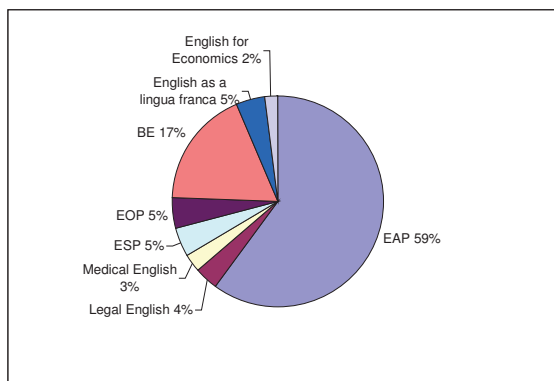


Figure 3. Focus of articles between 2002 and 2008 (source: author's own data).

In his further analysis of the topics of papers in *ESPj*, Hewings (2003) found that there has been a considerable increase in articles focusing on text or discourse analysis. According to him, the reason for this is twofold. Firstly, this kind of analysis provides reliable and relevant information about the target situation for ESP courses. For example, in order to teach how to give business presentations, ESP practitioners need to investigate real business presentations. Secondly, new tools and techniques have been developed to help analyse this kind of data. The most frequently used approaches within ESP are **genre analysis** and **corpus analysis**.

#### 2.2.4. Theoretical influences on the study of ESP

The development of ESP has been considerably influenced by developments in approaches to linguistic description and language teaching in general. In order to gain an overall picture of the main theoretical influences on ESP, Hewings (2003) analysed the references in the articles of *ESPj* from 1980 to 2001. Table 1 gives a summary of the most

influential authors and the topics of their studies, that were most often cited over this period. In the early days of ESP, that is, in the 60s and 70s, focus on EST was prevalent. Language teaching programmes for specific purposes, especially within EST, were designed and started at universities. Analysing references of articles in *ESPj* in the early 1980s, Hewings found studies by Swales (1985), Widdowson (1978, 1984) and Selinker, Tarone and Hanzeli (1981) on EST among the most widely used sources.

According to Hewings (2003), Henry Widdowson's influence has been felt throughout the history of ESP. At the beginning, it was his works on EST, and later his concept of **communicative language teaching** (Widdowson, 1978), and since the second half of the 1980s his seminal work, entitled *Learning Purpose and Language Use* (1984), which provided a theoretical framework for ESP that has been referred to constantly. In the 1980s another influential concept was Munby's **needs analysis** (1978). Although his book entitled *Communicative syllabus design* was not referred to in the later issues of *ESPj* the concept of needs analysis became one of the corner stones of teaching and course design in ESP. As can be seen in Figure 4, needs analysis is still an important subject of articles, which is indicated by the fact that 12% of articles published in *ESPj* dealt with issues of needs analysis between 2002 and 2008.

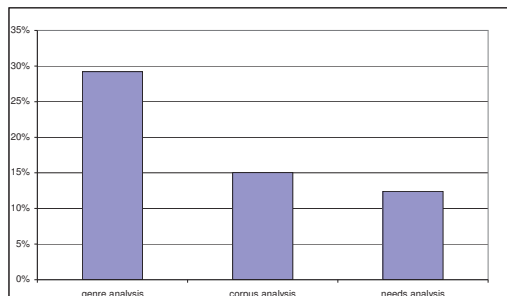


Figure 4. Research methods most frequently used between 2002 and 2008  
(source: author's own data)



Since the 1980s, researchers and practitioners in ESP have turned to a new approach to linguistic description. This new approach offered a broader concept of text analysis that focused on aspects of discourse. The theory underpinning **discourse analysis** was Halliday's SFL, which concentrates on the function of language as opposed to the individual structures and elements of language. A central concept of SFL is the social context, and it investigates how language acts upon, and is constrained by social contexts (Eggins, 2004; Halliday, 1994; Halliday & Hasan, 1985; Halliday & Matthiessen, 2004; Martin, 1992). The remarkable influence of Halliday's linguistic theory on ESP has also been documented in the frequent referencing of his work, even up until now (Hewings, 2003).

In the same vein, as interest in EAP, especially academic discourse and academic writing increased, seminal works like Swales' *Genre Analysis* (1990), Hyland's (2000) book on disciplinary discourses, and Bhatia's book entitled *Analysing genre* (1993), have been referred to frequently (Hewings, 2003). In addition to these books, references have often been made to their articles on academic writing (e.g., Bhatia, 1999; Hyland, 2002a, 2006; Swales, 2004; Swales & Feak, 1994).

The most recent theoretical influence in the field is that of **corpus linguistics**. Already since the second half of the 1990s papers with a corpus-based methodology have been published in the field of ESP (e.g., Conrad, 1996, 1999; Flowerdew, 2004, 2005). As shown in Figure 4, this method has become one of the most frequently used types of analysis in the last six years. Already earlier references to Sinclair's works on corpus linguistics and aspects of corpus analysis (see Table 1) revealed the general interest in this relatively new approach within ESP.

Author	Vols 1-5 1980-1986	Vols 6-10 1987-91	Vols 11-15 1992-96	Vols 16-20 1997-2001
Swales	• ESP programmes worldwide	• research articles – genre analysis	• academic writing	• academic writing
Widdowson	• EST • communicative language teaching	• communicative language teaching • learning purpose and language use	• communicative language teaching • learning purpose and language use	• communicative language teaching • learning purpose and language use
Munby	• needs analysis			
Dudley-Evans & St John	• team-teaching		• academic writing	• general ESP textbook
Hutchinson & Waters		• general ESP textbook	• general ESP textbook	• general ESP textbook
Selinker et al.	• EST interlanguage	• EST interlanguage		
Halliday		• theory for discourse analysis SFL	• theory for discourse analysis SFL	• theory for discourse analysis SFL
Hyland				• academic writing
Bhatia				• academic writing
Sinclair				• corpus linguistics

Table 1. A selection of authors and works referred to most frequently between 1980 and 2001 in *ESPj* (based on Hewings, 2003, p. 9)

### 2.2.5. Approaches to teaching ESP

The development of ESP has always been influenced by the shifts in approaches to linguistic description in general, the importance attached to the language and practices in the target situation, and, closely related to it, by the evolution of the concept of learner needs. All of these influences have left their mark on the methods and approaches to teaching ESP. Accordingly, five main approaches can be distinguished throughout the development of ESP depending on what stands in the focal point of teaching language for specific purposes: (1) ESP is defined from a linguistic point of view in the **language-centred approaches**; (2) in other approaches the **needs of learners** are the starting point for teaching ESP; (3) in the **skills-centred approach** language is viewed as a means of communication in ESP classes; (4) in later developments the focus shifted to the disciplines ESP serves and **multi-disciplinary** and content-based ESP programmes were designed with parallel language and subject matter learning objectives; (5) a totally different view of ESP proposed by Hutchinson and Waters

(1987) centred on the process of **language learning**. Although certain approaches were prevalent at particular stages in the development of ESP, there is no clear cut beginning and end of these stages chronologically. As illustrated in the timeline presented in Figure 5, the different approaches were developed, simultaneously and findings and benefits of all perspectives have been available to ESP teachers throughout the relatively short history of ESP. Especially in ESP today, institutions or practitioners usually use the benefits of earlier perspectives applying an **eclectic approach** (Nelson, 2000). Therefore, these different perspectives of ESP are not discussed here in a chronological order, but they are grouped according to their main focus. A summary of the main approaches to teaching ESP is given in Table 2.

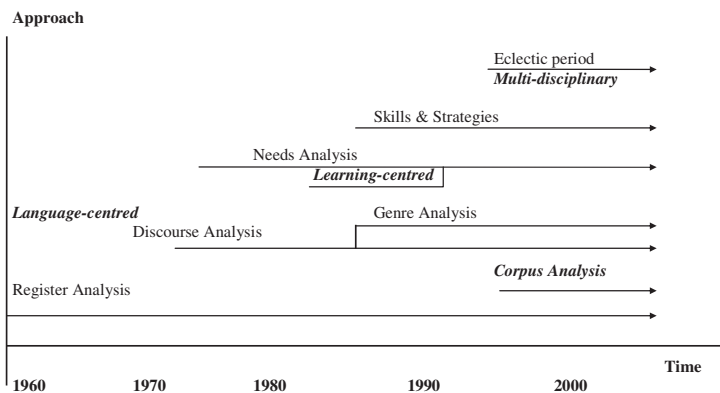


Figure 5. A time-line of approaches to ESP  
(based on Nelson, 2000, p. 35; terms in italics added)

### 2.2.5.1. Focus on learning

Hutchinson and Waters (1987) proposed the **learning-centred approach** to ESP. In contrast to language-centred and skills-centred approaches, they claimed that the learner must be considered at every stage in the course design process, and factors such as the target

situation and learning situation should be taken into account. The learning process should be in the focus of course design and throughout the course there is a constant dialogue between the learner and the teacher. In the case of a learning-centred approach, factors such as variety of activities in the lesson, students' reaction to the tasks, and learners' attitude to methodology throughout the course, and to the aim or topic of the course, are to be considered. The purpose of this approach is to maximise the potential of the learning environment, therefore the syllabus should be used flexibly, sensibly and sensitively, in order that it provides guidance and support, and does not stifle creativity. The starting point for ESP in the learning-centred approach is not so much the competence the learner needs in the target situation, but rather the way in which this competence can be acquired.

Although this approach brought new and relevant elements for consideration in ESP in the 1980s, it has not been taken up by many practitioners and researchers as an approach in its own right. Several factors regarding the learning process and learning situation are now included in the concept of learner needs (Dudley-Evans & St John, 1998, p. 125). The reason why this approach was not widely applied may be that it did not capture the real difference between teaching general English, and ESP, which lies more in the specificity of learner needs and language (Hyland, 2002b).

Focus	Author	Main aspect of the target situation	Approach to linguistic description	Approach to language teaching, methodology
<b>Focus on learning</b>	e.g., Hutchinson & Waters (1987)	• performance in the target situation is of secondary importance	• ESP is not about teaching specialised varieties of English	• learning-centred approach
<b>Focus on needs</b>	e.g., Munby (1978), Robinson (1991)	• needs analysis – target situation analysis	• language as a means of communication in specific target situations • described terminal situation language functions	• functional/notional approach
	e.g., Hutchinson & Waters (1987)	• needs analysis – means, lacks and learning	• sociolinguistic approach	
<b>Focus on skills and strategies</b>	e.g., Esteban & Cañado (2004) Powell (1996)	• skills and strategies	• universal interpreting processes underlie language use	• skills-centred approach • instrumental approach • case studies
<b>Focus on the discipline</b>	e.g., Dudley-Evans & St John (1998) Stoller (2004)	• sociolinguistics • concept of discourse communities	• multi-disciplinary approach – language is inseparable from subject	• content-based instruction • team-teaching
<b>Focus on language</b>	e.g., Zak & Dudley-Evans (1986)	• language, focus on written language	• register analysis	• language-centred approach
		• language, focus on written language	• rhetorical analysis – discourse analysis looks beyond the sentence level, language use rather than usage	
	e.g., Árvay & Tankó (2004) Swales (1990) Tompos (2001)	• language, broader concept of text, spoken language included • concepts of discourse and function included	• genre analysis - discourse analysis	• genre-based approach
	e.g., Chung & Nation (2004) Nelson (2000) Mudraya (2006)	• language, broader concept of text • concepts of discourse and function included	• corpus linguistic register analysis	• lexical approach • data-driven language learning

Table 2. Summary of approaches to researching and teaching ESP

### 2.2.5.2. Focus on needs

The assumption behind the concept of **needs analysis** is that the purpose of an ESP course is to prepare learners for the target situation, that is, the situation in which they will use English at their workplaces. Advocates of this view argue that the first step in ESP course design should be to identify the target situation and to carry out a careful analysis of the characteristics of that situation (Hutchinson & Waters, 1987). In the beginning these characteristics comprised linguistic features of the target situation, that is, the language functions required. Target situation analysis was developed in parallel with the functional/notional approach to language teaching (Nelson, 2000; Wilkins, 1976). According to Wilkins (1976), language constitutes functions, that is, the purposes to which language is used, and notions, that is, ideas to express. Munby (1978) proposed a list of target situations for communicative syllabus design. His list had a great influence on the development of course design in ESP (see Table 1). The main criticism against Munby's proposed list was that it was too long and elaborate to be applied in practice (Nelson, 2000; Robinson, 1991). Therefore, although it was of great theoretical value, it could not be used in the practice of ESP course design.

Concepts of needs analysis have changed as approaches to linguistic description and communicative competence have also changed. The notion of needs was broadened and it was not only defined as the target situation language functions, but new elements like means, necessities, lacks and wants (Hutchinson & Waters, 1987) were identified for ESP course design.

In later developments of ESP, the concept of need was broadened further as computers started to be used for the analysis of texts used in the target situation, and for processing responses to needs analysis (Jones, 1991; Nelson, 2000). Nelson (2000), for example, analysed a corpus of published BE materials in order to measure to what extent these were

suitable for BE students. As needs analysis is of special relevance to the present study, the concept of needs analysis will be discussed in more detail together with a description of course and materials design in ESP, in Section 2.2.7.

### 2.2.5.3. Focus on skills and strategies

The main concept behind the **skills-centred approach**, as formulated by Hutchinson and Waters (1987), is “that underlying all language use there are common reasoning and interpreting processes, which, regardless of the surface forms, enable us to extract meaning from discourse” (p. 13). Therefore, with this approach to ESP focus shifted from the language form to cognitive processes that underlie language use. As a consequence, interest in subject registers was abandoned, because these universal thinking processes were not found to be specific to registers. The emphasis was on teaching strategies that enable learners to cope with surface forms, for example, using context to guess the meaning of a word (Hutchinson & Waters, 1987).

In terms of materials, the **skills-centred approach** focused almost exclusively on reading skills and strategies in the beginning. By the 1980s this focus was broadened to include listening and speaking skills, too. At the beginning of the 1990s, several teaching materials, especially in BE, were published that concentrated on specific skills like giving presentations (Ellis & O’Driscoll, 1992; Kerridge, 1988; Powell, 1996), business meetings (Goodale, 1987; O’Driscoll & Pilbeam, 1987), negotiating (O’Connor, Pilbeam, & Scott-Barrett, 1992), socialising (Ellis, O’Driscoll, & Pilbeam, 1987) or telephoning (Bruce, 1987).

The other approach that focuses on skills and strategies is the **instrumental approach** which aims to teach language as a means of communication, in order to enable learners to carry out certain activities in the target language. The method applied within ESP to create a near real life context for teaching and practising these skills is the **case study** (Esteban & Cañado, 2004; Jackson, 2002, 2004; Howe, 1992a, 1992b). English for Business is a good

example of this approach, as the main aim of a course on English for Business is to equip learners with the ability to operate effectively on those occasions in their business life when they are required to use English.

#### **2.2.5.4. Focus on the discipline**

Dudley-Evans and St John (1998) considered ESP a multi-disciplinary activity in that the specificity in the ESP teaching practice is to be based on the insights of researchers of the disciplines or professions ESP serves. They proposed a **multi-disciplinary approach** for ESP that has two main aspects: firstly, ESP teachers must be willing to deal with other disciplines and, secondly, they need to draw on the insights of researchers in other disciplines. Sociological studies of professions and rhetorical studies of how different professions communicate help to understand the use and function of spoken and written texts in the particular disciplines and professions. In the case of English for Business texts on, for example, human resources management and management training, indicate what are common thought patterns, communication and cognitive styles of people in business. Similarly, cultural differences have to be accounted for, thus elements of cross-cultural communication training can be applied in ESP teaching. Dudley-Evans and St John (1998) also emphasised that the influence is not in one direction. ESP has had its impact on other disciplines as well. Communication skills training in L1, and writing in L1, are examples of the influence of ESP on other disciplines (Williams, Swales, & Kirkman, 1984).

A fairly new development within ESP is **content-based** instruction, or Content and Language Integrated Learning (CLIL) (Stoller, 2004). This method aims to integrate language and subject-learning objectives. Studies investigating the effects and outcomes of content-based ESP programmes have demonstrated that learners of these programmes achieved higher grades at language proficiency tests, and performed better in later language development courses than learners of non-content-based instruction (Song, 2006; Stoller, 2004). An



effective means of applying a content-based method is **team-teaching**, where a language teacher and a subject teacher work together on developing and teaching the course (Dudley, 1984; Stoller, 2004). Although there has been an increasing interest in this method both inside and outside the field of ESP, and it has gained acceptance worldwide, especially, in the US; as it poses organisational and financial challenges to educational institutions (e.g., colleges and universities), it has not become regular practice in ESP contexts.

#### 2.2.5.5. Focus on language

The concept of a **special language** has also been in the focal point of research and teaching of ESP and LSP (Kurtán, 2003). Petneki (2000), coming from a German for Specific Purposes background, defines LSP as one of the social dialects determined by occupation or profession. She concluded that LSP is part of general language and it can be found in technical texts and special situations. Furthermore, while Chambers and McDonough (1981) claimed that there must be a special language, otherwise we could not teach ESP, they also recognised that it is not a separate language, but rather a certain register of the given language.

Authors in the literature on ESP agree that at the first stage in the history of ESP the approach ESP practitioners used was a **language-centred approach**. More specifically, at this early stage, researchers of the field tried to identify lexical and grammatical features of varieties of English used in certain disciplines like Engineering, Aviation, Physics, etc. (e.g., Hüllen, 1981; Malcolm, 1987; Tarone, Dwyer, Gillette, & Icke, 1981; Zak & Dudley-Evans, 1986). Their assumption was that these varieties constituted specific registers, and learners of these special areas of English could be best taught by informing them of the key grammar structures and lexical items to be found in the texts of their respective disciplines or professions (Halliday, McIntosh, Strevens, 1964; Petneki, 2000). The most important method of identifying these specific linguistic features was **register analysis** (Dudley-Evans & St

Johns, 1998; Hutchinson & Waters, 1987; Nelson, 2000). The language teaching methodology based on the findings of early register analysis was described in a volume of papers looking at the history of ESP over the last 25 years, as follows:

Once upon a time, ESP was little more than the teaching of special vocabulary and certain structures. Instead of a text on ‘The Brown family’ with sentences such as: ‘This is Susan. Susan is a girl.’ the students read texts on ‘The Workshop’ with sentences such as ‘This is a hammer. A hammer is an instrument’. (Holmes, 2005, p. 239)

Register analysis was abandoned in the 1970s, but because of the successful application of new technologies and, especially, new methods corpus methodologies to the investigation of specific linguistic features of ESP (e.g., Conrad, 1996; Chung, 2003; Chung & Nation, 2003, 2004), it is gaining importance again. As shown in Figure 4 (see p. 32), 15% of the articles in *ESPj* in the period between 2002 and 2008 applied some kind of corpus analysis. A more detailed overview of the methods and benefits of corpus research to ESP will be given in Section 4.3.

There are two language teaching methodologies that are based on findings of corpus analyses. One of them is the **lexical approach** proposed by Lewis (1993), and the other is the **data-driven language learning** (DDL), first applied by Johns (1991a, 1991b). Neither of these is specific to ESP, as they can be applied in any language teaching situation. The lexical approach will be discussed in more detail in Section 2.3, and DDL will be presented in Section 4.2.

Another frequently used method of identifying linguistic and discourse features of texts used in certain disciplines is **genre analysis**. Such analyses brought insights into rhetorical patterns and discourse structures of genres relevant to specific disciplines and professions (e.g., Swales, 1990; Tompos, 2001). A widely researched genre in ESP/EAP has been the

research article (e.g., Árvay & Tankó, 2004; Swales, 1990). As illustrated in Figure 4 (see p. 32), the analysis of articles in *ESPj* between 2002 and 2008 indicate that genre analysis and the genre-based approach was used in nearly 30% of all articles, and a third of these dealt with research articles of different disciplines.

### **2.2.6. Lexis in ESP**

After reviewing the theoretical influences and different approaches to ESP, one specific area, lexis, will be discussed in more detail, as it is of special relevance to the present study. Before outlining the views and categories of lexis in ESP, however, a terminological issue needs to be considered. Lexical items of a language as a group are referred to as ‘vocabulary’ and also as ‘lexis’ in the literature. According to Altenberg and Granger (2001), the use of the one or the other term expresses a different viewpoint on the status of the lexical level in linguistic description. They enumerate three aspects that are emphasised in studies on **lexis**, which acknowledge the lexical level as an important aspect of language description. Firstly, these studies recognise that lexis and grammar are interrelated, and therefore, analyse lexicogrammatical associations in texts. Secondly, as studies on lexis examine ‘the company words keep’ (Firth, 1968, p. 179), this kind of analysis resulted in the discovery of a great variety of word combinations, for example, collocations, MWIs, or lexical bundles. Finally, studies focusing on lexis have revealed lexical differences in registers in terms of field, mode and tenor, for example, spoken lexis, ESP lexis, and informal lexis. As the present study is motivated by an interest in such lexical and lexicogrammatical features in written English EU discourse, the term ‘lexis’ will be used here to refer to lexical items in general.

In addition, it is also important to clarify what is meant by the terms ‘lexical item’ and ‘word’ in the present study. The working definition of ‘word’ was taken from the *Manual of WordSmith Tools* (Scott, 2004). Word is defined there as “a sequence of valid characters with a word separator at each end. Valid characters include all the letters from A to Z...” (p. 150).

A word separator is most often a space, but *WordSmith Tools* can handle other standard codes used by wordprocessors, for instance, carriage return, tabs, etc. Based on this definition word is understood in the present study as an orthographic unit (Moon, 2000, p. 43) rather than a unit of meaning. The term 'lexical item', however, is used to refer to "the smallest distinctive unit [...] which is mostly understood as a combination of a form and a meaning" (Sterkenburg, 2003, p. 404). Thus it includes single-word items and MWIs as well, as long as the meanings of MWIs is non-compositional, that is, it cannot be inferred from the meaning of its constituents.

The present study made use of further terms relating to different categories and concepts of lexis such as 'lemma', 'word type', 'token' and 'word family'. These will be defined and explained in detail in Sections 4.1.3. and 6.4

As regards lexis in ESP, according to Sager, Dungworth and McDonald (1980), "...the lexicon of special languages is their most obvious distinguishing characteristic" (p. 230). Furthermore, as outlined in Section 2.1, Halliday, McIntosh and Strevens (1964) also pointed out that often it is specific lexical items or their combinations that are signals of particular registers. In addition to acknowledging the distinguishing role of lexis in specialised texts, Swales (1990) also listed the forms in which lexis can be specific as he stated that "in addition to specific genres an established discourse community possesses specific lexis in several forms: using lexical items known to the public in technical ways, using highly technical terminology, using community-specific abbreviations and acronyms" (p. 26). Following a similar pattern, studies into the lexis in ESP established several categories of lexical items in specialised texts. The following sections will focus on lexis in specialised texts in general and MWIs in particular. Finally, issues of teaching lexis in ESP will be outlined.

### 2.2.6.1. Categories of lexis in specialised texts

Most taxonomies (e.g., Dudley-Evans & St John, 1998; Kurtán, 2003; Robinson, 1991; Viel, 2002) of lexis in ESP identify three main categories of lexis in specialised texts, such as **technical**, **semi-technical** and **general lexis**. Chung and Nation (2004) stated that technical lexis is subject-related, and therefore technical terms either occur with a much greater frequency in specialist domains, or occur only in specialised texts. Mudraya (2006) describes lexical items in this category as “words [that] are characterized by the absence of exact synonyms, resistance to semantic change and a very narrow range” (p. 239). In the literature of ESP, technical lexis is referred to by a range of different terms, for example, ‘specialised lexis’ (Baker, 1988), ‘specialist vocabulary’ (Kennedy & Bolitho, 1984). Technical lexis is also the subject of the discipline of terminology, where it is referred to using the labels ‘technical term’, ‘terminological unit’ or ‘terms’ (Kurtán, 2003; Chung & Nation, 2004).

The category of **semi-technical lexis** is defined by Baker (1988) as “a whole range of items that are neither highly technical and specific to a certain field of knowledge, nor obviously general in the sense of being everyday words which are not used in a distinctive way in specialised texts” (p. 91). She also pointed out that it is a “middle area between specialised and general” lexis (p. 92). Other researchers have referred to this group of lexical items as ‘subtechnical’ (Baker, 1988), ‘core vocabulary’ (Dudley-Evans & St John, 1998) or ‘academic vocabulary’ (Coxhead, 2000; Nation, 1990; Martin, 1976). The present study will use the term ‘semi-technical lexis’ for this group of lexical items as it accurately expresses the status of the lexical items in this category as being between the two categories of technical and general lexis.

According to Dudley-Evans and St John (1998), semi-technical lexis has two main types. Some of them appear more frequently in technical texts and others have a specialised meaning in texts belonging to specific disciplines. Examples of the second type include the

word *bug* which refers to a “small insect” in general English and to “a fault in the software” in computer science (Summers, 2003, p. 192), *float* which means “to stay or move on the surface of a liquid without sinking” in general English and “to sell shares in a company or business to the public for the first time” in the context of the stock exchange (Summers, 2003, p. 613), and *council* which refers to “a group of people that are chosen to make rules, laws, or decisions” in general English (Summers, 2003, p. 355), and to a particular institution of the EU in the EU context.

The third category comprises lexical items that are not specific, but can be found in general English with the sub-categories of **general content words** and **general function words**. A summary of the categories with examples is given in Table 3.

Categories of lexis in ESP	Definition	Example
1. technical lexis	• highly specialised lexical items with no semantic ambiguity	<i>call option, rapporteur, directive, dividend, equity</i>
2. semi-technical lexis	• general lexis with a higher frequency in specialised texts	<i>factor, method, project, management, part, analyse</i>
	• general lexis with specific, restricted meaning in certain disciplines	<i>bill, bug, table, harmony, wall, heart, council, float</i>
3. general lexis	• general content words • function words	<i>give, get, early, common, direct the, it, about, have, be</i>

Table 3. Categories of lexis in ESP

(based on Dudley-Evans & St John, 1998; Kurtán, 2003; Robinson, 1991; Viel, 2002)

### 2.2.6.2. Multi-word items in ESP

Earlier studies into the lexis in ESP (e.g., Moon, 1998, 2000; Nelson, 2000) have found that **MWIs** play an important role in technical writing and other specialised texts. In addition, several recent studies emphasise the importance of word combinations in spoken and written discourse in general (Cowie, 1992; Biber, Conrad, & Cortes, 2004; Moon, 1998, 2000). Cowie (1992), for example, in his analysis of journalistic prose, claimed that collocations as ready-made complex units had a significant role in newspaper articles, and therefore they should be focused on in language teaching. There is no generally accepted term in the

literature for MWIs. These word combinations have been studied under different terms like fixed expressions (Moon, 1998, 2000), lexical phrases (Nattinger & DeCarrico, 1992), prefabs, ready-made units (Cowie, 1992), using different criteria to define and identify MWIs and thus throwing light on different aspects of structures and functions of MWIs in discourse. In a comprehensive discussion of word combinations in English, Moon (2000) used the term multi-word item (MWI) and gave the following definition:

A multi-word item is a vocabulary item which consists of a sequence of two or more words (a word being an orthographic unit). This sequence of words semantically and/or syntactically forms a meaningful and inseparable unit. Multi-word items are the result of lexical (and semantic) processes of fossilisation and word-formation, rather than the result of grammatical rules. (p. 43)

She listed three criteria which help distinguish MWIs from other kinds of word strings. The first is **institutionalisation**. This is the degree to which the item is conventionalised in the language, that is, whether it recurs in language use. The second criterion is **fixedness** of the MWI, which determines whether the parts of it can be varied, or its word order can be changed. The third is **non-compositionality** which means that a MWI cannot be interpreted by word-by-word analysis, but its meaning is more than the sum of its components. The different types of MWIs identified by Moon (2000) are compounds, phrasal verbs, idioms, fixed phrases like *of course*, *at least* and prefabricated routines or prefabs such as *I'm a great believer in*, *the fact/point is*, *that reminds me*. She also claimed that the set of MWIs is open-ended and not static.

Discussing the role of MWIs, on the one hand, she drew on evidence of analyses of language corpora, on the other hand, on text analysis. On the basis of analyses of huge corpora (e.g., The Bank of English) she concluded that: "There are a lot of multi-word items in the language but a lot of them are very infrequent" (Moon, 2000, p. 52). According to her

findings of text analyses, the densities of MWIs in different text types indicated that the use of MWIs is dependent on the particular register. Moon illustrated the use of MWIs by a piece of technical writing, i.e. a specialised text from a handbook on painting. She found that technical terms, signalling of structure and clause relationships were, most commonly expressed by the use of MWIs in written technical registers.

Whereas studies into MWI by Moon (2000), Nattinger and DeCarrico (1992) and others (e.g., Pawley & Syder, 1983) focused on pre-selected word combinations, recent research has concentrated on frequently recurring word combinations that emerged from the analyses of specialised texts. This frequency-based approach resulted in a different type of MWI, the so called **lexical bundle** (Biber, 2009; Biber et al., 1999). The concept and characteristics of lexical bundles together with findings concerning lexical bundles in different registers are outlined in Section 4.3.3.1 and 6.4.3.

#### **2.2.6.3. Teaching lexis in ESP**

The importance of teaching lexis has been a controversial issue in ESP. Some researchers (Hutchinson & Waters, 1987) argued that teaching should focus on semi-technical lexis exclusively as this is the category of lexical items that causes difficulty in understanding and producing specialised and technical texts (Baker, 1988; Mudraya, 2006). Dudley-Evans and St John (1998), however, emphasised that it is also the ESP teachers' duty to help students master technical lexis.

The viewpoint on teaching lexis in ESP taken in the present study is the one suggested by Dudley-Evans and St John (1998), that is, both technical and semi-technical lexis should be dealt with by the ESP teacher. This might involve teaching the national, which in the author's case is Hungarian, equivalents for special lexical items, and also give practice in their frequent patterns. Looking at the behaviour and patterns of these lexical items in specialised



texts can provide a good foundation for materials and course design which constitutes one of the aims of the current study.

### **2.2.7. Course and materials design in ESP**

Widdowson (1984), in his distinction between ESP and 'General Purposes English' (GPE) offered two interpretations of 'purpose' in language teaching. He claimed that 'purpose' in the case of ESP meant the occupational or academic aims for which the language will be used eventually. Taking this distinction into consideration, Widdowson (1984) suggested that an ESP course be essentially a training operation which is designed to meet the immediate objectives of the learner, which were at the same time the aims of learning. These objectives make up the specific purposes that should be met by an ESP course. These purposes determine course design, that is, planning the contents of the language teaching programme and also the responsibilities of the language teacher.

Researchers in ESP agree that there are several tasks an ESP teacher has to fulfil. According to Hutchinson and Waters (1987), ESP teachers have to deal with needs analysis, course design, materials writing or adaptation of materials and evaluation. Dudley-Evans and St John (1998) added even more functions to the ESP teacher's responsibilities. One of them is the collaborator's role that refers to the necessary co-operation and consultation with a subject specialist (e.g., team teaching). The additional researcher's role implies not only carrying out needs analysis, but also discourse analysis and conversation analysis of the texts that students will use in the target situation.

The central role of needs analysis is a common element in all definitions of ESP (Dudley-Evans & St John, 1998; Hutchinson & Waters, 1987; Robinson, 1991) Hutchinson and Waters (1987) give six guiding questions that need to be answered "in order to provide a reasoned basis" (p. 21) for course design and materials writing. Figure 6 illustrates these questions and how the different factors affect ESP course design.

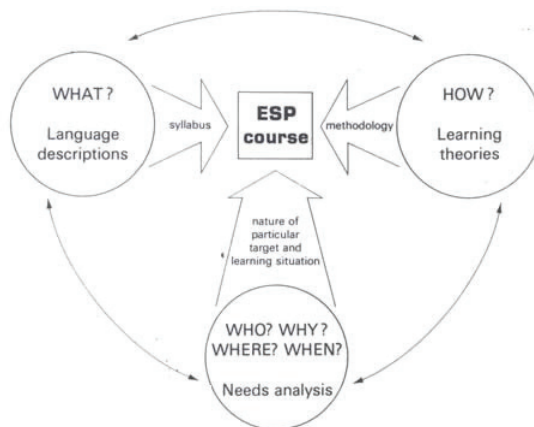


Figure 6. Factors affecting ESP course design (Hutchinson & Waters, 1987, p. 22)

Furthermore, Dudley-Evans and St John (1998) defined the concept of needs analysis as a process that provides professional and personal information about learners, learners' lacks and their needs from the course, language learning needs, and information about the environment in which the course will be run, and also information about means of communication and language of the target situation (p. 125). The focus of this study is on professional communication and language of the target situation. The needs analysis conducted within the framework of the present study, as part of corpus design, is described in more detail in Section 6.3.2.

### 2.3. Lexis in the centre of course design

Focus on lexis is a relatively new approach to course design. Although lexis-oriented course design is often applied in an ESP context, in some cases together with a syllabus structured around relevant topics of the given discipline, profession, or occupation (Kurtán, 2003), it is not exclusive to ESP. Willis (1990), based on corpus research by Cobuild (Sinclair, 1987; Sinclair & Renouf, 1988), proposed a lexical syllabus for teaching general

English. He suggested that learners should be provided with a corpus containing the most frequently occurring words with their commonest patterns in authentic texts. Learners then should infer grammatical patterns themselves. Furthermore, he claims that many grammatical constructions are better acquired as lexis. For example, tense, aspect, voice, conditionals and reported speech. According to Willis (1990), focus in a lexical syllabus should be on lexical items, analysis of language in use, and on raising awareness.

Lewis (1993) criticised the lexical syllabus for the following “dangers” (p. 109):

- (1) the most frequent words are usually function words with low semantic content and very complex patterns, e.g. *to, with, have*;
- (2) the lexical syllabus concentrates on the word as opposed to meaning or senses of words as the basic unit which causes confusion, rare meanings of frequent words are given priority over frequent meaning of less frequent words;
- (3) MWIs are undervalued. (p. 109)

He proposed a lexical approach instead in order to avoid the dangers he perceived in the application of a lexical syllabus. According to Lewis (1993), the lexical approach “is specifically not a lexical syllabus, and explicitly recognises word patterns for (relatively) de-lexical words, collocational power for (relatively) semantically powerful words, and longer multi-word items, particularly institutionalised sentences as requiring different and parallel pedagogical treatment” (p. 109). Lewis also emphasised the importance of MWIs in language teaching and noted that fluency does not come from knowledge of grammar rules but from learning phrases that are often used in certain contexts. A lexically-oriented approach to course design is of special relevance to the present study, as the findings of the analyses can be used as a starting point for course design with a focus on lexis.

## **2.4. Summary**

This overview of the theoretical influences and approaches to ESP revealed a tendency to focus on relevant differences in the language use of certain disciplines and professions. The conclusion that can be drawn here is that the specificity of language varieties used by different discourse communities in different social contexts should be studied in order to form the basis of English language courses that serve learners' communication needs in their relevant target situations. The methods ESP research applies to serve these needs, and to describe language use in specific contexts and for specific communicative purposes are register analysis, genre analysis, and discourse analysis. Quite recently, as a theoretical approach combined with its unique methodology, corpus analysis has also gained grounds among the theoretical influences in ESP research, and has yielded relevant findings in ESP theory and practice, as will be discussed in Section 4.3.

After reviewing the theoretical framework of the current study, the next chapter will review earlier analyses of the variety of the English language as used in communication within the European Union.

## Chapter 3: Earlier analyses of EU discourse

Most studies on the languages used in the EU context concentrate on issues of translation and terminology (e.g., Born & Schütte, 1995; Dróth, 2000; Fischer, 2006, 2007; Károly, 2007; Klaudy, 2001; Schäffner & Adab, 2001a, 2001b), language policy or international relations theories (e.g., Diez, 2001; Pym, 2000; Truchot, 2002). There are only a few studies that focus on the linguistic analysis of documents issued by institutions of the European Union (Born & Schütte, 1995; Laviosa, 2000; Pym, 1993). There are even fewer studies investigating English language EU documents from a language teaching perspective. This chapter will summarise the findings of research into English EU discourse in general, and highlight the findings of analyses with specific pedagogic aims.

### 3.1. Language in the EU – eurojargon and hybridity

In most studies on aspects of EU discourse authors have referred to the existence of a so called ‘eurojargon’, characterised by complex sentence structure, overuse of abstract nouns, complex noun phrases, nominalisation (Trosborg, 1997b), reduced meanings of certain lexical items, and limited inventory of grammatical forms (Pym, 1993).

Furthermore, EU text types have been referred to as **hybrid texts** (Schäffner & Adab, 2001a, 2001b; Trosborg, 1997b). Although further research is needed to define the exact characteristics of this text type, a provisional definition of hybrid texts is available:

A hybrid text is a text that results from a translation process. It shows features that somehow seem ‘out of place’/‘strange’/‘unusual’ for the receiving culture, i.e., target culture. These features, however, are not the results of lack of translational competence or examples of ‘translationese’, but they are evidence of conscious and deliberate decisions by the translator. Although the text is not yet fully

established in the target culture (because it does not conform to established norm and conventions), hybrid text is accepted in its target culture because it fulfils its intended purpose in the communicative situation (at least for a certain time). (Schäffner & Adab, 2001a, p. 175)

This definition was extended by Trosborg (1997b), as she considered not only translated texts as ‘hybrid texts’, but also text types which result from negotiations between cultures and conventions. As she formulated, hybrid texts “are arrived at as an outcome of negotiations between cultures and the norms and conventions involved as well as through translation” (p. 146). As examples of hybrid texts Trosborg mentioned “texts that are produced through collaboration in the European Community, the European Parliament, the United Nations, etc.” (p. 147).

Following Trosborg’s (1997b) definition, EU texts can be considered ‘hybrid texts’. Although they cannot all be looked upon as translations, they are certainly the results of negotiations and co-operation between cultures, as the members of the Commission and other bodies of the EU belong to different cultures and their job is to find common solutions to problems.

### **3.2. English EU discourse in an ESP context**

In order to establish a sound basis for language courses in English for the EU, research is needed into the lexical, syntactic and discourse characteristics of the register of English used within EU institutions. Furthermore, it is necessary to know more about the target situation and the context of this register.

López and Cañado (2001) conducted a needs analysis survey for ESP courses in the European Commission focusing on the teaching and learning process of participants in the courses provided at this particular EU institution. Results of their needs analysis show that

learners need English as a *lingua franca*, that is, they use English to communicate with professionals who are not native speakers of English, but they quite often communicate with English native speakers as well. Moreover, participants have a positive attitude to English and like learning the language, using a wide variety of methods focusing on listening and speaking skills. Although the findings of this analysis provide some information as to the methods learners of English prefer in an EU context, it does not inform ESP teachers about the target situation where these learners will use their English. Questions like what EU genres they use on a regular basis and what these EU texts are used for, remain unanswered.

The most detailed analysis of the register of English in EU documents has been provided by Trebits (2008, 2009a, 2009b). Examining the Corpus of EU English, a fairly small corpus of 200,000 running words containing official EU documents selected randomly and intuitively, she investigated lexis, conjunctions and phrasal verbs. As regards lexis in EU documents, she found that 46.5% of the word types, including several frequently used EU abbreviations in her corpus were, not in the BNC 3000 word list, and that there is little overlap between these frequent words and the words in the lists on the websites of *Euro-Jargon* and *EU Glossary* on the European Union portal. Based on these findings, she concluded that EU documents pose challenges for language learners at an intermediate (B1-B2) level.

In her study on conjunctions, she demonstrated that the number of conjunctions in the Corpus of EU English resembles the register of academic prose in the BNC (Trebits, 2009a). The most frequent types of conjunctions are additives, temporals and causals. The conjunctions that have been found strikingly more frequent in EU texts included conjunctions expressing causal relations such as *with a view to*, *in order to*, *so as to*, *to this end* and continuative conjunctions like *such as*, *in particular*, *namely*, *regarding*, *as regards*. She has also identified several conjunctions that are among the frequently occurring ones in the

written part of the BNC, but are absent from EU documents. These are *in spite of, as though, by comparison, next, in conclusion*.

Finally, as regards phrasal verbs in English EU documents, their number and use has been found similar to that of academic prose, in that both registers apply fewer phrasal verbs than, for example, fiction and news texts. As regards their meanings, phrasal verbs can have multiple meanings in EU documents, but they exhibit fewer meanings than in general English (Trebitts, 2009b).

Jablonkai's (2009a) study compared the lexis and lexical bundles in two EU-related registers. Comparing news texts reporting on EU-related issues, and EU texts such as press releases and legal texts issued by EU institutions, Jablonkai demonstrated that the discourse of these two registers differ considerably, and therefore, instruction for future EU professionals should include teaching materials that are created specifically based on the analysis of official EU documents. The corpora she used were rather small, containing 120,000 running words of news texts, and the same amount of EU texts selected randomly from the time period of January to August 2007.

### **3.3. Summary**

This chapter provided an overview of previous studies looking into EU discourse. The studies reviewed suggest that EU texts exhibit features that are perceived as 'strange' in the target culture, therefore, in general, EU texts are examples of hybrid texts. What specific characteristics and linguistic features this 'strangeness' is caused by has, however, not been described in detail yet. Furthermore, findings regarding English EU discourse from an ESP pedagogic perspective are all relevant for the ESP teacher preparing courses or materials for English for EU purposes. Nevertheless, the corpora used for these analyses were small, and were created based on a rather intuitive selection of texts. In order to provide more reliable findings based on a more comprehensive analysis of relevant lexis a larger corpus is needed



that is compiled based on a more principled selection of EU texts. This is what the current undertaking intended to do.

## **Chapter 4: Methodological framework: corpus linguistics**

As demonstrated in Section 2.2, an increasing number of language variation studies in ESP apply corpus linguistic methods in their analyses. Conrad (2002) noted in her article, evaluating the importance of corpus linguistics for discourse studies, that corpus linguistics had been found particularly useful for characterising the lexis of a specific field. Furthermore, several studies also applied corpus linguistic methods to describe certain text types, registers or the language of certain disciplines and professional fields (e.g., Biber, 2006; James, Davison, Heung-yeung, & Deerwester 1994; Nelson, 2000). As the present study is also investigating the lexis of a professional field with an empirical stance, the approach and methods offered by corpus linguistics were found appropriate.

The following overview of corpus linguistics starts with firstly, a general introduction to the field, giving a brief history, key concepts, units of analysis, approaches, reasons for its application, criticism, and the benefits of corpus research in text analysis. Secondly, the method of language teaching that draws heavily on findings and methods of corpus linguistics, that is, data-driven learning (DDL), is described. Thirdly, corpus research in ESP will be discussed in detail, highlighting the areas within ESP that have benefited most from corpus linguistic methods. Finally, issues in corpus design for ESP will be outlined and a *Model for Corpus Creation for ESP* will be proposed for the investigation into the language of specific disciplines and professional fields.

### **4.1. A corpus linguistic approach to text analysis**

The 20th century witnessed an unprecedented development in all fields of technology. These advances resulted, among others, in the innovation which has since almost become a household 'appliance', the computer. Computer technology has influenced all areas of

scientific research, and it has also contributed to the emergence of a new field within linguistics. Although the application of corpora in linguistic research has its roots in the 19th century, its real ‘career’ can be viewed in parallel with the development of computer technology. In the beginning, the computer was considered a tool making it possible to collect large amounts of data. Later, corpus linguistics developed its own methodology, and in recent years it has become a discipline in its own right. This also means that there are attempts to develop a theoretical framework for linguistic description specific to corpus linguistics. The theoretical stance of linguists who embarked on compiling the first corpora had always been a more empiricist one to linguistic analysis than the mainstream theoretical thinking of their time. How the theoretical stance of linguists towards corpora has developed since then, and what the ‘explosion’ in the field of corpus linguistics has brought to research in linguistics and other fields, are the main focus of this chapter.

#### **4.1.1. Brief history**

There were corpora used for the study of language as early as the 19th century. A German scholar, Kading, at the end of the 1890s, for instance, compiled a huge corpus of 11 million German words. He used his corpus to collate frequency distributions of letters, and sequences of letters, in German (McEnery & Wilson, 1996b). Field linguists, and linguists of the structuralist tradition, used corpus-based methodologies as well, especially for studying distinguishing features in phonetics and certain aspects of grammar (McEnery, Xiao, & Tono, 2006). In the field of language acquisition, corpora of parental diaries were compiled around the turn of the century and they were used also in the 1950s and 70s. Language pedagogy was another important area where corpora were used, especially to create lists of useful lexical items for foreign language learners in the first few decades of this century (McEnery & Wilson, 1996b; West, 1953).

There is, however, a clear methodological difference between the work of these early corpus linguists and their modern counterparts. Nelson (2000) enumerates the following characteristics of these early corpora: (1) they were almost exclusively corpora of written texts, (2) scholars were interested in forms rather than meaning, (3) they did not annotate and parse their corpora which caused problems with, for example, homonyms that were often classified as one word. It should also be added that as computers were not available at that time, early corpora were analysed manually, with paper and pencil methods using several thousand assistants in some cases (McEnery & Wilson, 1996a, 1996b; McEnery et al., 2006).

The development of this empirical approach towards the study of language was halted in the later 1950s. The main reason for this halt was Chomsky's (1957; 1962) view on corpora and the notion of empiricism in linguistics. Chomsky (1986), following a rationalist approach, gave primacy to intuition and introspection, and claimed that "the judgments of the native speaker will always provide evidence for the study of language" (p. 37). Furthermore, Chomsky also argued that "a distinction must be made between what the speaker of a language knows implicitly (what we may call his *competence*) and what he does (his *performance*). A grammar, in the traditional view, is an account of competence" (Allen & Van Buren, 1971, p. 7).

Consequently, Chomsky (1962) considered corpora to be inadequate for linguistic enquiry that should model language competence and not performance. He attacked corpus-related studies by saying that:

Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list. (p. 159)

The core of Chomsky's criticism is that a corpus is a collection of performance data and therefore it cannot serve as evidence for models of linguistic competence (McEnery & Wilson, 1996b).

Scholars of the 1950s and 60s challenged the value of early corpus linguistic studies as they were "time-consuming, expensive and error-prone" (McEnery & Wilson 1996a, p. 10). This criticism was, however, justified at that time, when we consider the lack of technology enabling researchers to run automated searches on their collection of language data (McEnery et al., 2006).

The theoretical stance at that time, however, did not favour studies based on empirical data in general. Linguists were in a situation that Sinclair (1991) described as follows:

Starved of adequate data, linguistics languished – indeed it became almost totally introverted. It became fashionable to look inwards to the mind rather than outwards to society. Intuition was the key, and the similarity of language structure to various formal models was emphasized. The communicative role of language was hardly referred to. (p. 1)

Although linguistic research on corpora has never ceased altogether, there is a clear discontinuity between the work of early and modern corpus linguists. The origins of modern corpus linguistics were there already at the end of the 50s, when Randolph Quirk started compiling the Survey of English Usage (SEU) Corpus, originally on paper. Soon after, Francis and Kucera in America created the Brown Corpus, "a standard sample of printed American English for use with digital computers" (Leech, 1991, p. 9). A similar British corpus was completed in 1978, under the leadership of Leech, with the help of Norwegian colleagues. Hence, it was named the London–Oslo/Bergen (LOB) Corpus (Svartvik, 1996). The era of modern corpus linguistics can be characterised by the extensive use of computers in compiling and analysing corpora. A part of the original SEU Corpus, which contains 50%

spoken and 50% written texts, was computerised by Svartvik late in the 1980s, thus becoming the first, and long unsurpassed, collection of spoken English with the name London-Lund Corpus of Spoken English. The revival of corpus studies, and later on the 'explosion' in the number and size of corpora in the study of language, can also be attributed to technological innovations, and the development of computer technology. Table 4 below illustrates how a direction, which was once thought to be a dead end of linguistics, has by now become mainstream (Svartvik, 1996), and even a discipline in its own right (Tognini-Bonelli, 2001).

<b>Year</b>	<b>Number of publications</b>
to 1965	10
1966-1970	20
1971-1975	30
1976-1980	80
1981-1985	160
1986-1991	320

Table 4. Number of publications in the field of corpus linguistics (Johansson, 1991, p. 312)

#### 4.1.2. Key concepts of corpus linguistics

In the following section, definitions of this young discipline, and the main elements of its methodology, namely, **corpus**, **concordance** and **annotation**, will be given and analysed. A fairly general definition of the scope of corpus linguistics can be found in the Introduction to the volume *English Corpus Linguistics*: “Corpus linguistics can be described as the study of language on the basis of text corpora” (Aijmer & Altenberg, 1991, p. 1). This definition, however, does not reflect what clearly distinguishes the two eras of early and modern corpus linguistics, that is, the use of computers. Without the development of computer technology corpus-related studies would not have become so widespread and popular and the types of analysis that are possible this way would not be feasible without computers.

#### 4.1.2.1. What is a corpus?

A simple way to look at corpora is to view them as collections of texts. The analysis of these corpora falls within the scope of corpus linguistics. However, research on spoken and written text is not limited to corpus linguistics, and not all collections of texts can be considered a corpus in the modern corpus linguistic sense. The actual function of the collection of texts should be taken into consideration as well. There are anthologies whose purpose is literary and the Corpus Juris of a king or emperor whose purpose is legal. The purpose of a **corpus** in modern corpus linguistics is for linguistic analysis, to investigate language use (Tognini-Bonelli, 2001).

A corpus that serves this purpose must fulfil some essential criteria. The following requirements are most widely accepted by researchers working in the field (e.g., Aarts, 1991; Biber et al., 1998; Knowles, 1996; McEnery & Wilson, 1996a; Sinclair, 1991; Tognini-Bonelli, 2001):

- authenticity
- representativeness
- sampling
- finite size
- machine-readable form
- standard reference

*Authenticity*: This is one of the corner stones of corpus work. The starting point for linguistic enquiry is the language in use (Aarts, 1991, p. 45). In order to capture language in use, all texts included in a corpus are assumed to be taken from genuine communication based on the assumption that “texts exist not for the sake of the form, but in order to communicate meaning” (Knowles, 1996, p. 52). The lack of authenticity leads, for example, to pedagogical grammars with prescriptive rules. These will have to be revised in the light of authentic corpus evidence (Tognini-Bonelli, 2001).

*Representativeness:* There seems to be general agreement in the literature that a corpus used for linguistic analysis should be representative of a certain population in order to make results of investigations on the corpus generalisable for the language use of that certain population. On the one hand, in the case of a 'general purpose' corpus this means that the sample we are analysing should be as large as possible, and a broad range of authors and genres should be included. Hence it is possible to make more precise statements of the language as a whole. On the other hand, one can choose to investigate one register, or a variety of a language, which necessitates a corpus constructed of the text types and genres relevant to the communication of that particular population (Biber et al., 1998; McEnery & Wilson, 1996a; Tognini-Bonelli, 2001).

*Sampling:* In order to be able to decide which texts to include in the corpus, researchers should define the target population the corpus aims to represent. Throughout the collection process these criteria will provide the rationale for decisions about sampling (Biber et al., 1998). An important issue that needs to be covered here is whether only whole texts, or excerpts of texts, should be included in the corpus. Although there are corpora which consist exclusively of extracts of texts, e.g. the Brown Corpus, which is a collection of texts not longer than 2000 words, the British tradition seems to consider the whole text as the unit of study for corpus work (Tognini-Bonelli, 2001). The theoretical consideration behind this is formulated by Sinclair as follows: "a corpus made up of whole documents is open to a wider range of linguistic studies than a collection of short samples" (Sinclair, 1991, p. 19). The issue of sample size will be taken up again in Section 4.4.6.2 as an important issue in corpus design.

*Finite size:* Corpora, more often than not, comprise a finite number of texts. This is, however, not always the case. There is the, so called, monitor corpus, the first one of which was designed by Sinclair's team, and which has an infinite size, as new texts are being added



to it continuously. This is mainly used for studies in lexicography. A monitor corpus has its drawbacks, as quantitative research done at different times is not comparable, as the size of the corpus is always changing (McEnery & Wilson, 1996a).

*Machine-readable form:* This criterion refers to corpora of the modern corpus linguistics era. Although there are a few corpora that are available in a book format, or spoken corpora which are available as actual recordings, the tendency is that corpora are created in a machine-readable format. This format makes it possible for corpora to be searched by computers. This automatic means of investigation is much quicker and less error-prone than any other methods (McEnery & Wilson, 1996a).

*A standard reference:* Most of the big corpora are available either on CD-ROM or via the Internet (e.g., David Lee's Corpus-based linguistics links at <http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLlinks.htm>). The advantage of widely accessible corpora is that findings are easily comparable, as long as methodology issues are described in detail, since they refer to the same set of data (McEnery & Wilson, 1996a).

On the basis of the criteria discussed above, the definition of a **corpus** that reflects these requirements was formulated by Oakes (1998) as follows: a corpus is "(iii) (more strictly) a finite collection of machine-readable texts, sampled to be maximally representative of a language or variety" (p. 251). It is important to bear all these requirements in mind when researchers start to compile their own corpora.

A further important element of corpus building is documentation, that is, details of the genres, text types, length of individual texts, etc. in the corpus, as findings can only be interpreted with sufficient information on the composition of the corpus used for the analysis. Figure 7 gives an example of this kind of information, based on the documentation of the Business English Corpus (BEC) compiled by Nelson (2000).

## WRITING TO DO BUSINESS

PART OF CORPUS	TOKENS	CONTENTS
ANNUAL REPORTS	34,537	3 annual reports
BUS PRESS RELEASES	21,656	29 business press releases
BUSINESS CONTRACTS	29,602	13 contracts/agreements
BUSINESS FAXES	23,105	114 faxes
BUSINESS LETTERS	26,793	94 letters
BUSINESS REPORTS	62,908	17 reports
COMPANY BROCHURES	23,239	13 company brochures
EMAILS	28,857	202 emails
JOB ADVERTISEMENTS	22,293	87 job advertisements
MANUALS	21,160	5 manuals
MEMOS	12,542	47 memos
MINUTES	34,805	15 sets of minutes
PRODUCT BROCHURES	26,175	19 product brochures
QUOTATIONS	8,997	21 quotations
MISCELLANEOUS	2,427	OHT, job description & agendas
<b>TOTAL</b>	<b>379,096</b>	

Figure 7. Documentation of the BEC (Nelson, 2000, p. 226)

### 4.1.2.2. Kinds of corpora

The criteria defining a corpus were discussed in Section 4.1.2.1, what follows is a short description of different kinds of corpora. As the development of corpora is an on-going process, and new kinds of corpora for specific purposes are being designed, the list is by no means exhaustive. Table 5 gives a few examples of the kinds of corpora discussed in this section.

With advances in computer technology, more and more facilities were developed to make use of corpora. The different kinds were developed for different purposes as the scope of corpus linguistics has been widening in the last few decades. The first generation of corpora can be considered as **sample corpora** (Sinclair, 1991, p. 23). These are collections of carefully selected texts, in some cases not even whole texts, only extracts from different genres (e.g., novels, letters, and talks). A **monitor corpus**, on the other hand, has an infinite number of texts. The aim of a monitor corpus is to provide a large and up-to-date collection of texts. The monitor corpus “like the language itself, keeps on developing” (Sinclair, 1991, p. 25). The other aspect in which a monitor corpus differs from a sample corpus is that texts are

not selected according to genres or other textual criteria. Researchers who wish to use parts of the monitor corpus for specific research can filter the texts in the corpus to serve their own purposes of study. Monitor corpora (e.g., the Cambridge International Corpus) are primarily used for lexicographic research. A drawback of this kind of corpus is, as mentioned earlier, that studies conducted with the help of monitor corpora are not comparable, because of their continuously changing size and contents (McEnery & Wilson, 1996a).

Scholars interested in syntactic or grammatical studies of the language most often use **annotated corpora**. In this kind of corpora additional information is built into the corpus. According to Leech (1991), there is a need for annotation to help automatic searches, as the computer, for example, will not be able to make a difference between *minute* as a noun, and *minute* as an adjective. He also claims that the original 'raw' corpus should be made available for researchers who "find annotations useless or worse" (Leech, 1991, p. 25). What annotation entails, and what this additional information can be, will be discussed in the next section.

A type of corpus that is used in contrastive and translation studies is called **parallel corpora**. These corpora contain the same texts in original and translated forms. These are also aligned, which makes it possible to compare equivalents in the two given languages. Another type of corpus that contains texts in more than one language is a **comparable corpus**. Comparable corpora contain texts collected according to the same criteria that often refer to similar circumstances of communication, for example, tourist brochures or job advertisements in different languages (Thompson, 2001). These corpora can be used to compare linguistic and discourse patterns across languages, and avoid the distortions introduced by translations (Hunston, 2002).

The most recent development within corpus linguistics is that it is not only texts produced by native speakers in naturalistic situations that can be parts of a corpus, but texts produced by language learners are also compiled into corpora. These **learner corpora** (e.g.,

The International Corpus of Learner English) are used to study second-language acquisition, for example, to compare the interlanguage of learners of different mother tongues.

The kind of corpus of special interest to the present study is the so called **specialised corpus**. A specialised corpus is defined by Hunston (2002) as:

A corpus of texts of a particular type, such as newspaper editorials, geography textbooks, academic articles in a particular subject, lectures, casual conversations, essays written by students etc. It aims to be representative of a given type of text. It is used to investigate a particular type of language. [...] There is no limit to the degree of specialisation involved, but the parameters are set to limit the kind of texts included. For example, a corpus might be restricted to a time frame, consisting of texts from a particular century, or to a social setting, such as conversations taking place in a bookshop, or to a given topic, such as newspaper articles dealing with the European Union. (p. 14)

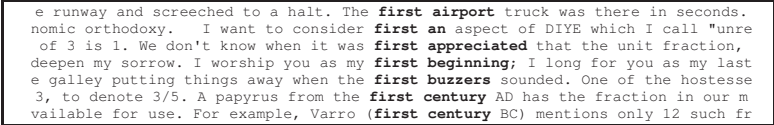
The advantages of specialised corpora for ESP will be enumerated in detail in Section 4.3.1.

Kind of corpus	Example	Description
<b>sample corpus</b>	• British National Corpus (BNC)	<ul style="list-style-type: none"> <li>• 100 million words</li> <li>• spoken and written British English texts of varying length</li> <li>• <a href="http://www.natcorp.ox.ac.uk">www.natcorp.ox.ac.uk</a></li> </ul>
	• Hungarian National Corpus	<ul style="list-style-type: none"> <li>• 187.6 million words</li> <li>• written Hungarian texts representing several contemporary varieties</li> <li>• <a href="http://corpus.nytud.hu/mnsz/">http://corpus.nytud.hu/mnsz/</a></li> </ul>
<b>monitor corpus</b>	• Cambridge International Corpus	<ul style="list-style-type: none"> <li>• over one billion words of spoken and written, American and British English, new texts added continuously</li> <li>• <a href="http://www.cambridge.org/elt/corpus">http://www.cambridge.org/elt/corpus</a></li> </ul>
	• The Bank of English Corpus	<ul style="list-style-type: none"> <li>• 524 million words</li> <li>• new texts added continuously</li> <li>• <a href="http://www.titania.bham.ac.uk/">http://www.titania.bham.ac.uk/</a></li> </ul>
<b>annotated corpus</b>	• Penn Treebank	<ul style="list-style-type: none"> <li>• 4.9 million words, POS tagged, syntactically parsed</li> <li>• <a href="http://www.cis.upenn.edu/~treebank/">http://www.cis.upenn.edu/~treebank/</a></li> </ul>
<b>parallel corpus</b>	• Chemnitz Corpus	<ul style="list-style-type: none"> <li>• English and German translations of literary and scientific texts</li> </ul>
	• Hungarian–English Professional Corpus	<ul style="list-style-type: none"> <li>• English texts and their Hungarian translations in several disciplines (Heltai, 2007)</li> </ul>
<b>comparable corpus</b>	• International Corpus of English	<ul style="list-style-type: none"> <li>• a 1 million-word corpus of varieties of English worldwide, each corpus contains the same proportion of different genres and text types</li> <li>• <a href="http://ice-corpora.net/ice/design.htm">http://ice-corpora.net/ice/design.htm</a></li> </ul>
<b>learner corpus</b>	• International Corpus of Learner English	<ul style="list-style-type: none"> <li>• over 3 million words of written English produced by learners of English from several different linguistic backgrounds</li> <li>• <a href="http://cecl.fltr.ucl.ac.be/Cecl-Projects/lcle/icle.htm">http://cecl.fltr.ucl.ac.be/Cecl-Projects/lcle/icle.htm</a></li> </ul>
	• Hungarian Corpus of Learner English	<ul style="list-style-type: none"> <li>• a 2.4 million-word corpus of argumentative essays and theses of Hungarian students (Károly &amp; Tankó, 2009)</li> <li>• <a href="http://real.mtak.hu/1666/">http://real.mtak.hu/1666/</a></li> </ul>
<b>specialised corpus</b>	• British Academic Spoken Corpus	<ul style="list-style-type: none"> <li>• 1,644,942 tokens of recordings and transcriptions of 160 lectures and 39 seminars in a range of departments, at both undergraduate and postgraduate level</li> <li>• <a href="http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/">http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/</a></li> </ul>
	• Professional English Corpus	<ul style="list-style-type: none"> <li>• 28-million-word corpus of English used by professionals in science, engineering, technology and other fields</li> <li>• <a href="http://www.perc21.org/corpus_project/index.html">http://www.perc21.org/corpus_project/index.html</a></li> </ul>

Table 5. Kinds of corpora

#### 4.1.2.3. Concordance, collocation and annotation

An additional facility machine-readable corpora can make use of is concordancing. With the help of specially designed software, words or phrases of a corpus can be viewed in context in the form of so called **concordance lists**. Concordance lists are very often used in lexical research, and when editing dictionaries. An example of a concordance list is shown in Figure 8.



e runway and screeched to a halt. The **first** airport truck was there in seconds. nomic orthodoxy. I want to consider **first** an aspect of DIYE which I call "unre of 3 is 1. We don't know when it was **first** appreciated that the unit fraction, deepen my sorrow. I worship you as my **first** beginning; I long for you as my last e galley putting things away when the **first** buzzers sounded. One of the hostesse 3, to denote 3/5. A papyrus from the **first** century AD has the fraction in our m available for use. For example, Varro (**first** century BC) mentions only 12 such fr

Figure 8. Concordance list (McEnery & Wilson, 1996b)

Concordancing is also used to identify the lexical items that often co-occur in a corpus. The concept applied to describe and analyse the typical company of individual lexical items is **collocation**. Collocational analyses have been used to describe frequently occurring lexical patterns in particular registers (e.g., Gledhill, 2000; Nelson, 2000, 2006; Shin & Nation, 2008). A detailed discussion of the concept of collocation together with the measures of its strength, will be given in Section 4.3.1.3.

In grammatical research **annotated corpora** are widely used. Corpus annotation or tagging, is the practice of adding explicit additional information to machine-readable text and the physical representation of such information (Oakes, 1998, p. 249). The additional information is most often the information of part-of-speech (POS) or tense in linguistics. In other fields of research, however, other additional information is necessary. In discourse analysis, for example, the functions of certain discourse stretches can be categorised and codes of these categories can be added to the corpus. Figure 9 illustrates a corpus with POS tags and Figure 10 gives examples of discourse tags used in the London-Lund Corpus.

```

Perdita&NN1-NP0; ,&PUN; covering&VVG; the&AT0; bottom&NN1; of&PRF; the&AT0;
lorries&NN2; with&PRP; straw&NN1; to&TO0; protect&VVI; the&AT0; ponies&NN2;
'&POS; feet&NN2; ,&PUN; suddenly&AV0; heard&VVD-VVN; Alejandro&NN1-NP0;
shouting&VVG; that&CJT; she&PNP; better&AV0; dig&VVB; out&AVP; a&AT0;
pair&NN0; of&PRF; clean&AJ0; breeches&NN2; and&CJC; polish&VVB; her&DPS;
boots&NN2; ,&PUN; as&CJS; she&PNP; 'd&VM0; be&VBI; playing&VVG; in&PRP;
the&AT0; match&NN1; that&DT0; afternoon&NN1; .&PUN;

```

Figure 9. POS tags (McEnery & Wilson, 1996b)

- "apologies" e.g. *sorry, excuse me*
- "greetings" e.g. *hello*
- "hedges" e.g. *kind of, sort of thing*
- "politeness" e.g. *please*
- "responses" e.g. *really, that's right*

Figure 10. Discourse tags by Stenström (1984) (McEnery & Wilson, 1996b)

#### 4.1.3. Units of analysis in corpus research

The starting point for most analyses of corpora is the frequency list created by corpus processing tools. These tools transform texts into lists by reducing all recurring tokens into types, that is “each instance (*token*) of the word THE is counted but the complete list displays THE only once as a *type*, usually together with its frequency (the number of tokens found)” (Scott & Tribble, 2006, p. 13). In some cases it is useful to further reduce the types into **lemmas**, that is, to group the different word forms of “the same stem and belonging to the same major word class” (Francis & Kučera, 1982, p. 1). In the present study, as generally in the literature, lemmas are represented by words in small capitals, for example NOTIFY refers to the group including the forms *notifying, notified* and *notifies*. Previous corpus studies creating word lists for pedagogic purposes have applied a further reduction of word types into **word families**. As there is no generally accepted representation specific to word families in the literature, the present study presents word families by their headwords in small capitals in italics like *ANALYSE*. Word families include not only inflected, but also derived forms of the same base word or headword. The notion of the word family is defined in detail in Section 6.4.1.1. All these notions refer to single-word items. Corpus analysis tools, however, are also capable of computing and counting MWIs. In general, they are referred to as ‘clusters’ (Scott

& Tribble, 2006) or 'n-grams' (Forchini & Murphy, 2008). The unit of analysis applied in the present study for investigating MWIs is the **lexical bundle** proposed by Biber et al. (1999). The difference between 'clusters', 'n-grams' and 'lexical bundles' is that a MWI has to occur with a certain frequency in the corpus – usually 20 or 40 times per million words – in order to qualify as a 'lexical bundle'. The notion of the 'lexical bundle' is defined in detail in Section 6.4.3.

#### **4.1.4. Approaches to corpus analysis**

The characteristics of corpus linguistic analyses are summarised by Biber et al. (1998) as follows:

- it is empirical, it analyses the actual patterns of use in natural texts;
- it utilises a large and principled collection of natural texts, known as a “corpus”, as the basis for analysis;
- it makes extensive use of computers for analysis, using both automatic and interactive techniques;
- it depends on both quantitative and qualitative analytical techniques. (p. 4)

The novelty of an approach with these characteristics is its scope and reliability of analysis, which would otherwise not be possible. Most of the advantages of this approach, as stressed earlier, result from the use of computers. Computers can store large amounts of data, keep a record of stages of analyses, and they can provide consistent and reliable findings.

In addition, Sinclair (1991), advocated new approaches that challenge traditions of linguistic description fundamentally. The reason of the need for new approaches became clear as findings of early computer-based investigations were found to be in conflict with linguistic categories and descriptions based on intuition and introspection. Sinclair stressed the importance of evidence for linguistic analysis, and the main source of evidence for him was the corpus.



Along these lines Tognini-Bonelli (2001) distinguished between two approaches to corpus work. On the basis of the theoretical stance, the collection of a corpus, and its application for scientific research, the two approaches are the **corpus-driven** and the **corpus-based** approach. The main characteristics of the two approaches are summarised in Table 6. The difference between the two approaches is primarily theoretical. Linguists applying a corpus-based approach, based on the Chomskian tradition, give primacy to subjective introspection over objective, empirical data. Aarts (1991), for example, based his decision on which types of phenomena to include in the grammar, and which to exclude, ultimately on intuition.

Tognini-Bonelli (2001) also claimed that researchers of the corpus-based approach use linguistic categories, and syntactic structures already defined, which makes it difficult to detect patterns that emerge from the data itself. An example of this is the wide-spread application of annotation or tagging where words and phrases in the corpus are classified, and this might, in her view, result in loss of information from the data.

As this theoretical distinction of approaches is fairly new, the distinction is not yet used in a consistent manner in the literature. Although reference is made to these two different approaches, studying the titles of corpus-related journal articles the term 'corpus-based' is prevalent irrespective of the actual theoretical stance of the author.

<b>corpus-based approach</b>	<b>corpus-driven approach</b>
<ul style="list-style-type: none"> <li>• corpus is used to exemplify, quantify theories</li> <li>• classical relationship between data and theory</li> <li>• theoretical statement pre-exists corpus evidence</li> <li>• insulates, reduces, standardises corpus data</li> <li>• information from syntactic patterns have priority, categories of the system are pre-determined</li> <li>• intuition is primary</li> <li>• Chomskian framework</li> </ul>	<ul style="list-style-type: none"> <li>• corpus is used as evidence</li> <li>• linguist revises theory if necessary in light of the corpus data</li> <li>• theory derives directly from corpus evidence</li> <li>• builds corpus data into theoretical categories</li> <li>• systematic interrelations of syntactic and lexical patterns determine the categories of the system</li> <li>• intuition is not comprehensively reliable, objectivity is necessary</li> <li>• Firthian framework</li> </ul>
<b>scholars applying this kind of methodology e.g.</b>	
Aarts, Biber, Halliday, Leech	Cermak, Francis, Kennedy, Sinclair, Tognini-Bonelli,

Table 6. Characteristics of the corpus-based and corpus-driven approaches towards corpus research (based on Tognini-Bonelli, 2001, pp. 65-99)

#### **4.1.5. Reasons for using corpora in linguistic research**

Linguists applying a corpus linguistic approach to their field of study emphasise different aspects of it as the factor behind their decision. In this section, based on Nelson (2000, pp. 207-209), these different reasons for using corpora are outlined.

- *Empirical data vs. introspection* (Biber et al., 1998; Cermak, 2002; Knowles, 1996; Nelson, 2000; Sampson, 1996; Sinclair, 1991)

Researchers, although not necessarily to the same extent, as we saw earlier, recognise the objective power of language corpora. One of the main reasons for applying corpora for investigation in the different fields is the objective, quantitative data it can provide (see Table 7). In connection with objectivity, verifiability of results, quantitative data, accountability and reliability are also mentioned in the literature as benefits of corpus-related research. Reliability is important from two further perspectives (Biber et al., 1998, p. 4). Firstly, computers are less likely to make mistakes in automatic analyses, than humans are. Secondly, the evidence produced by the empirical investigation of corpora of authentic, natural texts,

can result in unambiguous findings, for example, frequency data of particular lexical items, which is not possible by introspection.

- *Broad range of data* (Biber et al., 1998; Leech, 1991; Nelson, 2000; Sinclair, 1991; Svartvik, 1996)

The need for analysing large amounts of data is widely discussed in the literature. As these analyses look for what is typical in the language as a whole, or for a certain register or language variety, it is only feasible if the database that is analysed contains a large amount of data. With enormous corpora of more 100 million running words, these analyses became possible.

- *Access* (McEnery & Wilson, 1996a; Nelson, 2000; Sinclair, 1991; Svartvik, 1996)

As many of these corpora are available either on CD-ROMs or via the Internet, it is easy to access them. The advantages of this are twofold. Firstly, findings become comparable as studies can be conducted on the same corpus. Secondly, as researchers all over the world can access the data, non-native speakers will have the same possibilities as native speakers which was not the case with an intuition-based approach.

- *Broad scope of analysis* (Biber et al., 1998)

As mentioned earlier many of the developments within corpus linguistics can be attributed to advances in computer technology. This is not only true of, for example, the size of the corpora, but also of the types of analyses that can be conducted on the corpora. Concordance and annotating software provide great possibilities for an array of different studies.

- *Speed* (Biber et al., 1998)

Computers not only work more accurately than humans, but also they are much quicker in automatic analyses.

- *Pedagogic reasons* (Nelson, 2000; Tribble & Jones, 1997; Wichmann, Fligelstone, McEnery, & Knowles, 1997)

Pedagogic reasons such as, authenticity, face validity and motivation are also among the reasons why scholars choose corpus linguistic methodologies.

#### **4.1.6. Fields enjoying the benefits of corpus linguistics**

The enumerated advantages of corpus-related analysis were recognised by researchers of different aspects of linguistic study. The fields, mostly benefiting from these advantages, have been lexical studies in general, and lexicography, in particular. Over the last few decades, with the advances in computer technology more and more areas within and without linguistics can make use of corpus data. Table 7 presents a summary of the fields that use corpora, and it also indicates what benefits a corpus linguistic approach has brought to this specific field. The list is by no means exhaustive. Other areas include psycholinguistics where corpus research lead to the finding that speech errors are more common than one would think, and it also plays a role in analysis of language pathologies. A field, outside linguistics, that made use of corpus data is, for example, social psychology (McEnery & Wilson, 1996a).

<b>Field of linguistics</b>	<b>Benefits gained from corpus linguistics</b>
lexical studies, lexicography	<ul style="list-style-type: none"> <li>• quick analysis of sheer data</li> <li>• lexical patterns emerge which could not be analysed earlier (e.g. collocation, usage)</li> <li>• authenticity</li> </ul>
grammatical studies	<ul style="list-style-type: none"> <li>• patterns can be analysed</li> <li>• shed light on lexicogrammatical interdependences</li> <li>• authenticity, empirical data</li> <li>• representativeness</li> <li>• quantitative data</li> </ul>
speech research	<ul style="list-style-type: none"> <li>• broad range of data</li> <li>• authenticity, naturalistic speech</li> <li>• annotation makes comparisons between different categories possible</li> </ul>
language teaching	<ul style="list-style-type: none"> <li>• authenticity</li> <li>• representativeness</li> <li>• criticism towards non-empirically based teaching materials</li> </ul>
language varieties	<ul style="list-style-type: none"> <li>• corpora used as testbed for theories</li> <li>• representativeness</li> <li>• quantitative data</li> </ul>
semantics	<ul style="list-style-type: none"> <li>• objectivity</li> <li>• frequency data to establish categories (e.g. fuzzy categories)</li> </ul>
historical linguistics	<ul style="list-style-type: none"> <li>• reservations of representativeness as limited availability</li> <li>• frequency analysis</li> <li>• study the evolution of language through time</li> </ul>
stylistics	<ul style="list-style-type: none"> <li>• quantitative data</li> </ul>
contrastive studies, translation	<ul style="list-style-type: none"> <li>• semantic, pragmatic contrastive analysis</li> <li>• analysis of translationese</li> </ul>
pragmatics	<ul style="list-style-type: none"> <li>• limited - difficult to automate</li> <li>• role of certain words, phrases or pauses in conversation</li> </ul>
discourse analysis	<ul style="list-style-type: none"> <li>• limited – difficult to automate</li> <li>• co-reference</li> <li>• speech acts</li> </ul>
sociolinguistics	<ul style="list-style-type: none"> <li>• limited - tradition of elicited data</li> <li>• authenticity</li> <li>• quantitative data</li> </ul>

Table 7. Fields of linguistics that use corpora (based on Meyer, 2002; McEnery & Wilson, 1996a; Partington, 1998)

#### 4.1.7. Criticism of corpus linguistic studies

Although corpus linguistics has already become an accepted field within linguistics there are still aspects that are rightly criticised, and maybe other aspects where criticism is

unjustified. Nelson (2000, pp. 210-212) enumerated the most relevant points of criticism and their refutation by corpus linguists. These are outlined below:

- *Focus on performance vs. competence*

The criticism that corpus linguistics focuses on performance and not on competence brings us back to the theoretical stance of Chomsky. In his view, the aim of linguistics is to describe competence, or in Hallidayan terms, the system (Halliday, 1991). Much of the debate that goes on in the literature about the validity of corpus research is actually questioning the theoretical stance that it is the performance or instances that we can observe, and that should be the starting point of linguistic description (Sinclair, 1991; Stubbs, 2001a; Tognini-Bonelli, 2001). To reconcile these two theoretical stances is, however, outside the scope of corpus linguistics. As Nelson (2000) put it:

There is no denying that it is a lot easier to statistically count occurrences of words than it is to say *why* they are there in the first place, or why they occur in the pattern that they do. However, this is not a problem of corpus linguistic methodology per se, but a problem facing all linguistic analysis. Corpora give the opportunity to take advantage of the very best sources of information which can then be utilised to perform further analysis. (p. 210)

- *Lack of correspondence between native speaker intuition and corpus finding*

Another critical statement about corpus linguistics in the literature is that corpus findings do not correspond to native speaker intuition. Widdowson (2000), for example, stated that results of corpus analysis are “only a partial account of real language” (p. 7) as they do not reflect how native speakers intuitively think they use the language. The counter-argument here is based on the fact that humans are not observing their language use objectively and systematically, and they “tend to notice unusual occurrences” (Biber et al., 1998, p. 3), whereas corpus linguistics looks for what is typical. Stubbs (2001a) also added that corpus

linguistics sheds light on the fact that intuition is not a good starting point for linguistic description as he states that: “People do not talk as they believe they do, and corpus linguistics now often points out how radically intuition and use may diverge” (p. 151).

- *Emphasis on frequency figures*

As corpus linguistics looks for typical patterns, the statistical method to detect it is to look for what is frequent, that is, the frequency of certain items or patterns. We have to distinguish, however, raw frequency and interpretative significance, as they are not necessarily the same. An occurrence might be significant because it is not frequent in a corpus (Stubbs, 2001a). There are statistical methods and concordancing software to help corpus linguists to set apart what is frequent and significant, e.g. raw and normalised frequency counts suggested in Biber et al. (1998, p. 32). According to them, **raw frequency** is the actual frequency of occurrence of an individual word in a given corpus, and **normalised frequency** is the number of occurrences of an individual word per million words in respective corpora. Normalisation makes it possible to compare frequencies in corpora of different sizes (Biber et al., 1998).

In connection to the criticism of over-reliance on frequency, there is another aspect that is mentioned in the literature. According to this, unique instances in corpus linguistic studies might be overlooked. Stubbs (2001a), in his reply to Widdowson (2000), claimed that:

Corpus linguistics is not concerned with what happens to occur (at least once): indeed its methods are generally designed to exclude unique instances, which can have no statistical significance. It is concerned with a much deeper notion: what frequently and typically occurs. (p. 151)

- *Reliance on machines and automatic search*

It is argued that certain language phenomena, for example, collocations, are given disproportionate attention in research, because there is a well-suited technological tool for its investigation (Partington, 1998, p. 144). Other researchers (e.g., Owen, 1993) criticised corpus linguists for excluding intuition totally from linguistic research, and he claimed that they rely too much on machines and automatic search, which results in superficial findings and misinterpretation. The answer to this objection is that most corpus-related studies state that intuition plays a role when interpreting the data. What is important is that the bases for interpretation are the objective data from the corpus of natural texts, and primacy is given to the patterns that emerge from the data (Biber et al., 1998; Sinclair, 1991; Tognini-Bonelli, 2001). However, analysis based on corpus work should be complemented with methods from other approaches. We can also add that corpus-related research, a fundamentally quantitative type of analysis, “benefits as much as any field from (such) multi-method research, combining both qualitative and quantitative perspectives of the same phenomena” (McEnery & Wilson, 1996a, p. 77).

#### **4.1.8. Summary**

This chapter reviewed the main elements and approaches of an up and coming field of linguistics. The growing interest in corpus linguistic research, and relevant findings in the field, have proved that this is the right direction for linguistic description. Researchers, however, who conduct corpus studies, should not forget the limitations of corpus work, which were also discussed.

For research that involves analysing large amount of data and analysis that can be automated, corpus linguistics is the alternative to opt for. Corpus linguistics has long been a controversial area, but today it has become a widely-used way to investigate language. The



utility of corpus work is not really challenged any more, and the view given by Murison-Bowie (1996) expresses the prevailing view of the field:

The strong case suggests that without a corpus (or corpora) there is no meaningful work to be done. The weak case is that there are additional descriptive pedagogic perspectives facilitated by corpus-based work which improve our knowledge of the language and our ability to use it. (p. 182)

#### **4.2. Data-driven learning**

A recent application of findings and methods of corpus linguistics in language teaching is **data-driven language learning** (DDL) (Boulton, 2007; Cobb, 1997; Johns, 1991a, 1991b; Stevens, 1991). The first advocate for using corpus data and concordance lines in language classrooms was Johns (1991a, 1991b), who noted that DDL is an

approach to foreign language learning that takes seriously the notion that the task of the learner is to 'discover' the foreign language, and that the task of the language teacher is to provide a context in which the learner can develop strategies for discovery – strategies through which he or she can learn how to learn. (1991a, p. 1)

The concept that underlies the approach of DDL is that: "Research is too serious to be left to the researchers: that the language-learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data" (Johns, 1991a, p. 2).

In DDL classes students are given concordance lists and they analyse these authentic examples as "researchers". The theory behind the method is that DDL supports language learning, because students will remember better what they discover themselves (Hunston, 2002). In addition, Boulton (2007) claimed there is psychological evidence that for human beings in general it is easier to recognise patterns, than to work with abstract rules. Therefore

language learners should be given huge numbers of examples of real language use, rather than rules, and they should identify patterns themselves.

There have only been a few studies that looked into how effective DDL is. Cobb (1997) conducted a study with Arabic learners of English in order to establish empirical evidence for the effectiveness of using concordancing for learning new words. In his experiment, students who used online concordancing scored 12% higher on the final vocabulary tests.

Several studies reported on applying parallel corpora in language teaching in different settings, and with varied aims (Chujo, Utiyama, & Nishigaki, 2005; Jablonkai, 2007; Sass, 2007; St John, 2001). St John and Chujo et al. suggested ways and tasks for the application of concordancing with beginners. Sass (2007) argued for applying a Hungarian–English parallel corpus for language teaching. The proposed parallel corpus was used with Hungarian students for translation instruction at a tertiary level (Jablonkai, 2007). Results of these studies suggested that DDL can be used with learners at a beginner level, and at higher proficiency levels as well, and that learners usually found the method both meaningful and useful.

In order to prepare suitable and useful DDL exercises to language learners, appropriate corpora need to be compiled, and the selection of areas for DDL activities should be based on research. This is especially relevant to ESP, where language learners' attention should be drawn to the specific linguistic and discourse characteristics of their specific disciplines or professional fields. The following chapter will outline what corpus research has brought to ESP, and how the findings of such research have been used in the ESP teaching practice.

### **4.3. Corpus research and ESP**

In 2005 Mike Scott gave the title “Corpus Linguistics and ESP – is there a link?” to a presentation in São Paulo (Scott, 2005). During that presentation he proposed several links between the two disciplines, suggesting a firm ‘Yes’ as an answer to his rhetorical question. Since then these links have been strengthened by new perspectives, an abundance of

applications of corpus linguistics in ESP research, as well as and teaching, and has resulted in deeper insights into the discourse and lexis of several disciplines and professions. The next chapter will give details of these new perspectives, and will summarise the main findings of corpus research on ESP.

The areas within ESP, in which corpus research yielded relevant results, are the following: (1) **language knowledge**, especially lexis in ESP; (2) **context knowledge**, that is, knowledge of the social context in which specialised texts are used, and in which ESP learners will use their English (Tribble, 2000); (3) **discourse competence**, that is, describing discourse features of certain academic and professional genres; (4) **course and materials design** and, more specifically, applications of DDL, lexical approach, and evaluation of existing teaching materials; and (5) **cross-linguistic analysis**. In what follows, a brief overview will be given of the findings and methods of corpus research applied, in order to gain insights into specific features of ESP texts.

As regards the analytical frameworks applied in ESP corpus research, most studies apply a combination of quantitative and qualitative analyses. The most frequently used quantitative or frequency-based frameworks, include the multidimensional analysis by Biber (1988; Biber et al., 1998; Conrad & Biber, 2001; Reppen, Fitzmaurice, & Biber, 2002), the analysis of collocations (e.g., Gledhill, 2000; Nelson, 2000, 2006), lexical bundles (e.g., Biber & Barbieri, 2007; Biber et al., 2004; Cortes, 2004, 2006; Cortes & Csomay, 2007; Jablonkai, 2009a, 2009b), and key words of certain registers and genres (e.g., Tribble, 2000). Qualitative analyses applied in ESP are concordancing, and the analyses of semantic prosody of certain lexical items (e.g., Nelson, 2000, 2006). Several studies examine linguistic and discourse features of certain academic and professional genres, and contrast these to characteristics of general English (e.g., Jablonkai, 2009a; Nelson, 2000, 2006; Trebits, 2009a, 2009b).

The academic genres being studied covered several disciplines like Engineering, English, Computer Sciences, Biology, Medicine, Business Studies, Applied Linguistics, History and Law. These corpora usually comprised textbook extracts (e.g., Biber et al., 2004; Biber & Barbieri, 2007), lectures (e.g., Cortes & Csomay, 2007; Nesi & Basturkmen, 2006), and expert and student writings (e.g., Scott & Tribble, 2006; Cortes, 2004). Professional genres were represented by a smaller number of subject fields that included BE (e.g., Nelson, 2000) EU Phare projects (e.g., Tribble, 2000), and financial journalism (e.g., Forchini & Murphy, 2008). Corpora in these analyses contained genres like introduction to guest speakers, letters of application (Henry & Roseberry, 2001), job advertisements, tourist brochures (e.g., Thompson, 2001), and EU Phare project proposals (e.g., Tribble, 2000). In what follows, the five areas of ESP with relevant contributions from corpus research will be reviewed, with the description of important methodological issues.

#### **4.3.1. Analysis of the language of specific disciplines**

Although ESP research into the language of specific disciplines has drawn on the findings of analysis of large reference corpora like the BNC or the Bank of English (e.g., Mudraya, 2006; Nelson, 2000; Trebits, 2009a), there has been a very strong tendency among researchers in ESP to build their own specialised corpora for their own specific purposes.

Flowerdew (2004) presented the case for specialised corpora to examine academic and professional language, and argued that general corpora, although in some cases comprising a wide-variety of sub-corpora, are not suitable for the type of analyses needed to investigate a special language variety, for several reasons. Firstly, general corpora were created to be representative of the language as a whole; therefore, this kind of corpora reflect the importance and weight of specific genres and text types in British or American culture, and are not representative of certain disciplines and professional fields. Secondly, although some general corpora contain specialised sub-corpora, it is usually difficult to access such a sub-

corpus as the search fields were not designed for such purposes. Thirdly, some genres are very difficult to gain access to as they are not in the public domain, or even in some cases, are semi-confidential or even secret. In particular, spoken discourse data in general, and in business or other professional settings belong to this category (Nelson, 2000; 2006). Therefore, it would be too time-consuming to compile such data for general corpora. Finally, many general corpora comprise text extracts, rather than whole texts. This implies that such corpora are mainly suitable for the analysis of individual lexical or grammatical items, and cannot be used for genre-based analyses to identify the discourse function of particular lexical or grammatical items in different parts of the text.

As regards the size of corpora, there is a tendency to analyse small corpora, especially compared to the huge size of monitor or reference corpora, as described in Section 4.1. The size of corpora used for ESP ranges from 32,000 running words (Conrad, 1996) to 20 million running words (Chujo & Utiyama, 2006), and the most frequent size of specialised corpora is between 1 and 5 million running words (Coxhead, 2000; Mudraya, 2006; Nelson, 2000; Wang et al., 2008).

The analysis of the language of specific fields focused mainly on aspects of the lexis of the fields under study. The type of analysis that has yielded relevant findings include, compiling **word lists**, identifying **key words**, and analysing **collocations** in specialised corpora. The following sections will outline the results of these analyses. A summary of the literature reviewed will be presented in Table 12 at the end of the section. As these types of analyses are of special interest to the present study, aspects of methodology will also be discussed in detail.

#### **4.3.1.1. Word lists in ESP**

As outlined in Section 2.2.6.1, lexis in specialised texts is grouped into three main categories: technical, semi-technical and general lexis. This distinction has been created,

partly based on the meaning of lexical items, and partly on their frequency in specialised texts. With the development of corpus tools, however, it has become feasible to evaluate the established categories based on computerised analyses of specialised corpora. Based on empirical evidence, Nation (1990) identified four levels of lexis for language courses: (1) high frequency or **general service lexis**, (2) academic or **semi-technical lexis**, (3) **technical lexis** and (4) **low frequency lexis**. The lexis of these categories can be characterised in terms of frequency, range and text coverage (Chung & Nation, 2003, 2004; Nation & Hwang, 1995; Sutarsyah, Nation, & Kennedy, 1994). **Text coverage** is defined as the percentage of the tokens, that is, instances of words, in a corpus that are covered by the elements of a particular word list (Nation & Hwang, 1995). Earlier research on lexis in specialised texts found that low frequency words cover about 5% of the tokens in specialised texts, and technical words usually account for another 5% (Nation, 1990). Research into the text coverage of technical words in texts of particular disciplines, however, revealed that the coverage of technical lexis can be considerably higher, for example, in an anatomy text, the text coverage of technical words can reach as high as 31.2%, and in an applied linguistics text as high as 20.6% (Chung & Nation, 2003). The methodology for identifying academic lexis and defining the text coverage of academic words in a particular text will be discussed later in this section.

High frequency or **general service lexical items** in research into lexis in ESP are often represented by the *General Service List of English Words (GSL)* edited by West (1953). The aim of his compilation was to establish a list of lexical items that the learners of English as a foreign language should start with when learning English. Despite its age, some errors, and the fact that it had been created based on a written corpus, the GSL is still widely referred to, and applied as, the first most frequent 2000 words for EFL learners (e.g., Coxhead, 2000; Nation & Waring, 1997; Wang et al., 2008). Research into lexis in specialised texts has found that the GSL typically covers 70-75% of tokens of texts (Nation & Hwang, 1995), for

example, in Economics (Sutarsyah et al., 1994) and Applied Linguistics (Chung and Nation, 2003).

With the advance of computer technology and accessibility of corpora and corpus analysis software programmes, the task of creating word lists based on frequency lists of general and specialised corpora became feasible for individual researchers and teachers of ESP. Mudraya (2006), for example, created the Student Engineering Word List (SEWL) with 1260 word families based on her 2-million-running-word corpus of textbooks on basic engineering disciplines, such as Engineering Mechanics, Engineering Materials, Manufacturing Processes and Computer Programming. She aimed at developing a reliable English for Engineering syllabus for students in Thailand who needed to study from English-language textbooks for their engineering courses at a local university. After organising the initial frequency list of more than 18,000 word types by word families, the selection of the word families forming part of the final engineering word list was carried out on the basis of the cumulative frequency of the members of the word families. The cut-off point was set at 100 occurrences, or 0.005% of the whole corpus. The first ten headwords with frequency information are given in Table 8 below.

N	Headword	Frequency	%
1	<b>use</b>	10,313	0.52
2	<b>force</b>	9247	0.46
3	<b>form</b>	7075	0.35
4	flow	7045	0.35
5	<b>pressure</b>	7016	0.35
6	<b>show (v)</b>	7002	0.35
7	<b>determine</b>	6896	0.34
8	<b>figure/configure</b>	6650	0.33
9	section	6404	0.32
10	<b>line</b>	5812	0.29

Table 8. The ten most frequent word families in the SEWL with the elements of the GSL in bold (based on Mudraya, 2006)

Comparing the SEWL with the GSL shows that the word list for engineering students contains many elements of general lexis which are highlighted in bold in Table 8. This might

be useful for courses where students do not have a sound basis in General English before specialising in the language of a discipline, but in most cases learners of ESP already possess the basic general lexis. Therefore, a word list that focuses on the frequent lexical items that are specific to the given discipline is more useful for ESP course and materials design.

The way word list compilers controlled for specificity was that the word families of the GSL were excluded from among the frequently occurring word families in the specialised corpora of the discipline. The first example of such a word list is the **Academic Word List (AWL)** (Coxhead, 2000). The aim of the list was to replace similar earlier lists (Ghadessy, 1979; Xue & Nation, 1984) that had been compiled without the help of electronic corpora and also to serve as the basis of language courses for academic purposes. The corpus used for the analysis was the 3.5-million-word Academic Corpus, which contained more than 400 academic texts like journal articles, course books, and laboratory manuals. The corpus was made up of four sub-corpora: arts, commerce, law, and science, containing about 875,000 running words and each was subdivided into seven subject fields, for example, Education, History, Accounting, Economics, Criminal law, Rights and remedies, Biology, and Chemistry. The selection of word families was guided by the following principles: (1) ensuring **specialised occurrence** by including word families in the final AWL that are outside the GSL representing the first 2,000 most frequent English words; (2) requiring that word families represent the lexis of several academic disciplines by determining a minimum **range**, that is, a member of a word family had to have an occurrence higher than 10 in each of the main sub-corpora and had to occur 15 times or more in the 28 subject fields; (3) setting a minimum **cumulative frequency** of occurrence of a word family – that was defined as the sum of the frequencies of the members of the word family – at higher than 100 in the Academic Corpus.



In the course of the selection process priority was given to range over frequency in order to avoid bias towards longer texts and topics. It meant that word families with more members had to have a cumulative frequency of 100, whereas word families with a single member were included with a frequency less than 100. The least frequently occurring single-member word family was the word *forthcoming*, with a frequency of 80. Coxhead's final AWL (2000) contains 570 word families, and in order to assist the sequencing of teaching, it is presented in frequency-based sub-lists.

Coxhead (2000) evaluated the AWL by testing its text coverage in the Academic Corpus, another corpus of academic texts, and a corpus of fiction texts. Results of these analyses indicated that the AWL is a truly academic word list, as it accounted for 10.0% of all the tokens in the Academic Corpus, it covered 8.5% of the second academic corpus, and its coverage in the corpus of fiction texts was only about 1.4%.

Chen and Ge (2007) investigated the text coverage of the AWL in medical research articles. They concluded that, although word families in the AWL represent a high text coverage – slightly more than 10% – in medical research articles, only 51.2% of all word families in the AWL were frequently used in their corpus of medical research articles. Encouraged by the findings of Chen and Ge's research, Wang et al. (2008) established the Medical Academic Word List (MAWL) of 623 word families frequently used across various subfields of medicine.

Following Coxhead's (2000) methodology the compilation of the MAWL was based on a one-million-word corpus of medical research articles of 32 different sub-fields of medicine like Urology, Health Informatics, Gastroenterology, Surgery, etc., and the final word families were selected according to similar criteria defining (1) **specialised occurrence**, (2) **range**, and (3) **frequency**. **Specialised occurrence** was understood in the same way as in Coxhead's study, that is, only word families outside the 2000 word families of the GSL were included.

**Range** was defined as the minimum number of occurrences of members of word families in the 32 sub-fields at 16, that is, word families had to be applied in at least half of the sub-fields of medicine. The criterion **frequency** was set at 30 for the cumulative occurrence of word families in the whole corpus of medical research articles. Wang et al. (2008) argued, that because their corpus is approximately a third of Coxhead's Academic Corpus, the criterion frequency was set at the third of Coxhead's frequency requirement of 100. They also applied an additional step in the selection process of the final MAWL, which was consulting two experienced professors of English for Medical Purposes, who made decisions on the inclusion or elimination of controversial word families.

The analysis of the MAWL included testing its text coverage in the corpus of medical research articles and comparing it to the AWL. The MAWL was found to cover 12.24% of the total corpus which is slightly higher than the text coverage of the AWL in academic texts. The comparison of the two word lists showed that only 342 (54.90%) of the word families of the MAWL can be found among the word families of AWL. On the basis of these results, they argued that different disciplinary discourses operate with their own subject-specific lexis, which makes a general academic word list less valuable for individual disciplines.

#### **4.3.1.2. Key word analysis**

Based on the assumption that technical words appear in specialised, technical texts, or occur in higher frequencies in such texts, several studies applied a lexical analysis for exploring genres and registers relevant for their particular disciplines (Chung & Nation, 2004; Mudraya, 2006; Nelson, 2000, 2006; Tribble, 2000).

In their article on technical lexis for teaching purposes, Chung and Nation (2004) compared four approaches to identifying technical words in a text on anatomy. The three approaches they evaluated were: (1) clues like labels in diagrams, definitions given by the author, (2) applying a technical dictionary, (3) comparison of frequency counts of technical

words in the technical text with frequency counts of technical words in a general corpus. The control approach was a four step rating scale, which had a high degree of reliability. Their findings suggested that the third approach, based on frequency counts, proved most reliable for identifying technical words, although it did not recognise words like *neck*, *chest*, *skin* which were frequently used in the non-technical corpus as well.

On the basis of this frequency-based approach, several corpus analysis tools provide the function **Key word analysis** (e.g., *WordSmith Tools*, Scott, 2004; *AntConc*, Anthony, 2007) by which frequency counts of the lexical items in the specialised corpus can be automatically compared to their frequency counts in a general, reference corpus. These tools also apply statistical tests, for example, log likelihood, to evaluate the difference between frequency counts. The resulting key word lists contain positive key words, that is, words that occur with unusually high frequency in the specialised corpus, and negative key words, that are unusually infrequent in the specialised corpus (Nelson, 2000; Scott, 1997, 2000, 2004; Scott & Tribble, 2006; Tribble, 2000). A key word analysis is often performed as a first step in the analysis of specialised corpora, in order to provide the investigation with lexical items for further analysis (e.g., Flowerdew, 1998; Nelson, 2000, 2006; Tribble, 2000).

#### **4.3.1.3. Collocational analysis in ESP**

One of the types of analysis within corpus linguistics that probably benefited the most from the advances in computer technology, has been automated **collocational analysis**. Collocational analysis has been widely used in corpus studies with a lexicographic focus (e.g., Kilgarriff & Rundell, 2002; Krishnamurthy, 2008; Walker, 2009), and also in corpus research with pedagogic aims (e.g., Nelson, 2000, 2006; Shin & Nation, 2008; Ward, 2007). There have been, however, few attempts to analyse collocations in specialised texts for ESP teaching purposes. These studies concentrated on collocations in engineering (Ward, 2007), pharmaceutical (Gledhill, 2000), and business (Nelson, 2000, 2006) texts. In this section an

overview of the findings of these studies will be given after defining the concept of collocation in general.

Although the concept of collocation is widely applied in corpus linguistics, applied linguistics, and in language teaching, a widely accepted, clear-cut definition is not yet available. The concept itself was introduced by Firth (1968) in the 1950s, and it was elaborated by Sinclair (Renouf & Sinclair, 1992; Sinclair, 1987; Sinclair, 1991; Sinclair, Jones, Daley, & Krishnamurthy, 2004), and his colleagues over the coming decades.

Partington (1998), by looking at different definitions of **collocation**, identified three approaches to the concept that highlight different but related aspects of collocation. The ‘textual’ definition is illustrated by Sinclair’s definition: “Collocation is the occurrence of two or more words within a short space of each other in a text” (Sinclair, 1991, p. 170).

According to this definition, a lexical item collocates with another if it occurs within a certain collocational **span** or window with the **node**, the word under scrutiny, in a given text. The word which frequently co-occurs with the node is referred to as the **collocate** (Hunston, 2002; Scott & Tribble, 2006). There is no agreement in the literature about the size of an appropriate span. Most studies apply a span between 2 to 5 words from the node to left or right (Stubbs, 1995), and Sinclair (1991) suggested that real collocates cannot be found outside a span of 4:4, i.e. four words to left or right from the node. Including this parameter, Stubbs (1995) defines collocation as follows:

By collocation I mean a relationship of habitual co-occurrence between words (lemmas or word forms). A node may be observed to occur with various collocates within a certain span or window, say 4:4, i.e. four words to left or right. (p. 23)

This definition raises the question of how the node should be defined. Is it one particular **word form** like *criteria* or *implements*, or is it all word forms of the same **lemma** like IMPLEMENT for *implementing*, *implemented* and *implements*? Some linguists argue that

collocates of word forms should be analysed, as even the form of the node influences what words it is likely to co-occur with (Knowles & Zuraidah, 2004; Renouf, 1987; Sinclair, 1991; Tognini-Bonelli, 2001). Others, however, use the lemma as node, for example of lexicographic and pedagogic purposes (Kilgarriff & Rundell, 2002; Kilgarriff & Tugwell, 2001, 2002).

The second aspect Partington (1998) identifies is the ‘psychological’ or ‘associative’ nature of collocations. By this he meant that collocation is “part of a native speaker’s communicative competence” (p. 16), that is, meaning of a word is learnt by looking at the items co-occurring with it in texts in the course of constant exposure to their mother tongue. The definition exemplifying this aspect of collocation is given by Leech: “Collocative meaning consist of the associations a word acquires on account of the meanings of words which tend to occur in its environment” (Leech, 1974, p. 20 as cited in Partington, 2004, p. 16).

The third group of definitions identified by Partington (1998) are the statistical definitions of collocation that focus on the strength of the association between the node and its collocates. An example of that is Hoey’s (1991) definition that collocation is “the relationship a lexical item has with items that appear with greater than random probability in its (textual) context” (pp. 6-7).

This definition highlights the fact that collocations are not absolute, but probabilistic events (Altenberg & Granger, 2001; Halliday, 1966; O’Keefe, McCarthy, & Carter, 2007) and they are the result of habitual co-occurrence of words in frequent combinations as used by speakers and writers of a language. The strength of collocations can be measured by several statistical methods like raw frequency, observed/expected, t-score, z-score, MI score, MI3 score, log likelihood, log log and salience (e.g., Hunston, 2002; Kilgarriff & Tugwell, 2001; Kilgarriff and Rundell, 2002; McEnery et al., 2006; Scott, 2004). Here the focus is on three

methods two of which, MI score and t-score, are applied most frequently in corpus studies and salience that is applied by the tool, *Sketch engine* (Kilgarriff & Tugwell, 2001), that was used for the collocational analysis in the present study. Comparing the lists of the first ten collocates of the node word *conversation* by the measures MI score and t-score, Kilgarriff and Rundell (2002) found that there are striking differences between the two lists. As can be seen in Table 9, collocates significant by the MI score include rather infrequent words like *stilted* and *peppered*, whereas t-score gives preference to very frequent words like the indefinite article and prepositions. Kilgarriff and Rundell argued that neither of these measures was really helpful for lexicographers, who are interested in words that can typically be found towards the middle of the t-score lists.

Collocates by MI score	Collocates by t-score
<i>overhearing</i>	<i>with</i>
<i>phatic</i>	<i>a</i>
<i>overhear</i>	<i>had</i>
<i>eavesdrop</i>	<i>in</i>
<i>snatches</i>	<i>telephone</i>
<i>stilted</i>	<i>between</i>
<i>transcripts</i>	<i>our</i>
<i>overheard</i>	<i>about</i>
<i>topic</i>	<i>into</i>
<i>peppered</i>	<i>phone</i>

Table 9. Comparing collocates by MI score and t-score in the *Cobuild Online Collocation Sampler* (based on Kilgarriff & Rundell, 2002, p. 810)

Therefore, Kilgarriff and Tugwell (2002) proposed a new measure for selecting relevant collocates for lexicographic purposes they called **salience**. Salience is the product of Mutual Information and log frequency. Based on the experience of lexicographers, Kilgarriff and Tugwell argued that multiplying MI score by log frequency brings the words relevant for the lexicographers' work to the top of collocate lists, as illustrated by the Word sketch for the node *conversation* in Table 11. Although these researchers concentrated on aspects important for creating learners' dictionaries, considerations for choosing relevant collocations for

language teaching purposes are very similar. Therefore, the suggested statistical measure can also be used to select relevant and appropriate collocates for the language classroom.

An additional form of collocation focusing on what a word typically does grammatically is **colligation** (Hoey, 2000). According to Hoey (2000), colligation of a word characterises the “grammatical company a word keeps and the position it prefers” (p. 234). Illustrating the concept, Baker, Hardie, & McEnery (2006) state that nouns often colligate with adjectives, whereas verbs frequently colligate with adverbs. The concept of colligation can be applied to words and phrases as well, for example, the word *window* often colligates with prepositions like in *coming in **through** his window, went over **to** the back window, a window **on** the City* (p. 36).

An approach that combines grammatical and lexical aspects of collocation has been proposed by Kilgarriff and Tugwell (2002). In their corpus analysis tool, the *Sketch engine*, collocations are presented in the form of **word sketches**. Creating word sketches, the tool selects the words that occur in a certain span of the node based on their salience and groups collocates according to the grammatical relations they form with collocates. The grammatical relations are stored as a pre-determined list of relations, considering aspects that are important from a lexicographic point of view. Examples of grammatical relations include: (a) for verbs: subject, object, modifying adverbs; (b) for nouns: subject of, object of, modifying adjectives; (c) for adjectives: noun complements, modifying adverbs; (d) for all three parts-of-speech: prepositional complements, and/or relations (Kilgarriff & Rundell, 2002, p. 813). Grammatical relations in the database of the *Sketch engine* are illustrated with examples in Table 10 with the node words highlighted in bold.

Grammatical relation	Example
bare-noun	<i>the angle of <b>bank</b></i>
possessed	<i>my <b>bank</b></i>
plural	<i>the <b>banks</b></i>
passive	<i>was <b>seen</b></i>
reflexive	<i><b>see</b> herself</i>
ing-comp	<i><b>love</b> eating fish</i>
finite-comp	<i><b>know</b> he came</i>
inf-comp	<i><b>decision</b> to eat fish</i>
wh-comp	<i><b>know</b> why he came</i>
subject	<i>the <b>bank</b> refused</i>
object	<i>climb the <b>bank</b></i>
adj-comp	<i><b>grow</b> certain</i>
noun-modifier	<i>merchant <b>bank</b></i>
modifier	<i>a <b>big</b> bank</i>
and-or	<i><b>banks</b> and mounds</i>
predicate	<i><b>banks</b> are barriers</i>
particle	<i><b>grow</b> up</i>
prep+gerund	<i><b>tired</b> of eating fish</i>
PP-comp/mod	<i><b>banks</b> of the river</i>

Table 10. Grammatical relations in the *Sketch engine*  
(based on Kilgarriff & Tugwell, 2002, p. 129)

node + preposition <i>with</i> (PP_with)	node + preposition <i>at</i> (PP_at)	node as object of
<i>friend</i>	<i>table</i>	<i>overhear</i>
<i>stranger</i>	<i>dinner</i>	<i>steer</i>
<i>passenger</i>	<i>time</i>	<i>record</i>
<i>people</i>	<i>party</i>	<i>tape</i>

Table 11. Extract from word sketch for *conversation*  
(based on Kilgarriff & Rundell, 2002, p. 812)

At a theoretical level, the concept of collocation can be regarded as a key concept supporting the ‘**idiom principle**’ Sinclair (1991) put forward based on his lexicographic work on large – many-million-word corpora. As he put it: “The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments” (Sinclair, 1991, p. 110).



In Sinclair's (1991) view of language the 'idiom principle' goes hand in hand with the '**open choice principle**'. He suggested that the 'idiom principle' is central to the creation of texts and meaning, and that the register and lexical choices determine and limit further choices, leaving fairly little room for the 'open choice principle'. As he argued: "Once a register choice is made, and these are normally social choices, then all slot-by-slot choices are massively reduced in scope, or even, in some cases, pre-empted" (Sinclair, 1991, p. 110).

Results of collocational analysis in ESP seem to support Sinclair's idiom principle as analyses of collocations in specialised corpora found that collocations are highly discipline-specific (Gledhill, 2000; Ward, 2007), and collocates of words become more fixed in specialised texts (Gledhill, 2000; Nelson, 2000, 2006).

Summarising the kinds of relations between lexical items, Partington (2004) based on Stubbs (2001b), lists four different kinds:

- (i) collocation: the relationship between lexical item and other lexical items;
- (ii) colligation: the relationship between lexical item and a grammatical category;
- (iii) semantic preference;
- (iv) semantic prosody (p. 145).

This section focused on the first two kinds, collocation and colligation. The next section illustrates how the analysis of semantic preference and semantic prosody can give insights into the context of specialised texts.

Area within ESP	Subfield	Research	Application within ESP
language knowledge	specialised corpus	<p><b>Cheng-yu</b> (1993): corpus of English of Computer Science</p> <p><b>L. Flowerdew</b> (2004): benefits and caveats of using specialised corpora</p> <p><b>Krause</b> (2005): validated a small, specialised corpus for linguistic analyses</p>	<ul style="list-style-type: none"> <li>• provides specific linguistic characteristics</li> </ul>
	lexis in ESP	<p><b>James, Davison, Heung-yeung, &amp; Deerwester</b> (1994): lexical analysis of an English in Computer Sciences corpus</p> <p><b>Sutarsyah, Nation, &amp; Kennedy</b> (1994): general lexis and semi-technical lexis should be learnt before students can focus on the lexis of their special fields of study</p> <p><b>Coxhead</b> (2000): a new Academic Word List</p> <p><b>Fuentes</b> (2001): identified three main types of lexical behaviour: general academic, specialised/technical, genre-specific</p> <p><b>Bowker &amp; Pearson</b> (2002): practical guide to compiling specialised corpora and creating glossaries</p> <p><b>Chung &amp; Nation</b> (2003): proportion of technical lexis in anatomy and applied linguistics texts is higher than 5% (31.2% and 20.6%)</p> <p><b>Chung &amp; Nation</b> (2004): technical lexis can be identified best by comparing frequency data in a general and a specialised corpus</p> <p><b>Mudraya</b> (2006): most frequent words in a specialised corpus are sub-technical and non-technical, created SEWL</p>	<ul style="list-style-type: none"> <li>• helps create subject-specific glossaries</li> <li>• makes it possible to create specific word lists based on real language use data</li> <li>• helps distinguish academic lexical items from subject words</li> <li>• makes it possible to identify high frequency technical lexis</li> </ul>

Table 12. ESP corpus research into the analysis of language

#### 4.3.2. Analysis of specific contexts

**Context knowledge** in ESP refers to the knowledge of the social context in which specialised texts will be used, and in which ESP learners will ultimately use their English. It is an extension of the context knowledge as defined, by Tribble (2000), focusing on ESP writing

skills as the “knowledge of the social context in which the text will be read” (p. 90). Table 13 gives a summary of ESP corpus research into the analysis of specific contexts. Tribble (2000) compared how the multi-dimensional analysis by Biber (1988) and the key word analysis by Scott (1997, 2000), can be used to gain insights into aspects of language and context knowledge. The key word analysis as outlined in Section 4.3.1.2, is fairly straightforward and available to the individual researcher. The **multi-dimensional analysis**, on the other hand, is a rather complex analysis that involves tagging 67 different features in the corpus, such as ‘past tense’, ‘perfect aspect’ and ‘present tense’; nominalisations ending in *-tion*, *-ment*, *-ness*, *-ity*; public verbs (*complain*, *explain*, *promise*), private verbs (*believe*, *think*, *know*), and then analysing the frequencies of these features statistically by factor analysis in order to group salient factors into dimensions, and thus allowing comparisons to other registers (Biber, 1988; Biber et al., 1998; Conrad & Biber, 2001). Tribble (2000), based on his comparison of the two methods, concluded that although the multi-dimensional analysis provided a more comprehensive picture of a register or genre, the key word analysis was also a powerful means of identifying relevant lexical items, and concepts in relevant registers and genres for language teaching purposes, and it had more potential in an ESP setting.

Another way to learn about the context of the language of a special field is to analyse semantic preference and semantic prosody of the key words or selected lexical items. Stubbs (2001b) defined **semantic preference** as “the relation, not between individual words, but between a lemma or word-form and a set of semantically related words” (p. 65). To illustrate this category of relation, Partington (2004) gives the semantic preferences of *sheer*. The semantic sets the word *sheer* was found to collocate with included (1) “magnitude”, “weight” or “volume”, e.g. *the sheer volume of reliable information*; (2) “force”, “strength” or “energy”, e.g. *the sheer force of his presence*; (3) “persistence”, e.g. *sometimes through sheer insistence*; (4) “strong emotion”, e.g. *sheer joy of life* and (5) physical quality, e.g. *the sheer*

*glamour of evil* (p. 145). Furthermore, he demonstrated that there is also interaction between typical syntactic behaviour of words and their semantic preferences. In the example of *sheer*, the typical structure for the first two semantic sets, that is, “magnitude” and “force” words, were found to be “*the sheer* (noun phrase) *of* (noun phrase)”, and in the third semantic category the word *sheer* was found to be often preceded by prepositions expressing means or manner, e.g. *through, out of, by*. An example of the analysis of semantic preference in a specialised corpus is the result of Nelson’s (2000, 2006) analysis of his BEC. Nelson (2006) found that, for example, the word *package* had a preference for being connected to computers, and it also shared a preference related to finance, with words like *merger* and *market*.

Stubbs (1995) also demonstrated that lexical items have a tendency to co-occur with negative or positive words. This phenomenon is usually referred to as semantic prosody in the literature (Partington, 2004). In his analysis of the word *cause*, for example, Stubbs (1995) found that the most frequent collocations of it are negative abstract nouns like *anxiety, concern, crisis* and many examples are from the medical field like *cancer, blood, death, disease*. Furthermore, Nelson (2006) claimed that looking at the semantic prosody of words as used in a business context not only reveals insights into language use, but also provides information about the business world as such. The example he gave here were semantic prosodies of the words *boss* and *manager*. According to his findings *boss* has a tendency to be used with negative adjectives such as *meanest, old-fashioned*, whereas *manager* displayed positive collocates like *excellent* and *good* (Nelson, 2000, 2006).

In addition, Partington (2004) also analysed the relationship between semantic preference and semantic prosody. He suggested that in most cases semantic prosody can be considered a sub-category of semantic preference, a special case that includes “instances where a lexical item shows preference to co-occur with items that can be described as bad, unfavourable or unpleasant, or as good, favourable or pleasant” (p. 149). In his further

analysis, however, he noted that “semantic preference is a ‘narrower’ phenomenon – relating the node to another item for a particular semantic set – than prosody which can affect wider stretches of text” (p. 151). He also illustrated how semantic preference contributes to build semantic prosody and how prosody in turn restricts the preferential choices of the node word. Partington also found that both phenomena add to the cohesion of discourse.

Area within ESP	Research	Application within ESP
context knowledge	<p><b>Knowles</b> (1996): word maps of lexical items designate prominent concepts in the concept structure of the discourse community</p> <p><b>Nelson</b> (2006): semantic prosody in business texts gives insights into business culture, collocates become more fixed in the business context</p> <p><b>Tribble</b> (2000) key word analysis of EU project proposals</p>	<ul style="list-style-type: none"> <li>• provides means to analyse semantic prosody that helps understand context</li> <li>• allows for more top-down, qualitative, contextually-informed analyses</li> <li>• ethnographic information can also be searched for in corpora</li> <li>• corpus data can provide learners with both language knowledge and context knowledge</li> </ul>

Table 13. ESP corpus research into the analysis of specific contexts

#### 4.3.3. Analysis of specific discourses

Several corpus studies in ESP analysed certain genres or registers in order to gain insights into how the discourse of certain disciplines and professional fields are construed, and to develop ESP learners’ **discourse competence**, that is, enabling them to read and produce relevant texts for their specific subject areas. Flowerdew (1998) in her review of how corpus linguistics had been applied to analyse genres and discourse, argued for a **corpus-based approach** and the application of corpora annotated at a semantic and discourse level, in order to make findings of corpus analyses more applicable for pedagogic purposes.

Other studies in the discourse of registers and genres of several academic disciplines, however, applied a **corpus-driven approach**, and yielded relevant results in identifying discourse characteristics without annotating corpora. Many of these studies applied the

frequency-based framework of analysing MWIs in texts proposed by Biber et al. (1999). A summary of the literature on the corpus linguistic analysis of specific discourses in ESP is presented in Table 14 (see pp.103-104).

#### 4.3.3.1. Academic discourse

As outlined in Section 2.2.6.2, MWI have been studied under several terms, like ‘fixed expressions’ (Moon, 1998, 2000), ‘lexical phrases’ (Nattinger & DeCarrico, 1992), ‘pre-fabs’, and ‘ready made units’ (Cowie, 1992). MWI in these studies were selected based on “perceptual salience” rather than empirical evidence in real language use (Biber & Conrad, 1999). With the advent of computer technology and corpus linguistics, it became possible to investigate longer sequences of words in discourse statistically, that is, researchers could focus on what is frequent instead of examining what stands out. The type of MWI that is defined by its frequency in particular corpora is the lexical bundles. The concept of **lexical bundles** was introduced in *The Longman Grammar of Spoken and Written English* (Biber et al., 1999). Biber et al. (1999) distinguished lexical bundles from both collocations and idioms. According to them, **idioms** are the most idiomatic and invariable type of MWIs, and they are usually structurally complete units. **Collocations**, on the other hand, are statistical associations between two words that are variable and not idiomatic in the sense that in a collocation words can be associated with several other words and they retain their own meaning (see Section 4.3.1.3). Biber and Conrad (1999) defined lexical bundles as “the most frequent recurring lexical sequences;” [...] “which can be regarded as extended collocations: sequences of three or more words that show a statistical tendency to co-occur (e.g., *in the case of the, do you want me to, I said to him*)” (p. 183). The concept of lexical bundles has been used in several later studies (e.g., Biber & Barbieri, 2007; Biber et al., 2004; Biber & Conrad, 1999; Cortes, 2004; Hyland, 2008; Nesi & Basturkmen, 2006) to investigate common MWIs in discourse focusing mainly on registers in university and academic contexts.

Area within ESP	Subfield	Research	Application within ESP
discourse competence	-	<p><b>L. Flowerdew</b> (1998): reviews corpus linguistics in ESP, suggestions for how specialised corpora can be explored for discourse analysis</p>	<ul style="list-style-type: none"> <li>• findings of tagged, specialised corpora can inform language teaching better</li> </ul>
	academic discourse	<p><b>Conrad</b> (1996): defines characteristics of academic texts in Biology</p> <p><b>Williams</b> (1998): registers consist of central conceptual frameworks realised linguistically through relatively closed collocational networks</p> <p><b>Gledhill</b> (2000): analysis of collocations in pharmaceutical research articles</p> <p><b>Bondi</b> (2001): argumentative procedures in academic and economics textbooks and research articles</p> <p><b>L. Flowerdew</b> (2004): more qualitative analyses can be carried out on specialised corpora</p> <p><b>Biber, Conrad, &amp; Cortes</b> (2004): lexical bundles are building blocks of discourse</p> <p><b>Cortes</b> (2004): lexical bundles in student and professional writing in History and Biology</p> <p><b>Cortes</b> (2006): lexical bundles in History</p> <p><b>Nesi &amp; Basturkmen</b> (2006): lexical bundles in academic lectures</p> <p><b>Biber &amp; Barbieri</b> (2007): lexical bundles give a frame for new propositional assertions</p> <p><b>Cortes &amp; Csomay</b> (2007): lexical bundles in classroom teaching</p> <p><b>Hyland</b> (2008): lexical bundles show disciplinary varieties</p>	<ul style="list-style-type: none"> <li>• multidimensional analysis investigates many linguistic features to describe a language variety</li> <li>• informs about the 'aboutness' of texts</li> <li>• makes it possible to combine corpus analysis and genre analysis</li> <li>• general corpora cannot provide adequate linguistic evidence in academic and professional discourse</li> <li>• makes it possible to create dictionaries with pedagogic value based on the description of genre-specific corpora</li> <li>• makes it possible to identify lexical bundles that can be used to distinguish registers (spoken vs. written, subject matter)</li> <li>• many lexical bundles in academic lectures signal discourse relations</li> </ul>

Table 14. ESP corpus research into the analysis of specific discourses

Area within ESP	Subfield	Research	Application within ESP
discourse competence	professional discourse	<p><b>Collins &amp; Scott</b> (1997): business meetings can be characterised by the organisation of lexis</p> <p><b>Tribble</b> (2000): analysis of EU Phare project proposals</p> <p><b>Henry &amp; Roseberry</b> (2001): described linguistic characteristics of moves within two genres</p> <p><b>Upton &amp; Connor</b> (2001): analysis of moves and politeness strategies in business letters</p> <p><b>Forchini &amp; Murphy</b> (2008): 4-grams in the Financial Times Corpus</p> <p><b>Trebits</b> (2008; 2009a): lexis and phrasal verbs in EU texts</p> <p><b>Trebits</b> (2009b): conjunctions in EU texts</p> <p><b>Jablonkai</b> (2009a): lexical bundles in news texts and EU texts are significantly different</p>	<ul style="list-style-type: none"> <li>• multidimensional analysis provides the fullest linguistic analysis of a collection of a genre, but key words offer powerful means of defining relevant lexical items for teachers and students</li> <li>• corpus analysis can go beyond simple count of linguistic features – analysis of pragmatic strategies</li> <li>• corpora allow for a more thorough understanding of how language is used in particular contexts or in particular genres</li> </ul>

Table 14. cont. ESP corpus research into the analysis of specific discourses

#### 4.3.3.2. Functions and structures of lexical bundles

One of the early findings concerning lexical bundles was that they are present in written and spoken registers alike and they were considered “basic building blocks for constructing spoken and written discourse” (Biber & Conrad, 1999, p. 188). Moreover, further research found that in certain written registers like written course management – comprising syllabi and description of course assignments – lexical bundles are surprisingly common (Biber & Barbieri, 2007). These results were in contrast to previous analyses which regarded the use of MWIs a characteristic of spoken registers (Pawley & Syder, 1983).

As regards the structure of lexical bundles, previous studies found that lexical bundles are structurally complex, usually incomplete and not fixed (Biber & Conrad, 1999).



Comparing bundles across registers showed that the grammatical structure of lexical bundles is a distinct characteristic of registers. Biber and Conrad (1999) reported that the most frequently occurring lexical bundles in conversation have the pattern personal pronoun + verb phrase (clause-fragment), for example, *I don't know how, you might as well*, whereas in academic prose the two most important patterns are noun phrase and prepositional phrase fragments, for example, *one of the most, an increase in the*. Structural types of lexical bundles were further analysed and classified by Biber et al. (2004). Their framework, which was applied in the present study as well, is described in detail in Section 6.4.3.

Previous research also looked into the discourse functions of lexical bundles. By developing a detailed taxonomy, Biber et al. (2004) found that the three main functions lexical bundles serve in discourse include expressing stance, organising discourse and referring to, for example, specific attributes, time and place. The full taxonomy of functions of lexical bundles in text is outlined in Section 6.4.3.

Research on MWIs emphasised the important role of longer word combinations in language teaching (Cowie, 1992). Still, there are only a few studies that focus on the role of lexical bundles in language teaching (Cortes, 2004; Cortes, 2006; Scott & Tribble, 2006). Most of these investigations compare student writing with expert writing in academic disciplines. Scott and Tribble (2006), by looking at student and expert literary papers, concluded that such a comparison can be helpful for students to find what they lack in becoming proficient writers. Cortes (2004), by looking at texts in two disciplines, namely, History and Biology, also compared the use of lexical bundles in student and expert writing and found that students did not often use lexical bundles, or used them in a different way. In her later study, Cortes (2006) described a few tasks for the explicit teaching of lexical bundles, and analysed the effectiveness of the tasks. She concluded that having a few lessons that demonstrate the use of some examples of lexical bundles in expert writing will not result

in students using lexical bundles in a more appropriate way, but increases their interest in, and awareness of, these expressions, which is a useful step towards producing texts that are accepted by the respective discourse community.

#### **4.3.3.3. Professional discourse**

There are fewer studies focusing on professional genres, than on academic genres. The studies reviewed here cover two subject fields: business in general, and genres in the European Union context. The business genres analysed by corpus linguistic methods include several spoken genres like business meetings (Collins & Scott, 1997), and introduction to guest speakers (Henry & Roseberry, 2001), and written genres, like letters of application (Henry & Roseberry, 2001; Upton & Connor, 2001). The genres representing the EU context include project proposals (Tribble, 2000), and several other EU genres like press releases, legal text and reports (Jablonkai, 2009a, 2009b; Trebits, 2008, 2009a, 2009b). On the basis of these studies it has been demonstrated that corpus analysis can go beyond simply counting linguistic features and it can be used to analyse pragmatic strategies, like politeness strategies (Upton & Connor, 2001), and discourse structure of important genres (Henry & Roseberry, 2001; Upton & Connor, 2001). All the studies in the previous section applied corpora which comprised, exclusively or partly, textbooks or course books in the field under scrutiny. The relevance of the studies investigating professional genres for ESP teaching is that they inform about the target situation for which ESP learners have to be prepared for by looking into the discourse of genres they will read and write in their specific professions.

#### **4.3.4. Analysis for course and materials design**

As summarised in Table 15, corpus research has yielded relevant findings for course and materials design in ESP in three aspects. Firstly, several studies have demonstrated that investigating corpus data representing the language of the specific subject field in use

provides a sound basis for course and materials design for the ESP classroom (e.g., J. Flowerdew, 1994; L. Flowerdew, 2001; Fuentes, 2002; Jabbour, 1998). Furthermore, some of these studies have also shown how findings of corpus research can be integrated into the ESP teaching practice (e.g., Cortes, 2004, 2006; Fuentes, 2002; Mudraya, 2006; Trebits, 2009b) in the form of new approaches and tasks. Mudraya (2006), for example, based on the lexical analysis of the Student Engineering English Corpus, argued for integrating the lexical approach with DDL, as she considered corpora an invaluable tool for giving students insights into the unique collocational and usage patterns of lexical items in specialised texts. Moreover, Cortes (2004, 2006) did not only identify lexical bundles in two disciplines for the purposes of writing instruction, but she also looked into the effectiveness of her corpus-based materials. Her results suggested that exposure to lexical bundles alone does raise students' awareness of such constructs, but does not necessarily result in the appropriate use of lexical bundles. Secondly, some of these studies not only present DDL tasks that were found useful in the ESP classroom, but propose ways in which students themselves can compile corpora, for example, for academic writing instruction. Lee and Swales (2006) reported on an academic writing course at a doctoral level where students created their own corpora and compared their own writing to that of established writers in their specific disciplines. Participants in the course considered the use of corpora "confidence-building and empowering" (p. 71), and they found that using corpora had advantages over using grammar books and reference books. Finally, corpus research in ESP has also been used to evaluate existing teaching materials. Nelson (2000, 2006) compared a corpus of published course books of BE to a corpus of English in "real-life" business in order to test his hypothesis that "the lexis found in Business English published materials is significantly different from that found in real-life business" (Nelson, 2000, p. 1). He compiled his BEC from spoken and written texts like faxes, emails, reports, radio and TV interviews and newspaper articles which

are produced when “doing business” and “talking about business”. He also created a Published Materials Corpus (PMC), consisting of the most widely used BE course books. In this analysis, Nelson (2000, 2006) investigated the lexis of BE by focusing not only on single-word lexical items as in earlier word lists, but also examining word clusters, collocation and colligation of certain lexical items. He also analysed the semantic prosody selected words showed in the two corpora. Based on these analyses he found that the lexis in his PMC was “simpler, more concrete, less varied, more polite and much more focused on human interaction” (Nelson, 2000, p. 544), than the lexis found in “real-life” business.

Area within ESP course and materials design	Subfield	Research	Application within ESP
	-	<p><b>J. Flowerdew</b> (1994): specific purpose course design needs specific language in terms of language skills and subject matter</p> <p><b>Jabbour</b> (1998): ESP syllabus design will meet learners’ needs if language in use is taken as the basic element in the design</p> <p><b>L. Flowerdew</b> (2001): compared learner and expert corpora</p> <p><b>Fuentes</b> (2002): corpus-based language learning tasks in BE</p> <p><b>Cortes</b> (2004): exposure to frequent linguistic elements like lexical bundles does not result in appropriate use</p> <p><b>Cortes</b> (2006): teaching lexical bundles for writing instruction</p> <p><b>Mudraya</b> (2006): integration of lexical approach and corpus-based methodology</p> <p><b>Chujo &amp; Utiyama</b> (2006): nine statistical measures used to identify level-specific special lexis</p> <p><b>Trebits</b> (2009a): sample tasks for the use of phrasal verbs in EU texts</p>	<ul style="list-style-type: none"> <li>• specialised corpora can provide specific information for course design</li> <li>• corpus can be used for generating text-based language practice activities</li> <li>• helps identify lexical bundles for subject-specific writing courses</li> <li>• informs the design of appropriate special purpose teaching materials and courses</li> <li>• allows to create corpus-based teaching activities</li> <li>• allows the application of DDL</li> <li>• insights into which collocational, pragmatic or discourse features should be addressed in EAP materials</li> <li>• makes it possible to automatise the identification of technical lexis for different proficiency levels</li> </ul>

Table 15. Corpus research in ESP for course and materials design

<b>Area within ESP course and materials design</b>	<b>Subfield</b>	<b>Research</b>	<b>Application within ESP</b>
	data-driven learning	<b>Lee &amp; Swales</b> (2006): a corpus-informed EAP course for doctoral students <b>Mudraya</b> (2006): integration of lexical approach and corpus-based methodology	<ul style="list-style-type: none"> <li>• enhances learner autonomy, not only can students use available specialised corpora, but they can also compile their own corpora</li> </ul>
	evaluation of existing teaching materials	<b>Nelson</b> (2000): provided detailed lexical profile in BE, compares existing BE materials to real business language use	<ul style="list-style-type: none"> <li>• makes it possible to correct mistaken intuitions about what ESP materials should focus on</li> </ul>

Table 15. cont. Corpus research in ESP for course and materials design

#### 4.3.5. Analysis across languages

As shown in Table 16, only a few studies in ESP used corpora for comparing the discourse of certain genres across languages. Two types of corpora have been used for cross-linguistic analyses. Thompson (2001) used ‘comparable corpora’ for the comparative analysis of Chinese and English tourist brochures, and Swiss and English job advertisements. He argued that cross-linguistic comparisons can be used as awareness-raising resources in the ESP language classroom, and students’ attention should not only be drawn to specific language structures of interest, but also to contrasts at the discourse level. The other type of corpus designed to conduct cross-linguistic analyses is ‘parallel corpora’. An example of a parallel corpus in ESP is the Hungarian-English professional corpus containing texts of several subject fields like Agriculture, Environmental protection, Biology, etc., in English, with their Hungarian translations (Heltai, 2007). As Heltai claimed, the novelty of this parallel corpus is that, besides expert translation texts, it contains texts translated by students, and also working translations prepared under time constraints. Thus, the analysis of the corpus can shed light at the differences of these types of translations.

Area within ESP	Research	Application within ESP
cross-linguistic analysis	<p><b>Thompson</b> (2001): comparable corpora of job advertisements and tourist brochures (Chinese, Swiss, English)</p> <p><b>Heltai</b> (2007): compilation of a parallel, Hungarian-English, professional corpus</p> <p><b>Forchini &amp; Murphy</b> (2008): 4-grams in Italian and English comparable corpora</p>	<ul style="list-style-type: none"> <li>• allows for corpus-based cross-linguistic genre comparison</li> <li>• makes it possible to analyse translations of professional texts</li> <li>• comparable corpora can be applied to compare the use of MWI across languages</li> </ul>

Table 16. Cross-linguistic corpus research in ESP

#### 4.3.6. Summary

To conclude, the benefits of corpus linguistics for ESP, and also as seen from the perspective of the current investigation, can be summarised as follows:

- the use of specialised corpora provides a means to identify specific linguistic and discourse characteristics of relevant disciplines, registers and genres;
- corpus linguistics provides new methods for the analysis of lexis in ESP:
  - analysis of frequency and collocation of certain lexical items for subject-specific dictionaries, glossaries and word lists for pedagogical purposes,
  - more detailed analysis of the patterns and behaviour of lexical items in specialised texts;
- courses and materials based on findings of corpus research can more precisely cater for specific needs of ESP learners as regards target situation language use;
- corpus research has brought new methods to describe discourse patterns of academic and professional discourse (e.g., annotating politeness strategies, analysing lexical bundles);
- corpora of specialised texts can provide context knowledge, that is, information on the target professional culture and norms of the discourse community;
- specialised corpora and corpus analysis tools provide methods for the comparison of student and expert language use;
- specialised corpora and corpus analysis tools provide methods for comparing general purpose and specific purpose language use (e.g., key word analysis, collocational analysis, semantic preference, semantic prosody);
- corpus data provide the basis for new approaches and methods to teaching ESP (lexical approach, DDL).

The review of corpus linguistics and specialised corpora for ESP would not be complete without touching on the limitations of this kind of approach. Possible pitfalls of using specialised corpora have been widely debated in the literature. Representativeness, size and generalisability have been the most relevant issues of concern (Flowerdew, 2004). Nevertheless, the advantages corpus studies bring to ESP teaching and learning outweigh these limitations.

After reviewing the methods, findings, and benefits of corpus research to ESP, the next sections will discuss the first steps in the practice of corpus research, that is, corpus building. The sections will review theoretical considerations and practical steps in corpus design and corpus creation for ESP purposes.

#### **4.4. Issues in corpus design for ESP**

One of the key issues in corpus studies is the creation of the corpus itself. This is a matter of crucial importance, as all the conclusions drawn from an analysis of the corpus can only be interpreted in light of the collection of texts examined. How big should the corpus be? How many texts should be included? What genres and text types should be represented? Should the corpus be made up of whole texts or excerpts of texts of a predetermined size? If we are to include excerpts, how long should they be? These are some of the important questions researchers have to answer when they begin designing and creating their corpora for research and teaching purposes. Ever since corpora have been used for linguistic research, the 'how' and 'what' of corpus building has always been an issue. In the literature, theoretical considerations have been suggested (Biber et al., 1998; Clear, 1992; Leech, 1991; McEnery & Wilson, 1996a; Szirmai, 2005) and practical problems raised, whilst solutions have also been proposed (Leitner, 1992; Nelson, 1996; Sinclair, 2005) for systemising the method for compiling a corpus. This section will review these most important theoretical and practical considerations of corpus building in order to propose a *Model for Corpus Creation for ESP*.

The model aims to integrate earlier guiding principles and practices and include all the elements of the process that are essential in sound corpus research.

#### **4.4.1. Guiding principles for corpus design**

One of the most influential set of principles for corpus building was proposed by Clear (1992). In order to avoid the undisciplined collection of texts for linguistic analysis, and to allow comparability of corpora and results of corpus analyses, he suggested the following guiding principles for building a corpus of general English:

“P1: The notion of a ‘core’ of language is useful.” (p. 27)

Extending the idea of a ‘core vocabulary’ in applied linguistics to all levels of language use, he suggested that corpus building deal with the central and typical.

“P2: The corpus may be a sample corpus or a monitor corpus.” (p. 28)

The distinction between sample and monitor corpus is based on practice rather than theory. As outlined in Section 4.1.2.2, the term ‘monitor corpus’ refers to a corpus with a large (or in some cases unlimited) size, such as the Bank of English. A sample corpus, however, is of finite size, and the collection of texts is strictly controlled. Clear urged for a compromise between the two approaches, that is, one that creates an open-ended corpus on the one hand, but which complies with rigorous sampling principles, on the other.

“P3: The definition of a “text type” should be fairly clear and objective.” (p. 28)

Although there exist some intuitive text categories which are used to classify written and spoken language, but these are not supported by theoretical criteria. Taxonomies of text types, however, are not feasible, according to Clear, as a text is a “very complex socio-linguistic artefact” (p. 29).

“P4: The definition of “text types” should distinguish internal criteria from external.”  
(p. 29)



Clear introduced the internal and external criteria of text types that should also be a factor when building a corpus for linguistic analysis. Internal criteria are essentially linguistic criteria (e.g. the categories formal/informal are based on linguistic characteristics). External criteria, however, are non-linguistic criteria (e.g. the classification of texts according to the gender of the authors or the time of publication). In order to create valuable corpora we should apply both types of criteria.

“P5: The corpus will help us to discover new aspects of language use and will provide evidence to confirm (or refute) provisional hypotheses.” (p. 29)

In close connection with the previous principle, Clear drew attention to another important aspect of corpus building. There is a danger of formal bias in that certain texts may be selected because they have particular linguistic features: although the application of exclusively internal criteria may appear to support certain hypotheses of language use, the lack of external criteria in text selection means that findings cannot be considered as proof for the assumptions.

“P6: Decisions concerning corpus quality should be based whenever possible on assessment of existing corpus resources.” (p. 30)

A cycle of corpus creation is introduced under this principle. The value of a corpus is not only measured in how data based on them are processed and presented, but also in how methods of text collection are reviewed and refined. Clear suggested that corpus building should be based on experience gained from earlier corpora not only in respect of linguistic description, but also concerning methodology.

These principles refer to general English corpora. However, throughout the rest of this section special considerations that are necessary for specialised corpora will be the focus of discussion. In order to distinguish general corpora from specialised corpora, definitions of both categories will be briefly returned to. According to Hunston (2002), a general corpus

includes a wide variety of texts and text types. Specialised corpora, however, focus on one particular type of text or variety of language. Although general corpora are usually much larger than specialised corpora, they are not likely to be representative of any particular “whole”. On the other hand, specialised corpora aim to be representative of given types of text or particular kinds of language that researchers intend to investigate. Further details on the distinction between these two types of corpora will be given in Section 4.4.6.

This section will start with an overview of the main steps of corpus design and corpus creation. Then, using these steps as a starting point, a *Model for Corpus Creation for ESP* will be proposed. The model provides for theoretical issues such as questions of size, representativeness, and also covers sampling and practical issues concerning data collection, data entry and legal issues. In later chapters on research design and procedures of analysis it will also be outlined how the model was applied for the design and creation of the corpus used in the present study.

#### **4.4.2. An overview of the steps of corpus design and creation**

According to Sinclair (2005), corpus creation has two almost inseparable stages: design and implementation. However, in order to create a meaningful model for corpus studies, individual steps need to be analysed more closely. Therefore, for the purposes of the present study the whole process of corpus design and implementation is divided into seven main steps. A summary of all steps is given in Figure 11. These steps cover theoretical and practical considerations as well. Moreover, one single step also includes several procedures to consider and probably many decisions to make. The most important decision about the corpus is its ultimate purpose, that is, it should serve the aims of the research project it is designed for. The purpose of the corpus will influence all subsequent decisions, and it will be a decisive factor in all steps and aspects of corpus design, for instance, type of corpus, text categories, etc. Therefore the first step is to identify the purpose of the corpus.

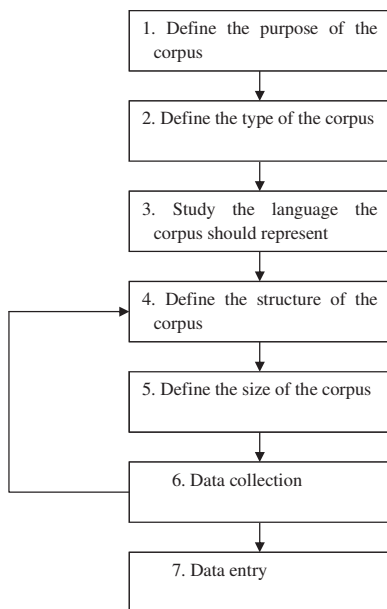


Figure 11. Main steps of corpus creation

The next two steps give the theoretical background for the corpus design. These decisions give the criteria according to which texts or excerpts of texts should be selected for the corpus.

Step 4, the decision on the structure and contents of the corpus, and Step 5, the decision on the size of the corpus, will be based on the above-mentioned decisions, but will also be influenced by practical constraints. This is what is indicated by the arrow coming from Step 6, data collection. Despite the careful preparation of theoretically sound selection criteria and contents by designers, all accounts of corpus creation report on the practical difficulties of data collection which can cause slight distortion in the proportion of text types and genres in a corpus, or in some cases result in the exclusion or inclusion of certain text categories and

other changes to the originally planned contents (Kennedy, 1998; Leitner, 1992; G. Nelson, 1996; M. Nelson, 2000).

The final step is data entry. This step does not only involve decisions on the methods of entering data into the database, but also on creating a storage and retrieval system.

These basic steps of corpus design and creation are developed into a model of corpus design and creation in the course of the present chapter. In the following sections this comprehensive model is outlined with detailed descriptions of the individual phases.

#### **4.4.3. A Model for Corpus Creation for ESP**

In the early days of corpus research, teams of linguists created the first corpora for linguistic studies. Since then technology has developed, and individual researchers (Cheng-yu, 1993; Ghadessy, Henry, & Roseberry, 2001), language teachers (Conrad, 1999; Hunston, 2002; Nelson, 2000) and translators (Bowker, 2000; Károly, 2003) can now create their own corpora for their own purposes, whether it be linguistic research in general, language teaching, or translation. In order to make these corpora comparable and the findings of these research projects really meaningful, the methods by which their corpora are created should be theoretically well-founded. Therefore, a disciplined and systematic method for collecting texts for corpus studies in the form of a *Model for Corpus Creation for ESP* is proposed.

In the literature of corpus research we find descriptions of earlier studies (Kennedy, 1998; Leitner, 1992; Szirmai, 2005), guiding principles, and practical advice (Clear, 1992; Sinclair, 2005) for corpus development. The proposed model aims to combine the theoretical and practical considerations that have to be part of systematic corpus building, and develop the whole process into a coherent sequence of phases. The model can be used as a guiding tool for corpus creation. It includes seven phases from designing a corpus for one's own research purposes to data entry that is the actual creation of the corpus.

Table 17 shows the model with all its elements. The seven phases correspond to the seven main steps of corpus design and creation. In the second column under the heading Considerations, issues that need to be considered or decided during the given phase are listed. In the third column, sources of information are given where data necessary for the decision making process can be found. In the last column, the most commonly occurring cases for each relevant issue are enumerated. A visual representation of the model can be seen in Figure 12.

In what follows, each phase is discussed in detail, giving its relevance in corpus design and creation, and outlining theoretical underpinning for the decisions that need to be made in the particular phase. The way the model was applied to the corpus design of the current study will be outlined in Section 6.3.

Phases of corpus design	Considerations	Sources of information	Examples
1. Define the aim and purpose of the corpus	<ul style="list-style-type: none"> <li>the aim of the research project</li> </ul>	<ul style="list-style-type: none"> <li>earlier corpus-based studies</li> <li>manuals of corpus-analysis software</li> </ul>	
2. Define the type of the corpus	<ul style="list-style-type: none"> <li>the planned type of analysis</li> </ul>	<ul style="list-style-type: none"> <li>earlier corpus-based studies</li> </ul>	<ul style="list-style-type: none"> <li>specialised corpus</li> <li>general corpus</li> <li>learner corpus</li> <li>monitor corpus</li> <li>pedagogic corpus</li> <li>parallel corpus</li> <li>comparable corpus</li> <li>historical corpus</li> </ul>
3. Study the language or language variety the corpus represents	<ul style="list-style-type: none"> <li>representativeness</li> <li>define a sampling frame</li> <li>set external criteria for text selection</li> <li>sampling methods: <ul style="list-style-type: none"> <li>random</li> <li>stratificational</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>findings of social sciences and other relevant fields of research</li> <li>findings of earlier linguistic research</li> <li>needs analyses</li> <li>statistical information</li> </ul>	<ul style="list-style-type: none"> <li>time period texts were produced</li> <li>mode of text: written, spoken, electronic</li> <li>the location texts were produced at</li> <li>types of texts, genres: letter, book, journal, etc.</li> </ul>
4. Define the structure of the corpus	<ul style="list-style-type: none"> <li>relevant text categories</li> <li>proportion of text categories</li> <li>balance of text categories</li> <li>diversity of topics</li> </ul>	<ul style="list-style-type: none"> <li>findings of Phase 3</li> </ul>	
5. Define the size of the corpus	<ul style="list-style-type: none"> <li>total number of running words in the corpus</li> <li>number of samples</li> <li>number of running words of samples</li> <li>whole texts</li> <li>feasibility of analysis</li> <li>comparability with earlier corpus-based studies</li> </ul>	<ul style="list-style-type: none"> <li>similar earlier corpus-based studies</li> </ul>	
6. Data collection	<ul style="list-style-type: none"> <li>adequate sources of texts</li> <li>feasibility, availability</li> <li>systematic collection and selection procedure</li> <li>legal and ethical issues</li> <li>confidentiality</li> </ul>	<ul style="list-style-type: none"> <li>language users</li> <li>publications</li> <li>World Wide Web</li> </ul>	
7. Data entry	<ul style="list-style-type: none"> <li>electronic versions</li> <li>mark up system</li> </ul>	<ul style="list-style-type: none"> <li>earlier corpus-based studies</li> <li>standards</li> </ul>	<ul style="list-style-type: none"> <li>methods: keyboarding, scanning</li> <li>transcription</li> <li>copy original text files</li> <li>convert html, pdf, doc into text file</li> </ul>

Table 17. *Model for Corpus Creation for ESP*

#### 4.4.4. Purpose, the guiding thread

The first and most important decision about a corpus is what it will be used for. Not all investigations into language and language varieties can be answered by using a corpus. Although an increasing number of different analyses can be carried out with the help of corpora as computer technology advances, there are certain research queries for which a corpus is the best research tool. Three main forms of corpus study can be distinguished: **lexical**, **syntactic** and **discoursal**. These forms of study can be used for describing a language, register, or genre, for example. The vast majority of corpus research is focused on **lexical issues** (e.g., Moon, 1998; Nelson, 2000; Sinclair, 1987; Stubbs, 2002). Learner dictionaries such as the *Collins Cobuild English Dictionary for Advanced Learners* (Sinclair, 2003), *Longman Dictionary of Contemporary English* (Summers, 2003) and *Oxford Collocations Dictionary* (Crowther, Dignen, & Lea, 2002) were published based on lexical analyses of monitor corpora. In addition to linguistic description of languages, registers or genres, many corpus-based studies have pedagogical purposes as well. Word frequency, collocational patterns (Renouf & Sinclair, 1992) in specialised corpora is often a research aim in ESP for pedagogic purposes (e.g., Coxhead, 2000; Mudraya, 2006; Nelson, 2000). Furthermore, corpus research can inform language teachers and researchers about the semantic associations of certain words (Nelson, 2000, 2006; Partington, 2004; Stubbs, 1995, 2001b). Secondly, both descriptive and pedagogic corpus research can focus on **syntactic** features such as verb forms and the use of prepositions or conjunctions. Reference grammar books, for example the *Collins COBUILD English Grammar* (Sinclair, 1990), are also the results of such syntactic studies. It should be noted, however, that many corpus studies focusing on the lexical and syntactic level of language also inform about lexicogrammatical features of a particular language or language variety. *The Longman grammar of spoken and written English* (Biber et al., 1999) is a very good example of such studies.

Finally, corpus-linguistic **discourse** studies focus on rhetorical patterns (Henry & Roseberry, 2001) or moves (Upton, & Connor, 2001) of specific genres; they may also analyse how texts are structured, describe the use of discourse markers, or suggest ways in which corpus work and discourse analysis can contribute to language teaching (Guillot, 2002). Earlier corpus studies, manuals of corpus analysis software (Scott, 2004), and comparisons of corpus analysis tools (Ari, 2006; Hockey, 2001; Reppen, 2001; Wiechmann & Fuhs, 2006) can be used as sources of information for deciding whether a corpus-linguistic methodology can best serve a particular research purpose.

The purpose for which the corpus is compiled will be an underlying decisive factor throughout the whole corpus design and implementation process. Closely related to it is the decision on the type of corpus. This is visualised by the two big boxes with the headings 'Purpose of the corpus study' and 'Type of corpus' in Figure 12, underlying and guiding all other phases of corpus creation.

#### **4.4.5. Representativeness**

The next underlying factor is representativeness. The issue of **representativeness** can also only be discussed in the light of the purpose of a particular corpus. A corpus is more than a collection of texts in that it is created to represent a language or a specified part of a language. As Leech (1991) put it: "In practical terms a corpus is 'representative' to the extent that findings based on its contents can be generalized to a larger hypothetical corpus" (p. 27).

What does this mean for corpus design? The appropriate way to create a corpus is based on what the corpus needs to represent. The representativeness of the corpus determines what type of analyses can be carried out with its help, and also the extent to which findings can be generalised. For example, results of an analysis on a corpus of exclusively written texts would not allow generalisation on language use as a whole. Furthermore, if we are investigating the



use of slang words in the conversations of teenagers, our findings will not be characteristic of conversations in general (Biber et al., 1998).

Although issues of representativeness are crucial in corpus linguistics, it is a truism that it is impossible to create a perfectly representative corpus (Hunston, 2002; Kennedy, 1998; Nelson, 2000). Leech (1991) was hoping for statistical and other models to measure the representativeness of a corpus, however, there has been no major development in this respect.

Other important issues in corpus creation are balance (Hunston, 2002; Kennedy, 1998; Nelson, 2000) and diversity (Biber et al., 1998). **Balance** refers to the “weighting between the different sections” (Kennedy, 1998, p. 62) in a corpus, whereas **diversity** is necessary in order to make the corpus represent the language or a language variety in its entirety with all its different registers, dialects, subject matters, and so on. Balance and diversity are important in the case of specialised corpora investigating the language of one specific discipline or professional field, genre or topic. In addition to representativeness, researchers in ESP (Fuentes, 2002; Hänchen, 2002; Krause, 2005) who apply smaller specialised corpora emphasise other criteria such as diversity of addressees, text types, topic relevance and up-to-datedness as relevant in corpus design.

#### **4.4.5.1. Sampling methods**

Increasingly sophisticated methods have been devised by corpus compilers to achieve representativeness, balance and diversity in corpus design. The most widely used method is sampling. Random sampling is a standard way of selecting subjects for analysis in many areas of science and the social sciences. Stratificational sampling as suggested by Biber (1993), however, often proved more representative for corpus building (McEnery & Wilson, 1996a). Clear (1992) gave the following summary of the issues and difficulties regarding sampling:

first – phenomenon to be sampled poorly defined ...

... second – there is no obvious unit of language which is to be sampled ...

... third – the sheer size of the population ensures that any attempt to account for the difficulty of setting up a sampling frame by gathering ever larger samples will not of itself advance our state of knowledge significantly – given current and foreseeable resources, it will always be possible to demonstrate that some features of the population is not adequately represented in the sample. (p. 21)

Therefore, it is the corpus linguist's duty to make every effort to define the population as precisely as possible, and when discussing the results of the analysis, researchers must constantly bear in mind the limitations of the particular study. This is the main reason why the starting point for stratificational sampling is the definition of the population to be investigated. This way a sampling frame, that is, the entire population of texts, has to be defined as closely as possible. The samples for the corpus will be taken from this population. There are several ways to define the sampling frame. In the case of general corpora, a comprehensive bibliographical index can be used. For example, for the creation of the LOB Corpus, the British National Bibliography and Willing's Press Guide were used (Biber, 1993; McEnery & Wilson, 1996a). Other approaches towards defining the population of texts include taking all the books and periodicals in a particular library or time period, and looking for statistical information on publications, such as how many books, journal articles, or pieces of legislation were published in a given time period. The sources commonly used in corpus design are represented by boxes in the second column on the left of the visual representation of the model in Figure 12.

In the case of corpora specialising in a certain language variety such as the technical language of a discipline, findings of that particular discipline or results of linguistic analyses of the language variety help define the sampling frame. Useful sources of information, especially for corpus studies in teaching language for specific purposes, are needs analyses that give information on language usage in particular target situations. As the application of needs

analyses is special for ESP, it is highlighted in the visual representation by broken lines. Needs analysis was also applied in the phase of corpus design in the present study. There are two reasons for involving a needs analysis in ESP corpus design. Firstly, as it is also discussed in Section 2.2.1, learner needs are a fundamental element in ESP course and materials design, and therefore they should be the starting point for designing corpora for ESP purposes as well. Secondly, Sinclair (1991) also emphasised that subject-specialist informants should be involved in corpus design, and one of the ways to draw information from experts in a field is to conduct a needs analysis survey. There are already corpus studies in ESP which in several ways, (interviewing subject specialists and/or using questionnaires), based their corpus creation on the analyses of the target situation needs.

Clear (1992) drew attention to the significance of extra-linguistic or external factors for the definition of text types. Therefore, social variables are also applied in the text selection procedure. For example, variables such as age, gender, level of education, and 'nativeness' of the language were defined and recorded in the text collection procedure for the International Corpus of English (ICE) (Nelson, 1996). At the end of Phase 3 of the proposed model, a set of external criteria have to be established in order to help develop the structure and contents of the corpus (see Table 17 and Figure 11).

In general, in large reference corpora, representativeness is sought through a rigorous selection procedure and by large amounts of data. In the case of specialised corpora, external selection criteria are used to measure the representativeness of the corpus. The number of these criteria should be limited, and the criteria themselves should be easily establishable in order to avoid complications at the text selection stage (Sinclair, 2005). In a recent study, Williams (2002) argues the case for including internal criteria for text selection for specialised corpora to avoid the subjectivity of the application of external criteria. He proposed lexical criteria, more

precisely, the corpus-directed study of collocational networks, to create subject specific groupings of texts within the corpus.

Another study where a combination of internal and external criteria was applied in the sampling procedure is the creation of the Tobacco Industry Documents Corpus (Kretzschmar, Darwin, Brown, Rubin, & Biber, 2004). The aim of the study was to investigate rhetorical manipulation ('deception') in documents produced by tobacco companies. During the three-step corpus creation the researchers employed a rigorous sampling strategy. In order to determine the representative text categories and their proportion, in the overall collection of texts they first drew a limited sample, an 'exploratory core sample'. The second step was to create a reference corpus of 500,000+ words which was a stratified random sample of all documents. Stratification took place by external criteria such as decade, source of text (only texts written by authors within the tobacco industry were included), and target audience, whilst internal criteria (such as named or unnamed audience) and certain surface features of the documents (interlinear editing and handwritten comments) were considered as potential indicators of manipulative intent. Finally, based on these criteria, a corpus was created of texts that were assumed to exhibit rhetorical manipulation.

#### **4.4.5.2. Structure of the corpus**

The next phase in the design process is to determine the structure of the corpus. In some projects, text categories were defined partly by intuition, for instance, ICE project, in others, especially in ESP, structure was determined based on surveys among subject specialists, for example, Krausse's environmental engineering corpus (Krausse, 2005). However, it is generally accepted in the literature that initial classifications and categorisations of texts might turn out to be too rigid for the purposes of a study, and thus corpus building procedures should allow

certain flexibility, for example, by including sub-categories or merging some categories (Kennedy, 1998; Leitner, 1992; Nelson, 1996).

Another example of applying stratificational sampling for a specialised corpus is Nelson's (2000) BEC. Nelson started out by defining the population of language users, that is, the discourse community of business people, for his research. The members of the particular discourse community for his research were native speakers of English "who use English in the pursuit, transaction and discussion of business, trade and commerce" (Nelson, 2000, p. 240). The extra-linguistic factors he used to closely define the population included gender, regional varieties, levels of respondent in business, type of business. The decisions regarding which text categories and genres to include in the corpus were based on the literature and earlier findings about the language of BE. Based on these considerations Nelson created the content specification of his ideal BEC.

Biber and his colleagues (1998) also drew attention to the importance of careful corpus design and flexibility:

It is important to be realistic. Given constraints on time, finances, and availability of texts, compromises often have to be made. Every corpus will have limitations, but a well-designed corpus will still be useful for investigating a variety of linguistic issues. (Biber et al., 1998, p. 250)

#### **4.4.6. The bigger the better?**

After designing the structure of the corpus, decisions on its overall size have to be made. The issue of size has been a heavily debated question among scholars in corpus linguistics. Recently, the main guiding principle of "bigger means better" (Leech, 1991, p. 9) seems to be being abandoned, and smaller corpora are also being compiled, especially, for specific purposes. Leech (1991) suggested that one of the factors that show development in corpus

linguistics is the increasing size of corpora besides the increasing power of computer technology which makes the analysis of corpora of huge sizes possible.

On the basis of their size, Leech (1991) distinguished three generations of corpora. Examples of the 'first generation' are the Brown Corpus and its British counterpart the LOB Corpus with one million running words, which seemed almost unsurpassable at that time. In the 1980s, the 'second generation' appeared with ten to thirty million running words. This generation can be represented by the Cobuild project (Sinclair, 1987), by the work of John Sinclair and his team at the University of Birmingham, and by the Longman/Lancaster English Language Corpus. The 'third generation' of corpora already consists of many hundreds of millions of words and they also apply advanced computer technologies. Sinclair (1991) proposed the creation of large corpora as he noted: "The only guidance I would give is that a corpus should be as large as possible, and should keep on growing" (p. 18). The main reason for arguing for big corpora is the uneven pattern of occurrence of words in texts. Therefore, because of the need of statistically significant number of tokens "in order to study the behaviour of words in texts, we need to have available quite a large number of occurrences" (Sinclair, 1991, p. 18).

Although the prevailing view has been that large corpora are best, this view was already challenged in the early 1990s by Leech and later by other researchers as well (Hunston, 2002; Leech, 1991; Kennedy, 1998). Leech (1991) enumerated four reasons why to focus on the size of the corpus only is "naive":

Firstly, a large collection of texts does not necessarily make a corpus. We should distinguish large archives of machine-readable texts from corpora which are carefully designed and systematically collected samples of texts. Corpora must fulfil a representative function as well.

Secondly, the vast increase in the size of corpora has taken place almost exclusively in the collection of written language.

Thirdly, copying texts into a corpus present legal problems like copyright and confidentiality.

Fourthly, the lack of available computer technology to analyse large corpora. (p. 10)

A fifth reason that can be added to Leech's concerns is the overwhelming amount of data that large corpora generate (Hunston, 2002). Kennedy (1998) pointed out that "although it is the case that for descriptive adequacy of low frequency phenomena such as collocations very large corpora are necessary, there is no point in having bigger and bigger corpora if you cannot work with the output" (p. 68).

Hunston (2002) suggested that in order to gain a manageable amount of information from corpora we have two options: (a) to use software to select data randomly or, on the basis of certain important characteristics, for example, frequency, or (b) to use a smaller corpus.

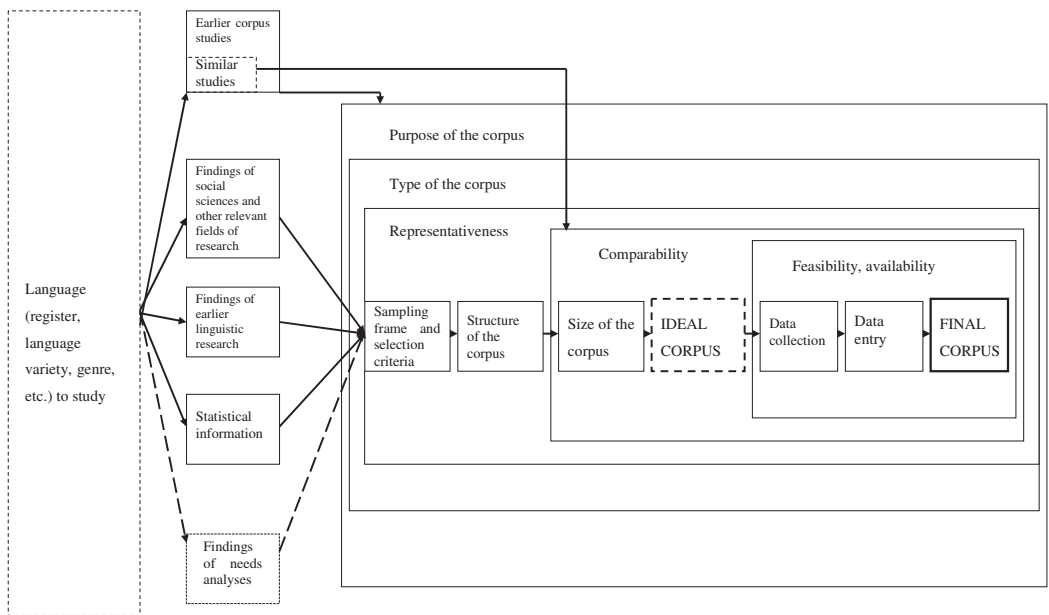


Figure 12. Visual representation of the *Model for Corpus Creation for ESP*



#### 4.4.6.1. Small can be beautiful too

In recent years a movement can be witnessed among scholars in corpus linguistics which emphasises the importance of small corpora (Ghadessy et al., 2001; Nelson, 2000). A distinction should, however, be made between 'general-purpose' (Leech, 1991) or general (Hunston, 2002; Sinclair, 1991) corpora and 'domain-specific' or specialised corpora. The aim of a **general corpus** is to describe the language in general. Therefore, it includes many text types, genres, written and spoken texts and texts of several subject matters. Although it cannot be totally representative of a particular language as a whole, they are often used to produce reference materials for language teaching and translation. These corpora are mainly used for lexicographic purposes. The British National Corpus (BNC) with 100 million words and the Bank of English with many hundred million words are examples of general corpora.

A **specialised corpus**, on the other hand, seeks to represent a special language variety. This can be a linguistically or socially determined variety of a given language. As outlined in Section 4.3.1, specialised corpora are usually smaller than general corpora. The aim of a specialised corpus is to investigate a particular linguistic variety or one single genre. The specific linguistic variety or genre is set at the beginning of corpus design and this delimits the types of texts one can include in such corpora. Examples of specialised corpora are CANCODE, a collection of informal registers of British English with 5 million words (Hunston, 2002), Tobacco Industry Documents Corpus with 4 million words (Kretzschmar et al., 2004) and the corpus of the English of computer science with 1 million words (Cheng-yu, 1993). As it is demonstrated in Section 4.3, specialised corpora for specific pedagogic purposes have also been created. A corpus of 40 letters of application and 20 'introduction to speaker' speeches were, for example, used to identify key lexical and rhetorical elements of the particular genre for teaching purposes (Henry & Roseberry, 2001). Mudraya (2006) built the Student Engineering

English Corpus of 2 million running words in order to develop a lexical syllabus for engineering students. In an innovative English language course for non-native doctoral students participants are asked to build and analyse their self-compiled corpora of research articles and their own writings. These corpora were of really small size with running words between 40,000 and 250,000 each (Lee & Swales, 2006).

Considering the size of a corpus involves not only a decision on the mere number of words in the collection, but it also includes decisions on the number of text types, the number of samples of each text type and also the size of each sample (Biber et al., 1998; Kennedy, 1998). These issues will be dealt with in the next section.

The conclusion that can be drawn from the above is that the overall size of a corpus is very much dependent on the purpose that the corpus will eventually be used for. What ultimately determines the number of texts and size of samples is what we would like to analyse and what the specific aim of the analysis is. An additional factor here is **comparability**. As there are also certain traditions in corpus linguistics as regards the size of corpora, this has to be taken into consideration when deciding on the number of running words in a corpus and the size of samples as well. A box with the heading 'Comparability' represents this factor in the model including all the following phases of the design and implementation process. The arrow pointing to the box of 'Comparability' comes from the source of information box 'Earlier corpus studies' and within that from studies which have similar research aims (see Figure 12).

#### **4.4.6.2. Sample size**

As regards the size of samples in a corpus, there are two main approaches to determine sample size in corpora. The first is to take extracts of a predetermined number of words from carefully preselected texts. These are very often extracts of 2000 words following the traditions of the LOB and Brown Corpus; for example, the International Corpus of English (Leitner, 1992;

Nelson, 1996) and the corpus of the English of Computer Science (Cheng-yu, 1993) used this method. Other scholars suggest that 2000-word extracts are not sufficient to represent a text type or genre as linguistic features are not evenly distributed in texts, and with random selection certain characteristics might be lost. Therefore, a sample size of 20,000 words was suggested as an appropriate size to provide statistically reliable results (Kennedy, 1998; Nelson, 2000). An example of this method is the BEC by Nelson (2000).

The second approach is to include whole texts. Sinclair (1991), a strong advocate for whole texts in corpora, argued:

The alternative is to gather whole documents. Then there is no worry about the marked differences that have been noted between different parts of a text. Not many features of a book-length text are diffused evenly throughout, and a corpus made up of whole documents is open to a wider range of linguistic studies than a short collection of samples. (p. 19)

To include whole texts is especially important when studying genre and discourse features, as results of earlier research in these fields suggested that certain linguistic characteristics are typical of certain parts of a text. An example of a corpus with whole texts is the Bank of English corpus (Renouf, 1987).

However, decisions and considerations of these phases lead to an ideal structure, content and size of a corpus which many corpus linguists consider almost impossible to achieve in reality (Biber et al., 1998; Nelson, 2000). It is a common feature in all kinds of corpus creation that theoretically-based decisions and criteria have to be applied in a flexible way because of practical constraints. Therefore a box with broken lines represents the 'Ideal corpus' in the model (see Figure 12). The broken line of the box symbolises the fact that ideal contents cannot be fully realised.

Next in the corpus creation process, following the theoretical considerations outlined above, are the implementation phases of data collection and data entry. Issues here include the difficulties of collecting samples of language from authentic sources, entering data into the corpus, and gaining permission for copying entire texts or extracts from copyright holders.

#### **4.4.7. Data collection**

One of the sources for a corpus study is **publicly available data**. These include, for example, newspapers, journals, sites on the Internet and magazines. Advantages of these sources are that (1) such texts are very often available in an electronic form; (2) they are easily accessible. One disadvantage, however, is that usually they only reflect certain limited aspects of language use. For example, in creating the BEC, Nelson (2000) found that these sources provided texts talking about business, but they did not reflect the actual process of doing business.

Therefore other sources have to be used as well. Texts from other sources are named as **private data** by Nelson (2000), and refer to all documents that are not publicly available. Private data include texts such as business letters, handwritten notes and recorded conversations. However, most difficulties occur when gathering this type of data, as they are not easy to access and special permissions are necessary for using them in research (G. Nelson, 1996; M. Nelson, 2000).

#### **4.4.8. Data entry**

The second issue that is related to practical considerations is **data entry**. This concerns the way in which texts will be converted into a machine-readable format. There are technically three ways to do this. The first is keyboarding, that is, typing the texts manually into the computer database, which is probably the most time-consuming method, but necessary if the document is not available in an electronic form. The second way is scanning. In this case it is

important to make sure that the scanned text matches the original, as in some cases (e.g., bad photocopies) the quality is not good enough for scanning. The third way is to adapt texts that are available in electronic format. Depending on the corpus analysis tool or the storage system, even these types of texts may have to be converted into a format that is used in the database. For example, texts on the Internet are frequently in html, pdf or doc format, and they have to be plain text files for the corpus (Kennedy, 1998; Nelson, 2000). The final format of texts, as regards character sets and language encoding, is also determined by the corpus analysis tool used for analysing the corpus. Some tools can only handle one type of character set, while others can decode several types. Settings functions allow users to prescribe which character set is to be used for encoding the texts in the corpus.

The methods mentioned above mainly concern written texts or already transcribed speeches. In the case of spoken language data the tedious job of creating the transcription has to be added to the steps of data entry.

Practical considerations of data entry involve not only decisions on the format of the data, but also decisions on the whole data storage system. When creating a database, one has to bear in mind that the data must be easily retrievable. A straight-forward way for small corpora is to use a file manager for this purpose. Sub-corpora can be created by putting the texts into the same folder, or by merging several text files into a single text file. Coded file names can also be helpful for selecting relevant texts from the whole corpus for certain analyses. More sophisticated corpora like the BNC and the Bank of English have their own storage and retrieval systems.

Corpus compilers can also encode their texts in order to signpost different parts of the texts such as word boundaries, hesitation in spoken texts, line numbers, and so on. The first system developed for the electronic encoding of texts originated from the publishing industry in the 1980s. The Standard Generalised Markup Language was used outside the publishing

industry for corpus linguistic purposes as well. In the 1990s, the need for an agreement on what features of a text should be encoded arose, and the *Text encoding initiative guidelines*, a complex application of SGML was designed. The TEI Guidelines help establish standards among scholars creating their electronic corpora, regardless of the language of the corpus. This flexible encoding system gives assistance to researchers on what to mark up in a text and how to mark it up. It gives the freedom to the individual scholar, however, to decide how detailed the markup in the given corpus should be (Kennedy, 1998; Guidelines for Electronic Text Encoding and Interchange, 2002).

Probably the most widely used coding of corpora are so called POS tags as mentioned also in Section 4.1.2.3. There are automatic taggers available, but in order to reliably tag a corpus automatic tagging has to be corrected manually.

#### **4.4.9. Ethical and legal issues**

As in all types of research, in corpus research there are also certain ethical and legal issues to consider. One of the most important considerations is getting permission for using texts for research purposes. The difficulties involved in gaining copyright permission are well recognised in the literature (Kennedy, 1998; Nelson, 2000; Renouf, 1987; Sinclair, 1991). The general experience was that although copyright holders were willing to grant permission, the whole process took quite a long time, and it is advisable to collect more samples for the different categories than originally planned, as copyright holders might be very slow to react to requests (Nelson, 1996; Kennedy, 1998).

Another important issue is confidentiality. Corpus compilers have to make every effort not to invade personal privacy in any way. In earlier studies, spoken samples were taken with concealed microphones or cameras, but this is now unacceptable. Participants need to be

informed about being recorded, and should have the right to listen to the recording before giving their consent to any analysis.

In corpus creation practical considerations are as important as theoretical underpinning and researchers in corpus linguistics should be aware of these practical difficulties. As Nelson (2000) expressed it: “Any attempt at corpus creation is therefore a compromise between the hoped for and the achievable” (p. 250). Therefore the two underlying factors ‘Feasibility and availability’ are added to the model including all the phases of implementation and the ‘Final corpus’ as well (see Figure 12).

#### **4.4.10. Summary**

This section gave an overview of issues that are to be considered for corpus design and creation for ESP. It proposed a *Model for Corpus Creation for ESP* to support a systematic and disciplined corpus design and implementation process. The model is based on the findings of corpus linguistics and guiding principles already proposed in the literature on corpus building. The proposed model was also used to guide the corpus design and creation stage of the present study. In what follows the specific research aims of this study will be presented with details of the research design, including the description of the main stages and procedures of the study.

## Chapter 5: Aims and research questions

The aims of the present study are:

- to identify the genres and specific documents that EU professionals working in different EU subject fields consider relevant and, therefore, can be regarded as being representative of written English EU discourse,
- to identify the lexical items that are typically associated with written English EU discourse,
- to analyse and describe the patterns of lexis in written English EU discourse and
- to formulate recommendations for course design and materials development for English language courses for EU studies and occupational purposes within the EU context.

Following on from these aims, the study is guided by these specific research questions:

1. What genres and specific documents can be regarded as representative of written English EU discourse?
  - 1.1. What EU genres and specific documents do Hungarian EU professionals use in their daily work?
  - 1.2. What are the genres common to Hungarian EU professionals working in different EU subject fields?
  - 1.3. What EU genres and specific documents do Hungarian EU professionals consider useful for the preparation of future EU professionals?
  - 1.4. How are EU texts used by Hungarian EU professionals?
2. What lexical items are especially associated with written English EU discourse?
  - 2.1. What are the most frequent lexical items used in a wide range of EU texts?



- 2.2. What collocational patterns emerge in written English EU discourse?
- 2.3. What are frequent MWIs in written English EU discourse?
- 3. What pedagogical implications do the findings have for teaching English for EU purposes with special emphasis on course design and materials development?

## Chapter 6: Research design and procedures of analysis

This chapter will give an overview of the stages, steps and procedures of the analysis. The analysis was conducted in two stages. The two stages combined the analysis of linguistic and pedagogic aspects of written English EU discourse, and included the application of quantitative, as well as qualitative analytical frameworks for the creation and investigation of the EEUD Corpus. A summary of the two stages, with the main steps and procedures, can be seen in Table 18.

### 6.1. Stage 1: Needs analysis and corpus creation

In the **first step** in Stage 1, a socially oriented perspective was taken in order to identify the relevant genres that may be regarded as representative of written English EU discourse from the perspective of Hungarian EU professionals working in different subject fields within the EU context. This investigation was conducted as a needs analysis survey in the form of questionnaire research. A detailed description of the process of questionnaire construction and administration can be found in Section 6.3. The **second step** was to create the EEUD Corpus on the basis of the results of the survey. Following the proposed *Model for Corpus Creation for ESP* described in Section 4.4, this step involved the design of an ideal corpus and the compilation of the final EEUD Corpus, based on the results of the needs analysis survey. A detailed discussion of the design and compilation of the final corpus can be found in Section 7.1. Texts for the corpus were collected from two major sources. The majority of texts – slightly more than 60% – were sent by the participants of the survey. Additional EU texts were downloaded from the official websites of EU institutions. Copyright issues needed to be tackled. However, as all texts were issued by EU institutions, it did not cause any difficulties. EU texts can be used for research purposes on the condition that the copyright of the European

Union is acknowledged. Therefore, a table listing all the texts in the final EEUD Corpus is presented in Appendix 4. In the phase of corpus compilation the texts received from the participants of the survey needed to be converted from their original format, mainly pdf or Word files, into plain text files. The sample EU texts were kept at their original length, but the reference sections, where different pieces of EU legislation are listed, were deleted.

<b>Procedures of research</b>	<b>Data collection tools and methods of analysis</b>	<b>Results</b>
<b>Stage 1: Needs analysis and corpus creation</b>		
Step 1: Questionnaire construction	<ul style="list-style-type: none"> <li>• interview</li> <li>• expert judgement</li> <li>• think aloud</li> </ul>	<ul style="list-style-type: none"> <li>• needs analysis questionnaire</li> </ul>
Step 2: Needs analysis survey	<ul style="list-style-type: none"> <li>• online needs analysis questionnaire</li> </ul>	<ul style="list-style-type: none"> <li>• list of common EU genres</li> <li>• representative English EU texts</li> <li>• uses of English EU texts</li> </ul>
Step 3: Corpus building	-	<ul style="list-style-type: none"> <li>• English EU Discourse Corpus</li> </ul>
<b>Stage 2: Corpus analysis</b>		
Step 1	<ul style="list-style-type: none"> <li>• analysis of frequency and range of lexical items</li> </ul>	<ul style="list-style-type: none"> <li>• EUWL</li> </ul>
Step 2	<ul style="list-style-type: none"> <li>• key word analysis</li> <li>• comparison of key words and EUWL</li> <li>• collocational analysis of selected items in EEUD and BNC Written</li> </ul>	<ul style="list-style-type: none"> <li>• key words of EEUD</li> <li>• selected lexical items for further analysis</li> <li>• collocations in EEUD</li> </ul>
Step 3	<ul style="list-style-type: none"> <li>• corpus-driven analysis of MWIs in EEUD</li> </ul>	<ul style="list-style-type: none"> <li>• lexical bundles in EEUD</li> </ul>

Table 18. Summary of the main procedures of research

## 6.2. Stage 2: Corpus analysis

In Stage 2, the study focused on the lexical and lexicogrammatical characteristics of EU texts. The analyses comprised quantitative elements that were conducted with the help of corpus analysis tools, and qualitative elements of manual analysis. As can be seen in Table 19, for most analyses the corpus investigation package *WordSmith Tools 4* (Scott, 2004), was used. *WordSmith Tools* was chosen because most studies conducting lexical analyses apply them, thus results can easily be compared to findings of previous research. The other reason was that it can carry out the analyses of this study at high quality (Nelson, 2000; Reppen, 2001).

Another tool used was the *Sketch engine* (Kilgarriff & Tugwell, 2001). This corpus analysis tool has been especially developed for lexicographic research. In the present study it was applied for the collocational analysis of selected lexical items.

Type of analysis	Aim of analysis	Results of analysis	Tool used for the analysis
Quantitative analysis	<ul style="list-style-type: none"> <li>• analysis of frequency and range of lexical items</li> <li>• validation of the EUWL</li> </ul>	<ul style="list-style-type: none"> <li>• frequency list of the EEUD Corpus</li> </ul>	<i>WordSmith Tools 4</i>
		<ul style="list-style-type: none"> <li>• text coverage of the EUWL in several registers and genres</li> </ul>	<i>Range programme</i>
	<ul style="list-style-type: none"> <li>• key word analysis</li> </ul>	<ul style="list-style-type: none"> <li>• words particularly associated with EU discourse</li> </ul>	<i>WordSmith Tools 4</i>
	<ul style="list-style-type: none"> <li>• collocation analysis</li> </ul>	<ul style="list-style-type: none"> <li>• collocations of selected lexical items</li> </ul>	<i>Sketch engine</i>
Qualitative analysis	<ul style="list-style-type: none"> <li>• identification of four-word clusters</li> <li>• semantic analysis of key words</li> <li>• manual analysis of concordance lines</li> <li>• semantic analysis of collocates of selected lexical items</li> </ul>	<ul style="list-style-type: none"> <li>• four-word lexical bundles</li> </ul>	<i>WordSmith Tools 4</i>
		<ul style="list-style-type: none"> <li>• main semantic categories of key words in EU discourse</li> </ul>	<i>WordSmith Tools 4</i>
		<ul style="list-style-type: none"> <li>• similarities and differences of collocational patterns in the EEUD Corpus and the BNC</li> </ul>	<i>Sketch engine</i>

Table 19. List of qualitative and quantitative analyses applied in Stage 2

A further corpus analysis tool was used for the creation of the EUWL. The programme is called *Range* (Heatley, Nation, & Coxhead, 2002). The main reason for applying it to this particular analysis was that all studies in ESP used *Range* for the creation of word lists (Coxhead, 2000; Trebits, 2008; Wang et al., 2008), therefore, results were easily comparable.

As can be seen in Table 18, Stage 2 of the corpus analysis was divided into three further steps, each of them focusing on one particular aspect of the analysis of the lexis of the texts in the EEUD Corpus. The analysis started with gaining an overall view of the most frequent lexical items in the EEUD Corpus, by creating the EUWL and identifying key words. Next, patterns of behaviour of selected frequent lexical items were analysed with the help of collocational analysis, and the analysis of semantic preference and semantic prosody. The patterns in the EEUD Corpus were compared to patterns identified in the written part of the

BNC. Finally, MWIs in the EEUD were identified with a corpus-driven approach as lexical bundles. A detailed description of the individual steps will be given in the following chapter.

### **6.3. Design and compilation of the English EU Discourse Corpus**

In the following a detailed description of the steps and procedures of the corpus creation is given. The main steps of corpus design and corpus creation applying the *Model for Corpus Creation for ESP* can be divided into theoretical issues such as questions related to size, representativeness and sampling, and practical issues concerning data collection, data entry and legal issues (Kennedy, 1998; McEnery & Wilson 1996a; Nelson, 2000; Sinclair, 1991).

A small, specialised corpus, as defined earlier, seeks to represent a special language variety. This can be a linguistically or socially determined variety of a given language. As the aim of this study is to investigate a particular linguistic variety, that is, written English EU discourse in English, the corpus for this study falls into the category of specialised corpora. The preliminary structure of the corpus was defined on the basis of similar specialised corpora in ESP research (Coxhead, 2000; Mudraya, 2006; Nelson, 2000; Wang et al., 2008). Based on the characteristics of these earlier specialised corpora, the preliminary structure of the EEUD Corpus can be described as follows: the corpus should contain about 1 million running words of whole texts; it should cover all the special EU subject fields like agriculture, monetary union, etc., and should include a wide variety of different EU genres used by EU professionals. As regards the time period texts represent, the corpus should comprise texts that have been issued since 2000.

Regarding the practical issues of corpus design, special attention should be paid to copyright issues in the course of the text collection procedure. A summary of the necessary steps and relevant sources of information and issues to consider in the course of corpus design and creation for the study is given in Table 20.

After defining the purpose of corpus creation and the type of corpus, the genres and specific documents for the corpus were defined. The selection process included the following phases:

1. Defining EU discourse, that is, defining the target situation and members of the discourse community as appropriate for the purposes of the study;
2. Conducting a needs analysis survey in order to identify common EU genres and to define the external criteria for the final text selection;
3. Compiling the corpus based on the results of the survey taking the common genres and the representation of the different special EU subject fields into consideration.

<b>Steps</b>	<b>Issues to consider, sources of information</b>
1. Define the aim and purpose of the corpus	<ul style="list-style-type: none"> <li>• analysing written English EU discourse for pedagogic purposes</li> <li>• specialised corpus</li> </ul>
2. Define the type of the corpus	
3. Study the population the corpus should represent	<ul style="list-style-type: none"> <li>• set external criteria for text selection based on: <ul style="list-style-type: none"> <li>• findings of EU studies</li> <li>• findings of earlier linguistic research into EU discourse</li> <li>• findings of translation in an EU context</li> <li>• needs analysis among participants as defined in 6.3.1</li> </ul> </li> </ul>
4. Define the structure and content of the corpus	<ul style="list-style-type: none"> <li>• time period: since 2000</li> <li>• based on the results of the needs analysis: <ul style="list-style-type: none"> <li>• relevant EU genres</li> <li>• proportion of EU genres</li> <li>• diversity of EU subject fields</li> <li>• balance of EU subject fields</li> <li>• representativeness</li> </ul> </li> </ul>
5. Define the size of the corpus	<ul style="list-style-type: none"> <li>• number of words: around 1 million</li> <li>• number of samples</li> <li>• whole texts</li> <li>• comparability with other ESP corpora</li> <li>• feasibility of analysis</li> </ul>
6. Data collection	<ul style="list-style-type: none"> <li>• adequate sources: official EU websites,</li> <li>• texts sent by respondents of the needs analysis</li> <li>• copyright issued clarified with EU Publication Office</li> </ul>
7. Data entry	<ul style="list-style-type: none"> <li>• methods: <ul style="list-style-type: none"> <li>• electronic versions converted from pdf, html and doc format into plain text</li> </ul> </li> </ul>

Table 20. Steps of the design and compilation of the EEUD Corpus

### **6.3.1. Defining written English EU discourse**

Written English EU discourse was defined for the purposes of the present study by, firstly, delimiting the language users, that is, members of the EU discourse community, using the following categories:

- Hungarian EU professionals who use official English EU documents in their jobs
- lecturers in EU studies and teachers of English for the EU
- EU translators and interpreters

Secondly, EU texts for the analysis were, more specifically, defined as English texts issued by EU institutions.

### **6.3.2. Needs analysis survey for corpus design**

As discussed in Sections 4.3 and 4.4.6.1, there have been several corpus studies in ESP investigating small, specialised corpora. An overwhelming majority of these corpora encompass extracts from textbooks and journal articles of the special disciplines and professional fields (Coxhead, 2000; Cortes, 2004; Gledhill, 2000; Hyland, 2008; James et al., 1994; Mudraya, 2006). These textbooks or research articles were selected based on recommended lists for undergraduates. Few studies applied some form of surveys among subject specialists in the special disciplines. Cortes (2004), for example, interviewed seven professors in order to select relevant journals and textbooks for her corpus. Gledhill (2000) conducted an ethnographic survey among professors at university to select texts for his corpus of pharmaceutical sciences.

Different ways to define corpus structure and selection criteria for corpus building was applied by Nelson (2000) and Trebits (2008). The ideal structure of Nelson's BEC was based on the literature of BE. Trebits (2008), when compiling her EU English Corpus considered the needs of a future EU expert, and selected EU texts from EU websites randomly and intuitively.

A further method of selecting relevant texts for a specialised corpus was used by Krausse (2005). In order to define the topics and genres for her Environmental Engineering Corpus, she

conducted a questionnaire survey among companies that employed graduates of the degree course for which her English classes were designed. She conducted the survey in two rounds. In the first round she gathered information to decide on topics and genres, and in the second round she asked for sample texts that were used in the companies' daily work routines. The response rate in the two rounds was very different. Although more than half of the companies sent back the questionnaires in the first round; in the second round, only eleven questionnaires were responded to.

On the basis of previous research practices and experience, a needs analysis type of questionnaire survey in one round distributed among members of the EU discourse community, was found appropriate for the purposes of the present research. As discussed in Section 2.2.7, needs analyses in general have an underlying and fundamental role in course and materials design in ESP. Among the several forms and focuses, needs analysis can take elements of the target situation analysis with a focus on the genres used in the target situation were deemed to be relevant. Thus the aim of the needs analysis survey was to map out the target situation for which students of EU studies need to be prepared for.

#### **6.3.2.1. Questionnaire construction**

The questionnaire survey method has been found to be the most appropriate for the purposes of this study, especially the application of an online questionnaire. The online questionnaire made it possible to reach Hungarian EU professionals working in Brussels, Luxembourg, or in other geographically distant places, too.

Designing a survey instrument involves several steps and decisions about question types, order of questions, format (Brown, 2001), and the language of the questionnaire. As the survey was conducted among potential respondents with a Hungarian mother tongue, the questionnaire was compiled in Hungarian. The final questionnaire has two main sections. The first section



contains questions on biographical data, like occupation, age, command of English, EU institutions they worked for, time spent at EU institutions with English as the working language, command of other languages, special subject field within the EU context, etc. The second section contained questions referring to specific English EU genres. In this section respondents were asked to rate the frequency and importance of the use of certain EU genres on the five-point Likert scale with options for frequency ranging from *never* or *rarely* or *monthly* to *weekly* or *daily*<sup>1</sup>, and for importance, with the following options *it has nothing to do with my job*, *not important*, *important*, *very important* and *indispensable for my job*<sup>2</sup>. This format, with modifications, was taken from a questionnaire used to identify genres in ESP (Tompos, 2001). Tompos used her questionnaire to validate a list of genres for ESP testing purposes. She applied her survey instrument to demonstrate that the text types and genres listed are recognised and used by the professional respondents from various fields such as agriculture, business and economics, engineering, law and medicine, in their work. As the questionnaire in the present research aimed to identify the genres Hungarian EU professionals, working in different EU subject fields, and at different EU institutions, recognise and use in their work, the application of a similar format was found appropriate and useful. The list of EU genres was established, based on the literature of translation for EU institutions, recommendations by EU experts in interviews, the official portals of EU bodies, like the European Commission and the European Parliament, and the official legal portal of the EU, EURlex. EU genres were grouped into five **text categories**. The categories *primary legislation*, that is, treaties (8 items), *secondary legislation*, that is, for example, regulations and decisions, and *EU case law*, such as

---

<sup>1</sup> The original wording is as follows: *soha, ritkán, havonta, hetente, naponta* (see Appendix 1 for a print out of the full original questionnaire.)

<sup>2</sup> The original wording is as follows: *semi köze a munkájához, nem fontos, fontos, nagyon fontos, elengedhetetlen a munkájához* (see Appendix 1 for a print out of the full original questionnaire.)

judgements of the court (13 items), and *legislative preparatory texts*, such as legislative proposals, ECOSOC opinions (12 items), were based on the text categories used in the EURlex portal. Two further categories, that is, documents related to *application for EU funds* (5 items) and *other documents* for information and other purposes (10 items), were added. Each subsection referring to one text category had the following type of questions in the same order: (1) Likert scale of frequency of use, (2) Likert scale of importance for respondent's job, (3) an open question asking for further genres or specific documents in this category that respondents used, but were not listed and (4) a question asking how the texts in this category are usually used, with ten options, and a blank space for adding other uses. There were three more open questions in this section of the questionnaire. One asking for specific examples of English EU documents that were used by the respondents around that time (question 13), one question asking for recommendations for EU documents that were thought to be relevant for students and future EU professionals (question 17), and the final question asking respondents to send English EU documents that they used in their work (question 35).

The final form of the questionnaire (see Appendix 1) contained altogether 35 questions, of which 12 are background questions, 23 are concerned with EU genres and specific documents, and their uses in the EU context.

#### **6.3.2.2. Establishing the validity and reliability of the instrument**

Research manuals emphasise the importance of the validation of survey instruments (Brown, 2001; Hatch & Lazaraton, 1991; Seliger & Shohamy, 1989). Although validation is accepted as a necessary step in research, “[t]here are no general principles of good pretesting, no systemization of practice, no consensus about expectations ...” (Converse & Presser, 1986, p. 52). Therefore, the decisions on the validation process in this study were based partly on the

literature, for example, concerning the choice of methods and general principles, and also on what was felt feasible and necessary under the given time and practical constraints.

The choice of validation methods were based on Cohen, Manion and Morrison (2000), who discuss theoretical considerations for validity and reliability, on Alderson and Banerjee (1996), who overview the literature of validation of research instruments, and on Brown (2001), Converse and Presser (1986), and Petric and CzárI (2003), who offer practical guidance for validation in general, and for questionnaire construction, in particular. Alderson and Banerjee (1996) recommended the combination of more methods of validation. Therefore, several qualitative methods like interviews and think aloud protocols with individuals of different backgrounds from the target population, and expert opinions, were used to test the instrument for reliability and validity. A step-by-step summary of questionnaire construction and validation is given in Table 22.

Researchers in the education and language teaching fields distinguish several types of validity of research instruments (Brown, 2001; Cohen et al., 2000). Considering the purpose, to collect relevant EU documents, and to identify relevant EU genres for corpus building, and the context of the research instrument of the current study, that respondents are in a stressful working environment, and access to and availability of, respondents are limited, the most relevant types of validity were considered to be content validity, construct validity, and response validity.

The procedures suggested for establishing **construct validity** are statistical procedures and comparison to theory (Brown, 2001; Petric & CzárI, 2003). The most widely used statistical procedure for testing validity of a questionnaire is factor analysis. Several researchers (Alderson & Banerjee, 1996; Petric & CzárI, 2003) have, however, expressed their concerns regarding the appropriateness of factor analysis for validating questionnaires. Alderson and Banerjee (1996) also question the validity of factor analysis and the interpretability of the results, and suggested

that “factor analysis in instrument validation has to be seen as tentative and suggestive” (Alderson & Banerjee, 1996, p. 29). According to Alderson and Banerjee (1996), testing construct validity should involve triangulation, that is, constructs should be tested from several perspectives. Therefore, instead of applying statistical procedures checking construct validity of the questionnaire was based on triangulation of data sources, that is, representatives of different groups of respondents were interviewed, and relevant constructs were compared to available literature.

The two constructs in the research instrument are, EU genres, and application of these genres. Establishing a list of relevant items of these constructs involved two types of interviews, and a literature search of the relevant fields. These steps are summarised in Step 1-3 in Table 22. The first interview was an informal focus group interview with four students of International Relations who also took a ‘Translation of English EU texts’ minor. The interview was of exploratory nature, with three main questions about EU genres and their uses, and perceived difficulties with these EU genres. The purpose of the interview was to compile an initial list of relevant EU genres and how these are used by the students, and to gain some insights into what difficulties students might have in connection with reading and understanding these EU genres. The initial lists of EU genres, and the way they are applied are given in Tables 21 and 23. The perceived difficulties included several language related issues, such as EU terminology, long, logically complex sentences, unknown legal terms, abbreviations, and the lack of Hungarian equivalents. Students also suggested that in some cases texts were probably not written by native speakers, and contained expressions that are not used in English. The interview was conducted in Hungarian, as this was the mother tongue of both the interviewees and the researcher.

Although a few genres were already collected in the interview with students, the most important step in this respect was the second round of interviews with EU professionals and

interns, as they could inform the researcher about relevant genres of the target situation. EU professionals included four EU experts who work at universities, research institutes, and at the prime minister’s office. They also teach EU studies in higher education institutions. The other EU professionals were two EU translators and an interpreter who worked at the European Commission. One of the interns worked at the EU department of the Hungarian Ministry of Foreign Affairs, and the other worked at EuroDirect, an information service about EU issues for the general public. The interviews were conducted in Hungarian. They were recorded with the permission of the participants. The purpose of these interviews was to extend the initial lists of relevant English EU genres and their application. The interviews were semi-structured, as there was a set list of questions, but participants were encouraged to elaborate freely on related issues. The interview protocol is presented in Appedix 2. The interviews were also used to test some questions and options of the final questionnaire for wording and format, thus ensuring **reliability** as well. The results of these semi-structured interviews, in the form of two extended lists, can be seen in Table 21 and 23.

<b>Initial list based on focus group interview</b>	<b>Extended list based on semi-structured interviews</b>	<b>Final list</b>
<ul style="list-style-type: none"> <li>• thesis writing</li> <li>• translation</li> </ul>	<ul style="list-style-type: none"> <li>• skimming</li> <li>• scanning</li> <li>• proofreading</li> <li>• interpreting</li> <li>• legal application</li> <li>• law harmonisation</li> <li>• creating word lists</li> <li>• research</li> <li>• writing articles</li> <li>• preparation for teaching</li> <li>• summarising text in Hungarian</li> </ul>	<ul style="list-style-type: none"> <li>• skimming for general information</li> <li>• as a template for writing</li> <li>• finding specific EU terms</li> <li>• scanning for specific information</li> <li>• translating the text</li> <li>• interpreting the text</li> <li>• summing it up in English</li> <li>• summing it up in Hungarian</li> <li>• collecting EU terminology</li> <li>• I do not use this type of document</li> <li>• other, please specify:</li> </ul>

Table 21. Uses of EU texts based on the interviews

The final list of EU genres was based on two main EU sources related to the categorisation and publication of EU documents. One of them was the Prelex Manual<sup>3</sup> that provides a list of types of documents for following official documents transmitted from one EU institution to the other. The second EU source was the EURlex, the official portal of EU legislation<sup>4</sup>. Creating the final list of EU genres on the basis of these official EU documents and texts categories also helped enhance the **reliability** of the instrument, as it used the officially established text categories that were very likely to be recognised by the potential respondents.

---

<sup>3</sup> The European Commission Prelex available at [http://ec.europa.eu/prelex/ct/sgv\\_manual\\_dsp\\_main.cfm?manualcat\\_id=documents&cl=en](http://ec.europa.eu/prelex/ct/sgv_manual_dsp_main.cfm?manualcat_id=documents&cl=en)

<sup>4</sup> EURlex available at [http://eur-lex.europa.eu/en/droit\\_communaire/droit\\_communaire.htm](http://eur-lex.europa.eu/en/droit_communaire/droit_communaire.htm)

Step	Participants/Source	Method	Questions	Results	Aspect of validation
Step 1	<ul style="list-style-type: none"> <li>• 4 fifth-year International Relations students with translating EU texts minor</li> </ul>	<ul style="list-style-type: none"> <li>• unstructured focus group <i>interview</i> conducted in Hungarian</li> </ul>	<ol style="list-style-type: none"> <li>1 What kind of English EU documents did you use?</li> <li>2 What did you use English EU documents for?</li> <li>3 What difficulties did you have with these documents?<sup>5</sup></li> </ol>	<ul style="list-style-type: none"> <li>• initial list of potentially relevant EU genres</li> <li>• initial list of uses of EU genres</li> <li>• perceived difficulties with reading EU documents</li> </ul>	<ul style="list-style-type: none"> <li>• construct validity</li> </ul>
Step 2	<ul style="list-style-type: none"> <li>• 9 members of the target population: <ul style="list-style-type: none"> <li>• 4 EU experts</li> <li>• 2 student interns in EU-related institutions</li> <li>• 2 EU translators</li> <li>• 1 interpreter at the European Commission</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• semi-structured <i>interviews</i> conducted in Hungarian (recorded)</li> </ul>	<ol style="list-style-type: none"> <li>1 What did you use English EU documents for?</li> <li>2 What difficulties did you have with documents issued by EU institutions?</li> <li>3 Please show me an example of an English EU document that you used!</li> <li>4 What linguistic preparation did you receive before starting your career?</li> <li>5 What type of linguistic preparation would you recommend for students of your profession?</li> <li>5.1 What documents, texts should students be made familiar with?</li> <li>5.2 What should an EU English language course focus on? (see interview protocol in Appendix 2)</li> </ol>	<ul style="list-style-type: none"> <li>• extended list of uses of EU genres</li> <li>• extended list of potentially relevant EU genres</li> <li>• testing some questions for format and wording</li> </ul>	<ul style="list-style-type: none"> <li>• construct validity</li> <li>• reliability</li> </ul>
Step 3	<ul style="list-style-type: none"> <li>• journal articles on translating EU documents</li> <li>• EURlex the legal portal of the European Union</li> <li>• Prelex Manual</li> </ul>	<ul style="list-style-type: none"> <li>• literature search</li> </ul>	<ul style="list-style-type: none"> <li>• established EU genres and genre names</li> </ul>	<ul style="list-style-type: none"> <li>• five text categories</li> <li>• list of documents and genres</li> </ul>	<ul style="list-style-type: none"> <li>• construct validity</li> <li>• reliability</li> </ul>
Step 4	<ul style="list-style-type: none"> <li>• 1 expert in research methodology</li> <li>• 1 expert in text analysis</li> </ul>	<ul style="list-style-type: none"> <li>• <i>expert opinion</i></li> </ul>	<ul style="list-style-type: none"> <li>• judge first draft of the questionnaire</li> </ul>	<ul style="list-style-type: none"> <li>• modification in question type, wording and order of questions</li> </ul>	<ul style="list-style-type: none"> <li>• content validity</li> </ul>
Step 5	<ul style="list-style-type: none"> <li>• 1 graduate of EU translator and interpreter course</li> </ul>	<ul style="list-style-type: none"> <li>• <i>think aloud</i> of the paper-based format of the questionnaire (recorded)</li> </ul>	<ul style="list-style-type: none"> <li>• pretest modified questionnaire</li> </ul>	<ul style="list-style-type: none"> <li>• modifications in wording and order of questions</li> </ul>	<ul style="list-style-type: none"> <li>• content validity</li> </ul>
Step 6	<ul style="list-style-type: none"> <li>• 1 assistant in EU context</li> <li>• 1 ESP and translation teacher</li> <li>• 1 business professional</li> </ul>	<ul style="list-style-type: none"> <li>• developmental <i>think aloud</i> of the online version of the questionnaire (recorded)</li> </ul>	<ul style="list-style-type: none"> <li>• pre-test modified questionnaire</li> </ul>	<ul style="list-style-type: none"> <li>• modifications in wording and order of questions</li> <li>• final version</li> </ul>	<ul style="list-style-type: none"> <li>• response validity</li> <li>• reliability</li> </ul>

Table 22. Main steps of questionnaire construction and validation

<sup>5</sup> The Hungarian version of the questions was as follows: 1. *Milyen angol nyelvű EU szövegeket használt?* 2. *Mihez használt angol nyelvű EU szövegeket?* 3. *Mi okozott nehézséget az EU szövegek használatát során?*

**Content validity** was established by asking two experts to give their opinion on the draft of the questionnaire. One of them was an expert in research methodology, and the other one on text and discourse analysis. They were asked to comment on the relevance of questions to the purpose of the questionnaire, possible wording and interpretation problems in questions and instructions, and on the type and order of questions. Content validity was also tested by the think aloud procedure with a member of the target population. Issues of wording of questions, the relevance of the items in the list of EU genres, were discussed, and additional ideas were invited in order to ensure that all genres relevant for the target population are covered. The content validity check resulted in major changes. Based on the **expert opinion**, the proportion of open questions was reduced, and the wording of the Likert scales was also modified. Experts also expressed their concerns as regards the overall length of the questionnaire, which was addressed in the **think aloud protocol**. The think aloud, using a paper-based format of the questionnaire, resulted in a few changes to the wording of questions in the background section, but the respondent did not find the questionnaire too long overall, and filled it out in a very straightforward manner that took her 20 minutes. As the final administration of the questionnaire was planned to take place using an electronic format, it was decided that the issue of overall length should be addressed in the final format again.



Initial list based on focus group interview	Extended list based on semi-structured interviews	Final list
<ul style="list-style-type: none"> <li>• Constitution</li> <li>• Documents in connection with CFSP</li> <li>• Press release</li> <li>• Treaty</li> </ul>	<ul style="list-style-type: none"> <li>• Annual report</li> <li>• Application form</li> <li>• Call for proposals</li> <li>• Commission communication</li> <li>• Commission leaflet</li> <li>• Commission legislative proposal</li> <li>• Common European format of CV</li> <li>• COREU</li> <li>• Council Decision</li> <li>• Decision</li> <li>• Declaration (e.g., Berlin declaration)</li> <li>• Directive</li> <li>• Fact sheet on the EU</li> <li>• Judgement of the Court of Justice</li> <li>• Opinion</li> <li>• PRAG</li> <li>• Presidency conclusions</li> <li>• Project contract</li> <li>• Recommendation</li> <li>• Resolution</li> <li>• Sessions of the European Parliament</li> <li>• Treaty (e.g., Single European Act, Treaty of Nice, Consolidated version of the treaties, Treaty of Rome)</li> </ul>	<ul style="list-style-type: none"> <li>• see final questionnaire in Appendix 1</li> </ul>

Table 23. EU genres from the interviews

**Response validity** and **reliability** were tested in the form of **think-aloud protocols** referred to as “participating pretest” by Converse and Presser (1986). In their practical guide for questionnaire construction they stated that “a minimum of two pretests are indispensable” (p. 65), the first one being a developmental pretest with still uncertain wording and order of the questions, and the second pretest for creating the final form of the instrument. Based on their recommendation there were altogether three think aloud sessions applied for validating the instrument of this study. The aims of the think aloud procedures were the following: (1) to check the appropriateness of wording of questions and instructions; (2) to obtain feedback on the questionnaire as a whole, concerning issues like length, division into sections, general flow, order and type of questions; (3) to check if scales in the questionnaire were used consistently

and if participants' verbal responses corresponded with their choices on the scales; (4) to test if the electronic environment caused any special difficulties in filling out the questionnaire.

These pretests were developmental in the sense that changes were made after each think aloud protocol, and the modified version of the questionnaire was used in the next think aloud. Two participants of the think aloud were members of the target population and the third one was a business professional with EU-related experience. Strictly speaking he was not considered as a potential respondent, but as there was limited access to EU professionals working in EU institutions or EU-related bodies, his insights were hoped to be useful, especially concerning filling out the questionnaire in a rather stressful environment. Overall, the three participants seemed interested in the topic, and were willing to participate. Although the purpose and method of thinking aloud was explained in detail and practised with the participants, they differed in their ability to verbalise their thoughts and explain their choices. Therefore, the protocols varied in length from 23 to 45 minutes.

These final think aloud sessions resulted in only a few changes to the wording of the questions. Two of the participants did not have any difficulty in registering and entering the electronic platform, and filling in the questionnaire electronically. The ESP and translation teacher, however, lacked the experience of using such platforms, and needed more explanation for the registration process. Therefore, it was decided that a username and password would be provided to all potential respondents, in order to make their access to the platform easier, and to save their time.

### **6.3.2.3. Participants**

The population defined as members of the EU discourse community, that is, potential users of English EU documents included Hungarian EU professionals who use English EU documents for their jobs, lecturers of EU studies, teachers of English for the EU, and EU translators and interpreters. The identified discourse community was not a closed and easily

accessible group, and there are no exact statistics about the number of Hungarians working with EU institutions, or in an EU-related field elsewhere, therefore availability and snowball sampling (Goodman, 1961; Heckathorn, 1997; Tompos, 2001) were considered to be the most effective and efficient means to access members of the EU discourse community who have relevant experience with English EU documents. This meant that a link to the electronic version of the questionnaire was sent to Hungarian EU professionals who were on the list of Hungarian EU experts available on the website of the Ministry of Foreign Affairs, and to the mailing list of Hungarians working in EU institutions, lecturers of EU studies at universities, the mailing list of a postgraduate international relations and EU studies programme, former students of an EU translation minor, and ESP teachers who were likely to teach EU English. Although these methods cannot be considered representative, the questionnaire survey yielded sufficient and relevant results, in view of the research questions. Altogether 429 e-mails were sent out, and the survey yielded 99 responses.

The characteristics of the respondents are summarised in Table 24. These characteristics are presented by the number of respondents, and percentages of all the respondents. The majority of respondents, almost 70%, considered themselves EU professionals. Slightly more than half of them work at an EU institution, and almost 80% gained work experience with one of the EU-related bodies. The time they spent in an EU-related job was fairly long. Almost three quarters of all respondents spent at least 3 years in an EU-related job, and more than half of them spent more than 5 years in such a position. As regards English as a working language in EU institutions, 46% said they worked at such an institution for more than a year. As for the special fields within the EU, almost 9% of respondents did not specify one particular field. Most of the respondents worked in special fields like Environment, Regional Policy, EU applications, Internal Market, Culture and Agriculture. In general, all EU subject fields were represented by a few respondents. Overall, the respondents showed relevant experience

regarding work in an EU context, and using English EU documents, and therefore, their responses shed light on the genres that can represent written English EU discourse, and therefore, can be used as the basis for corpus design.

<b>Characteristics</b>	<b>Categories</b>	<b>Number of respondents</b>	<b>Percentage of all respondents</b>
<b>Occupation</b>	EU professional	67	69.8%
	translator	12	12.5%
	ESP teacher	3	3.1%
	EU trainee	3	3.1%
	teacher of EU studies	1	1.0%
	interpreter	1	1.0%
	other	7	7.3%
	missing	2	2.1%
<b>Degree</b>	economist	38	30.9%
	arts	27	22.0%
	law	19	15.4%
	engineering	10	8.1%
	other	29	23.6%
	<b>Workplace</b>	EU institution	50
local authorities, governmental bodies	19	19.8%	
consultancy firm	8	8.3%	
university, research institute	7	7.3%	
EU-related governmental bodies	6	6.3%	
other	5	5.2%	
missing	1	1.0%	
<b>Command of English</b>	advanced	59	61.5%
	near native	29	30.2%
	intermediate	6	6.3%
	beginner	1	1.0%
	missing	1	1.0%
	<b>Knowledge of other languages</b>	intermediate	88
advanced	63	65.0%	
beginner	57	59.0%	
near native	10	10.0%	
<b>EU-related degree</b>	no EU-related degree	45	46.9%
	EU studies	24	25.0%
	postgraduate EU translation course	6	6.3%
	undergraduate EU translation minor or similar higher education programme	1	1.0%
	postgraduate EU interpretation course	1	1.0%
	SCIC exam	1	1.0%
	other	16	16.7%

Table 24. Characteristics of the respondents of the needs analysis survey

Characteristics	Categories	Number of respondents	Percentage of all respondents
<b>Time spent in an EU-related job</b>	more than 5 years	37	38.5%
	3 - 5 years	34	35.4%
	1 - 2 years	8	8.3%
	no time spent in EU-related job	7	7.3%
	1 - 3 months	4	4.2%
	7 - 12 months	4	4.2%
	4 - 6 months	1	1.0%
	less than 1 month	0	0.0%
	missing	1	1.0%
<b>Time spent at an EU institution where the working language was English</b>	I have not worked at such EU institution	37	38.5%
	3 - 5 years	32	33.3%
	1 - 2 years	10	10.4%
	7 - 12 months	4	4.2%
	1 - 3 months	4	4.2%
	4 - 6 months	3	3.1%
	more than 5 years	3	3.1%
<b>Work experience with EU-related bodies</b>	yes	76	79.2%
	no	18	18.8%
<b>EU subject field</b>	Agriculture	9	3.8%
	Audiovisual and Media	4	1.7%
	Budget	8	3.4%
	Competition	7	3.0%
	Consumers	3	1.3%
	Culture	10	4.2%
	Customs	4	1.7%
	Development	3	1.3%
	Economic and Monetary Affairs	8	3.4%
	Education, Training, Youth	8	3.4%
	Employment and Social Affairs	8	3.4%
	Energy	3	1.3%
	Enlargement	7	3.0%
	Enterprise	3	1.3%
	Environment	15	6.3%
	EU applications	14	5.9%
	EU in general	21	8.9%
	External Relations	8	3.4%
	External Trade	2	0.8%
	Fight against fraud	2	0.8%
	Fisheries and Maritime Affairs	1	0.4%
	Food Safety	6	2.5%
	Foreign and Security Policy	5	2.1%
	Human rights	5	2.1%
	Humanitarian aid	2	0.8%
	Information Society	3	1.3%
	Institutional affairs	7	3.0%
	Internal Market	13	5.5%
	Justice, freedom and security	10	4.2%
	other	10	4.2%
	Public Health	3	1.3%
	Regional Policy	14	5.9%
	Research and Innovation	7	3.0%
	Taxation	2	0.8%
Transport	2	0.8%	

Table 24. cont. Characteristics of the respondents of the needs analysis survey

### **6.3.3. Conducting the needs analysis survey**

The questionnaire was administered online in the first half of 2008. The Moodle platform of the Institute of Behavioural Studies and Communication Theory at the Corvinus University of Budapest was used to make the questionnaire accessible via the internet. Only respondents with a valid username and password were allowed to access the questionnaire, and potential respondents received these by an introductory e-mail. The e-mail contained a short explanation of the aims of the survey, and the way to access the online questionnaire. The text of the e-mail in the original language, that is Hungarian, is provided in Appendix 3. As mentioned above, this e-mail was sent to several groups of potential respondents, who were also asked to forward it to colleagues. Altogether 99 questionnaires were received, but three of them were disregarded as they did not contain sufficient answers to relevant questions. Therefore, the final number of responses included in the analysis is 96 questionnaires.

On the whole, the electronic version and online administration has been found very useful and effective, as it would have been a lot more difficult, costly, and time-consuming to reach this diverse and geographically distant population by more traditional means of questionnaire administration.

### **6.3.4. Methods of needs survey data analysis**

Analyses of responses involved quantitative and qualitative methods. The scores of frequency, importance and application of EU genres were analysed statistically. The options for the five-point Likert scale were from 1 to 5 in the electronic format of the questionnaire, as the platform did not provide other possibilities. Therefore, scores had to be given different weight in order to make them as close to nominal data as possible for frequency and importance of EU genres. The conversion tables for the different scores are given in Table 25.

Question type	Original scores	Adjusted scores for statistical analysis
<b>Frequency of use of EU genres</b>		
never	1	0
rarely	2	1
monthly	3	5
weekly	4	20
daily	5	100
<b>Importance of EU genres to respondent's job</b>		
it has nothing to do with my job	1	0
not important	2	1
important	3	5
very important	4	10
indispensable for my job	5	25

Table 25. Conversion table of frequency and importance scores

Answers to the open questions were grouped according to text categories. Genres and specific documents not in the original list of the questionnaire were evaluated based on the number of respondents mentioning or recommending it, and their availability and appropriateness for further study. The results of the needs analysis survey together with a detailed description of the final EEUD Corpus, will be given in Section 7.1.

## 6.4. Procedures and tools of corpus analysis

As outlined and defined in Section 4.1.3, analyses of corpora can apply several types of units of analysis, such as the token, the word type, the lemma, the word family, and the lexical bundle. The ones especially relevant to the present study are the word family, the lemma and the lexical bundle. The following sections will provide detailed discussion of these concepts, and how they were defined for the purposes of the present study. Moreover, these sections will also outline the three individual procedures of the corpus analysis.

### 6.4.1. Step 1: Establishing the EU Word List

The compilation of the EUWL started with the frequency list of the EEUD Corpus computed by *WordSmith Tools 4*. The unit of analysis for this part of the investigation was the

word family as defined in the next section. The individual procedures for the establishment and evaluation of the EUWL can be summarised as follows:

1. Creating the frequency list of EEUD Corpus by *WordSmith Tools 4*;
2. Organising the frequency list into word families, with the help of the lemmatising function of *WordSmith Tools 4*;
3. Selecting word families for the final EUWL on the basis of pre-set criteria and consultation with an ESP practitioner and an EU expert;
4. 513 word families of EUWL identified and stored;
5. validating the final EUWL by the *Range* program.

In what follows the notion of the word family is defined and a detailed description of the procedures of establishing the English EUWL will be given.

#### **6.4.1.1. The notion of the word family**

The most widely used unit of analysis in research into teaching lexis (e.g., Laufer, 1997; Laufer & Nation, 1995; Nation, 2004; Nation & Waring, 1997), defining necessary vocabulary size and text coverage for effortless comprehension of texts (e.g., Hirsch & Nation, 1992; Laufer, 1992; Nation, 1993; Nation, 2006; Ward, 1999), and developing word lists for general and specific language teaching purposes (e.g., Coxhead, 2000; Mudraya, 2006; Nation, 2001; Nation, 2004; Wang et al., 2008), is the word family. According to Bauer and Nation (1993) a **word family** includes a base word, its inflected forms, and transparent derivations. Perceived transparency refers to the assumption behind the idea of a word family that understanding a derived or inflected form of a word does not require extra effort from language learners, if they are familiar with the base word or a derived form, and some knowledge of the word-building processes in English. Transparency also implies that the meaning of the base word and derived forms must be closely related, for example, *hard* and *hardly* are not included in the same word family. The concept is also supported by empirical evidence, as research found that the word



family is a psychological unit in the mental lexicon (Coxhead, 2000; Nation, 2006). Word families are illustrated by an example from the GSL and the AWL in Table 26.

<b>Headword</b>	<b>Members of the word family</b>
<i>ABLE</i> (GSL)	<i>ability, abilities, inability abler, ablest, ably, unable</i>
<i>ANALYSE</i> (AWL)	<i>analysed, analysing analyser, analysers, analyses, analysis, analyst, analysts analytic, analytical, analytically analyze, analyzed, analyzes, analyzing</i>

Table 26. Examples of word families from the GSL and the AWL

For the creation of word families, Bauer and Nation (1993) defined seven levels of inflection and affixation, based on criteria including the frequency, productivity, predictability and regularity of affixes, which can be used for decisions on whether a certain word form can be included into a word family at a given level. These levels, with short descriptions and examples, are illustrated in Table 27. Bauer and Nation (1993) also stated that their levels are arbitrary, and further affixes can be included if they are found frequent or useful in a particular field. Therefore, at level 7 the table contains three additional prefixes that are relevant in EU texts.

<b>Levels</b>	<b>Description</b>	<b>Examples of affixes at this level</b>	<b>Example</b>
Level 1	Each form is a different word	-	<i>develop</i>
Level 2	Inflectional suffixes	plural, 3 <sup>rd</sup> person singular present tense, comparative, possessive	<i>develops developed developing</i>
Level 3	The most frequent and regular derivational affixes	<i>-able, -er, -ish, -ly, -ness, non-, un-</i>	<i>developable undevelopable developer undeveloped development semidevelopment</i>
Level 4	Frequent, orthographically regular affixes	<i>-ation, -ful, -ize, -ment</i>	<i>developmental</i>
Level 5	Regular but infrequent affixes	<i>-age, -atory, -ling, mid-, -ship, pro-, semi-, sub-</i>	<i>redevelopment</i>
Level 6	Frequent but irregular affixes	<i>re-, pre-, -ee, -ive</i>	<i>redevelopment predevelopment agri-development</i>
Level 7	Classical roots and affixes, compounds	<i>Euro-, agri-, ex-</i>	

Table 27. Levels of inflection and affixation (based on Bauer & Nation, 1993, p. 254)

Although the concept of the word family is widely applied in studies of lexis there are some limitations to its application. A major practical difficulty of the concept is the requirement of transparency, that is, which word forms should be recognised as belonging to a particular word family. Biber (2006) reported that he found “it extremely difficult to reliably group the [remaining] words into word families” (p. 242). Another difficulty is to define at which level word families should be interpreted, and if there are any affixes to include in the analysis of the lexis of a particular field. Despite these limitations, the present study applied the concept of the word family to make the results comparable to earlier analyses of ESP lexis. Potential discrepancies between what is included in individual word families were minimised by applying the 14 word family lists created by Nation (2006) on the basis of the BNC (Heatley et al., 2002).

#### **6.4.1.2. Compilation of the English EU Word List**

The corpus analysis programmes *Range* (Heatley et al., 2002) and *WordSmith Tools 4* (Scott, 2004) were used to develop the EUWL. First, a frequency word list was created by *WordSmith Tools 4*, and it was organised by word families using the lemmatiser function of the programme, which joins certain entries according to a pre-prepared list. This list was prepared on the basis of the 14 base word lists of the *Range* programme. For the fine tuning of the word list the *Range* programme was used. The programme counts the frequency of word types in several different files, and records the frequency of occurrence of individual word types in total and in each file. It also counts the number of files in which each word type occurs. Table 28 shows the output of the *Range* programme for a few examples from the EEUD Corpus. The programme can also be used with different word lists of word families, and it can count the cumulative frequency of a word family and provides information on the percentage of tokens, word types, and number of families of a word list in a corpus (Nation, 2001). The *Range*

software was also used to evaluate the final EUWL for text coverage in texts representing different registers and genres.

<b>Word type</b>	<b>Range</b>	<b>Total frequency</b>	<b>Frequency in File 1</b>	<b>Frequency in File 2</b>	<b>Frequency in File 3</b>
<i>Community</i>	34	3222	80	69	185
<i>framework</i>	34	1010	19	1	50
<i>implement</i>	32	227	12	2	4
<i>OJ</i>	32	874	25	28	55
<i>undertakings</i>	24	254	7	0	2

Table 28. Sample output of range and frequency by the *Range* programme

The three selection criteria used by Coxhead (2000) were adopted for the compilation of the final EUWL in this study, with some modifications. Firstly, **specialised occurrence** was ensured by eliminating the word families of the GSL from among the word types of the frequency list of the EEUD Corpus. Secondly, only word families used in a wide range of EU subject fields were selected. In her study Coxhead used a two-level criterion for defining **range** for the word selection. As her Academic Corpus was divided into sub-corpora of disciplines, and these were further divided into subject areas, she set a range for both levels for a member of a word family. On the other hand, Wang et al. (2008) set the criterion for range at 50%, that is, word families with a cumulative range of 16 or more of the total 32 sub-fields of medicine, were included in the final MAWL. As the aim of the EUWL was to provide a list of useful words for students of EU studies with an intermediate level of general English, the criteria for word selection were set slightly broader than in previous studies. Word families had to occur in 16 or more of the 34 EU subject fields, which correspond to a range of 47%. Thirdly, this study, as that of Wang et al. (2008), started out from the **cumulative frequency** criterion set by Coxhead (2000) at 100 in her 3.5-million-word corpus. Wang et al., however, argued that because their corpus of medical research articles had one million running words, which is approximately a third of that of Coxhead's corpus, they set the frequency criterion at 30 for inclusion into their word list, assuming a linear relationship between the number of running

words, and the number of word types in a corpus. Biber (2006), however, based on experiments of the stability of the distribution of lexis, found that the relationship between corpus size and the number of word types is not linear. According to his findings, half a corpus represents around 70% of the word types in the larger corpus. The simple formula suggested for adjusting the number of word types in corpora of different sizes is:

“[number] of word types of Corpus 1 = [number] of word types of Corpus 2 / square root of corpus size” (p. 256).

Thus the reformulated formula to calculate the number of word types of Corpus 2 is as follows:

number of word types in Corpus 2 = number of word types in Corpus 1 × Square root of relative corpus size of Corpus 2 to Corpus 1

Consequently, a corpus (Corpus 2) of half the size of another corpus (Corpus 1) has a number of word types of 0.707 times the number of word types of the full corpus (Corpus 1) as the square root of 0.5 is 0.707. Applying the formula to the required cumulative frequency of 100 applied by Coxhead to her three times bigger corpus than the corpus used in the present study, resulted in the adjusted cumulative frequency of 57 for inclusion in the EUWL as the square root of one-third is 0.57.

As a consequence, the final selection criteria were formulated as follows:

1. **Specialised occurrence:** The word families included in the final EUWL had to be outside the GSL representing the first 2,000 most frequent English words.
2. **Range:** A member of a word family had to occur in 16 or more of the 34 EU subject fields.
3. **Frequency:** The cumulative frequency of occurrence of a word family had to be higher than 57 in the EEUD Corpus.

In addition to these criteria, two experts were also consulted on finalising the EUWL. One of them was an EU expert, and the other was an experienced teacher of ESP. Involving experts was found to be necessary, as earlier studies on lexis in ESP also report on the difficulties of identifying subject-specific technical and semi-technical lexis (Chung & Nation, 2003, 2004; Mudraya, 2006), and emphasise the advantages of consulting ESP experts in the final stage of developing a word list (Bowker & Pearson, 2002; Wang et al., 2008).

In the selection process, range was considered secondary to frequency, because all the texts in the corpus were issued by EU institutions, and as such by definition, all represented an EU related subject field. Therefore, following consultation with the two experts, 28 word families were added to the final list, which met the first and third criteria fully. These were selected despite the fact that their range was less than 16, because their cumulative frequency was high, and were considered necessary for language learners for an EU context by the experts. Thus the word families that can be found in the least of the EU related subject fields are *ICT*, *INTEROPERABILITY*, *DEMOCRACY* and *STATUTORY*, with a range of 12 and a cumulative frequency of 166, 75, 103 and 272 respectively.

Evaluation of the final EUWL was carried out with the help of the *Range* programme (Heatley et al., 2002), by testing the text coverage of the list in several registers and genres of different sources. These sources included another corpus of English EU texts (Trebits, 2008), randomly selected pieces of EU legislation and EU press releases, randomly selected news texts with business, UK news, world news and European news topics, news releases of the UK government, two randomly selected pieces of UK legislation, and British and American literary texts. All the 20<sup>th</sup> century literary texts were downloaded from the Project Gutenberg's collection of texts ([http://www.gutenberg.org/wiki/Main\\_Page](http://www.gutenberg.org/wiki/Main_Page)), and the extract from Dickens's *Tale of Two Cities* came as a trial text with the *WordSmith Tools 4* (Scott, 2004), and the other texts were downloaded from the Internet.

#### **6.4.2. Step 2: Collocational analysis of selected lexical items**

The aim of the **collocational analysis** in the present study was twofold. Firstly, it was applied in order to detect specific lexical patterns characterising written English EU discourse. Secondly, it aimed to provide language learners with information about the different uses and meanings of certain lexical items. The present study focused on fifteen lexical items selected based on their frequency and range in the EEUD Corpus. Although most studies, especially in ESP, focus on nouns, in order to gain insights into the behaviour of lexical items of different word classes, the current study investigated nouns, verbs and adjectives as well. This section will outline the selection of the fifteen lexical items for collocational analysis, and the procedure of the actual analysis of the collocations of the selected items.

##### **6.4.2.1. Selection of lexical items for collocational analysis**

The method several corpus studies applied to identify typical lexis of a subject area is the **key word analysis** (Chung & Nation, 2004; Kennedy, 1998; Nelson, 2000; Tribble, 2000). Key word analysis in this study was carried out as a step in selecting lexical items for further analysis. The individual procedures resulting in the final set of fifteen lexical items for further analysis can be summarised as follows:

1. Identifying key words in the EEUD Corpus by the Key word function of *WordSmith Tools 4* with the BNC World as reference corpus
2. Manual categorisation of key words into semantic sets
3. Final list of fifteen lexical items based on the following considerations: (1) they are among the key words and the elements of the EUWL, or they were recommended by EU professionals in the initial interviews, (2) they represent different semantic sets identified among the key words, (3) there is a sufficient number of occurrences, both in the EEUD Corpus, and the BNC Written, (4) they include nouns, verbs and adjectives as well.

The key words of the EEUD Corpus were identified with the help of the Key word function of the *WordSmith Tools 4*. The method for identifying key words by *WordSmith Tools 4* is based on the following principles: (1) a word type with high frequency in a text or corpus is likely to be key in it; (2) high frequency alone does not make a word key, its frequency has to be high in relation to the frequency of the same word type in a reference corpus; (3) the frequency of the word type has to reach a threshold level, usually 2 or 3 occurrences (Scott & Tribble, 2006). The reference corpus used in the present study was the BNC World in the form of its frequency lists available at the website of the *WordSmith Tools 4* (Scott, 2004). The threshold level was set at 3 occurrences. The tool can also calculate whether the difference between the frequencies in the analysed corpus and the reference corpus are statistically significant by the log likelihood statistic. The key word analysis can provide positive key words, that are significantly more frequent in the analysed corpus, and negative key words that are significantly less frequent in the analysed corpus than in the reference corpus. The present analysis focused on positive key words, as its aim has been to identify lexical items that can be associated with written English EU discourse. Altogether the analysis resulted in 2515 positive key words and 1265 negative key words at the  $p=0.000001$  level. Further analysis concentrated on the first 300 positive key words.

The semantic analysis of the first 300 positive key words resulted in five semantic sets. In addition, a group of function words and a group of general words in an EU context were identified. The semantic sets include (1) words in connection with applications for EU funding; (2) institutions, members and countries within the European Communities; (3) words that express aspects of integration; (4) a further semantic set represents words describing different subject fields where the EU introduced common policies; (5) and one set comprising words representing different types of legal documents. Table 29 and Table 30 give examples of the

key words ranked according to their keyness in the EEUD Corpus, grouped into the relevant semantic sets. The ones included in the final list are in bold.

<b>EU funding</b>	<b>Communities</b>	<b>Integration</b>	<b>EU policies</b>	<b>Legal documents</b>
<i>application</i>	<b>European</b>	<i>accordance</i>	<i>customs</i>	<i>article</i>
<i>audit</i>	<i>Member</i>	<b>cooperation</b>	<i>SMEs</i>	<b>regulation</b>
<i>requirements</i>	<b>Commission</b>	<i>conformity</i>	<i>audiovisual</i>	<i>directive</i>
<i>eligible</i>	<i>states</i>	<i>coordination</i>	<i>energy</i>	<i>measures</i>
<b>projects</b>	<i>EU</i>	<i>common</i>	<i>protection</i>	<i>annex</i>
<i>monitoring</i>	<i>union</i>	<i>budget</i>	<b>policy</b>	<i>treaty</i>

Table 29. The first six key words in the specific semantic sets

<b>General words</b>	<b>Function words</b>
<i>referred</i>	<i>shall</i>
<i>implementation</i>	<i>the</i>
<i>programme</i>	<i>of</i>
<i>financial</i>	<i>or</i>
<i>framework</i>	<i>under</i>
<b>objectives</b>	<i>within</i>

Table 30. The first six key words in the categories of general words and function words

Another source of relevant lexical items for further analysis was recommendations by EU professionals interviewed in the early stages of the needs analysis survey. EU professionals were asked to recommend lexical items that might cause difficulties for students, and that had a different meaning in EU texts from their meanings in general English texts. The list contained the following words: *commitment, payment, provision, pillar, initiative, neighbourhood, objective, consistency, action plan, entrepreneurship, criteria*.

As can be seen in Tables 29 and 30 above, and in the recommended list, most lexical items are nouns. Therefore it was decided that the final list should include more nouns, but that it should not be restricted to nouns. The adjective *European*, ranking third overall, and being the first among words in the ‘Communities’ semantic set, was also selected. However, it was more difficult to find suitable verbs for the analysis. These were selected from further down the key word list. There were two more important aspects of the selection. First, words in the final list had to be among the words in the EUWL, and second, there had to be a sufficient number of



occurrences both in the EEUD Corpus and the BNC Written, in order to allow for collocational analysis. Thus the final list of lexical items for collocational analysis contains nine nouns: *co-operation/ cooperation, policy, commission, criterion, regulation, project, initiative, commitment, objective*; one adjective: *European*; and five verbs: *lay, notify, function, ensure, implement*.

#### **6.4.2.2. Analysis of collocational patterns in the EEUD Corpus**

The corpus analysis tool *Sketch engine* (Kilgarriff & Tugwell, 2001) was used for the collocational analysis of the selected fifteen lexical items. The tool provides collocations of lemmas based on their saliency (see Section 4.3.1.3), and organised into groups according to the grammatical relations they form with the node word. The collocations are provided in the form of **word sketches** as illustrated in Figure 13 by the word sketch of the verb NOTIFY.

The collocational analysis compared the collocational patterns of the selected lemmas in the EEUD Corpus and in the BNC Written. A POS tagged version of the BNC is available in the *Sketch engine*. It is also possible to upload one's own corpus into the *Sketch engine* which tags the uploaded corpus automatically by the *Tree tagger* software (Schmid, 1994). Therefore, first the EEUD Corpus was uploaded into the system, and next, the word sketches for the fifteen lemmas were created in both the EEUD Corpus and the BNC Written. The procedures of the collocation analysis can be summarised as follows:

1. The EEUD Corpus was uploaded into the *Sketch engine*. It was POS tagged by the *Tree tagger* automatic tagging software.
2. Word sketches for the fifteen lemmas in both the EEUD Corpus and the BNC Written were created.
3. Types and number of grammatical relations and collocates in the EEUD Corpus and the written BNC were compared for each lemma.

4. The collocates of five lemmas were compared by the semantic prosodies and semantic preferences they exhibit in the EEUD Corpus and the BNC Written.

**notify EEUD freq = 486**

<b>and/or</b>	<u>2</u>	<b>0.0</b>	<b>pp_to-i</b>	<u>21</u>	<b>3.1</b>	<b>subject</b>	<u>11</u>	<b>0.6</b>
publish	<u>2</u>	9.03	addressee	<u>2</u>	10.51	government	<u>2</u>	6.87
			manufacturer	<u>4</u>	8.66	party	<u>2</u>	5.69
<b>pp_under-i</b>	<u>9</u>	<b>15.0</b>	parliament	<u>2</u>	8.55	authority	<u>3</u>	5.2
number	<u>8</u>	8.15	party	<u>3</u>	6.26	body	<u>2</u>	4.89
<b>np_adj_comp</b>	<u>5</u>	<b>8.8</b>	<b>pro_subject</b>	<u>8</u>	<b>2.4</b>	<b>pp_of-i</b>	<u>6</u>	<b>0.3</b>
concerned	<u>5</u>	10.54	he	<u>2</u>	8.64	decision	<u>3</u>	5.52
			they	<u>2</u>	5.89	application	<u>2</u>	5.35
<b>object</b>	<u>329</u>	<b>8.0</b>	it	<u>4</u>	5.4			
body	<u>218</u>	11.46						
intention	<u>8</u>	9.31	<b>modifier</b>	<u>24</u>	<b>1.7</b>			
authority	<u>42</u>	8.85	immediately	<u>10</u>	11.33			
council	<u>7</u>	8.58	forthwith	<u>3</u>	11.3			
manufacturer	<u>6</u>	8.28	previously	<u>2</u>	9.7			
applicant	<u>7</u>	8.09				<b>passive</b>	<u>11</u>	<b>1.7</b>
access	<u>6</u>	7.54					<u>11</u>	5.28
measure	<u>8</u>	5.87						
decision	<u>2</u>	4.74						
			<b>pp_in-i</b>	<u>10</u>	<b>1.2</b>			
<b>pro_object</b>	<u>4</u>	<b>4.7</b>	advance	<u>4</u>	10.83			
him	<u>2</u>	10.54	accordance	<u>3</u>	6.67			
it	<u>2</u>	4.4						

Figure 13. Sample output of the *Sketch engine* of the collocates of the lemma NOTIFY

### 6.4.3. Step 3: Analysis of lexical bundles in the EEUD Corpus

A corpus-driven approach was applied in the analysis of the MWIs in the EEUD Corpus. Therefore, the unit of analysis was the lexical bundle, as defined by Biber and Conrad (1999). According to their definition, **lexical bundles** are “sequences of three or more words that show

a statistical tendency to co-occur” (p.183). Although Scott and Tribble (2006) argued that three and four-word lexical items are both good discriminators of registers and three-word lexical items have advantages especially for pedagogic purposes, the present study focused on four-word lexical bundles for two reasons. Firstly, most studies on lexical bundles analyse bundles with four words and therefore lexical bundles in written English EU discourse can be compared to bundles in other registers like university or academic registers (Biber et al., 2004; Biber & Barbieri, 2007; Cortes, 2004; Hyland, 2008). Secondly, as Cortes (2004) also argued, three-word lexical bundles are often part of four-word bundles, and four-word bundles are more frequent and give more variety for the structural and functional analysis, than five-word bundles. The individual procedures resulting in the final set of lexical bundles in the EEUD Corpus can be summarised as follows:

1. Four-word lexical items were automatically identified by the Cluster function of *WordSmith Tools 4*.
2. Lexical bundles were selected based on preset criteria: (1) a minimum frequency of 47 in the corpus and (2) occurrence in at least 10% of the 241 EU texts of the EEUD Corpus.
3. Qualitative analysis of the length, structural and functional characteristics of lexical bundles in the EEUD Corpus was carried out.
4. Characteristics of lexical bundles in the EEUD Corpus were compared to lexical bundles in other registers.

In the investigation, lexical bundles were selected from the automatically identified four-word lexical items. In order to be considered lexical bundles, the four-word lexical items had to recur at least 40 times per million (Biber & Barbieri, 2007). In order to avoid the impact of idiosyncratic use, lexical bundles were defined, in addition to their overall frequency in the

corpora, on their distribution in individual texts. Therefore, the following requirement was introduced: only recurring four-word lexical items occurring in at least 10% of the texts, that is, in 24 different EU texts (Biber & Barbieri, 2007; Cortes, 2006; Hyland, 2008) can be included in the analysis.

The qualitative analysis of the identified lexical bundles in the EEUD Corpus included their structural and functional analysis. The structural analysis applied the structural types of lexical bundles identified by Biber et al. (2004). According to this classification, there are three main structural types which include (1) lexical bundles that incorporate verb phrase fragments like *that's one of the, is based on the*, (2) lexical bundles that incorporate dependent clause fragments like *that this is a, to come up with*, (3) lexical bundles that incorporate noun phrase and prepositional phrase fragments like *at the end of, at the same time*. Each main structural type entails several sub-types as illustrated in Table 31.

<b>Structural types</b>	<b>Sub-types</b>	<b>Sample bundles</b>
<b>1. Lexical bundles that incorporate verb phrase fragments</b>	1.a 1st/2nd person pronoun + VP fragment	<i>I'm not going to</i>
	1.b 3rd person pronoun + VP fragment	<i>and this is a</i>
	1.c discourse marker + VP fragment	<i>I mean I don't</i>
	1.d Verb phrase (with non-passive verb)	<i>have a lot of</i>
	1.e Verb phrase (with passive verb)	<i>is based on the</i>
	1.f yes-no question fragments	<i>are you going to</i>
	1.g WH-question fragments	<i>what do you think</i>
<b>2. Lexical bundles that incorporate dependent clause fragments</b>	2.a 1st/2nd person pronoun + dependent clause fragment	<i>I want you to</i>
	2.b WH-clause fragments	<i>when we get to</i>
	2.c if-clause fragments	<i>if we look at</i>
	2.d to-clause fragment	<i>to be able to</i>
	2.e that-clause fragment	<i>that this is a</i>
<b>3. Lexical bundles that incorporate noun phrase and prepositional phrase fragments</b>	3.a Noun phrase with of-phrase fragment	<i>one of the things</i>
	3.b Noun phrase with other post-modifier fragment	<i>the way in which</i>
	3.c Other noun phrase expressions	<i>a little bit more</i>
	3.d Prepositional phrase expressions	<i>at the end of</i>
	3.e Comparative expressions	<i>as well as the</i>

Table 31. Structural types of lexical bundles (Biber et al., 2004, p. 381.)

In the functional analysis, the taxonomy of discourse functions of lexical bundles outlined by Biber et al. (2004) was applied. According to this taxonomy, lexical bundles serve three

main discourse functions in registers: (1) stance bundles express attitude or assessment like *you have to do, are likely to be, it is important that* (2) discourse organisers reflect the relationships between different parts of texts like *on the other hand, for the most part, in addition to the* (3) referential expressions refer to physical or abstract entities, or to other textual parts like *one of the most, a great deal of, beyond the scope of*. Each of these main categories has several sub-categories which are associated with more specific discourse functions. The main discourse function categories with their sub-categories, as developed by Biber et al. (2004), are shown in Table 32.

Categories	Sub-categories	Sample bundles
<b>I. Stance bundles</b>	<b>A. Epistemic stance</b>	<i>the fact that the, and I think that</i>
	<b>B. Attitudinal/ modality stance</b>	
	B1) Desire	<i>what do you want</i>
	B2) Obligation/ directive	<i>you don't have to, will be required to</i>
	B3) Intention/ Prediction	<i>it's going to be</i>
	B4) Ability	<i>it is possible to</i>
<b>II. Discourse organisers</b>	B5) Importance	<i>of the most important</i>
	<b>A. Topic introduction</b>	<i>in this chapter we</i>
	<b>B. Topic elaboration/ clarification</b>	<i>on the other hand</i>
	<b>C. Identification/ focus</b>	<i>one of the things, and this is a</i>
<b>III. Referential bundles</b>	<b>D. Conditions</b>	<i>if you do not, if you wish to</i>
	<b>A. Identification/ focus</b>	<i>is one of the, is referred to as</i>
	<b>B. Imprecision</b>	<i>or something like that</i>
	<b>C. Specification of attributes</b>	
	C1) Quantity specification	<i>a lot of people</i>
	C2) Tangible framing	<i>in the form of</i>
	C3) Intangible framing	<i>on the basis of</i>
	<b>D. Time/ Place/ Text reference</b>	
	D1) Place reference	<i>in the United States</i>
	D2) Time reference	<i>at the same time</i>
D3) General location reference or framing	<i>at the base of, at the bottom of</i>	
D3) Text-deixis	<i>as shown in figure</i>	
D4) Multi-functional reference	<i>in the middle of</i>	

Table 32. Discourse functions of lexical bundles  
(based on Biber et al., 2004, pp. 386-388; Biber, 2006, pp. 151-168)

For the analysis of the discourse functions of the lexical bundles in the EEUD Corpus, the Concordance function of the *WordSmith Tools 4* was used. The software provided the context by concordancing, and the functions the lexical bundles in question performed, were analysed manually. The final structural and functional analysis applied a few additional categories that will be described in Section 7.2.3, together with the results of the analysis. In order to ensure that the additional categories can be applied in a reliable way, and that the classification of lexical bundle types was carried out in a consistent manner, inter-rater reliability and intra-rater reliability of this qualitative analysis has been tested on a set of thirty lexical bundles. The thirty lexical bundles were selected randomly, and they were categorised again by an independent researcher and by the author within a period of six months after the first analysis. Rating for the inter-rater reliability test was preceded by a training session, where all categories were explained and examples were shown from earlier analyses. This also involved a session where ten lexical bundles were categorised by the researcher and the independent ESP expert together with detailed explanation. Finally, the Kappa statistic was performed to determine consistency among raters and ratings (Cohen, 1960; Sajtos & Mitrev, 2009). Statisticians proposed the following categories to interpret Kappa values: (1) values below 0.40 suggest fair to poor agreement, (2) values between 0.41 and 0.61 represent moderate agreement, (3) values between 0.61 and 0.80 represented substantial agreement, and (4) values of 0.80 and above represented excellent agreement (Landis & Koch, 1977; Sajtos & Mitrev, 2009).

The additional corpora used in the comparison of lexical bundles in the EEUD Corpus to lexical bundles of other registers, were the written genres of the *BNC Sampler* and the three written sub-corpora of the *BNC Baby*, namely, academic, fiction and news. Although each corpus contains a different number of texts, all corpora comprise around one million running words, which makes the comparison of their discourses more accurate. In order to provide the analysis with a comparable sample of lexical bundles, the frequency criterion to select lexical

bundles had to be lowered. Therefore, a four-word sequence was considered a lexical bundle for the comparison, if it occurred more than 20 times in the corpus, and in at least 10% of the texts, but in not less than five texts.

## Chapter 7: Results and discussion

In what follows, first, the results of the needs analysis survey will be presented, and the final composition and structure of the EEUD Corpus will be described in detail. Then, in the subsequent sections of the chapter the results of the corpus analysis will be presented and discussed.

### 7.1. Results of the needs analysis

The results of the needs analysis questionnaire survey were analysed by the *SPSS* statistical software, and the answers provided to the open questions were explored in a qualitative manner. It is important to note that in the light of the results of the needs analysis survey, the preliminary criteria for text selection for the corpus had to be slightly modified. The modification included the extension of the date of issue of the texts for two main reasons. Firstly, because respondents mentioned several documents issued in the 1980s as important or used for their daily work. Secondly, in order to have enough texts in the corpus representing all the subject fields of the EU, some texts issued in the 1980s and 90s also had to be included in the corpus. Thus the final text selection criteria for the corpus were formulated as follows:

- the text should represent one of the selected genres and
- the text should be issued after 1980.

The selection of individual texts was preceded by defining the genres for the corpus. The aim was to select genres which are regularly used by the members of the EU discourse community in their daily professional routine. To “measure” the perceived relevance of individual genres in the target situation, the scores of frequency and importance in the survey results were used. A joint frequency and importance score was computed by multiplying the two scores for each genre. The joint score indicates the perceived relevance of a particular genre



in the target situation. The means of the joint frequency and importance scores (MJ score) for all 48 different genres listed in the questionnaire can be seen in Table 33. The 27 genres that were included in the final EEUD Corpus are shaded in the Table. In addition to the 48 genres in the questionnaire, the open questions after each of the five text categories inquired about other genres that were not mentioned in the questionnaire. The other types of open questions asked about specific EU documents that were either used by the EU professionals in the near past or ones that they would recommend for students studying to become EU professionals.

<b>EU genre</b>	<b>MJ score<sup>6</sup></b>	<b>SDJS</b>
Regulation	1095.73	1184.72
Directive	1019.76	1183.06
Decision	830.91	1104.85
Commission working document	599.28	976.85
Commission proposal	561.35	970.47
Rules of procedure	314.64	757.73
Recommendation	282.54	701.05
Communication from the Commission	279.94	704.51
Opinion	276.20	745.24
Council common position	266.75	718.20
Press release	254.17	668.33
Call for proposals	239.87	661.92
Green paper	232.78	632.68
White paper	224.34	633.30
Resolution	217.27	663.43
Common position CFSP	213.78	665.28
EP legislative resolution	209.44	638.59
Application form	205.56	618.89
Project contract	200.17	621.64
Treaty (Consolidated Versions of the Treaty on European Union and of the Treaty establishing the European Community)	198.09	562.03
Judgment of the Court of Justice	153.42	519.81
Presidency conclusions	148.32	514.85
EP Initiative	142.50	527.34
Declaration	142.31	515.91
Project fiche	102.06	395.95
Community guidelines	98.29	388.05
Common strategy	98.28	449.73
Treaty (Treaty establishing the European Community)	79.08	367.96

Table 33. EU genres of the survey in order of their perceived relevance in the EU context

<sup>6</sup> The abbreviations in Table 33 stand for the following: MJ score - mean frequency and importance joint score, SDJS - standard deviation of joint score

Advocate General's Opinion	78.95	378.41
Joint action	70.29	369.20
Treaty (Treaty on European Union)	66.12	362.48
Commission Annual report	62.35	290.65
Eurostat news release	56.67	286.26
PRAG (Practical Guide to Contract Procedures)	53.69	276.21
Treaty (Accession of the Czech Republic, Estonia, Cyprus, Latvia, Lithuania, Hungary, Malta, Poland, Slovenia and Slovakia)	53.07	267.45
Opinion of the Committee of the Regions	46.39	274.75
General Report on the Activities of the European Union	40.96	268.38
Treaty (Treaty establishing a Constitution for Europe)	40.24	259.68
Regular report	40.16	267.92
Treaty (Treaty of Amsterdam)	36.92	260.47
Budgetary resolution	36.55	269.06
Treaty (Single European Act)	33.31	259.80
Treaty (Treaty of Nice)	33.19	259.78
Eurobarometer First Results	17.54	63.28
Report of the Court of Auditors	16.30	62.93
Opinion of the European Economic and Social Committee	15.26	63.72
Fact sheet on the EU	13.54	39.63
Recommendation of the European Central Bank	8.35	57.13

Table 33. cont. EU genres of the survey in order of their perceived relevance in the EU context

Additional genres and specific documents listed by respondents include *Council Minutes*, *Report from the Commission*, *Guideline for Applicants*, *Financial Regulation*, *EP Amendments*, *Reports of the Court of Auditors*, *EEA Agreement*, *Agreement between Norway and the EU on the Norwegian Financial Mechanism*, *EP Draft Reports*. These additional genres are given in italics in Table 35, constituting the final structure of the EEUD Corpus.

Most respondents recommended the *Lisbon Treaty* and the *Constitution* as documents that are particularly useful for students, and *Presidency conclusions* as an important genre. Many suggested that students should be familiarised with the founding treaties and general materials available on the EU official website, as documents of a special EU subject field are too specific, and require a great deal of specific technical background knowledge to be fully understood.

On the basis of these results the final selection criteria of genres to be included in the corpus can be summarised as follows:

- the genre is to be used on a regular basis by respondents in their daily work: i.e. the MJ score of this genre is to be higher than 95  
or
- the genre is to be recommended or sent by the respondents.

As many as 27 genres were selected based on the MJ score, and a further 13 genres were added on the basis of the recommendations of the respondents. The proportion of the different genres in the corpus was defined based on the MJ scores, that is, genres with higher MJ scores are represented by a higher number of tokens in the corpus. Therefore, genres of the text category secondary legislation, such as Regulations, Directives and Decisions, recommended by the majority of respondents account for a larger part of the corpus. The final composition of the EEUD Corpus can be seen in Table 35.

#### **7.1.1. How EU documents are used by professionals**

The questionnaire also explored what EU documents most often were used for in the EU context by professionals. Respondents were asked to tick the option or options that they found relevant, and more options could be chosen for the same category of text. The cumulative results for all genres in all five text categories are given in Table 34. As can be seen, most respondents scan English EU documents for specific information, and somewhat fewer of them skim EU documents to obtain general background information on a particular topic. Almost one third uses EU documents to find specific EU terms, and a little more than a fifth of the respondents uses EU texts as templates for writing their own texts. Around 13% of EU professionals use EU documents for collecting EU terminology, or for translation. Summarising EU documents does not seem to be a task that many EU professionals have to carry out, as only 5%, and slightly less than 5%, mentioned summarising documents in Hungarian and in English respectively.

<b>Purpose of using the text</b>	<b>Mean number of respondents (total N=96)</b>	<b>% of respondents</b>
scanning for specific information	51.2	53.3%
skimming for general information	38.6	40.2%
finding specific EU terms	26.0	27.1%
as a template for writing	21.4	22.3%
collecting EU terminology	13.0	13.5%
translating the text	10.2	10.6%
other	6.2	6.5%
summing it up in Hungarian	4.8	5.0%
summing it up in English	4.2	4.4%

Table 34. Purpose of using EU texts in general

More than half of the respondents, that is 53 out of 96, specified additional purposes for using EU documents. Several respondents mentioned that they used the texts in their daily for legal applications, writing opinions, planning, drafting national legislation, or preparing the government's position. Another additional application is using EU documents as reference in project proposals, in legal texts, or quoting them in translations. Some respondents were involved in drafting EU documents, others use these documents to inform the general public, or help them with understanding EU documents. EU documents are also used for writing job applications for one of the EU institutions. The information on the way EU professionals make use of EU documents has significant pedagogical implications: it can help create authentic types of tasks for the ESP classroom.

Overall, the survey operated very well as an instrument to gain insights into the target situation of learners of English for the EU, by collecting relevant EU texts for the corpus, by providing measures for the perceived relevance of certain EU genres for EU professionals' daily work, and by offering information about the ways EU documents are used in EU institutions. The majority of respondents sent at least one document, but many of them sent even more. The documents were either sent as a pdf or Word file in e-mail attachments, or as a link to the document on the internet. Although the questionnaire survey resulted in the gathering

of several EU texts, some of the respondents were not able to send samples, as the documents they used were confidential.

Text category	Genre	Length (in number of words)	Number of texts	% of tokens in the corpus
Primary legislation	Treaty	119,673	1	10.19
	<i>International agreement</i>	2,044	2	0.17
Secondary legislation	Regulation	114,015	14	9.71
	Directive	102,251	12	8.7
	Decision	86,622	23	7.37
	Opinion	33,780	8	2.88
	Common position CFSP	20,210	8	1.72
	Recommendation	18,852	11	1.6
EU case-law	Judgement of the Court of Justice	24,107	2	2.05
Preparatory documents	Commission legislative proposal	61,087	6	5.2
	Communication from the Commission	41,578	6	3.54
	Council common position	27,599	5	2.35
	Green paper	23,675	3	2.02
	White paper	22,477	2	1.91
	EP legislative resolution	20,431	12	1.74
	EP initiative	17,672	5	1.5
	<i>ECOSOC Opinion</i>	1,295	1	0.11
	<i>EP Draft Report</i>	1,524	1	0.13
	<i>EP Position</i>	556	1	0.05
Documents related to EU funds	Call for proposals	26,303	9	2.24
	Project contract	26,008	3	2.21
	Project fiche	10,470	2	0.89
	Application form	9,884	3	0.84
	<i>Guide for applicant</i>	41,405	5	3.52
	<i>Ex ante guide</i>	3,571	1	0.3
	<i>Grant agreement</i>	503	1	0.04
Other documents issued by EU institutions	Commission Working Document	64,532	6	5.49
	Rule of procedures	29,835	3	2.54
	Press release	26,580	30	2.26
	Resolution	24,113	15	2.05
	Declaration	19,919	10	1.7
	Presidency conclusions	13,348	2	1.14
	Community guidelines	11,767	2	1
	Common strategy	9,401	3	0.8
	<i>Report</i>	57,578	5	4.90
	<i>Presidency Note</i>	24,215	1	2.06
	<i>Council minutes and addenda to minutes</i>	15,134	14	1.29
	<i>Operation manual</i>	8,894	1	0.76
	<i>Press conference</i>	7,259	1	0.62
	<i>Commission Notice</i>	4,586	1	0.39
Total		1,174,753	241	100.00

Table 35. Final composition of the EEUD Corpus

### **7.1.2. Description of the EEUD Corpus**

In this section, a detailed description of the EEUD Corpus will be given. Aspects of balance for EU subject fields, proportion of sent and randomly selected texts, EU institutions that issued the texts, and the time period the texts were issued in, will be discussed in order to provide the background against which the final results of the corpus analysis can be interpreted. All the percentages given in the description refer to proportions of tokens, that is, the number of running words, in the full corpus. Finally, issues concerning copyright will be outlined.

#### **7.1.2.1. Sources of texts in the EEUD Corpus**

The texts that were included in the final version of the Corpus are all available on the internet, either on the EURlex, the legal portal of the European Union, or on the official websites of the individual institutions, or other EU-related bodies. Most of the texts were sent, recommended, or mentioned as used in the work of the respondents of the needs analysis survey. This means that either the given document they referred to was sent or recommended, or the title or a link to the document was sent by the respondents. As shown in Figure 14, these texts make up about two-thirds of the tokens in the corpus. The final one-third of the corpus is made up of texts that were selected randomly by the current researcher as examples of the selected EU genres with a MJ score higher than 95. In other cases only the genre was mentioned by the respondents, and thus the texts had to be randomly selected from official EU websites as examples of the recommended genres (e.g., *Council minutes and addenda to minutes*, *Guide for applicants*).

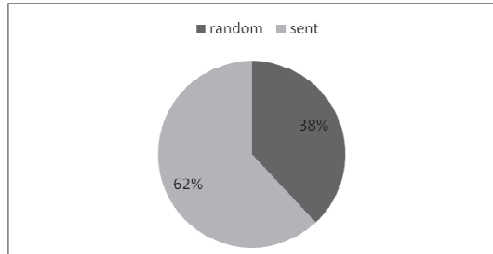


Figure 14. Proportion of tokens in texts sent by respondents and selected randomly by researcher

#### 7.1.2.2. Balance for EU subject fields

Although the whole corpus could have been compiled by using only texts that were sent by respondents, in order to ensure a balanced representation of all main EU subject fields in the Corpus some texts from the respondents had to be left out, and others had to be included from among the EU subject fields that were underrepresented by the already chosen texts. As can be seen in Figure 15 and Table 36, the Corpus is still biased to a certain extent to EU subject fields the respondents work in, for example, *Regional Policy*, *Agriculture*, and to topical EU issues, for example, *Competition*, *Institutional affairs*, *Audiovisual and Media*, *Foreign and Security Policy*. The subject field labelled *EU in general* contains texts that are not related to any of the special fields, or deal with all special areas within the EU, like treaties, for example. The final proportion of EU subject fields was considered acceptable for the purposes of further research as the bias this proportion displayed reflected the bias of the Hungarian EU professionals answering the questionnaire, and therefore represented the bias characteristic of the target situation of a particular period.

<b>EU subject field</b>	<b>Percentage of tokens in the corpus</b>	<b>Reference number in Figure 15</b>
EU in general	17.86%	1
Regional Policy	5.52%	2
Agriculture	4.09%	3
Competition	3.96%	4
Institutional affairs	3.72%	5
Audiovisual and Media	3.67%	6
Foreign and Security Policy	3.67%	7
Enterprise	3.05%	8
Economic and Monetary Affairs	3.02%	9
Education, Training, Youth	3.01%	10
Food Safety	2.79%	11
Justice, freedom and security	2.69%	12
Consumers	2.48%	13
Employment and Social Affairs	2.48%	14
Enlargement	2.45%	15
Research and Innovation	2.34%	16
Culture	2.27%	17
Fight against fraud	2.23%	18
Human rights	2.23%	19
EU applications	2.20%	20
Taxation	2.15%	21
Energy	2.05%	22
External Trade	2.01%	23
Information Society	1.99%	24
Customs	1.96%	25
Environment	1.94%	26
Internal Market	1.90%	27
Development	1.79%	28
Transport	1.72%	29
External Relations	1.64%	30
Fisheries and Maritime Affairs	1.54%	31
Budget	1.49%	32
Public Health	1.20%	33
Humanitarian aid	0.89%	34

Table 36. Proportion of tokens of EU subject fields in the EEUD Corpus



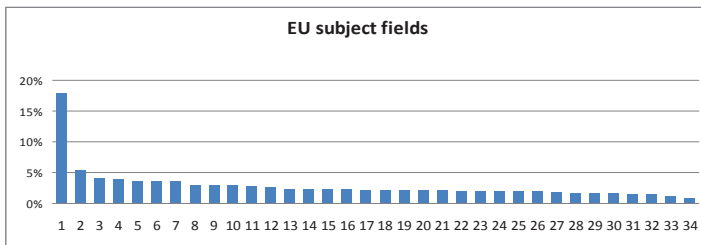


Figure 15. Proportion of EU subject fields in the EEUD Corpus (the figures presented horizontally are the reference numbers indicating the EU subject field shown in Table 36)

### 7.1.2.3. The time period represented by the EEUD Corpus

As regards the time period the corpus represents, it may be regarded as homogenous. The vast majority of texts in the corpus were issued after 2000. As shown in Figure 16, a few texts were issued in the 1990s, and only a handful of texts came out in the 1980s. As described in Section 7.1, the reasons for including texts issued in earlier decades were twofold. First, a few texts from the 1990s and 1980s, like the *Community Guidelines on state aid for small and medium-sized enterprises* from the year 1996 were given by respondents as being still used in their work. Secondly, in some cases there were not enough texts issued after 2000 in certain specific EU subject fields, such as Humanitarian aid and Human rights, therefore texts issued earlier had to be selected.

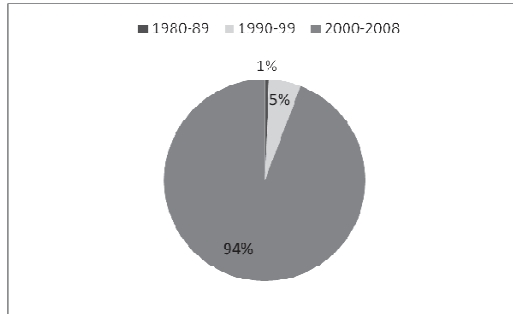


Figure 16. Year of issue of texts in the EEUD Corpus

#### 7.1.2.4. EU institutions represented in the EEUD Corpus

All three main EU institutions, that is, the Commission, the Council and the Parliament, are represented in the EEUD Corpus. As shown in Figure 17, half of the texts were issued by the European Commission. This is not surprising, as the European Union is represented by the Commission and it is the EU's main legislative and executive body. Around 20% of the texts were issued by the Council, and about 10% of the texts were produced jointly by the Council and the Parliament. A little more than 7% of the texts were issued by the European Court of Justice. The category 'Other' includes EU bodies like *European Data Protection Supervisor*, *Eurostat*, *European Central Bank* and *European Court of Auditors*.

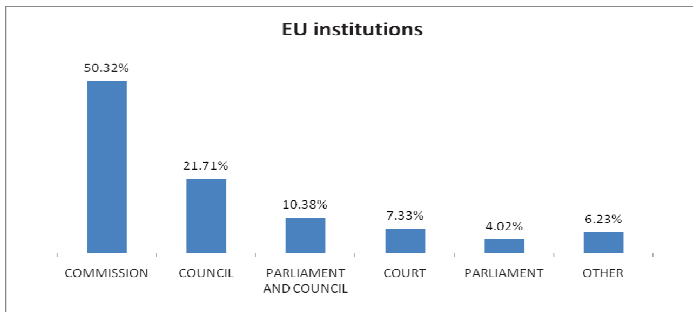


Figure 17. EU institutions represented in the EEUD Corpus

### **7.1.3. Limitations of the needs analysis and the EEUD Corpus**

There are three aspects that should be mentioned as limitations of the findings of the needs analysis survey and the final EEUD Corpus. Firstly, because the group of language users of English in the EU context is diverse and hard to define, and because there is only restricted access to potential respondents, the results of the survey, and thus the Corpus, cannot be considered fully representative. Nevertheless, the relatively high number of respondents and the fact that 74% of the respondents worked in an EU-related job longer than 3 years, and slightly more than half of the respondents worked for an EU institution at the time of the survey, ensure that the genres they regard as relevant can represent the language use they are generally exposed to within the EU context fairly accurately. Secondly, although efforts have been made to balance the proportion of texts representing the different EU-related subject fields, the Corpus is still biased towards the fields Hungarian EU professionals are involved in, and for the fields that enjoyed priority politically, economically, or in any other way during the time period of the survey. Thirdly, texts in the corpus are examples of written discourse, with the only exception being the written introductory statement of the press conference of the European Central Bank. This text was recommended by one of the respondents, and although it is written in advance, and is available in a written format on the internet, it is read out at the press conference, and as such also has features of spoken discourse. The reason for still including the text in the Corpus was that a text of less than 8000 running words in a corpus of slightly more than 1 million running words was not considered a threat regarding distorting the written data, but might yield important lexical items in the field of EU monetary and financial issues. These limitations need to be taken into account when interpreting the final results of the analysis.

#### **7.1.4. Conclusions concerning the EEUD Corpus**

The EEUD Corpus was planned to serve as a text collection representing the genres Hungarian EU professionals use in their daily work, and all the subject-fields of EU activities. Thus the EEUD Corpus provides a useful sample of written English language use within EU institutions for linguistic analyses with pedagogic aims. The final corpus, with 40 different EU genres represented by 241 individual texts, consists of a wide variety of genres representing the different EU subject fields in a balanced proportion, which makes it suitable for the analysis of written English EU discourse.

## 7.2. Results of the corpus analysis

As a first step in the analysis of the EEUD Corpus the frequent lexical items in EU texts were identified and analysed following the steps outlined in Section 6.4.1. Secondly, collocations of selected lemmas were examined as described in Section 6.4.2. Finally, lexical bundles were identified and their structural and functional characteristics were investigated as presented in Section 6.4.3. The next sections will summarise and discuss the findings of these analyses.

### 7.2.1. The EU Word List<sup>7</sup>

The final EUWL contains 513 word families that are made up of 2,457 lexical items. Table 37 presents two examples of the word families with its members in the EUWL. A list of all the headwords of the EUWL can be found in Appendix 5. The most frequent word families include *EUROPE*, *COMMISSION*, *COMMUNITY*, *REGULATION*, *FINANCE* and *IMPLEMENT*, and, as can be seen in Table 38, almost all of them occurred in all the EU subject fields. Examples of the least frequent word families are *CAMPAIGN*, *VULNERABLE*, *WORLDWIDE*, *HIGHLIGHT* and *ALIGN* and, as shown in Table 39, all of them occurred at least 57 times in the Corpus. Their range is at least 40%, which means that words with the lowest frequency occurred in at least 14 EU subject fields.

---

<sup>7</sup> This section is the written version of the presentation *From a specialised corpus to classrooms for specific purposes* given at the Corpus Linguistics 2009 conference at the University of Liverpool.

N	Headword	Cumulative family frequency	% of whole corpus	Members of the word family
1	<i>EUROPE</i>	7401	0.69%	<i>Europe</i> [600] <i>Europe's</i> [90] <i>cross-Europe</i> [1] <i>e-Europe</i> [11] <i>European</i> [6621] <i>European-based</i> [1] <i>European-wide</i> [1] <i>Europeans</i> [29] <i>intra-European</i> [3] <i>non-European</i> [20] <i>trans-European</i> [23] <i>transeuropean</i> [1]
3	<i>COMMUNITY</i>	3635	0.34%	<i>community</i> [3113] <i>communities</i> [355] <i>community's</i> [109] <i>community-based</i> [6] <i>community-wide</i> [10] <i>community-flagged</i> [6] <i>extra-community</i> [1] <i>intra-community</i> [20] <i>non-community</i> [15]

Table 37. Examples of EU word families

N	EU Word family	Cumulative family frequency	Range	% of whole corpus
1	<i>EUROPE</i>	7401	100.00%	0.69%
2	<i>COMMISSION</i>	5390	100.00%	0.50%
3	<i>COMMUNITY</i>	3635	100.00%	0.34%
4	<i>REGULATE</i>	2693	97.06%	0.25%
5	<i>FINANCE</i>	2693	100.00%	0.25%
6	<i>IMPLEMENT</i>	2285	100.00%	0.21%
7	<i>PROCEED</i>	2229	100.00%	0.21%
8	<i>EC</i>	2172	100.00%	0.20%
9	<i>TREATY</i>	1996	100.00%	0.19%
10	<i>POLICY</i>	1925	100.00%	0.18%
11	<i>EU</i>	1883	94.12%	0.17%
12	<i>REQUIRE</i>	1869	100.00%	0.17%
13	<i>AUTHORITY</i>	1864	100.00%	0.17%
14	<i>ESTABLISH</i>	1766	100.00%	0.16%
15	<i>DIRECTIVE</i>	1753	85.29%	0.16%

Table 38. The first 15 word families of the EUWL

N	EU word family	Cumulative family frequency	Range	% of whole corpus
499	<i>DERIVE</i>	64	67.65%	0.01%
500	<i>COMPULSORY</i>	63	52.94%	0.01%
501	<i>EEA</i>	62	47.06%	0.01%
502	<i>MANDATORY</i>	62	50.00%	0.01%
503	<i>ENTAIL</i>	62	67.65%	0.01%
504	<i>DISTINCTION</i>	61	61.76%	0.01%
505	<i>CORE</i>	60	61.76%	0.01%
506	<i>CIRCULATION</i>	60	52.94%	0.01%
507	<i>SCHEDULE</i>	60	52.94%	0.01%
508	<i>ROMANIA</i>	59	41.18%	0.01%
509	<i>CAMPAIGN</i>	59	58.82%	0.01%
510	<i>VULNERABLE</i>	57	47.06%	0.01%
511	<i>WORLDWIDE</i>	57	58.82%	0.01%
512	<i>HIGHLIGHT</i>	57	67.65%	0.01%
513	<i>ALIGN</i>	57	44.12%	0.01%

Table 39. The least frequent 15 word families of the EUWL

As can be seen in Table 40, the word families in the EUWL occurred in a wide range of the EU subject fields. Slightly more than 11% can be found in all the 34 subject fields, and 299 (58.3%) word families occurred in 25 or more subject fields. Altogether 417 (81.3%) word families of the EUWL are found in 20 or more of the EU subject fields in the corpus under study.

EU subject fields covered	Number of word families	% of total EUWL
34	57	11.11%
33	31	6.04%
32	39	7.60%
31	23	4.48%
30	23	4.48%
29	22	4.29%
28	22	4.29%
27	28	5.46%
26	19	3.70%
25	35	6.82%
24	18	3.51%
23	23	4.48%
22	29	5.65%
21	28	5.46%
20	20	3.90%
19	14	2.73%
18	15	2.92%
17	20	3.90%
16	20	3.90%
15	9	1.75%
14	6	1.17%
13	7	1.36%
12	5	0.97%
Total	513	100.00%

Table 40. Range of word families in EUWL

The detailed analysis of the word families of the EUWL found that these include legal words like *REGULATE* and *TREATY*, words in connection with funding like *FUND* and *RESOURCE*, as well as the main EU institutions like *COMMISSION*, *PARLIAMENT* and *PRESIDENCY*. In addition, the word list contains 15 (2.9%) abbreviations, for example, *DG*, *EC*, *OJ*, *SME* and 29 (5.7%) geographical names, which include all member states, the names of two cities: *BRUSSELS* and *LISBON*, and a few – 5 cases (1%) of – function words, such as *PRIOR*, *BEHALF* and *VIA*.

#### 7.2.1.1. Evaluation of the EUWL

The EUWL was evaluated for its specificity for EU discourse and relevance for English for EU purposes, by testing its text coverage, that is, the percentage of the tokens in a corpus that are covered by the elements of the EUWL, in texts representing several different registers



and genres. As shown in Table 41, the EUWL accounts for 18.03% of the tokens in the EEUD Corpus. It is a fairly high coverage compared to earlier ESP word lists, as the coverage of the AWL was reported to be 10% (Coxhead, 2000), and that of the MAWL was 12.24% (Wang et al., 2008) in their respective corpora. In addition to testing the EUWL on the EEUD Corpus itself, it was also tested on four EU texts representing two different genres included in the Corpus, namely, regulations and press releases. The four texts were published about a year later than other texts in the EEUD Corpus, and were selected at random. The EU word families account for 18.7% of the tokens of these EU texts (see Table 42), which is very similar to the coverage of that in the original EEUD Corpus. In Appendix 6, a 500-word extract of an EU legislative text also illustrates the text coverage of the EUWL. The words that are part of the word families of the EUWL are underlined in the sample text. The EUWL also reached a high coverage – 17.02% – in the EU English Corpus (Trebits, 2008), another corpus of EU texts, that was compiled according to different selection criteria than the EEUD Corpus. The EU English Corpus contains genres like information booklets, annual reports and sample test materials for recruitment competition for jobs in an EU institution, which are very different from the genres in the EEUD Corpus. Thus the high coverage reinforces the validity of the EUWL as a word list useful in understanding the lexical composition of EU texts in general.

<b>Texts</b>	<b>Tokens</b>	<b>Text coverage</b>	<b>Mean text coverage</b>
EEUD Corpus	1,076,460	18.03%	
EU English Corpus	197,620	17.02%	
<b>EU corpora in total</b>	<b>1,274,080</b>		<b>17.52%</b>
20 <sup>th</sup> century US short story	4,575	1.44%	
20 <sup>th</sup> century British play	7,873	0.17%	
20 <sup>th</sup> century British novel	105,578	1.39%	
19 <sup>th</sup> century British novel	1,013	2.37%	
<b>Literary texts in total</b>	<b>119,039</b>		<b>1.34%</b>

Table 41. Text coverage of EUWL in EU and literary texts

In order to establish that the EUWL is a truly EU-specific word list, it was also tested on literary texts, news texts, governmental and legislative texts. As can be seen in Table 41, on

average, elements of the EUWL accounted for 1.34% of four different literary genres of altogether almost 120,000 running words. Not surprisingly, this register seems to be the farthest from written English EU discourse. News texts, with slightly less than 5% coverage seem to apply a markedly different lexis than EU texts, which reinforces the findings of earlier research on contrasting the language in EU documents, and news texts on EU-related issues (Jablonkai, 2009a).

In order to avoid the impact of text length on the results, the same analysis was carried out on only the very first 500 tokens of the same texts. As can be seen in Table 42, these results do not show considerable differences in the tendencies revealed using whole texts.

<b>Texts</b>	<b>Tokens</b>	<b>Coverage %</b>	<b>Coverage of first 500 tokens %</b>
EU texts	5,326	18.70%	19.74%
News texts	3,567	5.33%	4.18%
UK government texts	2,017	13.29%	12.63%
UK legislation	13,497	19.25%	17.99%

Table 42. Text coverage of EUWL in different registers for the first 500 tokens

As shown in Table 42, the EUWL covers three times more tokens of governmental texts than news texts. Even higher text coverage was found in legislation texts. Not surprisingly, this shows a considerable similarity of written English EU discourse to these registers, especially as regards their use of lexis. This can firstly be explained by the similar formal, written style of these texts. Secondly, another reason may be the contents of the EEUD Corpus, as EU legal texts account for about 63% of the tokens in the whole corpus.

Furthermore, the word families of the EUWL were compared to word families in two other word lists. One of them was the AWL, which contains general academic lexical items that are widely used in various disciplines (Coxhead, 2000). The other one is the BNC 3000 containing high frequency word families of the BNC. It is considered a general word list with some bias to written language based on the composition of the BNC, which contains 90%

written and 10% spoken texts (Nation, 2004). The aim of the comparison was to test whether the word families of the EUWL can be considered EU-specific, and whether the EUWL adds to the coverage of EU texts by the two more general word lists.

The comparison of the BNC 3000 and the EUWL showed that the total BNC 3000 contains about 60% of the word families in the EUWL. As can be seen in Table 43, most of the EU word families can be found among the second 1000 most frequent word families of the BNC 3000 word list. Word families like *AMEND*, *CLAUSE*, *COHERENCE*, *COOPERATE* and *REINFORCE* are examples that can be found in the EUWL, but are not present in the BNC 3000. Contrasting the 513 word families of the EUWL to the 570 word families of the AWL showed that 323 word families overlapped. The words that can be found in both word lists include *COMMISSION*, *COMMUNITY*, *FINANCE*, *REGULATE*, *IMPLEMENT*, *PROCEED*, *POLICY*, *REQUIRE*, *AUTHORITY*, *ESTABLISH*. This means that almost 40% of the EU word families, for example *ACCESSION*, *ACQUIS*, *CROSS-BORDER*, *ENLARGEMENT* and *RAPPORTEUR*, can be considered EU-specific. These findings are very similar to those of Wang et al. (2008) on the comparison of the MAWL and the AWL. Consequently, these results strengthen their argument for the necessity of the development of subject-specific word lists for different disciplines.

<b>Word list</b>	<b>Overlap in number of EU word families</b>	<b>Overlap in % of all EU word families</b>
AWL	323	63.0%
BNC 3000	298	58.1%
BNC 1 <sup>st</sup> 1000	77	15.0%
BNC 2 <sup>nd</sup> 1000	167	32.6%
BNC 3 <sup>rd</sup> 1000	54	10.5%

Table 43. Comparison of the EUWL to the AWL and BNC 3000

A further argument for the application of the EUWL in ESP teaching is the high coverage in EU texts it provides. As shown in Table 44, the first 2000 word families of the GSL and the families of the EUWL together account for 93.5% of the EEUD Corpus, that is, already very close to the level of a 95% coverage, which is suggested as necessary for understanding a text

without a dictionary (Hirsch & Nation, 1992; Nation & Waring, 1997). Table 44 also shows that the EU-specific word families yield approximately 5% higher text coverage than the coverage of the GSL and the AWL together, which indicates that it is more beneficial to apply the word list specific to the EU subject field to the ESP teaching practice.

Word lists	Coverage of EEUD Corpus
GSL+EUWL	75.46 + 18.03 = <b>93.5%</b>
GSL+AWL	75.46 + 13.8 = <b>89.26%</b>
BNC 3000	88.54%

Table 44. Text coverage of general and specific word lists

### 7.2.1.2. Limitations of the EUWL

It needs be noted that there are a few aspects of lexis that are not covered by word lists in general, and by the EUWL, in particular. One aspect is hyponymy, that is, some lexical items, for example *COUNCIL*, are not included in the EUWL, because they are elements of the GSL, even though they have a specific meaning in the EU context. On the other hand, there are a few (11) elements which are included in the EUWL, despite their general use, as they were not part of the GSL, mainly because of its age, as it was compiled in the 1950s (West, 1953). Examples of these include *AUTOMATIC*, *WEBSITE*, or *INTERNET*. The second aspect of lexis that is not covered by the EUWL is MWIs. Although many lexical items are used as parts of MWIs, the concept of the word list of word families, as applied in previous studies in ESP, concentrates on single-word lexical items. Therefore, for ESP pedagogic purposes, the EUWL should be complemented by MWIs strongly associated with EU discourse.

### 7.2.1.3. Conclusions concerning the EUWL

On the whole, the analysis reinforced the findings of earlier corpus analyses conducted on the language of different disciplines, as it found a marked difference between the elements of the EUWL, and that of other general and academic word lists. The study revealed that there is a considerable specificity in the English written discourse within the European Union, as

represented by the EEUD Corpus that can be characterised by its stereotypical use of particular lexical items. Further analysis of the elements of the EUWL should focus on the meaning these lexical items express, and the patterns they regularly form in general language use, and in the specialised EU context.

### 7.2.2. Collocations in the EEUD Corpus

One way to complement the EUWL in order to obtain a more comprehensive view of the lexical composition of written English EU discourse is to provide collocational information for the elements of the word list. This chapter will describe the results of this kind of analysis on a small selection of lexical items of the EUWL. The collocational analysis is also used to compare the lexical behaviour of the selected items in the current corpus of specialised EU discourse, and in general English. As described in Section 6.4.2.1, the analysis of collocations in EU discourse focuses on 16 lexical items. The selection was motivated by two factors: (1) the selected lexical items are strongly associated with written English EU discourse, that is, they are among the key words of the EEUD Corpus, and they are members of the EUWL; (2) they pose some kind of difficulty to the users of EU documents, that is, they were mentioned as examples by EU professionals. Although originally it was 15 lemmas that the analysis focused on, in the course of the analysis it turned out that considering the spelling variants *co-operation* and *cooperation* separate items yields interesting insights into the behaviour of lexical items in the EEUD Corpus. Therefore the final list includes:

- (a) ten nouns: CO-OPERATION, COOPERATION, POLICY, COMMISSION, CRITERION, REGULATION, PROJECT, INITIATIVE, COMMITMENT, OBJECTIVE
- (b) one adjective: EUROPEAN;
- (c) five verbs: LAY, NOTIFY, FUNCTION, ENSURE, IMPLEMENT.

The analysis was carried out with the help of *Sketch engine* (Kilgarriff & Tugwell, 2001), which provides output for each lemma in the form of a table in which collocates are classified according to the grammatical relations they form with the node word. Sample outputs of the collocations of the lemma NOTIFY can be seen in Figure 13 in Section 6.4.2.2.

Previous research into the collocational behaviour of lexical items in specialised corpora found that collocations become more fixed in specialised texts than in general corpora (Gledhill,

2000; Nelson, 2000, 2006). It is important to note, though, that **fixedness** is interpreted in different ways by researchers. Gledhill (2000), with a more theoretical stance, argued that fixedness of collocations in science writing demonstrated the idiom principle at work. He noted that: “In some instances collocation involves terminology and reflects the recurrent semantics of the specialist domain. In other instances collocation reveals the dominant discourse strategies in the research article” (p. 130). Fixedness by Nelson (2006) was interpreted in relation to semantic prosody<sup>8</sup>, and semantic preferences the collocates of the analysed node words demonstrate in the specialised BEC and in the general BNC. He argued that collocates in the specialised environment become more fixed, that is, the percentage of collocates that are covered by semantic preferences is greater in the specialised corpus, suggesting greater collocational variety in the general corpus.

**Fixedness** in the present study was analysed from two angles. Firstly, the number of grammatical relations the selected lemmas form with their collocates were examined, resulting in the characteristic grammatical behaviour of each lemma in the specialised EEUD Corpus and the general BNC Written. Secondly, the semantic preferences and semantic prosodies of the lemmas in both corpora were investigated, and conclusions were drawn on the variety or fixedness of lexical behaviour of the selected lemmas in the EEUD Corpus. Results of these collocational analyses correspond only partially to the findings of earlier studies of collocations in specialised texts. The following sections will present these results, with a detailed discussion of the findings.

#### **7.2.2.1. Grammatical behaviour of selected lemmas**

As a first step, the number and type of grammatical relations the selected lemmas form with their collocates were compared in the BNC Written and the EEUD Corpus. In order to

---

<sup>8</sup> Under semantic prosody Nelson (2000, 2006) refers to both semantic prosody and semantic preference as defined in the present study.

make the data comparable, normalised frequencies, that is, the number of instances of particular grammatical relations per million words, were used, and only grammatical relations that were found with a frequency of 3 or higher per million words were included in the analysis. The cut-off point was adopted based on the comparative collocational analysis of the lemma DEAL in two registers by Biber et al. (1998). The grammatical relations with frequency data and examples of collocates of one lemma from groups with different characteristics – COOPERATION with more grammatical relations in the EEUD Corpus and COMMISSION with more grammatical relations in the BNC Written – are illustrated in Table 46 and Table 47 (see pp. 203-204). The grammatical relations are given in order of frequency, and the identical relations the lemmas form in the EEUD Corpus and the BNC Written, are shaded. Collocates are highlighted in bold in the examples. As can be seen in Table 46, there is only a single identical grammatical relation the lemma COOPERATION forms in the two corpora, and exhibits a much greater grammatical variety in the EEUD Corpus. However, in the case of COMMISSION, there is a greater variety of grammatical relations in the BNC Written with half of the relations being the same in the two corpora.



<b>Lemma</b>	<b>Normalised frequency in EEUD</b>	<b>Number of grammatical relations in EEUD</b>	<b>Normalised frequency in BNC Wr</b>	<b>Number of grammatical relations in BNC Wr</b>	<b>Number of identical grammatical relations</b>
COMMISSION n	126.5	6	104.5	12	6
COMMITMENT n	235.2	10	64.8	7	4
COOPERATION n	879.5	13	11.5	1	1
CO-OPERATION n	71.4	9	34.6	5	2
CRITERION n	275.3	9	45.8	6	4
INITIATIVE n	375.7	9	52.3	5	3
OBJECTIVE n	1099.0	10	67.5	5	3
POLICY n	1239.6	11	315.7	20	7
PROJECT n	983.0	10	182.0	12	5
REGULATION n	925.6	12	64.0	6	4
ENSURE v	1216.5	6	134.6	5	3
FUNCTION v	45.4	4	15.8	2	2
IMPLEMENT v	742.5	12	39.3	2	2
LAY v	712.8	8	150.7	12	7
NOTIFY v	361.6	10	9.0	1	1
EUROPEAN adj	1741.8	5	184.5	2	2

Table 45. Number of grammatical relations of selected lemmas in the EEUD corpus and the BNC Written

Comparing the variety of grammatical relations of all the selected lemmas, however, shows that there are only four lemmas – the ones shaded in Table 45 – namely, COMMISSION, POLICY, PROJECT and LAY that are part of more types of grammatical relations in the BNC Written, than in the EEUD Corpus. As can be seen in Table 45, all the other lemmas show a greater variety in their grammatical relations in the specialised EEUD Corpus. These findings seem to suggest that the grammatical behaviour of these lemmas is not fixed in the EEUD Corpus. It seems that these patterns are not primarily influenced by the relative frequency of lemmas, as all the lemmas analysed are key words in the specialised corpus, and therefore have a higher normalised frequency in the EEUD Corpus. Findings seem to indicate that lemmas with fairly general meanings, for example POLICY, PROJECT and LAY, or with several senses like COMMISSION, tend to be involved in fewer grammatical relations in the EEUD Corpus. Results suggest a certain degree of fixedness in the case of these lemmas. However, the majority of the analysed lemmas do not exhibit this kind of fixed grammatical behaviour. In

some cases the difference in the number of grammatical relations is not considerable, like with the lemmas ENSURE and COMMITMENT. There are, however, cases where the number of grammatical relations formed in the EEUD Corpus is more than double of that in the BNC Written. Examples of these include COOPERATION, IMPLEMENT, NOTIFY and EUROPEAN. It is interesting to compare the patterns the lemma COOPERATION, and its spelling variant COOPERATION exhibit. Findings suggest that the hyphenated variant is preferred in the general use of English, whereas the unhyphenated spelling is prevalent in written English EU discourse. It is indicated not only by the difference in the normalised frequencies of the lemmas, but also in the different grammatical behaviour of the two lemmas. Although there are more grammatical relations in the case of both variants in the EEUD Corpus, there is a marked difference in the exact number of grammatical relations in the two corpora. In the BNC Written there is about half as many grammatical relations as in the EEUD Corpus, in the case of the hyphenated spelling variant. In the case of the unhyphenated variant, however, there is only a single grammatical relation with a normalised frequency higher than 3 in the BNC Written. In marked contrast, there are 13 different grammatical relations the unhyphenated variant of the lemma forms with collocates in the EEUD Corpus. As shown in Table 45 and Table 46, these data reveal a considerable difference in the grammatical patterns the spelling variants exhibit in the two corpora.

COOPERATION n					
BNC Written			EEUD Corpus		
Grammatical relation	Frequency per million words	Example	Grammatical relation	Frequency per million words	Example
modifier	4.5	<i>mutual/close/economic cooperation</i>	modifier	395	<i>enhanced/judicial/close cooperation</i>
			object_of	209	<i>strengthen/reinforce cooperation</i>
			and/or	139	<i>exchanges and cooperation between appropriate cooperation and coordination</i>
			modifies	94.5	<i>cooperation objective/agreement</i>
			pp_in	89.3	<i>cooperation in criminal/civil matters</i>
			n_modifier	86.3	<i>police cooperation</i>
			pp_with	73.6	<i>cooperation with other/third countries</i>
			pp_between	72.2	<i>cooperation between higher education/cultural institutions</i>
			subject_of	52	<i>cooperation referred to in Article 42(6)</i>
			pp_on	24.5	<i>cooperation on issues</i>
			pp_at	6.7	<i>cooperation at European level</i>
			adj_subject_of	6.7	<i>cooperation is essential</i>
			pp_within	3.7	<i>cooperation within the meaning of Article 7(2)</i>

Table 46. Grammatical relations the lemma COOPERATION forms with collocates in the EEUD Corpus and the BNC Written

COMMISSION n					
BNC Written			EEUD Corpus		
Grammatical relation	Frequency per million words	Example	Grammatical relation	Frequency per million words	Example
modifier	55.85	<i>Audit Commission, Monopolies and Mergers Commission</i>	object_of	10.42	<i>inform the European Commission be charged a commission of 6%</i>
subject_of	19.52	<i>Royal Commission recommended</i>	possession	10.42	<i>The Commission 's powers of inspection</i>
and/or	15.02	<i>new regulatory commissions and consumer councils</i>	subject_of	10.42	<i>the Commission is, as a rule, exempt from all taxes</i>
object_of	14.06	<i>A Truth Commission was established</i>	modifier	5.95	<i>be provided by the joint Commission, the European Commission</i>
pp_obj_of	11.95	<i>banks may charge a commission President of the Cuban Commission of Human Rights</i>	modifies	5.21	<i>Commission approval</i>
modifies	7.19	<i>Special Commission report</i>	and/or	3.72	<i>the commission and other charges</i>
pp_obj_by	6.11	<i>proposed by the European Commission</i>			
pp_on	5.64	<i>the Royal Commission on Pollution</i>			
possession	5.39	<i>The Royal Commission 's proposals</i>			
pp_of	5.24	<i>commission of inquiry</i>			
pp_for	4.28	<i>the Commission for Racial Equality</i>			
pp_obj_to	3.50	<i>submitted evidence to the Royal Commission</i>			

Table 47. Grammatical relations the lemma COMMISSION forms with collocates in the EEUD Corpus and the BNC Written

In general, it seems that the grammatical behaviour of the selected lemmas is not so much influenced by their POS, as there are both nouns and verbs in the group of lemmas that form more grammatical relations in the EEUD Corpus, as well as in the group of lemmas that form more grammatical relations in the BNC Written. Furthermore, it is neither the absolute nor the relative frequency of the selected lemmas that affects their grammatical behaviour, as all the lemmas occur more frequently in the much larger BNC Written, and have a higher normalised frequency in the EEUD Corpus. Instead, it seems that their grammatical behaviour is influenced by their meanings, and the number of senses they express in the given context. Further analysis is therefore needed to describe the grammatical relations of lemmas in EU discourse, and in general English with a wider scope, in order to identify the factors and their interaction that influence the typical grammatical behaviour of particular lemmas in particular registers.

#### **7.2.2.2. Semantic preference and semantic prosody of selected lemmas**

Based on the findings of the comparison of grammatical patterns in the two corpora, six lemmas were chosen for further analysis in order to gain insights into fixedness in terms of semantic prosody. Two of the selected lemmas, COMMISSION and LAY, have fewer grammatical relations in the EEUD Corpus, and four of them, COOPERATION, CO-OPERATION, EUROPEAN and IMPLEMENT have more grammatical relations in the EEUD Corpus. Firstly, the collocates in the same grammatical relations were grouped into relevant semantic sets and summarised in a table format, as shown in Table 48 and Table 49, with the data of the lemmas CRITERION and EUROPEAN. Next, the identified semantic preferences and semantic prosodies were compared across the general BNC Written and the specialised EEUD Corpus.

CRITERION noun	
BNC Written	EEUD
<b>grammatical relation: object of</b>	
1. <u>meet</u> collocates: <i>satisfy, fulfil, meet, match, fit</i>	1. <u>meet</u> collocates: <i>fulfil, fulfil, meet, satisfy</i>
2. <u>set</u> collocates: <i>formulate, adopt, outline, define, establish</i>	2. <u>set</u> collocates: <i>set, agree, establish, lay</i>
3. <u>respect</u> collocates: -	3. <u>respect</u> collocates: <i>follow, respect</i>
4. <u>list</u> collocates: <i>list, specify</i>	4. <u>list</u> collocates: <i>list, specify, give</i>
5. <u>apply</u> collocates: <i>apply, use, employ</i>	5. <u>apply</u> collocates: <i>apply</i>
6. <u>evaluate</u> collocates: <i>assess, judge, review</i>	6. <u>evaluate</u> collocates: -
Other collocates: <i>invoke, exemplify, propose, interpret, identify, derive, alter</i>	Other collocates: <i>need, see, base, propose</i>
<b>Number of semantic preferences</b>	
5	5
Number of identical semantic preferences: 4	
<b>grammatical relation: pp for</b>	
1. <u>participation</u> collocates: <i>eligibility, inclusion, exclusion</i>	1. <u>participation</u> collocates: -
2. <u>evaluation</u> collocates: <i>selection, evaluation, assessment, diagnosis</i>	2. <u>evaluation</u> collocates: <i>selection</i>
3. <u>membership</u> collocates: <i>admission, acceptance, membership, entry, access</i>	3. <u>membership</u> collocates: <i>membership</i>
4. <u>distribution of funds</u> collocates: - Other collocates: <i>imposition, promotion, recognition, transfer, success, use, service</i>	4. <u>distribution of funds</u> collocates: <i>allocation</i> Other collocates:
<b>Number of semantic preferences</b>	
3	3
Number of identical semantic preferences: 2	
<b>Total number of semantic preferences:</b>	
8	8
Total number of identical semantic preferences: 6	

Table 48. Comparison of the semantic preferences of the lemma CRITERION

Findings of the comparison show that collocations in the EEUD Corpus indicate a certain degree of fixedness in Nelson's (2006) terms, that is, the proportion of collocates of the selected lemmas covered by semantic preferences is higher in the specialised corpus. The other factor

that also supports the concept of fixedness in collocational patterns in a specialised corpus is the fact that, in general, there is a larger number of semantic preferences identified among the collocates of the lemmas in the BNC Written. The number of semantic preferences of lemmas ranges from 5-13 in the EEUD Corpus, and 7-23 in the BNC Written. The difference between the number of semantic preferences in the two corpora ranges from 0 to 10, and it is greatest in the case of the lemmas with fewer grammatical relations in the EEUD Corpus, that is, in the case of COMMISSION and LAY, with a difference of 9 and 10 respectively.

Nelson (2006) noted that words in his BEC were found to be associated with business-specific semantic preferences, and also with semantic sets that are the same in both BE and in general English. The results of the analysis of the semantic preferences of the selected lemmas in the EEUD Corpus show similar patterns. The comparison of the number of identical semantic preferences lemmas are associated with in the two corpora shows that the analysed six lemmas have, in general, 2-10 identical semantic preferences, with an average of 5.67 per lemma and 2.125 per grammatical relation. The greatest number of identical semantic preferences were identified in the case of CRITERION (see Table 48) and CO-OPERATION with 6 and 5 identical semantic preferences respectively, which corresponds to an average of 3 identical semantic sets in the case of CRITERION, and an average of 2.5 identical semantic preferences in the case of CO-OPERATION, per grammatical relation. The collocational patterns of the lemma EUROPEAN is the most strikingly different in the two corpora, as it was found to have only one identical semantic set in the two corpora, as can be seen in Table 49.

EUROPEAN adj	
BNC Written	EEUD Corpus
<b>grammatical relation: modifies</b>	
semantic preferences: 1. <u>integration</u> collocates: <i>union, integration, unity</i>	semantic preferences: 1. <u>integration</u> collocates: <i>community, union, level, dimension, integration</i>
2. <u>institutions</u> collocates: <i>community, commission, parliament, court</i>	2. <u>citizens</u> collocates: <i>citizen, citizenship</i>
3. <u>counties</u> collocates: <i>country, nation</i>	3. <u>aims, policies</u> collocates: <i>objective, agenda, policy, strategy, area</i>
4. <u>sport</u> collocates: <i>championship, cup, champion, final</i>	4. <u>values and standards</u> collocates: <i>value, heritage, standard, convention</i>
5. <u>finance</u> collocates: <i>bank, currency, fund, mechanism</i>	5. <u>economic activity</u> collocates: <i>work, industry, sector</i>
Other collocates: <i>market, convention, tour, directive, partner, agency, act, language</i>	Other collocates: <i>film, year, study, network</i>
<b>Number of semantic preferences</b>	
5	5
Number of identical semantic preferences: 1	
<b>grammatical relation: and/or</b>	
1. <u>common</u> collocates: <i>single, common, joint</i>	1. <u>common</u> collocates: <i>common, enhanced, joint</i>
2. <u>geography</u> collocates: <i>eastern, western, central, continental, regional, northern, American, Japanese</i>	2. <u>public</u> collocates: <i>civil, public, social</i>
3. <u>time</u> collocates: <i>nineteenth-century, colonial</i>	3. <u>national, international</u> collocates: <i>non-national, national, international</i>
4. <u>economic</u> collocates: <i>monetary, economic</i>	4. <u>areas of co-operation</u> collocates: <i>audiovisual, cultural, cinematographic, industrial, intellectual</i>
Other collocates: <i>indoor, major, other, free, junior, environment, proposed, leading, modern, political</i>	5. <u>positive semantic prosody</u> collocates: <i>strong, high-quality</i>
	Other collocates: <i>territorial, added, key, genuine, active, creative, third, private</i>
<b>Number of semantic preferences</b>	
4	5
Number of identical semantic preferences: 1	
<b>Total number of semantic preferences</b>	
9	10
Total number of identical semantic preferences: 2	

Table 49. Comparison of the semantic preferences of the lemma EUROPEAN



As regards semantic prosody, findings suggest that a preference for positive or negative collocates can only be identified in a handful of cases in both corpora. The few examples of lemmas that exhibit semantic prosody include LAY in the BNC Written and IMPLEMENT in both corpora. LAY has a few collocates that can be considered negative in the BNC Written only. These collocates are the following: *fault*, *blame* and *difficulty*. In the EEUD Corpus this lemma is not associated with any semantic prosody. The other example is IMPLEMENT. It is associated with positive prosody in the EEUD Corpus, with collocates like *fully*, *successfully*, *properly*. In the BNC Written, however, it has positive and negative collocates as well. The positive ones are similar adverbs to the ones in the EEUD Corpus, for example, *well*, *fully*, *successfully*. The negative collocates are nouns such as *sanction* and *cut*. These findings suggest that the verb IMPLEMENT is associated with positive and negative words in general written English, and it has a more positive prosody in written English EU discourse, whereas the verb LAY is neutral in written English EU discourse, but has a slightly negative prosody in general written English. As regards the interaction of typical structure and semantic preference, these results also correspond to Partington's findings (2004), as in the case of IMPLEMENT the positive prosody is typically expressed by adverbs, and negative prosody is expressed by nouns. The tables summarising the collocates grouped into grammatical relations and semantic preferences of COMMISSION, IMPLEMENT and LAY can be found in Appendix 7.

### **7.2.2.3. Limitations of the collocational analysis**

One of the limitations of the collocational analysis carried out in this study originates from the corpus analysis tool, the *Sketch engine*. POS tagging is done automatically by *Sketch engine* with the *Tree tagger* software, which means that tagging is not perfect. Publications on the use of the Tree tagger cite 96% accuracy (Schmid, 1994), which is acceptable for the type of

analysis in this study. Furthermore, collocates were also checked manually in concordance lines in order to minimise errors from inaccurate tagging.

The second limitation that should be mentioned is the categories of semantic preferences used in the analysis of semantic preferences. As semantic preferences can be very diverse, there are no established semantic categories for the analysis of semantic preferences of collocates of particular lexical items. Therefore, the semantic preferences applied in the present study were formed in consultation with an independent ESP researcher. In addition, as Nelson (2006) also pointed out in his comparison of semantic preferences in his BEC and a general English corpus, these comparisons are partially misleading, as the contents and size of the compared semantic preferences might be different. Moreover, as semantic preferences and prosody were compared within the categories of grammatical relations, only collocates in the identical grammatical relations were included in the analysis. Nevertheless, the analysis provided some useful insights into how certain lexical items behave in specialised and general corpora.

A third limitation that needs to be noted is that the present study used the lemma as the unit of analysis. The literature on collocations emphasised that individual word forms of the node word often collocate with different words (Hoey, 2005; Renouf, 1987; Sinclair, 1991; Tognini-Bonelli, 2001). Consequently, the analysis of lemmas hides these differences. Considering, however, that the main aim of the analysis is to provide insights for pedagogic purposes, the use of lemmas as the unit of analysis was seen as more useful, as this is the usual way language learners are given information on lexical items, for example, in glossaries or dictionaries.

Finally, the collocational analysis in the present study was carried out on a very limited scope concentrating on a few lemmas only, which does not allow for generalisation of the results. In order to provide a clearer picture of the characteristic lexical behaviour and lexicogrammatical patterns in written English EU discourse, and to gain insights into how it differs from general English use, a much wider scale of analysis is needed.

#### **7.2.2.4. Conclusions concerning the collocational analysis**

Although the scope of the collocational analysis is rather limited, it yielded relevant insights into the lexical and lexicogrammatical patterning in the EEUD Corpus. Firstly, it was found that fixedness in specialised contexts is only partially true for collocations in the EEUD Corpus. The comparison of grammatical relations of the selected lemmas revealed that there is greater diversity in the EEUD Corpus in the grammatical patterns, than in the general BNC Written. On the one hand, it seems that absolute frequency is the cause of a greater variety of collocates, but findings appear to indicate that the same is not true for grammatical relations. The exact factors that determine variety in grammatical behaviour remain, however, uncovered. On the other hand, fixedness in terms of semantic preferences is a characteristic feature of the lexical patterns in the EEUD Corpus. The patterns of collocational behaviour in a specialised corpus identified by Nelson (2006) were found in the EEUD Corpus as well. This means that (1) semantic preferences cover a higher percentage of collocates in the specialised EEUD Corpus than in the general BNC Written; (2) there are semantic preferences that lemmas share across corpora, for example, in the case of CRITERION; and (3) there are subject-field-specific semantic preferences that can only be found in the EEUD Corpus, for example, in the case of EUROPEAN.

### 7.2.3. Lexical bundles in the EEUD Corpus<sup>9</sup>

This section will focus on the results of the corpus-driven analysis of MWIs in the EEUD Corpus. As outlined earlier, the unit of analysis was the lexical bundle as proposed by Biber and Conrad (1999) for a frequency-based framework for analysing different registers. Altogether 247 lexical bundle types were identified in the EEUD Corpus. The full list of lexical bundle types is provided in Appendix 8. As shown in Table 50, the most frequently occurring lexical bundle types, for example, *in accordance with the, of the European Union* and *referred to in article* occurred more than 600 times per million words, with the most frequent lexical bundle type occurring 783 times per million words. These bundle types are used in a wide-range of EU texts. Many of them are used in nearly half of all the texts in the EEUD Corpus. As many as 89 (36%) out of the 247 lexical bundle types occur in at least half of the 40 different EU genres, and are also used in at least half of the 34 EU subject fields. The least frequent lexical bundle types occur 47 times in the EEUD Corpus, which corresponds to the 40 per million cut-off point, and are used in 47 EU texts. Lexical bundle types like *in the course of, the principle of subsidiarity, the exchange of information*, are some examples of these. Lexical bundle types with the lowest range are used in 24 of the 241 EU texts in the EEUD Corpus, for instance, *of the internal market, the procedure referred to* and *within the scope of*. Table 50 shows the EU-specific bundles in the EEUD Corpus. In addition to the criteria set for lexical bundles in the present study, EU-specific bundles were defined as lexical bundles that occur in at least half of the EU genres and half of the EU subject fields. In order to ensure EU specificity, bundles that can be found among the 40 most frequent four-word lexical bundles of the BNC (Scott & Tribble, 2006, p. 140) were excluded from the list presented in Table 50.

---

<sup>9</sup> This section is an extended version of the article Jablonkai (in press).

N	EU-specific lexical bundle	Frequency in EEUD Corpus	No. of texts in which bundle type occurs	No. of genres in which bundle type occurs	% of all genres in EEUD Corpus	No. of subject fields in which bundle type occurs	% of all subject fields in EEUD Corpus
1	in accordance with the	920	110	32	80	33	97.0
2	of the European Union	750	139	32	80	33	97.0
3	Article # of the	698	110	34	85	34	100.0
4	# of the Treaty	628	77	24	60	32	94.1
5	the European Parliament and	549	103	28	70	32	94.1
6	of the Member States	443	97	31	77.5	33	97.0
7	Regulation EC No #	441	70	22	55	27	79.4
8	of the European Parliament	399	109	28	70	33	97.0
9	having regard to the	375	113	21	52.5	32	94.1
10	in accordance with Article	346	56	20	50	30	88.2
11	accordance with Article #	319	55	20	50	30	88.2
12	with a view to	307	100	31	77.5	32	94.1
13	Articles # and #	285	58	21	52.5	29	85.2
14	the implementation of the	281	82	31	77.5	34	100.0
15	in the field of	275	75	29	72.5	30	88.2
16	# and # of	271	75	28	70	30	88.2
17	European Parliament of	262	79	24	60	29	85.2
18	Parliament and of the	262	79	24	60	29	85.2
19	and of the Council	261	80	25	62.5	29	85.2
20	in Article # of	260	68	21	52.5	30	88.2
21	European Parliament and the	247	60	23	57.5	29	85.2
22	Parliament and the Council	212	53	21	52.5	27	79.4
23	of the European Communities	205	75	28	70	32	94.1
24	the basis of the	204	69	29	72.5	28	82.3
25	for the purposes of	203	56	21	52.5	28	82.3
26	the Commission and the	198	55	25	62.5	29	85.2
27	of the Council of	192	81	25	62.5	29	85.2
28	the Member States and	181	71	26	65	29	85.2
29	the Council of #	167	59	21	52.5	25	73.5
30	as set out in	165	55	22	55	27	79.4
31	of Article # of	165	45	21	52.5	21	61.7
32	and # of the	163	44	23	57.5	27	79.4
33	the framework of the	162	54	27	67.5	27	79.4
34	and the Member States	154	42	21	52.5	23	67.6
35	Member States and the	148	56	22	55	28	82.3
36	of # December #	144	63	21	52.5	29	85.2
37	of the EC Treaty	142	47	22	55	21	61.7
38	in the area of	141	46	24	60	25	73.5
39	by the Member States	139	53	22	55	28	82.3
40	set out in the	138	59	30	75	26	76.4
41	for the purpose of	138	48	20	50	27	79.4
42	Economic and Social Committee	135	53	20	50	28	82.3
43	in the Official Journal	134	72	20	50	28	82.3
44	of # June #	130	73	25	62.5	29	85.2
45	within the framework of	127	49	26	65	28	82.3
46	in the light of	126	60	27	67.5	31	91.1
47	in relation to the	125	50	23	57.5	27	79.4
48	in the Member States	123	45	23	57.5	26	76.4

Table 50. EU-specific lexical bundle types in the EEUD Corpus

N	EU-specific lexical bundle	Frequency in EEUD Corpus	No. of texts in which bundle type occurs	No. of genres in which bundle type occurs	% of all genres in EEUD Corpus	No. of subject fields in which bundle type occurs	% of all subject fields in EEUD Corpus
49	in order to ensure	122	59	26	65	31	91.1
50	with regard to the	121	50	25	62.5	28	82.3
51	the Council and the	120	53	23	57.5	26	76.4
52	in the framework of	115	49	26	65	22	64.7
53	# of the EC	111	42	22	55	20	58.8
54	to the European Parliament	111	41	20	50	25	73.5
55	to ensure that the	110	62	28	70	31	91.1
56	the objectives of the	103	42	24	60	19	55.8
57	the context of the	97	48	27	67.5	27	79.4
58	taking into account the	96	49	25	62.5	29	85.2
59	in line with the	96	43	25	62.5	25	73.5
60	to Article # of	89	40	20	50	25	73.5
61	of # July #	87	46	20	50	25	73.5
62	the European Union and	86	36	21	52.5	19	55.8
63	# December # on	81	48	21	52.5	28	82.3
64	on the implementation of	78	38	24	60	22	64.7
65	be taken into account	73	47	28	70	24	70.5
66	in order to achieve	67	48	23	57.5	27	79.4
67	the results of the	67	39	21	52.5	22	64.7
68	Member States of the	67	32	20	50	22	64.7
69	the light of the	64	40	22	55	25	73.5
70	the scope of the	60	42	24	60	22	64.7
71	the development of the	60	34	20	50	20	58.8
72	in the European Union	57	31	20	50	19	55.8
73	of # April #	55	40	20	50	23	67.6
74	as soon as possible	54	35	23	57.5	22	64.7
75	to contribute to the	53	28	21	52.5	19	55.8
76	take into account the	52	38	22	55	23	67.6
77	in the implementation of	50	33	20	50	21	61.7
78	the establishment of a	48	27	21	52.5	19	55.8
79	to be carried out	48	24	20	50	19	55.8

Table 50. cont. EU-specific lexical bundle types in the EEUD Corpus

Comparing the number and frequency of lexical bundles in different registers shows that EU texts use by far the most lexical bundle types – slightly more than double the number of lexical bundle types in academic prose – and these lexical bundles are used very frequently in the EEUD Corpus. As can be seen in Table 51, the total number of cases of lexical bundles in the EEUD Corpus is six times greater than in academic prose, around ten times greater than in fiction, and in a general English corpus, the written part of the *BNC Sampler*, and almost twenty times greater than in news texts. This finding harmonises with the results of an earlier analysis of lexical bundles in a smaller corpus of EU texts where it was found that the frequency of

lexical bundles in EU texts is slightly more than double the frequency of lexical bundles in news texts on EU-related issues (Jablonkai, 2009a). The results of the present study suggest that the language used in EU documents is very formulaic, and a relatively large proportion of EU texts are covered by lexical bundles. In order to illustrate the coverage of lexical bundles in EU texts, a randomly selected sample text, with all the lexical bundles in bold, is provided in Appendix 9. The tokens in lexical bundles amount to 27% of the total number of tokens in the sample text, that is, almost a third of the tokens in this text are used as part of lexical bundles. However, the findings of the comparison of lexical bundles in the EEUD Corpus and other corpora representing particular registers, should be regarded with caution as, the EEUD Corpus contains a greater variety of genres than the corpora of the *BNC Baby*, which could be the reason for the larger number of bundle types in the EEUD Corpus.

Corpus	Number of lexical bundle types	Normalised number of lexical bundle types (per million)	Tokens	Total cases	Texts
BNC Sampler	44	44	1,005,533	2076	84
BNC Academic	116	112	1,039,776	4354	30
BNC Fiction	84	85	988,485	2763	25
BNC News	41	42	979,911	1425	97
EEUD	268	228	1,174,753	27558	241

Table 51. Comparison of number and frequency of lexical bundles across registers

### 7.2.3.1. Structural analysis of lexical bundles in the EEUD Corpus

The analysis of the grammatical structures was based on the taxonomy of Biber et al. (2004). A few new categories were added to the original types in order to classify bundles which incorporated structures that had not been recognised in earlier research. A new sub-type was added to the types of lexical bundles incorporating dependent clause fragments based on Biber et al. (1999), who identified four such lexical bundles, namely, *as shown in figure*, *as we have seen*, *as we shall see* and *if there is a*. The lexical bundles in the EEUD Corpus with a similar structure and therefore classified as lexical bundles incorporating adverbial clause are, for example, *as set out in*, *as referred to in*.

Furthermore, there were two main structural types added. These are bundles incorporating adjectives and adverbs on the one hand, and lexical bundles with numbers only, on the other. Examples of the first type include, for instance, *European Economic and Social* for the adjective sub-type and *in so far as* for the adverb sub-type. Numbers in the lexical bundles are replaced by the hash symbol (#) by *WordSmith Tools 4*, and they are presented in this way following the practice of earlier studies (e.g., Forchini & Murphy, 2008). The structural type with numbers was divided further into a sub-type of lexical bundles incorporating numbers only, like in *# and # and* (frequency: 68) and others with numbers and prepositions like *# of the EC* (frequency: 111), *# EC of the* (frequency: 121) and *and # of the* (frequency: 163). Although there are altogether only 5 lexical bundle types incorporating numbers, most of them occur rather frequently (as shown in brackets) with the most frequent being *# and # of* with a frequency of 271. Similarly, a comparison of the 30 most frequently used four-word lexical items in the *BNC* and the *Financial Times Corpus*, found that there is a high frequency of four-word lexical items with numbers in the *Financial Times Corpus*, because of its specialised financial subject matter (Forchini & Murphy, 2008). In the *EEUD Corpus*, however, these numbers very often refer to the numbers of articles or paragraphs in legal texts, as shown in Examples (1) and (2):

- (1) *The means which OLAF has at its disposal for the purpose of pursuing those objectives are specifically listed, notably in Articles 4, 7 and 9 of the regulation.*
- (2) *Where reference is made to this paragraph Articles 5 and 7 of Decision 1999/468/EC shall apply, having regard to the provisions of Article 8 thereof.*

Table 52 gives an overview of all structural types of the lexical bundles in the *EEUD Corpus* with the new types and sub-types shaded. The numbers of lexical bundles in the main structural types are illustrated in Figure 18. The intra-rater reliability was found to be  $Kappa = 0.911$  ( $p < 0.001$ ), which indicates outstanding agreement between ratings, and the inter-rater reliability



was found to be Kappa = 0.783 ( $p < 0.001$ ), which indicates substantial agreement between raters (Landis & Koch, 1977; Sajtos & Mitrev, 2009).

Structural types	Sub-types	Number of bundle types	Sample bundles in EU discourse
<b>1. Lexical bundles that incorporate verb phrase fragments</b>	1d. Verb phrase (with non-passive verb)	10	<i>the Member States shall</i>
	1e. Verb phrase (with passive verb)	14	<i>be taken into account</i>
	<b>Sub-total</b>	<b>24</b>	
<b>2. Lexical bundles that incorporate dependent clause fragments</b>	2d. <i>To</i> -clause fragment	7	<i>to ensure that the</i>
	2e. <i>That</i> -clause fragment	1	<i>the fact that the</i>
	2f. Adverbial clause	8	<i>as referred to in</i>
<b>Sub-total</b>		<b>16</b>	
<b>3. Lexical bundles that incorporate noun phrase and prepositional phrase fragments</b>	3a. Noun phrase with <i>of</i> -phrase fragment	74	<i>the basis of the, Council of the European</i>
	3b. Noun phrase with other post-modifier fragment	9	<i>proposal from the Commission, the Treaty establishing the</i>
	3c. Other noun phrase expressions	24	<i>the European Union and, States and the Commission</i>
	3d. Prepositional phrase expressions	89	<i>accordance with the procedure, to in Article #</i>
	3e. Comparative expressions	2	<i>as well as the</i>
	<b>Sub-total</b>		<b>198</b>
<b>4. Lexical bundles that incorporate adjectives and adverbs</b>	4a. Adjectives	3	<i>the Economic and Social</i>
	4b. Adverbs	1	<i>in so far as</i>
<b>Sub-total</b>		<b>4</b>	
<b>5. Lexical bundles that incorporate numbers</b>	5a. Numbers	1	<i># and # and</i>
	5b. Numbers and preposition	4	<i># EC of the</i>
<b>Sub-total</b>		<b>5</b>	
<b>Total</b>		<b>247</b>	

Table 52. Structural types and examples of lexical bundle types in the EEUD Corpus

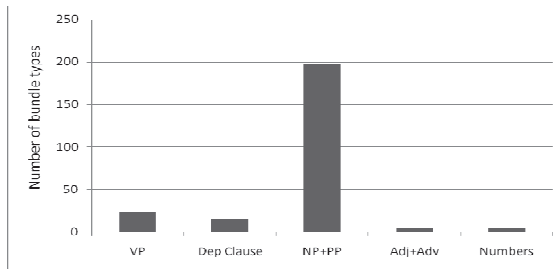


Figure 18. Structural distribution of lexical bundle types in the EEUD Corpus

The structural analysis of lexical bundles in the EEUD Corpus yielded similar findings to the earlier analysis of lexical bundles in English EU texts (Jablonkai, 2009a, 2009b). As shown in Figure 18, most lexical bundles in this study also contain noun phrases and prepositional phrases. As can be seen in Table 52, there are 198 bundles of this type which corresponds to 80% of all bundle types in the EEUD Corpus. Examples of this type include *Committee of the Regions, proposal from the Commission, the entry into force, on behalf of the*.

A closer look at the structural sub-types, as presented in Table 52, shows that the most common among the bundle types with noun phrases and prepositional phrases are bundles with prepositional phrase expressions with almost 36% of all bundle types. This sub-type includes bundles like *in relation to, in the Member States, in order to ensure, with regard to, for in Article #, entry into force*. The second most frequent structural sub-type is noun phrases with *of*-phrase fragments with 74 different bundle types corresponding to 30% of all bundle types.

As shown in Figure 18, the second most frequent main structural group is bundles with verb phrases, which account for almost 10% of all lexical bundles in the EEUD Corpus. As shown in Table 52, slightly fewer than half of these incorporate verb phrases with non-passive verbs and a slight majority is composed of verb phrases with passive verbs. The proportion of bundle types with verb phrases is relatively high in the EEUD Corpus compared to the proportion of similar bundle types in other written registers, like academic prose. Previous studies on lexical bundles (Biber, 2006; Biber et al., 2004; Biber & Conrad, 1999) found that lexical bundles with verb phrases occurred very frequently in spoken registers like conversation and classroom teaching. There were, however, no lexical bundles with verb phrases identified in the written register of academic prose, and textbooks were found to use only a few of them. Hyland (2008), on the other hand, based on his analysis of disciplinary differences in written academic prose, found that in contrast to applied linguistics and business texts, science and engineering texts made frequent use of lexical bundles with passive verbs. He found that

bundles with passive verbs in these technical type of text were used for guiding the reader through the text, or identifying the basis for an argument, for example, *is shown in Figure, are summarised in Table, is based on the* and *can be used to* (Hyland, 2008, p. 11).

In view of these findings, it may be claimed that written English EU discourse exhibits features of academic prose in general, in that it applies an abundance of lexical bundles incorporating noun phrases and prepositional phrases. Furthermore, the fact that the texts in the EEUD Corpus make use of a relatively high number of lexical bundles with verb phrases indicates that the variety of English in these texts resembles the language of written technical types of texts.

#### **7.2.3.2. Functional analysis of lexical bundles in the EEUD Corpus**

The functional analysis of the lexical bundles identified in the EEUD Corpus was carried out based on a revised version of the taxonomy of Biber et al. (2004). A few new categories needed to be added to the framework in order to be able to classify bundles which performed functions that had not been identified earlier in university and academic registers. Most of the newly added categories, however, were created based on the analysis of a smaller EU corpus (Jablonkai, 2009a), and have been found useful in the investigation of the EEUD Corpus as well. These categories were the new main category **Subject-specific** and two sub-categories of referential bundles, namely, the **Quality specification** and **Intertextual** sub-categories. The category of Subject-specific bundles was added based on Hyland's (2008) classification, as he introduced a category of lexical bundles that are related to the actual field of research. In a similar way the category Subject-specific was created for lexical bundles that refer to organisations or documents that are related to the European Union. On the basis of the entities they referred to, the main category was divided into four sub-categories such as **Organisations**, **Documents**, **Codes** and **Other**. A detailed description of these sub-categories will follow in Section 7.2.3.2.4.

One of the additional sub-categories was introduced for lexical bundles that express quality attributes, consequently, this sub-category, called **Quality specification**, was added to the main category of Referential specification bundles. The sentence in Example (3) illustrates bundles of this type:

- (3) *The aim of the programme shall be to contribute to the protection of children, young people and women against all forms of violence and to attain a **high level of health protection, well-being and social cohesion.***

The other sub-category that has already been found useful for the analysis of EU texts (Jablonkai, 2009a; 2009b) is the so called Intertextual referential bundle. This category was created for bundles which refer to other texts, as illustrated in Examples (4), (5) and (6):

- (4) *In cases of imperative need arising from changes in the situation and failing a review of the Council decision as referred to in paragraph 1, Member States may take the necessary measures as a matter of urgency **having regard to the general objectives of that decision.***
- (5) *The Community funds thus distributed shall be administered by the national agencies **provided for in Article 6(2)(b).***
- (6) *(a) the form, content and other details of complaints lodged **pursuant to Article 7 and the procedure for rejecting complaints;***

The third additional sub-category to the category of Referential bundles has not been applied in earlier analysis. The sub-category was created for lexical bundles that express the purpose of certain acts or documents was not applied in the earlier research on the smaller EU corpus (Jablonkai, 2009a). This sub-category, called **Purpose**, is illustrated by Examples (7), (8) and (9):

- (7) *The European Parliament may, acting by a majority of its component Members, request the Commission to submit any appropriate proposal on matters on which it considers that a Union act is required **for the purpose of implementing the Treaties.***
- (8) *In accordance with the principle of proportionality, as set out in that Article, this Directive does not go beyond what is necessary **in order to achieve this objective.***
- (9) ***In order to contribute** to the achievement of the objectives referred to in this Article:*

The results of the functional analysis of lexical bundles identified in the EEUD Corpus applying the revised taxonomy with the additional categories are given in Table 53 with the new categories shaded. The numbers of lexical bundles in the main functional categories are shown in Figure 19. The intra-rater reliability was found to be Kappa = 0.925 (p<0.001), and

the inter-rater reliability was found to be  $Kappa = 0.887$  ( $p < 0.001$ ). Both values indicate outstanding agreement (Landis & Koch, 1977; Sajtos & Mitrev, 2009).

Categories	Sub-categories	Number of bundle types	Sample bundles in EU discourse
<b>I. Stance bundles</b>	<b>A. Epistemic stance</b>	1	<i>the fact that the</i>
	<b>B. Attitudinal/ modality stance</b>		
	B2) Obligation/ directive	6	<i>the Commission shall be, shall be subject to</i>
	B5) Importance	2	<i>it is necessary to</i>
<b>Sub-total</b>		<b>9</b>	
<b>II. Discourse organisers</b>	<b>B. Topic elaboration/ clarification</b>	14	<i>as set out in, on the other hand</i>
<b>Sub-total</b>		<b>14</b>	
<b>III. Referential bundles</b>	<b>A. Identification/ focus</b>	21	<i>hereinafter referred to as</i>
	<b>B. Specification of attributes</b>		
	B1) Quantity specification	1	<i># of the total</i>
	B2) Tangible framing	4	<i>in the field of, in the area of</i>
	B3) Intangible framing	46	<i>in accordance with the, in the framework of</i>
	B4) Quality specification	1	<i>a high level of</i>
	B5) Purpose	6	<i>for the purposes of</i>
	<b>C. Time/ Place/ Text reference</b>		
	C1) Place reference	8	<i>in the Official Journal</i>
	C2) Time reference	25	<i>on # December#</i>
C3) Text-deixis	6	<i>as referred to in</i>	
C4) Multi-functional reference	3	<i>by the Commission in</i>	
C5) Intertextual	22	<i>pursuant to article #</i>	
<b>Sub-total</b>		<b>143</b>	
<b>IV. Subject-specific bundles</b>	A2) EU-related – Reference to a country/ organisation/institution	51	<i>of the European Communities, Economic and Social Committee, Committee of the Regions</i>
	A3) EU-related – Reference to a document	18	<i>the Treaty on European, the provisions of the</i>
	A4) Other	8	<i>in the implementation of</i>
	A5) Codes	4	<i># of the EC, # and # of</i>
	<b>Sub-total</b>		<b>81</b>
<b>Total</b>		<b>247</b>	

Table 53. Functional types and examples of lexical bundle types in the EEU Corpus

In general, the functional distribution of lexical bundles in EU texts corresponds to the findings of the earlier analysis of lexical bundles in EU texts (Jablonkai, 2009a, 2009b). As can

be seen in Figure 19, most of the lexical bundles in EU discourse are **Referential bundles**, specifying several attributes like quantity, quality, purpose, time, place, etc. The second largest group of bundles is related to the European Union, and they thus make up the category of **Subject-specific bundles**. Finally, there are only a few lexical bundles in the current corpus that express stance or serve as discourse organisers.

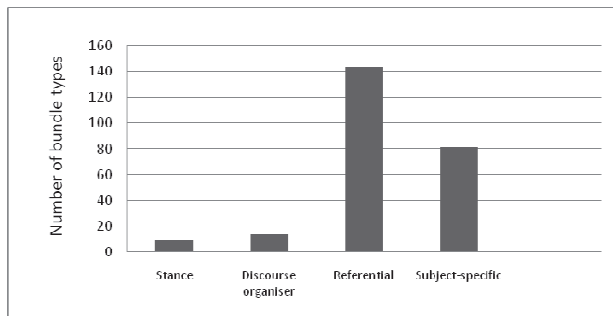


Figure 19. Functional distribution of lexical bundle types in the EEUD Corpus

Biber et al. (2004) found that there is a strong association between the form and function of lexical bundles in academic discourse. Based on the analysis of university registers, their results indicate that most stance bundles use dependent clause fragments, most referential bundles are composed of noun phrase and prepositional phrase fragments, and discourse organisers use all three structural types. Furthermore, they also claim that these patterns are register-specific, which means that particular registers make frequent use of lexical bundles composed of particular structures, for example, academic prose at the written extreme of the analysed registers applies mostly noun phrases and prepositional phrases for referential functions. In what follows, each functional category of lexical bundles is described in detail, discussing the interaction of structural and functional categories as well.

### 7.2.3.2.1. Stance bundles in the EEUD Corpus

There were only nine stance bundles identified in the EEUD Corpus. All of these belong to the category of impersonal **Stance bundles**, that is, they do not use personal pronouns. The stance functions Desire, Intention and Ability are not expressed with lexical bundles in the EEUD Corpus. Instead, bundles are used for expressing the functions **Epistemic** and **Attitudinal stance**, more specifically, **Obligation/ directive** and **Importance**. There is only one Epistemic stance bundle (*the fact that the*), there are six bundles performing the function Obligation/ directive (*the Commission shall be, shall enter into force, shall ensure that the, on the Commission to, shall be subject to, the Member States shall*), and two that perform the function Importance (*it is necessary to, be taken into account*). The impersonal Epistemic bundle, also noted by Biber et al. (2004), expresses the degree of certainty, as in Example (10):

(10) *The fact that the conversion loss is not included gives biomass an unfair advantage over wind and solar energy.*

Obligation bundles give directions, as in Examples (11) and (12):

(11) *The Authority shall ensure that the public and any interested parties are rapidly given objective, reliable and easily accessible information, in particular with regard to the results of its work.*

(12) *Welcomes the idea of setting up an internal market for the people, goods and services of the creative industry, and calls on the Commission to present Parliament with a Green Paper on this subject;*

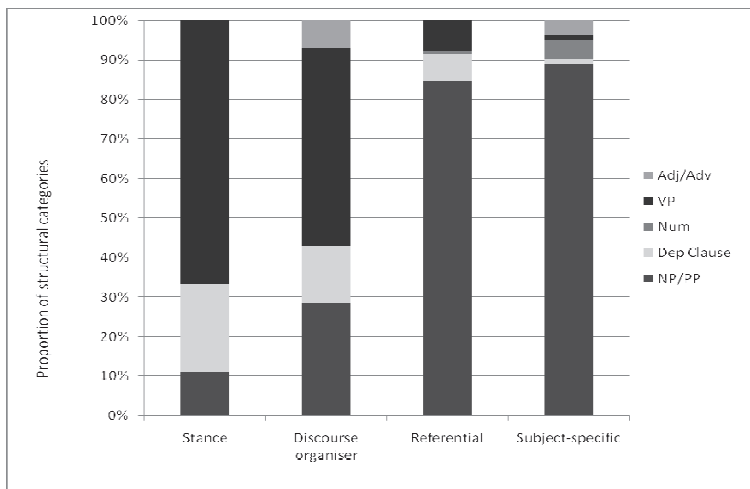


Figure 20. Interaction of structural and functional categories in the EEUD Corpus

Finally, Importance bundles mark how important the following proposition is, as shown in Example (13):

(13) *For a meaningful discussion on how to frame lobbying at EU level, it is **necessary to** define the basic framework on which the relationship between the EU institutions and lobbyists should be built.*

Regarding the most common structure of Stance bundles in the EEUD Corpus, these are most frequently expressed by verb phrases. As can be seen in Figure 20, 67% of Stance bundles are composed of verb phrases, and 22% contain dependent clause fragments.

#### 7.2.3.2.2. Discourse organisers in the EEUD Corpus

Findings show that the specific function of **Discourse organising** lexical bundles in the EEUD Corpus is **Topic elaboration and clarification**. These bundles are often used to provide additional sources, clarification or comparison, and contrast in the EEUD Corpus. Sentences



(14), (15) and (16) from the EEUD Corpus exemplify these uses of discourse organising bundles in an EU context:

(14) *In addition to the above considerations, the following should be noted:*

(15) *Recommendation 98/376/EC is amended as set out in the Annex.*

(16) *For the classification of fruit juices within the combined Nomenclature annexed to Regulation (EEC) No 2658/87, a distinction is to be made between, on the one hand, fruit juices containing added sugar of heading 2009 and, on the other hand, preparations for the manufacture of beverages including flavoured sugar syrups of heading 2106.*

Biber et al. (2004) found that discourse organising bundles are composed of all structural types in university registers. The pattern that emerges in EU discourse is similar as – except for numbers – four out of the five structural types are used in these bundles, as shown in Figure 20. The most frequent structural type of discourse organising bundles is lexical bundles with verb phrases. This structural type accounts for 50% of all discourse organising bundles. The proportion of lexical bundles with noun phrases and prepositional phrases is 29%, and there are a few bundles with dependent clause fragments, and adjectives and adverbs, with a proportion of 14% and 7% respectively.

#### 7.2.3.2.3. Referential bundles in the EEUD Corpus

The majority of lexical bundles identify entities, or give particular attributes, that is, they function as **Referential bundles** in different written registers. Referential bundles were found to be the largest category in textbooks, academic prose (Biber et al., 2004) and in a smaller EU corpus as well (Jablonkai, 2009a, 2009b). As regards their functions, most of the lexical bundles identified in this study also belong to the category of Referential bundles. As can be seen in Figure 21, Intangible framing bundles with 46 bundle types, Time bundles with 25 bundle types, Intertextual bundles with 22 bundle types, and Focus bundles with 21 bundle types are among the most common functional categories of referential lexical bundles in written EU discourse. Among the sub-categories of referential bundles Intertextual, Quality specification,

and Purpose bundles seem to be specific to EU texts, as these have not been identified in earlier analysis and they were added as new functional categories here. Lexical bundles belonging to these categories are described and illustrated in Section 7.2.3.2. Bundles in the new Intertextual category refer to other texts. The functional category **Text-deixis**, however, includes bundles that refer to different parts of the same text. Previous research has found text-deixis bundles frequently in textbooks and academic prose (Biber et al., 2004). Sentences (17) and (18) provide examples of text-deixis bundles in the EEUD Corpus:

(17) *The implementing rules for **paragraphs 1 and 2** shall be adopted by the Commission in accordance with the procedure referred to in Article 103(3).*

(18) *At the request of a Member State or of the Commission, or on its own initiative, the Board of Governors shall, in accordance with the same provisions as governed their adoption, interpret or supplement the directives laid down by it **under Article 7 of this Statute.***

Referential lexical bundles most often provide an intangible frame for propositions in the EEUD Corpus. These bundles specify abstract frames, and are used to establish logical frames and relationships in EU texts, as shown in Examples (19), (20) and (21):

(19) *However, as regards aid schemes **within the meaning of** Article 87 of the Treaty only, in addition to the conditions set out in the previous subparagraph, the public contribution corresponding to the expenditure included in a statement of expenditure shall have been paid to the beneficiaries by the body granting the aid.*

(20) *The European Parliament and the Council, **acting in accordance with** the ordinary legislative procedure, may lay down appropriate provisions for sea and air transport.*

(21) ***In the absence of** opposition, the European Council may adopt the decision.*

**Tangible framing bundles** describe concrete size and form in the EEUD Corpus, as in (22):

(22) *The Member State of refund shall take into account as a decrease on increase of the amount of the refund any correction made concerning a previous refund application in accordance with Article 13 or, where a separate declaration is submitted, **in the form of** separate payment or recovery.*

**Focus bundles** in the EEUD Corpus are used to identify sub-groups of people or countries that are in focus or identify specific pieces of EU legislation as particularly relevant, as shown in Examples (23), (24) and (25):

- (23) *Having regard to the Treaty establishing the European Economic Community and in particular Articles 28, 43, 113 and 235 thereof,*
- (24) *In its application, the Commission claims that the Court should annul the contested decision on the ground that it infringes Regulation No 1073/1999, in particular Article 4 thereof.*
- (25) *(c) for the Convergence objective only, the level of expenditure guaranteeing compliance with the additionality principle referred to in Article 15 and the action envisaged for reinforcing administrative efficiency as referred to in Article 27(4)(f)(i).*

The second most frequently performed function of lexical bundles in the EEUD Corpus is reference to time. Time is referred to by bundles as dates, periods, or in general as the end or time of certain events, as in Examples (26) and (27):

- (26) *(b) payments due under contracts, agreements or obligations that were concluded or arose before **the date on which** those accounts became subject to restrictive measures,*
- (27) *On this basis, the EU-Ukraine Action Plan was adopted in February 2005 **for a period of three years.***

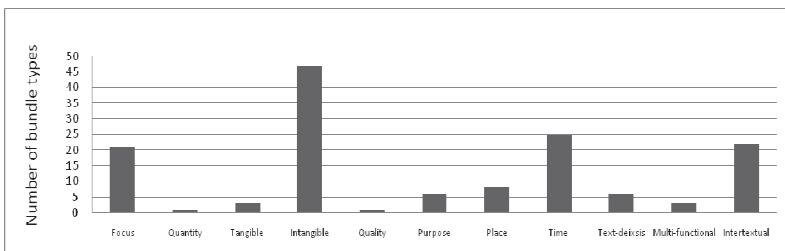


Figure 21. Distribution of Referential lexical bundle types across sub-categories in the EEUD Corpus

There are fewer bundles referring to the places of co-operation, or other activities, and the publication of different information, as in Sentence (28).

(28) *Extensions of the time limits set **out in Article 4(2), (3) and (5)** shall be adopted in accordance with the regulatory procedure with scrutiny referred to in Article 16(2).*

Biber et al. (2004) identified a functional sub-category of lexical bundles that serve more than one function in texts. Depending on the context these bundles refer to time, place, and/or

different parts of the text. There were only three lexical bundle types found to express several things in different contexts in the EEUD Corpus. Sentences (29) and (30) below show that the same lexical bundle refers to the date in the first example, and to the topic in the second:

(29) *The conference declares that the decision relating to the implementation of Article 16(4) of the Treaty on the Functioning of the European Union will be adopted by **the Council on the** date of the signature of the Treaty of Lisbon and will enter into force on the day that Treaty enters into force.*

(30) *The Commission will inform the European Parliament and **the Council on the** outcome.*

Finally, the only lexical bundle in EU discourse that refers to quantity, and therefore, classified under the **Quantity specification** category, is the one shown below in Example (31). More often than not quantity specification bundles refer to nouns from the field of finance like *cost*, *expenditure*, *refund*, *budgetary resources*, etc.:

(31) *The checks carried out for the period 2000-2006 shall cover at least **15 % of the total** eligible expenditure incurred on projects first approved during that period.*

Regarding the grammatical structures referential lexical bundles are composed of, the overwhelming majority, i.e., 85%, apply noun phrase and prepositional phrase fragments. As shown in Figure 20, nearly 8% apply verb phrase fragments, and a similar proportion of bundle types, that is 7%, incorporate dependent clause fragments. There is one single lexical bundle type that is composed of numbers. These findings are similar to the findings of Biber et al. (2004) on referential bundles in university registers, in that the majority of referential bundle types incorporate noun and prepositional phrases.

#### **7.2.3.2.4. Subject-specific bundles in the EEUD Corpus**

The functional category, **Subject-specific**, is a new category comprising of lexical bundles that are related to the subject field of the texts, in this case, to the European Union. There are four sub-categories based on the entities the lexical bundles refer to. Lexical bundles in the first and most common sub-category, **Organisations**, refer to countries and institutions,

especially, member states and EU bodies like the Commission, the Council, and the Economic and Social Committee, as shown in the Examples (32), (33) and (34) below:

(32) *Do you agree that the issue of multi-territory rights licensing must be addressed by means of a Recommendation of **the European Parliament and the Council**?*

(33) *However, interveners other than the **Member States and the institutions of the Union** may bring such an appeal only where the decision of the Civil Service Tribunal directly affects them.*

(34) *Having regard to the opinion of the **European Economic and Social Committee**,*

As shown in Figure 22, the second largest sub-category with 17 different bundle types is **Documents** including lexical bundles that refer to EU documents like treaties, the Official Journal or proposals as shown in Sentences (35) and (36):

(35) *They shall be published in the **Official Journal of the European Union** if the texts in the present languages were so published.*

(36) *HAVE AGREED upon the following provisions, which shall be annexed to the **Treaty establishing the European Community**:*

The sub-category, **Codes**, contains lexical bundles that specify certain articles and paragraphs of EU texts, as shown in Sentence (37):

(37) *Article **5 of the EC Treaty** thereof as interpreted by Protocol No 30 on the application of the principles of subsidiarity and proportionality annexed to the Treaty establishing the European Community, namely, its Point 9.*

Finally, there is a sub-category, **Other**, including lexical bundles expressing important EU-related concepts, as in Examples (38) and (39):

(38) *A. whereas, in accordance with the Treaties, the Community is called upon to play an active role in the field of health, whilst complying with **the principle of subsidiarity**,*

(39) *Monitoring report on the **implementation of commitments made in the accession negotiations by Latvia, 15 May 2003 Chapter 13.***

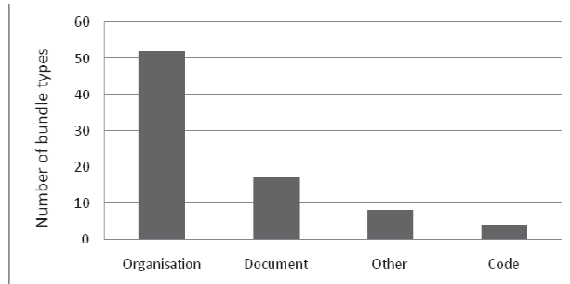


Figure 22. Distribution of Subject-specific lexical bundle types across sub-categories in the EEUD Corpus

The relationship of structure and function in the case of subject-specific bundles shows that a similar proportion of lexical bundles in the Subject-specific category operate with noun phrase and prepositional phrase fragments, as in the Referential bundles category, that is, 84% of subject-specific bundles use this grammatical structure. As findings revealed, there is only a single lexical bundle type with a verb phrase fragment, and one lexical bundle type with a dependent clause fragment, which indicates that subject-specific bundles do not frequently incorporate verb phrases and dependent clause fragments. As shown in Figure 20, there are, however, a few bundle types with adjectives and adverbs, and with numbers only.

### 7.2.3.3. Length of lexical bundles in the EEUD Corpus

In order to make findings comparable to previous findings of research on lexical bundles in various registers the present study – similarly to previous research in the field – focused on four-word lexical bundles. Several lexical bundles, however, were found to overlap or to be part of longer word combinations. Therefore, lexical bundles with an increasing number of words were identified in this study, applying the criteria that word combinations have to occur at least 20 times per million in at least 10% of the texts in the EEUD Corpus. The highest number is 12

words in bundles, because this is the longest cluster that *WordSmith Tools 4* can identify. The summary of the number of lexical bundles with different length is provided in Table 54.

The results of the analysis suggest that, although the number of bundle types decreases steadily with the increase in length, relatively long sequences with up to twelve words, are used repeatedly in EU texts. Concordances of these longer lexical bundles show that they are used frequently in the genres of secondary legislation, for example, regulations, decisions. This indicates that EU legal texts, especially, have the tendency to repeatedly use longer word sequences.

Length of lexical bundle types (in No. of words)	Number of lexical bundle types in the EEUD corpus	Examples
2	2232	<i>Member States, European Union, on the, in order,</i>
3	833	<i>accordance with, in accordance</i>
4	268	<i>in accordance with; with a view; set out in</i>
5	103	<i>having regard to the; in accordance with the; with a view to on the implementation of the; the procedure referred to in the Commission and the Member States; its publication in the Official Journal; shall enter into force on the</i>
6	61	<i>in accordance with the principle of proportionality;</i>
7	48	<i>does not go beyond what is necessary</i>
8	39	<i>published in the Official Journal of the European; having regard to the proposal from the Commission the opinion of the European Economic and Social Committee;</i>
9	33	<i>in accordance with the procedure laid down in article the council of the European Union having regard to the; the treaty establishing the European Community and in particular article</i>
10	27	<i>regard to the opinion of the European Economic and Social Committee; regard to the treaty establishing the European Community and in particular</i>
11	22	<i>having regard to the treaty establishing the European Community and in particular;</i>
12	19	<i>to the proposal from the Commission having regard to the opinion of</i>

Table 54. Frequency and examples of lexical bundles with different length in the EEUD Corpus

#### 7.2.3.4. Limitations of the analysis of lexical bundles

It should be noted that the large number of different bundle types in the EEUD Corpus could be the result of the wide variety of different EU genres that are included in the corpus. Therefore, further research should concentrate on (1) how the lexical bundles identified in this study are distributed in the different EU genres, and (2) what lexical bundles are specific to individual EU genres. Such analyses may provide a more realistic overall picture of the lexical patterns in English EU documents.



Limitations of the analysis of lexical bundles in the EEUD Corpus also concern the comparisons of results to earlier findings of lexical bundles in other registers. Lexical bundles are defined in a similar, but not exactly the same way, in these earlier analyses of different registers. Some researchers define the frequency cut-off point at 20 per million words (Biber et al., 2004; Hyland, 2008), others use a more conservative 40 per million words criterion (Biber & Barbieri, 2007; Jablonkai, 2009a, 2009b). Although these differences make the comparisons less accurate, findings still suggest meaningful tendencies.

#### **7.2.3.5. Conclusions concerning the analysis of lexical bundles**

This section presented the findings of the corpus-driven analysis of four-word lexical bundles in the EEUD Corpus. Although the investigation of lexical bundles in terms of their structures and functions revealed some similarities with the language of textbooks and academic prose, in view of the findings, written English EU discourse may be assumed to differ from other written registers in a few important aspects. Firstly, as regards the structure of lexical bundles in the EEUD Corpus, results show that English EU texts make use of a higher number of bundles with verb phrases. Secondly, lexical bundles appear in fairly high frequencies in the EEUD Corpus, which suggests that a substantial proportion of English EU texts consist of formulaic patterns. Finally, the results also indicate that four-word lexical bundles in the EEUD Corpus are often part of longer word combinations.

#### 7.2.4. Lexis in the EEUD Corpus

On the whole, the overall analysis of written English EU discourse yielded relevant findings for a clearer understanding of the use of English in an EU context. The most important findings on the lexical and lexicogrammatical features of English EU documents are summarised in Tables 55, 56 and 57. Firstly, the results concerning the **lexis** of English EU texts show that 75% of all tokens in the EEUD Corpus are covered by the GSL, that is, the first most frequent two thousand word families in English. This corresponds, in general, to findings of earlier analyses of specialised texts of other disciplines (Chung & Nation, 2003; Nation & Hwang, 1995; Sutarsyah et al., 1994), in which the GSL has been found to cover 70-75% of specialised texts of these disciplines. Word families of the AWL – representing the semi-technical lexis used frequently in texts of several academic disciplines – account for almost 14% of the tokens in the EEUD Corpus. A higher proportion of the tokens, i.e. 18%, is covered by the EUWL established as a result of the present study. Although there is a certain overlap between these two word lists, there are 190 word families in the EUWL that can be considered EU-specific. These EU-specific lexical items can be considered the frequent technical lexical items in written English EU discourse, which cover slightly more than 4% of the texts in the EEUD Corpus.

Focus of analysis	Results	Pedagogical applications
<i>Lexis in the EEUD Corpus</i>	<i>EU Word List</i>	
aim:	<ul style="list-style-type: none"> <li>• 513 word families</li> <li>• 2,457 lexical items</li> </ul>	<ul style="list-style-type: none"> <li>• starting point for course and teaching materials design for English for EU purposes</li> <li>• frequency order can serve as guidance for sequencing the teaching of lexis</li> </ul>
<ul style="list-style-type: none"> <li>• identify lexical items that are strongly associated with EU discourse</li> </ul>	<ul style="list-style-type: none"> <li>• legal words, EU institutions, words in connection with funding, geographical names, abbreviations</li> <li>• coverage of EU texts: around 18%</li> <li>• 190 EU-specific word families</li> <li>• GSL and EUWL together cover 94% of EU texts</li> </ul>	

Table 55. Summary of findings relating to the frequent lexical items in the EEUD Corpus

Secondly, the **collocational analysis** of selected lexical items of the EUWL shed some light on what makes English EU documents examples of **hybrid texts**. Hybridity refers to the characteristic of a text that it exhibits “features that somehow seem ‘out of place’/‘strange’/‘unusual’ for the receiving culture” (Schäffner & Adab, 2001a, p. 175). The collocational analysis revealed several differences between the collocates the selected lexical items typically co-occur with in general English texts, and in EU texts. Furthermore, the analysis found that there are also differences in the grammatical relations these lexical items frequently form in the EEUD Corpus and in the BNC Written. The different collocates or frequent untypical grammatical behaviour of lexical items might be perceived as ‘strange’ or ‘unusual’ features of EU texts, therefore, these features can be considered as elements contributing to the hybridity of English EU documents. Furthermore, findings of the collocational analysis revealed that in general lexical items form a greater variety of grammatical relations in the EEUD Corpus. The number of collocates, and the variety of semantic preferences the selected lexical items appear to have, however, is greater in the BNC Written, which suggests a certain degree of fixedness in the lexical aspects of collocation in the EEUD Corpus. As regards semantic preferences, results show that there are EU-specific semantic preferences, but there are also semantic preferences that lexical items share in the two corpora.

Focus of analysis	Results	Pedagogical applications
<p><i>Collocational analysis</i></p> <p>aim:</p> <ul style="list-style-type: none"> <li>• comparison of individual lemmas in the EEUD and the BNC Written;</li> <li>• provide information on relevant collocates and semantic preferences of lemmas</li> </ul>	<ul style="list-style-type: none"> <li>• most lemmas exhibit a greater variety of grammatical relations in the EEUD Corpus</li> <li>• proportion of collocates covered by semantic preferences is higher in the EEUD Corpus</li> <li>• fewer preferential semantic sets identified in the EEUD Corpus</li> <li>• shared and EU-specific semantic preferences</li> <li>• semantic prosody is not characteristic of lemmas in the EEUD Corpus</li> </ul>	<ul style="list-style-type: none"> <li>• <i>pedagogic collocational profile</i> of lexical items</li> <li>• source for tasks on usage of lexical items</li> </ul>

Table 56. Summary of findings relating to the collocational analysis of selected lemmas in the EEUD Corpus

Thirdly, the frequency-based analysis of MWIs in the EEUD Corpus seems to indicate that written English EU discourse applies a large number of four-word **lexical bundles**, and the frequency of these lexical bundles is higher than the frequency of lexical bundles in other registers, which suggests that the language use in written English EU discourse is rather formulaic. Regarding the structural distribution of bundles, the investigation revealed that most bundles in the EEUD Corpus incorporate noun and prepositional phrases, and that there is a relatively high proportion of bundles with verb phrases. The functional distribution of bundles in the EEUD Corpus suggests that lexical bundles most frequently perform referential functions, such as Intangible framing, Time, Intertextual and Identification. The second largest group of bundles is the Subject-specific category, with most lexical bundles referring to EU-related countries and institutions. The interaction of structural and functional characteristics shows that the majority of referential and subject-specific bundles incorporate noun phrases and prepositional phrases, whereas the majority of stance bundles are composed of verb phrases in the EEUD Corpus.

Focus of analysis	Results	Pedagogical applications
<p><i>Lexical bundles</i></p> <p>aim:</p> <ul style="list-style-type: none"> <li>characterise EU discourse by its frequent MWIs</li> </ul>	<p><i>List of lexical bundles</i></p> <ul style="list-style-type: none"> <li>247 lexical bundle types</li> <li>79 EU-specific lexical bundle types</li> <li>more types of lexical bundles and higher frequency of lexical bundles in EEUD Corpus than in other registers</li> </ul> <p>Structural characteristics:</p> <ul style="list-style-type: none"> <li>74% of lexical bundles incorporate noun phrase and prepositional phrase fragments</li> <li>30% of lexical bundles with <i>of</i>-phrase fragments</li> <li>relatively high proportion (10%) of lexical bundles incorporate verb phrase fragments</li> </ul> <p>Functional characteristics:</p> <ul style="list-style-type: none"> <li>lexical bundles are not likely to be used to express stance and do not often serve discourse organising functions</li> <li>58% of lexical bundle types function as referential bundles</li> <li>most frequently lexical bundles serve the following referential functions: Intangible framing, Time reference, Intertextual, Identification/ focus</li> <li>a third of lexical bundle types express subject-specific entities</li> <li>most of the subject-specific lexical bundles refer to countries and institutions</li> </ul> <p>Interaction of structural and functional characteristics:</p> <ul style="list-style-type: none"> <li>85% of referential and 89% of subject-specific bundle types incorporate noun phrase and prepositional phrase fragments</li> <li>67% of stance bundle types and 50% of discourse organising bundle types incorporate verb phrase fragments</li> </ul>	<ul style="list-style-type: none"> <li>explicit instruction of relevant lexical bundles help understand and produce EU texts</li> <li>list of EU-specific four-word lexical bundles can serve as starting point for teaching materials design</li> </ul>

Table 57. Summary of findings relating to lexical bundles in the EEUD Corpus

## **Chapter 8: Implications for teaching English for EU purposes**

The results of the present study have important implications for the teaching practice of English for EU purposes, in general, and for course and materials design, in particular. Firstly, findings on the use of English within the EU context suggest that there are several aspects of this particular language variety that future EU professionals should be prepared for. Secondly, the needs analysis survey, conducted in this study, resulted in a list of EU genres and particular documents that EU professionals use in their daily work and therefore, can represent the language use future EU professionals should be prepared for. Consequently, this list also served as the basis for the EEUD Corpus. Finally, the findings of the needs analysis also identified what these EU documents are generally used for by EU professionals in EU institutions.

On the basis of these findings, this chapter will put forth some recommendations for the application of the EEUD Corpus for teaching purposes and it will also focus on the analysed aspects of written English EU discourse with a practical stance by highlighting relevant areas for instruction and proposing sample tasks for the practice of teaching.

As presented in Table 58, the needs analysis survey resulted in the EEUD Corpus that served as the basis for the analysis of written English EU discourse. The EEUD Corpus contains 241 different English EU texts representing 40 different written EU genres used within the EU context. These genres (e.g., treaties, regulations, decisions, presidency conclusions, press releases) may be regarded as a list of common core EU genres that can serve as a starting point for a genre-based approach to teaching English for EU purposes. It should be noted, however, that the list reflects the genres that are used by Hungarian EU professionals. The results of the

needs analysis survey cannot be considered fully representative of either Hungarian EU professionals or EU professionals in general, as there are no precise statistics available on the number of professionals working in EU-related jobs. The resulting EEUD Corpus may be claimed to be balanced for EU subject fields, therefore, it can provide a basis for analysing the characteristics of written English EU discourse in general. Although the EEUD Corpus cannot be considered fully representative, findings relating to its linguistic, especially, lexical and lexicogrammatical aspects can be generalised to some extent as the comparison of findings of the lexical analysis from the EEUD Corpus to findings of the lexical analyses of other corpora of English EU texts showed considerable similarity (e.g., in the text coverage of the EUWL and the structural and functional distribution of lexical bundles). In addition to further linguistic analysis, the EEUD Corpus can also be used for teaching purposes not only as a source for developing paper-based teaching materials, but also as a source for concordance lines for DDL activities. Later in this chapter examples of such tasks will be presented.

Focus of analysis	Results	Pedagogical applications
<i>Needs analysis</i>	<i>English EU Discourse Corpus</i>	
aim:	<ul style="list-style-type: none"> <li>• around 1 million running words</li> </ul>	<ul style="list-style-type: none"> <li>• further analysis of written English EU discourse for pedagogic purposes</li> </ul>
<ul style="list-style-type: none"> <li>• identify relevant EU genres and texts for future EU professionals</li> </ul>	<ul style="list-style-type: none"> <li>• 40 different EU genres</li> <li>• 241 EU texts</li> <li>• fairly balanced for EU subject fields</li> </ul>	<ul style="list-style-type: none"> <li>• source for developing teaching materials</li> <li>• source for electronic DDL activities</li> </ul>

Table 58. Summary of findings relating to the needs analysis survey

Furthermore, the needs analysis also provided insights into what purposes English EU documents are used for by EU professionals. The findings suggest that EU professionals most often scan the documents for specific information or skim them for general information. In addition, specific EU terms are also searched for in particular EU documents and these documents are often applied as templates for writing. These uses can be applied in the teaching practice providing meaningful tasks for the ESP classroom.

The analysis of the EEUD Corpus also identified the lexical items especially associated with written English EU discourse in the form of the EUWL with 513 word families. The words in the EUWL are not specific to any one subject field of the EU's activities. The evaluation of the EUWL also demonstrated that the EUWL comprises word families that are used in a wide range of EU texts. Therefore, it can serve as reference for course and materials design for teaching English for EU purposes. At the same time, the EUWL provides guidelines for the sequencing of the teaching of lexical items as teaching can follow the frequency order of word families in the list. With the help of the EUWL, the EU-specific elements can easily be selected and can be used as the basis for traditional lexis teaching exercises and also for DDL activities focusing on the lexical and lexicogrammatical patterns specific to written English EU discourse. The mere list of EU-related lexis can be supplemented with information on the frequently used patterns of individual lexical items that can be explicitly taught to language learners as suggested by the lexical approach to language teaching (Lewis, 1993). Based on the results of the collocational analysis, such information can be provided in a straight-forward manner in the form of **pedagogic collocational profiles**. As shown in Table A1, Table B1 and in Appendix 10, this profile not only gives language learners guidance on relevant collocates, but it also presents frequent semantic preferences and constructions, that is, the grammatical relations the particular lemma frequently forms with relevant collocates.

In what follows, a few task types will be presented to illustrate the way findings of the present corpus analysis can be applied in teaching:



## Task type A

**Aim:** to raise learners' awareness of collocates of particular lexical items

**A.1 Instruction:** Study the collocational profile of the verb IMPLEMENT in Table A1 and underline the nouns in the table that are likely to be used with it in EU documents.

the accession criteria	function
opinion	the acquis
a reform	measures
the internal market	a directive
legislation	a summit
a timetable	policies
a programme	a debate

<b>IMPLEMENT verb</b>	
<b>construction</b>	<b>semantic groups</b>
IMPLEMENT + noun	<p>1. <u>legislation</u> collocates: <i>measure, rule, regulation, provision, directive, legislation, recommendation, decision, convention</i> The Commission shall <b>implement</b> this Regulation in accordance with the Financial Regulation.</p> <p>2. <u>plans</u> collocates: <i>reform, strategy, programme, project, policy, commitment, budget, plan</i> Many European policies and programmes are <b>implemented</b> at regional and local levels.</p> <p>3. <u>approach</u> collocates: <i>approach, principle</i> The forthcoming proposal for a new Directive <b>implementing</b> the principle of equal treatment outside employment will be addressed.</p> <p>4. <u>activity</u> collocates: <i>action, tool, operation</i> By way of derogation from paragraph 1 , in-kind contributions , depreciation costs and overheads may be treated as expenditure paid by beneficiaries in <b>implementing</b> operations under the following conditions:</p>
IMPLEMENT + adverb	<p>1. <u>positive</u> collocates: <i>properly, effectively, fully, successfully, actively</i> The Commission, in its role of guardian of the Treaty, is responsible for ensuring that Community legislation is properly transposed into national law and properly <b>implemented</b> and enforced by national authorities in the Member States.</p> <p>2. <u>negative</u> <i>not</i> However, Albanian legislation does not yet protect these rights sufficiently and is not fully <b>implemented</b>. Other collocates: <i>systematically, as, directly</i></p>

Table A1. Extract from the collocational profile of the verb IMPLEMENT (see full pedagogic collocational profile in Appendix 10)

**A.2 Instruction:** Study the collocational profile of the verb IMPLEMENT in Table A1 and add five more nouns that are often used together with it in EU documents?

- a.
- b.
- c.
- d.

**Key to Task type A.1:**

the accession criteria	function
opinion	the acquis
<b>a reform</b>	<b>measures</b>
the internal market	<b>a directive</b>
<b>legislation</b>	a summit
a timetable	<b>policies</b>
<b>a programme</b>	a debate

**Key to Task type A.2:** a. rule, b. convention, c. plan, d. action (for further examples see Table A1)

## Task type B

**Aim:** to help learners identify and use frequent collocates of particular lexical items in context

**B.1 Instruction:** Study the concordance lines and add an appropriate adverb to the sentences below.

storage, transport and retail sale, will be RTD activities. An ethical review will be to ensure that Albania can properly fight against money laundering if properly provisions of the agreement are being properly laws. (2) For the purposes of effectively The new strategy needs to be effectively transposed in a timely manner and effectively capacity building to design and effectively relationship, the Union will strive to fully Presidencies will make every effort to fully guarantees on freedom of expression are not yet these rights sufficiently and is not fully on a business permit system is not fully determined to pursue reforms and to fully partner in ensuring that they are successfully represents a key instrument for successfully	<b>implemented</b> <b>implemented</b> <b>implement</b> <b>implemented</b> <b>implemented</b> <b>implementing</b> <b>implemented</b> <b>implement</b> <b>implement</b> <b>implemented</b> <b>implemented</b> <b>implement</b> <b>implemented</b> <b>implemented</b> <b>implemented</b> <b>implementing</b>	systematically and in a consistent manner systematically for proposals dealing with the new public procurement legislation . Some progress has been made on upgrading . Where the Agency and/or the Commission the relevant provisions of the Amsterdam and monitored. The existing legislation by Member States. The process of evaluating sound trade and integration policies, as the Joint EU-Africa Strategy as well as the Action Plan on Simplification, for fully, particularly regarding the print . As regards children's rights , the government , because, mainly, of the lack of implementing the work programme, in particular by . This plan will be presented before the EPAs, and whereas regional integration
--	--	---

- a. This reform could significantly enhance the fight against money laundering if \_\_\_\_\_ **implemented**.
- b. Member States are determined to pursue reforms and to \_\_\_\_\_ **implement** the work programme.
- c. The 2006 law on a business permit system is not \_\_\_\_\_ **implemented**, because, mainly, of the lack of implementing regulations.
- d. The new strategy needs to be \_\_\_\_\_ **implemented** and monitored.
- e. In this way, this policy covering all sectors of the food chain will be **implemented** \_\_\_\_\_ and in a consistent manner.

### Key to Task type B.1:

- a. This reform could significantly enhance the fight against money laundering if properly implemented.
- b. Member States are determined to pursue reforms and to fully implement the work programme
- c. The 2006 law on a business permit system is not fully implemented, because, mainly, of the lack of implementing regulations.
- d. The new strategy needs to be effectively implemented and monitored.
- e. In this way, this policy covering all sectors of the food chain will be implemented systematically and in a consistent manner.

**B.2 Instruction:** Study the concordance lines and list verbs that are likely to be used with the noun CRITERION? Compare your answers with the verbs listed in the collocational profile of the word in Table B1.

<p>Have any additional eligibility evidence that they comply with the selection on excellence and in accordance with the Commission. 30. Experience has shown that the checks can be considered as fulfilling the adopt the list of regions fulfilling the and 2 and of Member States fulfilling the the NUTS level 2 regions fulfilling the correct, that the estimated costs meet the identifying other areas of crime that meet their entirety, subject to the necessary no longer satisfy the regional eligibility eligible provided that they satisfy the 2003 of 18 February 2003 establishing the verification of application of the selection such tasks are carried out in line with the explanation of how it had regard to the allocation divide"; stresses the need to apply quality restrictive manner, the Court has applied these Member States apply different basic must be granted on the basis of transparent can demonstrate its conformity with the or under its responsibility according to the NUTS level 2 regions fulfilling the shall be amended in accordance with the worldwide aim at respecting sustainability concerns 20. Asks the Member States to respect decision, the Commission shall respect the</p>	<p><b>criteria</b> been set for this call? Check that you  <b>critereion</b> set out in the Article 176 of the  <b>criteria</b> set by the Governing Board of the EIT;  <b>criteria</b> set out in sectoral Directives which conformity  <b>criteria</b> needed to count toward the minimum number  <b>criteria</b> under paragraph 1 and of Member States  <b>criteria</b> under paragraph 3. This list shall be valid  <b>criteria</b> laid down in Article 2 of Protocol No 6  <b>criteria</b> for eligible as established in the call  <b>criteria</b> specified in this paragraph. It shall act  <b>criteria</b> being met. 34. Declaration on Article 179  <b>criteria</b> of the Convergence objective and which  <b>criteria</b> set out in the previous paragraph:  <b>criteria</b> and mechanisms for determining the Member  <b>criteria</b> established by the monitoring committee  <b>criteria</b> established for such tasks. Existing procedures  <b>criteria</b> established in accordance with paragraph  <b>criteria</b> for health-related websites; 18. Encourages  <b>criteria</b> of legality, legitimacy and necessity in  <b>criteria</b> for establishing whether a food is safe  <b>criteria</b> applied in a non-discriminatory way to  <b>criteria</b> laid down in the harmonised standards,  <b>criteria</b> laid down by the monitoring committee and  <b>criteria</b> laid down in Article 2 of Protocol No 6  <b>criteria</b> laid down in Article 3(6). 4. With effect  <b>criteria</b> in investing policies. Aid effectiveness  <b>criteria</b> of legal clarity and legal security for  <b>criteria</b> referred to in paragraph 3. Article 6 The</p>
--	---

**Key to Task type B.2:** fulfil, meet, set, respect, apply (for further examples see Table B1).

<b>2. CRITERION noun</b>	
<b>construction</b>	<b>semantic group</b>
verb + CRITERION	<p>1. <u>meet</u> collocates: <i>fulfil, fulfill, meet, satisfy, conform to</i> conforms to the <b>crit</b>erion of independence defined below</p> <p>2. <u>set</u> collocates: <i>set, agree, establish, lay</i> Unlike Article 60(1)(c) of Regulation (EC) No 44/2001, which <b>estab</b>lishes three <b>crit</b>eria, the conflict-of-laws rule should proceed on the basis of a single criterion;</p> <p>3. <u>respect</u> collocates: <i>respect</i></p> <p>4. <u>list</u> collocates: <i>list, specify, give</i> Paragraph 3 <b>spec</b>ifies the <b>crit</b>eria that may be used by the courts to decide whether it should apply the mandatory provisions of another Member State.</p> <p>5. <u>apply</u> collocates: <i>apply</i> Other verb collocates: <i>need, see, base, propose</i></p>
noun + CRITERION	<p>1. <i>award, selection, evaluation, quality</i> (a) the operation meets the <b>sel</b>ection <b>crit</b>eria for the operational programme</p> <p>2. <i>performance, efficiency</i> The payment for the services delivered is based on the meeting of the agreed <b>per</b>formance <b>crit</b>eria. Other noun collocates: <i>eligibility, convergence, sustainability, allocation, risk</i> The <b>eli</b>gibility <b>crit</b>eria are given in the work programme.</p>
adjective + CRITERION	<p>1. <i>basic, minimum</i> However, these Member States apply different <b>bas</b>ic <b>crit</b>eria for establishing whether a food is safe.</p> <p>2. <i>same, common</i> Any common price policy shall be based on <b>com</b>mon <b>crit</b>eria and uniform methods of calculation.</p> <p>3. <i>political, technical, environmental, economic, cultural</i> Meetings focused on <b>pol</b>itical accession <b>crit</b>eria and rule of law issues. Other adjective collocates: <i>transparent, objective, indicative, relevant, certain, general</i> The plan shall be based on objective and <b>tra</b>nsparent <b>crit</b>eria, including those listed in Annex III.</p>
CRITERION + for	<p>1. <u>evaluation</u> collocates: <i>selection</i></p> <p>2. <u>membership</u> collocates: <i>membership</i> ... and of the need for further significant progress to respond to the other issues and <b>crit</b>eria for <b>mem</b>bership included in the Commission 's Opinion ...</p> <p>3. <u>distribution of funds</u> collocates: <i>allocation</i></p>
mostly in plural	

Table B1. Collocational profile of the noun CRITERION

### Task type C

**Aim:** to give practice with frequent collocates of particular lexical items

**Instruction:** Study the collocational profile in Table B1 and supply the missing words. The same word should be used in one set of concordance lines.

C1. missing word: \_\_\_\_\_

shall be based on objective and \_\_\_\_\_ **criteria** , including those listed in Annex III, taking programmes. 2.2 Objective and \_\_\_\_\_ **criteria** for resource allocation 64. Within global objective and \_\_\_\_\_ resource allocation **criteria** based on needs and performance will guide consistent basis and in line with \_\_\_\_\_ **criteria** , as part of integrated impact assessments must be granted on the basis of \_\_\_\_\_ **criteria** applied in a non-discriminatory way to be selected on the basis of \_\_\_\_\_ **criteria** . Information provided by the UN Task Force

C2. missing word: \_\_\_\_\_

work programme, to check the \_\_\_\_\_ **criteria** and any other additional conditions that your proposal eligible? The \_\_\_\_\_ **criteria** are given in the work programme. See also consortium. Have any additional \_\_\_\_\_ **criteria** been set for this call? Check that you that your proposal meets the \_\_\_\_\_ **criteria** that apply to this call and funding scheme measures which fall within the \_\_\_\_\_ **criteria** and main scope of, or receive assistance applications that fulfil the \_\_\_\_\_ **criteria** will be considered for a grant. If an application financial requirements, such as \_\_\_\_\_ **criteria** and financial capacity, with regard to SELECTION CRITERIA The following \_\_\_\_\_ **criteria** define the scope of the call and apply

C3. missing word: \_\_\_\_\_

situation in Albania in terms of the political **criteria** for \_\_\_\_\_ ; - analysis the situation in Albania on the basis of the economic **criteria** for \_\_\_\_\_ ; - reviews Albania's capacity progress to respond to the other issues and **criteria** for \_\_\_\_\_ included in the Commission

C4. missing word: \_\_\_\_\_

no longer \_\_\_\_\_ the regional eligibility **criteria** of the Convergence objective and which by LIFE+ if they \_\_\_\_\_ the eligibility **criteria** set out in Article 3: (a) operational activities eligible provided that they \_\_\_\_\_ the **criteria** set out in the previous paragraph

**Key to Task type C:** 1. transparent, 2. eligibility, 3. membership, 4. satisfy, meet

The distributional patterns and general characteristics of lexical bundles in the EEUD Corpus also have several implications for ESP pedagogy. Firstly, the structural analysis of lexical bundles found that certain verb phrases, and noun and prepositional phrases, are prevalent in the EEUD Corpus, therefore, these grammatical structures, for example, noun phrases with *of*-phrase fragment, prepositional phrases and verb phrases with passive verbs, should be given more emphasis in the teaching practice. Secondly, the functional analysis revealed that nearly a third of the lexical bundles refer to EU-specific entities. Therefore, they can be used to compile glossaries of useful terms for students. Finally, it was found generally that lexical bundles occur very frequently in the EEUD Corpus. Therefore, explicit instruction in these recurrent word combinations may increase the efficiency of courses and teaching materials of English for EU purposes. The explicit teaching of lexical bundles in written English EU discourse should (1) raise students' awareness of lexical bundles, (2) focus on the function of lexical bundles, and (3) provide examples and practice of lexicogrammatical patterns of lexical bundles (Cortes, 2004; Cortes, 2006; Neely & Cortes, 2009; Trebits, 2009a). In awareness raising tasks, students should be provided with excerpts from EU texts with the lexical bundles highlighted, as illustrated in Task type D. Based on these excerpts, students can draw conclusions on the relevance of lexical bundles in EU texts, and can also identify the functions of the specific bundles. The next step should be to expand on the different functions lexical bundles can have in EU texts. Most bundles in the EEUD Corpus appear to make reference to physical or abstract entities as referential bundles, therefore tasks concentrating on referential bundles should be included in a syllabus of English for EU purposes. Task type E and F present two ways to provide students practice with referential bundles. In Task type E, students are given concordance lines of specific bundles and they are asked to identify the functions these bundles have in the given contexts. In a follow-up "fill-in-the-blank" exercise, as illustrated in Task type F,

students should complete sentences with the lexical bundles that perform the functions specified in brackets. Finally, concordance lines can also be used to provide students with examples of the usage patterns of certain lexical bundles. Task type F focuses on the most frequent lexical bundle type in the EEUD Corpus, that is, *in accordance with the*. Students should study how this lexical bundle is used in a sentence, what words are likely to precede and follow it, and where they are usually placed within the sentence. On the following pages, these task types provide examples of how the findings of the present study can be applied in the ESP classroom.



### Task type D

**Aim:** to raise learners' awareness of lexical bundles and to identify their functions in EU documents

**Instruction:** Examine the following excerpts from two different EU documents. The highlighted expressions are lexical bundles, which are frequently used in many EU documents. Can you identify their functions in the context?

Excerpt 1

The European Parliament,

[...]

– having regard to Rule 51 of **its Rules of Procedure**,

– **having regard to the** report of the Committee on Transport and Tourism (A6 0038/2005),

1. Approves the Commission proposal as amended;
2. Calls on the Commission to refer the matter to Parliament again if it intends to amend the proposal substantially or replace it with another text;
3. Instructs its President to forward its position **to the Council and Commission**.

Excerpt 2

[...]

With regard to Article 60 of Regulation (EC) No 1083/2006 and **in the light of** the experience gained, **it is necessary to** lay down the obligations which the managing authorities should have with regard to beneficiaries in the phase leading to the selection and approval of the operations to be funded, **with regard to the** aspects which the verifications of the expenditure declared by the beneficiary should cover, including administrative verifications of the applications for reimbursement, and on-the-spot verifications of individual operations and **with regard to** the conditions to be observed when on-the-spot verifications are carried out on a sample basis.

[...]

### Key to task type D:

*with regard to the, in the light of:* express logical relations in text (Intangible framing bundles)

*having regard to the:* refers to other texts (Intertextual bundles)

*it is necessary to:* expresses importance

*its Rules of Procedure, to the Council and, Council and the Commission:* refer to EU related entities

### Task type E

**Aim:** to practise identifying the function of lexical bundles in EU documents

**Instruction:** Study the concordance lines of the lexical bundles and identify what they make reference to. Choose from the list below:

Time, Place, Purpose, Quality, Reference to another document

E1) Reference to: \_\_\_\_\_

COUNCIL OF THE EUROPEAN UNION, Union and in particular Article 15 thereof,	<b>having regard to the</b>	Treaty on the European Union and in conclusions of the European Councils
COUNCIL OF THE EUROPEAN UNION, regard to Rule 45 of its Rules of Procedure, –	<b>having regard to the</b>	Treaty on European Union and in particular report of the Committee on Culture and
of the Diversity of Cultural Expressions, –	<b>having regard to the</b>	Commission communication on a European
the Diversity of Cultural Expressions, 2005,	<b>having regard to the</b>	Council conclusions of 13 and 14 November

E2) Reference to: \_\_\_\_\_

may not be made out more than 14 days before	<b>the date on which</b>	the plants, plant products or other objects leave
shall not cover expenditure incurred before	<b>the date on which</b>	the Commission received the application.
obligations that were concluded or arose before	<b>the date on which</b>	those accounts became subject to restrictive
tasks may under no circumstances begin before	<b>the date on which</b>	the specific contract enters into force. I.2.3
may implementation commence before	<b>the date on which</b>	the Contract enters into force. Execution of the
obligations that were concluded or arose before	<b>the date on which</b>	those accounts became subject to restrictive

E3) Reference to: \_\_\_\_\_

is a crucial factor to achieve	<b>a high level of</b>	social cohesion. The Community will
out by Member States should add	<b>a high level of</b>	credibility to the annual accounts of
safety framework, which can deliver	<b>a high level of</b>	public health and consumer
The Union shall endeavour to ensure	<b>a high level of</b>	security through measures to prevent
animal feed production, establish	<b>a high level of</b>	consumer health protection and

### Key to Task type E:

- 1) Reference to another document
- 2) Reference to time
- 3) Reference to quality

### Task type F

**Aim:** practise the use of lexical bundles in context

**Instruction:** Complete the sentences with a suitable lexical bundle from the list. The functions the bundles have in the sentences are given in brackets. Choose the appropriate ones from the list below.

*a high level of, for the purposes of, in the Official Journal, the date on which, having regard to the*

F1)

The principal objective of a European Food Authority will be to contribute to \_\_\_\_\_ (quality) consumer health protection in the area of food safety, through which consumer confidence can be restored and maintained.

F2)

Notwithstanding the results of any audits performed by the Commission or the European Court of Auditors, the final balance paid by the Commission for the operational programme may be amended within nine months of \_\_\_\_\_ (time) it is paid or, where there is a negative balance to be reimbursed by the Member State, within nine months of \_\_\_\_\_ (time) the debit note is issued.

F3)

The Council of the European Union,

[...]

\_\_\_\_\_ (reference to another document) Treaty establishing the European Community, and in particular Article 269 thereof,

\_\_\_\_\_ (reference to another document) opinion of the Court of Auditors,

\_\_\_\_\_ (reference to another document) opinion of the Economic and Social Committee,

Whereas: ...

#### Key to Task type F:

F1) a high level of

F2) the date on which, the date on which

F3) Having regard to the

Having regard to the

Having regard to the

The lexical bundles *for the purposes of* and *in the Official Journal* are not used in any of the sentences.

## Task type G

**Aim:** to identify patterns of use of particular lexical bundles

**Instruction:** Study the concordance lines containing the lexical bundle *in accordance with the*. Notice the words and phrases that precede and follow this lexical bundle. Can you identify patterns? What are possible positions of this lexical bundle in a sentence?

paragraph 1 shall also be used to establish, to be drawn up by the Commission which the Commission has conducted Member State concerned, it may be decided, established in an implementing Regulation of the financial contribution may be approved, above-mentioned period may be extended, an implementing Regulation may specify, and the Commission are in agreement, referred to in paragraph 1 shall be adopted shall not vote. 3. Where the measures are them forthwith. Where the measures are not Member States of this information each year. s they relate to plant-health checks carried out European Parliament and the Council, acting European Parliament and the Council, acting shall adopt the list of Members drawn up European Parliament and the Council, acting Union and to rule on proceedings, brought favour certain undertakings or activities shall be supplemental to domestic action, level, the Community may adopt measures, concerned shall endeavour to reach agreement at Union level shall be implemented either participants in Community programmes present a proposal to improve this Regulation the Commission shall adopt a Regulation	<b>in accordance with the in accordance with the in accordance with the in accordance with the in accordance with the In accordance with the in accordance with the in accordance with the In accordance with the In accordance with the in accordance with the in accordance with the in accordance with the in accordance with the in accordance with the in accordance with the in accordance with the in accordance with the in accordance with the</b>	Treaty, whether the Member State from which procedure laid down in Article 18. Article 24 procedures analogous to those in Article 39 of same procedure, that the Community financial procedure laid down in Article 18. 6. In the same procedure, depending on the outcome of same procedure, if examination of the procedure laid down in Article 18, cases in procedure laid down in Article 18, this procedure laid down in Article 18. 3. For the opinion of the Committee, the Commission opinion of the Committee or if no opinion is procedure laid down in Article 17, Member fourth subparagraph of paragraph 8. ordinary legislative procedure, shall establish ordinary legislative procedure, may adopt proposals made by each Member State. ordinary legislative procedure, may amend conditions laid down in the fourth paragraph requirements of the Treaty, in particular relevant provisions of the Kyoto Protocol and principle of subsidiarity as set out in Article 5 procedure and time limits set out in Article procedures and practices specific to agreements concluded with these countries. revised regulatory framework for the procedure referred to in Article 23(2) for a
---	---	---

### Key to Task type G:

a) verb + *in accordance with the*

For example verbs like act, adopt, approve, be, carry out, conduct, decide, draw up, establish, implement, specify

The verb *act* in –ing form appears extremely frequently as it also forms a lexical bundle with *in accordance with*: *acting in accordance with*, as in the following sentence:

*The European Parliament and the Council, acting in accordance with the ordinary legislative procedure, may adopt rules designed ...*

b) noun + *in accordance with the*

For example nouns like action, activities, agreement, measure, programme, regulation

c) *in accordance with the* + noun

For example nouns like opinion, paragraph, principle, procedure

The nouns *principle* and *procedure* co-occur with this lexical bundle very frequently as they form lexical bundles with it: *accordance with the principle* and *accordance with the procedure*

d) In most examples *in accordance with the* is in mid-sentence position, but it also appears in initial position.

## Chapter 9: Conclusion

The present study was motivated by an interest in the variety of English used in the documents published by institutions of the European Union. Previous studies examining the language use within the EU pointed out that the number of EU documents published first in English has increased since the early 1990s (e.g., Truchot, 2002). In addition, research into the language of English EU documents have revealed some language features that distinguish EU discourse from other registers (Jablonkai, 2009a; Pym, 1993; Trebits, 2008, 2009, 2009b; Trosborg, 2007b). Consequently, EU professionals who use English as their L2 need proper preparation and sufficient practice in order to function adequately in the EU context. Therefore, as there has been little research on English EU discourse for pedagogic purposes, the present study aimed to obtain a clearer understanding of the English language use characterising the EU context from an ESP pedagogic perspective.

The present study took a lexically-oriented approach following the Hallidayan tradition of the description of a language variety. Hence lexis was in the focal point of this study and it started out by identifying lexical items strongly associated with a wide range of EU genres and subject fields. This study, however, did not investigate lexical items in isolation. Patterns of lexis were examined in the form of collocations at a semantic and grammatical level and in the form of longer word combinations as lexical bundles examined for their grammatical structure and discourse functions. For the purposes of the present study EU discourse was defined from an ESP pedagogic perspective by relevant and frequently used EU genres that were selected based on a needs analysis survey among Hungarian EU professionals.

The study offers original and novel contributions to several fields of research. These include at a theoretical level: corpus linguistics in ESP research, the methodology of corpus

building for ESP, register analysis in ESP, genre-based approach to ESP, and at a more practical level ESP pedagogy.

Corpus research has gained importance in ESP since the 1990s. The focus of corpus linguistics on real language use made it an appropriate approach for ESP, which also focuses on real language use as represented by texts applied by the respective discourse community. The reason for its focus on real language use lies, first and foremost, in the growing recognition within ESP that tasks and materials need to be based on the analysis of authentic texts used in the target situation language learners learn English for. The present study continues this direction in ESP research and provides the following models for **corpus linguistics in ESP research**. Firstly, the steps and procedures of the present study can be applied as a model for gaining an overall picture of the lexical and lexicogrammatical features of different professional fields. Secondly, at the level of corpus methodology, the *Model for Corpus Creation for ESP* proposed here can be used as guidance for corpus building for ESP research purposes. The novelty of the *Model for Corpus Creation for ESP* lies in its summarising nature pinpointing the relevant steps and questions of a principled design and systematic compilation of corpora for the analysis of the language of particular disciplines or professional fields.

As regards **register analysis in ESP**, findings reinforce earlier research findings regarding particular lexical and lexicogrammatical features of English in specialised texts in some aspects. Firstly, the proportion of high frequency general and technical lexis in written English EU discourse – as represented by the EEUD Corpus – with 75% and 4% respectively is similar to findings of earlier analyses of the lexis in specialised texts where these types of lexis account for 70-75% and 5% (Nation, 1990). Secondly, concerning collocational patterns, results indicate a certain degree of fixedness regarding the number of collocates and semantic preferences of the selected lexical items as it has been suggested by earlier analyses of collocations in specialised texts (Gledhill, 2000; Nelson, 2000, 2006). Finally, written English

EU discourse – as represented by the EEUD Corpus – appears to exhibit characteristics similar to technical type of texts and the registers of textbooks and academic prose in that it also makes frequent use of lexical bundles that incorporate noun and prepositional phrases (Biber, 2006; Hyland, 2008).

However, there are some aspects from which findings of the present research indicate that there are areas of ESP where existing models and descriptions of professional discourses need to be revised. Firstly, the high frequency of lexical bundles in the EEUD Corpus that indicates a rather formulaic language use which is not characteristic to this extent of other written registers analysed earlier. Secondly, the relatively high proportion of lexical bundles with verb phrases which seems to contradict findings of earlier analysis of written registers. Finally, as regards the fixedness in collocational patterns, it was found that the grammatical relations collocates form with the selected lexical items show greater variety in the specialised EEUD Corpus, which seems to challenge the general idea of fixedness of collocations in specialised texts. It should be noted, however, that the collocational analysis was carried out on a very limited set of lexical items, therefore, research with a wider scope is needed to confirm the results of the present study. Overall, these distinctive features of written English EU discourse might contribute to a better understanding of the hybrid nature of EU texts in general.

One of the aims of **genre-based studies in ESP** has been to identify particular genres that members of certain discourse communities recognise as typical, and frequently used in their respective professional settings (e.g., Tompos, 2001) The present study followed this line of research by identifying a list of common core genres of written English EU discourse. The list can be used as a starting point for both materials design in ESP and further research into written English EU discourse.

At a more practical level, the study contributes to **ESP pedagogy** in two aspects. Firstly, it provides informed decisions and a firm basis for **course and materials design** for courses of

English for EU purposes by establishing the EUWL and the list of EU-specific lexical bundles. These results can serve as the basis for developing teaching materials or student glossaries as they include the typical lexical items that characterise this particular discourse type. Furthermore, the results also provide information on the patterns of usage of these lexical items. Secondly, based on the research findings, recommendations were made for the practical aspects of course and materials design by proposing methods and types of task to integrate the findings into the practice of teaching English for EU purposes. In addition, the findings are also transferable to the ESP teaching practice by using sets of raw concordance lines for data-driven language learning.

The limitations of the various research procedures are discussed at the end of the respective sections. It should be noted, however, that in general the findings of the present study refer to written communication within the EU context and, consequently, an obvious overall limitation of the analysis is its exclusive focus on written English EU discourse. Further research is thus needed on spoken communication within EU institutions in order to complement our present knowledge of the professional discourse in the EU context by the distinctive characteristics of spoken English EU discourse. Such a comprehensive analysis may provide a more realistic picture of the English language use within the EU for course and materials design for teaching English for EU purposes.

Finally, the results of the study also suggest that the **EEUD Corpus**, which served as the basis for the present research, is a suitable collection of English EU texts for the linguistic and pedagogic analysis of written English EU discourse. In addition, it needs to be noted that the EEUD Corpus has fulfilled its main aim not only as a source for linguistic description of the variety of English used for written communication within the European Union for the purposes of the present study, but it has already been used in the larger ESP practice, as it recently served



as a basis for designing an entire course book for teaching English for EU purposes with the title *EU English – Using English in EU Contexts* (Trebits & Fischer, 2009).

## References

- Aarts, J. (1991). Intuition-based and observation-based grammars. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics* (pp. 44-63). London, New York: Longman.
- Aijmer, K., & Altenberg, B. (Eds.). (1991). *English corpus linguistics: Studies in honour of Jan Svartvik*. London, New York: Longman.
- Alderson, J. C., & Banerjee, J. (1996). *How might impact study instruments be validated?* Unpublished manuscript.
- Allen, J. P. B., & Van Buren, P. (Eds.). (1971). *Chomsky: Selected readings*. London, New York: Oxford University Press.
- Altenberg, B., & Granger, S. (Eds.). (2001). *Lexis in contrast: corpus-based approaches*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Anthony, L. (2007). *AntConc 3.2.1*. Retrieved August 1, 2007, from [http://www.antlab.sci.waseda.ac.jp/antconc\\_index.html](http://www.antlab.sci.waseda.ac.jp/antconc_index.html)
- Ari, O. (2006). Review of three software programs designed to identify lexical bundles. *Language learning & Technology*, 10(1), 30-37.
- Atkinson, D., & Biber, D. (1994). Register: A review of empirical research. In D. Biber, & E. Finegan, (Eds.), *Sociolinguistic perspective on register* (pp. 351-385). New York, Oxford: Oxford University Press.
- Árvay, A., & Tankó, Gy. (2004). A contrastive analysis of English and Hungarian theoretical research article introductions. *IRAL*, 42, 71-100.
- Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.
- Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: an analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language*, 4(2), 91-105.
- Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253-279.
- Beaugrande, d. R. (1993). 'Register' in discourse studies: a concept in search of a theory. In M. Ghadessy (Ed.), *Register analysis* (pp. 7-25). London, New York: Pinter Publisher.
- Beaugrande, d. R. (1997). *New foundations for a science of text and discourse: Cognition, communication and freedom of access to knowledge and society*. New Jersey: Ablex Publishing Corporation.

- Beaugrande, d. R., & Dressler, W. U. (1983). *Introduction to textlinguistics*. London, New York: Longman.
- Bhatia, V. K. (1993). *Analysing genre*. London: Longman.
- Bhatia, V. K. (1999). Integrating products, processes and participants in professional writing. In C. N. Candlin, & K. Hyland (Eds.), *Writing: Texts, processes and practices* (pp. 21-39). London: Longman.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243-257.
- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.
- Biber, D. (2006). *University language*. Amsterdam, Philadelphia: John Benjamin Publishing Company.
- Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics*, 14(3), 275-311.
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263-286.
- Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard, & S. Oksefjell (Eds.), *Out of corpora* (pp. 181-190). Amsterdam, Atlanta GA: Rodopi.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, New York: Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.
- Bondi, M. (2001). Small corpora and language variation: Reflexivity across genres. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small corpus studies and ELT* (pp. 135-174). Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Born, J., & Schütte, W. (1995). *Eurotexte Textarbeiten einer Institution der EG [Eurotexts Text production in an EU institution]*. Gunter Narr Verlag: Tübingen.

- Boulton, A. (2007). *DDL is in the details ... and in the big themes*. Paper presented at the Corpus Linguistics 2007 Conference, University of Birmingham, Birmingham.
- Bowker, L. (2000). Towards a methodology for exploiting specialized target language corpora as translation resources. *International Journal of Corpus Linguistics*, 5(1), 17-52.
- Bowker, L., & Pearson, J. (2002). *Working with specialized language*. London, New York: Routledge.
- Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.
- Bruce, K. (1987). *LBES: Telephoning*. Harlow: Longman.
- Čermak, F. (2002). Today's corpus linguistics: Some open questions. *International Journal of Corpus Linguistics*, 7(2), 265-282.
- Chambers, F., & McDonough, J. (1981). How many people? Opposing views on the function and preparation of the ESP teacher. In *The ESP teacher: role, development and prospects*. ELT Documents 112 (pp.71-81). Oxford: Pergamon Press and The British Council.
- Chen, Q., & Ge, G. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, 26(4), 502-514.
- Cheng-yu, F. (1993). Building a corpus of the English of computer science. In J. Aarts, P. de Haan, & N. Oostdijk, (Eds.), *English language corpora: Design, analysis, and exploitation*. Papers from the 13<sup>th</sup> International Conference on English Language Research on Computerized Corpora, Nijmegen 1992 (pp. 73-78). Amsterdam: Rodopi.
- Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.
- Chomsky, N. (1962). Untitled paper given at the Third Texas Conference on Problems of Linguistic Analysis in English, 1958, p. 159. Austin: University of Texas.
- Chomsky, N. (1986). *Knowledge of language: its nature, origin and use*. New York: Praeger.
- Chujo, K., & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System*, 34(2), 255-269.
- Chujo, K., Utiyama, M., & Nishigaki, C. (2005). Japanese - English parallel corpus application and CALL: A powerful tool for vocabulary learning. *Foreign Language Education and Technology Conference 2005*, Utah BYU. Retrieved October 2, 2007, from Brigham Young University, Foreign Language Education and Technology Web site: [http://fleet5.byu.edu/\\_files/proceedings.pdf](http://fleet5.byu.edu/_files/proceedings.pdf).

- Chung, T. M. (2003). Corpus comparison approach for term extraction. *Terminology*, 9(2), 221-246.
- Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language*, 15(2), 103-116.
- Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System*, 32(2), 251-263.
- Clear, J. (1992). Corpus sampling. In G. Leitner (Ed.), *New directions in English language corpora* (pp. 21-31). Berlin, New York: Mouton de Gruyter.
- Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System*, 25(3), 301-315.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.
- Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education*. London, New York: Routledge Falmer.
- Collins, H., & Scott, M. (1997). Lexical landscaping in business meetings. In F. Bargiela-Chiappini, & S. Harris (Eds.), *The language of business* (pp. 183-208). Edinburgh: Edinburgh University Press.
- Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics*, 22, 75-95.
- Conrad, S. M. (1996). Investigating academic texts with corpus-based techniques: An example from Biology. *Linguistics and Education*, 8, 299-326.
- Conrad, S. M. (1999). The importance of corpus-based research for language teachers. *System*, 27(1), 1-18.
- Conrad, S., & Biber, D. (Eds.). (2001). *Variation in English multi-dimensional studies*. Harlow: Longman.
- Converse, J. M., & Presser, S. (1986). *Survey questions. Handcrafting the standardized questionnaire*. London: SAGE Publications.
- Cook, G. (1989). *Discourse*. Cambridge: Cambridge University Press.
- 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. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, 17(4), 391-406.

- Cortes, V., & Csomay, E. (2007). Positioning lexical bundles in university lectures. In M. C. Campoy, & M. J. Luzón (Eds.), *Spoken corpora in applied linguistics* (pp. 57-76). Frankfurt am Main: Peter Lang.
- Coulthard, M. (1985). *An introduction to discourse analysis*. London: Longman.
- Council Regulation No 1 determining the languages to be used by the European Economic Community. (1958). *Official Journal of the European Union* 17, 1958.10.6., 385-386. Retrieved on May 2, 2009 from <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31958R0001:EN:HTM>.
- Cowie, A. P. (1992). Multi word lexical units and communicative language teaching. In P. J. L. Arnaud, & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 1-12). London: Macmillan.
- Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, 34(2), 213-238.
- Crowther, J., Dignen, S., & Lea, D. (Eds.). (2002). *Oxford collocations dictionary*. Oxford: Oxford University Press.
- Diez, T. (2001). Europe as a discursive battleground. *Cooperation and Conflict*, 36(1), 5-38.
- Dróth, J. (2000). Legyen egységes az Európai Unió magyar nyelvű terminológiája! Az EU adminisztratív és közigazgatási nyelvezetének magyar fordítása [The Hungarian terminology of the European Union should be uniform! Hungarian translation of the EU's language of public administration]. *Magyar Nyelvőr*, 124(3), 287-297.
- Dudley-Evans, T. (1984). The team-teaching of writing skills. In R. Williams, J. Swales, & J. Kirkman (Eds.), *Common ground: Shared interests in ESP and Communication Studies* (ELT Documents 117 (pp.135-143). Oxford: Pergamon Press and The British Council.
- Dudley-Evans, T., & St John, M. J. (1998). *Developments in English for Specific Purposes A multi-disciplinary approach*. Cambridge: Cambridge University Press.
- Eggs, S. (2004). *An introduction to systemic functional linguistics*. London, New York: Continuum.
- Ellis, M., & O'Driscoll, N. (1992). *LBES: Giving presentations*. Harlow: Longman.
- Ellis, M., O'Driscoll, N., & Pilbeam, A. (1987). *LBES: Socializing*. Harlow: Longman.
- Esteban, A. A., & Cañado, M. L. P. (2004). Making the case method work in teaching Business English: A case study. *English for Specific Purposes*, 23(2), 137-161.
- EURlex The documentary holdings, Retrieved August 25, 2008 from [http://eur-lex.europa.eu/en/droit\\_communataire/droit\\_communataire.htm](http://eur-lex.europa.eu/en/droit_communataire/droit_communataire.htm)

- Favretti, R.R., Tamburi, F., & Martelli, E. (2001). Words from Bononia Legal Corpus. *International Journal of Corpus Linguistics*, 6, 13-34.
- Firth, J. R. (1968). A synopsis of linguistic theory, 1930-55. In F. R. Palmer (Ed.), *Selected papers of J. R. Firth 1952-59* (pp. 168-205). London: Longmans.
- Fischer, M. (2006). Translation(policy) and terminology in the European Union. Paper presented at the *Terminology and Society - the Impact of Terminology on Everyday Life. International Conference on Terminology*, NL-TERM Lessius University College Antwerpen.
- Fischer, M. (2007). Fordításpolitika és terminológia az Európai Unióban [Translation policy and terminology in the European Union]. In P. Heltai (Ed.), *Nyelvi modernizáció – szaknyelv, fordítás, terminológia [Linguistics modernisation – LSP, translation, terminology]*, a MANYE 2006. évi XVI. kongresszusának köteté [Proceedings of the 16th Hungarian Applied Linguistics Congress], (pp. 806-811). MANYE - Szent István Egyetem, Pécs - Gödöllő.
- Flowerdew, J. (1994). Specific language for specific purposes: Concordancing for the ESP syllabus. In R. Khoo (Ed.), *LSP problems and prospects* (pp. 97-113). Singapore: SEAMEO Regional Language Centre.
- Flowerdew, L. (1998). Corpus linguistics techniques applied to text linguistics. *System*, 26(4), 541-552.
- Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry & R. L. Roseberry (Eds.), *Small corpus studies and ELT* (pp. 363-380). Amsterdam Philadelphia: John Benjamins Publishing Company.
- Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor, & T. A. Upton (Eds.), *Discourse in the profession* (pp. 11-33). Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/EAP: Countering criticism against corpus-based methodologies. *English for Specific Purposes*, 24(3), 321-332.
- Forchini, P., & Murphy, A. (2008). N-grams in comparable specialized corpora. *International Journal of Corpus Linguistics*, 13(3), 351-367.
- Francis, N., & Kučera, H. (1982). *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.

- Fuentes, A. C. (2001). Lexical behaviour in academic and technical corpora: implications for ESP development. *Language learning & technology*, 5(3), 106-129.
- Fuentes, A. C. (2002). Exploitation and assessment of a business English corpus through language learning tasks. *ICAME Journal*, 26, 5-32.
- Ghadessy, M., Henry, A., & Roseberry, R. L. (Eds.). (2001). *Small corpus studies and ELT*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Ghadessy, P. (1979). Frequency counts, word lists, and materials preparation: A new approach. *English Teaching Forum*, 17, 24-27.
- Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, 19(2), 115-135.
- Goodale, M. (1987). *Language of meetings*. Hove: Language Teaching Publications.
- Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32(1), 148-170.
- Guidelines for electronic text encoding and interchange by the Text Encoding Initiative Consortium. (2002). Retrieved March 10, 2007, from [www.tei-c.org](http://www.tei-c.org)
- Guillot, M. (2002). Corpus-based work and discourse analysis in FL pedagogy: a reassessment. *System*, 30(1), 15-32.
- Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, & R. y. H. Robins (Eds.), *In memory of J.R. Firth* (pp. 148-62). London: Longman.
- Halliday, M. A. K. (1976). The form of a functional grammar. In G. Kress, (Ed.), *System and function of language* (pp. 7-25). Oxford: Oxford University Press.
- Halliday, M. A. K. (1978). *Language as a social semiotic - the social interpretation of language and meaning*. London: Arnold.
- Halliday, M. A. K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics* (pp. 30-43). London, New York: Longman.
- Halliday, M. A. K. (1994). *An introduction to functional grammar*. London: Arnold.
- Halliday, M. A. K., & Hasan, R. (1985). *Language, context, and text: aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.
- Halliday, M. A. K., & Matthiessen, C. (2004). *An introduction to functional grammar*. London: Arnold.
- Halliday, M. A. K., McIntosh, A., & Stevens, P. (1964). *The linguistic sciences and language teaching*. London: Longman.



- Hänchen, R. (2002). *Die Französische Marketingsprache: Eine diachrone Untersuchung ihrer Terminologie anhand der Revue Française du Marketing (1960-2000)* [Language of the French marketing. A diachronic analysis of the terminology in Revue Française du Marketing (1960-2000)]. Frankfurt: Peter Lang.
- Hatch, E., & Lazaraton, A. (1991). *The research manual*. Boston: Heinle & Heinle.
- Heatley, A., Nation, P., & Coxhead, A. (2002). *Range and frequency programs software*. Retrieved September 1, 2005, from <http://www.victoria.ac.nz/lals/staff/paul-nation.aspx>
- Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2), 174-199.
- Heltai, P. (2007). Párhuzamos szaknyelvi korpusz munkálatai a Szent István Egyetemen [Work on the parallel LSP corpus at Szent István University]. In M. Silye (Ed.), *PORTA LINGUA 2007 Szaknyelvvoktatásunk - határokon átívelő híd [LSP a bridge across borders]* (pp. 285-294). Debrecen: Debreceni Egyetem, SZOKOE [Hungarian Association of Teachers and Researchers of Languages for Specific Purposes].
- Henry, A., & Roseberry, R. L. (2001). Using a small corpus to obtain data for teaching a genre. In Ghadessy, M., Henry, A., & Roseberry, R.L. (Eds.), *Small corpus studies and ELT* (pp. 93-134). Amsterdam, Philadelphia: John Benjamins Publishing.
- Hewings, M. (2003). A history of ESP through 'English for Specific Purposes'. *ESP World*, 1(3). Retrieved April 26, 2007, from [http://esp-world.7p.com/Articles\\_3/Hewings\\_paper.htm](http://esp-world.7p.com/Articles_3/Hewings_paper.htm)
- Hirsch, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8(2), 689-696.
- Hockey, S. (2001). Concordance programs for corpus linguistics. In C. R. Simpson., & J. M. Swales, (Eds.), *Corpus linguistics in North America Selections from the 1999 Symposium* (pp. 58-97).
- Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.
- Hoey, M. (2000). A world beyond collocation: New perspectives on vocabulary teaching. In M. Lewis (Ed.), *Teaching collocations*. Hove: Language Teaching Publications.
- Hoey, M. (2005). *Lexical priming*. New York: Routledge.
- Holmes, J. (2005). Text analysis and the teaching of language items. In M. A. A. Celani, T. Deyes, J. Holmes, & M. Scott (Eds.), *ESP in Brazil: 25 years of evolution and reflection* (pp. 239-303). São Paulo and Campinas, Brazil: EDUC and Mercado de Letras.
- Howe, B. (1992a). *Portfolio: Case studies for Business English*. Harlow: Longman.

- Howe, B. (1992b). *Portfolio: Case studies for Business English teacher's guide*. Harlow: Longman.
- Hüllen, W. (1981). Movements on earth and in the air: A study of certain verbs occurring in the language of international pilots. *English for Specific Purposes*, 1(2), 141-154.
- Hunston, S. (2002). *Corpora in applied linguistics*. Oxford: Oxford University Press.
- Hutchinson, T., & Waters, A. (1987). *English for Specific Purposes A learning-centred approach*. Cambridge: Cambridge University Press.
- Hyland, K. (2000). *Disciplinary discourses*. Harlow: Pearson Education Ltd.
- Hyland, K. (2002a). Genre: Language, context, and literacy. *Annual Review of Applied Linguistics*, 22, 113-135.
- Hyland, K. (2002b). Specificity revisited: how far should we go now? *English for Specific Purposes*, 21(4), 385-395.
- Hyland, K. (2006). *English for Academic Purposes: An advanced resource book*. Abingdon, Oxon, New York: Routledge.
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4-21.
- Jabbour, G. (1998). *Corpus linguistics, contextual collocation and ESP syllabus creation: A text analysis approach to the study of medical research articles*. Unpublished PhD, University of Birmingham, Birmingham, United Kingdom.
- Jablonkai, R. (2007). *A nyelvtanuló mint kutató [Language learners as researchers]*. Paper presented at SZOKOE [Hungarian Association of Teachers and Researchers of Languages for Specific Purposes] Conference, Janus Pannonius University, Pécs.
- Jablonkai, R. (2009a). In the light of: A corpus-based analysis of two EU-related registers. *WoPaLP*, 3, 1-27. Available at: <http://langped.elte.hu/W3Jablonkai.pdf>.
- Jablonkai, R. (2009b). *Lexikai csoportok az angol sajtónyelvben és EU szaknyelvben [Lexical bundles in English journalism and in the language of the EU]*. In T. Váradi (Ed.), III. Alkalmazott Nyelvészeti Doktorandusz Konferencia kötet [Proceedings of the III. Applied Linguistics PhD Conference] (pp. 13-27). Budapest: MTA Nyelvtudományi Intézet. Available at: <http://www.nytud.hu/alknyelvdok09/proceedings.pdf>
- Jablonkai, R. (in press). English in the context of European integration: a corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes. Special Issue: ESP in Europe*.

- Jackson, J. (2002). The China strategy: A tale of two case leaders. *English for Specific Purposes*, 21(3), 243-259.
- Jackson, J. (2004). Case-based teaching in a bilingual context: Perceptions of business faculty in Hong Kong. *English for Specific Purposes*, 23(3), 213-232.
- James, G., Davison, R., Heung-yeung, A. C., & Deerwester, S. (1994). *English in computer science A corpus-based lexical analysis*. Hong Kong: Longman Asia.
- Johansson, S. (1991). Times change, and so do corpora. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics* (pp. 305-314). London, New York: Longman.
- Johns, T. (1991a). From printout to handout: Grammar and vocabulary learning in the context of data-driven learning. *English Language Research Journal*, 4, 27-46.
- Johns, T. (1991b). Should you be persuaded: two samples of data-driven learning materials. *English Language Research Journal* 4, 1-16.
- Jones, C. (1991). An integrated model for ESP syllabus design. *English for Specific Purposes*, 10(3), 155-172.
- Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk, & H. G. Widdowson (Eds.), *English in the world* (pp. 11-30). Cambridge: Cambridge University Press.
- Károly, K. (2003). Korpusznyelvészet és fordításkutatás [Corpus linguistics and translation research]. *Fordítástudomány*, 5(2), 18-26.
- Károly, K. (2007). *Szövegtan és fordítás. [Text linguistics and translation]*. Budapest: Akadémiai Kiadó.
- Károly K., & Tankó Gy. (2009). *Magyarországi Angolnyelv-tanulói Korpusz. [Corpus of Hungarian Learners of English]* Budapest: ELTE BTK Angol-Amerikai Intézet.
- Kennedy, C., & Bolitho, R. (1984). *English for Specific Purposes*. London: Macmillan.
- Kennedy, G. (1998). *An Introduction to corpus linguistics*. London: Longman.
- Kerridge, D. (1988). *LBES: Presenting facts and figures*. Harlow: Longman.
- Kilgarriff, A., & Rundell, M. (2002). Lexical profiling software and its lexicographic applications - a case study. *Proceedings of the Tenth EURALEX International Congress*, 2 (pp. 807-818). Copenhagen: Center for Sprogteknologi. Retrieved August 25, 2009, from <http://www.kilgarriff.co.uk/publications.htm>
- Kilgarriff, A., & Tugwell, D. (2001). WORD SKETCH: Extraction and display of significant collocations for lexicography. *Proceedings of the ACL workshop on collocation*:

- computational extraction, analysis and exploitation*, (pp. 32-38). Toulouse, France. Retrieved August 25, 2009, from <http://www.kilgarriff.co.uk/publications.htm>
- Kilgarriff, A., & Tugwell, D. (2002). Sketching words. In M. Corréard (Ed.), *Lexicography and natural language processing: A festschrift in honour of B. T. S. Atkins* (pp. 125-137). EURALEX. Retrieved August 25, 2009, from <http://www.kilgarriff.co.uk/publications.htm>
- Klaudy, K. (2001). Mit tehet a fordítástudomány a magyar nyelv "korszerűsítéséért"? [What can translation studies do for "modernising" the Hungarian language?] *Magyar Nyelvőr*, 125(2), 145-152.
- Knowles, G. (1996). Corpora, databases and the organization of linguistic data. In J. Thomas, & M. Short (Eds.), *Using corpora for language research* (pp. 14-26). London: Longman.
- Knowles, G., & Zuraidah, M. D. (2004). The notion of a "lemma": Headwords, roots and lexical sets. *International Journal of Corpus Linguistics*, 9(1), 69-81.
- Krausse, S. (2005). Testing the validity of small corpus information. *ESP World*, 4(1(9)), Retrieved 26 April, 2007, from [http://www.esp-world.info/Articles\\_9/testing\\_the\\_validity.htm](http://www.esp-world.info/Articles_9/testing_the_validity.htm)
- Kretzschmar Jr., W. A., Darwin, C., Brown, C., Rubin, D. L., & Biber, D. (2004). Looking for the smoking gun. *Journal of English Linguistics*, 32, 31-47.
- Krishnamurthy, R. (2008). Corpus-driven lexicography. *International Journal of Lexicography*, 21(3), 231-242.
- Kurtán, Zs. (2003). *Szakmai nyelvhasználat*. [Technical language use]. Budapest: Nemzeti Tankönyvkiadó.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Laufer, B. (1997). The lexical plight in second language reading. In J. Coady, & T. Huckin *Second language vocabulary acquisition* (pp. 20-34). Cambridge: Cambridge University Press.
- Laufer, B. (1992). How much lexis is necessary for reading comprehension. In P. J. L. Arnaud, & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126-132). London: Macmillan.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307-322.

- Laviosa, S. (2000). TEC: A resource for studying what is "in" and "of" translational English. *Across Languages and Cultures*, 1(2), 159-177.
- Leckie-Tarry, H. (1993). The specification of a text: register, genre and language teaching. In M. Ghadessy (Ed.), *Register analysis* (pp. 26-42). London, New York: Pinter Publisher.
- Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available corpora to self-compiled corpora. *English for Specific Purposes*, 25(1), 56-75.
- Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics* (pp. 8-29). London, New York: Longman.
- Leitner, G. (1992). International Corpus of English: Corpus design – problems and suggested solutions. In G. Leitner (Ed.), *New directions in English language corpora* (pp. 21-31). Berlin: Mouton de Gruyter.
- Lewis, M. (1993). *The lexical approach*. London: LTP.
- López, Á. L., & Cañado, M. L. P. (2001). Needs analysis of ESP learners in the Commission of the European Union. In S. Posteguillo, I. Fortanet & J. C. Palmer (Eds.), *Methodology and new technologies in LSP* (pp. 293-305). Universitat Jaume.
- Louhiala-Salminen, L., Charles, M., & Kankaaranta, A. (2005). English as a lingua franca in Nordic corporate mergers: Two case companies. *English for Specific Purposes*, 24(4), 401-421.
- Malcolm, L. (1987). What rules govern tense usage in scientific articles? *English for Specific Purposes*, 6(1), 31-43.
- Martin, A. V. (1976). Teaching academic vocabulary to foreign graduate students. *TESOL Quarterly*, 10(1), 91-97.
- Martin, J. R. (1992). *English text*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- McArthur, T. (2003). World English, Euro-English, Nordic English? *English Today*, 19(1), 54-58.
- McEnery, T. & Wilson, A. (1996a). *Corpus linguistics*. Edinburgh: Edinburgh University Press.
- McEnery, T., & Wilson, A. (1996b). *Web-based course on corpus linguistics web pages to be used to supplement the book "Corpus linguistics"*. Retrieved February 1, 2005, from <http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies*. New York: Routledge.

- Meyer, C. (2002). *English corpus linguistics: an introduction*. Cambridge: Cambridge University Press.
- Moon, R. (1998). *Fixed expressions and idioms in English A corpus-based approach*. Oxford: Clarendon Press.
- Moon, R. (2000). Vocabulary connections: Multi-word items in English. In N. Schmitt, & M. McCarthy (Eds.), *Vocabulary description, acquisition and pedagogy* (pp. 40-63). Cambridge: Cambridge University Press.
- Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes*, 25(2), 235-356.
- Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.
- Murison-Bowie, S. (1996). Linguistic corpora and language teaching. *Annual Review of Applied Linguistics* 16, 182-199.
- Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston: Heinle and Heinle Publishers.
- Nation, I. S. P. (1993). Vocabulary size, growth and use. In R. Schreuder, & B. Weltens (Eds.), *The bilingual lexicon*. (pp. 115-134). Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Nation, P. (2001). Using small corpora to investigate learner needs. In M. Ghadessy, A. Henry & R. L. Roseberry (Eds.), *Small corpus studies and ELT theory and practice* (pp. 31-45). Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Nation, P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards, & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 3-13). Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59-82.
- Nation, I. S. P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, 23(1), 35-41.
- Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt, M. McCarthy (Eds.), *Vocabulary description, acquisition and pedagogy*. Cambridge: Cambridge University Press.
- Nattinger, J., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

- Neely, E., & Cortes, V. (2009). A little bit about: analyzing and teaching lexical bundles in academic lectures. *Language Value*, 1(1), 17-38. Retrieved on 25 January, 2010 from <http://www.e-revistas.uji.es/languagevalue>
- Nelson, G. (1996). The design of the corpus. In S. Greenbaum (Ed.), *Comparing English worldwide: The International Corpus of English*. Oxford: Clarendon Press.
- Nelson, M. (2000). *A corpus-based study of Business English and Business English teaching materials*. Unpublished PhD, University of Manchester, Retrieved November 15, 2005 from <http://www.kielikanava.com/thesis.html>
- Nelson, M. (2006). Semantic association in Business English: A corpus-based analysis. *English for Specific Purposes*, 25(2), 217-234.
- Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics*, 11(3), 283-304.
- Nickerson, C. (2005a). English as a lingua franca in international business contexts. *English for Specific Purposes*, 24(4), 367-380.
- Nickerson, C. (Ed.). (2005b). English as a lingua franca in international business contexts. *English for Specific Purposes*, 24(4), 367-452.
- Oakes, M. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
- O'Connor, P., Pilbeam, A., & Scott-Barrett, F. (1992). *Negotiating*. Harlow: Longman.
- O'Driscoll, N., & Pilbeam, A. (1987). *LBES: Meetings & Discussions*. Harlow: Longman.
- O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom*. Cambridge: Cambridge University Press.
- Owen, C. (1993). Corpus-based grammar and the Heineken effect: Lexico-grammatical description for language learners. *Applied Linguistics*, 14(2), 167-187.
- Partington, A. (1998). *Patterns and meaning: using corpora for English language research and teaching*. Amsterdam: John Benjamins Publishing Company.
- Partington, A. (2004). Utterly content in each other's company: semantic prosody and semantic preference. *International Journal of Corpus Linguistics*, 9(1), 91-118.
- Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards, & R. Schmidt (Eds.), *Language and communication* (pp. 191-230). London: Longman.
- Petneki, K. (2000). A szaknyelvoktatás néhány elméleti és módszertani kérdése [Theoretical and methodological issues in teaching LSP]. *Modern Nyelvoktatás*, 6(2-3), 61-69.

- Petric, B., & Czárli, B. (2003). Validating a writing strategy questionnaire. *System*, 31(2), 187-215.
- Powell, M. (1996). *Presenting in English*. Hove: Language Teaching Publications.
- Project Gutenberg. Retrieved April 5, 2009, from [http://www.gutenberg.org/wiki/Main\\_Page](http://www.gutenberg.org/wiki/Main_Page)
- Pym, A. (1993). *Epistemological problems in translation and its teaching. A seminar for thinking students*. Calceit: Caminade.
- Pym, A. (2000). The European Union and its future languages: Questions for language policies and translation theories. *Across Languages and Cultures*. 1(1), 1-17.
- Renouf, A. (1987). Corpus development. In J. Sinclair (Ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Harper Collins.
- Renouf, A., & Sinclair, J. M. (1992). Collocational frameworks in English. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics* (pp. 128-143). London: Longman.
- Reppen, R. (2001). Review of MonoConc Pro and WordSmith tools. *Language learning & Technology*, 5(3), 32-36.
- Reppen, R., Fitzmaurice, S. M., & Biber, D. (Eds.). (2002). *Using corpora to explore linguistic variation*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Robinson, P. (1991). *ESP Today: a Practitioner's Guide*. Hemel Hempstead: Prentice Hall International.
- Sager, J. C., Dungworth, D. and McDonald, P.F. (1980). *English special languages Principles and practice in science and technology*. Oscar Brandstetter Verlag KG. Wiesbaden.
- Sajtos, L., & Mitev, A. (2009). *SPSS kutatási és adatelemzési kézikönyv. [SPSS manual for research and data analysis]*. Budapest: Alinea.
- Sampson, G. (1996). From central embedding to corpus. In J. Thomas, & M. Short (Eds.), *Using corpora for language research* (pp.14-26). London: Longman.
- Sanders, T., & Sanders, J. (2006). Text and text analysis. *Encyclopedia of Language & Linguistics*, 597-607.
- Sass, B. (2007). A Hunglish korpusz mint oktatási segédeszköz [The Hunglish corpus as a teaching tool]. In P. Heltai (Ed.), *Nyelvi modernizáció – szaknyelv, fordítás, terminológia [Linguistic modernisation – LSP, translation, terminology]*, a MANYE 2006. évi XVI. kongresszusának kötete [Proceedings of the 16th Hungarian Applied Linguistics Congress], (pp. 969-973). MANYE - Szent István Egyetem, Pécs - Gödöllő.
- Schäffner, C., & Adab, B. (2001a). The idea of the hybrid text in translation: Contact as conflict. *Across Languages and Culture*, 2(2), 167-180.



- Schäffner, C., & Adab, B. (2001b). The idea of the hybrid text in translation revisited. *Across Languages and Culture*, 2(2), 277-302.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International conference on new methods in language processing* (pp. 44-49). Manchester, UK. Retrieved November 10, 2009, from <http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf>
- Scott, M. (1997). PC analysis of key words – And key key words. *System*, 25(2), 233-245.
- Scott, M. (2000). Focusing on the text and its key words. In L. Burnard, & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 103-121). Frankfurt am Main: Peter Lang.
- Scott, M. (2004). *Oxford WordSmith Tools Version 4.0*. Oxford University Press. Retrieved December 10, 2005, from: <http://www.lexically.net/wordsmith/>
- Scott, M. (2005). *Corpus linguistics and ESP: is there a link?* Paper presented at the 19<sup>th</sup> National ESP Seminar, PUC, São Paulo. Retrieved May 25, 2006 from [http://www.lexically.net/downloads/corpus\\_linguistics/](http://www.lexically.net/downloads/corpus_linguistics/)
- Scott, M., & Tribble, C. (2006). *Textual patterns*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Seliger, H. W., & Shohamy, E. (1989). *Second language research methods*. Oxford: Oxford University Press.
- Selinker, L., Tarone, E. & Hanzeli, V. (Eds.), (1981), *English for academic and technical purposes: Studies in honour of Louis Trimble*. Rowley, Mass: Newbury House.
- Shin, D., & Nation, P. (2008). Beyond single words: the most frequent collocations in spoken English. *ELT Journal*, 64(4), 339-348.
- Sinclair, J. M. (Ed). (1987). *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. London: Harper Collins Publishers.
- Sinclair, J. (Ed.). (1990). *Collins COBUILD English Grammar*. London: Harper Collins Publishers.
- Sinclair, J. (Ed.). (2003). *Collins Cobuild English dictionary for advanced learners*. London: Harper Collins Publishers.
- Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

- Sinclair, J. (2005). Corpus and text - basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: a guide to good practice* (pp. 1-16). Oxford: Oxbow Books. Retrieved March 21, 2007, from: <http://ahds.ac.uk/linguistic-corpora/>
- Sinclair, J. M., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter, & M. McCarthy (Eds.), (pp. 141-160). London, New York: Longman.
- Sinclair, J. M., Jones, S., Daley, R., & Krishnamurthy, R. (2004). *English collocation studies the OSTI report*. London: Continuum.
- Song, B. (2006). Content-based ESL instruction: Long-term effects and outcomes. *English for Specific Purposes*, 25(4), 420-437.
- Sterkenburg, van P. G. J. (Ed.). (2003). *A practical guide to lexicography*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- St John, E. (2001). A case for using a parallel corpus and concordancer for beginners of a foreign language. *Language Learning & Technology*, 5(3), 185-203.
- St John, M. J. (1996). Business is booming: Business English in the 1990s. *English for Specific Purposes*, 15(1), 3-18.
- Stevens, V. (1991). Concordance-based vocabulary exercises. *English Language Research Journal*, 4, 47-62.
- Stoller, F. L. (2004). Content-based instruction: Perspectives on curriculum planning. *Annual Review of Applied Linguistics*, 24, 261-283.
- Stevens, P. (1988). ESP after twenty years: A re-appraisal. In M. Tickoo (Ed.), *ESP: State of the art* (pp. 1-13). Singapore: SEAMEO Regional Language Centre.
- Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1), 1-33.
- Stubbs, M. (2001a). Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics*, 22(2), 149-172.
- Stubbs, M. (2001b). *Words and phrases*. Oxford: Blackwell.
- Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, 7(2), 215-244.
- Summers, D. (Ed.). (2003). *Longman dictionary of contemporary English New edition*. Harlow: Longman.
- Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus-based study. *RELC Journal*, 25(2), 34-50.

- Svartvik, J. (1996). Corpora are becoming mainstream. In J. Thomas, & M. Short (Eds.), *Using corpora for language research* (pp. 3-13). London and New York: Longman.
- Swales, J. (1985). *Episodes in ESP a source and reference book on the development of English for Science and Technology*. Oxford: Pergamon.
- Swales, J. M. (1990). *Genre analysis*. Cambridge: Cambridge University Press.
- Swales, J. M. (2004). *Research genres: Exploration and applications*. Cambridge: Cambridge University Press.
- Swales, J. M., & Feak, C. B. (1994). *Academic writing for graduate students*. Ann Arbor: The University of Michigan Press.
- Szirmai, M. (2005). *Bevezetés a korpusznyelvészetbe [Introduction to corpus linguistics]*. Budapest: Tinta Könyvkiadó.
- Tarone, E., Dwyer, S., Gillette, S., & Icke, V. (1981). On the use of the passive in two astrophysics journal papers. *English for Specific Purposes*, 1(2), 123-140.
- The European Commission Prelex. Retrieved August 25, 2008 from  
[http://ec.europa.eu/prelex/ct/sgv\\_manual\\_dsp\\_main.cfm?manualcat\\_id=documents&cl=en](http://ec.europa.eu/prelex/ct/sgv_manual_dsp_main.cfm?manualcat_id=documents&cl=en)
- Thompson, G. (2001). Corpus, comparison, culture: Doing the same things differently in different cultures. In M. Ghadessy, A. Henry & R. L. Roseberry (Eds.), *Small corpus studies and ELT* (pp. 311-334). Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Tompos, A. (2001). *A genre-based approach to ESP testing*. Unpublished PhD, University of Pécs.
- Trebits, A. (2008). English lexis in the documents of the European Union – A corpus-based exploratory study. *WoPaLP*, 2, 38-54. Retrieved September 1, 2009, from  
<http://langped.elte.hu/Wopalpindex.htm>
- Trebits, A. (2009a). Conjunctive cohesion in English language EU documents – A corpus-based analysis and its implications. *English for Specific Purposes*, 28(3), 199-210.
- Trebits, A. (2009b). The most frequent phrasal verbs in English language EU documents – A corpus-based analysis and its implications. *System*, 37(3), 470-481.
- Trebits, A., & Fischer, M. (2009). *EU English – Using English in EU contexts*. Budapest: Klett Kiadó.

- Tribble, C. (2000). Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals. In L. Burnard, & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 75-90). Frankfurt am Main: Peter Lang.
- Tribble, C., & Jones, G. (1997). *Concordances in the classroom: using corpora. A resource guide for teachers*. London: Longman.
- Trosborg, A. (1997a). Text typology: Register, genre and text type. In A. Trosborg, (Ed.), *Text typology and translation* (pp. 3-23). Amsterdam: John Benjamins Publishing Company.
- Trosborg, A. (1997b). Translating hybrid political texts. In A. Trosborg, (Ed.), *Text typology and translation* (pp. 145-158). Amsterdam: John Benjamin Publishing Company.
- Truchot, C. (2002). *Key aspects of the use of English in Europe*. Strasbourg: Council of Europe.
- Upton, T. A., & Connor, U. (2001). Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes*, 20(4), 313-329.
- Viel, J. C. (2002). The vocabulary of English for Specific and Technological Occupational Purposes. *ESP World*, 1. Retrieved December 10, 2005, from [http://www.esp-world.info/Articles\\_1/vocabulary.html](http://www.esp-world.info/Articles_1/vocabulary.html)
- Walker, C. (2009). The treatment of collocation by learners' dictionaries, collocational dictionaries and dictionaries of Business English. *International Journal of Lexicography*, 22(3), 281-299.
- Wang, J., Liang, S., & Ge, G. (2008). Establishment of a Medical English word list. *English for Specific Purposes*, 27(4), 442-458.
- Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language*, 12(2), 309-323.
- Ward, J. (2007). Collocation and technicality in EAP engineering. *English for Academic Purposes*, 6(1), 18-35.
- West, M. (1953). *A general service list of English words*. London: Longman.
- West, R. (1994). Needs analysis in language teaching. *Language Teaching*, 27(1), 1-19.
- Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (Eds.). (1997). *Teaching and language corpora*. Harlow: Longman.
- Widdowson, H. G. (1978). *Teaching language as communication*. Oxford: Oxford University Press.
- Widdowson, H.G. (1984). *Learning purpose and language use*. Oxford: Oxford University Press.
- Widdowson, H. G. (1996). *Linguistics*. Oxford: Oxford University Press.

- Widdowson, H.G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21(1), 3-25.
- Wiechmann, D., & Fuhs, S. (2006). Corpus linguistics resources Concordancing software. *Corpus Linguistics and Linguistic Theory*, 2(1), 109-130.
- Wilkins, D. (1976). *Notional syllabuses*. Oxford: Oxford University Press.
- Williams, G. C. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3, 151-171.
- Williams, G. (2002). In search of representativity in specialised corpora. *International Journal of Corpus Linguistics*, (7)1, 43-64.
- Williams, R., Swales, J., & Kirkman, J. (Eds.). (1984). *Common ground: Shared interests in ESP and communication studies*. Oxford: Pergamon Press and The British Council.
- Willis, D. (1990). *The lexical syllabus: A new approach to language teaching*. London: Collins Cobuild.
- Xue, G., & Nation, P. (1984). A university word list. *Language Learning and Communication*, 3(2), 215-229.
- Zak, H., & Dudley-Evans, T. (1986). Features of word omission and abbreviation in telexes. *English for Specific Purposes*, 5(1), 59-71.

## Appendices

1. Needs analysis questionnaire
2. Interview protocol
3. Text of the e-mail message to potential respondents
4. Titles of EU texts in the English EU Discourse Corpus
5. EU Word List
6. Sample EU text to illustrate the text coverage of EUWL
7. Results of collocational analysis of the lemmas COMMISSION, LAY, IMPLEMENT
8. Full list of lexical bundle types in the EEUD Corpus
9. Sample EU text to illustrate the text coverage of lexical bundles
10. Pedagogic collocational profile of the verb IMPLEMENT