"EAP vocabulary in native and learner writing : from extraction to analysis : a phraseology-oriented approach"

Paquot, Magali

Abstract

This thesis deals with the phraseology of English for Academic Purposes (EAP) vocabulary in learner writing. The seven chapters of the thesis are organised as follows: Chapter One aims to characterize English for Academic Purposes. It gives a general account of its distinctive linguistic features before focusing on vocabulary needs in academic settings. A distinction is made between receptive and productive vocabulary and the term #academic vocabulary' is defined in the light of its particular nature and role in academic discourse. It then offers a review of corpus-based studies of vocabulary in professional academic discourse, learner writing and native student writing. Chapter Two deals with the fuzzy boundaries of the phraseological spectrum. It successively addresses the questions of the categorization of phrasemes and their defining criteria before presenting the typology and definitions adopted in this thesis. It then sheds light on the dual nature of the term "collocation" an...

Document type : *Thèse (Dissertation)*

Référence bibliographique

Paquot, Magali. *EAP vocabulary in native and learner writing : from extraction to analysis : a phraseology-oriented approach.* Prom. : Granger, Sylviane

UNIVERSITE CATHOLIQUE DE LOUVAIN
FACULTE DE PHILOSOPHIE ET LETTRES

EAP vocabulary in native and learner writing: From extraction to analysis

A phraseology-oriented approach

MAGALI PAQUOT

Dissertation submitted for the Degree of
Doctor of Philosophy and Letters
Université catholique de Louvain

Supervisor: Professor Sylviane Granger

September 2007

Acknowledgments

I would like to start off by expressing my deepest and most sincere gratitude to my PhD supervisor, Professor Sylviane Granger, for her catching enthusiasm, her intellectual perceptiveness and her unfailing expert guidance. I am greatly indebted to you, Sylviane, for giving me the opportunity to join the renowned Centre for English Corpus Linguistics four years ago now and for truly believing in me despite some advice to the contrary! I have been lucky enough to undertake research in an environment where writing a PhD also means collaborating with many fellow researchers on up-and-coming projects, going to thought-provoking conferences, organizing seminars, phraseology conferences and summer schools on learner corpora, as well as lecturing and offering guidance to undergraduate students. Not only have you enabled my PhD project to come to full fruition, you have also opened up avenues and opportunities for my future career in the field of academic research, and I want to thank you for that.

I am also very grateful to my colleagues and friends at the Centre for English Corpus Linguistics - Céline, Christelle, Claire, Jennifer, Fanny, Gaëtanelle, Marie-Aude, Stéphanie, and Sylvie – for making the Centre for English Corpus Linguistics such an inspiring and intellectually-stimulating research centre. I also wish to thank them for their moral and intellectual support and for all the entertaining lunchtimes we spent together talking about everyday life…and work.

Many thanks are due to Yves Bestgen for helping me with the statistical aspects of this thesis and for spending valuable time answering my numerous emails about statistical tests, comparisons of means, and p values.

I also wish to thank all the members of my Comité d'encadrement – Fanny Meunier, Jean Heiderscheidt and Jean-René Klein - for their invaluable advice and kind words of encouragement.

I gratefully acknowledge the support of the Communauté française de Belgique, which funded my doctoral dissertation within the framework of the 'Action de recherche concertée' project entitled 'Foreign Language Learning: Phraseology and Discourse' (No. 03/08-301).

I cannot finish off without thanking the many colleagues who kindly gave me access to a corpus or software tool: Hilary Nesi who gave me access to a preliminary version of the *British Academic Writtten English corpus*, JoAnne Neff van Aertselaer who provided me with a corpus of Spanish editorials, Liesbet Degand and Julien Perrez who made the *Nederlands door Nederlandstaligen* corpus available to me, John Osborne who sent me a corpus of French EFL student essays which I would like to study in my future work, and Sebastian Hoffmann who granted me access to the BNCWeb.

These last few words are for Arnaud: thank you, Arnaud, for making it all worthwhile.

List of abbreviations and acronyms

AKL	Academic Keyword List (my own list)
AWL	Academic Word List (Coxhead 2000)
BAWE	*British Academic Written English* corpus
BNC	*British National Corpus*
BNC-AC	*British National Corpus – Academic* sub-corpus
BNC-AC-HUM	*British National Corpus – Academic – Humanities* sub-corpus
BNC-SP	*British National Corpus – spoken*
CALL	Computer-Assisted Language Learning
CA	Contrastive Analysis
CIA	Contrastive Interlanguage Analysis
CLAWS	Constituent Likelihood Automatic Word-tagging system
CNN	Corpus Nederlands door Nederlandstaligen
CODIF	*COrpus de DIssertations Françaises*
EAP	English for Academic Purposes
EFL	English as a Foreign Language
EGAP	English for General Academic Purposes
ESAP	English for Specific Academic Purposes
ESL	English as a Second Language
ESP	English for Specific Purposes
GSL	General Service List (West 1953)
ICLE	*International Corpus of Learner English*
ICLEv2	*International Corpus of Learner English* (second edition)
IL	Interlanguage
L1	First language
L2	Second/foreign language
LDOCE4	Longman Dictionary of Contemporary English (fourth edition)
LGSWE	Longman Grammar of Spoken and Written English
LOCNESS	*LOuvain Corpus of Native Speaker Essays*
LogL	Log-likelihood statistical test
MED2	Macmillan English Dictionary (second edition)
MI & MI^3	Mutual information
MLD	Monolingual learners' dictionary
MWU	multi-word unit
NS	Native speaker
NNS	Non-native speaker
pmw	Per million words
POS	Part-of-speech
SLA	Second Language Acquisition
SLDMs	Second Language Discourse Markers (cf. Siepmann 2005)
STUD-US-ARG	Sub-part of *LOCNESS* (American students, argumentative essays) used in this thesis
TRILLED	*TRILingual Louvain corpus of EDitorials*
USAS	UCREL Semantic Analysis System
VIEW	Variation in English Words and Phrases
WST4	WordSmith Tools 4
X^2	Chi-square statistical test

List of tables

List of figures

0. Introduction

This thesis originates in my research within the framework of the Concerted Research Action (ARC) project entitled 'Foreign Language Learning: Phraseology and Discourse' (No. 03/08-301) funded by the Communauté française de Belgique. The project has three main objectives:

- Theoretical objective: the project aims to study both developmental and cross-linguistic influence on English language learning in two relatively neglected areas, i.e. phraseology and discourse, at advanced proficiency levels.
- Methodological objective: the project seeks to demonstrate the importance of corpus data and methodology for analysing features of written interlanguage lexis and discourse.
- Applied objective: the project aims to reinforce the link between theory and practice in Foreign Language Learning (FLL) research, by taking classroom practice into account in FLL research and integrating research findings in teaching practice.

Within the framework of this project, this thesis deals with the **phraseology of English for Academic Purposes (EAP) vocabulary in learner writing**.

It meets the **theoretical** objective of the project by focusing on words and phrases that serve specific rhetorical functions in academic prose and which can thus be described as being at the interface between phraseology and discourse. In addition, it seeks to distinguish between features which are shared by several learner populations and are therefore likely to be developmental or teaching-induced and those which are L1-specific, and therefore possibly transfer-related.

The **methodological** objective is achieved by exploiting corpus data and corpus linguistics tools for a number of purposes. This thesis first proposes a new methodology based on a corpus-driven approach to select EAP vocabulary. It then makes use of a number of corpus-handling tools to analyse texts produced by native speakers of English and EFL learners and highlights their advantages and limitations. It also relies on statistical packages to test statistical significance and shows that a number of tests are available for multiple corpus comparisons.

This thesis also partly meets the project's **applied** objective by providing new pedagogical materials based on the findings that emerged from the corpus analyses.

The initial aim of this thesis was to focus on EFL learners' use of EAP vocabulary. For lack of detailed descriptions of the phenomenon in native academic prose, however, it was necessary to start with thorough analyses in native academic writing.

The seven chapters of the thesis are organised as follows:

Chapter One aims to characterize English for Academic Purposes. It gives a general account of its distinctive linguistic features before focusing on vocabulary needs in academic settings. A distinction is made between receptive and productive vocabulary and the term 'academic vocabulary' is defined in the light of its particular nature and role in academic discourse. It then offers a review of corpus-based studies of vocabulary in professional academic discourse, learner writing and native student writing.

Chapter Two deals with the fuzzy boundaries of the phraseological spectrum. It successively addresses the questions of the categorization of phrasemes and their defining criteria before presenting the typology and definitions adopted in this thesis. It then sheds light on the dual nature of the term 'collocation' and argues that 'to continue to thrive', the phraseological approach and the distributional approach to collocation will need to agree on a common terminology. Finally, it focuses on the relationship between discourse and phraseology as phrasemes with specific rhetorical functions have been reported to be typical of academic discourse.

Chapter Three reviews major findings about the influence of the first language on single words and phrasemes in the foreign language. It then focuses on some methodological flaws of transfer studies which cast doubt on the validity of several of the findings discussed in the first part of the chapter.

Chapter Four is devoted to a detailed description of the corpora, methods and software programs used in this thesis. We make use of corpora of professional academic prose, learner writing and native student writing and compare them within the framework of the *Integrated Contrastive Model.* We adopt a quantitative approach to phraseology and describe the extraction procedure used after discussing a number of parameters that can influence the results of a co-occurrence analysis.

Chapter Five is dedicated to the selection of words that are typical of academic texts and which will provide a basis for the comparison of native speaker and EFL learner academic writing. It proposes a new methodology based on the criteria of keyness, range and evenness of distribution.

Chapter Six aims to test our working hypothesis that upper-intermediate to advanced EFL learners, irrespective of their mother tongue backgrounds, share a number of linguistic features that characterize their academic writing. It focuses on words and phrasemes that expert writers and EFL learners use to serve typical organizational or rhetorical functions. The function of exemplification is presented in detail so as to serve as an illustration of the type of data and results obtained when examining the whole range of lexical strategies available to EFL learners, as opposed to expert writers, when they want to establish cohesive links in their essays. Other functions were similarly described but the numerous analyses conducted are not presented in detail in this thesis as they would become tedious for the reader. Instead, the focus is placed on the general interlanguage features that emerge from these analyses, which fall into five categories: aspects of overuse and underuse, register-awareness, phraseological patterns, semantic misuse and sentence position. Important pedagogical implications related to teaching practices and the role of corpora in materials design are then examined.

Chapter Seven is largely methodological in nature. It first seeks to demonstrate how learner corpus data can be used to select interlanguage features worthy of investigation in transfer studies. It then tries to implement Jarvis's (2000) unified framework to investigate L1 influence on the basis of corpus data and investigates their potential to uncover new types of evidence of transfer. It proposes to employ statistical techniques based on comparisons of means – ANOVA tests and the Dunnett post-hoc test - to operationalize the first two effects of L1 influence described in the framework. The major advantages and limitations of the approach are highlighted.

The thesis ends with a general conclusion, which briefly summarizes the major findings and offers several avenues for future research.

For the reader's convenience, appendices are included in a separate companion volume.

1. English for Academic Purposes and EAP vocabulary

1.1. Introduction

As stated by Hyland and Hamp-Lyons (2002:1), "the field of English for Academic Purposes [EAP] has developed rapidly over the past 25 years to become a major force in English language teaching and research". This rapid development stems from at least two reasons. First, English has become the lingua franca of academic communication and knowledge dissemination (cf. Witt 2000; Pecman 2004). University students thus need to have good command of English if they want to have access to a wide range of publications in their disciplines. Second, an increasing number of students undertake tertiary studies in English-speaking countries and host institutions need to provide them with special tuition in English use in academic settings[1]. Biber argues that "[s]tudents who are beginning university studies face a bewildering range of obstacles and adjustments, and many of these difficulties involve learning to use language in new ways" (Biber 2006:1). This is especially true of EFL learners.

Section 1.2 first proposes a definition of English for Academic Purposes and answers one of the major criticisms that have been levelled at the field. It then reviews some of the major distinctive linguistic features of academic discourse. Vocabulary has often been reported to cause major problems to EAP students and is the focus of section 1.3. After discussing theoretical and pedagogical aspects of vocabulary in EAP reading and writing, section 1.4 reviews a selected list of corpus-based studies of vocabulary in EAP writing.

1.2. Delimiting the scope of English for Academic Purposes

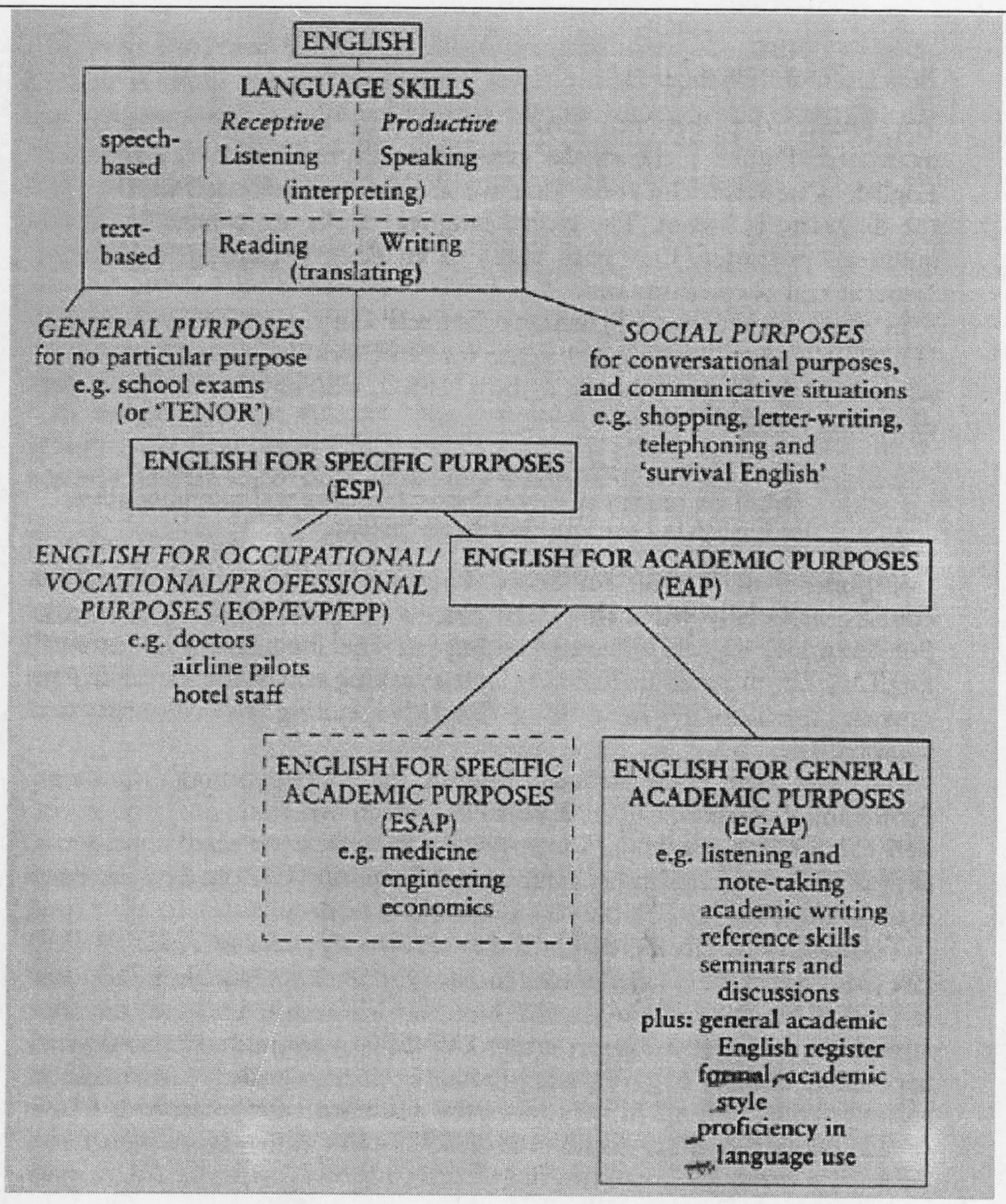
English for Academic Purposes is generally defined as being concerned with "the teaching of English for use in academic contexts, to students for whom English is an additional language, and who are preparing to begin a course of academic studies, or who are currently engaged on a course" (Thompson 2006). Jordan (1997:4) describes EAP as a sub-discipline of the larger field of **English for Specific Purposes** (ESP) and contrasts it with the teaching of **English for Occupational/Vocational/Professional Purposes** (EOP/EVP/EPP). One of the strongest links between these three fields is "the emphasis that practitioners give to needs analysis as a systematic way of identifying the specific sets of skills, texts, linguistic forms, and communicative practices that a particular group of learners must acquire" (Hyland and Hamp-Lyons 2002: 5). The field of EAP further divides into **English for General Academic**

[1] There are also an increasing number of English-taught degrees in Europe, Asia, etc.

Purposes (EGAP) and **English for Specific Academic Purposes** (ESAP) as illustrated in Figure 1.1 (see also Flowerdew L. 2002). EGAP is concerned with "the teaching of the skills and language that are common to all disciplines" (Dudley-Evans and St. John 1998:41) and focuses on "a *general academic English register*, incorporating a *formal, academic style*, with *proficiency in the language use*" (Jordan 1997:5). ESAP refers to the teaching of the subject-specific English that is needed for a particular academic subject, e.g. chemistry, together with its disciplinary culture.

The status and usefulness of EGAP has been questioned by Hyland who believes that "academic literacy is unlikely to be achieved through an orientation to some general set of trans-disciplinary academic conventions and practices" (Hyland 2000:145) and that "students actually have to readjust to each discipline they encounter (ibid). In a later article, he further argues that "[d]isciplines have different views of knowledge, different research practices, and different ways of seeing the world, and as a result, investigating the practices of those disciplines will inevitably take us to greater specificity" (Hyland 2002a:389). He thus concludes that "the teaching of specific skills and rhetoric cannot be divorced from the teaching of a subject itself because what counts as convincing argument, appropriate tone, persuasive interaction, and so on, is managed for a particular audience" (ibid:390).

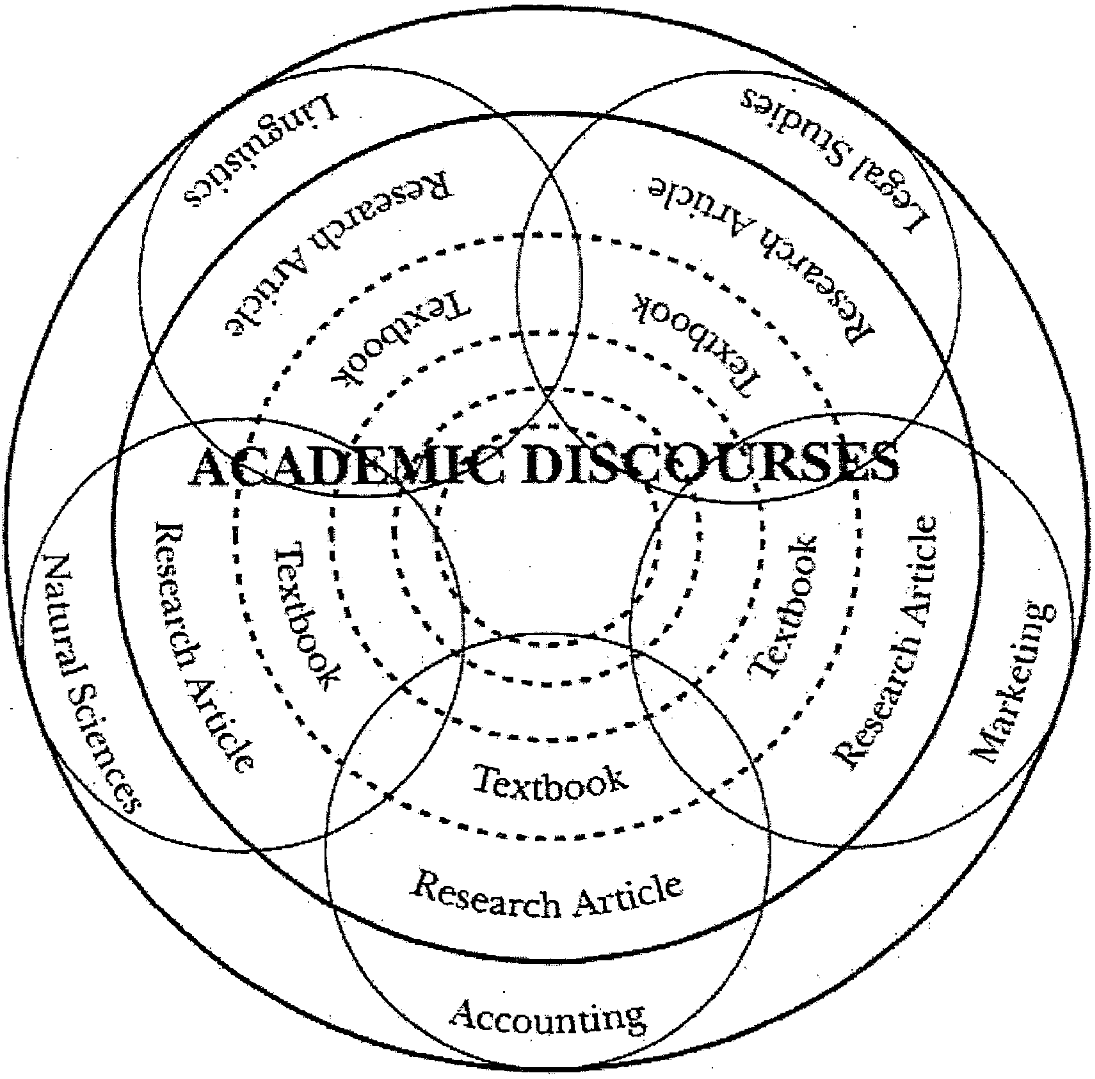
Figure 1.1: Jordan's (1997: 3) description of the many purposes of English



With the recent development of specialised genre-based corpora (see section 1.4.1), the field of academic discourse research has witnessed the burgeoning of studies on **variability** within academic texts. Studies have shed light on differences between several genres within the same academic discipline (e.g. Conrad 1996). Others have described differences in the same genre across several disciplines (e.g. Hyland 2000; Peacock 2006; Fløttum et al. 2006a) and even sub-disciplines. Ozturk (2007), for example, analyses the textual organization of research article introductions in two sub-disciplines of applied linguistics, namely second language acquisition and second language writing research. Some studies have also compared the use of linguistic features across text sections (e.g. Biber and Finegan 1994; Martínez

2003; Moreno 2004). These numerous studies rather support Hyland's (2002a) case for specificity. They have also made scholars such as Bhatia note that "one may find it difficult to conceptualise academic discourse as a single entity with an identifiable common core; it may be more realistic to represent the variations quite legitimately in terms of academic discourses, in the plural rather than the singular" (Bhatia 2002: 34) as in Figure 1.2.

Figure 1.2: Bhatia's (2002) variations in academic discourse



On the other hand, Biber's (1988) study of variation across speech and writing has shed light on distinctive characteristics of academic texts on the basis of an analysis of a large number of linguistic features (e.g. private verbs, *that*-deletion, time adverbials, conjuncts, agentless passive structures, third person pronouns, nominalisations) and their co-occurrences. His multi-dimensional study has shown that academic texts typically have an informational and non-narrative focus; they require highly explicit, text-internal reference and deal with abstract, conceptual or technical subject matter (cf. Biber 1988: 121-160). Similarly, the *Longman Grammar of Spoken and Written English*[2] (LGSWE) (Biber et al. 1999) provides a comprehensive description of the range of distinctive grammatical features of academic prose, in comparison to conversation, fiction and newspaper reportage. Common features of academic prose include a high rate of occurrence of nouns, nominalisations, noun phrases with modifiers, attributive adjectives, derived adjectives, activity verbs, verbs with inanimate subjects, agentless passive structures and linking adverbials. By contrast, first and second person pronouns, private verbs, *that*-deletions and contractions occur very rarely in academic texts. See Appendix 1.1 for a comprehensive list of grammatical features that are especially common in academic prose.

By highlighting a number of shared features across academic texts, Biber's (1988) multi-dimensional study as well as later corpus-based analyses of academic prose have conferred legitimacy to the concept of 'academic discourse' or 'E(G)AP'. At the same time, a large number of studies in EFL teaching and learning have proved the usefulness of E(G)AP programmes at university. They have pointed out that students often fail to recognize and appropriately use the conventions and linguistic features of academic prose (e.g. Johns 1997; Chang and Swales 1999; Hinkel 2002). Studies in second language writing have also established that learning to write L2 academic prose requires an advanced linguistic competence, without which learners simply do not have the range of lexical and grammar skills required in academic writing (Jordan 1997; Nation and Waring 1997; Hinkel 2002; Hinkel 2004; Reynolds 2005). A recent questionnaire survey of almost 5000 undergraduates from all 26 departments at the Hong Kong Polytechnic University has shown that students experience difficulties with writing skills that are necessary for studying content subjects through the medium of English (cf. Evans and Green 2006). Almost 50% of the students

[2] The LGSWE is a reference grammar of English based on the *Longman Spoken and Written English Corpus*, comprising approximately 5 million words for each genre (see Biber et al 1999: 24-35). The academic prose sub-corpus consists of academic books and research articles from a number of academic disciplines, e.g. medicine, sociology, law.

reported that they encountered difficulties in using appropriate academic style, expressing ideas in correct English and linking sentences smoothly.

Hinkel (2002:257-265) argues that the exclusive use of a process-writing approach, the relative absence of direct and focused grammar instruction and the lack of academic vocabulary development contribute to a situation in which EFL learners are simply not prepared to write academic texts. She provides a list of priorities in curriculum design and writes that, among the top-tier priorities, "NNSs need to learn more contextualized and advanced academic vocabulary, as well as idioms and collocations to develop a substantial lexical arsenal to improve their writing in English" (Hinkel 2002:247). Two decades ago, the same view was already formulated by Saville-Troike in an article on second language teaching for academic achievement: "Vocabulary knowledge is the single most important area of second language competence when learning content through that language is the dependent variable" (1984:199).

1.3. Vocabulary in academic discourse

Vocabulary has been shown to vary relative to text type or genre of writing (cf. Carter 1998; Tribble 1998; Ljung 2002) and to be a strong indicator of whether writers have adopted the conventions of the relevant discourse community (cf. Nation 2001:178). Several studies have investigated the type of vocabulary foreign language learners need to succeed at university. Section 1.3.1 discusses vocabulary needs in EAP reading and makes a distinction between three types of vocabulary, namely core vocabulary, academic vocabulary and technical terms. Section 1.3.2 distinguishes between receptive and productive vocabulary and redefines the term of 'academic vocabulary' in the light of its particular nature and role in academic discourse.

1.3.1. Vocabulary needs in EAP reading

There has been a continuing interest in whether there is a threshold which marks the boundary where vocabulary knowledge becomes sufficient for adequate reading comprehension (e.g. Laufer 1992; Hirsch and Nation 1992; Hu and Nation 2000; Cobb and Horbst 2001). Research by Laufer (1989)[3] has shown that **at least 95 per cent coverage** is needed to ensure reasonable reading comprehension of a text. It has also been shown that text coverage is

[3] Quoted in Nation (2001:146)

largely dependent on text type, length of text and homogeneity of text (cf. Nation 2001:146): the 2,000 most frequent words of English, for example, provide poorer coverage of academic texts than they do of fiction and conversation.

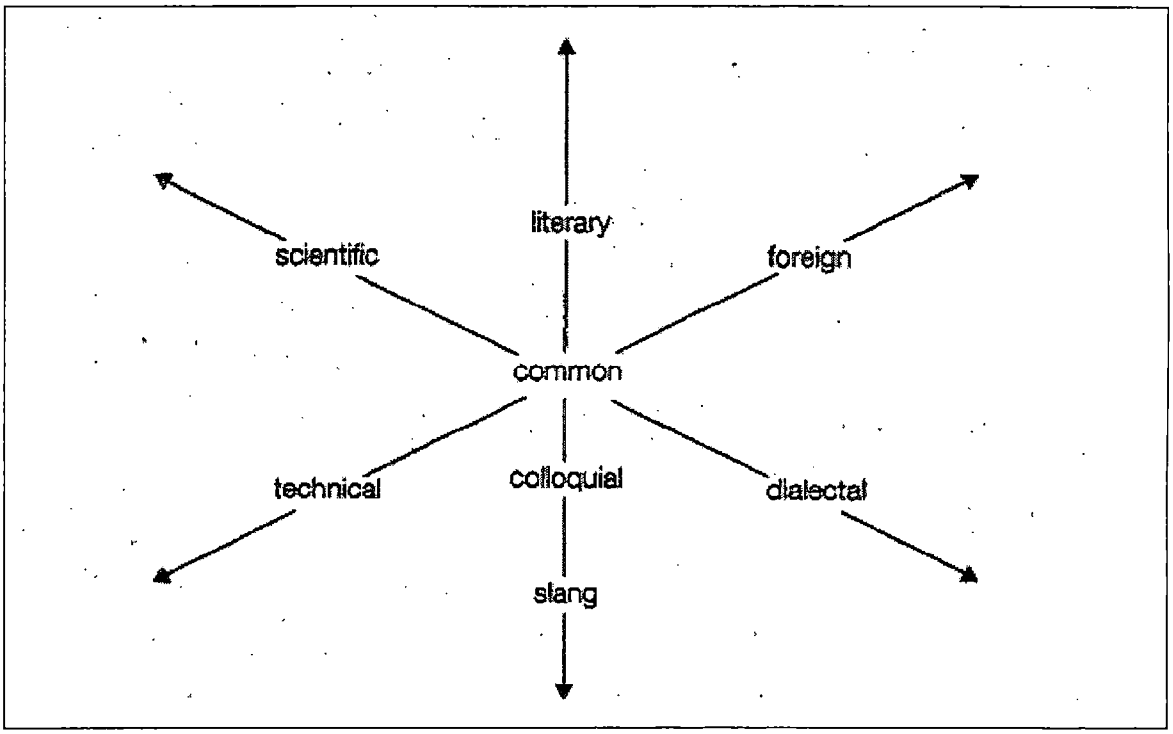
Several studies of vocabulary needs in EAP have thus focused on which vocabulary will provide 95% coverage of academic texts. Research by Nation and his colleagues (see Nation 2001:187-205) suggests that the 95% criterion can be achieved for academic texts if readers have knowledge of the following sets of words: a core vocabulary, an 'academic' vocabulary and technical or domain-specific terms. These three categories are described in the following sections.

1.3.1.1. Core vocabulary

A **core** or **basic** or **nuclear** vocabulary consists of words that are of high frequency in most uses of the language. It includes the most useful function words (e.g. *a, about, be, by, do, he, I, some* and *to*) and content words like *bag, lesson, person, put* and *suggest.* Stubbs (1986:104) describes nuclear words as an essential common core of "pragmatically neutral words" (cf. Figure 1.3). The author lists 5 main reasons for their pragmatic neutrality (1986:104-106):

1. Nuclear words have a "purely conceptual, cognitive, logical or propositional meaning, with no necessary attitudinal, emotional or evaluative connotations" (ibid 104).
2. Nuclear words have no cultural or geographical associations.
3. Nuclear words give no indication of the field of discourse from which a text is taken, i.e. its domain of experience and social settings.
4. Nuclear words are also neutral with respect to tenor and mode of discourse: they are not restricted to formal or informal usage and to specific medium of communication, e.g. written or spoken language.
5. Nuclear words are used in preference to non-nuclear words in summarizing tasks.

Figure 1.3: Stubbs's organization of English vocabulary (1986:102)



The best-known list of core words is West's (1953) *General Service List of English Words* (GSL). It was created from a 5 million word corpus of written English and contains around 2,000 word families. Percentage figures are given for different word meanings[4] and parts of speech of each headword. Frequency was not the only factor taken into account in making the selection: other criteria such as necessity and stylistic level were also used (West 1953:ix-x). West also wanted the list to include words that are often used in the classroom or that would be useful for understanding definitions of vocabulary outside the list. The GSL has had a wide influence for many years and served as a resource for writing graded readers and other material.

A number of criticisms have been levelled at the GSL, most particularly at its coverage and age. Engels (1968) criticizes the low coverage of the second 1,000 word families. While the first 1,000 word families cover between 68 and 74 percent of the running words in the ten 1,000 word texts he analyzes, the second set of word families in the GSL provides coverage of less than 10 per cent of the texts. It has also been argued that, because of changes in the English language and developments in curriculum design, the GSL contains many words that are considered of limited utility today (e.g. *crown, coal, ornament* and *vessel*) but does not contain highly frequent words such as *computer, astronaut* and *television* (see Nation and

[4] The *General Service List* and the *Cambridge English Lexicon* (Hindmarsh 1980) are the only two lists that give separate information on the different meanings of words.

Hwang 1995:35-36; Leech et al. 2001:ix-x; Carter 1998:207). However, several researchers have pointed out that, for educational purposes, it still remains the best of the available lists because of "its information on frequency of each word's various meanings, and West's careful application of criteria other than frequency and range" (Nation and Waring 1997:13).

In a variety of studies, the GSL provided coverage of up to 92% of fiction texts (e.g. Hirsh and Nation 1992), and up to 76% of academic texts (cf. Coxhead 2000)

1.3.1.2. Academic vocabulary

Most studies of academic reading comprehension have shown that it is neither high-frequency words nor technical terms that pose most difficulty to university students. The receptive vocabulary problems students most frequently encounter are predominantly related to what has commonly been referred to as **sub-technical** or **academic** vocabulary, i.e. a rather formal vocabulary with a middle frequency of occurrence across texts of various disciplines (cf. Cohen et al 1988; Corson 1997). A number of academic word lists have been compiled to meet the specific vocabulary needs of students in higher education settings (Campion and Elley 1971; Praninskas 1972; Lynn 1973; Ghadessy 1979; Xue and Nation 1984). The *Academic Word List* (Coxhead 2000) is the most widely used today in language teaching, testing and materials development. It is now included in academic vocabulary textbooks, CALL materials, the new *Longman Exams Dictionary* and in the latest edition of the *Collins Cobuild Advanced Dictionary of American English.*

The *Academic Word List* (AWL) was created from a corpus of 414 academic texts by more than 400 authors, totalling around 3.5 million words. The Academic Corpus includes journal articles, chapters from university textbooks and laboratory manuals. It is divided into four sub-corpora of approximately 875,000 words representing broad academic disciplines: arts, commerce, law and science. Each sub-corpus is further subdivided into seven subject areas as shown in Table 1.1.

Table 1.1: Composition of the Academic Corpus (Coxhead 2000:220)

	Discipline				
	Arts	Commerce	Law	Science	Total
Running words	883,214	879,547	874,723	875,846	3,513,330
Texts	122	107	72	113	414
Subject areas	Education History Psychology Politics Psychology Sociology	Accounting Economics Finance Industrial relations Management Marketing Public policy	Constitutional Criminal Family and medicolegal International Pure commercial Quasi-commercial Rights and remedies	Biology Chemistry Computer science Geography Geology Mathematics Physics	

Like the *General Service List*, the *Academic Word List* is made up of word families. Each family consists of a headword and its closely related affixed forms according to level 6 of Bauer and Nation's (1993) scale, which includes all the inflections and the most frequent and productive derivational affixes (see section 1.3.1.4 for a discussion and Table 1.4 below for examples). Coxhead (2000) selected word families to be included in the AWL on the basis of three criteria:

1. **Specialised occurrence**: a word family could not be in the first 2,000 most frequent words of English as listed in West's (1953) *General Service List.*
2. **Range**: a word family had to occur in all 4 disciplines with a frequency of at least 10 in each sub-corpus and in 15 or more of the 28 subject areas.
3. **Frequency**: a word family had to occur at least 100 times in the Academic Corpus.

The resulting list consists of **570 word families** which are reasonably frequent in most academic texts and which are not closely connected with any particular subject area or discipline. The AWL covers at least 8.5 per cent of the running words in academic texts. By contrast, it accounts for a very small percentage of words in other types of texts such as novels, which suggests that the AWL's word families are closely associated with academic writing (Coxhead 2000:225).

The AWL is divided into **10 sublists ordered according to decreasing word family frequency**. Some of the most frequent word families included in sublist 1 are represented by the headwords *analyse, benefit, context, environment, formula, issue, labour, research, significant* and *vary*. Examples of the least frequent word families as found in sublist 10 are *assemble, colleague, depress, enormous, likewise, persist* and *undergo.*

1.3.1.3. Technical terms

Domain-specific or **technical** terms are words whose meaning requires scientific knowledge. They are typically characterized by semantic specialisation, resistance to semantic change and absence of exact synonyms (cf. Mudraya 2006:238-239). As explained by Nation (2001:203), several practitioners consider that it is not the English teacher's job to teach technical terms. Technical words are best learned through study of the body of knowledge that they are attached to. Language teachers do not know the scientific fields and may have a great deal of difficulty with technical words. By contrast, learners who specialize in the field may have little difficulty in understanding these words (Strevens 1973:228).

Since technical terms are highly subject-specific, it is possible to identify them on the basis of their frequencies of occurrence, range and distribution (see sections 5.3.1.2 and 5.3.1.3) and to use them as a way of characterizing text types (Yang 1986). Technical terms occur with very high or at least moderate frequency within a very limited range of texts (Nation and Hwang 1995). In biology, for example, we find words like *alleles, genotype, chromatid, cytoplasm* and *abiotic*. These words are very unlikely to occur in texts from other disciplines or subject areas. Technical vocabulary is difficult to quantify. However, according to Coxhead and Nation (2001), technical dictionaries contain probably 1,000 headwords or less per subject area. Research suggests that knowledge of domain-specific or technical terms gives about 5% additional coverage of academic texts in a specific discipline (cf. Coxhead and Nation 2001).

1.3.1.4. Criticisms levelled at the approach

A first criticism is levelled at the commonly held assumption that learners need a general service vocabulary supplemented by an academic vocabulary and technical words in order to understand academic texts reasonably well. Ward (1999) investigates the size of vocabulary EAP engineering students need in order to read engineering textbooks in English and argues that, since there is often limited time available to facilitate student textbook reading, "there seems on the face of it an inherent contradiction in using a general list for learners with specific purposes" (Ward 1999:310). He then devises an engineering word list of 2,000 word families which contains both technical terms and all the general words necessary and shows that it provides 95% coverage in many basic engineering texts. Another example of such an approach can be found in Mudraya (2005).

Second, vocabulary categories have been described as if they are clearly separable but the boundaries between them are fuzzy (cf. Yang 1986; Mudraya 2005; Beheydt 2005). As Nation and Hwang (1995:37) remark, "any division is based on an arbitrary decision on what numbers represent high, moderate or low frequency, or wide or narrow range, because vocabulary frequency, coverage and range figures for any text or group of texts occur along a continuum." Chung and Nation (2003) investigate what kinds of words make up technical vocabulary in anatomy and applied linguistics texts. They classify technical terms on a four-level scale designed to measure the strength of the relationship of a word to a particular specialised field. Table 1.2 illustrates the resulting scale for the anatomy vocabulary. Chung and Nation consider items at steps 3 and 4 to be technical terms; items at steps 1 and 2 are not. They show that a considerable number of technical words belong to the 2,000 most frequent word families of English and the AWL. In the anatomy texts, 16.3% of the word types at step 3 are from the GSL or AWL. This percentage increases to 50.5% in the applied linguistics texts. A major output of this study is that a word can only be described as general service, academic or technical in context. This leads the authors to point out that, since most technical vocabulary needs to be learned productively by learners specializing in that area and since the technical use of a word "involves a collocation or a grammatical form that differs from its other uses" (ibid:113), technical words should be learned together with common collocation and grammatical patterns.

Table 1.2: Chung and Nation's (1993) rating scale for finding technical terms (2003:105)

Step 1 Words such as function words that have a meaning that has no particular relationship with the field of anatomy, that is, words independent of the subject matter. Examples are: *the, is, between, it, by, 12, adjacent, amounts, common, commonly, directly, constantly, early* and *especially*
Step 2 Words that have a meaning that is minimally related to the field of anatomy in that they describe the positions, movements, or features of the body. Examples are: *superior, part, forms, pairs, structures, surrounds, supports, associated, lodges, protects.*
Step 3 Words that have a meaning that is closely related to the field of anatomy. They refer to parts, structures and functions of the body, such as the regions of the body and systems of the body. Such words are also used in general language. The words may have some restrictions of usage depending on the subject field. Examples are: *chest, trunk, neck, abdomen, ribs, breast, cage, cavity, shoulder, girdle, skin, muscles, wall, heart, lungs, organs, liver, bony, abdominal, breathing.* Words in this category may be technical terms in a specific field like anatomy and yet may occur with the same meaning in other fields and not be technical terms in those fields.
Step 4 Words that have a specific meaning to the field of anatomy and are not likely to be known in general language. They refer to structures and functions of the body. These words have clear restrictions of usage depending on the subject field. Examples are: *thorax, sternum, costal, vertebrae, pectoral, fascia, trachea, mammary, periosteum, hematopoietic, pectoralis, viscera, intervertebral, demifacets, pedicle.*

The arbitrariness of word frequency counts also accounts for the presence of topic-dependent headwords such as *adult, sex* and *transport* in the AWL. It should also be noted that not all researchers define academic or sub-technical vocabulary in relation to general service words. Researchers such as Cowan (1974), Martin (1976), Baker (1988), King (1989) and Cohen et al (1988) do not distinguish between general service words and academic or sub-technical words (see section 1.3.2.2 for more information).

1.3.2. Vocabulary needs in EAP writing

Most studies of vocabulary needs in EAP have focused on reading comprehension and more generally, on receptive vocabulary. Results of these studies, however, have often influenced vocabulary teaching for **both receptive and productive purposes**. Most recent textbooks on academic vocabulary, for example, focus on the receptive and productive use of words that belong to the AWL (cf. Obenda 2004; Schmitt and Schmitt 2005; Huntley 2006). However, it can be questioned whether learners need the same vocabulary for academic reading and writing. Already in 1937, West argued that "both as regards Selection and still more as regards detailed Itemization, there is a need of a divorce between receptive and productive work" (1937:437) and regretted that teachers were giving "composite lessons aiming at teaching reading and speaking simultaneously, whereas reading and speaking are the Hare and the Tortoise. Reading and speech bear the same relation to each other as musical appreciation and actual execution on the piano. The one is *Recognition of a lot*; the other is *Skill in using a little*" (ibid). In this section, the major differences in learners' needs of receptive vs. productive vocabulary are first discussed. The concept of academic vocabulary is then re-examined in the light of learners' specific vocabulary needs for productive purposes.

1.3.2.1. Receptive vs. productive vocabulary

When the terms receptive and productive are applied to vocabulary, they "cover all aspects of what is involved in knowing a word" Nation (2001:26). Table 1.3 shows that, at the most general level, knowing a word involves **form, meaning** and **use**. Aspects of receptive and productive knowledge however differ widely. From the point of view of productive knowledge and use, knowing a word involves:

- Being able to pronounce the word correctly
- Being able to spell it correctly
- Being able to construct it "using the right word parts in their appropriate forms" (Nation 2001:28)
- Being able to produce the word to express the intended meaning
- Being able to produce synonyms, hyponyms, hypernyms and antonyms for the word
- Being able to use the word in appropriate patterns
- Being able to use this word with words that commonly occur with it
- Being able to "decide to use or not the word to suit the degree of formality of the situation" (ibid)

Learning a word productively thus involves a wide range of aspects of knowledge and use, which have been found to be much more difficult than receptive learning (cf. Nation 2001:28-30). Learning vocabulary for productive purposes requires more time and repeated effort. **Selection** is thus a key issue in teaching vocabulary for academic writing and speaking. It is questionable whether all words from the AWL should be the focus of productive learning as is currently done in textbooks (e.g. Schmitt and Schmitt 2005; Huntley 2006) and CALL materials (see, e.g., Gillett's website about vocabulary in EAP[5]; Luton's Exercises for the Academic Word List[6] and Haywood's AWL Gapmaker[7]).

Table 1.3: What is involved in knowing a word (Nation 2001:27)

Form	spoken	R	What does the word sound like?
		P	How is the word pronounced?
	written	R	What does the word look like?
		P	How is the word written and spelled?
	word parts	R	What parts are recognisable in this word?
		P	What word parts are needed to express the meaning?
Meaning	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	constraints on use (register, frequency, etc.)	R	Where, when, and how often would we expect to meet this word?
		P	Where, when, and how often can we use this word?
Note: in column 3, R = receptive knowledge, P = productive knowledge			

Another question relates to the **commonly held assumption that learners already know general service words** and that teaching can focus on the AWL. A case study of the use of the general service noun *example* in a 150,000 word corpus of argumentative essays written by French learners has shown that learners do not know the range of patterns in which the word can be used and tend to rely almost exclusively on the conjunct *for example* (cf. Paquot 2007). A second reason why Coxhead's criterion of non-appearance in the GSL is not really appropriate when it comes to productive purposes is that lexical items may be included in the 2,000 most frequent words but used differently in EAP. For example, Partington (1998:98) has shown that a *claim* in academic or argumentative texts is not the same as in news reporting or legal report. Frequency can also be an issue. Nouns such as *example, problem, reason, argument* and *result* appear with particular high range and frequency in

[5] http://www.uefap.com/vocab/vocfram.htm
[6] http://web.uvic.ca/~gluton/awl/index.htm
[7] http://www.nottingham.ac.uk/~alzsh3/acvocab/awlgapmaker.htm

academic corpora but are not considered as EAP vocabulary by Coxhead as these items are already in the GSL.

The AWL, as well as most word lists for learners of English, groups words into **families**. Other examples include the GSL, the *University Word List* (Xue and Nation 1984) and recent domain-specific word lists such as those developed by Ward (1999) and Mudraya (2005). Coxhead (2000:218) argues that this decision is supported by psycholinguistic evidence suggesting that morphological relations between words are represented in the mental lexicon (cf. Nagy et al 1989; Bertram et al. 2000). Although word families are useful for receptive purposes, not all members of a word family are likely to be as helpful in academic writing. For example, under the headword *item*, which has a relative frequency of 36 occurrences per million words in the Micro-Concord academic corpus (see section 5.2.1.1), we find the noun *itemisation* and word forms of the verb *itemise*. However, the verb *itemise* has a relative frequency of 1 occurrence per million words and the noun *itemisation* does not appear at all in the same corpus.

Additional problems relate to **meaning** and **part-of-speech**. Table 1.4 shows several word families of the AWL: the only information provided is that the words in italics are the most frequent form of their family. The headwords are the stem form of the words. Meanings and parts-of-speech are not differentiated. We do not know whether the word forms *issue* and *issues* (under the headword *issue*) are more often used as nouns or verbs in EAP. Similarly, the word family headed by the headword *stress* includes *stressed, stresses, stressful, stressing* and *unstressed* regardless of the fact that, for example, the noun *stress* can refer to 'a continuous feeling of worry' or to 'the special attention or importance given to a particular idea, fact or activity' (LDOCE4).

Table 1.4: Word families in the AWL

link	proceed	issue	evident	item	*stress*	utilise
linkage linkages linked linking links	procedural *procedure* procedures proceeded proceeding proceedings proceeds	issued *issues* issuing	evidenced *evidence* evidential evidently	itemisation itemise itemised itemises itemising *items*	stressed stresses stressful stressing unstressed	utilisation utilised utilises utilising utiliser utilisers *utility* utilities utilization utilize utilized utilizes utilizing

Like most other word lists of general service, academic or technical vocabulary, the AWL is based on **single words**. In a discussion of the factors that would need to be considered in the development of a resource list of high frequency words, Nation and Waring (1997:18) however suggest that some multi-word items should be included into word lists as they behave like high-frequency words. They further propose that set expressions be included under one of their constituent words. Such a view is supported by current trends in language acquisition and foreign language learning that stress the importance of prefabricated language over 'slot-and-filler' models of language (cf. Lewis 2002; Wray 2002; Schmitt 2004).

1.3.2.2. The nature and role of academic vocabulary

As discussed in section 1.3.1.2, 'academic' or 'sub-technical' vocabulary has often been used to refer to a rather formal vocabulary common to a wide range of academic texts but not so common in non-academic texts. Its definition has thus mainly been based on two criteria, namely **frequency** and **range** (cf. Coxhead 2000). The terms[8], however, have also been used in a more restricted sense to refer to words that "have in common a focus on research, analysis and evaluation – those activities which characterize academic work" (Martin 1976:92). In this thesis, the term 'academic vocabulary' will be used with this restricted sense. In this section, we will examine the nature and role of such a vocabulary in academic texts and discuss the problems academic words pose to learners of English.

Academic words "most probably occur because they allow academic writers to do the things that academic writers do. That is, they allow writers to refer to others' work (*assume, establish, indicate, conclude, maintain*); and they allow writers to work with data in academic ways (*analyse, assess, concept, definition, establish, categories, seek*)" (Nation 2001: 18). Martin (1976) reports that the vocabulary of the research process primarily consists of verbs, nouns and their co-occurrences (e.g. *state the hypothesis and expected results; present the methodology; plan, design the experiment; develop a model*). The vocabulary of analysis includes high-frequency verbs and two-word verbs that are "often overlooked in teaching English to foreign students but which graduate students need in order to present information in an organized sequence" (Martin 1076:93), e.g. *group, result from, derive, bring about, cause, base on, be noted for*. Adjectives and adverbs make up a large proportion of the vocabulary of evaluation.

[8] Other terms have also been proposed: non-technical terms (Goodman and Payne 1981), semi-technical vocabulary (Farrell 1990) and specialised non-technical lexis (Cohen et al. 1988).

Academic words have also been shown to play an important part in **discourse organization and cohesion** (cf. Halliday and Hasan 1976: 274-292; McCarthy 1991: 78-84; Partington 1998: 89-106; Nation 2001: 210-216). Very broadly speaking, it can be said that "these words provide a semantic-pragmatic skeleton for the text. They determine the status of the (more or less technically phrased) propositions that are laid down in it, and the relations between them" (Meyer 1997:9). Winter (1977), quoted in Carter (1998:83-85), distinguishes between three types of words that are commonly used to create cohesion or structure in discourse and that are basic to the understanding of academic texts:

- Subordinators that connect clauses (e.g. *except that, although, as far as, unless, whereas*);
- Sentence connectors which "make explicit the clause relation between the matrix clause and the preceding clause or sentence" (Winter 1977:15), e.g. *therefore, anyway, hence, for example, thus*)
- Words which serve to establish semantic relations in the connection of clauses or sentences in discourse.

The third group largely consists of **nouns** that are inherently unspecific and require lexical realization in their co-text, either beforehand or afterwards. Francis (1994) refers to this type of lexical cohesion as **advance and retrospective labelling** as they allow the reader to predict the precise information that will follow when they occur before their lexical realization and they encapsulate and package a stretch of discourse when they occur after their realization. Flowerdew J. (2003) refers to these abstract nouns as 'signalling' nouns. Examples of the most common labels found by Francis (1994) are *approach, area, aspect, case, matter, move, problem, stuff, thing* and *way.*[9] These nouns have traditionally been referred to as content words. However, when we encounter them in a text, we often need to do "something similar to what we do when we encounter words like *it*, *he* and *do* in texts: we either refer to the bank of knowledge built up with the author, look back in the text to find a suitable referent, or [look] forward, anticipating that the writer will supply the missing content" (Carter and McCarthy 1988: 206-207).

Within the category of labels, Francis further isolates a set of nouns which are "metalinguistic in the sense that they label a stretch of discourse as being a particular type of

[9] Francis's (1994) examples are not retrieved from academic texts but are from *The Bank of English* corpus (see www.titania.cobuild.collins.co.uk), and in particular, from a sub-selection of news articles.

language" (Francis 1994: 89). **Metalinguistic labels** are of four types "though there is some blurring and overlap between them" (ibid: 90):

1. **Illocutionary** nouns are nominalizations of verbal processes and usually acts of communication, e.g. *advice, answer, argument, assertion, claim, observation, recommendation, remark, reply, response, statement, suggestion*;
2. **Language-activity** nouns "refer to some kind of language activity or the results thereof. They are similar to illocutionary nouns, but they do not have cognate illocutionary verbs (though they may have cognate verbs)" (ibid: 91), e.g. *comparison, contrast, definition, description, detail, example, illustration, instance, proof, reasoning, reference, summary*, etc.;
3. **Mental process** nouns refer to "cognitive states and processes and the results thereof" (ibid: 92), e.g. *analysis, assumption, attitude, belief, concept, conviction, finding, hypothesis, idea, insight, interpretation, opinion, position, theory, thesis, view*, etc.
4. **Text** nouns refer to the formal textual structure of discourse, e.g. *phrase, words, quotation, excerpt, section, term*, etc.

As pointed out by Nation (2001: 212), the strength of labels "as discourse organising vocabulary [is] that they have a referential function and variable meaning like pronouns but, unlike pronouns, they can be modified by demonstrative pronouns, numbers, and adjectives, they can occur in various parts of a sentence and they have a significant constant meaning."

McCarthy has also shown that, as well as representing text segments, labels or what he calls **discourse-organising words**, "additionally give us indications of the larger text-patterns the author has chosen, and build up expectations concerning the shape of the whole discourse" (1991: 76). The author illustrates this claim with lists of words that typically occur in the problem-solution (cf. Table 1.5) and the claim-counterclaim patterns.

Table 1.5: The problem-solution pattern (McCarthy 1991:79)

Problem	concern, difficulty, dilemma, drawback, hamper, hinder, hindrance, obstacle, problem, snag
Response	change, combat (vb), come up with, develop, find, measure(s), respond, response
Solution/result	answer, consequence, effect, outcome, result, solution, (re)solve
Evaluation	(in)effective, manage, overcome, succeed, (un)successful, viable, work (vb)

These text-patterns have been referred to as clause relations (Winter 1994; Hoey 1994) or macro patterns (McCarthy and Carter 1994).

Focusing on receptive skills, McCarthy has also pointed out that "the language learner who has trouble with such words may be disadvantaged in the struggle to decode the whole text as efficiently as possible and as closely as possible to the author's designs. If the discourse-organising words are seen as signals of the author's intent, then inability to understand them or misinterpretation of them could cause problems" (McCarthy 1991:76). It may be hypothesized that EFL learners may encounter even more serious difficulty in using these discourse-organising words. However, very few studies have investigated their use in learner writing except for Flowerdew L. (1998; 2003) and Flowerdew J. (2006) (see section 1.4.2).

In addition to their problematic discourse-organising function, academic words, and more specifically signalling nouns, have been described as "likely to be problematic for non-native, as well as native speakers" (Flowerdew J. 2003:330) for a number of other reasons. First, they refer to **abstract** ideas and processes and introduce additional **propositional density** to a text (cf. Corson 1997). Second, the **polysemy** of some academic words adds to the difficulty (cf. Beheydt 2005) as well as their **higher register** (cf. Sonck-Mercier et al. 1991). Scarcella and Zimmerman (2005:127) also show that **mastery of derivative forms** makes academic words particularly difficult for L2 learners and report that learners often fail to analyze the different parts of complex academic words.

Several studies which have focused on academic words and their rhetorical functions in discourse have also pointed out that those words "should not be taught in isolation but in context and as central elements in typical collocations" (Baker 1988:103). Similarly, Francis shows that "there is a tendency for the selection of a label to be associated with common collocations. Many labels are built into a fixed phrase or 'idiom' (in the widest sense of the word), representing a single choice. Frequent collocations include, for example, 'the move follows ...', '... rejected/denied the allegations', '... to solve a problem', and '...to reverse the trend', where the retrospective label is found in predictable company (...). Even where the collocations are less fixed, the label occurs in a compatible lexical environment" (Francis 1994:100-101). Baker (1988) thus suggests that a frequency-based selection of academic single words should be supplemented with a **collocational study** of the resulting words so as to ensure that homographs be treated differently according to their various senses and that the nominal compounds and specialised multi-word units be identified as such (cf. Yang 1986).

1.3.3. Conclusion

Recent research on vocabulary needs in EAP has mainly revolved around the design of academic word lists for receptive purposes. Academic vocabulary has often been defined in terms of its frequency of occurrence and its distributional properties. We agree with Beheydt that "a more precise description of the construct of academic vocabulary is in order. That description is to be guided by the needs of learners and the criteria for word selection must be dictated by the practical problems encountered by users" (Beheydt 2005:243). It has been shown in section 1.3.2.2 that what characterizes a large proportion of academic words is their **discourse-structuring and cohesive function**. However, it remains an open question whether "it is possible to delimit a procedural vocabulary of such words that would be useful for readers/writers over a wide range of academic disciplines involving varied textual subject matters and genres" (McCarthy 1991:78).

EFL learners, as well as teachers and textbook developers, would greatly benefit from the elaboration of a productive counterpart to the *Academic Word List* as learners' needs and difficulties are clearly not the same in production as in reception. This would be reflected in the selection criteria of a productively-oriented academic word list. Although frequency remains an important criterion, it is only half of the story. A productively-oriented academic word list should also give L2 learners the lexical means necessary to do the things that academic writers do, e.g. stating a topic, hypothesizing, contrasting, exemplifying, explaining, evaluating, etc. Such a list would introduce new words together with information on how to use them, especially their patterns of use and phraseology as "particular collocations and grammatical patterns may be associated with particular functions of words" (Hoey 1993:82).

One of the aims of my dissertation is to propose a methodology for the selection of words that should be part and parcel of a productively-oriented academic word list.

1.4. Corpus-based studies of vocabulary in EAP

A corpus can be broadly defined as a collection of naturally occurring spoken or written data in electronic format, "selected according to external criteria to represent, as far as possible, a language or language variety as a source of linguistic research" (Sinclair 2005). Computer corpora are analysed with the help of software packages such as WordSmith Tools 4 (Scott 2004) which includes a number of text-handling tools to analyse textual data in quantitative and qualitative terms (see section 4.2.2.1.). Wordlists are used to study the frequency and distribution of the vocabulary – single words but also word sequences – used in one or more corpora. Wordlists for two corpora can be compared automatically so as to highlight the vocabulary that is particularly salient in a given corpus, i.e. its keywords or key word sequences. Concordances are used to analyse the co-text of a linguistic feature, i.e. its linguistic environment in terms of preferred co-occurrences and grammatical structures. Frequency is a key issue as corpus-based studies aim to provide automated descriptions of what is frequent and typical in the corpus under examination. The research paradigm of corpus linguistics is thus ideally suited for studying the linguistic features of academic discourse as it can highlight which words, phrases or structures are most typical of the genre and how they are generally used.

Corpus-based informed research has long been undertaken on academic sub-sections of general corpora (e.g. the 15 million word academic sub-corpus of the *British National Corpus*, cf. section 4.1.2.1). The last fifteen years or so, however, have seen a steady growth in the number and types of corpora that can be exploited in EAP. There has been a growing tendency among researchers to study the distinctive linguistic features of specific disciplines or genres, and with this trend various ESAP corpora have been compiled, e.g. the *Hyland Corpus* consisting of 240 research articles from eight disciplines (Hyland 1999). Similarly, bilingual corpora of academic texts are now becoming available to compare academic discourse and conventions across languages (cf. Siepmann 2005; Pecman 2004).

Student writing, and more particularly EFL learner writing, has often been the focus of EGAP corpus-based studies (cf. L. Flowerdew 2002:97). Two major learner corpora are the *International Corpus of Learner English* (cf. section 4.1.1) and the *Hong Kong University Science and Technology Learner Corpus*[10]. Learner corpus data have been compared with native student writing or sometimes with professional writing (see section 4.1 for a discussion). In order to investigate student ESAP writing, new corpora are currently being compiled: the *British Academic Written English* (BAWE) corpus[11] and the *Michigan Corpus of Upper-Level Student Papers* (MICUSP)[12] both consist of highly graded papers written by native and non-native students in several faculties. Academic spoken English has also become a major centre of interest to EAP practitioners with the development of the *Michigan Corpus of Academic Spoken English* (MICASE) and its British equivalent, viz. the *British Academic Spoken English* (BASE) corpus.

[10] See Pravec (2002) for a survey of English learner corpora.
[11] For more information, see http://www2.warwick.ac.uk/fac/soc/celte/research/bawe/
[12] For more information, see http://www.micusp.org/index.php?page=home

These relatively new types of corpora in the field of EAP have been used to analyse vocabulary in professional, student and EFL learner discourse across EAP genres, disciplines and media. The following sections review major findings of EAP corpus-based studies of **vocabulary in academic writing**. For more information on vocabulary in academic speech, see Biber (2003/2004/2006), Biber et al (2002), Biber et al (2003), Nesi (2002), Nesi and Basturkmen (2006), and Simpson and Mendis (2003).

1.4.1. Corpus-based studies of vocabulary in academic professional writing

Corpus-based studies have shed light on a number of distinctive linguistic features of academic discourse as compared with other genres such as conversation or news. As already mentioned above, Biber (1988) and Biber et al. (1999) show, for example, that nouns, nominalisations, derivational suffixes and linking adverbials are particularly frequent in academic prose while private verbs (e.g. *like, love, want, feel, hope*), *that*-deletions and contractions occur very rarely. Result and inference adverbials have been found to account for the largest proportion of linking adverbials in academic prose, directly followed by appositive, contrast/concession, enumerative/additive and summative adverbials (cf. Conrad 1999). Although the majority of linking adverbials are single adverbs, prepositional phrases (e.g. *for example, in other words, in addition, in conclusion, as a result*) and clausal linking adverbials (e.g. *that is, that is to say, what is more, to conclude*) are also relatively common in academic prose (Conrad 1999:11-12). In a study of adverbial marking of stance in speech and writing, Conrad and Biber also report that it is relatively common for academic prose to "overtly flag propositions for their degree of certainty or actuality" (Conrad and Biber 2000:66) with a relatively wide range of epistemic stance markers.

Biber et al (1999) compare the use of lexical bundles, i.e. "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (1999: 990), in academic prose and conversation and find that the structural types of lexical bundles in academic prose are radically different from those in conversation. Table 1.6 shows that almost 90% of all lexical bundles in conversation are composed of clause segments. A large proportion of these lexical bundles begin with a first person pronoun as subject together with a stative main verb (e.g. *I don't know what, I thought that was*). In academic prose, over 60% of all lexical bundles are parts of noun phrases or prepositional phrases (e.g. *the use of, the fact that*). When a lexical bundle is structurally complete, it is typically a prepositional phrase that functions as a linking adverbial, e.g. *on the other hand, in the same way*.

Table 1.6: Proportional distribution of four-word lexical bundles across the major structural patterns in academic prose and conversation (Biber et al 1999:996)

	CONV	ACAD	example
Patterns most widely used in conversation			
personal pronoun + lexical verb phrase (+complement clause)	44%	-	*I don't know what*
pronoun/NP (+ auxiliary) + copula *be* (+)	8%	2%	*it was in the*
(auxiliary +) active verb (+)	13%	-	*have a look at*
yes-no and *wh*-question fragment	12%	-	*can I have a*
(verb +) *wh*-clause fragment	4%	-	*know what I mean*
Patterns most widely used in academic prose			
noun phrase with post-modifier fragment	4%	30%	*the nature of the*
preposition + noun phrase fragment	3%	33%	*as a result of*
anticipatory *it* + VP/adjectiveP (+ complement clause)	-	9%	*it is possible to*
passive verb + PP fragment	-	6%	*is based on the*
(verb +) *that*-clause fragment	1%	5%	*should be noted that*
Patterns used in both registers			
(verb/adjective +) *to*-clause fragment	5%	9%	*are likely to be*
other expressions	6%	6%	
Total	**100%**	**100%**	

Biber et al. (1999:1014-1024) also describe the functions of lexical bundles. In academic prose, the lexical bundles consisting of a noun phrase followed by a post-modifying fragment are typically used to provide physical description, including identification of place, size and amount (*the shape of the, the position of the, the total number of the*); to mark existence or presence (*the presence of the*); to identify a variety of abstract qualities (*the nature of the*), to describe processes or events lasting over a period of time (*the development of the, the course of the*) or how a process occurs (*the way in which, the extent to which*) and to identify relationships among entities (*the difference between*). Most lexical bundles consisting of a prepositional phrase with an embedded *of*-phrase fragment functioning as post-modifier of the noun, mark abstract, logical relations (*as a result of, in the case of, in the absence of*). Lexical bundles beginning with an anticipatory *it* typically report the writer's stance (*it is possible to, it is important to*).

A high number of lexical bundles built around a verb phrase in academic prose consist of a passive voice verb followed by a prepositional phrase which typically marks a locative or logical relation, rather than introducing an agentive *by*-phrase. Examples include *are shown in table, is shown in figure* and *is based on the.* Lexical bundles with adjectival subject predicatives are used to identify causative relations (*may be due to*) or comparative relations (*is similar to the*). Lexical bundles with predicative adjectives controlling a *to*-clause are used to indicate possibility or certainty (*is likely to be, is not possible to*) while those with passive voice verb predicates controlling a *to*-clause are used to identify findings or known information (*has been shown to, was found to be*). Lexical bundles beginning with the subordinator *as* are used for deictic reference to other discourse segments (*as shown in figure, as we have seen*).

Unlike the clause-initiating bundles in conversation, those in academic prose have either the demonstrative *this*, or existential *there*, as subject, and copula *be* as the main verb. Lexical bundles with *this* as subject link the information that follows to the preceding discourse (*this is not to say that*) while those beginning with existential *there* are used for informational packaging purposes. Typical examples are lexical phrases about statistical significance or correlation such as *there was no significant difference between.*

Siepmann (2005) conducts a contrastive analysis of what he terms 'second-level discourse markers' (SLDMs) in English, French and German academic and journalistic texts. SLDMs are defined as "medium frequency fixed expressions or collocations of two or more printed words acting as a single unit. Their function is to facilitate the process of interpreting coherence relation(s) between elements, sequences or text segments and / or aspects of the communicative situation" (Siepmann 2005:52). They thus fall within the scope of what McCarthy referred to as 'procedural vocabulary' (see section 1.3.3). Siepmann shows that correspondences between SLDMs across languages cannot be inferred from structural similarities. He therefore develops a multilingual functional taxonomy which comprises 22 categories such as concession markers, exemplifiers, inferrers and digression markers (see Appendix 1.2).

Three categories of SLDMs are analysed in detail by Siepmann: exemplifiers, inferrers and reformulators and resumers. Exemplifiers, for example, were found to consist of semantically and pragmatically similar sets in English, French and German and to exhibit only a small degree of variation. Variation was most often found among German exemplifiers, which "may result from a general German tendency to ad-hoc formulation which stands in marked contrast with English and French reliance on stock phrases" (Siepmann 2005:141). Exemplifiers were also shown to occur with considerably higher frequency in French than English or German. One of the major findings of this study is the existence of 'collocational combinations' or 'long-distance collocations' between second-level discourse markers, such as *with this in mind* + *let us turn to* or *turning to* + *we find that.*

Siepmann's (2005) study can be praised for its originality and its significant theoretical implications (see section 2.5). It however suffers from methodological flaws in selection and identification of SLDMs[13]. SLDMs are distinguished from 'first-level' markers by means of a frequency criterion: they typically occur with a frequency of 3 to 50 tokens per 10 million words while first-level markers are much more frequent. Table 1.7 shows that *however, nevertheless, on the other hand, in contrast* and *by contrast* are classified as 'first-level' markers while *the fact remains that, note that, worse* and *the same goes for* are SLDMs. First-level markers are not analysed in Siepmann's (2005) study. Thus, while analyzing exemplifiers, Siepmann examines infrequent - if not rare - markers such as *as with; to paint an extreme example, consider; as an example* and *for the sake of illustration* (see Appendix 1.2) but does not consider *for example* and *for instance*. This decision is highly problematic for at least two reasons. First, we only get a partial picture of the lexical means used to serve rhetorical or organizational functions in academic prose. Second, Siepmann's conclusions about frequency differences across languages can be seriously challenged. The author claims that exemplifiers occur with considerably higher frequency in French than English or German and concludes that "there is thus empirical support for the hitherto unfounded claim (see Vinay and Darbelnet 1958:222) that, on average, French writers make more extensive use of connectors than their English or German counterparts" (Siepmann 2005:141). No such conclusion can be drawn on the sole basis of an analysis of SLDMs. First-level markers need to be taken into consideration as well.

Table 1.7: First-level and second-level discourse markers in a 10-million word academic corpus (based on Siepmann 2005:51)

Discourse marker token	Frequency
First-level discourse markers	
however	8897
nevertheless	805
on the other hand	751
in contrast/by contrast	902
Second-level discourse markers	
the fact remains that	13
note that	172
worse	43
the same goes for	5

Appendix 1.2 lists the SLDMs that the author uses to illustrate the 22 categories of his taxonomy. They include rare sequences such as *on inspection; for the sake of illustration; let us now see why; such instances could be multiplied; to round off the picture; the corollary of this is that; consider, for a digressive page or two, ...; how, then, may we explain; a moment's*

[13] See Stubbs (2006) for a critical review of the book.

reflection suggests that; it is a fair guess that; let us say that; to further confound the picture, etc. Siepmann's aim at exhaustiveness is arguably another methodological flaw for a study with **pedagogical and lexicographic objectives**. In addition, the register appropriateness of several of these examples can clearly be questioned.

As for Meyer (1997), he has used the academic section of the LOB corpus (see section 5.3.1.1) to investigate the contexts, senses and syntactic frames of verbs that centre around the acquisition of knowledge in the process of academic investigation, e.g. *find, note, observe, show, conclude, infer, prove*, as well as nouns and adjectives derived from them. The author reports that "there is a lot of ordinary language in technical discourse" (Meyer 1997:368) and that 'coming to know' verbs show "all the vaguenesses, polysemies, and ambiguities of everyday language" (ibid). These verbs, however, are "not concerned with trivial matters, but are used to discuss matters lying at the very heart of the scholarly process" (ibid). These findings further support our view that general service words need to be part of an EAP course (see section 1.3.2.1) and that great emphasis should be placed on phraseology in EAP (see 1.3.3).

A number of lexical studies have also focused on the vocabulary of academic genres and disciplines (see Table 1.8). Tribble (2000), for example, studies the keywords of project proposals to reveal salient features that are functionally related to that genre (see section 5.3.1.1 for more information on keywords); Bondi (2004) analyses the discourse functions of contrastive connectors in academic abstracts in economics, history and sociology; Howarth (1996) examines the phraseology of high-frequency verbs in social science texts while Chujo and Utiyama (2006) test several statistical measures to extract technical vocabulary from commerce and finance texts. Other studies have concentrated on the similarities and differences of vocabulary across genres and disciplines. Biber (2006) compares the distribution of word types, parts-of-speech, high-frequency words, specialized vocabulary and lexical bundles across a wide range of university genres[14]. Thompson and Tribble (2001) investigate the use of citation practices in PhD theses written in two departments (Agricultural Botany and Agricultural Economics) and Harwood

[14] It should be noted however that Biber does not use the term 'genre' but 'register' in his numerous studies.

Table 1.8: A selected list of discipline- or genre-based EAP corpus-studies

Study	Discipline	Genre	Linguistic features
Baker (1988)	Medicine	Research articles	Sub-technical vocabulary
Biber (1988)	Natural science, medicine, mathematics, social science, politics/education, humanities, technology/engineering	Research articles, reports, conference proceedings, book sections	Multidimensional analysis
Biber (2006)	Business, education, engineering, humanities, natural science, social science	Textbooks, course packs, course management, institutional writing, classroom teaching, etc.	Vocabulary use, expression of stance, lexical bundles
Bondi (2004)	History, economics and sociology	Abstracts	Contrastive connectors
Butler et al. (2004)	Mathematics, science and social studies	Textbooks	Lexical diversity, word frequencies, specialized and general vocabulary, etc.
Charles (2003)	Politics/international relations and materials science	Doctoral theses	The construction of stance through nouns
Charles (2006)	Politics/international relations and materials science	Doctoral theses	Phraseological patterns in reporting clauses used in citation
Chujo and Utiyama (2006)	Commerce and finance	Book sections and research articles*	Technical vocabulary
Conrad (1996)	Biology	Textbook sections and research articles	Multidimensional analysis
Cowie (1997)	Arts, belief and religion and social sciences	Book sections	Collocations
Curado Fuentes (2001)	Computer science, information science, telecommunications, audio-visual communication	Research articles, textbooks and technical reports	Lexical behaviour of words and collocations
Dahl (2004)	Economics, linguistics and medicine	Research articles	Metadiscourse
Fløttum et al (2006b)	Economics, linguistics and medicine	Research articles	'Let us' imperatives and metatextual expressions
Freddi (2005)	Linguistics	Introductory chapters of textbooks	Lexico-grammar of argumentation
Gledhill (2000)	Pharmaceutical science	Research articles	Collocations
Groom (2005)	History and literary criticism	Research articles and book reviews	It v-link ADJ that It v-link ADJ to-inf
Harwood (2005)	Physics, computing science, economics and business and management	Research articles	First person pronouns
Hiltunen (2006)	Literary criticism and law	Research articles	Coming-to-know verbs

Howarth (1996/1998)	Social science	Book chapters and research articles*	High-frequency verbs and their phraseology
Hyland (1998)	Microbiology, marketing, astrophysics, applied linguistics	Research articles	Metadiscourse
Hyland (2002b)	Mechanical engineering, physics, biology, electronic engineering, philosophy, marketing, applied linguistics, sociology	Research articles	Reporting practices
Hyland (2002c)	idem	Research articles, textbooks and reports	Imperatives
Hyland (2002d)	idem	Research articles	First person pronouns
Luzón Marco (1999)	Physics, biology, chemistry, geology, medicine*	Books and articles*	Procedural vocabulary
Luzón Marco (2000)	Medicine	Research articles	Collocational frameworks
Moreno (2004)	Business and economics	Research articles	Retrospective labelling
Mudraya (2005)	Engineering	Textbooks	Lexical syllabus for engineering students
Oakey (2002)	Social science, medicine and engineering	Research articles and book abstracts*	Lexical phrases
Peacock (2006)	Philosophy, marketing, applied linguistics, physics, mechanical engineering, sociology, biology, electrical engineering	Research articles	Boosters
Pecman (2004)	Chemistry, physics and biology	Research articles, abstracts, reports, etc.	EAP-specific phraseology
Stotesbury (2003)	Humanities, social sciences, natural sciences	Research article abstracts	Evaluation
Swales et al (1998)	Statistics, linguistics, experimental geology, philosophy, history, chemical engineering, art history, literary criticism, political science, communication studies	Research articles	Imperatives
Thompson (2001)	Agricultural botany and agricultural economics	Doctoral theses and research articles	Citation and modal verbs
Thompson (2005)	Agricultural botany	Doctoral theses	Intertextual reference
Thompson and Tribble (2001)	Agricultural botany and agricultural economics	Doctoral theses	Reporting practices
Tribble (2000)	-	Project proposals	Keywords
Vold (2006)	Economics, linguistics and medicine	Research articles	Epistemic modality

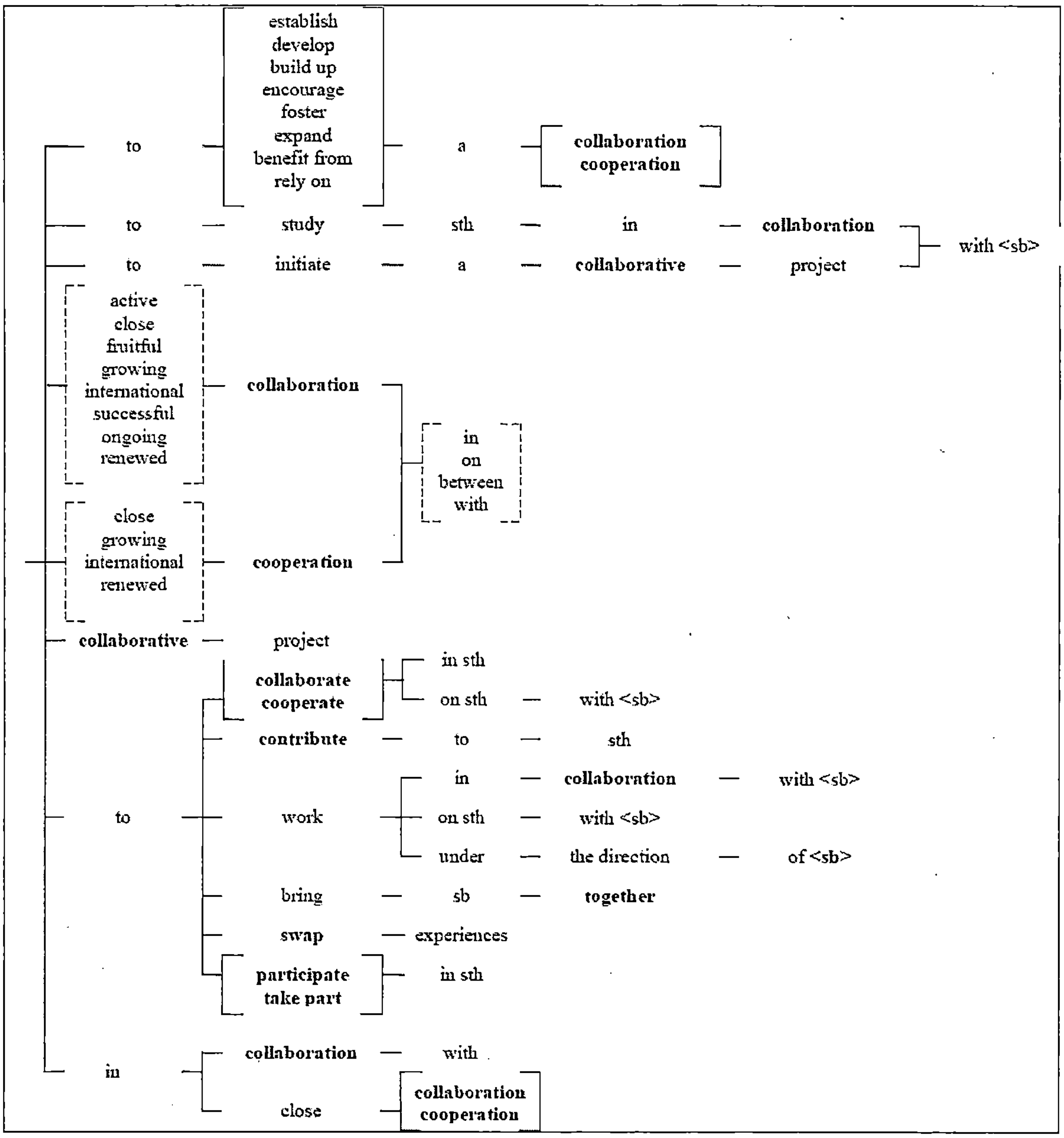
*not analysed separately in the study

(2005) reports on the use of personal pronoun *I* and inclusive and exclusive *we* in research articles across four disciplines. Table 1.8 shows that a number of these discipline- or genre-based studies have also focused on the phraseological preferences of academic prose. Cowie (1997) examines verbal collocations of a set of four abstract nouns – *attention, mind, consideration* and *thought* – in a 585,000 word corpus of arts, belief and religion and social science texts. A large proportion of the verb + noun co-occurrences analysed are restricted collocations. These collocations, however, are "restricted in their lexical structure but unrestricted in their application" (Cowie 1997:55). Cowie further argues that "it is precisely collocations of this kind which form the phraseological core of the vocabulary needed for academic written communication in English" (ibid).

Howarth (1996/1998) conducts an analysis of verb + noun collocations in a corpus of social science texts and shows that non-technical collocations are much more numerous than technical collocations. A large proportion of these non-technical collocations consist of a verb in a figurative sense (e.g. *apply, reach, obtain*) and an abstract noun denoting a recurrent concept in academic discussion (e.g. *approach, conclusion, finding, idea, method, result*). Howarth suggests that these collocations are an essential part of the procedural vocabulary of the academic discourse. The author further argues that it is "not idioms that learners need for effective communication" (Howarth 1996: 156), at least in academic settings. Learners need the lexical means that will allow them to conform to "the native stylistic norms for a particular register", which "entails not only making appropriate grammatical and lexical choices but also selecting conventional [multi-word units] to an appropriate extent" (Howarth 1998: 186).

Pecman (2004) postulates the existence of a language shared by scientists throughout the disciplines and analyses EAP-specific phraseological translation equivalents in a bilingual French-English corpus of research articles, abstracts and technical reports in chemistry, physics and biology. Examples of phraseological translation equivalents in General Scientific Language are *to [invalidate/refute] a hypothesis – [démentir/contredire] une hypothèse; working hypothesis – hypothèse de travail; in case of/in the event of – dans l'hypothèse de/où; to [advocate/defend] a theory – défendre une théorie*. The author classifies phraseological units into an ontology of 125 central concepts typical of general scientific discourse (e.g. OBSERVATION, RESULT, DEFINITION, METHODS, CONDITION, PROBABILITY, HYPOTHESIS), which results in the construction of collocational frameworks consisting of phraseological units belonging to the same semantic category. Figure 1.4 illustrates the collocational frameworks collected for the concept |COLLABORATION| in English.

Figure 1.4: Expressions of the concept |COLLABORATION| in English (from Pecman 2004:367)



As for Gledhill (2000), he examines the collocational behaviour of a set of **grammatical items** (e.g. *be, to, of*) in a 499,370 word corpus of pharmaceutical sciences articles. He argues that the differences in wording between different sections, viz. title, abstract, introduction, methods, results and discussion, must be interpreted "in terms of the textual and interpersonal functions of the text rather than simply in terms of propositional information" (Gledhill 2000:207). For example, most infinitive clauses of projection (clauses introduced by *to*, e.g. *has been shown to* ... + non-finite verb) occur in introductions, while projection is typically finite (*it has been shown that* + finite verb) in Abstracts and Discussion sections. Other grammatical items share associated phraseological roles throughout text sections, e.g. the construction of nominal groups (where *of* is a significant item) or the reformulation of immediately neighbouring discourse (*this*). Gledhill thus suggests that "while a grammatical item in the general language may have a largely unpredictable set of contexts, the corpus allows us to infer a very specific phraseology and system of lexico-grammatical relations for these words" (Gledhill 2000:207-208).

Some of these studies have also shown that the form, frequency and function of phraseological patterns may differ across genres and disciplines (e.g. Oakey 2002). They have also suggested that phraseological patterns correlate closely with the various communicative purposes that they serve in different genres or disciplines (e.g. Groom 2005; Charles 2006). Most studies, however, have highlighted the large extent of **common phraseological patterns across genres and disciplines** (e.g. Curado Fuentes 2001; Pecman 2004) and tend to support Gledhill's view that "there is a shared scientific voice or 'phraseological accent' which leads much technical writing to polarise around a number of stock phrases" (Gledhill 2000:204).

1.4.2. Corpus-based studies of vocabulary in learner academic writing

Learner corpora are a relatively recent addition to the wide range of existing corpus types. They consist of electronic collections of foreign or second language learner texts assembled according to explicit design critera (cf. section 4.1.1). As Granger explains, learner corpus data offer a number of advantages over other types of learner data: "they are usually quite large and therefore give researchers a much wider empirical basis than has ever been available before; they are stored in electronic format and can therefore be submitted to a wide range of automated methods and tools which make it possible to quantify learner data, to enrich them with a wide range of linguistic annotations (e.g. morpho-syntactic tagging, discourse tagging, error tagging) and to manipulate them in various ways in order to uncover their distinctive lexico-grammatical and stylistic signature" (Granger to appear). The method of *Contrastive Interlanguage Analysis*, which consists in two types of comparison, i.e. comparisons of learner language and one or more native speaker reference corpora (L2 vs. L1), and comparisons of different varieties of learner language (L2 vs. L2) (see section 4.2.1), has often been used by learner corpus researchers to bring out learners' specific features.

Unlike in corpus-based studies of native academic writing, learner corpus research has mainly focused on EGAP, and more precisely on argumentative writing (cf. L. Flowerdew 2002). Table 1.9 shows that only a few learner corpus-based studies have examined learner writing in ESAP settings, e.g. Flowerdew (1998), Hyland (2002b/c/d). Similarly, a large majority of studies have investigated English as a Foreign Language (EFL) rather than English as a Second Language (ESL) learner writing[15]. Notable exceptions are Hinkel's analyses of essays written by ESL students who have spent up to 9 years in US universities (cf. Hinkel 2002; 2003a; 2003b). One of the major findings of these studies is that ESL students may become fluent in English conversational discourse but "continue to have a restricted repertoire of syntactic and lexical features common in the written academic genre" (Hinkel 2003a: 1066). This repertoire largely consists of constructions prevalent in speech as well as everyday vocabulary items. Hinkel attributes this to the fact that, for most ESL university students, the greatest amount of exposure to English usage takes place in conversational discourse.

The advent of EFL learner corpus research can be said to have taken place with the publication of Granger's (1998) *Learner English on Computer*, a collection of pioneering papers on learner language largely based on the *International Corpus of Learner English* (cf. section 4.1.1). A large proportion of these studies compare corpora of texts produced by speakers of many different language groups with each other and with similar native productions to identify distinctive and shared features of a wide variety of interlanguages[16] (Ringbom 1998; Virtanen 1998; Petch-Tyson 1998; Aarts and Granger 1998; Biber and Reppen 1998; Gillard and Gadsby 1998). These studies all point to the fact that the "English of advanced learners from different countries with a relatively limited variation of cultural and educational background factors share a number of features which make it differ from NS language" (Ringbom 1998: 49). They all coincide in that the patterns of use in EFL learner writing are very similar to those typical of native conversation, but completely different from those found in academic texts.

Learners have proved to overuse many lexical and grammatical features that are typical of **speech**, such as personal pronouns and Germanic high-frequency verbs but significantly underuse many of the characteristics of academic prose, such as a high proportion of nouns and prepositions. In addition, they favour the use of **general and/or vague nouns** (*people, thing, problem*) to the detriment of many EAP-specific words such as *issue, belief, argument*

[15] EFL learning takes place in a setting in which English is not spoken in the local community: a French-speaking learner in Belgium is generally defined as a EFL learner. EFL learning generally takes place in a setting with formal language instruction. By contrast, ESL learning typically takes place in a setting in which English is the language spoken in the local community: the learning of English by a Japanese student in the USA would be an example of an ESL learner (cf. de Bot et al 2005:7).

[16] The term 'interlanguage' was introduced by Selinker and refers to "a separate linguistic system based on the observable output which results from a learner's attempted production of a TL [target language] norm" (Selinker 1972:214).

(cf. Granger and Rayson 1998). The latter findings are further supported by Petch-Tyson's (1999) investigation of retrospective labels in the writing of Dutch, French, Finnish and Swedish EFL learners in which she shows that learners typically underuse retrospective labels, particularly illocutionary, language activity and mental process labels (see section 1.3.2.2). The author argues that "these EFL writers are not equipped with the type of lexical knowledge necessary for the type of writing task they are undertaking" (1999:60).

Other studies in Granger (1998) focus on areas of difficulty in learner writing that necessitate more manual analysis, and more specifically on **collocational competence, semantic appropriacy** or **discourse features**. As a result, they are more limited in scope and analyse the writing of learners from the same mother-tongue background. Lorenz (1998), for example, examines the use of adjective intensifiers in four corpora of argumentative essays written by German teenagers (16-18), German university students of English (20-25), British teenagers (15-18) and British undergraduates (19-23). These four corpora make it possible to conduct a cross-sectional study[17] of the use of adjective intensifiers by German EFL learners. Lorenz shows that the most prominent function of adjective intensifiers for both German and British populations lies in intensifying qualities which are metalinguistic, i.e. "which mark off those parts of the text that the writer believes to be particularly important, different and interesting" (Lorenz 1998:59). However, German learners, and especially advanced German learners use far more intensifiers than the native speakers. The impression of **overstatement** thus created is further emphasized by learners' use of intensifiers in places where they are "semantically incompatible, communicatively unnecessary or syntactically undesirable" (ibid 64). German learners make use of adjective-intensifier collocations with a gradational irregularity. In example 1.1, the adjective *delicious* is ungradable and would normally call for a maximizer such as *absolutely*. Second, they use intensifiers when they are not essential to the argument. In example 1.2, the intensifiers even distract from the main point: emphasis is "misdirected and immaterial to the writer's concern about genetic engineering" (ibid 62). Third, they use adjective-intensifier collocations in theme position rather than in rheme position, "where one would expect to find the elements that are new, relevant and noteworthy enough to be intensified (ibid 62). This makes the subject noun phrase unnaturally heavy (example 3).

[17] Cross-sectional studies make use of data from several groups of learners representing, for example, different classes or grade levels, as representative of what learners can do at different proficiency levels. They are often used instead of longitudinal studies because of the difficulty of following the same learners over the years.

1.1. *Soon I experienced that in England you can have* ***very delicious*** *food – in Indian, Chinese or Mexican restaurants, where you can fully trust in the cooks.* (example from Lorenz 1998:60)

1.2. *It is possible to create any kind of baby – either* ***very intelligent*** *or* ***absolutely stupid****, either superior or inferior.* (example from Lorenz 1998:60)

1.3. *A* ***very difficult*** *question is whether one should have a hero at all, as I don't think that there is any person in the whole, wide world to whom one should look up admiringly.* (example from Lorenz 1998:60)

Altenberg and Tapper (1998) make use of three corpora to analyse Swedish learners' use of adverbial connectors: a corpus of argumentative essays written by Swedish learners, a corpus of essays written by British students and, to be able to compare the use of connectors in the Swedish learners' essays with native Swedish usage, a corpus of papers written in Swedish by native students of Swedish. Swedish learners tend to underuse resultative and contrastive conjuncts (e.g. *hence, therefore, thus, however, though, yet*), which all belong to the formal registers except for *though*. The authors suggest that the most likely explanation for this underuse is that the Swedish learners are less familiar with formal English adverbial connectors and that they use a less formal style than the British students. A comparison with native Swedish usage allows them to rule out the L1-induced transfer as an explanation for this underuse as well as for Swedish learners' stronger preference for sentence-initial position of adverbial connectors.

The problem with studies based on essays written by learners of a single mother-tongue background is that it is not possible to make a distinction between features that are specific to this L1 and those that are shared by a wide range of EFL learner populations (cf. section 3.3.3). It is not clear whether semantic incompatibility of adjective-intensifier collocations is a feature specific to German learners or whether it is shared by other learner populations. Comparisons are however possible with other studies that investigate the same linguistic features in other L1 corpora. Altenberg and Tapper compare their findings with data from Granger and Tyson's (1996) study of French learners' use of connectors and show that the same items are over- and underused by the two learner populations. They conclude that lack of register awareness is a feature that Swedish learners share to some extent with advanced French learners as both groups seem to be unaware of the functional and stylistic restrictions of adverbial connectors.

Table 1.9 shows that several studies have recently investigated the use of adverbial connectors and discourse features by Cantonese, Taiwanese, Italian, Japanese, French, Hungarian and Swedish learners (e.g. Bolton et al 2002, Chen 2006, Damascelli 2004, Milton and Tsang 1993, Narita and Sugiura 2006). Most of them support Granger and Tyson's and Altenberg and Tapper's findings that EFL learners experience difficulty with adverbial connectors. EFL learners have been found (1) to use adverbial connectors when they are unnecessary; (2) to favor a limited set of what Tankó (2004) calls 'pet' adverbial connectors; (3) to show a marked preference for sentence-initial position of adverbial connectors and (4) to misuse certain adverbial connectors most probably because they are not fully aware of their semantic and stylistic properties. A number of these studies suggest that "many patterns that are felt to be deviant do not seem to be L1-motivated at all" (Lorenz 1999b:56) and attribute EFL learners' patterns of overuse, underuse and misuse to the overtly simplistic approach to teaching connectors by means of semantically broadly distinguished lists of supposedly interchangeable connectors (see also Crewe 1990).

High frequency verbs and their collocational patterning have also been examined in several studies which focused on EFL learners from different mother-tongue backgrounds (e.g. Howarth 1996; Kaszubski 2000; Nesselhauf 2005; Granger et al 2006). Nesselhauf (2003a/2005), for example, investigates verb + object noun combinations in the German sub-part of the ICLE corpus. She shows that the most frequent type of errors in those combinations involve wrong choice of verb (e.g. **carry out races*). Among the nouns that are most often used with deviant verbs are the signalling nouns *action, aim, attitude, problem, question, statement, step* and *conclusion*. Other types of errors were also found to occur quite frequently:

- the wrong choice of noun, e.g. **close lacks* for *close gaps*;
- the production of a completely wrong combination, e.g. **hold children within bounds*;
- prepositional errors, e.g. **raise the question about*; and
- determiner errors, e.g. **get the permission.*

Nesselhauf also shows that verbs are not only frequently misused in collocations but they also appear to be a major source of error in free combinations, e.g. **plead for emancipation* [*fight for emancipation*] (cf. Nesselhauf 2005:204).

As the highest rate of errors occurs in combinations with a medium degree of restriction, Nesselhauf suggests that "whereas learners are mostly aware of the restriction in combinations where the verb only takes a few nouns, they are less aware of restrictions in combinations where the verb takes a wider range of nouns (such as *exert, perform,* or *reach*)" (2003a: 233). Nesselhauf (2003a) also stresses the influence of the mother tongue in V + N combinations, which accounts for 56% of the erroneous collocations. She investigates L1 influence on correct versus incorrect combinations and finds that non-congruent combinations are consistently – "that is independently of the degree of restriction of the combination – far more difficult for the learner than the congruent[18] ones" (2003a: 236). Nesselhauf (2005) however re-examines her previous findings and writes that "it is by no means the case that non-congruent collocations always pose problems for advanced learners or that congruent collocations never pose any problems" (2005: 223). In some cases, she found that "there was no exploitation of L1 influence, even though it would have been useful" (2005: 240).

Granger et al (2006) analyse the use of the high-frequency verb *make* and compare its **phraseological rate**, i.e. its proportion of phraseological uses, in five learner corpora, all subparts of the ICLE. The learner corpora consist of essay writing by higher intermediate to advanced EFL learners of French, Spanish, Italian, German and Dutch mother-tongue backgrounds. Our data show that the Spanish and Italian learners, and to a lesser extent, the French learners, not only display a lower phraseological rate of the verb *make*, but also display many more erroneous collocations than the German and Dutch learners. Here are a few examples:

1.4. *The political class* ***makes *a large use of*** **[makes good use of]** *words. They are very good in speaking and convincing people.* [ICLE-IT][19]

1.5. ... *we are showing that women are* ****making way*** **[making their way]** *in this men's society.* [ICLE-SP]

1.6. ... *the students make an appeal* ****towards*** **[to]** *the Ministry of Education.* [ICLE-SP]

These findings suggest that there may be a correlation between higher proficiency and a greater use of collocations as the Dutch and German sub-corpora represent higher degrees of proficiency than the French, Spanish and Italian corpora[20] (see also Kaszubski 2000 for a discussion of phraseological competence and language proficiency). The above examples also show that advanced learners are aware of a large number of phraseological uses of *make* but their knowledge is incomplete. They know that variations of *make use of* may include the adjectives *good, full* or *heavy* but seem unaware that *large* is not a possibility. Similarly, *make*

[18] "word-for-word equivalence of a collocation in the learners' L1 and the L2" (Nesselhauf 2005: 236).

[19] Note that, in this thesis, learners' sentences are reprinted as they appear in ICLE: no corrections have been made to their writing.

[20] A number of texts written by learners from the 11 mother tongue backgrounds found in the first version of the *International Corpus of Learner English* have recently been rated according to the descriptors for writing found in the *Common European Framework of Reference for Languages*. Results show that learner essays rate from B2 to C2, with a majority of C1 essays, and that the proportion of B2, C1 and C2 texts differs between the 11 mother tongue backgrounds (cf. section 4.1.1).

way requires the presence of a possessive determiner and *make an appeal* requires the preposition *to*, not *towards*.

In her study of V + N combinations in German learner writing, Nesselhauf (2005) also points to the fact that "the unavailability of pragmatic chunks for the learners also appears to be the underlying reason for a number of deviant collocations which are used to structure the body of the essay, (to introduce examples, for instance)" (2005: 141), e.g. *Only have a look at, If you have a look at, Let us have a look at, A first argument I want to name for this*. De Cock (2003) conducted a careful analysis of recurrent word combinations in French learner speech and writing. She points to learner quantitative and qualitative **stylistic deficiency** in writing as learners tend to overuse recurrent types, most of which contribute to the speech-like character of their writing. Learners are also typically unaware of "the more common, less salient and frequently used L2 multi-word building blocks" (65). De Cock shows that French learners (1) misuse English sequences, e.g. *on the contrary*; (2) underuse multi-word units which have no literal L1 counterpart, e.g. *sort of;* and (3) use idiosyncratic combinations, e.g. *according to me*. Finally, learners are shown to display less variety in the way they organize their written discourse and to underuse multi-word units that have been shown to be typical of formal academic writing.

Table 1.9: A selected list of learner corpus-based studies

Authors	EFL/ESL	EGAP/ESAP	L1	Linguistic features
Aarts and Granger (1998)	EFL	EGAP	Dutch, Finnish, French	Tag sequences
Ädel (2006)	EFL	EGAP	Swedish	Metadiscourse
Agerström (2000)	EFL	EGAP	Swedish	Hedges
Aijmer (2001)	EFL	EGAP	Swedish	'I think'
Aijmer (2002)	EFL	EGAP	Swedish, French and German	Modality
Altenberg and Granger (2001)	EFL	EGAP	French and Swedish	'make'
Altenberg and Tapper (1998)	EFL	EGAP	Swedish	Adverbial connectors
Biber and Reppen (1998)	EFL	EGAP	French, Spanish, Chinese, Japanese	Complement clauses
Bolton et al. (2002)	EFL	EGAP	Cantonese	Adverbial connectors
Bostrom Aronsson (2005)	EFL	EGAP	Swedish	Thematic features and it-clefts
Caldwell (2002)	EFL	ESAP	Xhosa (South Africa)	Lexical vagueness
Chen (2006)	EFL	EGAP	Taiwanese	Adverbial connectors
Chi et al (1994)	EFL	EGAP	Cantonese	Collocational problems
Cobb (2003)		EGAP	French (Quebec)	Vocabulary frequencies, prefabs and writer/reader invisibility
Damascelli (2004)	EFL	EGAP	Italian	Adverbial connectors
De Cock (2003)	EFL	EGAP	French	Prefabs
Flowerdew (1998)	EFL	ESAP	Cantonese	Cause and effect
Flowerdew (2003/2004)	EFL	ESAP	Cantonese	Problem/solution
Flowerdew (2006)	EFL	EGAP	Cantonese	Signalling nouns
Gilquin (2000/2001)	EFL	EGAP	French	Causative constructions
Granger (1997a)	EFL	EGAP	Swedish, Finnish and French	Passive constructions
Granger (1997b)	EFL	EGAP	French, Swedish and Dutch	Participle clauses
Granger (1998)	EFL	EGAP	French	Prefabricated patterns and collocations
Granger et al (1994)	EFL	EGAP	French	Lexical errors
Granger and Tyson (1996)	EFL	EGAP	French, German	Adverbial connectors
Granger and Rayson (1998)	EFL	EGAP	French	Vocabulary
Granger et al (2006)	EFL	EGAP	French, Spanish, Italian, German, Dutch	'make'
Hägglund (2001)	EFL	EGAP	Swedish	Phrasal verbs

Hasselgård (to appear)	EFL	EGAP	Norwegian	Stance and thematic choice
Hewings and Hewings (2002)	ESL	ESAP	-	Clauses with an anticipatory 'it' and extraposed subject
Hinkel (2002)	ESL	EGAP	Chinese, Japanese, Korean, Indonesian, Vietnamese, Arabic	Linguistic and rhetorical features
Hinkel (2003a)	ESL	EGAP	Chinese, Japanese, Korean, Indonesian	Adverbial markers and tone
Hinkel (2003b)	ESL	EGAP	Chinese, Japanese, Korean, Indonesian, Arabic	Syntactic and lexical features
Howarth (1996/1998)	EFL	EGAP	German, Greek, Cantonese, Taiwanese, Japanese, etc.	Verbs
Hyland (2002c)	EFL	ESAP	Cantonese	Imperatives
Hyland (2003)	EFL	ESAP	Cantonese	Pronouns, directives, direct questions
Hyland and Milton (1997)	EFL	EGAP	Cantonese	Expressions of possibility and certainty
Kaszubski (2000)	EFL	EGAP	Polish	High-frequency verbs
Leńko-Szymańska (2002)	EFL	EGAP	Polish	Size of active vocabulary
Liu and Shaw (2001)	EFL	EGAP	Chinese	'make'
Lorenz (1998/1999a)	EFL	EGAP	German	Causal links
Lorenz (1999b)	EFL	EGAP	German	Adjective intensification
McEnery and Kifle (2002)	-	EGAP	?? (Eritrean students)	Epistemic modality
Meunier (2000)	EFL	EGAP	French	Noun phrase complexity
Milton (1998)	EFL	EGAP	Cantonese	Quantitative lexico-grammatical differences
Milton (1999)	EFL	EGAP	Cantonese	Prefabs
Milton and Tsang (1993)	EFL	EGAP	Cantonese	Adverbial connectors
Mulak (2000)	EFL	EGAP	Swedish	Because clauses
Narita and Sugiura (2006)	EFL	EGAP	Japanese	Adverbial connectors
Neff (to appear)	EFL	EGAP	Spanish	Interpersonal discourse phrases
Neff et al (2003)	EFL	EGAP	Spanish, Dutch, Italian, French and German	Modal and reporting verbs
Neff et al (2004)	EFL	EGAP	Spanish	Stance
Neff et al (2007)	EFL	EGAP	Spanish	Errors
Nesselhauf (2003b)	EFL	EGAP	German and French	High-frequency verbs and transfer
Nesselhauf (2004b)	EFL	EGAP	German	Support verb constructions

Nesselhauf (2003a/2005)	EFL	EGAP	German	High-frequency verbs
Petch-Tyson (1998)	EFL	EGAP	French, Dutch, Swedish and Finnish	Reader/writer invisibility
Petch-Tyson (1999)	EFL	EGAP	French, Dutch, Swedish and Finnish	Demonstratives
Ringbom (1998)	EFL	EGAP	French, Spanish, Finnish, Finland-Swedish, Dutch, German	Vocabulary frequencies; high-frequency verbs
Ringbom (1999)	EFL	EGAP	French, Spanish, Finnish, Finland-Swedish, Swedish, Dutch and German	High-frequency verbs
Shaw (2004)	EFL	EGAP	Swedish	Appropriacy, accuracy and coherence
Siepmann (2005)	EFL	EGAP	German	Second-level discourse markers
Tankó (2004)	EFL	EGAP	Hungarian	Adverbial connectors
Tapper (2005)	EFL	EGAP	Swedish	Adverbial connectors
Tono (2004)	EFL	EGAP	Japanese	Verb subcategorization patterns
Virtanen (1998)	EFL	EGAP	Finnish, Finland-Swedish, Swedish, Belgium-French, Dutch, German, Spanish	Direct questions
Wiberg (2000)	EFL	EGAP	Swedish	Involvement
Wiktorsson (2001)	EFL	EGAP	Swedish	Prefabs and register differences
Wiktorsson (2003)	EFL	EGAP	Swedish	Idiomatic expressions

Learner anomalous use of modal verbs, modal adjuncts, personal and impersonal metadiscourse, stance markers, etc. has been found to be indicative of **pragmatic inappropriacy**. Aijmer's (2001) investigation of patterns with *I think* in argumentative essays written by advanced Swedish EFL students shows that learners often use the expression to organize their texts and make their claims more persuasive. Similarly, Ädel (2006) reports that personal metadiscourse, i.e. metadiscourse items that refer explicitly to the writer and/or reader, serves a wider range of rhetorical functions in Swedish learner writing than in British and American student writing, e.g. exemplifying, arguing, anticipating the reader's reaction, concluding. Hewings and Hewings (2002) show that MBA student dissertations written by non-native speakers of English are characterized by overstatement (see also Lorenz 1998): they repeatedly use phrases such as '*It is true that* ...' and '*It is a fact that* ...'; the modals *must* and *have to* as in '*It must be emphasized* ...'; and adjectives such as *necessary*, *crucial* and *essential* (see Neff to appear for similar findings about Spanish learners). Hyland and Milton's (1997) investigation of expressions of doubt and certainty shows that Cantonese learners use a more restricted range of epistemic modifiers and have considerable difficulty conveying the appropriate degrees of qualification and confidence. EFL learners have also been found to use more questions and directives (e.g. the modals *should* and *must*, imperatives) to organize discourse and to follow conversational patterns in employing more yes / no questions (cf. Virtanen 1998; Ädel 2006; Hyland 2003). More generally, EFL learners display more features of writer/reader visibility than native speakers of English (cf. Petch-Tyson 1998; Neff et al 2004).

A few studies have investigated the lexical means used by EFL learners to serve specific **rhetorical or discourse functions**. Flowerdew (1998) analyses cause and effect markers in native expert and Cantonese EFL learner writing and shows that *because* and *therefore* are not only significantly overused by Cantonese non-native learners but their grammatical patterning also differs (e.g. the non-native use of double connectors as in '*because ... so that*'). By contrast, Cantonese learners underuse a whole set of lexical means and, more especially multi-word units, to mark causativity, e.g. *as a result of, responsible for* and causative verbs. Flowerdew also compares the use of the nouns *reason, cause, effect* and *result* and shows that non-native learners tend to use these signalling nouns in a more restricted range of grammatical patterning. For example, they do not use the noun *reason* in an adverbial group introduced by *for* and make use of *reason* and *cause,* as well as *effect* and *result,* with little discrimination between the nouns. In another study, Flowerdew (2003) investigates the lexical means used to introduce a problem-solution pattern. She focuses her attention on the signalling noun *problem* and examines its collocational preferences and grammatical structures. She finds that 'problem' often collocates with a causative verb but that the range of causative verbs is far more restricted in Cantonese learner production. Causative verbs appear to be a major area of difficulty for non-native learners in the problem/solution pattern. Flowerdew also highlights other deficiencies in the verbal domain, especially **restricted range of vocabulary** and **semantic inappropriacy**. In her studies, Flowerdew has often insisted on a restricted use of EAP-specific vocabulary and its phraseology in Cantonese learner writing compared with the larger repertoire found in professional academic writing.

1.4.3. Corpus-based studies of vocabulary in NS student writing

NS student writing has often been studied within the framework of learner corpus research. Learner corpus researchers have used corpora of NS student writing as reference corpora to analyse EFL learner writing, e.g. Lorenz (1998), Granger (1997), Ringbom (1998), Petch-Tyson (1999), Tapper (2005), Virtanen (1998)[21]. Cutting (2000) compares written errors of German, Dutch, French, Italian and Spanish exchange students with those of English native university students. He shows that international students have many more problems with vocabulary and use more colloquial vocabulary than their native counterparts. The native students do not make so many errors in terms of formality. When errors occurred, however, "they tended to cluster, as if once the student had started writing informally, it was difficult to break out of it" (Cutting 2000:104). In their study of modal and reporting verbs in the expression of writer stance, Neff et al (2003:224) report that the English native speaker students "show a rather balanced use of the reporting verbs *say, state, show* and *argue*, whereas in some of the non-native groups there is a heavier reliance on one reporting verb (*say*) and a much reduced repertoire of other verbs which might allow the EFL writers to report authors' propositions with greater or lesser grades of certainty or doubt."

By contrast, Hyland and Milton's (1997) investigation of expressions of doubt and certainty shows that both Cantonese learners and more novice native students, i.e. British school leavers, experience considerable difficulty in conveying the appropriate degrees of qualification and confidence. They state that NS and NNS students respond to these difficulties "by mixing informal spoken and formal written forms and transfer conversational uses of academic genres" (Hyland and Milton 1997:192). They report that the verbs *think* and *know* in EFL learner texts and *believe*, *seem* and *think* in the native speaker corpus account for

[21] See section 4.1 for a discussion of reference corpora in learner corpus research.

almost two thirds of all forms. They conclude that "the limited use of epistemic verbs and a preference for predominantly speech forms indicates the novice writers' imperfect grasp of appropriate academic register. Apparently possessing only a rudimentary understanding of formal academic expectations, neither student group is able to employ "expert" forms in making claims. To avoid violating academic expectations, NNSs appear to seek a solution by employing more modal verbs and NSs by overusing epistemic adverbs" (ibid: 192).

Other learner corpus-based studies have compared learner data with both native student and professional writing to distinguish between features that characterize EFL learner writing and those that are more typical of novice writing, e.g. Meunier (2000), Aijmer (2002), Hasselgård (to appear). Neff et al (2004) compare the expression of writer stance in various corpora of argumentative texts written by four EFL groups, American university students and native professional writers. They show that "all of the student writers (native and non-native) have the novice-writer characteristic of excessive visibility" (Neff et al 2004:152). In both native student and EFL learner texts, *I think* was often used to accompany metadiscourse markers. By contrast, *I feel* was used excessively, in comparison to professional texts, only in American student writing.

As Nesi et al. (2004:440) have pointed out, "far more academic writing is produced for assessment purposes than for publication purposes, but because of the lack of a suitable corpus, research into the generic features of published academic writing vastly outweighs research into the generic features of assessed student writing." In order to investigate student writing in the disciplines, two corpora of ESAP student writing are currently being compiled. The *British Academic Written English* (BAWE) corpus and the *Michigan Corpus of Upper-Level Student Papers* (MICUSP) both consist of highly graded papers written by native and non-native students in several faculties. These two corpora are still under development and very few studies have used them so far. Ädel and Garretson (2006) analyse citation practices across the disciplines represented by student texts in the MICUSP and compare results with those described in Hyland's (1999) study of attribution across disciplines. They examine the use of the most common reporting verbs for each of the comparable disciplines and show that there is very little overlap between the lexical preferences of the students and the expert writers. They also report that "the frequent use of verbatim quotes could be the aspect that most marks the MICUSP writers as students; they appear, proportionally, to give more weight to the words of their authoritative sources" (Ädel and Garretson 2007:280).

Cortes (2002) examines four-word lexical bundles (see section 1.4.1) in a 360,704 word corpus of freshman composition consisting of 54 final portfolios, which included six different papers: two descriptions, two rhetorical analyses, a research proposal and a research paper. She shows that "contrary to any intuition which may consider freshman student's writing following a conversational style (loaded with contractions and expressions connected to narration), the highest number of bundles identified in the freshman composition corpus is nominal or phrasal rather than clausal, following the pattern in published academic prose" (Cortes 2002:137) (cf. section 1.4.1 and Appendix 1.1). The bundles used, however, often served as temporal or location markers and were thus not bundles exclusively used in academic prose. Table 1.10 also shows that patterns most widely used in academic prose such as 'passive verb + PP fragment' and 'verb (+ that-clause fragment) were not found in student writing. Cortes (2004) examines four-word lexical bundles in published articles from history and biology journals and in student essays in those disciplines. She shows that many lexical bundles used by expert writers are rarely or never used by students, e.g. referential bundles (*the beginning of the, the shape of the, with the number of*), and text organizers (*the extent to which, the degree to which, with respect to the*). On the other hand, they tend to repeat bundles several times in a single paper, a finding which the author compares with the case of repetitiveness in the use of fixed expressions produced by non-native speakers of English in written essays as reported by Granger (1998). Students sometimes also use lexical bundles to convey functions different from those identified in expert writing. For example, they use the bundle *at the same time* to express addition instead of simultaneity.

Table 1.10: Distribution of four-word lexical bundles by structural patterns (Cortes 2002:138)

Patterns most widely used in conversation	**Conv.**	**Acad. prose**	**example**	**Freshman writing**	**Example**
Personal pronoun + lexical verb phrase (+ complement clause)	44%		*I don't know what*	2%	*I will use this*
Pronoun/NP (+ auxiliary) + copula be (+)	8%	2%	*it was in the*	3%	*it is as if*
(auxiliary +) active verb (+)	13%		*have a look at*		
Yes-no and wh-question fragment	12%		*can I have a*		
(verb +) wh-clause fragment	4%		*know what I mean*		
Patterns most widely used in academic prose					
Noun phrase with post modifier fragment	4%	30%	*the nature of the*	35%	*the side of the*
Preposition + noun phrase fragment	3%	33%	*as a result of*	32%	*as a result of*
Anticipatory it + VP/-adjective phrase (+ complement clause)		9%	*it is possible to*	5%	*it is difficult to*
Passive verb + PP fragment		6%	*is based on the*	-	
(verb +) that-clause fragment	1%	5%	*should be noted that*	-	
Patterns in both registers					
(verb/adjective +) to-clause fragment	5%	9%	*are likely to be*	2%	*to appeal to the*
Other expressions	6%	6%		21%	
TOTAL	**100%**	**100%**		**100%**	

1.5. Conclusion

> "Vocabulary is made up of units of learning effort. In setting out any vocabulary we have to enumerate all those things which the pupil has to learn. In a reading vocabulary, we may count the word 'home' as one unit, because a child who knows the meaning of 'home' will find no difficulty in understanding 'at home', 'not at home', 'feel at home', 'come home', 'bring home'. But in speaking vocabulary, each of these things has to be learnt, and must therefore be listed and counted." (West 1937:436)

This chapter first delimited the scope of English for Academic Purposes and answered Hyland's (2002a) criticism about its lack of specificity by showing that academic prose is characterised by a common core of distinctive linguistic features and insisting on EFL learners' need for EAP courses. It then focused on vocabulary needs in EAP reading and writing and highlighted the importance of a rigorous and empirically based **selection** of vocabulary in academic settings. While a distinction between a general service word list and an academic word list is valuable for receptive purposes, it is not totally adequate for productive purposes. Numerous 'general service' words have important discourse functions in EAP and their productive use is not fully mastered by L2 learners, even at an advanced level. It was thus argued that EFL learners, as well as teachers and textbook developers, would greatly benefit from the elaboration of a **productive counterpart** to the *Academic Word List* as learners' needs and difficulties are clearly not the same in production as in reception. We propose to take up the challenge in chapter 5 of this thesis.

A review of selected corpus-based studies of vocabulary in **academic writing** has pointed to the paramount importance of academic words – nouns, verbs, adjectives, adverbs and even function words - and their phraseological patterns to serve specific referential, rhetorical and organizational purposes that are characteristic of academic discourse, and more generally, of scientific knowledge, irrespective of differences across genres and disciplines. They represent a common core of academic phraseological patterns or collocational frameworks. One important finding emerging from these studies is that the phraseology of academic discourse is highly **conventionalized**. Many studies acknowledge the existence of an EAP-specific phraseology characterized by word combinations that are essentially semantically and syntactically compositional, e.g. *as a result of, in the presence of, the aim of this study, the extent to which, it has been suggested, it is likely that*. These word combinations can be described as "lexical extensions" (Curado Fuentes 2001:115) of academic words and form an essential part of the procedural vocabulary of academic prose.

EAP vocabulary was typically found to cause major difficulties to EFL learners, especially retrospective and advance labels, adverbial connectors, modal verbs and prefabricated language. Corpus-based studies of **EFL learner writing** in academic settings have shown that some of the linguistic features that characterize learner language are shared by learners from a wide range of mother-tongue backgrounds whilst others are unique to one particular learner population. Most shared features can be assumed to be **developmental** in that EFL learners are all involved in a learning process of writing in a foreign language but others may also be partly explained by the fact that they are **novice writers** in their mother-tongue as well. The unique features, on the other hand, may be due to **transfer** from the learners' mother-tongue. Several studies point to the potential influence of the mother-tongue, in particular as regards collocations and pragmatic conventions. Some shared or unique features have also been found to be **teaching-induced**.

These studies also tend to confirm Ringbom's view that non-native features at a rather advanced level of proficiency are "less due to errors than to an insufficient and imprecise, though not necessarily erroneous, use of the resources available in English" (Ringbom 1998: 51). One notable exception is collocational errors such as those reported in Nesselhauf (2005). A high number of studies have underlined learners' difficulty with collocations and other phraseological units. Most of these, however, were based on a limited amount of data often representing the writing of learners from one mother-tongue background. Moreover, these studies have often focused on high-frequency verbs. Very few studies have analysed academic words and their phraseological patterns in learner writing except for Flowerdew (1998/2003).

Novice native writers also seem to experience difficulty with academic language, and more particularly with its highly conventionalized phraseology. Howarth postulates the existence of a continuum of phraseological competence that would "encompass mature NS writers at one extreme and weak NNS writers at the other, with NS and NNS students of varying levels of proficiency in between, and some overlap between native and non-native writers (Howarth 1999: 151). A promising area of research thus lies in the investigation of patterns of difficulty shared by English L1 students and EFL learners to separate linguistic features that are characteristic of novice writing from those features that have commonly been attributed to EFL writing.

2. Lifting a corner of the veil on the phraseological spectrum

> "There is no independent linguistic discipline phraseology similar to semantics, syntax, or morphology, each of which studies a particular component of the language. Phraseology is rather a particular field of interest that concentrates on a particular type of linguistic signs and has to deal with everything, starting with semantics and ending with phonetics (...). That is why phraseology is so difficult, but so appealing!" (Mel'čuk 1995:227)

2.1. Introduction

The emergence of phraseology is generally situated between 1940 and 1960 in the former Soviet Union, with the work of Vinogradov and Amosova (cf. Cowie 1998a). In her *Manual de Fraseología Española*, Corpas Pastor (1996: 11) also stresses Julio Casares's (1877-1964) contribution to Spanish phraseology. González Rey (2002) cites Charles Bally for being the father of French phraseology. However, it was only in the 1980s and 1990s that phraseology became the centre of renewed interest to the point of becoming "a significant focus of research, especially perhaps in Europe" (Cowie 1994: 3168), among specialists in lexicology and lexicography (Hausmann 1989; Cowie et al. 1983), lexical semantics (Cruse 1986), vocabulary in language teaching (Alexander 1984; Carter 1987; Carter and McCarthy 1988), psychology (Gibbs 1990), psycholinguistics and language acquisition (Peters 1983; Wray 2002), vocabulary in second language acquisition (Nation 2001), languages for academic and specific purposes (Clas 1994; Cowie 1997; Gledhill 2000) and natural language processing (Church et al. 1991; Smadja 1993, Sag et al. 2002) (see Gónzalez Rey (2002:19-32) for a detailed account of the history of phraseology in Europe).

Phraseology stands at the intersection between semantics, morphology, syntax and discourse (cf. Granger and Paquot to appear). This intermediary position has made Montoro del Arco (2006) argue that phraseology can only enjoy some autonomy by defining its status in relation to the many linguistic disciplines that can provide information on its object of study. Other phraseologists advocate a multidisciplinary approach to phraseology which would not only include its linguistic analysis (diachronic studies, stylistics, terminology, teaching, etc.) but also its psycholinguistic and sociolinguistic study (cf. González Rey 2002).

Despite (or because of) its blossoming within the last decades, there is no consensus on the definition of phraseology and its object of study. Phraseologists differ in their delimitation of the **scope** of phraseology and the types of **word combinations** that they include in the field (cf. Granger 2005). They also disagree about the **terminology** and **typology** of word combinations. Even when they use the same terminology, their definitions are often dissimilar. These observations make it all the more necessary to delimit and define the phraseological spectrum. This chapter successively addresses the questions of the categorization of phrasemes[22] and their defining criteria before presenting the typology and definitions adopted in this thesis. It then zooms in on collocations as these phrasemes enjoy a special status within the phraseological spectrum. Finally, it focuses on the relationship between discourse and phraseology as phrasemes with specific rhetorical functions have been reported to be typical of academic discourse (see section 1.4.1)

2.2. Categorizing the phraseological spectrum

> "Like all other scientists, linguists wish they were physicists. They dream of performing classic feats like dropping grapefruits off the Leaning Tower of Pisa, of stunning the world with pithy truths like "F=ma", and in general of having language behave in an orderly way so that they could discover the Universal Laws behind it all. Linguists have a problem because language just ain't like that. Physical laws are very basic, general-purpose constituents of the universe, so the Creator was forced to keep them elegant and potently simple. Language, by contrast, was recently invented by Man for the sole purpose of giving his Fellow Man the low-down; for this reason language is inextricably bound to humans, human communication, and the circumstances of human communication." (Becker 1975:60)

As already stressed by Cowie, phraseology is "a field bedevilled by the proliferation of terms and by the conflicting uses of the same term" (1998b: 210). The term 'phraseology' itself is not devoid of ambiguity (cf. Gónzalez-Rey 2002:20-21). A partial explanation for this ambiguity can be found in the long-running debate over the status of phraseology when compared with disciplines such as semantics, morphology, syntax and pragmatics. Another issue that provides an explanation to this ambiguity is the scope of the object of study of phraseology: Should idiomaticity be a defining criterion for phrasemes? Should sentence-level units such as proverbs and maxims be considered as phrasemes? Should compounds be part of the phraseological spectrum? What is the status of collocations? What are grammatical collocations?

As these questions constitute major points of disagreement, it is not surprising that linguists have not been able to reach a consensus on a superordinate term to refer to the object

[22] Note that in section 2.2, we will not be systematic in our use of a superordinate term to refer to the object of study of phraseology and use the term proposed by the authors reviewed.

of study of phraseology. Numerous terms have been proposed, among others, word-combinations (Zgusta 1971; Cowie 1994; Howarth 1996), multi-word units (Cowie 1988; De Cock 2003), fixed expressions (Alexander 1984; Gramley and Pätzold 1992; Moon 1998a)[23], phrasemes (Mel'čuk 1995, 1998)[24], multi-word items (Moon 1998b), set phrases (Mel'čuk 1995, 1998), phraseological units (Gläser 1998)[25], multi-word lexemes (Tschichold 2000), multi-word expressions (Sag et al. 2002), formulaic sequences (Wray 2002), and multi-lexemic expressions (Guenthner and Blanco 2004). The picture gets even worse when specific types of phrasemes are considered (see Wray (2002:9) for a list of terms used to refer to different types of phrasemes).

Similarly, models of phrasemes abound in the literature and include, among others, those of Zuluaga (1980), Gläser (1986; 1998), Cowie (1988; 1994), Gramley and Pätzold (1992), Nattinger and DeCarrico (1992), Lewis (1993; 2002), Corpas Pastor (1996), Burger (1998), Mel'čuk (1995; 1998), Ruiz Gurillo (1997), Moon (1998a), Wray (2002) and Sag et al. (2002). Some typologies are based on a narrower definition of phraseology than others (e.g. Ruiz Gurillo 1997) or focus on a specific sub-type of phraseological units, e.g. Nattinger and DeCarrico (1992) and Gross (1996). Some models are developed to describe the phraseological spectrum for lexicological or lexicographical purposes (cf. Moon 1998a); others are used to teach phraseological units (cf. Nattinger and DeCarrico 1992; Lewis 1993; 2002; Willis 2003) or to account for them in psycholinguistic terms (cf. Wray and Perkins 2000; Wray 2002). Finally, numerous 'ad hoc' descriptions have also been proposed within the field of natural language processing (NLP) (cf. Sag et al. 2002, Tschichold 2002). In this section, we shall only review typologies that offer a wide perspective on the phraseological spectrum.

Most classifications give prominence to one or more of five features of phrasemes:

a) their internal structure (e.g. noun + verb, verb + preposition + noun),
b) their function within the sentence or syntactic function: some phrasemes are equivalent to sentences; others function below sentence-level,
c) their degrees of compositionality or idiomaticity,
d) their function in the language
e) their degrees of syntactic flexibility and collocability.

[23] Fr. 'expression figée' (Gross 1996) and Sp. 'expresión fija' (Zuluaga 1980)
[24] Fr. 'phrasème' (Gréciano 1997), Ge. 'Phraseme' (Gréciano and Rothkegel 1997)
[25] Sp. 'unidad fraseológica' (Zuluaga 1980; Corpas Pastor 1996), Ge. 'Phraseologismus' (Gläser 1986; Burger 1998)

Differences between the typologies largely correspond to differences in the selection of the features used to categorize phrasemes and in the hierarchization of selected features.

This section is not intended as a comprehensive survey of the many typologies of phraseological units proposed in different fields. Instead, it focuses on a restricted set of typologies which are deeply rooted in lexicology and lexicography. It then briefly describes two types of models which come from English Language Teaching research and psycholinguistics respectively.

2.2.1. Typologies and lexicology/lexicography

Although typologies based on the criteria of idiomaticity and syntactic flexibility have been proposed (e.g. Zuluaga 1980)[26] in lexicology and lexicography, most typologies are primarily based on the criterion of **syntactic function**. Following the classical Russian phraseological theory, the models proposed by Gläser (1998)[27], Cowie (1988; 1994), Gramley and Pätzold (1992), Mel'čuk (1995; 1998) and Corpas Pastor (1996) share their primary categorizing criterion, viz. the distinction between phrasemes that function syntactically at or below the level of the simple sentence and those which function pragmatically as autonomous utterances (cf. Cowie 1998a:4).

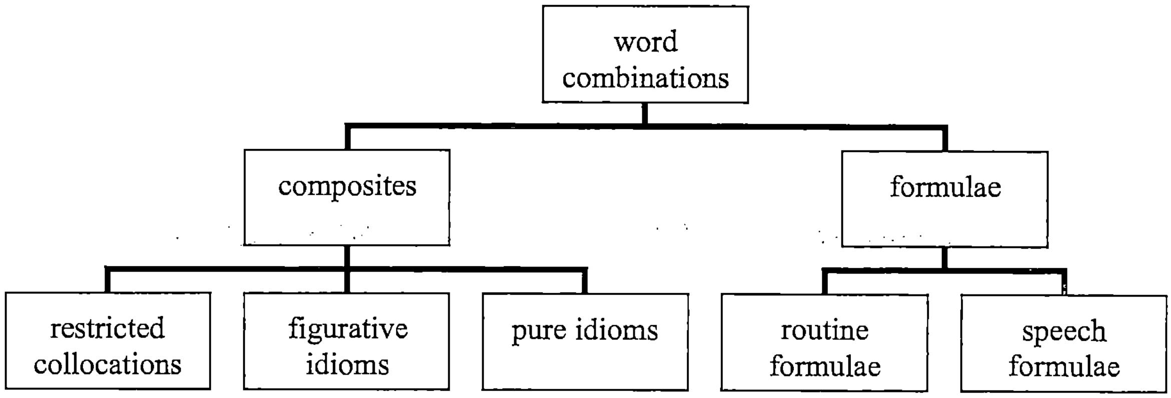
One of the most influential typologies in English lexicology and lexicography is that of Cowie (1988; 1994), which distinguishes between two main types of **word combinations** or **multi-word units** "according to the kinds of meaning which their members convey and to the structural level at which they operate" (Cowie 1988: 132). Thus, syntactic function is directly linked to semantico-pragmatic features. **Composites** function as constituents of sentences, contribute to their referential or propositional meaning and, as such, have become **semantically** specialized. As shown in Figure 2.1, they include restricted collocations (e.g. *jog one's memory, a chequered career, entertain an idea*), figurative idioms (e.g. *a close shave, do a U-turn, catch fire*) and pure idioms (e.g. *kick the bucket*) (cf. Cowie et al. 1983: xii-xiii). **Formulae** are "largely a reflection of the way they function in discourse" (Cowie 1988: 132) and, as such, have become **pragmatically** specialized. Cowie (2001: 11) further subdivides formulae into routine formulae and speech formulae. While the former are used to perform speech-act functions such as greetings, compliments, invitations, etc. (e.g. *Good*

[26] A major drawback of these types of models is that categories consist of a very heterogeneous set of phrasemes. For example, Zuluaga's (1980) category of idiomatic fixed expressions include compounds (prensa amarilla, En. 'sensationalist press'; lit. 'yellow press'), similes (terco como una mula, En. 'stubborn as a mule'), idioms (tomar el pelo, En. 'take the mickey out of someone'), etc.

[27] Gläser's model was first proposed in her book *Phraseologie der englischen Sprache* in 1986.

morning, see you soon), the latter are used to organize the speakers' messages and "employed in organizing turn-taking, indicating a speaker's attitude to other participants, and generally ensuring the smooth conduct of interaction" (Cowie 1988: 133), e.g. *if you please, are you with me?, you know what I mean.*

Figure 2.1: Cowie's (1988; 2001) classification of word combinations



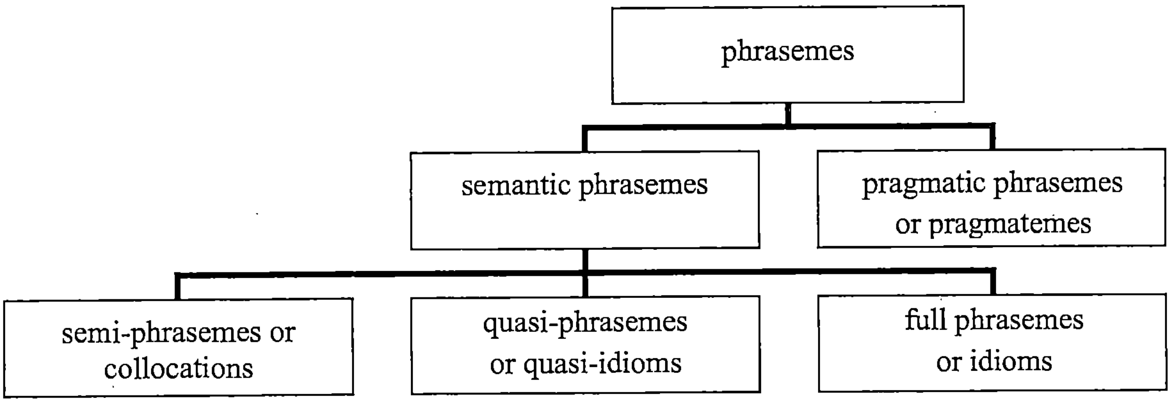
Following the Russian tradition, Cowie situates composites on a continuum from free or open combinations to pure idioms. The notion of **continuum** is central to Cowie's typology and the author insists that there is no clear dividing-line between the different types of composites. **Free combinations** (e.g. *drink one's tea, dismiss an employee*) are not part of the phraseological spectrum as the selection restrictions on the choice of object nouns can be stated in terms of features denoting general properties (cf. Cruse 1986):

> In *dismiss an employee*, for example, the verb can be recombined with nouns having the features 'human', 'employed', and 'subordinate', a specification which will account for the acceptability of *dismiss a secretary* or *dismiss a cleaner*, and for the oddness of **dismiss one's boss* and **dismiss a guarddog.* (Cowie 1994:3169)

Restricted collocations (e.g. *perform a task,, break one's jouney, wholesome fare*) are characterized by the arbitrary restriction on the collocational range of one of their elements (see section 2.3.5 for more detail on collocability) and the figurative or specialized meaning of the element hence selected. They include verb-noun combinations with a delexical verb such as *have an influence on someone/something*. **Figurative idioms** have a figurative meaning but also preserve a literal interpretation (e.g. *do a U-turn, close ranks, die a natural death*). They resist substitution of their components. The last group includes pure idioms such as *spill the beans, blow the gaff* and *kick the bucket.* **Pure idioms** have a figurative meaning and do not preserve a literal interpretation.

The model of **set phrases** or **phrasemes** proposed by Mel'čuk (1995; 1998) within the meaning-text theory is very similar to that of Cowie except for the fact that Mel'čuk does not lay emphasis on the link between the syntactic criterion and that of semantic vs. pragmatic specialization. Mel'čuk defines a **semantic phraseme** as a set phrase in which "the meaning is chosen freely (it is not bound by the situation), but the expression for this meaning is not chosen freely: its selection is completely or in part bound by this meaning" (Mel'čuk 1995:181). Semantic phrasemes include full phrasemes or idioms, quasi-phrasemes or quasi-idioms and semi-phrasemes or collocations (cf. Figure 2.3). The category of **full phrasemes** comprises all semantic phrasemes whose signified does not include either of the signifieds of their components in a semantically dominant position: the signifieds of *red* and *herring* are not included in the signified of *red herring*. Other examples are *to take somebody to the cleaner's, of course, to spill the beans* and *to throw up*. The signified of **quasi-phrasemes** includes the signifieds of their components plus an added signified. Thus, the signifieds of *bed* and *breakfast* are comprised in *bed and breakfast* together with an additional and unpredictable signified so that the phraseme can either mean 'the providing of a room for a night and breakfast in the morning, for example in a hotel' or 'a private house or small hotel where you can sleep and have breakfast' (LDOCE4). Other examples include *to start a family, to give the breast, bacon and eggs, shopping centre*. **Semi-phrasemes** or **collocations** are semantic phrasemes whose global signified is constructed out of the signified of one of their components A and a signified 'X' such that their other component B expresses 'X' only contingent on A. Thus, the collocation *strong coffee* is constructed out of the signified of *coffee* and a signified X that the adjective *strong* only expresses contingent to the noun *coffee*. This formulation covers four major types of collocations which are described in section 2.4.1.2.

Figure 2.2: Mel'čuk's (1998) typology of phrasemes


phrasemes
semantic phrasemes
pragmatic phrasemes or pragmatemes
semi-phrasemes or collocations
quasi-phrasemes or quasi-idioms
full phrasemes or idioms

The meaning of a **pragmatic phraseme** is bound by the situation in which it occurs. In a pragmatic phraseme, the situation "precludes free choice of possible meanings (and sometimes, for a chosen meaning, of possible expressions): It prescribes what to say and maybe how to say it" (ibid 179-180). In other words, the situation "phraseologically binds" the pragmatic phraseme. Pragmatic phrasemes include all ready-made expressions, even if they are wholly compositional semantically and syntactically: they are "non-compositional pragmatically" (Mel'čuk 1998:29). Examples of pragmatic phrasemes are greetings, conversational formulae, typical phrases used in letters or academic texts (e.g. *emphasis is mine*), signs in restaurants (e.g. *Caesar salad: all you can eat*) or in public buildings (*no talking please*), proverbs, quotations and sayings.

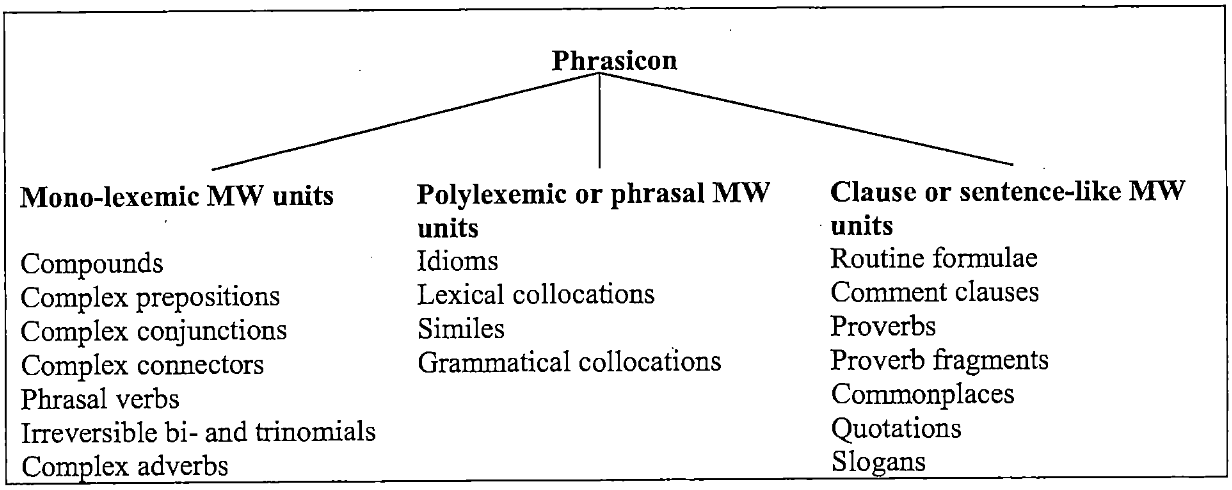
Phraseological models such as the ones proposed by Cowie (1988; 1994) have been criticized **for mixing formal and functional criteria** to make a primary distinction between formulae and composites. De Cock (2003:77) writes that "there does not seem to be a systematic one-to-one correspondence between form and function." The author illustrates this point with the word combinations *sort of, at least, as a result, in other words* and *in fact* and argues that they share the form of composites, i.e. they function structurally below the level of the sentence, and the function of formulae, i.e. they perform pragmatic or discourse-structuring functions. Although the link between form and function is less explicit in this model, Mel'čuk's typology is open to the same types of criticisms as that of Cowie: phrasemes such as *of course, by and large, in short,* and *as well as* are classified as idioms on the basis that their global signified does not include the signifieds of their components but they also perform pragmatic or discourse-structuring functions that are not taken into account in such a typology.

To avoid linking form and function, De Cock (2003) proposes two models of the **phrasicon** (i.e. the whole set of multi-word units): a structural model and a functional model. In her **structural model**, the author distinguishes between three major types of **multi-word units** according to the level at which they operate in the sentence (cf. Figure 2.3):

- **Mono-lexemic multi-word units** are equivalent to single words and thus fill only one grammatical function slot. They include compounds (e.g. *black hole, the oldest profession*), complex prepositions (e.g. *in addition to, such as*), complex conjunctions (*even though*), complex connectors (e.g. *in fact, by the way*), phrasal verbs (*crop up, hand in*), irreversible bi- and trinomials (e.g. *bed and breakfast, part and parcel*) and complex adverbs (e.g. *of course, kind of, at least*). The distinction made between complex adverbs and complex connectors is quite problematic. It creates some overlap between the two categories and makes it difficult to use the typology: it is not quite clear why the author decided to classify *of course* as a complex adverb while categorizing *in fact* as a complex connector.

- **Polylexemic** or **phrasal multi-word units** include idioms (e.g. *to bark up the wrong tree*), lexical collocations (e.g. *rancid butter, sorely miss*), similes (e.g. *as blind as a bat, to eat like a bird*) and grammatical collocations (e.g. *depend on* something). De Cock (2003) explains that, as they can be described as verb/noun/adjective phrases with an open post-modifying slot (e.g. *approve of* ***something/someone***), grammatical collocations cannot be considered to be mono-lexemic.
- **Clausal and sentential multi-word units** include routine formulae (e.g. *how are you, as far as X is concerned*), comment clauses (e.g. *you know, I mean*), proverbs and proverb fragments, commonplaces (e.g. *you never know*), quotations and slogans.

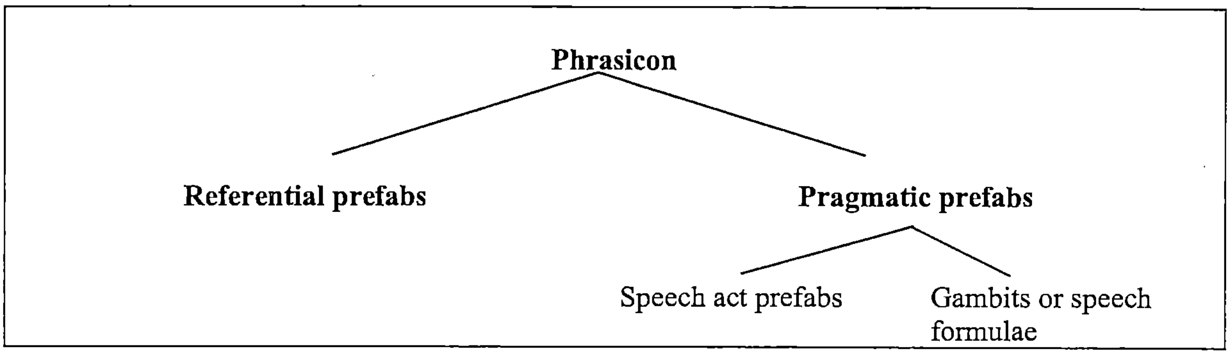
Figure 2.3: De Cock's (2003:83) structural model



In her **functional model**, De Cock (2003) divides the phrasicon into referential and pragmatic prefabs (cf. Figure 2.4). **Referential prefabs** are multi-word units "whose primary function is to refer to concrete things of all kinds and to abstract things such as actions, states, events, processes or qualities in the extra-linguistic world" (De Cock 2003:81). They include idioms, lexical and grammatical collocations, phrasal verbs, compounds and some mono-lexemic fixed expressions. **Pragmatic prefabs** are defined in terms of "what 'signalling' function they are used to perform in communication" (ibid). De Cock (2003) further categorizes pragmatic prefabs as **speech act prefabs**, i.e. Cowie's 'routine formulae' and

gambits, i.e. Cowie's speech formulae. Unlike Cowie, the latter category includes both sentence-like phraseological units used to signal speakers' attitudes towards their utterances and interlocutors (e.g. *I mean, you know*) and phraseological units that function at or below sentence-level to organise speakers' messages (e.g. *first of all, in a nutshell, on the one hand ... on the other hand*).

Figure 2.4: De Cock's (2003:83) functional model



Another attempt at meeting the criticisms levelled at models which use both formal and functional criteria is Burger (1998), who proposed a typology of phrasemes primarily based upon the communicative functions they serve[28]. He distinguished between referential, communicative and textual phraseological units (cf. Figure 2.5).

Referential phraseological units are divided into two sub-categories according to a syntactico-semantic criterion:

- **Nominative phraseological units** are constituents of the sentence and refer to objects, phenomena or facts of life (e.g. *Schwarzes Brett* 'billboard' or *jemanden übers Ohr hauen* 'to rip somebody off'). This category broadly corresponds to Cowie's 'composites' and Gläser's (1998) 'nominations'. Following the Russian tradition and phraseologists such as Cowie and Mel'čuk (see above), nominative phraseological units are sub-divided into **idioms, partial idioms** and **collocations**.

[28] It is noteworthy that, in the same year as Burger proposed his functional model of phraseological units, Butler grouped recurrent sequences (see sections 1.4.1 and 2.5) according to Halliday's metafunctions and Moon's *Fixed Expressions and Idioms in English* included a discussion of the text functions of phrasemes largely based on Halliday's language model as well. All these approaches to the phraseological spectrum share the view of functional grammarians that "language structure can only be satisfactorily explained in terms of the communicative functions which language serves, and the psychological and social conditions of language use" (Butler 1998:14). The reader is referred to Butler (2003) for a discussion of the place of multi-word sequences in recent models of Functional Grammar.

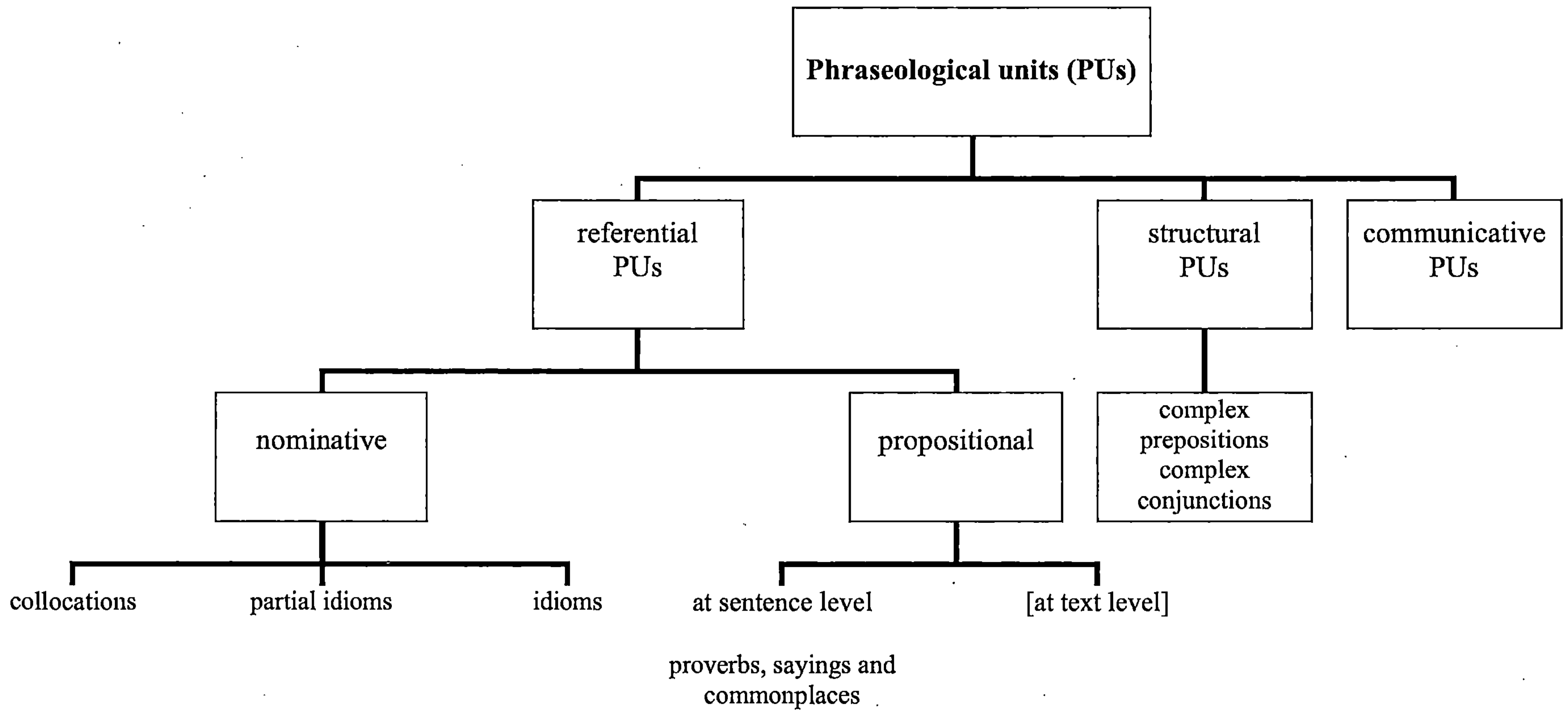
- **Propositional phraseological units** generally function at the level of the sentence but a few propositional PUs function at the level of the text. They refer to a statement or an utterance about these objects or phenomena (*Morgenstund hat Gold im Mund* 'the early bird catches the worm'). They include proverbs and idiomatic sentences, two broad categories that are classified as 'formulae' or 'pragmatic phrasemes' in models such as the ones proposed by Cowie and Mel'čuk that use both the criteria of function in discourse and function in the sentence.

Communicative phraseological units or **routine formulae** fulfil an interactional function: they are typically used as text controllers to initiate, maintain and close a conversation or to signal the attitude of the addressor. Examples are *Guten Morgen* (En. *Good morning*) and *Ich meine* ... (En. *Well, I mean*...).

The categories of referential phraseological units and communicative phraseological units are not original and bear marked similarities to Cowie's composites and formulae or Mel'čuk's semantic and pragmatic phrasemes. Burger's (1998) originality lies in a third category of **structural phraseological units** which includes all word combinations that establish grammatical relations, e.g. *in bezuf auf* (En. *concerning*) and *sowohl ... als auch* (En. *as well* ... *as* ...). They usually correspond to a unit smaller than the phrase and can have the function of a preposition (*an Hand von* 'on the basis of', *im Laufe* 'in the course of', *im Hinblick auf* 'in view of, with regard to') or of a conjunction (*wenn auch* 'even if', *um zu* 'in order to'). Although he is one of the few authors who create a category for this specific set of phraseological units, Burger regards textual phraseological units as the smallest and the least interesting[29] one and does not give more details about it.

[29] "Von den drei Gruppen ist [strukturelle Phraseologismen] die kleinste und am wenigsten interessante." (Burger 1998:37)

Figure 2.5: Burger's (1998) classification of phraseological units



2.2.2. Typologies and English Language Teaching

Very few comprehensive typologies of the phraseological spectrum have been proposed in English Language Teaching (ELT) research. Nattinger and DeCarrico (1992) concentrate on lexical phrases, i.e. a subset of phraseological units which have been assigned pragmatic or socio-interactional functions (cf. section 2.3.4). Attempts at describing the phraseological spectrum have rarely been made in a principled way but have rather made use of different criteria on an ad hoc basis. Lewis (1993) is a case in point. The author distinguishes between three main categories of 'multi-word units', which he defines on the basis of the criteria of **form, function** and **frequency/statistical significance** respectively:

Polywords are usually made up of two or three orthographic words; may belong to any word category and are "frequently found in dictionaries" (Lewis 1993:92). Polywords is the 'messiest' category in Lewis's words and include compounds (e.g. *taxi rank, record player* and *continuous assessment*), phrasal verbs (e.g. *put off, look up* and *look up to*) and grammaticalized prepositional phrases (e.g. *of course, on the other hand, by the way*).

Institutionalised expressions "allow the language user to manage aspects of the interaction: they are **pragmatic** in character" (Lewis 1993:94). They include 'short, hardly grammaticalized utterances' (e.g. *Not yet. Certainly not. Just a moment, please*), sentence heads or frames (e.g. *Sorry to interrupt, but can I just say ... ; That's all very well, but ...*) and full sentences with readily identifiable pragmatic meaning.

Collocations are defined as "describing the way individual words co-occur with others" (Lewis 1993:93):

> Possible two-word combinations vary from the totally unexpectedly novel – free collocation – to the rigidly institutionalized or ossified form – fixed collocation[30]. (...), this is not a dichotomy but a spectrum between fixed and free poles. (Lewis 1993:93)

Lewis (2001) rephrases his definition of collocation as follows:

> Collocation is the way in which words co-occur in natural text in statistically significant ways (Lewis 2001:132).

His definition thus includes a ragbag of word combinations such as *submit a report* and *examine thoroughly* that are widely recognized as collocations, compounds (*radio station, fire escape*), discourse markers (*To put it another way*), grammatical collocations (*aware of*),

[30] Quite surprisingly, fixed collocations are then referred to as "one kind of polyword" (Lewis 1993:93).

phrasal verbs (*turn in*), binomials and trinomials (*backwards and forwards*) and incomplete fixed phrases (*a sort of*…):

> If we define collocation as the way words occur together, it is easy to see that the definition is very wide, and will cover many different kinds of item. Certainly, all of the following are collocations in the sense that we readily recognise that these groups of words are regularly found together:
>
> 1.*a difficult decision* (adjective + noun)
>
> 2.*submit a report* (verb + noun)
>
> 3.*radio station* (noun + noun)
>
> 4.*examine thoroughly* (verb + adverb)
>
> 5.*extremely inconvenient* (adverb + adjective)
>
> 6.*revise the original plan* (verb + adjective + noun)
>
> 7.*the fog closed in* (noun + verb)
>
> 8.*To put it another way* (discourse marker)
>
> 9.*a few years ago* (multi-word prepositional phrase)
>
> 10. *turn in* (phrasal verb)
>
> 11. *aware of* (adjective + preposition)
>
> 12. *fire escape* (compound noun)
>
> 13. *backwards and forwards* (binomial)
>
> 14. *hook, line and sinker* (trinomial)
>
> 15. *On the other hand* (fixed phrase)
>
> 16. *A sort of* ... (incomplete fixed phrase)
>
> 17. *Not half!* (fixed expression)
>
> 18. *See you later/tomorrow/on Monday* (semi-fixed expression)
>
> 19. *Too many cooks* ... (part of a proverb)
>
> 20. *To be or not to be* ... (part of a quotation)
>
> Lewis (2001:133-134)

Despite their relative lack of theoretical soundness and terminological consistency[31], descriptions of phraseological units in ELT deserve praise for placing emphasis on the important role played by institutionalised expressions or lexical phrases in language. For example, Willis (2003:144-148) discusses in detail frames and sentence stems. **Frames** consist of discontinuous sequences of words that can be filled by a whole range of lexical items, depending on the context. Examples include *from a(n)* adj. *point of view*; *are not ... but; whatever ... are necessary*. Frames can fulfil a number of functions such as providing a framework for a whole sentence (e.g. *not only* X *but (also)* Y; *the* ___*er* X, *the* __*er* Y) (cf. also the notion of 'collocational framework' in section 2.4.2.2.3). **Sentence stems** provide an introduction to a sentence (e.g. *Do you mind if* …?; *What I mean is* …). They are sometimes referred to as "form/function composites since the form strongly signals the function it fulfils"

[31] The reader is referred to Gouverneur (in preparation) for a detailed review of the treatment of phraseological units in ELT studies as well as in ELT materials.

(Willis 2003:148). In 1983, Pawley and Syder described the dual status of a (lexicalized) sentence stem[32] in the following terms:

> On the one hand, its potential occurrence and meaning is predicted by the productive rules of syntax and semantics. These account for its status as a grammatical string and specify its internal structure and structural relationship to other sequences; they do not, however, mark the sequence as having a special status among the set of grammatically possible strings. On the other hand, the dictionary entry for the same sequence (…) should note its status as a lexical item, a (somewhat) arbitrary selection as a standard expression or name for a culturally authorized concept; that is, it should record the fact that the sequence **is an actually occurring, nativelike form, a 'common usage' having an institutionalized function**[33], in contrast to other sequences which do not have this status. (Pawley and Syder 1983:216)

In academic discourse, for example, sentence stems are often used for hedging (*It seems that…*; *Our results suggest that* …) and to introduce a research topic (*The aim/purpose/goal/object of this study is to analyse/investigate/establish* ...).

2.2.3. Typologies and psycholinguistics

Another type of model of the phraseological spectrum has recently been put forward by Wray (1999; 2000; 2002) and Wray and Perkins (2000). This model differs from previous classifications in that it is deeply rooted in a **psycholinguistic** perspective on language. The authors are primarily concerned with developing an explanatory model of **formulaicity** or **formulaic language**. Such an approach explains the different type of angle adopted in the definition of their object of study. Wray defines a **formulaic sequence** as follows:

> a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (Wray 2002:9)

[32] It should be noted that Pawley and Syder's (1983) definition of 'sentence stem' is broader than that of Willis (2003) as it includes full sentences: "A sentence stem consists either of a complete sentence, or, more commonly, an expression which is something less than a complete sentence" (Pawley and Syder 1983:210).
[33] My emphasis.

The model proposed by Wray and Perkins is an attempt at relating two types of functions identified in the literature for formulaicity in language. The first function of formulaic sequences is that they **perform communicative acts** (see Cowie's definition of routine formulae in section 2.2.1). Wray and Perkins (2000) identify three types of socio-interactional function for formulaic sequences relating to the addressor:

- **manipulate the others** (e.g. commands: *Keep off the grass;* requests: *Could you repeat that please?;* politeness markers: *I wonder if you'd mind …*);
- **assert one's own identity** (e.g. being taken seriously: *You're never going to believe this, but ...;* separating from the crowd: *I wanna tell you a story*);
- **assert group identity** (e.g. overall membership: *Praise the Lord*!; place in hierarchy - threats: *I wouldn't do that if I were you*; place in hierarchy - forms of address: *Your Highness*).

The second type of explanation for formulaicity relates to their potential for **reducing processing efforts**. Wray and Perkins subcategorize these processing functions of formulaic sequences into three types:

- **processing short-cuts** (e.g. standard ideational labels with agreed meanings such as *personal computer* and *bullet point*);
- **time-buyers** such as fillers (*if you want my opinion*), turn-holders (*and another thing* ...), and discourse shape markers (*There are three points I want to make. Firstly ... Secondly ... Thirdly / Lastly ...*);
- sequences that are used to **manipulate information** (e.g. mnemonics and rehearsal for memorization).

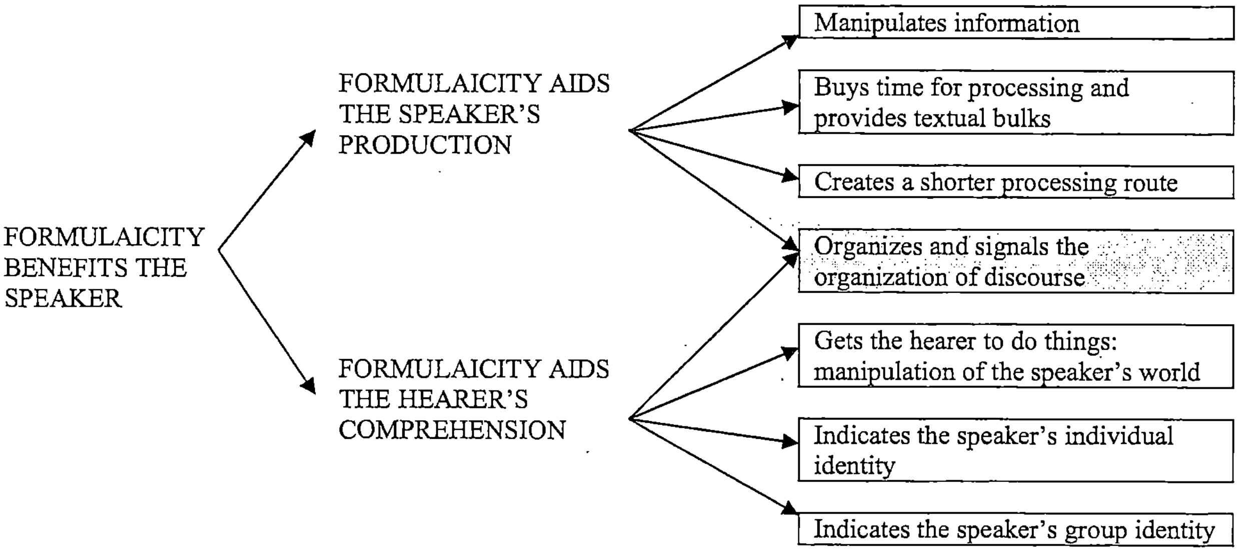
According to Wray (2000), one way of accommodating the two functions within a single model is to view them in terms of their "easing of either the speaker's or the hearer's processing pressures" (478), which makes it possible to represent them as "two intersecting parts of the same strategy" (ibid).

Figure 2.6 shows that the use of formulaic sequences **benefits the speaker** in two different ways. Formulaicity aids the speaker's production by reducing processing efforts. On the other hand, it can also support the "speaker's interactive goals through maximizing the chances of hearer comprehension" (ibid). The grey part of the diagram represents specific types of formulaic sequences which simultaneously facilitate speaker production and hearer comprehension, i.e. **discourse markers**:

> They both anchor the structure of the speaker's output, so that it is easier to sequence ideas fluently, and, simultaneously, signal to the hearer where it will be most appropriate, and inappropriate, to begin a turn and what the overall character of the speaker's message is. (Wray 2000:478)

Unlike Burger (1998), Wray and Perkins assign a prominent role to word combinations that are used to organise and signal the organization of discourse.

Figure 2.6: Wray's (2000:478) roles of formulaic sequences in benefiting the speaker


FORMULAICITY BENEFITS THE SPEAKER
FORMULAICITY AIDS THE SPEAKER'S PRODUCTION
FORMULAICITY AIDS THE HEARER'S COMPREHENSION
Manipulates information
Buys time for processing and provides textual bulks
Creates a shorter processing route
Organizes and signals the organization of discourse
Gets the hearer to do things: manipulation of the speaker's world
Indicates the speaker's individual identity
Indicates the speaker's group identity

2.2.4. Conclusion and typology adopted in this thesis

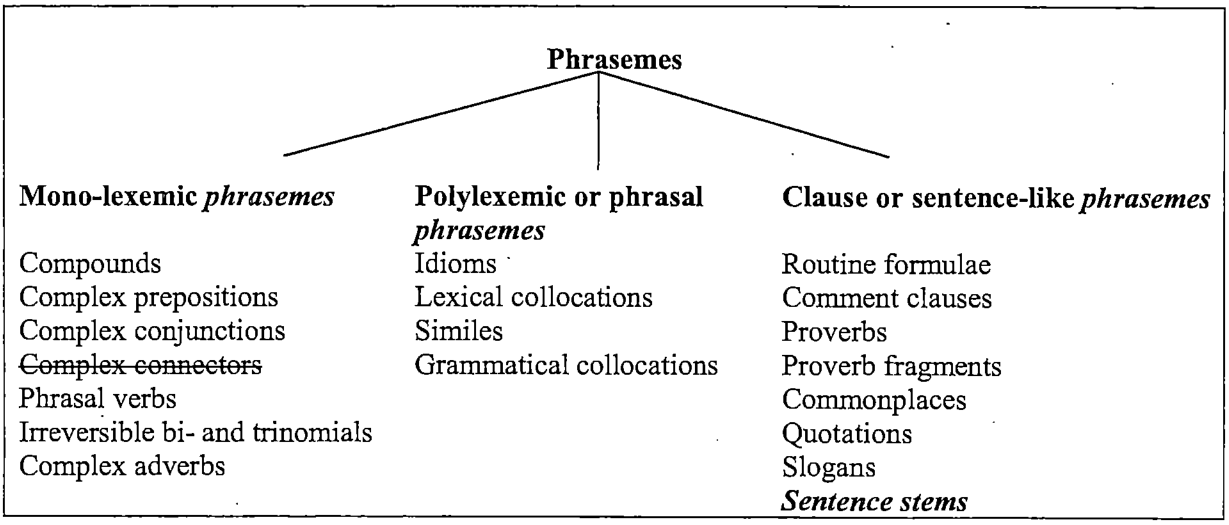
As already stressed by Cowie (1998: 210), "[c]ategorization is notoriously difficult in phraseology because of the bristling array of variables – syntactic, pragmatic, stylistic, semantic – which the material is constantly throwing up." Differences between typologies are largely explainable in terms of the importance attached by the authors to each of these variables, their delimitation of the phraseological spectrum and their objectives in classifying phrasemes. Categorizing phrasemes according to their form has been said to lend itself "rather better to purely descriptive accounts (...) than to explanatory ones with the result that the outcomes tend to be less consequential than those of other approaches" (Wray 2002:48). However, descriptive accounts of phrasemes are still very much needed in language teaching, lexicography, and natural language processing (cf. Tschichold 2000; Michiels 2002). They have the advantage of grouping together phrasemes which have the same syntactic function and, thus, of giving useful information on the use of these word combinations (cf. Pecman 2004:114).

Typologies proposed by Cowie (1988; 1991) and Mel'čuk (1995; 1998) place emphasis on the forms of phrasemes and the functions they perform in discourse. Following Wray (2002:48), we believe that "cross-associations such as these between form and function (...) are probably nearer the truth than single-parameter categorizations, but their additional complexity also clouds the picture, with the assignment of items to one or another subcategory becoming difficult at times." Consequently, an approach such as the one adopted by De Cock (2003) seems more appropriate to the description of phrasemes. As described in section 2.2.1, the author proposes two models – a formal model and a functional model - of the phrasicon to avoid linking form and function.

In this thesis, we adopt De Cock's (2003) **structural** model of the phrasicon except for the category of **complex connectors** which was shown to partly overlap with that of complex adverbs. Following ELT researchers such as Lewis (1993; 2001) and Willis (2003), we also add a new sub-category of sentence stems (in bold italics in Figure 2.7). **Sentence stems** can be broadly defined as routinized introductory fragments of sentence which constitute sequences of two or more clause constituents. They thus fall in the general category of clause or sentence-like phrasemes.

So far, the objects of study of phraseology have been interchangeably referred to as 'phrasemes', 'phraseological units', 'formulaic sequences' and 'multi-word units' to follow the terminology of the typologies reviewed. In this thesis, we prefer the more neutral **phraseme** to the other three terms: by using 'unit' and 'sequence', it is believed that too much emphasis is placed on 'holisticity' or 'unity' while it will be seen in the next sections that not all types of phrasemes can be easily described as units. Another argument in favour of 'phraseme' is that the term is built in accordance with other well-established linguistic terms such as morpheme, lexeme, etc. Phraseology can be loosely defined as "**the study of the structure, meaning and use of word combinations**" (Cowie 1994: 3168). Lacking in precision, however, this definition does not help circumscribe the field and is thus only intended as a first working definition. We shall propose our own definition of phraseology in section 2.3.7 after a review of the major defining criteria of phrasemes.

Figure 2.7: The structural model adopted in this thesis (based on De Cock 2003)



Phrasemes

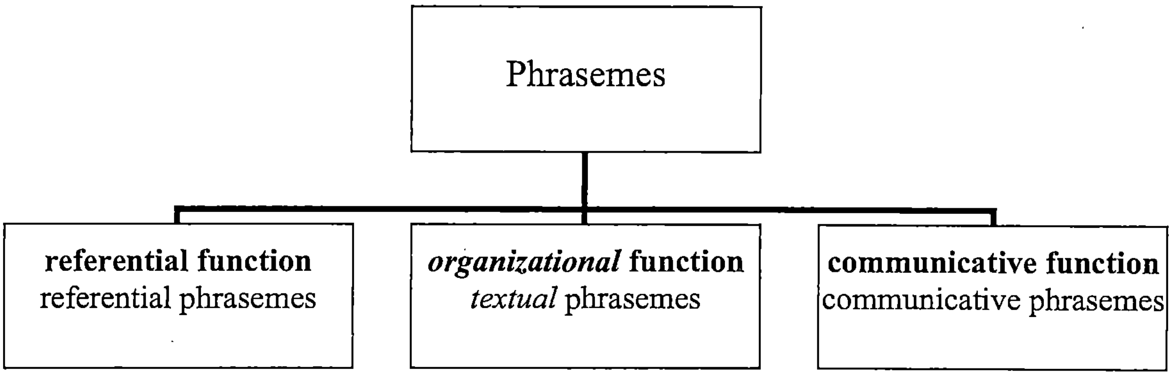
Mono-lexemic *phrasemes*	**Polylexemic or phrasal *phrasemes***	**Clause or sentence-like *phrasemes***
Compounds	Idioms	Routine formulae
Complex prepositions	Lexical collocations	Comment clauses
Complex conjunctions	Similes	Proverbs
~~Complex connectors~~	Grammatical collocations	Proverb fragments
Phrasal verbs		Commonplaces
Irreversible bi- and trinomials		Quotations
Complex adverbs		Slogans
		Sentence stems

As regards the **functional** model, we believe that a tripartite model such as the one adopted by Burger (1998) is more useful for the classification of phrasemes than a typology which distinguishes between semantically vs. pragmatically specialized phrasemes. Figure 2.8 presents the functional typology adopted in this thesis. Phrasemes are classified into three broad functional categories:

- **Referential phrasemes** are used to convey a content message: they refer to objects, phenomena or facts of life. They include collocations, idioms, similes, irreversible bi- and trinomials, compounds, phrasal verbs, etc.
- **Textual phrasemes** are used to organize the content (i.e. referential information) of a text or any type of discourse. We do not use Burger's term of 'structural phrasemes' as our category of textual phrasemes is broader than Burger's category, which is limited to complex prepositions and complex conjunctions. Textual phrasemes also include a wide range of sentence stems that are typically used to serve organizational or rhetorical functions.
- **Communicative phrasemes** are used to express feelings or beliefs towards a propositional content or to explicitly address interlocutors, either to focus their attention, include them as discourse participants or influence them. They include routine formulae, attitudinal formulae, commonplaces, etc.

More information about the different types of phrasemes included in each category will be given in section 2.3.7 after a review of their major defining criteria.

Figure 2.8: Functional typology adopted in this thesis (based on Burger 1998)



2.3. A common core of defining criteria

Gonzalez Rey (2002) lists 20 features that can be used to set phrasemes apart from other word combinations (see Table 2.1). This section is not intended as a review of all possible features of phrasemes but rather provides a detailed discussion of 7 criteria that serve as the basis of the typologies reviewed in section 2.2, namely polylexicality, institutionalization and lexicalisation, compositionality, semantico-pragmatic function, collocability and syntactic flexibility. Special attention will be paid to the way some of these criteria are used to distinguish between nonce combinations (see section 2.3.5 for a justification of the use of 'nonce' instead of 'free' combinations), collocations and idioms.

Table 2.1: Gonzalez Rey's (2002:52) distinguishing features of phrasemes

la polylexicalité ; la fréquence; le figement ou la fixité; le défigement; désautomatisation ou délexicalisation; l'institutionalisation; l'idiomaticité; la figuralité; l'iconicité; l'opacité; l'ambiguïté; l'écart ou déviation; la moulabilité ou reproductibilité; la répétition; la reproduction; les différents registres; la réductibilité; l'arbitrariété, la motivation et la démotivation ; la valeur métaphorique ; la remétaphorisation ; les éléments expressifs et les procédés productifs

Two additional criteria worthy of note are frequency and register. We will discuss **frequency** together with institutionalization in section 2.3.2 and will come back to this criterion in section 2.4.2 when introducing the distributional approach to collocations and in section 4.2.3 when describing quantitative methods for the extraction of phrasemes from corpora. The relationship between phrasemes and **register** will be addressed in section 2.4 when zooming in on collocations and section 2.5 when considering the links between phraseology and discourse.

2.3.1. Polylexicality

Phraseology has been defined in section 2.2.4 as "the study of the structure, meaning and use of word combinations" (Cowie 1994: 3168). In Saussurean terms, meaning relations not only exist between members of a paradigm *in absentia* (paradigmatic relations such as synonymy, hyponymy and antonymy) but can also be viewed from a sequential or syntagmatic point of view (cf. Van Roey 1990:72). **Syntagmatic relations** hold between items *in praesentia*, which occur in a linear sequence:

> D'une part, dans le discours, les mots contractent entre eux, en vertu de leur enchaînement, des rapports fondés sur le caractère linéaire de la langue, qui exclut la possibilité de prononcer deux éléments à la fois. Ceux-ci se rangent les uns à la suite des autres sur la chaîne de la parole. Ces combinaisons qui ont pour support l'étendue peuvent être appelées *syntagmes*. Le syntagme se compose donc toujours de deux ou plusieurs unités consécutives (par exemple: *re-lire; contre tous; la vie humaine; Dieu est bon; s'il fait beau temps, nous sortirons,* etc.). (Saussure 1982:170-171)

Not all syntagmatic relations are phraseological. In addition to criteria related to institutionalization and lexicalisation, compositionality, semantico-pragmatic function, collocability and syntactic flexibility (see following sections), restrictions are generally placed on what constitutes phraseological syntagmatic relations.

Polylexicality is generally described as one of the first necessary conditions for inclusion in the phraseological spectrum (cf. Gross 1996; Mejri 2005; Montoro del Arco 2006)[34]. As stated by Gross (1996),

> La première condition nécessaire pour que l'on puisse parler de figement est que l'on soit en présence d'une **séquence de plusieurs mots** et que ces mots aient, par ailleurs, une **existence autonome**[35]. (Gross 1996:9)
>
> [The first necessary condition in order to speak of phrasemes is to be in the presence of a sequence of several words which also happen to exist autonomously.] [36]

This definition, however, raises more issues than it settles as it shifts the emphasis on what constitutes a **word** and what constitutes an **autonomous existence**. Gross uses the criterion of

[34] One notable exception is Zuluaga (1980) who includes single words such as *Salud* (En. 'cheers') and *Adiós* (En. 'bye bye') in the phraseological spectrum on the basis that these words display pragmatic fixedness (Sp. 'fijación pragmática').

[35] Emphases are mine.

[36] 'Polylexicality' has also sometimes been used to refer to a semantic feature of phrasemes. Mejri (2005), for example, writes « En fait, il s'agit [la polylexicalité] d'une caractéristique propre aux SF [séquences figées], qui, contrairement aux dérivés par exemple, se distinguent par un signifiant pluriel (= *poly*) formé de plusieurs unités lexicales employées d'une manière autonome hors du cadre de la séquence (= *lexical*) ».

'autonomous existence' to distinguish fixed sequences from derived words, i.e. roots and their derivational affixes. Several 'words', however, never occur outside phrasemes. Moon (1998a:78-79) lists a number of examples: *kith* in *kith and kin*, *dint* in *by dint of*, *amok* in *run amok*, *cahoots* in *be in cahoots with*, *fro* in *to and fro*, *kibosh* in *put the kibosh on*, *fettle* in *in fine/good fettle*, *umbrage* in *take umbrage*, *retrospect* in *in retrospect* and *amends* in *make amends*.

The term 'word' is often used to refer to a **word form**, i.e. "any sequence of letters (and a limited number of other characteristics such as a hyphen and apostrophe) bounded on either side by a space or punctuation mark" (Carter 1998:4). It can also be used to refer to a **lemma**, i.e. a headword (the non-inflected word form) as well as its inflected and reduced forms (cf. Nation 2001:7). Thus, *take, takes, taken, taking* and *taken* are concrete realizations, that is, word forms, of the lemma TAKE[37]. Lemma and lexeme are sometimes used as synonyms (cf. McEnery et al 2006:35-36). The term **lexeme**, however, has also been used to refer to lexical items which consist of more than one 'word', e.g. phrasal verbs (*to give up, to drop in*) and idioms (*kick the bucket*) (cf. Carter 1998:7). To avoid confusion, we shall therefore not use 'lexeme' in this thesis except when discussing the literature. One distinction which is often made is that between grammatical words and lexical words. **Grammatical or function words** include articles (*a, the*), pronouns (*I, you, he*), prepositions (*of, with, in*), conjunctions (*and, but*) and other types of closed class words. **Lexical or content words**, i.e. "words which have 'lexical meaning'" (Van Roey 1991:13) include nouns (*dog, table, computer*), verbs (*eat, write, sleep*), adjectives (*beautiful, sad*) and adverbs (*happily, fortunately, then, however*)[38]. They carry higher information content than grammatical words, which syntactically structure lexical words (cf. Carter 1998:8).

Several definitions of phrasemes proposed in the literature use the term 'word' without explicitly specifying what it refers to (e.g. Cowie's (1994) definition above or Wray's (1999) definition in section 2.2.3). Other definitions describe phrasemes as combinations of **at least two graphic words**, setting their upper limit at the level of the complex sentence (see, e.g., Corpas Pastor 1996; Montoro del Arco 2006). This orthographic criterion, however, creates internal inconsistencies, especially as far as the status of **compounds** is concerned. Compounds such as *black market* and *red tape* are categorized as phrasemes according to the

[37] Lemmas are often written in small capital letters in the literature. In this thesis, most references will be made to single words as lemmas unless specified otherwise. We will therefore only use small capital letters in ambiguous contexts where it is necessary to distinguish lemmas from word forms.

[38] Note that adverbs such as *then* and *however* are sometimes classified as grammatical words. In this thesis, all adverbs are regarded as content words.

orthographic criterion while others such as *blackmail* and *blackbird* are not. Similarly, a word such as *horse-riding* is classified as a phraseme when spelt *horse riding* but does not belong to the phraseological spectrum when spelt *horse-riding*. Gross (1996) summarizes the difficulties caused by the orthographic criterion in the following words:

> Reste le problème des séparateurs entre les différents éléments lexicaux, qu'on ne doit pas réduire à un simple problème de graphie. On admettra comme séparateurs le trait d'union, l'apostrophe et le blanc. Faut-il accepter la soudure ? Des séquences comme *vin* et *aigre* sont assurément des mots autonomes, faut-il de ce fait considérer le mot *vinaigre* comme une suite figée ou un nom simple ? Si la perspective n'est pas exclusivement formelle, on ne peut répondre à cette question de manière équivoque. (Gross 1996:10)

Barkema (1996b:133) tries to make a distinction between phrasemes and compounds on the basis of a syntactic criterion. He distinguishes between **constructions**, i.e. phrasemes, which have "at least two grammatical function slots realised by lexical items other than the definite or indefinite article" (e.g. *blind alley*) and **compounds** which have "one function slot only" (e.g. *ice-cream, baby carriage, sidewalk*). De Cock (2003:36), however, highlights the difficulty of making a clear-cut distinction between constructions and compounds and argues that *blind alley* could also be seen to fill one grammatical slot.

It is argued here that **compounds**, whether written separately or not, can legitimately be considered as phrasemes as they are morphologically made up of at least two elements which have independent status outside their combinations. The fact that they constitute an important part of the field of morphology does not preclude them from being an object of study of phraseology as well. The syntagmatic relationship between the two elements can be described along a cline of institutionalization, lexicalization, compositionality and flexibility, four criteria which play an important role in the description of phrasemes (see following sections). Each discipline will thus throw light on different aspects of compounds and this double perspective can only be beneficial to the study of this specific type of syntagmatic relations.

Including compounds in the phraseological spectrum also causes difficulties as they must be distinguished from phrases. In a number of articles on the distinction between phrases and compounds, Giegerich draws a parallel between the phrase/compound distinction and "the distinction drawn in formal grammar between the syntax and the lexicon as sites for the concatenation of linguistic units" (Giegerich 2006:5). Broadly speaking, syntax produces highly productive and regular phrases while compounds originate in the lexicon. The author, however, argues that associative Adj. +N[39] (e.g. *dental decay, avian influenza, vernal equinox, tropical fish*) and attribute-head N+N constructions[40] (e.g. *country house, olive oil, pine cone, ice cream, arm-chair*) straddle the lexicon-syntax divide (Giegerich 2004; 2005; 2006). These constructions have an "irreconcilably hybrid status in English: their structural characteristics identify them as objects of the lexicon, while at the same time they may behave as though they were syntactic constructions" (Giegerich 2005:4). Thus, associative Adj. + N constructions can be said to enjoy lexical status on both syntactic and semantic grounds. First, adjectives in these constructions cannot occur in predicative position and show a restricted distribution (e.g. *vernal equinox* but not **vernal flowers* or **vernal weather*). Second, several semantic relationships can develop between the elements of associative Adj. + N constructions: the constructions may express an argument-predicate structure inherited from a predicate contained in the noun (e.g. *presidential election, papal visit*) or "express the less structured relationship of 'associated with', 'to do with' (which is then often augmented by encyclopaedic knowledge on the speaker's part)" (Giegerich 2005:579), e.g. *papal emissary*. Some of these constructions, however, are also available for the pro-*one* construction, which is clearly the outcome of a syntactic operation. For example, head nouns of *dental building, mental hospital* and *financial advisor* can be replaced by the pronoun *one* in the following examples:

2.1. *Is this the medical building or the dental one?*
2.2. *Is this the general hospital or the mental one?*
2.3. *Is he a legal advisor or a financial one?*

Giegerich (2005: 588) concludes that "there are actually individual associative AdjNs (*dental building, mental hospital* etc.) which are simultaneously lexical entities ('compounds') in some respects and syntactic entities ('phrases') in other respects".

Gramley and Pätzold define 'multi-word units' or 'lexical phrases' as "well-established lexical combinations which consist of one or more word forms or lexemes" (1992:53). They most probably combine 'word form' and 'lexeme' in their definition in order to account for different types of phrasemes:

[39] Associative adjectives express a property which "does not apply directly to the denotation of the head nominal, but rather to some entity associated with it" (Pullum and Huddleston 2002: 556). In *dental decay*, for example, *dental* does not describe the nature of the decay but identifies what is decaying.

[40] The core meanings of attribute-head N+N constructions are the 'made of' relationship (e.g. *steel bridge, fruit juice, vegetable oil*) and the 'belonging to' or 'associated with' relationship (e.g. *oak leaf, table leg, mountain peak, rose petal*).

- Some phrasemes typically show syntagmatic relations between two or more word forms. For example, the phraseme '*how do you do?*' used as a polite greeting when you meet someone for the first time cannot be realized by other word forms than '*how*' + '*do*' + '*you*' + '*do*'.
- A larger proportion of phrasemes are combinations of word forms in which one or more constituents can be replaced by other realizations of the same lemma. Thus, **kick a bucket* and **kick the buckets* are incorrect realizations of the idiom '*kick the bucket*' as they involve the indefinite instead of the definite article and the plural instead of the singular form of BUCKET respectively. However, all word forms of the lemma KICK can be used in this idiom. Other examples include RAIN *cats and dogs* and SPILL *the beans*.[41]
- Collocations can be defined as syntagmatic relations between two lexemes. Thus, *make a decision, makes a decision, making a decision, made a decision, make decisions, a decision was made*, etc. are all instantiations or realizations in discourse of the collocation MAKE + DECISION. As argued by Nesselhauf (2005:25), "this does not mean, however, that it is assumed that all theoretically conceivable instantiations of a certain collocation exist, let alone that they are equally common."

Gramley and Pätzold's use of the terms 'word form' and 'lexeme' in their definition of multi-word units can arguably be criticized as being a mix of two different levels of abstraction, word forms being realizations of lexemes in discourse. It may be more useful to define phrasemes as **syntagmatic relations between lemmas whose lexical realizations can be restricted in usage** (see section 2.3.7 for our complete definition of phrasemes). Taking into account the potential (in)flexibility of a phraseme, its different elements can be inflected to meet the specific needs of a given communicative situation. In the same way as a word form can be defined as an instantiation of an abstract lemma, we can distinguish between a phraseme and its realizations in discourse. The term 'phraseme' will be used both to refer to the abstract unit of language and its realizations in discourse.

The many examples given in this section suggest that phrasemes are typically combinations of **lexical words**. However, they can also consist of combinations of at least one content word and one or more grammatical words. Examples include *by the way, so long, of*

[41] Even phrasemes that are generally described as 'totally fixed' may present some degrees of variation. In fact, Moon's (1998) corpus-based study of fixed expressions and idioms has shown that there is much more instability in their forms than previously thought.

course (not), as a result (of) and *depend on*. While most of these examples are commonly regarded as phrasemes (cf. section 2.2), word combinations such as *depend on* and *afraid of* have not been included in traditional typologies of phrasemes (e.g. Cowie 1994; Gläser 1998). There is, however, a case for including this specific type of word combinations, which have often been called **grammatical collocations,** [42] into the phraseological spectrum. First, they involve syntagmatic relations between two lemmas. They consist of a content word which arbitrarily selects another word, and in that sense, they resemble lexical collocations. In addition, syntagmatic relations between a content word and a function word can be described along a similar cline to that of nonce combinations (e.g. *to run* ***out of*** *the house*), collocations (e.g. *interest* ***in****, to regard sb. /sth.* ***as*** ..), and idioms (*to make out, to get about, to get across, to give up*)[43]. Finally, grammatical collocations are often analysed as part of the **valency** (or 'valence') of words. It is noteworthy that Allerton (1994:4878) defines valency as a kind of **'lexico-syntactic property'** and not as a syntactic dependency, which seems to suggest that syntax alone cannot account for all the properties of the valency of a word:

> a particular kind of dependency property of lexical items. This kind of **lexico-syntactic property**[44] involves the relationship between, on the one hand, the different subclasses of a word-class (such as verb) and, on the other, the different structural environments required by those subclasses, these environments varying both in the number and in the type of elements. Valency is thus seen as the capacity a verb (or a noun, etc.) has for combining with particular patterns of other sentence constituents, in a similar way to that in which valency of a chemical element is its capacity for combining with a fixed number of atoms of another element.

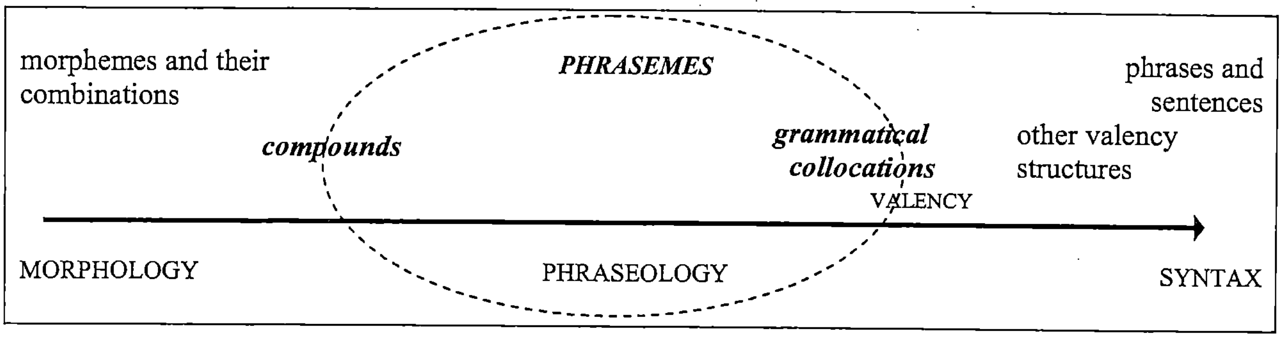
It will be seen in section 2.4.2.2 that the distributional approach to collocations and corpus linguistics played an important part in including in the phraseological spectrum aspects of the environment of words that were usually considered parts of their valency. Compounds were described above as being the object of study of both morphology and phraseology. Figure 2.9 shows that, like compounds, grammatical collocations can be described as being situated at the interface between two disciplines: syntax and phraseology.

[42] See section 2.4.1.2.1 for other uses of the term 'grammatical collocation'.

[43] Idiomatic syntagmatic relations between a content word and a function word constitute 'phrasal verbs'. They are not referred to as 'idioms' in this thesis as this category has been described above as being composed of polylexemic phrasemes (cf. section 2.2.4).

[44] My emphasis.

Figure 2.9: Syntagmatic relations



2.3.2. Institutionalization and lexicalization

The terms institutionalization and lexicalization come from the discipline of word formation. The two phenomena have been given many different definitions (cf. Bauer 1983; Barkema 1996b; Corpas Pastor 1996) and are not always clearly distinguished (cf. Pawley and Syder 1983; Blasco Mateo, 2002). Bauer (1983: 48) regards these two phenomena as consecutive stages in the development of a morphologically complex word. A lexeme becomes institutionalized when "the nonce formation starts to be accepted by other speakers as a known lexical item" (ibid). At this stage, a lexeme loses its potential ambiguity and becomes current with a specific, though transparent, meaning:

> Thus, for example, there is nothing in the form *telephone box* to prevent it from meaning a box shaped like a telephone, a box which is located at/by a telephone, a box which functions as a telephone, and so on. It is only because the item is ***familiar***[45] that the speaker-listener knows that it is synonymous with *telephone kiosk*, in the usual meaning of *telephone kiosk*. (Bauer 1983: 48)

Put differently, institutionalized lexemes "still form part of a synchronically productive series, differing only from potential words in that, by being used, they have come to have a specific reference" (Bauer 2001:46). **Institutionalization** is thus a diachronic and socio-linguistic process by which a new lexical item integrates, with its particular form and meaning, into a language community's existing stock of words. A word combination becomes institutionalized as soon as it is recognized by the members of a language community as a 'bound' or 'preferred' sequence (cf. Pawley and Syder's 'native-like selection'), with its particular syntactic and semantic features, and enters their mental lexicon. As already stated by Howarth (1996: 37) or Moon (1998a: 7), institutionalization is a necessary defining

[45] My emphasis.

criterion for all types of phrasemes: phraseology is the study of all word combinations that are more or less familiar to a language community.

For a lexeme or a word combination to be accepted by a speech community and consequently institutionalized, it has to be used and repeated (Bally 1909:66; Coulmas 1979; Corpas Pastor 1996:21-23). This is certainly the reason why degrees of institutionalization have often been assessed by a **quantitative** criterion, especially in corpus linguistics (see sections 2.4.2 and 4.2.2). However, authors such as Moon (1998c) and Barkema (1993) have shown that not all types of phrasemes can be frequently found in corpora. Assessing degrees of institutionalization by means of a quantitative criterion is only valid for specific subtypes of phrasemes, and more specifically for collocations.

As for **lexicalization**, it is a diachronic linguistic process by which lexical constructions are withdrawn from analytic access and accessed holistically (cf. Lehmann 2002). In other words, it is a process by which a construction becomes "subject to one or more types of idiosyncratic restriction, i.e. restriction that cannot be formulated in terms of general rules" (Barkema 1997: 28), e.g. *as drunk as a lord, open* someone's *eyes to* something, *spoil* somebody *rotten, hold the key,* etc. As Moon (1998a: 39) explains, lexicalization "results from a three-way tension" between the "criterion of institutionalization, the lexicogrammatical criterion of fixedness, and the qualitative criterion of non-compositionality." Thus, lexicalization cannot be assessed independently of flexibility, collocability and compositionality (see sections 2.2.3, 2.2.5 and 2.2.6).

The process of lexicalization has often been opposed to that of **grammaticalization**, i.e. "the process whereby content words become more grammatical (and already grammatical morphemes even more grammatical)" (Hoffmann 2004:188)[46] but the two processes have much in common (cf. Campbell 2001). Lehmann (2002) argues that they are better described as orthogonal to each other. Both "do not concern signs in isolation, but signs in their paradigmatic and syntagmatic relations" (Lehmann 2002:15). They are reduction processes in that they "constrain the freedom of the speaker in selecting and combining the constituents of a complex expression" (ibid: 17), but in a different sense. Grammaticalization reduces the autonomy of the unit, "shifting it to a lower, more strictly regulated grammatical level" while lexicalization "reduces the inner structure of the unit, shifting it into the inventory" (ibid). Lehmann summarizes the basic differences between grammaticalization and lexicalization as follows:

[46] See Campbell and Janda (2001) for a survey of the numerous definitions of grammaticalization in the literature.

> Let [XY]z be a complex construction which undergoes grammaticalization or lexicalization. Then the differences between the two processes consist in two aspects. First, in grammaticalization there may be a constituent of Z, e.g. X, which is the focus of the process and which is changed into a grammatical formative by it. In lexicalization, there is no such constituent; the lexicalization affects Z as a whole. From this it follows that lexicalization necessarily concerns an internally complex unit, whereas we may reasonably speak of grammaticalization even with respect to simple units.
> Second, in grammaticalization the internal relations of Z become more strict and constrained. This regards, in particular, the relation between X and Y or between X and Z. Again, in lexicalization the internal relations of Z become irregular and get lost. (Lehmann 2002:15)

These differences have two major implications according to Lehmann (2002). First, while lexicalization can only apply to complex units, grammaticalization may simultaneously affect complex units and one of their constituents. Thus, grammaticalization may simultaneously have had an impact on the complex preposition *in view of* and on its nominal element, making *view* lose the features that defined it as a noun. Second, lexicalization plays a role as the 'preparatory phrase of grammaticalization'. Lehmann argues that the first process which affects (complex) prepositions and conjunctions is lexicalization, not grammaticalization. Thus, before undergoing a process of grammaticalization, the complex preposition *in view of* has lost its compositional meaning. This may explain why only a fraction of complex prepositions will eventually grammaticalize (see Hoffmann 2004 for another explanation, based on the principle of analogy, for the grammaticalization of complex prepositions).

Institutionalization, lexicalization and grammaticalization are "not of an all-or-none kind, but of a more-or-less kind" (cf. Lipka 1994:2165; Hoffmann 2004). The processes result in degrees of institutionalization, lexicalization and grammaticalization in synchrony. This continuum from not institutionalized/lexicalized/grammaticalized to institutionalized/ lexicalized/ grammaticalized word combinations was already described by Bally (1909: 68) in the following terms:

> On peut donc dire que la combinaison des mots entre eux varie d'aspect dans les limites formées par deux cas extrêmes: 1) l'association se désagrège aussitôt après sa formation, et les mots qui la composaient recouvrent leur entière liberté de se grouper autrement ; 2) les mots, à force d'être employés ensemble pour l'expression d'une même idée, perdent toute autonomie, ne peuvent plus se séparer et n'ont de sens que par leur réunion. On comprend qu'entre ces deux extrêmes il y a place pour une foule de cas intermédiaires qui ne se laissent ni préciser ni classer.

Phrasemes such as idioms, phrasal verbs, irreversible bi- and trinomials, proverbs and commonplaces are lexicalized word combinations. On the other hand, most collocations (e.g. *blond/auburn hair, chestnut horse, prove guilty, conduct an experiment*) are not fully lexicalized as they are not semantically, syntactically or collocationally totally 'fixed'. Finally, complex prepositions and complex conjunctions, for example, reflect the process of grammaticalization (cf. Moon 1998a:217) as well as verbs in support-verb constructions (e.g. *take a step, do a favour*) (cf. section 2.4.1.2.2 and Batoux 2003).

2.3.3. Degrees of compositionality

Barkema defines the compositionality of a construction as "the extent to which its meaning is the combinatorial result of the basic or derived senses of the lexical items in the construction and the syntactic relations in the constituent that contains these lexical items." (1996b: 138) This definition is thus based on two preliminary notions:

1. **'Basic' sense**: "The basic sense of a lexical item is the sense from which other senses can be systematically derived by means of extension. It is usually the first sense that comes to a speaker's mind when s/he comes across a lexical item in isolation" (Barkema 1996b:158). Barkema seems to suggest that a 'basic sense' is identifiable by its psychological saliency or prototypicality, a feature which has been described in the literature as particularly difficult to evaluate (Tsohatzidis 1990; Violi 2000; Gilquin 2005).
2. **'Derived' sense**: "A derived sense of a lexical item is every sense, other than the basic sense, that a word has in isolation." (ibid: 158)

Barkema does not seem to acknowledge the difficulty of distinguishing between basic and derived senses of a lexical item as they contribute in the same way to his definition of compositionality. Neither does he address the central issue of separating senses that a word has in isolation from those that it acquires in context. However, this fundamental distinction is arguably the source of the vexed question of where to draw the line between collocations and nonce combinations (see sections 2.3.3.3 and 2.4.1.2).

Barkema distinguishes four levels of compositionality[47]: 'fully compositional' constructions, 'pseudo-compositional' constructions, 'fully non-compositional' or 'idiomatic' constructions and 'partly non-compositional' or 'partly-idiomatic' constructions:

- **'Fully compositional' constructions** have a meaning "that is entirely the combinatorial result of the senses of its lexical items and the syntactic structure of the

[47] See also Grant and Bauer (2004) for a discussion of compositionality, figurativeness and idiomaticity.

constituent that contains these lexical items" (Barkema 1996b:159), e.g. *That girl is young*.

- In **'pseudo-compositional' constructions**, only part of their meaning is "the combinatorial result of the senses of all of its lexical items and the syntactic structure of the constituent that contains these lexical items" (160), e.g. *bed and breakfast*, in which 'bed' and 'breakfast' have their basic senses but contribute partly to the meaning of the expression. Other meaning properties such as 'system of accommodation', 'pay', 'room', 'hotel', etc. cannot be inferred from the two independent words. All lexical items of a pseudo-compositional construction are nevertheless interpretable in their basic or derived senses.
- **'Fully non-compositional'** or **'idiomatic' constructions** have a meaning "that is not the combinatorial result of the senses of its lexical items and the syntactic structure of the constituent that contains these lexical items" (ibid 159). They do not have lexical items with basic or derived senses that form part of their meaning, e.g. *a thorn in one's side, put the cat among the pigeons*.
- **'Partly non-compositional'** or **'partly idiomatic' constructions** contain at least one lexical item with a basic sense, e.g. *rain cats and dogs*. Basic or derived senses of only some lexical items in the sequences contribute to their meanings.

As discussed in the following sections, degrees of compositionality have often been used to distinguish between nonce combinations, collocations and idioms (cf. Cowie 1988; Howarth 1996; Nesselhauf 2005). It has also been used to draw a distinction between different types of idioms (e.g. Nunberg et al. 1994; Grant and Bauer 2004) or different types of collocations (e.g. Howarth 1996) (see section 2.4.1.2).

2.3.3.1. Full compositionality of 'ad hoc' or 'nonce' combinations

'Nonce' combinations are word combinations "coined by the speaker/writer on the spur of the moment to cover some immediate need" (Bauer 1983: 45). Thus, they can only be fully compositional constructions: their meaning is quite transparent and easily derivable from their elements, e.g. *open a window, cut bread, black bird*, etc. It does not mean that fully compositional constructions can only be combinations of basic or derived senses. For example, *graveyard* in '*that graveyard of political careers*' (Barkema 1996b: 138) does not have a basic or derived sense but an 'extended' sense, i.e. "a sense which is the unique result of a violation of restriction rules that results in the activation of adaptation rules" (ibid: 159).

An extended sense can only occur in context. Barkema underlines three factors which will play a role in sense adaptation:

1. The basic sense of the lexical item,
2. The speaker's stereotyped beliefs about the referents of this basic sense,
3. The syntactic environment in which it occurs. (159)

Although the listener/ reader must activate adaptation rules to understand the meaning of the lexical item in context, the construction nevertheless remains entirely compositional.

2.3.3.2. Non-compositionality of idioms

The criterion of '(non-)compositionality' has repeatedly been used to separate idioms from other phrasemes (Fraser 1970; Aisenstadt 1979; Cruse 1986; Hausmann 1989; Gramley and Pätzold 1992; Gläser 1998; Mel'čuk 1998; Grossman and Tutin 2002, 2003; Grant and Bauer 2004) and is widely regarded as the most reliable criterion to do so (cf. Skandera 2004: 27). Although phraseologists may have a more or less inclusive definition of the category of idioms or they may name it differently (cf. Mel'čuk's 'full phrasemes'), they all share the view that the distinctive feature of idioms is that their meaning cannot be derived from the meanings of their constituents.

This does not mean that all idioms are completely 'unmotivated' or 'opaque'. Many idioms can be interpreted in figurative, often metaphorical terms: if a listener/reader knows that the literal meaning of '*to make/do a U-turn*' is 'to make a turn in a car, on a bicycle, etc so that you go back in the direction you come from'[48], s/he will then be able to interpret figuratively the expression as 'to make a radical change of ideas, plans, etc' in a sentence such as the following:

> *2.4. The leader of the opposition accused the Prime Minister of* ***doing a U-turn*** *on its promise to increase education spending.* (LDOCE4)

Other examples are *to play second fiddle* ('to play a less important part'), *to keep the ball rolling* ('to continue something, such as a conversation or a plan'), and *to clip somebody's wings* ('to restrict someone's freedom, activities, or power') (cf. Barkema 1996b: 140). It is for this particular reason that Nunberg et al. (1994), Cowie (1998a), Grant and Bauer (2004) make a distinction between idioms whose elements presumably carry identifiable parts of

[48] All definitions in this thesis come from the fourth edition of the Longman Dictionary of Contemporary English (LDOCE4), 2003.

their idiomatic meaning, i.e. Cowie's 'figurative idioms' (e.g. *spill the beans* 'reveal secrets') and idioms whose interpretation cannot be distributed over their parts, i.e. Cowie's 'pure idioms' (e.g. *kick the bucket* 'die').

It is, however, very difficult to establish a boundary between 'figurative' and 'pure' idioms: degrees of 'opaqueness' or 'figurativeness' highly depend on the judgment of individual speakers, a judgment largely based on their linguistic and cultural experience. Moreover, compositionality should not be equated with "the degree to which the phrasal meaning once known, can be analyzed in terms of the contributions of the idiom parts" (Nunberg et al. 1994: 498). It is highly debatable whether *spill the beans* can be interpreted figuratively on the basis of its literal meaning and decomposed into two meaningful elements, i.e. *spill* 'reveal' and *beans* 'secrets'. As rightly stressed by Sinclair (2004 [1996]: 31), "once established, it is dangerously easy to reverse the procedure and assume that the metaphorical extension is obvious." Figurativeness will thus be considered as "a particular mode of interpretation rather than a particular kind of linguistic unit" (Allerton 2004: 98). No distinction will be made between 'figurative' and 'pure' idioms in this thesis. It is however recognized that such a distinction might become necessary in future research, especially in the field of second language acquisition, as it is most probable that figurative and pure idioms are not processed, interpreted and memorized in the same way (cf. Grant and Bauer 2004).

Finally, it should be noted that non-compositionality is also characteristic of other types of phrasemes such as multi-word verbs (*take off, make out*), compounds (*red tape, red herring*), proverbs (*Don't make a mountain out of a molehill; in for a penny, in for a pound*), pragmatic phrasemes or formulae (*How do you do*?), etc.

2.3.3.3. Compositionality in collocations

While nonce combinations are fully compositional and idioms are fully non-compositional, collocations are more difficult to situate on a scale of compositionality. Constructions such as *foot the bill* ('to pay for something, especially something expensive that you do not want to pay for' LDOCE4), Fr. *peur bleue* ('a bad fright') and Fr. *colère noire* ('blind rage') are semantically close to idioms. However, as their base (see 2.4.1.1) is still interpretable as an independent semantic constituent, they are categorized as partly non-compositional (or partly idiomatic) collocations.

By contrast, constructions such as *utter contempt, strong coffee, heavy smoker* and *severe migraine* lie at the opposite end of the scale of compositionality. Their status as collocations or nonce combinations depends on whether the respective senses of the adjectives *utter, heavy* and *severe* are described as derived senses or as **'collocational meanings'** (cf. Firth 1957) in these constructions. Thus, *heavy smoker* will be referred to as a nonce combination if the adjective *heavy* is believed to be used in this construction with a sense that it has in isolation and which could be paraphrased as 'great in amount or degree' or 'very much'. By contrast, the construction will be called a collocation if our conclusion is that the adjective can only express this specific sense in context. A practical approach to this question is to test whether the co-occurrence restrictions imposed on the adjective *heavy* used with the sense of 'very much' are only of a semantic nature or also usage-determined (see section 2.3.5 on the criterion of collocability and section 2.3.5.3 for a discussion of the application of this criterion to the construction *heavy drinker*).

Constructions such as *make a decision, give a look, do somebody a favour* and *take a step* represent a third group of collocations that need to be accounted for by yet another type of compositional analysis. They are hardly classifiable into Barkema's (1996b) 4-level scale of compositionality: the meanings of *make, give, do* and *take* in these collocations are not basic, derived or idiomatic in nature. Rather, they are **delexical** or grammaticalized meanings of these verbs: they serve to 'grammaticalize' their agentive bases (see section 2.4.1.2).

The examples illustrated above stress the **scalar nature** of compositionality in collocations. Collocations can nevertheless be distinguished from idioms on the basis of this criterion as they share the following characteristics:

1. Collocations do not form single non-compositional semantic units;
2. They are not figurative as wholes;
3. One of their elements is used in a basic or derived sense;
4. Their components contribute independently to their overall meanings.

Howarth (1996) compares the unique[49] collocation *foot the bill* with the idiom *fill the bill* ('to be exactly what you need' LDOCE4) in order to illustrate these major differences between collocations and idioms:

> Comparing *foot the bill* with *fill the bill* (which is fully idiomatic), one can see that *bill* in the first refers to an actual bill of payment, while in the second it makes no analysable individual contribution to the overall meaning. Even though the verb *foot* in the sense of 'pay' collocates with no other noun (it contributes to a 'unique collocation'), it can be shown to have independent semantic status. (1996: 38)

[49] 'Unique collocations' or 'bound collocations' (Cruse 1986: 41) are collocations in which one of the elements only occurs in these particular word combinations.

The reader is referred to Koike (2002) and section 2.4.1.2 for more detail on the semantic characteristics of collocations.

2.3.4. Function in the language

Typologies such as those inspired by the Russian tradition (see section 2.2.1) typically distinguish between phrasemes which have a **referential** meaning, i.e. Cowie's composites, and phrasemes which have a **pragmatic** function in discourse, i.e. Cowie's formulae. Formulae are recognized by the members of a language community as preferred ways of performing certain functions, e.g. checking comprehension (e.g. *all right*?), shifting a topic (e.g. *by the way*), refusing (e.g. *no way, of course not*), expressing sympathy (e.g. *I'm (very) sorry to hear, how awful*), calling for brevity (e.g. *Get to the point*!), greeting (e.g. *good morning, how are you*?), etc.

In their pedagogically-oriented analysis of **lexical phrases**, Nattinger and DeCarrico (1992) make a distinction between three types of pragmatically specialized phrasemes. First, social interactional markers function as interaction 'facilitators' and include formulae used to perform speech-act functions such as greeting, thanking, offering, etc. (e.g. *Thank you, You're welcome, I'm sorry*, etc.), as well as formulae "employed in organizing turn-taking, indicating a speaker's attitude to other participants, and generally ensuring the smooth conduct of interaction" (Cowie 1988: 133) (e.g. *you know, I mean*, etc.). The second category includes lexical phrases that "mark topics about which learners are often asked" (Nattinger and DeCarrico 1992:63) and are often used in everyday conversation, e.g. 'autobiography' *my name is ...*; 'food' *I'd like to make a reservation (for ...)*; 'time' *it's ... o'clock.* Third, discourse devices, i.e. textual phrasemes in this thesis, structure the discourse: "their function is to signal, for instance, whether the information to follow is in contrast to, is in addition to, or is an example of, information that has preceded" (ibid: 60).[50] For example, the function of a pattern such as '*it + (modal) + passive verb (of saying/thinking) + that-clause*' (e.g. *it is said/thought that ...; it can be claimed/assumed that ...*) (cf. Granger 1998: 154) is basically to introduce a claim, an idea, a counter-argument, etc., which will help organise the discourse structure around a general pattern such as claim – counter-claim, cause – consequence, problem –solution, etc. (cf. Hoey 1993; Flowerdew L. 1998b; 2003). Other examples are as *a result (of* X*), and then, I think that* X*, as a matter of fact, in other words, not only* X *but also* Y*, as far as I can tell, my point (here) is that,* etc.

[50] Cowie included discourse devices or 'organizational formulae' in his category of 'speech formulae'.

2.3.5. Restricted collocability

> It is the native speaker's experience of his own language that tells him that 'weak tea' is a normal collocation and that 'feeble tea' is not; and consequently that if someone uses the latter combination, then it is for some special effect. Unfortunately for the foreign learner of English, there is no way in which he can be led to 'construct' the collocation 'weak tea' rather than 'feeble tea'. He can learn it only from experience, like the native speaker. (Mackin 1978:150)

Collocability can be defined as "the linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than that of its 'synonyms' because of constraints which are not on the level of syntax or conceptual meaning[51] but on that of usage" (Van Roey 1990: 48).[52] Collocability is generally tested by replacing an element of a phraseme by a synonym, near-synonym or antonym, e.g. *strong* (*powerful) *coffee, heavy* (*fat) *smoker, rancid* (*sour) *butter, to run* (*conduct) *a machine*, etc. This is made explicit in Barkema's (1996b) definition of the term:

> the degree to which it is possible to substitute a lexical item from an open class in a construction with alternatives from the same class: thus a noun is substituted by other nouns, a verb by other verbs, etc. These alternatives can be synonyms, near-synonyms and antonyms. If the collocability of a construction is restricted, this is a matter of arbitrariness, i.e. the restriction is not linguistically motivated. (Barkema 1996b: 145)

Barkema (1996b) distinguishes between collocationally 'open', 'limited' and 'closed' constructions. In 'collocationally open' constructions, all content words can be substituted by any number of alternatives from the same classes, e.g. *day after day* ($X^{\text{time reference}}$ after $X^{\text{time reference}}$) (cf. Lewis's notion of 'frames' in section 2.2.2). In 'collocationally limited' constructions, the number of alternatives is arbitrarily limited "to a few or to only one" (ibid 146), e.g. *strong tea/coffee* but not **powerful tea/coffee*. Finally 'collocationally closed' constructions admit no alternative, e.g. *red tape, by and large*. Barkema further distinguishes between collocationally 'partly' and 'entirely' open, limited or closed constructions. A construction is collocationally 'partly' open, limited or closed if only one part of the construction is collocationally open, limited or closed. An example of a collocationally partly open construction is 'DETERMINER$^{\text{possessive}}$ *cup of tea*'. Similarly, a construction is collocationally 'entirely' open, limited or closed if the whole construction is collocationally

[51] "The conceptual or cognitive meaning of a word is that aspect of lexical meaning which has to do with the language user's knowledge and experience of the extra-linguistic world, of reality." (Van Roey 1990: 24)

[52] Note that Van Roey (1990) uses the term 'collocation' to refer to both the linguistic phenomenon described here and its realizations in discourse. We prefer to use the term 'collocability' when we discuss the phenomenon.

open, limited or closed. An example of a collocationally entirely open construction is '$\text{NOUN}^{\text{singular, quantity}}$ *by* $\text{NOUN}^{\text{singular, quantity}}$'.

The test of substitutability has often been used in order to distinguish collocations from nonce combinations (cf. Aisenstadt 1979; Howarth 1996; Nesselhauf 2005). Testing a phraseme's potential for collocability by means of alternatives such as (near-)synonyms or antonyms can, however, be quite problematic as "problems of use and interpretation intrude and make the procedure less than objective" (Nattinger and DeCarrico 1992: 183). Barkema's intermediary category of 'collocationally limited' constructions presents the most serious problems:

1. How large is the category of 'near-synonyms'?
2. How many is 'a few' alternatives?
3. Where is the limit between collocationally open and limited constructions?

As Nesselhauf (2005: 27) points out, "arbitrary restriction on commutability is used to mean widely different things, and it is also not made clear what exactly is meant."

Collocability can only be fully apprehended if described in relation to other levels of co-occurrence restrictions. Allerton (1984) distinguishes between four levels of co-occurrence restrictions, i.e. syntactic, semantic, locutional and pragmatic co-occurrence restrictions. Van Roey (1990) makes a similar distinction between grammatical, semantic and usage-determined selectional restrictions but he does not describe an independent pragmatic level. Allerton's 'locutional level' and Van Roey's 'usage-determined selectional restrictions' correspond to the notion of collocability as they describe the fact that "language simply seems to dictate, for no good semantic reason that such-and-such combination does, or does not occur" (Allerton 1984: 28). A clear differentiation between **semantic** and **locutional or usage-determined co-occurrence restrictions** should therefore help to describe collocability in nonce combinations, collocations and idioms more accurately. These two types of restrictions are thus explained in the following two sections before examining the role of collocability in nonce combinations, collocations and idioms in section 2.3.5.3.

2.3.5.1. Semantic co-occurrence restrictions

Allerton states that "individual words are only combined when the ideas they express come together at some point in the physical and/or mental experience of the speaker"(1984: 20-21). Certain features of a lexical item's co-occurrence restrictions can be 'logically' predicted from its lexical meaning and semantic traits. Cruse calls these characteristics **'selectional restrictions'**, i.e. "a logically inescapable concomitant of the propositional traits of a lexical item" (1986: 278). Some word combinations are thus simply excluded for semantic reasons: *?The spoon died* and *?Arthur's exam results died* are unacceptable sentences because the verb 'to die' requires an organic, alive and possibly mortal subject (Cruse's examples, p. 278).

Coseriu (1964) introduces the concept of 'lexical solidarity' in an attempt to explain the combinability of lexemes on the basis of their semantic features. He distinguishes between three types of lexical solidarities: affinity, selection and implication. Affinity and selection are quite useful to delineate the scope of semantic co-occurrence restrictions:

1. **'Affinity'** is the lexical solidarity that a lexeme entertains with all lexemes belonging to a lexical class characterized by a very general semantic trait such as [± edible], [± human], [± animate], e.g. the object of *eat* should be edible.
2. **'Selection'** is slightly more restrictive: a given lexeme can be combined with all lexemes which share a similar, though less general, semantic content. Examples of selection are *prune* + the lexical field of 'tree/bush' or *drive* + the lexical field of 'vehicle'.

These examples clearly show that there is no such thing as a 'free combination'. A word combination is first restricted at syntax level (**He are leaving*). Then it needs to fulfil semantic requirements (**He is pregnant*). This explains why the term 'free combination' is not used in this thesis for word combinations such as *eat an apple, drink your tea, Arthur died,* etc. The term **'nonce' combination** is preferred, by analogy with Bauer's definition of a 'nonce' formation, i.e. "a new complex word coined by the speaker/writer on the spur of the moment to cover some immediate need" (1983: 45) as it stresses the fact that words occur together when there is a need for their combination in communication .

2.3.5.2. Usage-determined or locutional co-occurrence restrictions

Semantic co-occurrence restrictions are explainable on the basis of a **semantic class** (cf. Coseriu's 'affinity') or a **lexical field** (cf. Coseriu's 'selection'). Usage-determined or locutional co-occurrence restrictions, on the other hand, cannot be generalized and should be described for every single lexeme. Only **usage** can explain why we say *strong coffee* and not **powerful coffee*, *wide trousers* and not **broad trousers*, or why we prefer to speak of *blond hair* rather than of *yellow hair* or of *a chestnut horse* rather than of *a brown horse*. Phrasemes can show varying degrees of usage-determined co-occurrence restrictions:

1. They can be restricted at the level of the lexical field but with a few exceptions: the verb 'to commit' can combine with almost any noun referring to an illegal action or crime, e.g. *to commit a crime, murder, rape, arson, suicide, adultery, an offence, a sin,* etc, but *to commit delinquency* would seem odd to an English native speaker (cf. Nesselhauf 2005: 30).
2. Lexical items might show arbitrary co-occurrence restrictions among a subset of near synonyms. Van Roey (1990: 48) shows that the specialized sense of 'to regulate or control the affairs of' is expressed by different verbs according to the nouns with which they co-occur, e.g. *to direct an operation, to manage/run a company, to lead a party, to conduct an orchestra.* Another example is *to pay attention/heed* but *to take notice.*
3. The lexical company a word keeps can be quite exclusive: *rancid* is often used with *butter*, the French adjective *diluvien* is typically found with the noun *pluie* and *shrug* can only collocate with *shoulders*. Besides, *shrug* is so closely related to *shoulders* that the verb keeps the same meaning even if it used without the noun. This level of restriction more or less corresponds to Coseriu's (1964:155) notion of 'implication' for which he gave the examples of Fr. *nez aquilin* ('aquiline nose') and *cheval alezan* ('chestnut horse'): the nouns *nez* and *cheval* are implied by the adjectives *aquiline* and *alezan*; their meanings are encompassed by the meaning of the adjectives.

This third category shows that semantic and locutional co-occurrence restrictions are not always easily differentiated. Nesselhauf states that "part of the problem lies in the fact that senses cannot be clearly distinguished either" (2005: 31) (see also section 2.3.3.3). This is the reason why it would be "theoretically possible to make sense distinctions that are so fine that all combinations could be considered semantically motivated" (ibid.). Van Roey (1990) rejects this view in the following terms:

> Some would probably claim that if, in the case of adjectives denoting "gone rotten or bad", *rancid* is preferred to other items in combination with *butter*, this is because there is a specific kind of "rottenness" of *butter*, different from that of *eggs* or *apples* for instance. Here semanticists are on very unsure ground. But surely in many cases no such claim can be made. To take only the case of *blond*: as Palmer puts it (1981, 76), "we should not talk about **a blond door* or **a blond dress*, even if the colour were exactly that of blond hair". (Van Roey 1990:49)

Thus, English native speakers prefer to speak of *rancid butter* rather than of **rotten butter* only on the basis of their knowledge of 'language as a norm'[53] (cf. Corpas Pastor's 1996 category of phrasemes fixed in the norm, i.e. collocations).

2.3.5.3. Collocability in nonce combinations, collocations and idioms

Collocability, as defined by Van Roey (1990), is generally used to set the category of collocations apart from nonce combinations. When a native speaker of English wants to express the fact that somebody smokes or drinks a lot, he usually prefers to say that this person is a *heavy smoker/drinker* rather than a *big smoker/drinker.* By contrast, he will speak of a *big eater* rather than a *heavy eater.* Since no semantic restrictions can be used to explain this difference in preference, these constructions are categorized as collocations in this thesis. Table 2.2 gives another example of the arbitrary nature of collocability.

Table 2.2: Collocability of synonyms (based on Cowie 1994: 3169)

	test	experiment	task	survey
perform	√	√	√	X
carry out	√	√	√	√
conduct	√	√	X	√

Mel'čuk (1995; 1998) attempts to describe these lexical preferences with lexical functions. A lexical function is "a very general and abstract meaning that can be expressed in a large variety of ways depending on the lexical unit to which this meaning applies" (Mel'čuk 1995: 186). Examples of lexical functions are:

- **Magn** which expresses the meaning of 'intense(ly)' or 'very' and functions as an intensifier, e.g. **Magn**(shave$_N$) = *close, clean*; **Magn**(easy) = *as pie, as 1-2-3*; **Magn**(to condemn) = *strongly*
- **Oper** which expresses the meaning of 'do/perform', e.g. **Oper**$_1$(cry) = *to let out* [ART~]
- **Real** which conveys the meaning of 'fulfil the requirement of X' or 'do with X what you are supposed to do with X', e.g. **Real**$_1$(car) = *to drive* [ART~]; **Real**$_1$(accusation) = *to prove* [ART~]

[53] Hausmann expresses the same view in the following terms : « Si ces combinaisons [*regarder un arbre, les nudités noueuses des arbres*] appartiennent à la langue en tant que système, les vraies collocations relèvent plutôt de la langue en tant que norme (ou comme dit Bally après Vaugelas, de l'usage tyrannique). » (Hausmann 1979: 191)

Not all lexical functions describe syntagmatic relations however. Lexical functions also describe paradigmatic relations, e.g. the lexical function S 'the one who/which undergoes' as in S_2*(hotel) = guest*, S_2*(prison) = prisoner*, etc.

Unlike collocations, nonce combinations are entirely explainable by grammatical and semantic rules: they do not display usage-determined co-occurrence restrictions. Nesselhauf (2005: 33) uses the criterion of commutability or collocability in order to distinguish between collocations and idioms. She claims that a verb-noun combination is a collocation if the noun is semantically autonomous and only the verb shows restricted commutability but it is an idiom if both the noun and the verb show restricted commutability. I would argue, however, that the relationship between the elements of an idiom cannot be described as a lexical preference or as a usage-determined constraint since there is no selection of a lexical item by another one in these phrasemes. The fact that their elements cannot be easily substituted is due to their non-compositionality. The term 'restricted collocability' will therefore not be used to describe idioms.

2.3.6. Degrees of flexibility

> "The myth of fixedness is perpetuated." (Sinclair 2004 [1996]: 30)

Following Barkema (1996a; 1996b), the term 'flexibility' relates to the extent to which word combinations allow for the whole range of syntactic variation that is normally possible without losing their phraseological status (e.g. passivisation, pluralisation, negation, insertion, deletion, pronominalisation). The terms 'fixedness' and 'frozenness' are avoided in this thesis as scholars have often used them as generic terms to refer to what we consider to be two different characteristics of phrasemes: syntactic flexibility and collocability (cf. Zuluaga 1980; Fernando 1996; Moon 1998a).

Barkema analyses the syntactic flexibility of word combinations and distinguishes between 'fully flexible constructions', 'semi-flexible constructions' and 'inflexible constructions' on the basis of four criteria (1996a: 71):

- Addition: the introduction of an additional syntactic function such as a modifying adverb, e.g. *an appallingly tough nut to crack*,
- Term selection: the substitution of one of the lexical items by its plural or comparative form, e.g. *harder nuts to crack*,

- Permutation: the permutation of two elements of the construction, e.g. *a nut too tough to crack the first time out,*
- Interruption: the introduction of elements, e.g. *a hard nut, as always, to crack.*

Thus, 'fully flexible constructions', 'semi-flexible constructions' and 'inflexible constructions' can take all variations, a limited number of variations, and no variations that are grammatically possible respectively.

Flexibility, together with collocability, has often been used to delineate the class of idioms. Fernando defines idioms as "indivisible units whose components cannot be varied or varied only within definable limits" (1996: 30), i.e. Barkema's inflexible or semi-flexible constructions. He further claims that "all idioms are not grammatically regular." Degrees of flexibility have also been used to distinguish between different types of idioms. For example, Fraser (1970) proposes a 'frozenness hierarchy' of idioms based on their transformational deficiencies (permutation, insertion, etc.). Similarly, Nunberg et al. (1994) use the criterion of flexibility in order to distinguish between two semantically distinct types of idioms: semantically analysable[54] 'idiomatically combining expressions' such as *spill the beans* or *touch a nerve*, and opaque 'idiomatic phrases' such as *kick the bucket* (see section 2.3.3). They argue that the idiomatically combining expressions' potential for modification, quantification, topicalization, ellipsis, etc. offers strong evidence that they are semantically compositional, i.e. that their parts carry meaning:

- Modification by means of adjectives or relative clauses: *leave no legal stone unturned, Your remark touched a nerve that I didn't even know existed* (Nunberg et al. 1994: 500)
- Quantification: *touch a couple of nerves, We could ... pull yet more strings* (ibid 501)
- Topicalization: *Those strings, he wouldn't pull for you; Those windmills, not even he would tilt at; His closets, you might find skeletons in.*
- Ellipsis: *My goose is cooked, but yours isn't.*

Some of Nunberg et al.'s (1994) examples might be more typical of playful creativity in spoken language. It is nevertheless true that different degrees of transformational deficiency can be acknowledged within the category of idioms (Carter 1998: 70-72; Fernando 1996: 32), from the almost invariant *kick the bucket* to the more flexible *break somebody's heart.*

[54] As explained by Poulsen (2005:91-99), 'analysability' is an independent parameter that refers to the extent to which speakers are aware of the contribution that individual elements make to the whole construction while 'compositionality' refers to the degree to which the meaning of the whole was predictable from the meanings of the individual parts when the phrase was originally coined.

Degrees of transformational deficiency are also noticeable in other types of phrasemes such as proverbs (**Birds of a feather flocked together*), routine formulae (**How did you do?* to be contrasted with the many instantiations of the productive frame 'wh-word AUX PRO VERB', e.g. *where on earth have you been? Why on earth didn't you ask me*), irreversible bi- and trinomials (**hook, lines and sinker*) and complex adverbs (**for examples*). Although various degrees of flexibility can be found in collocations (Grossman and Tutin 2002; Fernando 1996: 70-72; Schenk 1995), from the more restricted *heavy smoker* (**the smoker is heavy*) to the less restricted *catch a bus*, semi-flexible or inflexible collocations are the exception rather than the rule. Most collocations (e.g. *carry out an experiment*) are fully flexible constructions as they allow for addition (*carry out a controlled experiment*), inflection (*carry out experiments*), permutation (*an experiment was carried out*) and interruption (*they carried out, as always, funny experiments*) (see section 1.4 for more detail on collocations).

Definitions of phrasemes based on the criterion of flexibility have been criticized as early as 1909 by Charles Bally. However, it is only with the development of corpora that linguists such as Moon (1998a) were able to "show up clearly the fallacy of the notion of fixedness of form" (ibid 47). Following Svensson (2002), syntactic inflexibility will not be retained as a necessary defining criterion for phrasemes. Lack of flexibility is, however, an 'indication' of phraseological status as is 'marked syntax', i.e. the breaking of conventional grammatical rules in word combinations. Moon (1998a: 81-83) gives numerous examples of 'ill-formedness' in phrasemes, which she categorizes under the following headings:

- Odd phrase structures, ellipsis, inflections and archaic mood: *at all, be that as it may, every which way, let alone, go for broke, etc.*
- Strange uses of word classes, especially non-nominal words used as nouns and adjectives as adverbs: *all of a sudden, at the ready, do the dirty on so., for free, ifs and buts, in general, once in a while, on the up and up, play fair, etc.*
- Component words which deviate from their usual syntactic behaviour, especially countable nouns used without determiners in the singular or verbs "used in aberrant transitivity patterns" (82): *in all weathers, put pen to paper, rain cats and dogs, sweat blood, under lock and key, in case, etc.*

2.3.7. Conclusion and definitions adopted

In an attempt to circumscribe the field of phraseology more precisely, a review of defining criteria has been conducted. On this basis, we can now propose a definition for phraseology. In this thesis, phraseology is defined as **the study of all syntagmatic relations between at least two lemmas, contiguous or not, written separately or not, which are typically syntactically closely related constituents and constitute "'preferred' ways of saying things" (Altenberg 1998:122) for the language user since:**

- **They form a functional (referential, textual or communicative) unit** (cf. section 2.3.4); **and**
- **They display arbitrary lexical restrictions** (cf. section 2.3.5); **and/ or**
- **They are characterized by a certain degree of semantic non-compositionality** (cf. section 2.3.3); **and/or**
- **They display arbitrary restrictions on the word forms that can be used to instantiate at least one of the lemmas involved** (cf. section 2.3.1); **and/or**
- **They display a certain degree of syntactic fixity** (cf. 2.3.6).

In a number of studies, Barkema examined the phenomena of institutionalization, lexicalization, compositionality, collocability and flexibility and reached the conclusion that there was a **general lack of correlation** among these defining criteria: word combinations which share, for example, similarities in terms of their degrees of compositionality may be far apart in terms of their flexibility (see, e.g. Barkema 1996b). This observation led him to avoid defining general categories such as nonce combinations, collocations and idioms and describe every single word combination on a multi-dimensional scale, e.g. the phraseme '*bed and breakfast*':

> Lexicalised expression:
> - pseudo-compositional;
> - inflexible;
> - collocationally partly limited (because of *breakfast/board*), partly closed (because of *bed*). (Barkema, 1996b: 152)

Barkema's description is certainly the most faithful to the multifaceted character of word combinations. It is however extremely difficult to use for practical purposes as it makes it impossible to classify results in broad categories of combinations sharing a set of common features. As one of the major objectives of this thesis is to describe and characterise EFL learners' use of phrasemes that serve discourse functions, it seems essential to classify phrasemes in broad categories. Similarly, in a second language acquisition (SLA) perspective, it seems particularly important to distinguish between different categories of phrasemes as it is most probable that different types of word combinations cause different kinds of difficulties to learners of English. We shall therefore propose definitions of the major types of phrasemes on the basis of the defining criteria discussed above and, as illustrated in Figure 2.10 below, classify them into the three categories of the functional model first presented in section 2.2.4, i.e. referential phrasemes, textual phrasemes and communicative phrasemes.

I) Referential phrasemes

Referential phrasemes were broadly defined in section 2.2.4 as phrasemes used to refer to objects, phenomena or facts of life. They include various types of phrasemes such as collocations, idioms, similes, irreversible bi- and trinomials, compounds, phrasal verbs and grammatical collocations. Collocations and idioms are often described on a cline from free combinations to idioms. Table 2.3 summarizes the major features of nonce combinations (NC), collocations (C) and idioms (I).

Table 2.3: Contrasting nonce combinations, collocations and idioms

	NC	C	I
Institutionalization	-	+	+
Lexicalization	-	Rather -	+
Compositionality	+	Rather +	-
Restricted collocability	-	+	-
Flexibility	+	Rather +	Rather -

This table clearly shows that institutionalization and restricted collocability are the only criteria which allow us to separate collocations from nonce combinations. By contrast, collocations and idioms are distinguishable on the basis of non-compositionality of idioms and restricted collocability of collocations. The fact that the elements of idioms cannot be easily substituted is a direct consequence of their non-compositionality. It is not referred to as 'restricted collocability' as the phenomenon is not usage-determined (see 2.3.5). We therefore propose the following definitions:

- **'Ad hoc'** or **'nonce' combinations** are not subject to idiosyncratic restrictions: they are semantically fully compositional, syntactically fully flexible and collocationally open. They are only governed by grammatical rules and semantic constraints. As a result, they do not belong to the phraseological spectrum.

- **(Lexical) collocations** are usage-determined or preferred syntagmatic relations between two lexemes in a specific syntactic pattern. Both lexemes make an isolable semantic contribution to the word combination but they do not have the same status: one of the two lexical items is arbitrarily selected and semantically determined by the other (see section 2.4 for more detail on collocations and a second version of our definition of collocations).

- The category of **idioms** is restricted to phrasemes that are constructed around a verbal nucleus and which are characterized by their semantic non-compositionality. Their semantic non-compositionality can be the result of a metaphorical process. Lack of flexibility and marked syntax are further indications of their idiomatic status. Examples include *to spill the beans, to let the cat out of the bag, to bark up the wrong tree* and *to kick the bucket.*[55]

The other types of referential phrasemes are irreversible bi- and trinomials, similes, compounds and phrasal verbs:

- **Irreversible bi- and trinomials** are fixed sequences of two or three word forms that belong to the same part-of-speech category and are linked by the conjunction 'and' or 'or', e.g. *bed and breakfast, kith and kin, cash and carry, sooner or later* (see Alexander 1984 and Gramley and Pätzold 1992).

- **Similes** are sequences of words that function as stereotyped comparisons. They typically consist of sequences following the frames '*as* ADJ *as* (DET) NOUN' (e.g. *as busy as a bee, as old as the hills, as easy as pie*) and 'VERB like a NOUN' (e.g. *to swear like a trooper, to sell like hot cakes, to eat like a bird*).

- **Compounds** are morphologically made up of two elements which have independent status outside these word combinations. They can be written separately (*black hole, easy going*), with a hyphen (*brother-in-law, law-abiding*) or as one orthographic word (*schoolmaster, bittersweet*). They resemble single words in that they carry meaning as a whole and are characterized by high degrees of inflexibility, i.e. set order and non-interruptibility of their parts (cf. Cruse 1986; Bussmann 1996:91).

- **Grammatical collocations** are restricted combinations of a lexical and a grammatical word, typically verb/noun/adjective + preposition, e.g. *depend on, cope with, a contribution to,*

[55] Other structural types of phrasemes that present different degrees of idiomaticity are classified under categories such as compounds (e.g. *red tape, black hole*), idiomatic sentences (e.g. *the die is cast!*) and formulae (e.g. *of course*).

afraid of, angry at, interested in. The term 'grammatical collocation' is borrowed from Benson et al. (1997) but our definition is slightly more restricted as these authors also use the term to refer to other valency patterns, e.g. avoid + *-ing* form (see section 2.4.1.2. for more detail on grammatical collocations), which we do not consider to be part of the phraseological spectrum.

- **Phrasal verbs** are combinations of verbs and adverbial or prepositional particles. Examples include *blow up, make out, crop up* and *hand in.*

II) Textual phrasemes

Textual phrasemes are typically used to structure and organize the content (i.e. referential information) of a text or any type of discourse. They include grammaticalized sequences such as complex prepositions and complex conjunctions, as well as similar sequences that are undergoing a process of grammaticalization (cf. section 2.3.2), linking adverbials and sentence stems.

- **Complex prepositions** typically consist of "two simple prepositions with an intervening nominal element" (Hoffmann 2004:195), e.g. *with respect to, in addition to*. Other structural patterns are, for example, adverb + preposition (e.g. *apart from, ahead of, instead of*) or adjective + preposition (e.g. *irrespective of, prior to*). They are grammatical elements whose main task "lies in the structuring of text on an informational level of organization" (ibid). Fully grammaticalized complex prepositions have a non-compositional meaning and are completely inflexible: they resist interruption of their parts (e.g. *in view of*). However, degrees of collocability and grammaticalization can also be found in complex prepositions (e.g. Sp. *con ganas/deseo(s) de* or *con vistas/miras a* 'in view of'). In fact, it may be as difficult to draw a line between complex prepositions and nonce combinations that follow the structure 'Prep + Noun + Prep', as to separate collocations from lexical nonce combinations (cf. Hoffmann 2004; Montoro del Arco 2006).

- **Complex conjunctions** are inflexible sequences such as *so that, as if, even though, rather than, as soon as, except that* and *given that*. Different degrees of collocability characterize complex conjunctions, from the fully grammaticalized *even though* to more flexible patterns such as *as long/soon as* or *whether or not*. The reader is referred to Gross (1996) for more information on complex conjunctions in French (Fr. 'locutions conjunctives') and Montoro del Arco (1996) for more detail on complex conjunctions in Spanish.

- **Linking adverbials** include various types of phrasemes such as grammaticalized prepositional phrases (*in other words, by the way*), adjectival phrases (*last but not least*), adverbial phrases (*more accurately*), finite clauses (*that is to say; what is more*), non-finite clauses (*to conclude, to summarize*). Linking adverbials have a conjunctive role, that is, to "tell us how the speaker or writer understands the semantic connection between two utterances, or parts of utterances" (Downing and Locke 2002:63). Table 2.4 gives examples of linking adverbials for each category of Quirk et al's (1972:661-677) classification of conjuncts according to the function(s) they serve in clause and sentence connection.

Table 2.4: A classification of linking adverbials based on Quirk et al. (1972)

Listing: enumeration	*in the first place; first of all; on the one hand ... on the other hand; to begin with; to conclude*
Listing: addition	*in the same way; in addition; what is more*
Summative	*in conclusion; in sum; to conclude; to sum up; to summarize*
Apposition	*in other words; for example; for instance: that is*
Result	*as a consequence; in consequence; as a result*
Inferential	*in other words; in that case*
Contrastive: reformulatory	*more accurately; more precisely; in other words*
Contrastive: replacive	*on the other hand*
Contrastive: antithetic	*on the contrary; in contrast; by contrast; by comparison; on the other hand*
Contrastive: concessive	*in any case; in spite of that; at any rate; all the same*
Transitional: discoursal	*by the way*
Transitional: temporal	*in the meantime; in the meanwhile*

- **Textual sentence stems** are routinized fragments of sentences which are used to serve specific textual or organizational functions. They consist of sequences of two or more clause constituents, and typically involve a subject and a verb, e.g. *the final point is ...; another thing is ...; it will be shown that ....; I will discuss ....* They typically have an empty slot for the following object or complement and "form the springboard of utterances leading up to the communicatively most important – and lexically most variable – element" (Altenberg 1998:113).

III) Communicative phrasemes

Communicative phrasemes were broadly defined above as phrasemes that are used to express feelings or beliefs towards a propositional content or to explicitly address interlocutors, either to focus their attention, include them as discourse participants or influence them. They include routine formulae, attitudinal formulae, proverbs, commonplaces, slogans, etc.

- **Routine formulae** or **speech act formulae** (cf. De Cock 2003) are relatively inflexible phrasemes which are recognized by the members of a language community as preferred ways of performing certain functions such as greetings, compliments, invitations, etc. (e.g. *Good morning!, see you soon!, take care!, nice to meet you!*). They display different degrees of compositionality, from semantically fully compositional routine formulae such as *happy birthday* and *Merry Christmas* to non-compositional phrasemes such as *how do you do*?

- **Attitudinal formulae** are phrasemes used to signal speakers' attitudes towards their utterances and interlocutors. They can take the form of prepositional phrases (*in fact*), finite (*don't take this personally*) and non-finite clauses (*to be honest; to tell the truth; strictly speaking*). As such, they function as disjuncts to "represent a comment by a speaker or writer on the content of the clause as a whole" (Downing and Locke 2002:62). They can also be realized by sentence stems such as *it is necessary that* or *I think that.* The category of attitudinal formulae includes a large proportion of what Cowie (1998) referred to as speech formulae (see Cowie's typology in section 2.2.1).

- **Proverbs, commonplaces** and **idiomatic sentences** are free utterances or self-contained statements. Unlike Burger (1998), proverbs are classified as communicative phrasemes as "what seems to be important is not so much their meaning, but the ability of a speaker to use a proverbial formula at the right moment" (Hamm 2004: 75). Gramley and Pätzold (1992:77) define commonplaces and proverbs in the following terms:

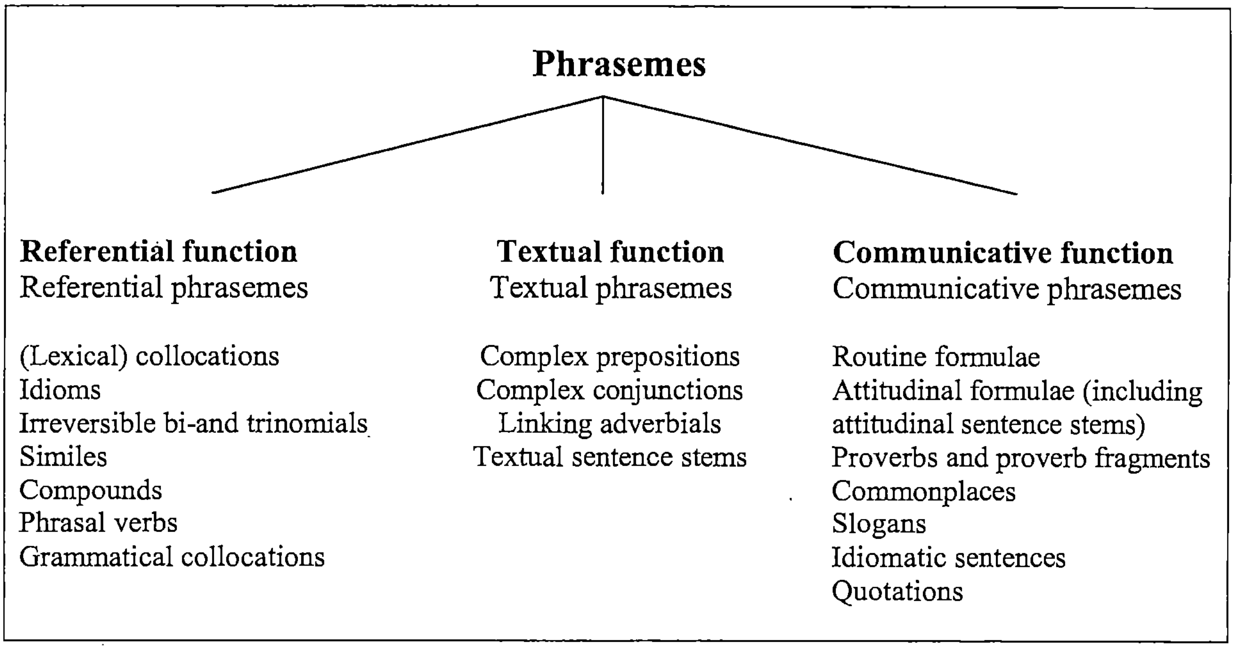
> Commonplaces are complete sentences, fall into three classes, i.e. tautologies (*Enough is enough, orders are orders*), truisms (*We only live once*) and sayings based on everyday experience (*Accidents happen; You never know; It's a small world*), claim universal validity and are non-metaphorical. Proverbs are traditional, express general ideas and show non-literal meaning (metaphorical, metonymic); they can be added to, transformed and abbreviated. Proverbs are equivalent to a sentence and are also prototypically characterized by certain metrical, structural and prosodic features. (Gramley and Pätzold 1992: 77)

Unlike commonplaces, idiomatic sentences have a non-literal meaning, e.g. *the die is cast*! Moreover, unlike proverbs, they cannot be modified.

A functional classification of phrasemes such as the one proposed here and illustrated in Figure 2.10 gives further indication that there is no one-to-one relationship between form and function (cf. De Cock 2003). Communicative phrasemes are mainly clause or sentence-like phrasemes but proverb fragments can also take the form of phrases (e.g. *the early bird*). Referential phrasemes include polylexemic (e.g. collocations, idioms, similes, irreversible bi-

and trinomials) as well as mono-lexemic phrasemes (compounds and phrasal verbs). Textual phrasemes can take the form of mono-lexemic phrasemes (e.g. complex prepositions and conjunctions) as well as of sentence stems which function at the level of the clause.

Figure 2.10: The phraseological spectrum



Phrasemes

Referential function Referential phrasemes	**Textual function** Textual phrasemes	**Communicative function** Communicative phrasemes
(Lexical) collocations Idioms Irreversible bi-and trinomials Similes Compounds Phrasal verbs Grammatical collocations	Complex prepositions Complex conjunctions Linking adverbials Textual sentence stems	Routine formulae Attitudinal formulae (including attitudinal sentence stems) Proverbs and proverb fragments Commonplaces Slogans Idiomatic sentences Quotations

2.4. *Zooming in on collocations*

> "Collocations make up the lion's share of the phraseme inventory,
> and thus deserve our special attention." (Mel'čuk 1998:24)

The review of the defining criteria of phrasemes proposed above has already hinted at the **'in-between' nature of collocations**: they are often described as word combinations situated on a continuum between nonce combinations and idioms. They tend to be compositional in meaning and syntactically flexible, two features that they share with nonce combinations. On the other hand, they are characterized by their limited collocability and as such, belong to 'language as norm'. Most linguists of the last decade consider that collocations constitute an essential part of phrasemes, e.g. Howarth (1996), Corpas Pastor (1996), Cowie (1998), Mel'čuk (1998), González Rey (2002) and Allerton et al. (2004) to the point of saying that "collocations make up the lion's share of the phraseme inventory" (Mel'čuk 1998). However, their transitional status has made some phraseologists take the view that collocations are not part of the phraseological spectrum, e.g. Casares (1992 [1950]), Zuluaga (1980), Alexander (1989), Ruiz Gurillo (1997), Bosque (2001), etc.

These divergent positions as regard the status of collocations in the phraseological spectrum have made it impossible for linguists adopting the 'phraseological approach' to collocations (see 2.4.1) to agree on a generally accepted definition of the term. Similarly, collocations do not seem to have a proper place in linguistic theories[56]. In addition, the term 'collocation' is not only used to refer to a group of phrasemes that share some well-defined linguistic characteristics. As already stressed by Nesselhauf (2004:1), 'collocation' is "used by researchers in many different fields, and the definition is usually adapted to the different aims and methods of their investigations." Another influential approach to collocations which developed alongside the phraseological approach is the 'distributional' (cf. Evert 2004) or 'frequency-based' (cf. Nesselhauf 2004) approach, which goes back to Firth and usually describes collocations as a probabilistic or statistical phenomenon. It was adopted and refined by researchers such as J. McH. Sinclair and G. Kjellmer who were primarily concerned with the computational analysis of collocations for lexicological or lexicographical purposes.

This section reviews the many definitions of collocations proposed within the phraseological and distributional approaches, describes several classifications of collocations provided by researchers whose studies are clearly anchored in the phraseological approach to collocations and assess the major contributions of the distributional approach. Finally, it discusses how and where the two approaches meet and argues that "to continue to thrive" (Nesselhauf 2004:20), the two approaches will need to agree on a common terminology.

2.4.1. The phraseological approach

The approach to collocations discussed in this section has strongly been influenced by the classical Russian phraseological theory and is therefore often referred to as the "phraseological approach" (cf. Nesselhauf 2004). Representative members of this first line of development are Cowie, Howarth, Mel'čuk, Benson and Hausmann. They all regard collocations as a sub-set of phrasemes which they situate on a continuum between nonce combinations and idioms (cf. section 2.2.1) but differ on the criteria used to define and organize collocations and on where to place the boundaries of the category.

[56] The reader is referred to Bartsch (2004: 40-50) for an overview of the place of collocations in Transformational Generative Grammar, Dependency grammar and Cognitive grammar.

2.4.1.1. Definitions of 'collocation'

> « [D]e par sa nature, la collocation demeure un concept difficilement formalisé, aucune définition ne satisfait tout le monde. » (Williams 2001)

Table 2.5 presents 10 representative definitions of the concept of 'collocation' within the phraseological approach. These definitions share a number of features. First, collocations are combinations of words that function below the level of the sentence. Second, there is a syntactic relationship between the elements, which is either used as a defining criterion for collocations (cf. Hausmann 1989, definition n°2; Van der Meer 1998, definition n°10) or, more frequently, left implicit. A large proportion of linguists have made use of the criteria of compositionality (cf. 2.3.3) and restricted collocability (cf. 2.3.5) to define collocations. They have often used restricted collocability to separate collocations from free combinations and compositionality to distinguish between collocations and idioms[57]. By contrast, they differ in their understanding of other characteristics of collocations and in the importance they attach to each of these.

While Hausmann (1989), for example, defines collocation as a relationship between two words, other linguists such as Cowie (1994) and Van der Meer (1998) include combinations of more than two words in their definitions (cf. definitions n°4 and n°10). More recently, Tutin and Grossman (2002) re-affirmed the **binary** nature of collocations but described it as a relationship between a word and another element, which can be a word or a phrase, e.g. *un brouillard <u>à couper au couteau</u>* (cf. definition n°9). The nature of the elements is also a topic of debate but most linguists today share the view that a collocation is a relationship between lexemes (cf. 2.3.1).

Phraseologists also hold different views on the relationship between the elements of a collocation. Hausmann (1989), for example, considers that the relation between the two elements of a collocation is oriented or hierarchically ordered, a property that he calls **directionality**. He argues that the two elements of a collocation do not have the same status. Semantically autonomous, the 'base' of a collocation is selected first by a language user for its independent meaning. The second element, i.e. the 'collocator'[58], is selected by and semantically dependent on the 'base'.

[57] One notable exception is Nesselhauf (2005) who uses the criterion of commutability or collocability both to distinguish between free combinations and collocations and between collocations and idioms (see section 2.4.1.2).

[58] Hausmann uses the term 'collocatif' and not 'collocat' in French. We therefore follow Nesselhauf (2005: 17) and translate the term 'collocatif' into 'collocator' and not 'collocate'.

> En effet, dans la collocation *célibataire endurci*, le signifié de la base (*célibataire*) est autonome. La base n'a pas besoin du collocatif (*endurci*) pour être clairement définie. Il en va tout autrement pour le collocatif qui ne réalise pleinement son signifié qu'en combinaison avec une base (*célibataire, pécheur, âme,* etc.). La base complète la définition du collocatif, alors que le collocatif se contente d'ajouter une qualité à une base en elle-même suffisamment définie. (Hausmann, 1979 : 191-192)

Hausmann (1979) also questions the nature of the **base** and concludes that "la partie du discours qui est la plus près du monde des choses et des êtres est sans contexte le substantif. C'est autour des données exprimées par les substantifs que le locuteur formule sa pensée" (ibid : 192). Thus, in collocations such as verb + noun or adjective + noun, the noun is the 'base' and the other element is the 'collocator'. Unlike Hausmann, Cruse (1986) and Van der Meer (1998), for example, do not regard collocations as an oriented relationship between two elements.

The type of elements involved in the combination is also at the centre of a controversy: linguists such as Hausmann (1989) and Cowie (1994) consider that collocations are necessarily combinations of **lexical** items while Benson et al. (1997) include combinations of lexical and **function** words, e.g. *differ about/on* (see section 2.4.1.2). Combinations of lexical and functional words are regarded by many linguists as not belonging to the phraseological spectrum. Heid (2002), for example, insists that syntactic and collocational properties of words should be clearly distinguished.

Another point of contention is the **formulaic nature** of collocations. Van der Meer (1998) describes collocations as 'conventional building blocks'. However, as stressed by Wray (2002:51), "it is far from being the case that commentators see collocational associations as 'formulaic' in any useful sense at all." Schmid, for example, argues that "it is the hallmark of collocations (as opposed to idioms) that they are not fully entrenched as linguistic units or *gestalts* but only partially" (2003:253). More psycholinguistic experiments such as those presented in Schmitt (2004) are clearly needed to test the way collocations are stored and retrieved by both native and foreign language users.[59]

[59] Most psycholinguistic studies so far have focused on rather fixed sequences of words such as communicative phrasemes (cf. Wray 2002, Davis and Lunsford 2005; Maclagan et al. 2005).

Table 2.5: Definitions of 'collocation' within the phraseological approach

1	Cruse (1986:40) "The term collocation will be used to refer to sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent. (...) the constituent elements are, to varying degrees, mutually selective. The semantic integrity or cohesion of a collocation is the more marked if the meaning carried by one (or more) of its constituent elements is highly restricted contextually, and different from its meaning in more neutral contexts."
2	Hausmann (1989:1010) « On appellera collocation la combinaison caractéristique de deux mots dans une des structures suivantes : a) substantif + adjectif (épithète) : *confirmed bachelor, célibataire endurci* b) substantif + verbe : *his anger falls, la colère s'apaise* c) verbe + substantif (objet) : *to withdraw money, retirer de l'argent* *d)* verbe + adverbe : *rain heavily, pleuvoir à verse* e) adjectif + adverbe : *seriously injured, grièvement blessé* f) substantif + (prép.) + substantif : *a gust of anger, une bouffée de colère* La collocation se distingue de la combinaison libre (...) par la combinabilité restreinte (ou affinité) des mot combinés (...). La collocation se distingue d'autre part des locutions (...) par son non-figement et par sa transparence. »
3	Van Roey (1990:48) "Collocation is the term we shall now use both for the linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than that of its "synonyms" because of constraints which are not on the level of syntax or conceptual meaning but on that of usage, as also for the word combinations which represent this phenomenon. Collocational meaning then is the meaning of a word as far as it is determined by the lexical company it keeps."
4	Gramley and Pätzold (1992:61) "[The term collocation] refers to combinations of two lexical items which make an isolable semantic contribution, belong to different word classes and show restricted range."
5	Cowie (1994:3169) "Collocations are associations of two or more lexemes (or roots) recognized in and defined by their occurrence in a specific range of grammatical constructions. HEAV- + RAIN is one such abstract composite, realized in the patterns *heavy rain* and *rain heavily* (Mitchell 1971). Though transparent and (usually) lexically variable (cf. *light rain, a heavy shower*), they are characterized by arbitrary limitation of choice at one or more points, as in *light exercise / heavy exercise* (Cowie 1981)."
6	Howarth (1996:47) "Combinations in which one component is used in its literal meaning, while the other is used in a specialized sense. The specialized meaning of one element can be figurative, delexical or in some way technical and is an important determinant of limited collocability at the other. These combinations are, however, fully motivated."
7	Corpas Pastor (1996:66)[60] "... we will call 'collocations' (...) phraseological units which consist of two lexical items in a syntactic relationship. Collocations are not autonomous speech acts or utterances. They are fixed in the norm and display restricted collocability due to repeated use. These restrictions are generally semantic in nature: the base not only selects its collocate but it also selects a specialized meaning of its collocate, often an abstract or figurative meaning."

[60] ... denominaremos colocación (....) a las unidades fraseológicas formadas por dos unidades léxicas en relación sintáctica, que no constituyen por sí mismas, actos de habla ni enunicados; y que, debido a su fijación en la norma, presentan restricciones de combinación establecidas por el uso, generalmente de base semántica: el colocado autónomo semánticamente (la base) no sólo determina la elección del colorativo, sino que, además, selección en éste una acepción especial, frecuentemente de carácter abstracto o figurativo.

8	Mel'čuk (1998:30) “A collocation **AB** of language **L** is a semantic phraseme of **L** such that its signified ‘X’ is constructed out of the signified of one of its two constituent lexemes – say, of **A** – and a signified ‘C’ [‘X’ = ‘A ⊕C’] such that the lexeme **B** expresses ‘C’ only contingent on **A**.”
9	Tutin and Grossmann (2002) « Une **collocation** est l'association d'une **lexie (mot simple ou phrasème) L** et d'un **constituant C** (généralement une lexie, mais parfois un syntagme par exemple *couper au couteau* dans *un brouillard à couper au couteau*) entretenant une relation syntaxique telle que : - C (le collocatif) est sélectionné en production pour exprimer un sens donné en cooccurrence avec L (la base). - Le sens de L est habituel. »
10	Van der Meer (1998:315) “The *prototypical* collocation could tentatively be defined as a combination of two or more lexical units: 1. with meanings also occurring independently elsewhere (in other combinations) 2. which are used *non-metaphorically*, 3. which combination occurs repeatedly and normally in a language (cf Carter 1987:47), as *conventional building block*, 4. which the language user has available as a whole, to express *conventional established concepts*, 5. whose constituent words are typically in a *grammatical modifier – modified relation* (including that of verb-object), 6. whose constituent words (in spite of point 1) naturally select each other because the sense definition of the modifier includes the modified (and sometimes vice versa) in a non-banal way (*semantic motivation*), 7. which typically function as *part of a larger group* and not as a complete utterance (sentence) itself.”

The definitions presented in Table 2.5 also differ in being more or less inclusive. The definitions provided by Cruse (1986) and Van Roey (1990) are relatively broad, while at the other end of the scale, Van der Meer's (1998) definition is very restrictive. It excludes, for example, most collocations in which one lexical item is used in its literal meaning while the other is used in a figurative, technical or delexical sense (cf. Howarth 1996; definition n°6), e.g. *pay attention, obtain a warrant, have access to.* Typical collocations according to Van der Meer's (1998) definition are *hazardous waste* and *perpetrate a crime.*

2.4.1.2. Classifications of collocations

As illustrated above, phraseologists have proposed definitions of ‘collocation’ which vary according to the criteria on which they choose to place emphasis. They have also put forward a range of classifications of collocations which fall into three categories: (1) classifications based on their syntactic patterns, (2) classifications based on their semantic features and (3) classifications based on their collocability.

2.4.1.2.1. Syntactic classifications

Hausmann (1989:1010) uses a syntactic criterion to define collocations (cf. Hausmann's definition, Table 2.5) and only regards as collocations those combinations that appear in six syntactic patterns made of two lexical items, namely noun + adjective (*confirmed bachelor*), noun + verb (*his anger falls*), verb + adverb (*rain heavily*), adjective + adverb (*seriously injured*) and noun + (*preposition*) + noun (*a gust of anger*). As shown in Table 2.6, Benson et al. (1997) refer to these types of collocations as **lexical collocations** and distinguish them from **grammatical collocations** which not only include combinations of a content word and a preposition (*angry at*) or a grammatical structure (*necessary* + *to*-INF) (cf. section 2.3.1.) but also verb patterns (G8 collocations).

Table 2.6: Benson et al's (1997) syntactic classification

Grammatical collocations	
G1 : noun + prep.	*blockade against*
G2 : noun + *to*-inf.	*a compulsion to* do it
G3 : noun + *that*-clause	*surprise that*
G4 : prep. + noun	*in advance, under sb's aegis*
G5: adjective + prep.	*angry at*
G6: adjective + *to*-inf.	*necessary to* work
G7: adjective + *that*-clause	*afraid that*
G8 collocations consist of 19 English verb patterns	SVO *for* O, SVOO, SVOC
Lexical collocations	
L1 : verb + noun	*make an impression*
L2 : verb meaning essentially *eradication* and/or *nullification* + noun	*reject an appeal*
L3 : adjective + noun	*strong tea*
L4: noun + verb	*bombs explode*
L5 indicate the unit that is associated with the noun	*a pride of lions*
L6: adverb + adjective	*sound asleep*
L7: verb + adverb	*affect deeply*

Unlike combinations of a content word and a grammatical word, verb patterns clearly do not fall into the boundaries of phraseology: if the base is left unspecified and the relationship is established between grammatical categories (e.g. ADJ + NOUN), the construction is not the object of study of phraseology but syntax. Even broad definitions of the field, or its object of study, such as the one proposed by Gries (to appear a), require at least the presence of one specific lexical item:

> As to the first criterion [*the nature of the elements*], the definition of a phraseologism I will adopt is among the broadest conceivable ones. I consider a phraseologism to be the co-occurrence of a form or a lemma of a lexical item and any other kind of linguistic element, which can be, for example,
>
> - another (form of a) lexical item (*kith and kin* is a very frequently cited example of a nearly deterministic co-occurrence of two lexical items, as is *strong tea*);
> - a grammatical pattern (as opposed to, say, a grammatical relation), i.e. when a particular lexical item tends to occur in / co-occur with a particular grammatical construction (the fact that the verb *hem* is mostly used in the passive is a frequently cited case in point). (Gries to appear a)

Unlike Gries (to appear a), we do not consider combinations of a lexical item and a grammatical structure as phrasemes. Woolard (2000) refers to these syntactic constraints on the use of lexis as **word grammar**. They are better studied within the framework of valency theories (cf. section 2.3.1).

2.4.1.2.2. Semantic classifications

Cowie (1991; 1992) argues that, to be regarded as a (restricted) collocation, a verb-object combination has to exhibit semantic specialization of the verb, which can be of three types:

- The verb has a **weakened or delexical meaning** and the sense of the collocation is conveyed by the noun. The function of the verb is to 'grammaticalize' the agentive noun, e.g. *to do sb. a favour, to make a choice, to give a look, to take a step, to launch an appeal.*[61] These collocations often have an equivalent verb form (e.g. *to make a choice – to choose; to make proposals – to propose*) and have often been referred to as 'support verb constructions'.
- The verb develops an **abstract or figurative meaning** in combination with the noun: *deliver a speech, abandon a principle, call for action, dismiss an idea.*
- The verb is used in a **technical or semi-technical sense**: *enact measures, draft the legislation, nominate a member.*

As Koike (2002:18) has suggested, specialization can lead to neutralization: different verbs lose their independent lexical meaning and are almost used as synonyms, e.g. *put money into, pour money into, throw money* at or Spanish *dar* (En. 'to give'), *soltar* (En. 'to let go'), *lanzar* (En. 'to throw') *una carcajada* ('a good laugh').

[61] Note that these collocations, which are often called 'support verb constructions', are sometimes excluded from collocations (cf. Batoux 2003; Van der Meer 1998).

Together with the test of semantic specialization of the verb, Cowie (1991; 1992) applies the test of commutability or 'collocability' (cf. section 2.3.5) both on the verb and the noun to show whether they are the only items that can be used to express the global meaning of the collocation (e.g. *run* is not commutable in *run a deficit*) or whether they belong to a short set of synonyms (*abandon/give up a principle*) (see section 2.4.1.2.3. for more information on classifications based on the criterion of commutability).

While Cowie focuses on verb-noun collocations, Mel'čuk's (1995; 1998) semantic classification applies to collocations in general. On the basis of his semantic definition of collocation, Mel'čuk distinguishes between four main types of collocations which are covered by his formulation **'B expresses 'C' only contingent on A'**. His classification of collocations is given in Figure 2.11. Collocations that meet criterion 1(a) correspond to Cowie's category of verb-noun collocations with a delexical verb, i.e. a **light or support verb** in Mel'čuk's terms. Examples include *[to] do [N] a FAVOUR, [to] give a LOOK, [to] take a STEP, [to] launch an APPEAL, [to] lay SIEGE [to N]*. Criterion 1(b) refers to collocations in which the collocator, often an adjective, develops a **specialized meaning** that it only has in combination with the base (or with a few similar lexemes): *black COFFEE, French WINDOW, Fr. BIERE bien frappée* 'well chilled (lit. 'beaten') beer'. Criterion 2(a) is fulfilled by collocations in which the collocator, often an adjective or an adverb, keeps its 'dictionary-like' sense but cannot be replaced in the collocation by any of his synonyms, e.g. *strong (*powerful) COFFEE*. Mel'čuk includes in this group collocations in which the collocator is used as an **intensifier** such as *heavy (*weighty) SMOKER, deeply MOVED, [to] ILLUSTRATE vividly* although it is highly debatable that *heavy* in *heavy smoker* keeps its 'dictionary-like' sense (cf. section 2.3.5.3). The fourth type includes collocations such as *the HORSE neighs, aquiline NOSE, rancid BUTTER,* or *artesian WELL* in which the collocator is **bound to the base** by including the base in its meaning.

Figure 2.11: Mel'čuk's (1998) definition and classification of collocations

A collocation **AB** of language **L** is a semantic phraseme of **L** such that its signified 'X' is constructed out of the signified of one of its two constituent lexemes – say, of **A** – and a signified 'C' ['X' = 'A ⊕C'] such that the lexeme **B** expresses 'C' only contingent on **A**.

The formulation '**B** expresses "C" only contingent on **A**' covers four major cases, which correspond to the following four major types of collocations:

1. Either 'C' ≠ 'B', i.e. B does not have (in the dictionary) the corresponding signified;
   AND [(a) 'C' is empty, that is, the lexeme **B** is, so to speak, a semi-auxiliary selected by **A** to support it in a particular syntactic configuration;
   OR (b) 'C' is not empty but the lexeme **B** expresses 'C' only in combination with **A** (or with a few other similar lexemes)];
2. OR 'C' = 'B', i.e. **B** has (in the dictionary) the corresponding signified;
   AND [(a) 'B' cannot be expressed with A by any otherwise possible synonym of **B**;
   OR (b) 'B' includes (an important part of) the signified 'A', that is, it is utterly specific, and thus **B** is 'bound' by **A**].

Finally, Tutin and Grossman (2002) propose a definition based on Mel'čuk's (1998) definition (cf. Table 2.5 above) and distinguish three types of collocations on the basis of a semantic criterion, i.e. *opaque, transparent* and *regular* (Fr. 'régulière') collocations. In **opaque collocations**, the sense of the collocator in combination with the base is different from its 'usual' sense while the base keeps its meaning, e.g. *peur bleue, colère noire, nuit blanche*. **Transparent collocations** are regarded as 'prototypical collocations': they are characterized by a collocator which is 'interpretable' but hardly predictable, e.g. *faim de loup, brouillard à couper au couteau, grièvement blessé*. **Regular collocations** possess several characteristics that situate them very close to free combinations. This category includes combinations in which the collocator includes the sense of the base (*nez aquilin, année bissextile, l'âne brait*) and combinations in which the verb has a very generic sense and a high level of commutability. For example, *grand* seems to be the usual intensifier that qualifies nouns of emotion (*grande tristesse*).

2.4.1.2.3. A classification based on the criterion of 'commutability'

Howarth (1996) proposes a classification of verb-noun collocations based on their degrees of restricted collocability or 'commutability' (see Table 2.7). He distinguishes five levels of restrictedness according to the number of elements that are restricted in their commutability, the nature of this or these element(s), and the degree of restriction. At level 1, the lexical set of noun substitutes is fairly open to the point that blockages are described as **possibly semantically motivated rather than arbitrary**. Only the small set of synonymous verbs is evidence of some degrees of restricted commutability. The dividing line between nonce combinations and 'level 1' collocations is thus very fine and Howarth acknowledges that "criteria at the free end of the spectrum are far harder to control and require more intuitive judgment" (ibid: 103). Level 2 is characterized by combinations in which the possible nouns belong to **lexical sets whose collocability is arbitrarily blocked** with a set of verbs; these verbs can also be described as arbitrarily blocked with a set of nouns. For example, while *bill* and *amendment* collocate with the verbs *introduce, table,* and *bring forward*, the noun *mention* only collocates with *introduce* and *table.* Howarth creates two different categories for collocations in which there is **complete restriction of the noun** (level 3) and collocations in which there is **complete restriction of the verb** (level 4). Howarth's argument is that, since the analysis mirrors the process of language production by taking the noun in a collocation as its starting point, the most immediate indication of restrictedness, and therefore of higher degree of restrictedness, is the lack of choice of verb. At level 5, finally, there is **complete restriction on the choice of both the verb and the noun.**

Table 2.7: Howarth's (1996:102) classification of collocations

1. freedom of substitution in the noun; some restriction on the choice of verb
   an open set of nouns
   a small number of synonymous verbs
   adopt/accept/agree to a proposal/suggestion/recommendation /convention/plan etc

2. some substitution in both elements
   a small range of nouns can be used with the verb in that sense
   there are a small number of synonymous verbs
   introduce/table/bring forward a bill/an amendment

3. some substitution in the verb; complete restriction on the choice of the noun
   no other noun can be used with the verb in that sense
   there are a small number of synonymous verbs
   pay/take heed

4. complete restriction on the choice of the verb; some substitution of the noun
   a small range of nouns can be used with the verb in that sense
   there are no synonymous verbs
   give the appearance/ impression

5. complete restriction on the choice of both elements
   no other noun can be used with the verb in the given sense
   there are no synonymous verbs
   curry favour

Howarth's definition of a collocation (cf. definition n°6 in Table 2.5) places emphasis on the **specialized meaning** of one lexical element of the collocation which can be **figurative, technical** or **delexical**. He therefore classifies verb + noun collocations according to their degree of commutability as described in Table 2.7 and their specialized meaning. His results, as reported in Table 2.8, suggest that there is no direct correlation between the semantic specialization of the verb and the collocation's level of commutability.

Table 2.8: Howarth's (1996:118) categorization of collocations

category	figurative	delexical	technical
1	assume importance require qualifications	get satisfaction give evidence of	
2	assume a role follow a procedure	give emphasis on have a chance	carry a motion consider a bill
3	bring up children reach a conclusion	have access to make an application	bring an action receive Royal Assent
4	pay attention put sth to use	do one's best take precautions	obtain a warrant publish a bill
5		make an investment have a bearing on	

Nesselhauf (2005: 28-29) rightly criticizes Howarth's decision to take the restricted commutability of nouns into account in his classification. She argues that a consequence of the independent status of the noun in a noun + verb collocation is that it has the same status as the noun in a nonce combination and should therefore not be used to distinguish collocations from nonce combinations. She proposes to use restricted commutability of the verb as the only defining criterion for collocations:

> Collocations (e.g. *shrug one's shoulders, make a decision*)
> The noun can be used without arbitrary restriction in the sense in which it is being used, but the verb is, in the given sense, to some degree arbitrarily restricted to certain nouns. (Nesselhauf 2005:33)

She distinguishes five groups of 'combinatory possibilities' of verbs in verb-noun combinations. The first group includes all verbs combinable with (virtually) any noun such as *want* sth. or sb. (*want a pen, a car, a baby, peace, fun,* etc.). The second group consists in verbs that are combinable with a large group of nouns which belong to a semantic class, e.g. *kill* + [+ ALIVE] (man, dog), while in group three, verbs are combinable with a small but well-delimitable semantic group of nouns, e.g. *drink* + [liquid] (*drink water*). These three groups are semantically motivated and are classified as **'free combinations'**. By contrast, verbs in groups four and five present restrictions which are arbitrary to some degree: verbs in group four are combinable with a sizable group of nouns but there are exceptions, e.g. *commit*

+ [something wrong or illegal], but ?*commit a lie, deceit, delinquency* while those in group five only combine with a small set of nouns, e.g. *shrug shoulders, run a risk, foot the bill.*

2.4.2. The distributional approach

> "The use of the word 'meaning' is subject to the general rule that each word when used in a new context is a new word." (Firth 1957:190)

Alongside the traditional or phraseological line of development, there is also a parallel approach to collocations, the distributional approach (cf. Evert 2004), also known as the 'frequency-based approach' (Nesselhauf 2004), whose most representative members are Firth's successors and disciples in the United Kingdom. This section reviews the definitions of 'collocation' used within this approach and assesses the significance of its contribution to the study of collocations. It will also introduce two recent methodological and theoretical developments anchored in the distributional approach to collocations.

2.4.2.1. Definitions of 'collocation'

The technical meaning of 'collocation' is generally attributed to Firth, although the term can be traced further back (cf. Anderson 2006:60). Firth adopted the term in the context of his theory of meaning and argued that the meaning of a lexical item also includes 'meaning by collocation', which he defined as "an abstraction at the syntagmatic level (...) not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation of *night* (Firth 1957:196)." A second of Firth's most famous examples is the collocational meaning of *silly ass* in colloquial English which can be described as "its habitual collocation with an immediately preceding *you silly*, and with other phrases of address or of personal reference" as in Firth's examples:

- *An ass like Bagson might easily do that.*
- *He is an ass!*
- *You silly ass!*
- *Don't be an ass!* (Firth 1957:195)

However, Firth never clearly defined what he meant by 'collocation' and his uses of the term have often been confusing and contradictory (cf. Nesselhauf 2004: 2-3).

The distributional concept of collocation was later developed and refined by members of the 'Neo-Firthian' school centred around M.A.K Halliday. Herbst (1996) refers to this line of development as the 'textual approach'. The distributional concept of collocation has also become central to what Herbst (1996) called the 'statistically-oriented approach' and De Cock (2003) labelled the 'probabilistic' or 'corpus linguistic' approach to collocations, i.e. an emergent corpus-based lexicographic tradition in the United Kingdom whose leading member was J. McH. Sinclair. As a result, the concept of collocation evolved in slightly different directions. Table 2.9 gives some of the most influential definitions of the term within the distributional approach from Halliday's probabilistic definition onwards.

Some of the points of contention underlined in the phraseological approach (cf. section 2.4.1.1) are also found in the distributional approach to collocation, namely the **number of words** involved in a collocation and the level at which the collocation operates. Thus, Jones and Sinclair (1974:19) define 'collocation' as the "co-occurrence of **two** items in a text within a specified environment" (cf. definition n°3, Table 2.9) whereas Sinclair (1991:170) broadens the definition to '**two or more** words' (cf. definition n°5, Table 2.9). Similarly, Halliday (1966:157) and Mitchell (1975) tend to regard collocation as an abstract tendency which operates at the level of the **'lexeme'** or the 'root' while the definitions used by Sinclair (1991) and Kjellmer (1987) are based on **word forms**.

Table 2.9: Definitions of 'collocation' within the distributional approach

1.	Halliday (1961:276) "Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c."
2.	Mitchell (1975:117) "Invoking the inescapable condition of cognitive equivalence and considering that the same complex element is present in *he works hard, a hard worker, hard-working,* and *hard work* – the ambiguity of the last association of forms is immaterial – we see, firstly, that such a composite element comprises simpler elements occurring elsewhere in other company, ie, *hard* and *work* considered separately, and, secondly, that the composite element can exhibit its own distribution *qua compositum*. Such an abstract composite element as *hard work* we shall term a 'collocation'."
3.	Jones and Sinclair (1974:19) ""Collocation" is the co-occurrence of two items in a text within a specified environment. "Significant" collocation is regular collocation between items, such that they co-occur more often than their respective frequencies and the length of text in which they appear would predict."
4.	Kjellmer (1987:133) "A collocation is a sequence of words that occurs more than once in identical form (in the Brown corpus) and which is grammatically well-structured". *to be, one of, had been, have been, would be, will be, in which, has been, out of, United States*

5.	Sinclair (1991:170) "Collocation is the occurrence of two or more words within a short space of each other in text. The usual measure of proximity is a maximum of four words intervening. Collocation can be dramatic and interesting because unexpected, or they can be important in the lexical structure of the language because of being frequently repeated. This second type of collocation, often related to measures of statistical significance, is the one that is usually meant in linguistic discussions. (...) Collocation in its purest sense, as used in this book, recognizes only the lexical co-occurrence of words. This kind of patterning is often associated with grammatical choices as well, leading to the wealth of idioms and fixed phrases that are found in everyday English. (...) In this book, the attention is concentrated on lexical co-occurrence, more or less independently of grammatical pattern or positional relationship. (...) there is no restriction to the number of words involved. Collocation is a contributing factor to idiom."

Two points of contention are specific to the distributional approach to collocations: span size and minimum threshold of occurrence. First, Kjellmer (1987) considers that collocations are '**sequences** of words' (cf. definition n°4, Table 2.9) while most other researchers agree that collocations can be found **across a given span**, i.e. a distance in terms of orthographic words between a node and its collocates (see below for definitions of the terms). The **optimum span size** is also a matter of debate: Jones and Sinclair (1974:21-22) argue that the optimum span size is four words to the left and right of the node (4:4) while Clear (1993) uses a 2:2 window (see section 4.2.2.1.2 for more information on span size). A related issue is whether the collocational window can span across sentence boundaries. The second point of contention concerns the use of a **minimum threshold of occurrence**. Halliday (1966:159) does not make use of a minimum threshold; Kjellmer (1987:133) defines a collocation as a "sequence of words that occurs more than once" (cf. definition n°4, Table 2.9) and the minimum frequency threshold is fixed at 3 in the OSTI report (cf. Krishnamurthy 2004) (see section 4.2.3.1 for more information on span size and minimum threshold of occurrence).

A last fundamental difference concerns the relationship between a node and its collocates. Jones and Sinclair (1974) and Sinclair (1991) analyse collocation phenomena "more or less independently of grammatical pattern or positional relationship". Thus, a node and its collocates can be found in a **syntagmatic** relation (e.g. *make – decision*) as well as in a **paradigmatic** association (e.g. *hospital – nurse – doctor*). By contrast, Mitchell (1975) and Greenbaum (1974) study collocations in **grammatical patterns**. It will be seen in section 4.2.3.1.1 that this latter approach has been very influential in today's lexicography with the development of software such as the Sketch Engine.

2.4.2.2. Major contributions of the distributional approach

> "It should be stressed here that the use of numerical methods is normally only the first stage of a linguistic investigation." (Sinclair 2004 [1996]:28)

Researchers within the distributional approach have not managed to reach a consensus on a definition of 'collocation'. However, they have contributed to the study of collocations in three significant ways. First, they have proposed using a specific terminology to describe collocations in corpora. Second, they have studied collocations within a particular theoretical framework that has been described as a "distinctive vision of language study". Third, they have advocated the study of the environment of a word in corpora not only in terms of its collocates but also of its colligations, semantic preferences and semantic prosody.

2.4.2.2.1. Terminology and description

A number of technical terms to describe lexical items and their environment in corpora have been proposed within the framework of the distributional and more precisely the corpus approach to collocations following J. McH. Sinclair. This terminology is widely used today, also by linguists working in other traditions (cf. Nesselhauf 2004:5). The term **node** refers to the item whose lexical behaviour is being studied; a word occurring in close proximity to the node is called its **collocate**; all the collocates of a node constitute its collocational **range** (cf. McIntosh 1961; Greenbaum 1974); the term **span** refers to a distance in terms of number of words on either side of the node within which collocates are investigated. Note that a node and its collocates are given the same status in the distributional approach: collocation phenomena are not oriented or hierarchically ordered as is sometimes the case in the phraseological approach (cf. Hausmann's definition of base and collocator in section 2.4.1.1).

2.4.2.2.2. A distinctive vision of language study

Researchers within the distributional approach have also developed a "distinctive vision of language study" (Stubbs 1993:1), from Firth's "contextual theory of meaning" to Sinclair's (1991) "idiom principle", Hunston and Francis's (2000) "pattern grammar" and more recently, Hoey's (2004; 2005) "lexical priming", in which collocations play a major role (see below). Stubbs (1993:2) summarises this "distinctive vision of language" by citing nine principles that have been central to British contextualism, and which Sinclair has developed in detail:

I. Linguistics is essentially a social science and an applied science.
II. Language should be studied in actual, attested, authentic instances of use, not as intuitive, invented, isolated sentences.
III. The unit of study must be whole texts.
IV. Text and text types must be studied comparatively across text corpora.
V. Linguistics is concerned with the study of meaning; form and meaning are inseparable.
VI. There is no boundary between lexis and syntax; lexis and syntax are interdependent.
VII. Much language use is routine.
VIII. Language is used to transmit the culture.
IX. Saussurian dualisms are misconceived. (Stubbs 1993:2-3)

Within this framework, Sinclair has demonstrated that collocation is a ubiquitous phenomenon in language and has put forward the "idiom principle":

> 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. To some extent, this may reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort; or it may be motivated in part by the exigencies of real-time conversation. (Sinclair 1991:110)

On the basis of corpus-based data, Sinclair and his followers have also shown that:

- different forms of a lexeme may pattern differently;
- different meanings of a word have very different frequencies;
- differences in structure and collocational range are often in close correlation with the different senses of a word;
- different meanings and usages of a word often occur in very uneven distribution;
- introspection does not give evidence about usage and can hardly predict the results of a collocational study (cf. Sinclair 1991:38; 1999a; 1999b).

As for intuition[62], Sinclair, however, insists that "there is one process where intuition can be safely trusted. In the evaluation of corpus evidence the researcher has virtually no option but to yield to the organising influence of his or her intuition" (Sinclair 2004b:45).

[62] Sinclair sometimes uses intuition where introspection would seem to be more appropriate (cf. Sinclair 2004b). However, as argued by Rundell on the corpora list (<http://torvald.aksis.uib.no/corpora/2001-4/0080.html >), intuition and introspection should be clearly distinguished:

> *I know intuition is a dirty word in some circles, but I think we need to *completely* distinguish it from introspection (i.e .where you just try to retrieve data from your own mental lexicon - this of course IS demonstrably unreliable). Could we say in this context intuition is the faculty by which humans interact with and interpret corpus data? All I know is, you don't get far without it in lexicography.*

2.4.2.2.3. Collocation, colligation, semantic preference and semantic prosody

The environment of a word can be studied at different levels of abstraction. While collocation is the co-occurrence of words, **colligation** is today commonly understood as "the grammatical company a word keeps" (Hoey 2004:28). Stefanowitsch and Gries define colligations as "linear co-occurrence preferences and restrictions holding between specific lexical items and the word-class of the items that precede or follow them." (2003:210). Thus, the word *involvement* is said to colligate with prepositions but to collocate with *in* and *with*. Hoey (2004) further extends the notion to cover all types of grammatical preferences of a word:

> I define colligation as
> a. the grammatical company a word keeps (or avoids keeping) either within its own group or at a higher rank.
> b. the grammatical functions that the word's group prefers (or avoids).
> c. the place in a sequence that a word prefers (or avoids). (Hoey, 2004: 28)

Hoey gives the example of the word *tea* to illustrate the first type of colligation: *tea* typically occurs as premodification to another noun, e.g. *tea chest, tea pot, tea bag, tea break*, but avoids co-occurrence with markers of indefiniteness (*a, another*, etc.). To illustrate the second type of colligation, Hoey compares the grammatical functions, i.e. part of subject, object, complement or adjunct, of the word *consequence* with that of the nouns *question, preference, aversion* and *use* and shows that there is a negative colligation between *consequence* and the function of object. *Consequence* also illustrates the third type of colligation: the noun typically occurs as part of the theme rather than the rheme.

Hoey's definition of a colligation shares several features with Hunston and Francis's (2000) concept of **pattern** under which the authors subsume both collocation and colligation and which they define as follows:

> The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it. (Hunston and Francis 2000: 37)

However, the approach advocated by Hunston and Francis (2000) differs significantly from collocational and colligational analysis of words in that pattern grammar starts with a set of patterns around the major word classes and investigates which words typically occur in these patterns rather than the other way round. For example, the pattern '*it v-link ADJ of n to-inf*' is associated with adjectives evaluating the action indicated by the *to*-infinitive clause. Thus, in the following sentences, the two adjectives used in this specific pattern are *typical* and *kind*:

*2.5. It was **typical** of him to see politics in ethical terms.* (BNC)
*2.6. She said, "Anyhow, it was **kind** of him to stay."* (BNC)

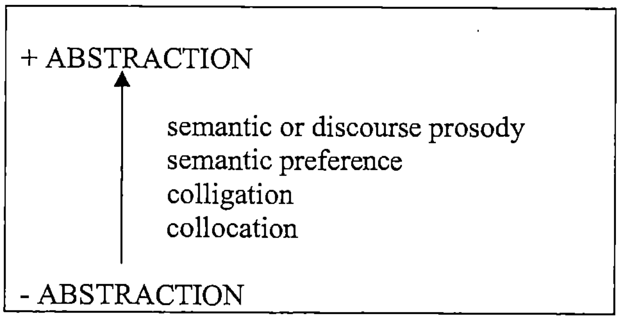
The adjectives used with this pattern belong to three general meaning groups: those associated with positive evaluation (e.g. *brave, clever, courageous, fair, generous*); those expressing negative evaluation (e.g. *absurd, arrogant, cruel*); and those evaluating the typicality of the action (e.g. *typical, untypical, characteristic*) (cf. Hunston and Francis 2000:99-100).

Collocates of a particular word have repeatedly been found to constitute semantic sets. The relationship between a lexical item and a lexical set of semantically related words is called **semantic preference** by Sinclair (1996, 1998) and Partington (2004). For example, Partington (2004:148) observes that collocates of the maximizers *utterly, totally, completely* and *entirely* share the semantic preference of 'absence/change of state', e.g. *totally uneducated* and *completely lacking.*

The "proximity of a consistent series of collocates" (Louw 2000:57) may establish yet another form of meaning, i.e. **semantic prosody**, whose primary function is "the expression of the attitude of its speaker or writer towards some pragmatic situation" (ibid) (see also Louw 1993). Partington (2004:150-151) illustrates the interdependence of semantic preference and semantic prosody by using Stubbs's (2001) example of the verb *undergo*, which collocates with, and thus shows semantic preference for, items from the lexical sets of 'change' (e.g. *dramatic changes, a historic transformation*), 'medicine' (e.g. *treatment, brain surgery*), 'testing' (e.g. *examinations*) and 'involuntariness' (e.g. *must, forced to, required to*). All these semantic preferences imbue the item *undergo* with a very strong unfavourable semantic prosody. Other often cited examples of words with negative semantic prosody include *happen, set in* (cf. Sinclair 1991) and *cause* (cf. Stubbs 1995).

Collocation, colligation, semantic preference and semantic prosody can be presented on a continuum from the least abstract to the most abstract relationship between a word and its environment as shown in Figure 2.12. Collocation is "precisely located in the physical text, in that even the inflection of a word may have its own distinctive collocational relationship" (Sinclair 1998:16). To conduct a colligational analysis, a word class has to be assigned to each collocate of the word under study and collocates have to be grouped by their word class. Semantic preference and semantic prosody are abstractions at the level of semantic fields and speaker attitude respectively. The systematicity of these relationships between a word and its environment has led Sinclair and his colleagues to postulate the existence of an extended unit of meaning "where collocational and colligational patterning (that is lexical and grammatical choices respectively) are intertwined to build up a multi-word unit with a specific semantic preference, associating the formal patterning with a semantic field, and an identifiable semantic prosody, performing an attitudinal and pragmatic function in the discourse" (Tognini-Bonelli 2002:79).

Figure 2.12: Different degrees of abstraction



A last but not unrelated type of unit unveiled by corpus-based studies of words and their environment is the **collocational framework**, which Renouf and Sinclair (1991:128) define as "a discontinuous sequence of two words, positioned at one word remove from each other; they are therefore not grammatically self-standing; their well-formedness is dependent on what intervenes." Examples include '*a* + ? + of', '*an* + ? + *of*', '*be* + ? + *to*', and '*too* + ? + *to*'. In a 10-million-word corpus of written British English, for example, the first twenty collocates of the collocational framework '*too* + ? + *to*' are *late, much, young, easy, small, close, tired, weak, good, old, early, hard, busy, ready, dark, big, long, poor, proud* and *far* (ibid 132). On the basis of an exploratory analysis in two sections of the Birmingham Collection of English Text, Renouf and Sinclair (1991:143) argue that "two very common grammatical words, one on either side, offer a firm basis for studying collocations." They show that collocational frameworks are characterized by different degrees of productivity and that the choice of collocates is governed by both elements of the framework. Moreover, they demonstrate that the words that occur in a given framework belong to semantic groupings.

In summary, it appears that the distributional approach to collocations not only contributed to the descriptive study of these items by providing a terminology and a methodology (see section 4.2.2 for more information on corpus analysis of collocations) but also developed a theoretical framework in which collocations and related constructions at different levels of abstraction play a major role. In the next section, we will briefly introduce recent methodological and theoretical developments anchored in the distributional approach to collocations.

2.4.2.3. Recent methodological and theoretical developments

The last few years have seen the emergence of two methodological and theoretical developments. First, Stefanowitsch and Gries (2003) have proposed a new distributional approach to the environment of words, i.e. collostructional analysis, which merges the theoretical positions held by Construction Grammar and similar theories, and the methodological framework of corpus-based collocational analysis. Second, Hoey (2004) has formulated the theory of 'lexical priming', i.e. the first theory of language that places lexis at its centre and attempts to give a psychological explanation for the ubiquitous phenomenon of collocation. These two recent developments are briefly discussed in this section.

Stefanowitsch and Gries (2003) situate their work in the framework of Construction Grammar, more specifically the versions developed by Lakoff (1987) and Goldberg (1995). This theory sees the *construction*, i.e. "a pairing of form with meaning/use such that some aspect of the form or some aspect of the meaning/use is not strictly predictable from the component parts or from other constructions already established to exist in the language" (Goldberg 1996: 68) as the primary unit of grammar. Stefanowitsch and Gries (2003) propose to apply collocational analysis within a constructional view of language, i.e. **collostructional analysis** (a blend of *construction* and *collocational analysis*), with the aim of providing "an objective approach to identifying the meaning of a grammatical construction and of determining the degree to which particular slots in a grammatical structure prefer, or are restricted to, a particular set or semantic class of lexical items" (ibid 211). Collostructional analysis can be performed on single words (e.g. *cause*) and what the authors call 'variable idioms' (e.g. the [X *think nothing of* $\mathrm{V}_{\mathrm{gerund}}$] construction), partially filled and unfilled argument structure constructions (e.g. the *into*-causative, the ditransitive) and tense, aspect and mood (e.g. lexemes attracted by the progressive form, the imperative or past tense).

They argue that collostructional analysis allows for a more refined analysis than collocational analysis. For example, previous collocational analyses have shown that the verb *cause* has a negative semantic prosody (cf. Stubbs 1995). A collostructional analysis of the verb confirms this claim while showing that there are fundamental differences between the three constructions in which the verb occurs with respect to the type of lexemes found (cf. Figure 2.13).

Figure 2.13: A collostructional analysis of the word cause (Stefanowitsch and Gries 2003:222, Table 6)

TRANSITIVE		PREPOSITIONAL DATIVE		DITRANSITIVE	
Collexemes	Coll. strength	Collexemes	Coll. strength	Collexemes	Coll. strength
problem (18)	3.30E-18	harm (3)	4.37E-10	distress (1)	4.54E-04
damage (7)	2.52E-10	damage (2)	5.47E-05	hardship (1)	4.54E-04
havoc (3)	8.74E-09	modification (1)	6.56E-04	discomfort (1)	5.19E-04
cancer (4)	4.39E-07	inconvenience (1)	8.43E-04	inconvenience (1)	5.84E-04
injury (5)	7.12E-07	famine (1)	9.37E-04	problem (2)	8.57E-04
injustice (3)	9.84E-07	delight (1)	1.59E-03	pain (1)	3.24E-03
stampede (2)	5.08E-06	problem (2)	1.83E-03	difficulty (1)	7.83E-03
congestion (2)	1.01E-05	disruption (1)	2.06E-03	night up (1)	1.89E-02
extrusion (2)	1.01E-05	accident (1)	1.66E-02		
change (6)	1.43E-05				

In a transitive construction, the verb *cause* occurs exclusively with external states and events, a feature which is also predominant in prepositional dative constructions:

2.7. *A mild recession may **cause** far more economic damage than a one-day stockmarket fall of, say, 25%, but it is much less unsettling.* (BNC)

2.8. *Factory worker Robert Brooks, 33, of Pleasant Row, Hyson Green, Nottingham, admitted six charges of **causing** grievous bodily harm to his son and one charge of child cruelty.* (BNC)

By contrast, in the ditransitive construction, the verb *cause* collocates primarily with mental states and experiences, a difference which has been missed by traditional collocational analyses.

2.9. *The knowledge **caused** her genuine distress and, in the face of Constance's increasing truculence, she turned to Louise for counsel.* (BNC)

The second development deeply anchored in the distributional approach to collocations is theoretical in nature. Corpus linguists have shown the pervasiveness of collocations (see, e.g. Sinclair 1991) while other authors have discussed the native-like character or naturalness of collocations (cf. Pawley and Syder 1983). Hoey (2005) however observes that theories of language account only for what is possible in a language and not for what is **natural**. In *Lexical Priming: A new theory of words and language*, Hoey formulates a new theory of language that places **lexis** at its centre and gives a psychological explanation for the ubiquitous phenomenon of collocation. He thus adopts a psychological definition of collocation and postulates that this phenomenon can be observed in corpora:

> So our definition of collocation is that it is a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution. The definition is intended to pick up on the fact that collocation is a psycholinguistic phenomenon, the evidence of which can be found statistically in computer corpora. (Hoey 2005:5)

Hoey's definition of collocation follows Jones and Sinclair's (1974) proposal of a 4-word span to extract significant collocates and Sinclair's (1991) conception of collocation as a relation between word-forms rather than lemmas.

Hoey uses the notions of **priming** and **nesting** to explain the pervasiveness of collocations:

> We can only account for collocation if we assume that every word is mentally **primed** for collocational use. As a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context. The same applies to word sequences built out of these words; these too become loaded with the contexts and co-texts in which they occur. I refer to this property as **nesting**, where the product of a priming becomes itself primed in ways that do not apply to the individual words making up the combination. (Hoey 2005:8)

Priming is not a permanent feature of a word or a word sequence; it is an individual construct that changes constantly as new contexts are discovered through contact with language. Individual primings are then harmonised through contact with others.

Priming goes beyond collocations to more complex relations between words such as colligations and semantic prosodies. It is described as the "driving force behind language use, language structure and language change" (Hoey 2005:12). This leads to a number of hypotheses that Hoey summarises in the first chapter of his book and that are further developed in the next chapters:

1. Every word is primed to occur with particular other words; these are its collocates.
2. Every word is primed to occur with particular semantic sets; these are its semantic associations.
3. Every word is primed to occur in association with particular pragmatic functions; these are its pragmatic associations.
4. Every word is primed to occur in (or avoid) certain grammatical positions, and to occur in (or avoid) certain grammatical functions; these are its colligations.
5. Co-hyponyms and synonyms differ with respect to their collocations, semantic associations and colligations.
6. When a word is polysemous, the collocations, semantic associations and colligations of one sense of the word differ from those of its other senses.
7. Every word is primed for use in one or more grammatical roles; these are its grammatical categories.
8. Every word is primed to participate in, or avoid, particular types of cohesive relation in a discourse; these are its textual collocations.

9. Every word is primed to occur in particular semantic relations in the discourse; these are its textual semantic associations.
10. Every word is primed to occur in, or avoid, certain positions within the discourse; these are its textual colligations. (Hoey 2005:13)

These hypotheses clearly show that Hoey's objective is to account psychologically for a number of phenomena, most of which were previously described by Sinclair and his followers.

Hoey insists that primings are constrained by **register** and **genre**. He gives the example of the word *research* which is primed in the mind of academic language users to occur with *recent* in academic discourse and news reports of research. The collocation is not primed to occur in other text types or other contexts. Priming is thus described as sensitive to the textual, generic and social contexts in which a lexical item is encountered. Following Firth (1951), Hoey argues that it is part of an individual's knowledge of a word that it is used in certain environments (collocations, colligations, etc.) in certain text types.

Finally, Hoey's conception of priming also carries direct implications for the way corpus data are used within this new theory of words and language. Corpora cannot provide evidence of which primings are present for any language user. They can, however, show the types of data a language user might encounter and indicate the kinds of feature for which lexical items might be primed. They can thus reveal which types of primings are likely to be shared by a large number of language users. In Hoey's words, corpora "can serve as a kind of laboratory in which we can test for the validity of claims made about priming" (Hoey 2005:14).

2.4.3. Conclusion

The phraseological approach and the distributional approach to collocations developed separately while pursuing different objectives. As a result, their respective use of the term 'collocation' differs significantly: within the phraseological approach, the term stands for an arbitrarily restricted word combination (e.g. *confirmed bachelor*) as well as for the linguistic phenomenon such a word combination represents while it refers to a probabilistic or statistical phenomenon in the distributional approach. Our survey has also shown that 'collocation' does not have a shared meaning within each tradition: the many definitions proposed within one approach share at best one or two features with each other (cf. sections 2.4.1.1 and 2.4.2.1). This leads to a situation in which results from various studies are hardly comparable[63].

Several researchers have also tried to encompass the different meanings of 'collocation' in a single definition as illustrated in the following two definitions:

> a collocation is any holistic lexical, lexico-grammatical or semantic unit normally composed of two or more words which exhibits minimal recurrence within a particular discourse community. (Siepmann 2005:438)
>
> *collocations* must be defined as 'combinations of lexemes exhibiting a medium degree of observable recurrence, mutual expectancy and idiomaticity'. (Schmid 2003:249)

Siepmann (2005) proposes a definition that relies on three different traditions to collocations: (1) the adjective *holistic* comes from psycholinguistic approaches to collocations[64]; (2) the expression *lexico-grammatical or semantic unit normally composed of two or more words* refers to criteria used to define collocation within the phraseological approach and (3) *minimal recurrence* is a frequency-based criterion typical of the distributional approach. The outcome of such an amalgam is a broad definition which includes all sorts of phrasemes (e.g. *the car holds the road well, shall I break this note into a smaller one, an empty parking space, pauvre hère* ('miserable wretch'), *far be it from me to* + INF), to the point of making 'collocation' a vacuous term.

Corpus-based studies of word combinations need to use different terms to refer to (1) recurrent or statistically prominent word combinations in corpora and (2) phrasemes. In this thesis, the term **collocation** will be exclusively used for arbitrarily lexically restricted word combinations. Following Schmid (2003) and Evert (2004), we will use the term **co-occurrence** to refer to the combinations of lexical items within a given span:

> [I]t is not clear what is gained by calling co-occurrences of words *collocations*, when the term *combination*, or indeed co-*occurrence* itself, covers the same range of phenomena. (Schmid 2003:239)

[63] Idiosyncratic definitions of 'collocation' have also been proposed but they were not reviewed in this section as they have not received wide currency. An example is van der Wouden's (1997) definition (quoted in Gledhill 2000: 10-11):

> "I will use the term COLLOCATION as the most general term to refer to all types of fixed combinations of lexical items. In this view idioms are a special subclass of collocations, to wit those collocations with a non-compositional, or opaque semantics." (van der Wouden 1997: 9)

[64] It was however shown in section 2.4.1.1 that the formulaic nature of collocations remains a major point of contention.

The term **co-occurrent** will be used to refer to each item in a co-occurrence. We will also make a distinction between casual and significant co-occurrences[65]. **Significant co-occurrences** are regular co-occurrences between two items, such that they co-occur more often than their respective frequencies and the length of text in which they appear would predict. Statistical measures for determining the degree of significance are described in section 4.2.3. **Casual co-occurrences** are statistically non-significant co-occurrences.

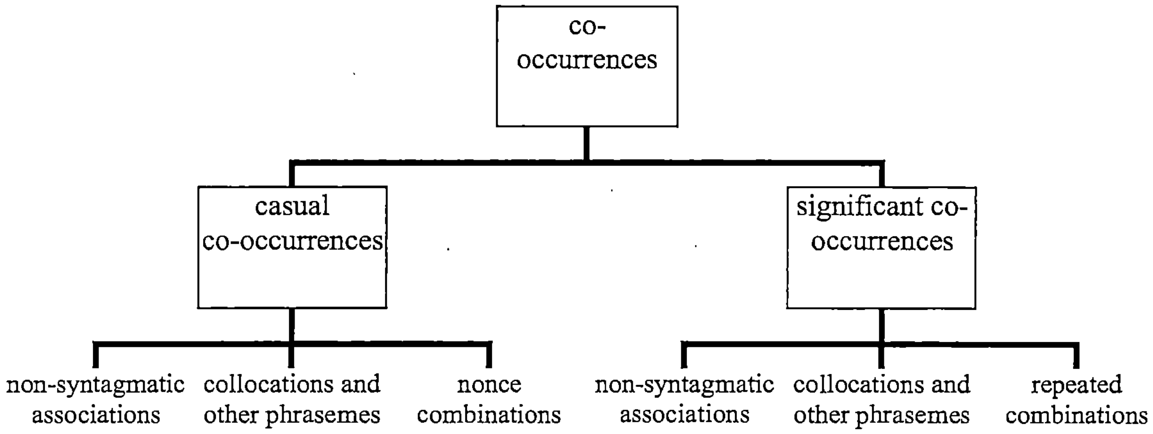
As already underlined by Evert (2004), there is no adequacy between significant co-occurrences and collocations. Consequently, statistics cannot be used to define collocations (cf. Williams 2003; Schmid 2003). They can, however, give prominence to recurrent word combinations in the corpus under study which will be the focus of a linguistic analysis. A linguistically oriented analysis of co-occurrent lexical items will typically involve the following steps. First, an association measure separates statistically significant co-occurrences from casual co-occurrences as illustrated in Figure 2.16. Second, significant co-occurrences are analysed linguistically: co-occurrent items will fall under three main categories:

- **Repeated combinations**, i.e. syntagmatic combinations that are not phraseological in nature but appear repeatedly in the corpus under study (e.g. the combination *buy + car* in a text about cars);
- **Non-syntagmatic associations** which include paradigmatic relations such as *doctor – hospital* as well as all sorts of combinations of words that are not found in a syntagmatic relation but are nevertheless extracted by a co-occurrence analysis, e.g. *a-of, fall-three, noteworthy-casual.*
- **Collocations and other phrasemes.**

It is essential to bear in mind that not all phrasemes used in a corpus will be extracted by a co-occurrence analysis. Figure 2.14 shows that different types of word combinations will also be found among casual co-occurrences. The types of co-occurrences retrieved by statistical techniques are very much dependent on the association measure used (see section 4.2.2.3), register, text type (see section 2.5) and corpus size. Consequently, a collocation or other phraseme can be missed for a number of reasons: the association measure is not adequate, the phraseme is not frequent in a particular register or text type or the corpus is not big enough to retrieve all phrasemes.

[65] This distinction is based on that of significant vs. casual collocations first proposed by Sinclair and his colleagues in the OSTI report (cf. Krishnamurthy 2004).

Figure 2.14: Co-occurrences vs. collocations and other phrasemes


co-occurrences
casual co-occurrences
significant co-occurrences
non-syntagmatic associations
collocations and other phrasemes
nonce combinations
non-syntagmatic associations
collocations and other phrasemes
repeated combinations

We define a **restricted collocation** or **collocation** as a usage-determined or preferred syntagmatic relation between two lexemes in a specific syntactic pattern. Both lexemes make an isolable semantic contribution to the word combination but they do not have the same status: the 'collocator' is arbitrarily selected and semantically determined by the 'base.' The category of collocations includes **support verb constructions**, which are defined as collocations in which the collocator, i.e. the verb, has a weakened or delexical meaning (cf. 2.4.1.2), e.g. *to do a favour, to make a choice, to give a look, to take a step, to launch an appeal.* Although collocations are defined as relations between two lexemes, they are not restricted to these two lexemes and include the other elements associated with them (cf. Nesselhauf 2005:25). Thus, *make an issue of* will be referred to as a collocation and not only *make + issue.* The term **grammatical collocation** is used to refer to restricted combinations of a lexical and a grammatical word (cf. section 2.3.7).

Attempts at categorizing collocations have been made by phraseologists such as Cowie, Howarth, Mel'čuk, Hausmann and Nesselhauf (cf. 2.4.1.2). Some authors have used a syntactic criterion; others have preferred a semantic distinguishing feature or used the criterion of commutability to distinguish between different types of collocations. Semantic classifications have repeatedly been criticized on the basis that it would be "theoretically possible to make sense distinctions that are so fine that all combinations could be considered semantically motivated" (Nesselhauf 2005:31). Categorisations based on the criterion of commutability have been criticized for the arbitrariness of the thresholds used to separate restricted from unrestricted commutability. For these reasons, no attempt will be made in this thesis to distinguish between different types of collocations.

2.5. Where phraseology and discourse meet

The review of **traditional** typologies of phrasemes conducted in section 2.2.1 has shown that while the interactional or **pragmatic** properties of phrasemes have already received full recognition (e.g. Cowie's routine and speech formulae, Mel'čuk's pragmatic phrasemes, De Cock's pragmatic prefabs), their **discoursal** functions have rarely, and incompletely, been used to categorize phrasemes. Burger (1998) proves the exception by proposing a functional typology which includes a category of structural phraseological units (cf. section 2.2.1). However, this category is very restricted and mainly consists of complex prepositions and complex conjunctions.

Corpus-based studies have recently highlighted the important role played by **lexical bundles** which are "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber et al 1999:990) (cf. section 1.4.1). These sequences are largely semantically and syntactically compositional and have traditionally been considered as falling outside the limits of phraseology. They have however revealed themselves to be pervasive in language. Biber and Conrad describe them as providing "basic building blocks for constructing spoken and written discourse" and argue that these lexical building blocks "tend to be used frequently by speakers or writers within a **register**"[66] (Biber and Conrad 1999:185). Biber et al. (2003) propose an initial taxonomy of lexical bundles based on their typical **discourse** functions. They group bundles into four major functional categories while insisting that some lexical bundles are multi-functional:

- **Referential** bundles "make direct reference to elements in the physical world or the textual context" (ibid:79), e.g. time markers (*the end of the, at the same time*);
- **Text organizers** "reflect relationships between prior and coming discourse" (ibid), e.g. contrast/comparison (*on the other hand*), inferential (*on the basis of*) framing (*in the presence of, in the case of*);
- **Stance** bundles "express attitudes or assessments of certainty towards the following proposition" (ibid), e.g. epistemic-impersonal (*it is possible to*), obligation (*going to have to*), intention (*let's have a look at*);

[66] My emphasis. Biber and his colleagues use the term 'register' to refer to genre (textbook vs. article) and medium (written vs. spoken discourse).

- **Interactional** bundles are "usually situational formulas associated with a specific situation, or conversational expressions used as strategies for conversational interactions" (ibid), e.g. reporting (*I said to him*), imprecision tags (*or something like that*), politeness markers (*thank you very much*).

Biber and Conrad (1999) further describe lexical bundles as "extended collocations: sequences of three or more words that show a statistical tendency to co-occur", a definition largely relying on the (corpus-based) distributional approach to collocations (cf. section 2.4.2). Although they draw a structural and semantic distinction between idioms and lexical bundles, the authors do not assess the phraseological nature of lexical bundles.

The link between discourse and phraseology is not established in traditional studies of **metadiscourse**, i.e. "the cover term for the self-reflexive expressions used to negotiate interactional meanings in a text, assisting the writer (or speaker) to express a viewpoint and engage with readers as members of a particular community" (Hyland 2005:37) (cf. Vande Kopple 1985; Crismore et al. 1993). This is particularly unfortunate as **textual** and **communicative phrasemes** play a major role in metadiscourse. Table 2.10 shows Hyland's (2005) interpersonal model of metadiscourse: 40% of the examples of interactive and interactional metadiscourse markers given by the author are phrasemes (in italics).

Table 2.10: An interpersonal model of metadiscourse (Hyland 2005:49, Table 3.1)

Category	**Function**	**Examples**
Interactive	**Help to guide the reader through the text**	**Resources**
Transitions	express relations between main clauses	*in addition*; but; thus; and
Frame markers	refer to discourse acts, sequences or stages	finally; *to conclude; my purpose is*
Endophoric markers	refer to information in other parts of the text	*noted above; see Fig.; in section 2*
Evidentials	refer to information from other texts	*according to X*; Z states
Code glosses	elaborate propositional meanings	namely; e.g.; *such as; in other words*
Interactional	**Involve the reader in the text**	**Resources**
Hedges	withhold commitment and open dialogue	might; perhaps; possible; about
Boosters	emphasize certainty or close dialogue	*in fact*; definitely; *it is clear that*
Attitude markers	express writer's attitude to proposition	unfortunately; *I agree*; surprisingly
Self mentions	explicit reference to author's	I; we; my; me; our
Engagement markers	explicitly build relationship with reader	consider; note; *you can see that*

The links between phraseology and discourse are better established, or at least suggested, in studies whose aims are to psycholinguistically account for phrasemes, to describe them for pedagogical or lexicographical purposes or to examine the phraseology of a specific genre. As shown in section 2.2.3, Wray's (2000; 2002) **psycholinguistic** analysis of formulaicity assigns a prominent role to formulaic sequences such as *as a result of* and *as a consequence of*. By organizing and signalling the organization of discourse, these sequences aid both the speaker's production and the hearer's comprehension. Nattinger and DeCarrico (1992) examine phrases that have been assigned pragmatic function, viz. **lexical phrases**, and focus on the category of **discourse devices** (see section 2.3.4). They argue that important pedagogical implications can be drawn from their study as it shows, for example, that there are major differences between lexical phrases in spoken versus written transactional discourse. Thus, exemplifiers such as *take a look at* and *it's like X* are more commonly found in spoken discourse than in written texts where *for example* is preferred. Their study, however, lacks empirical foundation and the lexical phrases described as typical of written discourse are highly questionable:

> [T]he actual lexical phrases used for these 'purpose' and 'maintenance' functions in writing are usually somewhat distinct from those used in conversation. For example, to nominate or shift a topic in conversation, common lexical phrase markers are such as *(by the way) do you know/remember X?* and *guess what?* These, however, do not usually appear in transactional discourse. Instead, we are more likely to find topic markers and shifters like *let me start with X, what I'd like to do is X*, and so on. (Nattinger and DeCarrico 1992: 83).

A place where links between phraseology and discourse have already been quite successfully developed is in **lexicological** and **lexicographical corpus-based** studies which are interested in establishing correlations between the form, function and frequency of phrasemes (cf. Moon 1998a:216). Two studies are worth mentioning here. First, Moon (1998a) proposes a classification of Fixed Expressions and Idioms (FEIs), viz. "holistic units of two or more words" (Moon 1998a:2), based on their **text functions**. Besides the categories of informational and evaluative FEIs, three discourse-related categories are distinguished:

- **Situational** FEIs are "typically found in spoken discourse as they are responses to or occasioned by the extralinguistic context: they may also be illocutionary speech acts" (ibid: 225). Examples include *long time no see, knock it off!, excuse me!, see you, good morning, I beg your pardon, no problem, walls have ears.*
- Two-third of the **modalizing** FEIs indicate epistemic modality, that is, they represent the speaker or writer's commitment to the truth value of the proposition: *I kid you not, you know what I mean, to all intents and purposes, at any price, I mean, if need be, in effect, no doubt, on no account, up to a point.*

- **Organizational** FEIs are of two types. A first group consists of FEIs which "control the continuity of text content" (ibid: 234): they indicate logical connections and relations such as purpose, reason, cause, result. Examples include *thanks to, in the light of, on the grounds that, in the event, in spite of,* and *with a view to*. The second type of organizational FEIs organize texts at a metadiscoursal level: they function as sequencers (*to begin with, in the first/second place*), boundary markers (*so much for*), signals of additional information, clarification, suggestion (*in addition, for example, in other words*), signals of counter-arguments, contrasts and denials (*on the other hand, on the contrary, as against*), signals of summaries and conclusions (*in a nutshell, to cut a long story short*), quotation markers (*in X's words, as X puts it*) or to comment on the selection of lexis itself (*to put it mildly, for want of a better word*).

Moon (1998a) also examines the many ways by which FEIs can give textual **cohesion**. She shows that organizational FEIs provide "grammatical cohesion, either referentially by tying texts to contexts in time and space, or conjunctively by showing the logical connections between propositions or signalling kinds of information, and so on" (Moon 1998a:279). Examples of cohesion through **conjunction** include *in turn, in other words, in fact, so that* and *at last.*

Second, Siepmann (2005) studies what he calls 'second-level discourse markers' (SLDMs), i.e. medium frequency fixed expressions or collocations whose function is to "facilitate the process of interpreting coherence relation(s) between elements, sequences or text segments and / or aspects of the communicative situation" (Siepmann 2005:52) in English, French and German academic and journalistic texts (see section 1.4.1). He carefully examines the syntactic, semantic and pragmatic (see Appendix 1.2) properties of SLDMs and shows that these lexical items are distributed across the entire phraseological cline. At one end are totally fixed opaque SLDMs such as *be that as it may be*; at the other extreme lie sentence-integrated markers such as *(NP) provides a good example.* He thus argues that "multi-word discourse markers can be described as collocations or fixed expressions" (Siepmann 2005:49).

Table 2.11 provides an overview of the wide range of lexico-grammatical realizations of SLDMs in English. Siepmann (2005) distinguishes between three major categories: (1) set expressions, (2) sentence fragments and (3) sentence-integrated markers. The SLDMs in the category of **set expressions** form a fairly small group. **Sentence fragments** have been classified into 15 categories, most of which are based on clause patterns.

Table 2.11: An overview of lexico-grammatical realizations of SLDMs in English (Siepmann 2005:55)

Category	Example
A. Set expressions	
1.1. Structurally incomplete set expressions	so far so good be that as it may be with hindsight/ with the benefit of hindsight
1.2. Structurally complete set expressions	to this we now turn this is not the whole story
B. Sentence fragments	
1. anticipatory it + verb / adjective phrase (+ complement clause fragment)	it will be seen that it seems arguable that It is worth noting that
2. existential clause (+ complement clause fragment)	there is no denying the fact that there are good reasons for believing that
3. personal pronoun (I/we/one) + (auxiliary) verb phrase (+ complement clause fragment)	one must acknowledge that I must point out that we find that
4. noun phrase + [+ (...)] +copular *be* + that-clause	a first point is that a further difficulty (for such an approach) is that my guess is that
5. adverbial clause fragments	as has been noted earlier
6. participial clauses	turning to (..., we find that)
7. *with* + verbless clauses	with this in mind
8. infinitive clauses or infinitive clause fragments	to return to (NP) to sum up
9. sentence adverbs	interestingly (enough)
10. imperatives (and hortatives) + noun / prepositional phrase (fragment)	let us first look at see further in
11. verbless clause fragments	One final point on (NP) An example:
12. noun/pronoun + verb phrase + clause fragment	the same goes for mention should be made of (NP)
13. variable prepositional phrases (phrasal constraints)	in this case/ in the present case in this connection/ in that connection
14. phrasal constraints	far be it from me to suggest far be it from me to claim
15. *here is / are* + complement	here are a few examples
C. Sentence-integrated markers	
1. noun phrase + V (+ ...) (active and passive)	we have (here) a good example is provided by (NP) <> (NP) provides a good example
2. sentence-like units	such instances could be multiplied

The category of **sentence-integrated markers** includes sequences that are "usually centred around a two-element association realizing a succession of two or more clause constituents" (Siepmann 2005: 61), e.g. *I will define (NP) as follows; we use the term to refer; a moment's reflection suggests*. The distinction drawn between sentence fragments and sentence-integrated markers sometimes lacks coherence: SLDMs such as *one must acknowledge that, a* *first point is that* and *here are a few examples* are classified as sentence fragments while *we have (here)* and *a good example is provided by* are sentence-integrated markers. However, Siepmann's analysis of SLDMs is one of the most comprehensive studies of the lexico-grammatical realizations of what are referred to in this thesis as textual phrasemes.

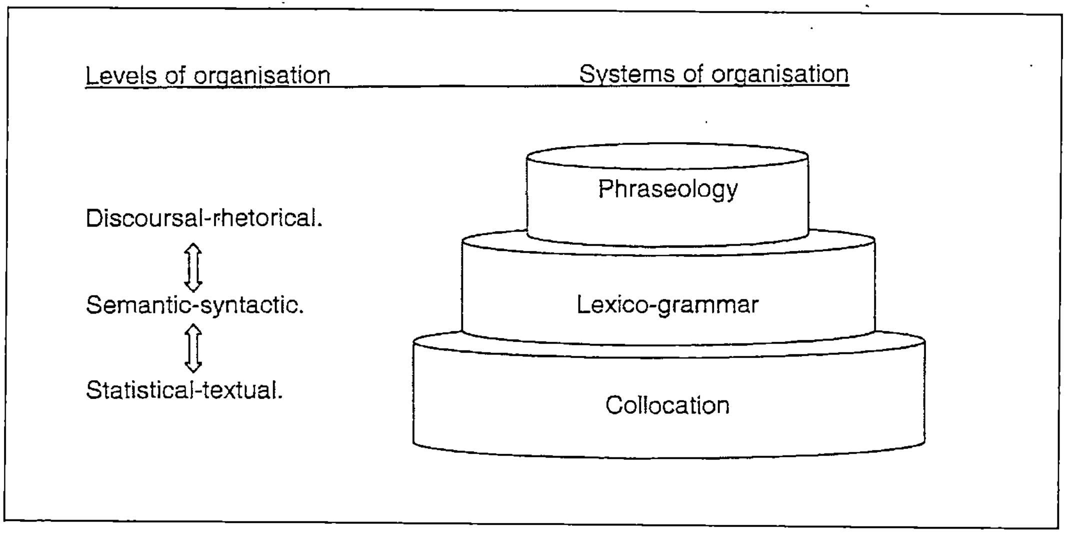
In the last few years, **genre-based** studies of collocations in academic discourse have given fresh impetus to the study of the **phraseology-discourse interface** by investigating the relation between phrasemes and the most typical rhetorical functions of a specific text type or text section (cf. section 1.4.1). They have shown that much of the language involved in a particular discourse community is highly conventionalized in nature and have highlighted the importance of multi-word units to the lexical profile of a given genre (cf. Luzón Marco 2000). They have also suggested that "cohesive mechanics of the discourse community appear to be stronger than previously imagined, even if these are largely invisible" (Gledhill 2000:131). These findings have made Gledhill (2000) argue for a **rhetorical** or pragmatic definition of 'phraseology':

> Phraseology is the 'preferred way of saying things within a particular discourse'. The notion of phraseology implies much more than inventories of idioms and systems of lexical patterns. Phraseology is a dimension of language use in which patterns of wording (lexico-grammatical patterns) encode semantic views of the world, and at a higher level idioms and lexical phrases have rhetorical and textual roles within a specific discourse. Phraseology is at once a **pragmatic dimension of linguistic analysis**, and a **system of organization**[67] which encompasses more local lexical relationships, namely collocation and the lexico-grammar. I claim that the phraseological analysis of a text should not only involve the identification of specific collocations and idioms, but must also take account of the correspondence between the expression and the discourse within which it has been produced. (Gledhill 2000:202)

Gledhill proposes to break phraseology down into sub-systems which correspond to different levels of description and organization. The left part of Figure 2.15 represents increasingly sophisticated levels of textual description. Gledhill argues that phraseology consists of expressions with specific discoursal functions which are recognisable at a higher level of organization than units which are simply seen as fixed from a syntagmatic or semantic point of view (e.g. fixed phrases and idioms) and which thus correspond to a semantico-syntactic system of explanation. Fixed phrases and idioms involve in turn "a more complex level of organization than collocations [*co-occurrences in this thesis*], which are simply recognized as textually recurrent expressions" (Gledhill 2000:203).

[67] My emphasis.

Figure 2.15: Gledhill's (2000:203) rhetorical definition of phraseology



Traditional typologies of phrasemes clearly need to take stock of a number of theoretical developments which have given a prominent descriptive and explanatory role to the interface between phraseology and discourse. A much broader range of phrasemes than usually acknowledged by the many approaches surveyed in this chapter lie at this interface and challenge the traditional boundaries of phraseology. As a consequence, in this thesis, the category of **textual phraseme** not only includes complex prepositions, complex conjunctions and complex adverbs but it also encompasses all sorts of semantically and syntactically compositional simple and multiple clause constituents (cf. Altenberg 1998), sentence stems and "regular form-meaning pairings" (cf. Pawley and Syder 1983) which are typically used to organize the content (i.e. referential information) of a text or any type of discourse.

2.6. Conclusion

This chapter has presented a synthesis of the theoretical developments that will provide an analytical framework for our phraseology-oriented approach to EAP-specific vocabulary. This framework is presented in Figure 2.16. Although traditional typologies are often too restrictive, they are useful tools for the **description and categorization** of phrasemes. In this thesis, we adopt slightly revised versions of De Cock's (2003) and Burger's (1998) models to account for both the structure and function of a wide range of phrasemes from totally fixed idioms to syntactically and semantically compositional clause constituents.

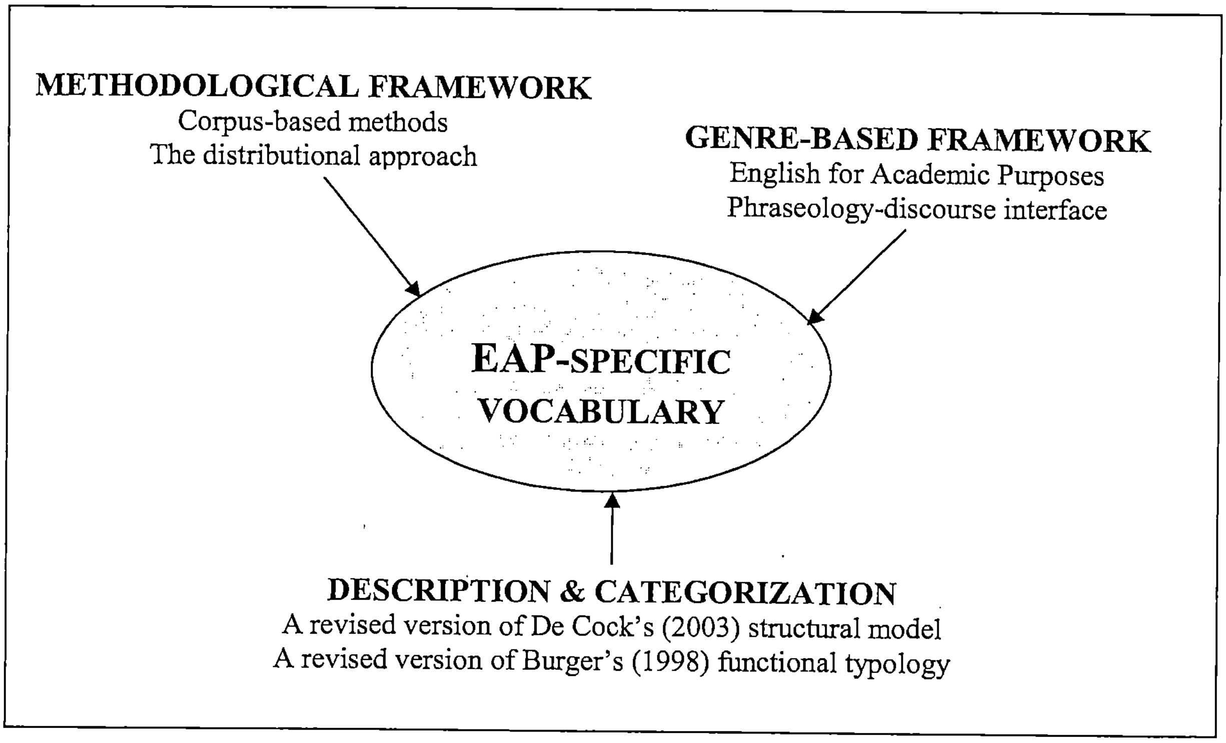
The structural and functional typologies adopted in this thesis can be described as typologies of English for General Purposes (EGP) phrasemes (cf. Gläser 1986). One of the main objectives of this thesis is to describe the phraseology of a language for specific purposes, viz. **English for Academic Purposes**, in professional and student writing, and more particularly the phraseology of EAP-specific words which have been shown to 'provide a semantic-pragmatic skeleton' for academic discourse (cf. section 1.3.2.2). It is most probable that academic words serve this semantico-pragmatic function in well-defined lexico-grammatical patterns and that a large proportion of these phrasemes will fall into the category of **textual phrasemes**.

The analytical framework adopted in this thesis also largely benefits from **genre-based studies of collocations** which contributed substantially to the establishment of a **phraseology-discourse interface** as discussed in section 2.5. These studies have shown that collocations are used to serve the most typical rhetorical or organizational functions of a specific academic text type or section and have documented their pervasiveness in academic discourse.

The methodological framework adopted in this thesis will be described in detail in section 4.2.2 and is based on **corpus-based methods** as employed in the **distributional approach** to collocations. Due to the various meanings of the term 'collocation', however, phraseological studies often suffer from terminological vagueness, if not inconsistency, and are hardly comparable. We have thus proposed in section 2.4.3 to reserve the term **collocation** for arbitrarily restricted word combinations and to use the term **(significant) co-occurrence** to refer to the probabilistic phenomenon described within the distributional approach to collocation. More generally, we have argued for a clear distinction between a **method of**

extraction of word combinations from corpora and a **linguistic analysis** of the results thereof by distinguishing between statistically defined or frequency-based terms. and linguistics terms.

Figure 2.16: A phraseological approach to EAP-specific vocabulary


METHODOLOGICAL FRAMEWORK
Corpus-based methods
The distributional approach
GENRE-BASED FRAMEWORK
English for Academic Purposes
Phraseology-discourse interface
EAP-SPECIFIC VOCABULARY
DESCRIPTION & CATEGORIZATION
A revised version of De Cock's (2003) structural model
A revised version of Burger's (1998) functional typology

3. Transfer, lexis and methodology

3.1. Introduction

This chapter is not intended as yet another PhD chapter devoted to transfer, its history, its numerous definitions, its manifestations, its specificities in multilingual settings, etc. The reader is referred to Odlin (1989), Ellis (1994:299-345) and Odlin (2003) for excellent syntheses[68]. Rather, it proposes to meet two objectives. First, it seeks to review major findings about the influence of the first language on EFL learners' use of words and phrasemes in L2. Second, it aims to scrutinize the methodologies used in transfer studies, highlighting their strengths and limitations.

Odlin's (1989) much-cited definition of 'transfer' is used as a working definition in this thesis although it will be seen in chapter 7 that it is hardly operationalizable:

> "Transfer is the influence resulting from similarities and differences between the target language and any other language that has been previously (and perhaps imperfectly) acquired." (Odlin 1989:27)

Transfer is a general cover term for a number of different kinds of influence from previously acquired languages (cf. Ellis 1994:341). However, only transfer from the first language will be considered in this chapter (see Cenoz et al 2001 for a review of transfer effects of previously acquired languages other than the mother tongue on a third language). The focus will also be restricted to L1 influence on English as a Foreign Language (EFL).

3.2. L1 influence on IL words and phrasemes

This section first deals with L1 influence on interlanguage[69] (IL) lexis and reviews major findings about borrowing and lexical transfer. It then focuses on available findings about transfer effects on IL referential, textual and communicative phrasemes.

[68] See also Singleton (1987a) for a discussion of the history of transfer.
[69] See note 16 for a definition of 'interlanguage'.

3.2.1. Transfer effects on IL lexis

Ringbom (1987:115) describes the full range of effects that previously acquired languages may have on IL lexis as a continuum, whose end points involve slightly different underlying processes. The overt manifestations of these two different processes in production are subsumed under the terms of "borrowing" and "lexical transfer" respectively.

3.2.1.1. Borrowing

As Figure 3.1 shows, the purest form of 'borrowing' is **complete language shift**, viz. the use of L1 material in an unmodified form in (spoken) interlanguage. This phenomenon has often been referred to as 'code switching' (e.g. Poulisse and Bongaerts 1994; Söderberg Arnfast and Normann Jørgensen 2003). Unlike in complete language shifts, activated lexical items from L1 or some other language are modified morphologically or phonologically by L2-procedures in hybrids, blends and relexifications. **Hybrids** are forms consisting of morphemes from different languages. For example, hybrids produced by Swedish-speaking learners of English typically consist of Swedish words to which an English-bound morpheme has been added:

3.1. She ***fylls*** 50 year (Sw. ***fylla*** 50 = "have one's fiftieth birthday")
3.2. All these wooden ***golves*** must be cleaned (Sw. ***golv*** = 'floor') (examples from Ringbom 1987:153-154)

By contrast, **blends** involve forms of the target language to which material from a previously acquired language has been added. The following are examples of blends produced by Swedish-speaking learners of English, where a Swedish ending is added to what is otherwise an English word:

3.3. If I found gold, I would be ***luckly*** (Sw. ***lycklig*** = "happy") (example from Ringbom 1987:153-154)

Blends and hybrids are not always easily distinguished. In fact, Ringbom does not make a distinction between these two types of borrowing when he lists the lexical errors made by EFL learners (ibid: 153-154). Other researchers often use the term 'coinage' to refer to both hybrids and blends (e.g. Gabryś-Barker 2006). In **relexifications**, lexical items from L1 or some other language are modified phonologically to fit what learners perceive as norms in the target language. Thus, Swedish learners may produce the verb *spride* from Sw. *sprida* (En. "spread") by analogy with recurring correspondences between English and Swedish words (e.g. Sw. *glida* – Eng. *glide*; Sw. *rida* – En. *ride*). Another example of a relexification is *fale* in the following sentence:

> 3.4. I don't believe it's your ***fale*** that you have put the cheque in wrong envelope" (Sw. ***fel*** = "fault")

In summary, **borrowing** "involves not only the 'online' consultation of a lexicon or of lexicons other than that of the language in which communication is taking place, but also the use of lexical knowledge relative to this latter language in order to 'camouflage' the alienness of the borrowed items" (Singleton 1999:181)

3.2.1.2. Lexical transfer

Lexical transfer differs from borrowing in that learners assume "an identity of semantic structure" between lexical items in a previously acquired language and lexical items in the target language. Figure 3.1 shows that lexical transfer may typically result in loan translations and semantic extension. Ringbom (1987:116) regards cognates as an intermediate category between transfer and borrowing. In this thesis, however, cognates are classified as manifestations of lexical transfer on the basis that the underlying process which leads to learners' use of cognates in L2 clearly involves the assumption of formal and semantic cross-linguistic similarity.

3.2.1.2.1. Loan translations

Loan translations are typically compound words or phrases in the target language resulting from the literal translation of each element of compounds from another language. In Ringbom's study, for example, a Finnish-speaking learner of English used *fire sticks* (based on Finnish *tulitikut* = *tuli* 'fire' and *tikku* 'splinter') instead of *matches*. Other examples include *green things* (Danish *grøntsager* = 'vegetables') (cf. Faerch and Kasper 1986:50) and *home animals* (from the Finnish word for domestic animals) (cf. Ringbom 1986:158). Over-extension of rules such as English prefix *un-* corresponds to French *in-* or English suffix *–less* equals Dansih *–løs* may also be interpreted as loan translations, e.g. **employless* (Danish arbejdløs) (cf. Faerch and Kasper 1986).

Figure 3.1: Overt cross-linguistic influence in production (based on Ringbom 1987:117)

LEXICAL TRANSFER				BORROWING
Loan translations	**Semantic extension**	**Cognates (as seen in false friends)**	**Hybrids, blends and relexifications**	**Complete language shift**
Semantic properties of one item transferred in a combination of lexical items:	Semantic properties extended to L2-word:	Formal cross-linguistic similarity between items with varying semantic relationships:	Morphological or phonological modification of item accommodated to L2-norm:	No modification of item accommodated to L2-norm
"*child wagon*" for "pram" (Sw. *barnvagn*)	"He bit himself in the *language*" (Fi. *kieli* = both "tongue" and "language")	(a) Wholly different meaning: "At the time he works in a *fabric*" (Sw. *fabric* = "factory") (b) Similar, but in no context identical meaning: "The next day we *grounded* a club" (Sw. *grunda* = "found") (c) In some, but not all contexts identical or near-identical meaning: "The *hound* is the best friend of man" (Sw. *hund* = "dog", occasionally also "hound")	"In the morning I was tired and in the evening I was piggy" (Sw. pigg = "refreshed"); He is good at mathematics but he succes [sic] in the other *amnys*, too." (Sw. *ämne* = "subject")	"I am usually very *pigg* after the diet." (Sw. *pigg* = "refreshed")

3.2.1.2.2. Semantic extension

Another type of lexical transfer involves **semantic extension**. Ringbom illustrates this phenomenon with a Finnish learner's use of *language* instead of *tongue* in the following sentence:

> 3.5. He bit himself in the ***language*** (Fi. kieli = both 'language' and 'tongue').

The procedure involved here has been called 'under-differentiation'[70]: the learner uses a previously acquired word in the target language, extending its meaning to include all the semantic properties of its L1 equivalent. Another example of semantic extension is provided by Agustín Llach et al (2006) who report that a Spanish learner of English used *fathers* for *parents*. As shown in Figure 3.2, the singular form of the Spanish noun '*padre'* corresponds to English '*father'*. When used in plural, the Spanish word has two translation equivalents in English.

Figure 3.2: Sp. *padre* vs. En. *father*

Spanish	English
'padre'	'father'
'padres'	'parents'
	'fathers'

Ringbom (1986:154) suggests that wherever possible beginning foreign language learners try to operate with simplified translation equivalents. They often acquire one of the equivalents before the others and make use of this 'primary counterpart' (Arabski 1979)[71] in both appropriate and inappropriate contexts. As Hasselgren (1994)[72] puts it,

> "when an L1 item is translatable by two or more L2 items, the learner will often select one of these and consistently let it do the job of both or all of them, spreading its area of meaning to cover the semantic space of the L1 source item" (Hasselgren 1994:251).

[70] See Weinreich's (1968: 18-19) classification of non-correspondence types between L1 and L2 items into (1) item substitution, (2) underdifferentiation, and (3) overdifferentiation.

[71] Quoted in Kellerman (1984), Ringbom (1987) and Selinker (1992): Arabski J. (1979) *Errors as indicators of the development of interlanguage*. Katowice: Universytet Slaski.

[72] Hasselgren (1994) refers to the phenomenon of 'semantic extension' as 'spreading'.

Viberg has studied phenomena of semantic extension in learner language in a number of studies (e.g. Viberg 1998; 2002). Like other researchers such as Jiang (e.g. Jiang 2004a; 2004b), Viberg refers to these phenomena as instances of **semantic transfer**. In his 1998 study, Viberg makes use of an elicitation task, i.e. an instruction-giving test[73], to investigate Spanish, Finnish and Polish-speaking learners' use of Swedish verbs of placement: *sätta, ställa* and *lägga*. These three verbs are primary equivalents of the English verb *put* but they are used in different contexts:

> - The verb stalla must be used when an object is placed in such as way that the vertical dimension is dominant:
>
> Jag ställer vasen på bordet. I'll put the vase on the table.
>
> - When the vertical dimension is not salient, the verb lägga must be used:
>
> Jag lägger paraplyet på bordet. I'll put (lay) the umbrella on the table.
>
> - The verb sätta is basically used when something is placed in a fixed position (e.g. attached in a file, AmE ring binder):
>
> Jag sätter in räkningarna i pärmen. I'll put the bills in the file. (Viberg 1998:184-185).

Figure 3.3 shows that Spanish and Finnish have a single equivalent to Swedish *sätta, ställa* and *lägga*. By contrast, the system of verbs of placement is partly equivalent in Polish. There is a partial equivalent to Swedish *ställa*, namely *stawiać* which shares the semantic feature + VERTICAL. Polish *kłaść* corresponds to Swedish *lägga* except that the feature –VERTICAL is not obligatory. This verb is semantically unmarked and can replace *stawiać* if the vertical dimension is not a salient feature of a situation. There is no Polish equivalent to Swedish *sätta*.

Figure 3.3: The principal verbs of putting in Swedish, Polish, Spanish and Finnish (Viberg 1998:187)

Semantic features	Swedish	Polish	Spanish	Finnish
- VERTICAL	lägga	kłaść	poner	panna
+ VERTICAL	ställa	stawiać		
FIXED	sätta			

Results indicate that learners tend to neutralize the semantic distinctions between the Swedish verbs of putting *sätta, ställa* and *lägga*. Neutralization of Swedish-specific semantic contrasts is strongly related to the type of semantic differentiation available in the L1. Generalization of

[73] "The test was organized around a video film showing a number of actions. (...) Most of the actions are intended to be illustrative of one of the Swedish verbs of putting. The learner serving as an informant was first asked to pick a fellow student to interact with. The informant was then shown the actions on the video one by one and assigned the task of instructing the fellow student, who was situated where the video could not be seen, to carry out the same actions." (Viberg 1998:187-189)

one verb of putting is found primarily among Spanish and Finnish learners who lack a corresponding L1 contrast. Spanish and Finnish-speaking learners typically over-extend the meaning of *ställa*. By contrast, Polish learners have an excellent command of the semantic field of this verb, which reveals a strong positive influence from the L1.

Jiang (2002) makes use of semantic judgment tasks to investigate semantic transfer and shed some light on the organisation of the bilingual lexicon[74]. Chinese-speaking learners of English were asked to decide whether two English words were related in meaning. Two types of related word pairs served as input: English word pairs sharing the same Chinese translation (e.g. *chance* and *opportunity* translate into the same Chinese word) and English word pairs that did not share the same L1 translation (e.g. *achievement* and *success* translate into two different words in Chinese). Jiang (2004b) is a replication of this study in which Korean-speaking learners of English were asked to perform the same task. In both studies, learners were found to respond to the same-translation pairs significantly faster than to the different-translation pairs, which Jiang interprets as evidence of continued L1 semantic mediation in L2 processing among L2 learners.

Translation equivalence, i.e. "the relation that holds between words which are regularly used as translations of each other and are presented as such in bilingual dictionaries" (Van Roey 1990:73), has been the object of numerous contrastive studies (e.g. Van Roey 1990; Salkie 2002; Viberg 2004/2005; Mudraya et al. 2005). Several studies have stressed learners' difficulties with partial translation equivalence but they have rarely systematically relied on interlanguage data to substantiate their claims. Thus, contributors to Swan and Smith (2001) report numerous English word-pairs confusion that can be attributed to a lack of semantic differentiation in the mother tongue. For example, Italian learners are said to use *why* and *because* (both rendered in Italian by 'perchè') or *also* and *even* (both rendered in Italian by 'anche') interchangeably and Malay/Indonesian speakers are reported to confuse *open* and *start* as well as *follow* and *accompany*. These statements might represent attested learner difficulties. However, the lack of detail about methodology and data used is a serious weakness of the book (cf. section 3.3 for a discussion of methodological issues in transfer studies).

[74] The reader is referred to Duyck (2004), Costa (2005) and Kroll and Tokowicz (2005) for a review of current research on the organisation of the bilingual lexicon.

Ringbom's category of semantic extension includes all instances of L1 semantic properties extended to L2 words which are not formal translation equivalents. Formal cross-linguistic similarities between items with varying degrees of semantic identity fall into the category of cognates.

3.2.1.2.3. Cognates

Figure 3.1 shows that Ringbom's category of cognates consists of false friends only, "where an underlying cross-linguistic similarity between words leads to errors" (Ringbom 1987:116). As the following examples reveal, Ringbom (1987) does not consider the criterion of etymological relatedness a necessary condition for cognates. This conception of 'cognates' thus broadly corresponds to Van Roey's (1990) use of 'common words'.

3.6. The child is ***locked*** to bed by telling him some stories (Sw. *locka* = "tempt")
3.7. Many people die every day because they are ***offers*** of the violence (Sw. *offer* = "victim")
3.8. I ***true*** that most of the teachers are good (Sw. *tro* = "think") (examples from Ringbom 1987: 124-125)

These examples help understand why Ringbom regards cognates as an intermediate category between transfer and borrowing. It is questionable, however, whether these "false friends" are not instances of borrowing, and more specifically hybrids (cf. section 3.2.1.1). It may be argued that *locked, offers* and *true* in examples 3.6 to 3.8 do not differ significantly from *fylls* and *golves* in examples 3.1 and 3.2. It may be pure coincidence that learners' attempts result in English forms in examples 3.6 to 3.8 but in non-English forms in examples 3.1 and 3.2. The word *true* in example 3.8 clearly supports this hypothesis. Ringbom (2006:38) explains that "[e]quivalence between individual items is difficult to perceive without an existing underlying functional equivalence between categories". In production, it seems even more implausible that learners assume an identity of semantic structure between L1-L2 word pairs that differ in their part-of-speech. The English word 'true' can be an adjective, an adverb or a noun but not a verb. However, no definite explanation can be provided for cases such as those illustrated above unless we resort to verbal reports in which learners "comment on their own productions and elicit thereby explanations for why they are or are not making errors" (Ellis and Barkhuizen 2005:22).

Following researchers such as Granger and Swallow (1988), Van Roey (1990) and Granger (1993), Ringbom (2007) makes use of a different definition of cognates:

> “Cognates in two languages can be defined as historically related, formally similar words, whose meanings may be identical, similar, partly different or, occasionally, even wholly different. Words with different meanings where the formal similarity is purely accidental, as in English *pain* – French *pain*, cannot be considered cognates.” (Ringbom 2007:73)

This definition differs from Ringbom’s (1987) in two ways. First, it makes use of the etymological criterion. Second, it is no longer restricted to false friends but also includes cognates whose meanings are identical. Granger (1993) categorizes cognates into ‘good cognates’, which have the same meaning, and ‘deceptive cognates’, which have partial or totally different meanings, as illustrated in Figure 3.4.

Figure 3.4: Granger’s (1993:44) categorization of cognates

GOOD COGNATES	(F) somptueux = (E) sumptuous (Lat. sumptuosus)
DECEPTIVE COGNATES	
Totally deceptive	(F) actuel ≠ (E) actual (Lat. actualis)
Partially deceptive	(F) expérience = (E) experience = (E) experiment (Lat. experientia)

Granger and Swallow (1988) examine the possible nature of the deceptiveness and explain that the difficulties inherent in deceptive cognates are not restricted to conceptual gaps:

> “Two cognates may have the same referential meaning and yet differ from a **collocative** point of view, showing a greater predilection for certain words or groups of words than for others, from a **connotative** point of view, that is to say in the associations which they call up, and from a **stylistic** point of view, in that they belong to different registers of language.” (Granger and Swallow 1988:112)[75]

They distinguish two forms of collocational restrictions depending on whether there is equivalence between the two cognates in a particular meaning, or whether the equivalence is limited to a number of arbitrarily restricted collocations (cf. Table 3.1). We will come back to transfer of L1 collocational restrictions on the target language in section 3.2.1.2.3.

[75] My emphasis.

Table 3.1: Cognates and collocational restrictions (based on Granger and Swallow 1988:112-114)

1. Equivalence between the two cognates in a particular meaning, expect in the case of certain restricted collocations in either or both of the languages, where the cognate is not used	**- maintenir / maintain** maintenir: conserver dans le même état ; faire ou laisser durer maintain : to continue or retain ; keep in existence Ex: *maintenir* des prix, coutumes, privilèges to *maintain* prices, customs, privileges But : *maintenir* une décision : to *stand by* a decision *maintenir* sa candidature: *not withdraw* one's application
	- faux / false faux: qui n'est pas vraiment, réellement ce qu'il paraît être false : not real or genuine but intended to seem real Ex: *fausse* barbe, *fausse* dents, *faux* plafond, *false* beard, *false* teeth, *false* ceiling, ... but: *fausse* fenêtre: *blind* window *false* bottom: *double* fond
2. Equivalence limited to a number of restricted collocations	**- assurance / assurance** Fr. assurance: convention par laquelle on s'assure = En. insurance Ex: police *d'assurance*: *insurance* policy *assurance* incendie: fire *insurance* but: assurance-vie: life *assurance* (also *insurance*)
	- tissu / tissue tissue: suite ininterrompue (de choses regrettables ou désagréables) = En. mass, string Ex : Son allocution n'était qu'un *tissu* de contradictions, d'obscénités His speech was a *mass* of contradictions, a *string* of obscenities but: un *tissue* de mensonges: a *tissue* of lies (or pack)

Examples of cognates that differ in their **connotative** meaning, i.e. "secondary features, either of a conceptual or emotive or evaluative nature, which form a kind of "halo associatif" around the word" (Van Roey 1990: 38), include:

- The pair En. *régime* – Fr. *regime* in which the English word has a pejorative connotation, "generally referring as it does to a system of government of which the speaker disapproves and not, like the French, to any type of government" (Granger and Swallow 1988:115);
- The pair En. *face* – Fr. *face*, where the French word usually connotes pejoratively, while its English counterpart is neutral.

Granger and Swallow (1988) suggest that French-English cognates often differ in their **stylistic** meaning, i.e. "the meaning of a word in so far as it is determined by the situation or the circumstances in which it is used" (Van Roey 1990: 42). Granger (1993) explains that many cognates which are core words in French are subject-core or non-core in English (cf. Table 3.2). As a result, French-speaking learners have a tendency to overuse less frequent English subject-core or non-core words that are directly activated by the French cognate.

Table 3.2: English-French cognates (Granger 1993:53)

FRENCH CORE	ENGLISH SUBJECT-CORE or NON-CORE	ENGLISH CORE
abandonner	*abandon*	*give up*
aider	*aid*	*help*
avare	*avaricious*	*mean*
courageux	*courageous*	*brave*
descendre	*descend*	*go/come down*
fatigue	*fatigue*	*tiredness*
liberté	*liberty*	*freedom*
monter	*mount*	*go/come up*
obtenir	*obtain*	*get*
profond	*profound*	*deep*

The role of cognates in foreign language teaching and learning has attracted much attention (e.g. Meara 1993; Granger 1993). As Carroll explains, cognates "present a certain paradox for learning theory: on the one hand, they appear to *facilitate* learning, i.e. unknown words which form cognate-pairs with known words appear to be easier to recognize, represent and deploy than new words which do not. On the other hand, cognate-pairing also appears to *hinder* long-term learning in that the so-called "false cognates" (les faux amis) cause erroneous production, may lead to misrepresentation of the meaning of the input, and may be difficult to overcome" (Carroll 1992:94)[76]. A number of studies have stressed the facilitative effect of cognates on **comprehension** skills (e.g. Ringbom 1987; Moss 1992; Jiménez et al 1996; Blonski Hardin 2001). As for **production**, Meara (1993:283) reports that several studies conducted in the 80s pointed to learners' tendency to avoid cognates. Other studies have shown that deceptive cognates may account for a large proportion of learners' lexical errors in their compositions or translations (e.g. Ringbom 1987:124-126; Hasselgren 1994). As Scarcella and Zimmerman (2005:127) comment, however, there are gaps in the literature concerning the effect of cognate knowledge on IL production. This situation is compounded by problems of comparability of results as the term 'cognate' is either used to refer to formal similarity (e.g. Carroll 1992; Singleton and Little 1991; Allerton and Wieser 2005) or etymologically related, formally similar words (e.g. Granger 1996b).

[76] My italics

Several factors have been reported to influence learners' use of cognates, among which, proficiency, frequency in L2 and individual differences. As for **proficiency**, Melka (1997: 97) states that "[a]s a beginner, the L2 learner has a tendency to generalize equivalences and to use this principle in comprehension and production. In later stages of learning, the same learner hesitates to produce cognates whose meaning he or she is not sure about. Reception will be maximal then, and production will be avoided". Hammer and Monod (1976:16) claim that "[i]ncorrect use of deceptive cognates is probably the least enduring type of interference between two languages"[77]. On the other hand, Granger (1993:49) writes that partially deceptive cognates "are, in the experience of many language teachers including myself, one the most enduring types of interference, giving rise to errors in the most advanced learning stages". Ringbom stresses the importance of **frequency** in L2 and states that "[h]igh-frequency deceptive cognates are easily confused at early stages of learning" (Ringbom 2007:76). Allerton and Wieser (2005:73) suggest that if the false friend item is more frequent than the correct translation equivalent, "the danger of it being used as a false friend seems to be greater". Thus, the English high-frequency word *small* may often be erroneously used to translate Ge. 'schmal' instead of its English translation equivalent *narrow*. **Individual differences** have also been suggested to play a part in learners' extensive use vs. non-use of cognates (cf. Meara 1993: 286-287).

There are many unsubstantiated, and often contradictory, claims with regard to the role of cognates (especially in production) and factors influencing their use. Studies that systematically investigate learners' productive use of cognates are relatively few. Granger (1996b), Scarcella and Zimmerman (2005) and Allerton and Wieser (2005) are three notable exceptions. Granger (1996b) makes use of a corpus of French-speaking learner essays to test the following two hypotheses based on the claims of some of the researchers cited above:

I. French-speaking learners of English have a tendency to overuse Romance words to the detriment of Germanic words
II. The use of deceptive cognates is the least enduring type of interference (Granger 1996b:112)

Results contradict hypothesis I. French learners are not found to overuse Romance words when compared to native speakers of English. Neither do they overuse Romance lexical verbs. On the contrary, they overuse a handful of high frequency Germanic verbs, an overuse which has also been reported in the writing of other L1 learner populations and is thus

[77] Quoted in Granger (1996:111-112): Hammer P. and M. Monod (1976) *English-French Cognate Dictionary*. The University of Alberta, Edmonton.

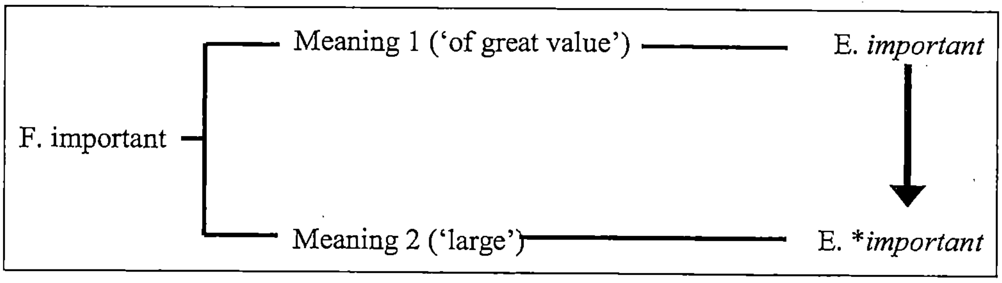
interpreted as a developmental feature. Similarly, Scarcella and Zimmerman (2005) investigate learners' use of what they call "academic words/cognates", i.e. Spanish-English cognates that belong to Coxhead's academic word list (cf. section 1.3.1.2), in expository essays written by Spanish-speaking and Asian language-speaking ESL students. They report that Spanish learners of English do not use more academic words/cognates than Asian students. In addition, Spanish students are also found to do worse on a test of academic word derivatives. Although Spanish students may have an advantage over learner populations whose L1 is not a romance language because many English words have Spanish cognates, this facilitating effect is not reflected in Scarcella and Zimmerman's findings.

Granger's (1996b) hypothesis II is also rejected: approximately one third of the 750 lexical errors analysed in this study are instances of false friends. However, results show a clear difference between totally and partially deceptive cognates. Although there are still some instances of totally deceptive cognates (e.g. *to achieve* – French 'achever' used instead of *to finish*; *to suppress* – Fr 'supprimer' used in the meaning of *to do away with*), partially deceptive cognates constitute the overwhelming majority of errors. Thus, in example 3.9, the adjective *important* is used in the meaning of *large*.

3.9. Their population is as ****important*** as the rest of Europe (E: large; F: important).

Granger (1996b:116) explains that French learners rely on the equivalence of Fr. *important* and En. *important* in the meaning of 'of great value' and mistakenly infer equivalence in the other meaning of 'large' as illustrated in Figure 3.5.

Figure 3.5: French learners' erroneous use of En. 'important' (Granger 1996b:116)



Allerton and Wieser (2005) describe different types of German-English false friends and make use of a translation task to investigate which of these actually cause more difficulties for German-speaking learners of English. Unlike in Granger (1996b), partial false friends do not seem to result in more errors. The number of correct answers is proportionally lower for the total false friends. The results of these two studies however are not comparable as they rely on two different definitions of a 'false friend' and make use of different tasks. Unlike Granger (1996b), Allerton and Wieser (2005) use the term to include any type of formal cross-linguistic similarity.

3.2.1.3. A restricted view of lexis and transfer effects

Ringbom's (1987) classification of lexical transfer effects is incomplete for at least two reasons. First, it deals almost exclusively with **single words** except for the category of loan translations. This category is very restricted in scope and consists solely of IL compound words or phrases resulting from the literal translation of each element of L1 compounds[78]. Second, it focuses exclusively on **lexical errors** despite its stated aim of representing 'overt cross-linguistic lexical influence in production'. Other manifestations of transfer have been identified in the literature, more particularly, facilitation (or positive transfer), avoidance (or underproduction), and overuse (cf. Ellis 1994: 301-306). However, Ringbom argues that in cases where, for example, lexical transfer leads the learner to a fully acceptable word, "a researcher can seldom establish that the use of a word has been the result of lexical transfer" (Ringbom 1987:115). This view will be challenged in chapter 7, in which we report the results of an investigation of transfer effects on overused phrasemes in essays produced by French-speaking learners of English.

In the next section, we will thus concentrate on transfer effects on phrasemes. Special emphasis will be placed on studies which have adopted a broader approach to transfer and investigated other L1 manifestations than errors.

3.2.2. Transfer effects on IL phraseology

This section focuses on findings about the influence of the first language on phrasemes. It first reviews findings about L1 influence on EFL learners' use of referential phrasemes. It then reviews the very restricted number of studies which have addressed the question of transfer effects on learners' use of textual and communicative phrasemes.

[78] Ringbom (2001) proposes a new classification of lexical transfer manifestations, in which the new category of 'calques' includes loan translations of compounds, phrasal verbs and idioms.

3.2.2.1. Transfer effects on referential phrasemes

Referential phrasemes – more particularly phrasal verbs, idioms and collocations - have clearly been the most extensively studied types of phrasemes. There are a number of studies which focus exclusively on, or devote much attention to, the influence of the first language on EFL learners' use of referential phrasemes. This section first reviews findings about transfer effects on learners' use of phrasal verbs. It then focuses on L1 influence on idioms in learner production and finally concentrates on L1-induced effects on learners' use of collocations. The numerous studies reviewed in this section are summarized in Table 3.3, which gives information on the type of phrasemes examined, the number of subjects investigated as well as their mother-tongue background and the task(s) used to answer the research question.

3.2.2.1.1. Phrasal verbs

Transfer effects have often been conceived of too narrowly, in terms of the 'transfer' of detectable features of a previously acquired language into another language (cf. Kellerman 1995). As Ellis explains, however, "the absence of a structural feature in the L1 may have as much impact on the L2 as the presence of a different feature" (Ellis 1994:311-312). Underuse or avoidance of English phrasal verbs has been reported for learners whose first languages do not have phrasal verbs (e.g. Dagut and Laufer 1985; Laufer and Eliasson 1993; Sjöholm 1998; Liao and Fukuya 2004). Dagut and Laufer (1985) report that Hebrew learners are more likely to choose one-word verbs where English speakers choose phrasal verbs that have the same meaning (e.g. *postpone* vs. *put off, reprimand* vs. *tell off*) and argue that the avoidance of phrasal verbs can be explained in terms of an indirect influence from the mother tongue, because the phrasal verb structure does not exist in Hebrew.

From Dagut and Laufer's (1985) conclusion, Hulstijn and Marchena (1989) draw the indirect implication that Dutch learners of English would not avoid phrasal verbs because these types of verb structures exist in Dutch. However, Hulstijn and Marchena show that even though Dutch learners do not avoid phrasal verbs as a category, they tend to avoid some figurative phrasal verbs that seem too Dutch-like (e.g. *go off, bring up, break out*). These results suggest that avoidance may not only result from L1-L2 structural differences but also from similarities that tend to be perceived as unlikely by Dutch learners. In addition, Dutch learners tend to "adopt a play-it-safe strategy, preferring one-word verbs with general, multi-purpose meanings over phrasal verbs with specific, sometimes idiomatic, meanings" (Hulstijn and Marchena 1989:241) (cf. section 3.2.2.2.4 for a discussion of Kellerman's studies of core vs. idiomatic meanings).

The two studies mentioned above are based on the assumption that avoidance presupposes some sort of prior knowledge of the target feature and a choice to rather use an alternative which is perceived as less difficult (cf. Kleinmann 1977). However, Kamimoto et al. (1974:259) argue that in both studies, "the methods used to establish this prior knowledge seem more hopeful than sound." Liao and Fukuga (2004) also criticize Dagut and Laufer's selection of the phrasal verbs investigated, which "depended on the researchers' impression from their teaching experiences". As a result, they argue that Hebrew learners' underproduction of phrasal verbs may just as well have resulted from their pure ignorance of the phrasal verbs (cf. section 3.3.2 for a discussion of the selection of items investigated in transfer studies).

Laufer and Eliasson (1993) investigate Swedish-learners use of phrasal verbs and compare their results with Dagut and Laufer's (1985) findings about Hebrew-speaking learners' avoidance of phrasal verbs. They find that, unlike Hebrew learners, Swedish learners do not underuse phrasal verbs, a verb structure which exists in Swedish. In addition, semantic opacity does not seem to induce learners' avoidance of congruent Swedish-English phrasal verbs. This result stands in sharp contrast with Hulstijn and Marchena's (1989) finding that Dutch learners tend to avoid 'too Dutch-like' English phrasal verbs. Laufer and Eliasson (1993) thus conclude that the best predictor of avoidance in their study is L1-L2 difference.

Sjöholm's (1998) objective is to distinguish between cross-linguistic influence and other factors such as proficiency, L2 exposure and semantic opaqueness on learners' avoidance of phrasal verbs. The author compares the use of phrasal verbs by Swedish- and Finnish-speaking EFL learners in order to investigate the role of cross-linguistic influence on the interlanguage of learners from two very different language backgrounds. As Ringbom explains, "[w]hile Swedish is a Germanic language with a vocabulary and structure very close to Norwegian and Danish, and fairly close to German, Finnish is a Fenno-Ugric language wholly unrelated to the Indo-European language family" (Ringbom 2006:37). In addition, the Swedish language system comprises constructions that are almost identical to English phrasal verbs while Finnish does not have such verb structures. Learners are grouped into four different proficiency level categories and two 'L2 exposure' groups (i.e. learners who had stayed in an English-speaking country less than ten months or more). The main conclusions of Sjöholm's (1998) study are:

1. Finnish learners of English tend to avoid idiomatic (opaque) phrasal verbs in the early stages of learning and avoidance is indirectly caused by L1.
2. Swedish learners of English display a slight tendency to avoid idiomatic (opaque) phrasal verbs that lack L1 counterparts.
3. Swedish learners tend to accept Swedish-based phrasal verbs in the early and advanced stages, but tend to avoid them in the intermediate stages (U-shaped curve). This tendency is more distinct with opaque phrasal verbs. This avoidance behaviour is believed to be due to indirect influence from L1.
4. The semantic feature opacity among Swedish-based phrasal verbs paired with L1-L2 similarity may cause avoidance in the intermediate stages among Swedes.
5. The semantic feature transparency among phrasal verbs combined with little exposure to the L2 (= young learners) may cause an initial over-use of these transparent phrasal verbs among Finnish learners.
6. Extensive exposure to natural input in the target language culture tends to increase the acceptance of opaque (idiomatic) phrasal verbs (especially among Finnish learners), but also tends to even out the differences in the choice pattern between the two language groups. (Sjöholm 1998:228-231)

The significant influence of proficiency, semantic opacity and L2 exposure on learners' use of English phrasal verbs is further documented in Liao and Fukuga (2004), who investigate Chinese learners' use of phrasal verbs. They find that intermediate learners produce fewer phrasal verbs than advanced learners, for whom "learning seems to have counteracted the effects of the L1-L2 difference" (Liao and Fukuga 2004:211). Their findings partially support the idea that L1-L2 differences are a good predictor of avoidance in L2 acquisition. They also report "a developmental manifestation of interlanguage from avoidance to nonavoidance" (ibid:212), noting that "the two notions of interlanguage development and L1-L2 structural difference (...) are not mutually exclusive or contradictory" (ibid:213). Incorporating the findings of Dagut and Laufer (1985), Hulstijn and Marchena (1989) and Laufer and Eliasson (1993), Liao and Fukuga (2004:212) even propose a model which "seems to suggest that, regardless of whether learners have phrasal verbs in their L1 (Dutch) or not (Chinese), they seem to go through the same developmental process from avoidance to nonavoidance of phrasal verbs". It is debatable, however, whether they do not oversimplify and overinterpret findings here.

Table 3.3: A selected list of studies of EFL learners' use of referential phrasemes including transfer-related claims or findings

Study	Types of word combinations	Subjects	Task
Kellerman (1977)	70 idioms and idiomatic sentences	Dutch learners	Correctness judgement task
Kellerman (1978)	The verb *break* in word combinations ranging from free combinations to idioms	a. 35 native speakers of Dutch b. 50 native speakers of Dutch c. 210 Dutch-speaking students and school-children d. Dutch university students - 50 first year - 31 third year	a. judgment of dissimilarity between pairs of meaning (36 sentences with the verb *break*) b. card sorting (17 sentences with *break*) c. transferability experiment I (9 sentence with *break*) d. transferability experiment II (17 sentences with *break*)
Kellerman (1986)	The noun *eye* in word combinations ranging from free combinations to idioms	35 Dutch first-year students of English	- translation test - similarity test - frequency test
Dagut and Laufer (1985)	phrasal verbs	Hebrew-speaking learners	- a multiple-choice test - a verb translation test - a verb-memorizing test
Irujo (1986)	45 idioms - 15 identical in form and meaning to their Spanish equivalents - 15 similar to their Spanish equivalents - 15 different from the corresponding Spanish idioms	12 Venezuelan advanced learners of English	COMPREHENSION - definition test - multiple-choice test PRODUCTION - discourse-completion (cloze) test - Spanish-English translation tests
Hulstijn and Marchena (1989)	phrasal verbs	Dutch-speaking learners	- a multiple-choice test - a verb translation test - a verb-memorizing test
Biskup (1992)	23 collocations (presumably verb – noun collocations and adjective-noun collocations)	very advanced EFL learners - 34 Polish university students - 28 German university students	- translation task (from L1 to English) Answers evaluated by three native speakers of English

Laufer and Eliasson (1993)	phrasal verbs	Swedish-speaking learners	- a multiple-choice test - a translation test
Irujo (1993)	45 idioms - 15 identical idioms - 15 similar idioms - 15 different idioms	12 bilingual speakers of Spanish and English	Translation task 45 paragraphs
Bahns and Eldaw (1993)	15 verb + noun collocations	58 German advanced learners	- German-English translation task - cloze task Evaluated by three native speakers of English
Chi et al. (1994)	collocational inappropriateness of delexical verbs (*have, make, take, do* and *get*)	first year Hong-Kong university students	- 1,000,000 words from the HKUST Learner Corpus - synchronic; written - assignments (reports + letters)
Hasselgren (1994)	Task A: semantic extension & collocational context Task B: 8 intensifier adverbs or adjectives Task C: 4 verb- noun combinations	Norwegian students of English (first year university and upper sixth-form) A control group of native English students 50-60 participants per L1 group	Second study: Task A: fill-in-the-blank translation task (8 sentences) Task B: fill-in-the-blank task (8 sentences) Task C: fill-in-the-blank task (4 sentences)
Granger (1998b)	1.a & 1.b. –ly intensifying adverb + adjective collocations 2. formulae: sentence-builders	1.a. French EFL advanced learners 1. b. 56 French EFL learners 56 native-speakers of English	1.a. & 2: corpus data ICLE-FR: argumentative essays (251,318 words)[79] LOCNESS: essays produced by English native students[80] 1. b. word combining test (for 11 amplifiers) Control group (56 native speakers of English)

[79] Cf. section 4.1.1.
[80] Cf. section 4.1.2.2.

Sjöholm (1998)	phrasal verbs	Setting: Finland - 638 Finnish-speaking students - 608 Swedish-speaking students - Four levels of proficiency per L1 - One group of students who had stayed abroad for ten months or more per L1	Multiple-choice test (with each item containing two correct alternatives, a phrasal verb and a 'synonymous' one-word verb along with two distractor verbs) Same test given to 15 native speakers of English
Abdullah and Jackson (1999)	80 idioms	120 advanced Syrian learners of English	COMPREHENSION - multiple-choice test - an English-into-Syrian Arabic translation test PRODUCTION - a Syrian Arabic-to-English translation test
Laufer (2000)	20 idioms	56 Hebrew-speaking university students	Test 1: fill-in translation task Test 2: passive knowledge test
Huang (2001)	40 word combinations (10 free combinations, 10 restricted collocations, 10 figurative idioms, 10 pure idioms)	60 Taiwanese students	fill-in-the-blank test
Altenberg and Granger (2001)	verb-noun combinations involving *make*	French EFL learners Swedish EFL learners English-speaking native students	Corpus data: - ICLE-FRENCH: 169,190 words; 285 essays - ICLE-SWEDISH: 169, 608 words; 296 essays - LOCNESS (native English): 168,325 words; 207 essays
Mahmoud Mohammed (2002)	idioms	230 Arabic-speaking second-year university students	Essays and term papers read to collect 124 (correct and incorrect) idioms
Nesselhauf (2003a)	1072 verb – object noun combinations (846 free combinations, 213 restricted collocations and 13 idioms)	advanced German EFL learners	Corpus data: - 32 essays from the German sub-corpus of ICLE (control: use of dictionaries, combined with some corpus analysis and some native speaker judgments)
Nesselhauf (2003b)	Locutional combinations or collocations: noun + verb combinations	- German EFL learners - French EFL learners + control group: native English	Corpus data: - German sub-corpus of ICLE: 154,191 words; 318 essays - French sub-corpus of ICLE: 154,008 words; 281 essays

	with the verbs *make* and *take*	speakers	- LOCNESS (native English): 154,886 words; 252 essays
Bulut (2004)	20 idioms	18 Turkish teachers of English	Idiom-recognition test
Liao and Fukuga (2004)	phrasal verbs	70 Chinese learners of English - 30 advanced learners of English (graduate students) - 40 intermediate learners of English (10 graduate students & 30 college students) Not all students took the three tests	multiple-choice, translation, and recall test 15 pairs of phrasal verbs / one-word verbs
Koya (2005)	68 verb – noun collocations	130 Japanese students (4 proficiency level groups)	Task A: vocabulary size test Task B: fill-in-the-blank translation task Task C: multiple choice task
Mahmoud (2005)	collocations	Arabic-speaking university students (third-year)	42 essays - 420 collocations - 20% grammatical collocations - 80% lexical collocations
Leśniewska (2006)	16 collocations adjective intensification	- 113 advanced Polish learners of English - 61 native speakers of English	a. fill-in-the-blank task (10 minutes) b. acceptability and saliency judgement task (5 minutes)
Piasecka (2006)	23 idioms	Polish learners of English - 59 university students - 31 teacher training college students of English	- contextualized fill-in-the-blank translation task - decontextualized discrete items translation task

3.2.2.1.2. Idioms

To the writer's knowledge, there is no study which investigates L1 influence on idioms as they are defined in section 2.3.7. The definition adopted in this thesis is reprinted here for the reader's convenience:

> The category of **idioms** is restricted to phrasemes that are constructed around a verbal nucleus and which are characterized by their semantic non-compositionality. Their semantic non-compositionality can be the result of a metaphorical process. Lack of flexibility and marked syntax are further indications of their idiomatic status. Examples include *to spill the beans, to let the cat out of the bag, to bark up the wrong tree* and *to kick the bucket.*

Studies which have analyzed EFL learners' use of 'idioms' have generally given a much broader definition to the term. Mahmoud Mohammed (2002), for example, defines an idiom as "a group of words which, as a whole, has a different meaning from the meaning of the individual words it contains". Thus, his category of idioms not only includes idioms as defined in this thesis but also compounds (e.g. *a man of straw*), complex prepositions (e.g. *in line with*), idiomatic sentences and commonplaces (e.g. *silence is golden*). The same holds true for Kellerman (1977) who examines sentences including various types of phrasemes such as *dyed-in-the-wool, take the bull by the horns, behind her back, victory in the bag* and *in cold blood* and Irujo (1986) who analyzes phrasemes as different as *point of view, to have on the tip of my tongue, to break the ice, a vicious circle, the black sheep of the family, the coast is clear* and *what's eating him?* Laufer (2000) does not define 'idioms' but she seems to adopt a more restricted definition that covers compounds and idioms as we defined them in section 2.3.7.

Idioms, in the sense of the word adopted by researchers such as Irujo (1986) and Mahmoud Mohammed (2002), have often been described as language-specific items "that are generally not transferred, even if it would be possible to do so and produce correct TL [target language]" (Kellerman 1977:101-102). In a correctness judgement task, Kellerman (1977) found that first-year English students were more likely to reject 'Dutch-like English' idioms as incorrect, thus indicating a reluctance to transfer them to English. Third-year students, by contrast, were found to be more successful at distinguishing correct English idioms similar to Dutch ones, thus displaying some acquired knowledge of what is possible in English. However, they were also reported to show "a tendency (admittedly slight) to be more generous towards erroneous Dutch-based expressions" (ibid: 126).

A number of empirical studies support Kellerman's (1977) findings. Abdullah and Jackson (1999) report that, in a Syrian Arabic-to-English translation test, some of the Syrian participants avoided giving the identical English equivalent to the Syrian idiom, "assuming that Syrian Arabic idioms could not have the same meanings in identical English forms" (Abdullah and Jackson 1999:96). Similarly, Bulut (2004) finds that, while processing idioms, even when these idioms have a perfect match in the mother tongue, Turkish teachers of English tend to rely on contextual cues, treating those cognate idioms as 'false friends'.

Other studies, by contrast, have shown that avoidance of idioms is not "a uniform phenomenon". Idioms are not avoided as a category (cf. Laufer 2000). **Identical or cognate idioms**, i.e. L2 idioms which have exact L1 translation equivalents, have been shown to be the easiest to comprehend and produce (e.g. Irujo 1986; Irujo 1993). **Similar or 'false cognate' idioms**, i.e. L2 idioms which have partial translation equivalents (e.g. En. 'to kill two birds with one stone' – Sp. 'matar dos pájaros de un tiro' [= to kill two birds from one shot]) appear to be the least used idioms together with **non-idioms in L1**, i.e. idioms which do not have an L1 counterpart (e.g. Hebrew does not have an idiomatic expression corresponding to English 'it's not my cup of tea') (cf. Laufer 2000). However, similar idioms also seem to provide the most opportunity for (negative) transfer effects (e.g. Irujo 1986; Irujo 1993; Abdullah and Jackson 1999; Laufer 2000; Mahmoud Mohammed 2002). Almost all incorrect idioms produced by Venezuelan EFL learners are due to literal translation of their Spanish idiom counterparts (cf. Irujo 1993). Similarly, L1 influence is held responsible for Arabic-speaking learners' production of lexico-grammatical errors in otherwise similar idioms, e.g. '**a drop in **an** ocean*' instead of '*a drop in **the** ocean*' (cf. Mahmoud 2002)[81]. Transfer effects are also found in learners' comprehension of similar idioms. For example, Syrian learners have been shown to interpret the English idiom *to bite the dust* as similar in meaning to the Arabic idiom 'to eat the dust', which means 'to be very poor/hungry' (cf. Abdullah and Jackson 1999:98).

Different idioms, i.e. formally totally different idioms in the two languages which express the same meaning and thus function as translation equivalents, are often the most difficult to understand and produce but are less susceptible to transfer (e.g. Abdullah Jackson 1999; Irujo 1986). As Irujo suggests, "[w]hen differences are slight, the tendency may be to generalize and ignore those differences. When differences are so great that two forms have

[81] Idioms which differ in article usage only are classified as identical idioms by Irujo (1986). Perhaps surprisingly, Irujo does not report problems with idioms which differ in article usage in the translations produced by Venezuelan learners.

nothing in common, there would be no reason to try to use one form to produce the other, so little transfer would occur" (Irujo 1986:298). All these findings have been interpreted as evidence of L1 use in the processing of idioms.

Very few studies have investigated the influence of task and learner variables (cf. section 4.1.1) on learners' use of idioms. As for the role of proficiency, results seem to point in opposite directions. Kellerman (1977) reports first-year students' general tendency to reject Dutch-like English idioms. By contrast, Irujo (1993) examines the translations of idioms produced by advanced Spanish learners of English and compares them with Irujo's (1986) findings for less advanced Spanish learners. She shows that there is more interference in the interlanguage of less advanced Spanish learners. A number of factors may influence these contradictory results. First, the tasks used are very different. Second, proficiency levels may not be comparable across the two studies. As a result, it is not possible to distinguish between proficiency and task effects. A third variable is the mother tongue, which may play a part in learners' perceptions of cross-linguistic similarities (cf. section 3.2.3). Other variables have hardly been examined. Abdulllah and Jackson (1999) comment that transfer with cognate and false cognate idioms is especially likely "when the L2 idiom is learnt out of context and in the first-language environment" (Abdullah and Jackson 1999:105). The use of the first language in producing English idioms seems to very by individual (cf. Irujo 1986), which may be explained by personality or cognitive-style relating to risk-taking (Abdullah and Jackson 1999: 97).

3.2.2.1.3. Collocations

Collocations have attracted much attention from teachers and researchers alike. The available findings about EFL learners' use of English collocations indicate that these word combinations cause them serious difficulties (e.g. Bahns and Eldaw 1993; Chi et al 1994; Lennon 1996; Howarth 1996; Howarth 1998; Källkvist 1998; Lorenz 1999a; Kaszubski 2000; Nesselhauf 2005). They also generally support the view that partly restricted collocations are the most problematic ones in L2 production[82] and the most likely to reflect L1 transfer effects. Nesselhauf (2003a) conducts a careful analysis of verb-noun combinations, i.e. free combinations, (restricted) collocations and idioms, in essays written by German learners of English and investigates how many of the wrong or questionable combinations are likely to

[82] Huang (2001) is an exception, considering pure idioms as the most challenging type of word combinations to EFL learners.

have been influenced by the first language. She shows that "whereas the influence of the L1 on verb-noun combination mistakes is considerable in general, it is greatest in collocations" (Nesselhauf 2003a:235). As Table 3.4 shows, L1 influence may account for 45% of mistakes in general and for 56% of the mistaken collocations. L1-induced errors include verb mistake (e.g. **make homework* instead of *do homework*; German *Hausaufgaben machen*), noun mistake (e.g. **close lacks* instead of *close gaps*; German *Lücken schliessen*), usage mistake or what Ringbom (1991) refers to as 'loan translation' or 'calque' (e.g. **train one's muscles* instead of *to exercise*; German *seine Muskeln trainieren*), preposition mistake (e.g. **draw a picture from* instead of *draw a picture of*; German *ein Bild zeichnen von*; both *of* and *from* frequently correspond to German *von*) and article mistake (e.g. **get the permission* instead of *get permission*; German *die Erlaubnis bekommen*) (cf. Nesselhauf 2003a:235).

Table 3.4: L1 influence on mistakes and questionable combinations (adapted from Nesselhauf 2003a:235)

	Free combinations	**Restricted collocations**	**Idioms**	**Total**
Number of mistakes and questionable combinations	220	59	4	283
L1 influence likely	92	33	2	127
Percentage	**42%**	**56%**	**50%**	**45%**

Similarly, Biskup (1992) examines the use of collocations by German and Polish EFL learners and finds L1 influence to be an important factor in learners' production of deviant collocations. In accordance with the findings about 'semantic extension' reported in section 3.2.2.2, Biskup reports that interference errors are particularly frequent when the semantic field of a given L1 lexical item is wider than the field covered by its corresponding L2 item. Chi et al. (1994) find transfer effects on the erroneous selection of delexical verbs and Mahmoud (2005) suggests that transfer is responsible for a number of misguided uses of prepositions in grammatical collocations.

Other researchers, by contrast, have tended to minimize the effects of L1 influence on collocational mistakes, arguing that "the real problem is that the speaker lacks knowledge of the verb which collocates with the following noun phrase" (Lennon 1996:28), or to report no such effect. Leśniewska (2006), for example, examines adjective intensifiers produced by Polish advanced learners of English and claims that, except for rare cases, "neither the deviant forms, the non-existent words, nor the inappropriate extensions of the collocability range made by the Polish learners display the influence of Polish" (Leśniewska 2006:69). Yet other studies have investigated learners' errors without taking the L1 variable into account.

Disregarding the mother tongue backgrounds of his subjects, Howarth (1996; 1998) interprets a large proportion of erroneous verb-noun combinations produced by nine teachers of English and one teacher of German from seven countries (Botswana, Germany, Greece, Hong Kong, Japan, Taiwan, and Thailand) as cases of direct confusion between two L2 delexical verbs (e.g. **do attempts* instead of *make attempts*). This is quite unfortunate as Nesselhauf (2003b) has shown that (exclusive) L1 influence is especially frequent with delexical verbs.

Errors may, however, often stem from both interlingual (i.e. L1-induced) and intralingual factors, i.e. "the intrinsic properties of the word which may affect its learnability, properties which are related to the word's form and meaning" (Laufer 1997:141). Nesselhauf (2003b) argues that "the L1 and L2 influences on mistakes interact in various ways. In a considerable number of mistakes produced, both L1 influence and L2 influence are discernable. Both types of influence are, for example, likely in *take charge of* (... for *take care of*), which seems to have been influenced by French *prendre en charge* as well as by the fact that the meaning of *take charge of* in English is similar to the intended one of *take care of*" (Nesselhauf 2003b:280).

The influence of the mother tongue on advanced EFL learners' production of collocations may also manifest itself in subtler ways. Learners have been found to rely heavily on those English collocations which are congruent with word combinations in their L1 (e.g. Granger 1998b; Nesselhauf 2003b). Thus, the few adverb-adjective collocations used by French learners in Granger's (1998b) study typically have a direct translation equivalent in French. Examples include *closely linked* (Fr. *étroitement lié*), *deeply rooted* (Fr. *profondément enraciné*), and *severely punished* (Fr. *sévèrement puni*). Conversely, non-congruency may at least partly explain learners' patterns of underuse. Thus, Altenberg and Granger (2001) state that "while the high degree of congruence between the English and the Swedish causative + adjective structures accounts for Swedish learners' overuse of this construction, the much more blurred correspondence between the English and French structures helps explain French learners' underuse" (2001:182).

Although a large proportion of the studies reviewed here have put forward L1 influence as the most likely explanation for a number of interlanguage features, very little is really known about the extent of transfer effects on EFL learners' use of collocations. Interlingual factors need to be studied in parallel with other factors that have been shown to influence EFL learners' use of word combinations, including lack of knowledge (e.g. Lennon 1996), learners' preference for 'core' items (cf. Hasselgren 1994; Granger 1998b) and intralingual factors (e.g. semantic opacity and delexical meaning) (cf. Koya 2005). The role of proficiency also needs to be investigated. Most studies of EFL learners' use of collocations have focused on the production of **advanced** learners, even though 'advanced' may represent very different proficiency levels across studies. One notable exception is Koya (2005) who groups Japanese EFL learners of English into four populations according to their score at a vocabulary test: Group A (2000 word level), Group B (3000 word level), Group C (4000 word level) and Group D (5000 word level). Koya reports a gradual increase of L1 transfer in collocations produced by learners from Group A to Group C and a slight decline at the 5000 vocabulary level, thus suggesting that some knowledge has been acquired.

3.2.2.2. Transfer effects on communicative and textual phrasemes

Communicative phrasemes include routine formulae, attitudinal formulae, proverbs and proverb fragments, commonplaces, idiomatic sentences, etc (section 2.3.7). There is no study, to the writer's knowledge, which examines L1 influence on proverbs, commonplaces and idiomatic sentences. Kellerman (1977) however claims that these phrasemes are not transferable as they are largely language-specific[83].

The influence of L1 **routine formulae**, i.e. phrasemes such as *Thank you* and *How do you do?*, on EFL learners' use of similar formulae in L2 has scarcely been investigated. On the other hand, there is a whole body of research which focuses on the transfer of L1 social rules of speaking, i.e. **pragmatic transfer**, and investigates politeness strategies, degrees of involvement and directness, etc. However, this field of study, with its general orientation towards content rather than form and its clear focus on speech acts, clearly lies outside the scope of this thesis. The reader is referred to Kasper (1992), Bou Franch (1998), Kasper and Rose (1999) and Charlebois (2004) for more information on pragmatic transfer.

Transfer effects on sentence stems which mainly function as **attitudinal formulae**, i.e. phrasemes used to signal speakers' attitudes towards their utterances and interlocutors (e.g. *to be honest; I think that* ...), have been highlighted in several studies. These studies do not make a distinction between communicative and textual sentence stems and the findings reported here also concern **textual formulae**. Granger (1998b) investigates French learners' use of what she refers to as 'sentence-builders', i.e. "phrases that function as macro-organizers in the text" (Granger 1998b:154). She examines two types of frames or sentence stems:

[83] Kellerman made the same point about idioms but it was shown in section 3.2.2.2.2 that L1 may play a part in learners' production of idioms in L2.

- **Passive frame**
  it + (modal) + passive verb (of saying / thinking) + *that*-clause
  Examples: *it is said / thought that ... ; it can be claimed / assumed that*
- **Active frame**
  I or *we/one/you* (generalized pronoun) + (modal) + active verb (of saying / thinking) + *that*-clause
  Examples: *I maintain / claim that ...; we can see / one could say that ....* (Granger 1998b:154)

Results show that "[w]hile the learners made a similar use of the passive structure to the native writers – both quantitatively and qualitatively – they massively overused the active structure" (ibid: 154-155). Granger suggests that this overuse may at least partly be explained in terms of L1 influence since "French uses many more phatic introductory phrases than English" (ibid: 155). Siepmann (2006:266-270) replicates Granger's (1998b) study of sentence builders in German EFL student writing and finds that German learners are much less inclined to overuse the active frame than French students, which makes the author conclude that "[i]t thus seems reasonable to assume that, in the present case, target-language behaviour is strongly influenced by first-language background" (Siepmann 2006:267). Siepmann also reports that German learners generally tend to "fight shy of structures which lack a 'direct' equivalent in their mother tongue" (ibid: 271).

Transfer effects on **sentence stems** have been most extensively studied by Neff and her colleagues. Neff et al (2004) show that Spanish learners use *we must* + reporting verb with an illocutionary force "which presents the writer as if forcing the reader to accept the following proposition which may seem slightly face-threatening" (Neff et al 2004:154). This use may be partly explained by a transfer of Spanish, in which the deontic *deber* can mean either *must* or *should.* Neff et al (2003) show that writers with an L1 Romance language rely heavily on clusters of *we* + modal verb + verb of mental/verbal process, which closely correspond to Romance languages' use of the first person plural to address readers (see also Neff 2006:66). As Neff van Aertselaer (to appear) explains, Spanish expert and novice writers' preference for passive structures in the present tense most probably reflects a transfer of Spanish discourse strategies, i.e. a translation of Spanish *se* impersonal passive phrases, e.g. *it is said* (Es. *se dice*). Neff and Bunce (to appear) further show that attempted literal translations of *se* impersonal passives phrases frequently result in grammatical errors in Spanish graduate students' academic writing as illustrated in the following examples:

3.10. *It has been introduced a new plan* ('Se ha introducido un proyecto nuevo')

3.11. *It will be observed the image*

Apart from the above studies, very few studies have investigated transfer-effects on EFL learners' use of **textual phrasemes**, i.e. phrasemes typically used to structure and organize the content (i.e. referential information) of a text or any type of discourse. Some studies on learners' use of **connectors** have commented on the potential L1 influence on learners' overuse or misuse of multi-word connectors or linking adverbials. Field and Yip (1992:25) suggest that Chinese learners' frequent erroneous use of *on the other hand* may be L1-induced. Granger and Tyson (1996) argue that French learners' overuse and misuse of *on the contrary* is probably due to an over-extension of the semantic properties of Fr. 'au contraire', which can be used to express both a concessive and antithetic link. Similarly, Gilquin (to appear) explains French learners' overuse of *even if* by the high degree of correspondence between this linking adverbial and its French counterpart *même si*. Siepmann (2006:257) also attributes German-speaking writers' overuse and misuse of *that is* to transfer effects.

The influence of the first language on text structure and organisation has typically been the focus of **contrastive rhetoric**, which Connor defines as "an area of research in L2 acquisition which identifies problems in writing by referring to the features of L1" (cf. Mieko 1997). However, contrastive rhetoric has generally been concerned with L1 influence at a macro-level of organisation, i.e. "in the writer's choice of rhetorical strategies and content" (Connor 2002:494) rather than in the writer's selection of words and phrasemes.

3.2.3. Constraints on transfer of phrasemes

This section is not intended as an exhaustive review of the factors that may promote or inhibit transfer (cf. Ellis 1994:315-332 for such a discussion). Rather, it focuses on the very few (and often contradictory) findings about constraints on transfer – more particularly, markedness and prototypicality, psychotypology and proficiency - that have emerged from studies focusing specifically on EFL learners' use of phrasemes.

As for markedness and prototypicality, in a number of studies based on acceptability tests and translation tasks, Kellerman (1977; 1978; 1979; 2000) suggests that L2 learners seem to work on the hypothesis that there are constraints on how similar the L2 can be to the L1, and these constraints seem to hold, even when the two languages are closely related and the structures congruent. Kellerman (1978) investigates the 'transferability' of the different meanings of the Dutch verb *breken* into its English cognate *break*. He shows that while Dutch learners of English accept the structures that are the least 'marked' in their mother tongue ('he broke his leg', 'the cup broke'), they tend to reject what they perceive as 'language-specific' items ('his voice broke when he was thirteen', 'some workers broke the strike'). **'Marked'** in this context means "semantically odd, or syntactically less producible or less frequent when compared with 'normal' forms" (Kellerman 1979:46). In the 2000 study, Kellerman expands on these findings and argues that the dimension of **'prototypicality'** largely determines Dutch learners' judgements about the transferability of the different usages of *breken* into *break.*

Although Kellerman acknowledges that learners' intuitions about what can be transferred in an L2 may not accurately reflect what they actually do when using the target language, his findings suggest that the further word combinations are situated from the central core of phraseology, i.e. semantically opaque, syntactically and collocationally inflexible multi-word units, the more potentially transferable they may be. This conclusion is challenged by Nesselhauf in a study of learners' multi-word combinations with the two verbs *take* and *make* in which she claims that "it does not seem to be the case that transfer decreases with the degree of idiomaticity of a combination [...] but rather that locutional combinations [restricted collocations] – at least in the case of the verb-noun combinations with the two verbs investigated – are the type of combination that is most susceptible to transfer" (Nesselhauf 2003b: 278). However, the author makes this claim on the basis of erroneous collocations only. For such a claim to be warranted, the author should arguably have also examined the potential L1 influence on native-like free combinations, collocations and idioms produced by EFL learners.

A related issue is that of **language distance** and **psychotypology**, i.e. learners' perceptions about language distance (cf. Ringbom 1987; Ringbom 2006). Comparing German and Polish learners, Biskup (1992) is able to observe that while German students tend to produce transfer errors resulting from assumed formal similarity, Polish students, "perceiving the distance between Polish and English, do not assume that there can be much formal similarity between these two languages" (Biskup 1992:91). Their errors reflect assumed semantic similarity instead and are either loan translations or extension of L2 meaning on the basis of the L1 word. Exactly the same types of findings are reported by Ringbom (1987) in his study of Swedish-speaking vs. Finnish-speaking EFL learners' use of lexis:

> "Thus the errors due to Swedish present exactly the opposite picture to the errors due to Finnish: Swedish influence results from formal cross-linguistic similarities between words, whereas Finnish influence manifests itself in either loan translations or, more commonly, in transfer of the semantic properties of a formally different L1-word." (Ringbom 1987:126)

The general conclusion that can be drawn from these findings is that while "the search for similarities is an essential process in [language] learning" (Ringbom 2006:5), it may not be conducted similarly by learners from different mother tongue backgrounds.

Textual phrasemes have been defined in section 2.3.7 as phrasemes "typically used to structure and organize the content (i.e. referential information) of a text or any type of discourse." The influence of the first language on text structure and organisation is typically the focus of **contrastive rhetoric**, which Connor defines as "an area of research in L2 acquisition which identifies problems in writing by referring to the features of L1" (cf. Mieko 1997). However, contrastive rhetoric has generally been concerned with L1 influence at a macro-level of organisation, i.e. "in the writer's choice of rhetorical strategies and content" (Connor 2002:494) rather than in the writer's selection of words and phrasemes.

Very few studies have investigated transfer-effects on EFL learners' use of textual phrasemes. Some studies on learners' use of **connectors** have commented on the potential L1 influence on learners' overuse or misuse of multi-word connectors or linking adverbials. Field and Yip (1992:25) suggest that Chinese learners' frequent erroneous use of *on the other hand* may be L1-induced. Granger and Tyson (1996) argue that French learners' overuse and misuse of *on the contrary* is probably due to an over-extension of the semantic properties of Fr. 'au contraire', which can be used to express both a concessive and antithetic link. Similarly, Gilquin (to appear) explains French learners' overuse of *even if* by the high degree of correspondence between this linking adverbial and its French counterpart *même si.* Siepmann (2006:257) also attributes German-speaking writers' overuse and misuse of *that is* to transfer effects.

Findings about L1 influence on **textual formulae** are, to the writer's knowledge, largely anecdotal. Siepmann, for example, reports that German learners generally tend to "fight shy of structures which lack a 'direct' equivalent in their mother tongue" (ibid: 271). Several studies have focused on sentence stems that are arguably best described as communicative phrasemes (e.g. *we can say that, we must ..., it can be said ...*) even though they may also be used to organize discourse. These studies will be reviewed in the next section.

3.2.4. United to unlock the mysteries of L1 influence

> "Learners clearly cannot be regarded as 'phraseologically virgin territory': they have a whole stock of prefabs in their mother tongue which will inevitably play a role – both positive and negative – in the acquisition of prefabs in L2" (Granger 1998:158)

This section has shown that there are numerous studies which devote attention to transfer effects on learners' use of single words and phrasemes in L2. However, it also supports Jarvis's (2000) view that "it is still unsettling to witness the level of confusion that remains in the field, particularly in the form of conflicting claims made about the nature of L1 influence and its interaction with other factors" (Jarvis 2000:248). Very few facts about L1 influence are well-established.

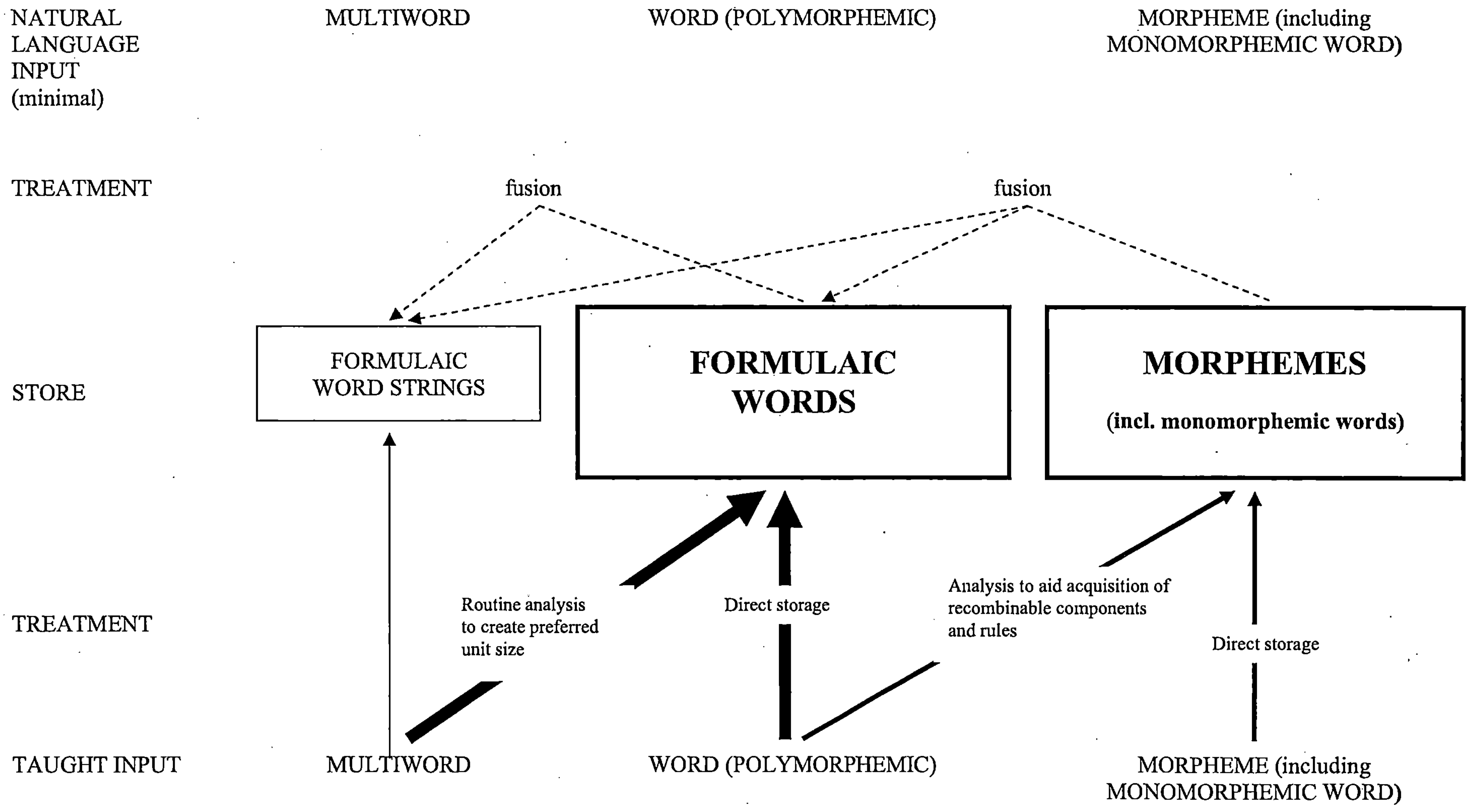
Most of the studies reviewed focus on what we can call 'transfer of form' (e.g. borrowing), 'transfer of meaning' (cf. semantic transfer, semantic extension) or 'transfer of form/meaning mapping' (e.g. cognates). Next to knowledge of form and meaning, however, knowing a word also involves knowing in what patterns, with what words, when, where and how to use it (cf. section 1.3.2.1). Research into learners' use of cognates has highlighted transfer effects on style and register (cf. Granger and Swallow 1988; Van Roey 1990; Granger 1996b). Studies focusing on learners' use of phrasemes have brought to light transfer effects on collocational restrictions and lexico-grammatical patterns. However, much remains to be done regarding **'transfer of use'**. L1 influence on collocations has never been investigated within a broader research framework which would encompass the stylistic and register-related aspects of collocations. Hoey (2005) argues that collocations are primed to occur in specific textual, generic and social contexts (cf. section 2.4.2.3) and that the "transfer of primings from earlier to later languages is (…) unavoidable" (Hoey 2005:183). Does this mean that learners will transfer L1 features into L2 within a specific register? Learner corpus research tends to suggest the contrary: EFL learners' writing is often described as too oral-like (cf. section 1.4.2) and their speech seems to be too formal (cf. Channell 1994:21; Guest 1998; De Cock 2003). There is arguably a need for studies which would compare L1 transfer across registers.

Wray (2002) proposes a model of the creation of the lexicon in classroom-taught L2 (after childhood) which is unique in considering the influence of the settings and types of L2 input on vocabulary development and in distinguishing between acquisitional patterns of morphemes, single words and formulaic word strings or phrasemes (cf. Figure 3.6). Despite the many uncertainties and controversies surrounding cross-linguistic influence, there is one thing that seems unquestionable from the transfer studies reviewed above, i.e. the influence of the first language on EFL learners' acquisition, comprehension and use of L2 lexis. It is therefore rather unfortunate that Wray (2002) does not incorporate the first language into her model.

Studies on the bilingual lexicon and bilingual processing, by contrast, have focused on the links between L1 and L2, and more specifically, on how we access words from different languages in our lexicon. As a result, they have been particularly interested in phenomena of code-switching or borrowing (cf. Poulisse and Bongaerts 1994; Poulisse 1997). However, the object of study of this research field largely remains the 'single word' despite De Bot's (2004) claim that "in work on bilingualism and multilingualism the structure and workings of the lexicon are at the heart of current research" (De Bot 2004:17). When the question of phrasemes is addressed, it is often to comment that they do not differ from single words and are treated equally in the mental lexicon. One example is De Bot (1992:10) who proposes a model of bilingual processing and writes without further consideration that "[i]t is assumed that idioms and phrases can be entries in the mental lexicon" (De Bot 1992:10). Evidence of the lack of interest in phraseology is also available from the widely held contention that syntagmatic responses on a word association test are indicative of a lower level of proficiency and that a shift from syntagmatic to paradigmatic responses suggest lexical development. This view has however recently been challenged by Wolter (2001; 2006) on the basis that it "underestimates the complicated process of acquiring syntagmatic connections" (Wolter 2006:746).

In short, it is regrettable that research fields as complementary as SLA, (learner) corpus research, teaching and psycholinguistics lead isolate existence and rarely interact. As a result, they can only provide partial theories of first language transfer, and incomplete models of the development of L2 vocabulary or the organization of the bilingual lexicon.

Figure 3.6: The creation of the lexicon in classroom-taught L2 (after childhood) (Wray 2002:208)


NATURAL LANGUAGE INPUT (minimal)
MULTIWORD
WORD (POLYMORPHEMIC)
MORPHEME (including MONOMORPHEMIC WORD)
TREATMENT
fusion
fusion
STORE
FORMULAIC WORD STRINGS
FORMULAIC WORDS
MORPHEMES
(incl. monomorphemic words)
TREATMENT
Routine analysis to create preferred unit size
Direct storage
Analysis to aid acquisition of recombinable components and rules
Direct storage
TAUGHT INPUT
MULTIWORD
WORD (POLYMORPHEMIC)
MORPHEME (including MONOMORPHEMIC WORD)

3.3. Investigating transfer: lost in 'methodological fog'[84]?

> "There has been considerable progress in the study of transfer during the last hundred or so years, especially during the years since World War II. Yet the controversies that have accompanied this progress make it clear that the findings of transfer research must be interpreted cautiously." (Odlin 1989:151)

In section 3.2, the focus has been placed on major findings about L1 influence on learners' use of words and phrasemes in L2. The numerous studies referred to have been discussed without describing in much detail the type of evidence and method used. Some methodological aspects, however, somehow cast doubt on the validity of several of the conclusions drawn. In this section, we focus on three central methodological factors which have proved to be vexed issues in transfer studies: the type and amount of data used to substantiate researchers' claim on L1 transfer, the selection of candidate items to serve as a basis for transfer studies (e.g. grammatical constructions, lexical items, morphemes) and the criteria used to assess L1 influence.

3.3.1. Data

Transfer studies, and more generally SLA studies, crucially depend on good datasets to work from (cf. Myles 2005:374). However, lack of rigour in data description and analysis has been identified as a serious weakness of the field (e.g. Jansen 2000). In this section, we discuss three limitations of a large proportion of transfer studies and the problems they may pose: (1) lack of detailed data description, (2) limited amount of data and (3) the types of datasets used to substantiate their claims about L1 influence.

3.3.1.1. Lack of detailed data description

As Ellis and Barkhuizen (2005:25) have pointed out, "SLA researchers have often been neglectful in providing detailed information about the situational background of the learners they study." In addition, they have often failed to provide detailed data description. Kamimoto et al. (1992) reconsider Schachter's (1974) study into the acquisition of English relative clauses by speakers of Persian, Arabic, Chinese and Japanese and express concern that such a 'second language classic' lacks rigour in data description. Schachter makes use of written

[84] Kellerman (1984:102)

compositions by learners studying English at the American Language Institute, in the University of California, but does not give any detail about IL productions. No information is provided about, for example, topic or how much time was allocated to the students to carry out the writing task. Quantitative data such as total number of words investigated for each group, relative frequency of relative clauses per L1 population or mean lengths of the compositions produced by Persian, Arab, Chinese and Japanese students are sorely lacking. It is therefore not possible to verify whether Schachter's finding that Chinese and Japanese learners use fewer relative clauses is correct or "whether RC [relative clause] frequency goes hand-in-hand with overall essay length" (Kamimoto et al. 1992:274).

There are many examples of text-based transfer studies which do not make use of textual data analysis (e.g. total number of words, total number of occurrences of the item under study, relative frequencies, mean length). Mahmoud (2002; 2005) investigates Arab learners' use of idioms and collocations in essays but does not provide any quantitative information about these texts. In his 2002 study on idioms, for example, Mahmoud writes that "relevant data were collected from paragraphs, essays and term papers written by Arabic-speaking second-year university students majoring in English" and that "a total of 124 idioms (excluding phrasal verbs and binomials) were found in 3220 pieces written by 230 students." However, the total number of words and the mean length of the essays are not given. It is thus impossible to make sense of "the small number of idioms" found in learner essays.

The study by Chi et al (1994) illustrates another problem relating to textual data analysis. The authors examine erroneous collocations with delexical verbs (*have, make, take, do* and *get*) in approximately 1,000,000 words from the HKUST Learner Corpus (cf. Table 3.4) and provide the frequency of lexical collocational errors for the five verbs (cf. Table 3.5). However, the total number of occurrences of these verbs is not given. As a result, it cannot be checked whether percentages of errors per verb are similar or not.

Table 3.5: Frequency of lexical collocational errors of the five verbs (Chi et al. 1994:159)

Verb	Number of occurrences
have	12
do	16
make	44
take	46
get	49

Jansen (2000:40) points up that "[a]pparently, when it comes to supporting a particular theoretical point, describing the data in only very general terms is deemed sufficient (by the authors and presumably also by their reviewers)." The main objective of SLA researchers is to build "models of the underlying mental representations and developmental processes which shape and constrain second language (L2) productions" (Myles 2005:374). It seems that data serves as a 'means to an end' here and is not assigned the major role it deserves. Lack of detailed data description and insufficient information on the learners under study make it very difficult to **compare** studies and make sense of apparently contradictory results. For each transfer study, it is often possible to find another study which takes the opposite view and provides counter-evidence. Thus, Kellerman (1977) provides little evidence of transfer effects in the acceptability judgements tasks performed by less advanced students who even tend to reject Dutch-like English idioms. By contrast, Irujo (1993) offers evidence for L1 influence in the translations produced by less advanced Spanish learners (cf. section 3.2.2.1.2). Similarly, L1 influence on learners' use of collocations is found in a large majority of studies except for Leśniewska (2006) (cf. section 3.2.2.1.3). Another corollary of this lack of detailed data description in transfer studies is that **replication studies** are often impossible to conduct (cf. Lightbown 1984:249)[85].

These weaknesses point to the necessity to control for and eventually manipulate certain variables in SLA studies. Ellis and Barkhuizen (2005:30) list the following variables that minimally need to be considered in producing a full description of the participants in a study:

- the learner's social and situational background
- the situational context in which the writing took place
- the genre
- the topic(s)
- timing
- availability of reference tools.

We will come back to learner and task variables in section 4.1.1 when describing the learner corpus data used in this thesis.

[85] But see Cobb (2003) for a succesful attempt at replicating three learner corpus-based studies.

3.3.1.2. Types of data

> "... competence can only be examined by investigating some kind of performance and (...) the key methodological issue is what kind of performance provides the most valid and reliable information about competence" (Ellis and Barhuizen 2005:21)

Ellis and Barkhuizen (2005:15-50) classify the types of data used by SLA researchers into three broad categories: (1) non-linguistic performance data, (2) samples of learner language, and (3) reports from learners about their own learning. As shown in Table 3.6, **non-linguistic performance data**[86] involve measuring learners' reaction times to linguistic stimuli, non-verbal measures of learners' comprehension of linguistic input (e.g. matching a sentence with the correct picture) and measures of learners' intuitions[87] about the grammaticality or acceptability of sentences. These tasks are supposed to "provide evidence about what learners *know* as opposed to what they *do*, about competence rather than performance" (Ellis 1994:673). Transfer studies have made extensive use of learners' intuitions, especially in the form of **correctness judgment tests** (e.g. Kohn 1986; Laufer 2003; Leśniewska 2006), **semantic-relatedness judgment tests** (e.g. Jiang 2002; 2004b) and **transferability judgment tests** (e.g. Kellerman 1978; 1986).

Laufer (2003) uses a test of correctness judgment consisting of 35 Russian sentences among which 17 include incorrect collocations that are modelled on Hebrew to investigate the influence that a prolonged contact with L2 Hebrew has on L1 collocational knowledge of Russian immigrants to Israel. Kellerman (1977; 1978; 1986) makes use of native speakers' intuitions regarding their L1 to demonstrate that learners have perceptions of their own language and that these perceptions influence what they are willing to transfer to L2 (cf. section 3.2.3). For example, Kellerman (1978) investigates the transferability of the various meanings of the Dutch verb '*breken*' by conducting two transferability experiments in which he asks Dutch students of English to decide whether they would translate the sentences containing '*breken*' using the English verb '*break*'.

Kohn (1986) argues that the choice of judgment tests is motivated by the memory-output paradox:

> "The problem here is that on the one hand knowledge can never be observed directly, but only inferred from its manifestation in output; on the other hand,

[86] This category broadly corresponds to Ellis's (1994:670) data type of 'metalingual judgments'.

[87] Ellis and Barkhuizen make use of the term 'intuition' where 'introspection' should arguably be employed (cf. note 62).

> however, a straightforward interpretation of output data in terms of knowledge is not possible either, due to the influence of ever-present retrieval constraints caused by, e.g. stress, fatigue, distraction, insecurity. A way out of this dilemma is to infer a learner's knowledge from an output produced under conditions for which retrieval constraints are assumed to be minimal." (Kohn 1986:28)

Similarly, Kellerman justifies using judgment tests by arguing that SLA studies need 'clean' data "dissected away from all irrelevances" (Kellerman 1986:36). However, judgment tests have also been challenged on a number of grounds. One of the most controversial issues is "what exactly they do measure and whether they provide consistent measurements" (Ellis and Barhuizen 2005:19). As a result, Ellis and Barkhuizen argue that, despite having figured strongly in SLA research, measures of learners' intuitions "should not serve as the primary data for studying L2 acquisition" (ibid: 21). They add that "the primary data for investigating L2 acquisition should be samples of learner language" (ibid).

Samples of learner language consist of three main categories: (1) naturally-occurring samples, (2) clinically elicited samples and (3) experimentally-elicited samples (cf. Table 3.6). **Clinical experimentation** has sometimes been used in transfer studies. It involves "the use of tasks where learners are primarily concerned with message conveyance, need to utilize their own linguistic resources to construct utterances, and are focused on achieving some non-linguistic outcome" (Ellis and Barkhuizen 2005:23). For example, Håkansson et al. (2002) make use of specially designed tasks with the aim of eliciting oral narratives from 20 German L2 learners with Swedish as their L1 to inform a study on word order acquisition.

Table 3.6: A classification of data types in SLA research (based on Ellis and Barkhuizen 2005:15-50)

1. NON-LINGUISTIC PERFORMANCE DATA
   - Measures of learners' reaction times to linguistic stimuli
   - Non-verbal measures of learners' comprehension of linguistic input
   - Measures of learners' intuitions about the grammaticality or acceptability of sentences
2. SAMPLES OF LEARNER LANGUAGE
   - Naturally-occurring samples
     - *Oral samples*
       - Pencil-and-paper
       - Audio recording
       - Video recording
     - *Written samples*
   - Clinically elicited samples
     - *Eliciting general samples*
       - Communicative 'gap' tasks
       - Open role plays
       - Text reconstruction tasks
       - Picture composition tasks
       - Oral interviews
     - *Eliciting focused samples*
   - Experimentally-elicited samples
     - *Discrete point tests*
       - Traditional language exercise formats
       - Cloze procedure
       - Elicited imitation
       - Elicited translation
     - *Prompts*
       - Sentence completion
       - Discourse completion
       - Question-and-answer
3. REPORTS FROM LEARNERS ABOUT THEIR OWN LEARNING
   - Self-report
     - *Questionnaires*
     - *Interviews*
     - *Personal learning histories*
   - Self-observation
     - *Diaries*
     - *Stimulated recall*
   - Self-revelation using think-aloud
   - Self-assessment

Until very recently, transfer studies have strongly favoured samples of learner language involving **experimental elicitation**[88], i.e. "the use of some kind of exercise, where learners attend primarily to form, are guided in the form to be produced and thus are focused on displaying usage of a specific linguistic form" (Ellis and Barkhuizen 2005:23) (cf. Table 3.3). Thus, Sjöholm (1998) uses a traditional language exercise format, i.e. a multiple-choice test (with each task involving two correct alternatives, i.e. a phrasal verb and a synonymous one-word verb, together with two distractor verbs) to investigate how patterns of avoidance develop in Swedish-speaking and Finnish-speaking learners' acquisition of English phrasal verbs. Irujo (1986) makes use of a cloze test, a definition test, a multiple-choice test and a Spanish-English translation test to investigate learners' avoidance in the production of idioms. **Fill-in-the-blanks (or cloze) tests** have been used in a number of transfer studies, including Bahns and Eldaw (1993), Hasselgren (1994); Laufer (2000), Huang (2001), Cieślicka (2006) and Leśniewska (2006).

One of the most favoured tasks however has been **elicited translation** (e.g. Ringbom 1978; Biskup 1992; Irujo 1993; Bahns and Eldaw 1993; Abdullah and Jackson 1999; Laufer 2000; Cieślicka 2006; Piasecka 2006). This is arguably most unfortunate as the danger of tasks in which learners have to translate a sentence in the L1 into the L2 is that "it leads to extensive L1 transfer when, in more natural language use, this would not occur" (Ellis and Barkhuizen 2005:38)[89]. Results of a translation task may be more indicative of the strategies of translation adopted by learners than of their underlying linguistic competence (cf. Latkowska 2006:212). One of these strategies is word-for-word rendition, which is most problematic for studies focusing on learners' use of phrasemes. These limitations call for cautious (and parsimonious) use of evidence from translation tasks in transfer studies, which may nevertheless be of value "provided it is used selectively to examine relevant phenomena and that its findings are interpreted with care and in relation to data obtained from other sources" (Latkowska 2006:222).

Researchers such as Corder (1973) and Kellerman (1984) have argued for the use of tightly controlled elicitation tasks in interlanguage studies that force learners to make a choice and, as a result, reveal facets of their IL behaviour interesting to the researcher (cf. Selinker 1992:160). However, there is considerable disagreement on the validity of experimentally elicited data. Quite a few SLA researchers today seem to share Ellis and Barkhuizen's view

[88] Ellis and Barkhuizen define 'elicitation' as "the use of specially designed instruments to obtain production samples from the learner" (2005:23).

[89] The reader is referred to Källkvist (1999) for a study of task-type effects on learner interlanguage in which she compares data from free compositions, retellings and translations.

that "experimental elicitation may only tell us what learners can produce under conditions of experimental elicitation and may or may not reflect what they can do under more neutral conditions of language use" (Ellis and Barkhuizen 2005: 36).

Studies based on experimentally-elicited samples of learner language have often made use of more than one task. Thus, Dagut and Laufer (1985) and Hulstijn and Marchena (1989) make use of a multiple-choice test, a verb translation test and a verb-memorizing test. Koya (2005) analyses Japanese EFL learner language on the basis of a fill-in-the-blank translation task and a multiple-choice test. While it is encouraging to observe that the need for multiple types of data seems to be acknowledged, it may be argued that triangulation should rely on different types of samples of learner language. As Ellis and Barkhuizen put it, "there is a need to supplement experimentally elicited data with clinically elicited and / or naturally occurring data, or both" (Ellis and Barkhuizen 2005:49).

Naturally-occurring samples of learner language are defined by Ellis and Barkhuizen (2005:23) as samples "produced in a real-life situation in order to satisfy some communicative or aesthetic need". Ellis and Barkhuizen (2005) regard **learner corpora** as a major source of naturally-occurring samples of learner language. Learner corpora consist of electronic collections of (near-)natural language learner texts (typically argumentative and literary essays) assembled according to explicit design criteria (see section 4.1.1). They constitute a relatively recent type of learner data in SLA studies, the use of which has been advocated by a number of scholars, e.g. Altenberg and Granger (2001), Housen (2002), Granger (2002; 2004; to appear), Myles (2005) and Gries (to appear b).

It is debatable whether learner corpora, and perhaps especially written learner corpora, can really be labelled naturally-occurring samples of learner language. They are arguably best described as **clinically elicited** data. While learners are concerned with message conveyance in essay writing, they also need to attend to form since essays are either produced in classroom settings (e.g. exams) or as homework to be handed in. In fact, naturally-occurring samples of learner language are very difficult to collect in classroom settings. The best types of approximation to naturally occurring language may be found, for example, in letters written to a pen friend or in computer-mediated communication with native students via the internet (cf. Meunier in preparation).

As Granger (to appear) stresses, one of the main advantages of learner corpus data is that "it brings to the SLA field a much wider empirical basis than has ever previously been available" (cf. section 3.3.1.3 for a discussion of amount of data in SLA studies). However, a review of the latest issues of *Studies in Second Language Acquisition, Language Learning* and *Second Language Research* shows that learner corpus-based SLA studies today remain the exception rather than the rule. Granger (to appear) welcomes the fact that Ellis and Barkhuizen (2005) include a whole chapter on learner corpora as "a clear sign that this new resource will soon be accepted as a bona fide data type in SLA research". However, the 'in-between' status of learner corpus data in SLA is made apparent from the very fact that Ellis and Barkhuizen, two SLA specialists, commissioned a corpus linguist to write this chapter.

A number of explanations may be put forward to account for SLA researchers' lack of enthusiasm for learner corpus data. First, they consist in **a less controlled type of data** than experimental elicitation tasks. Second, there has been a common misconception about learner corpora that variables affecting learner output are difficult to control. In carefully designed learner corpora such as the *International Corpus of Learner English*, a database including information about learner and task variables for each text can be used by researchers to compile sub-corpora that match certain criteria (e.g. all English texts written under exam conditions by Dutch female learners also studying French) (cf. section 4.1.1), thus allowing for strict control of variables.

It must however be recognized that learner corpus researchers have had a tendency to make use of the L1 variable exclusively[90] and **not to pay attention to other variables**. This tendency probably stems from the fact that learner corpus researchers are typically **corpus linguists and/or language teachers**. As a result, they have been primarily interested in investigating learner interlanguage data which represents the type of learners they meet everyday in the classroom (e.g. a majority of French-speaking learners). Learner corpus researchers are usually **not SLA specialists**, which may explain why learner corpus research so far has remained "rather descriptive, documenting differences between learner and native language rather than attempting to explain them" (Myles 2005:380).

It also seems that learner corpus data has been reduced to computer-aided error analysis by the SLA community as the following recent quote by Meara suggests:

> The published work that has come out of these corpora [learner corpora] still remains very much at the level of error analysis and error taxonomy (Meara 2002:400).

With such an affirmation, Meara sweeps away a whole body of research already epitomized in Granger's (1998) edited volume (cf. selected review in section 1.4.2).

[90] Some learner corpus-based studies have also pooled interlanguage productions from different L1 populations (e.g. Howarth 1996; 1998).

Myles offers yet another reason why SLA researchers have generally been quite reluctant to use learner corpora. She argues that learner corpus studies are "too closely dependent on what corpora are at hand, and what software tools are available" (ibid:388). It is undeniable that the field would greatly benefit from the compilation of "more learner corpora – particularly longitudinal ones – representing a much wider range of genres, tasks and learners in a wider range of languages" (Granger to appear) as well as from the design of specialized software tools. However, available written and spoken learner corpora remain largely underexploited in SLA studies.

Several SLA researchers, however, have accepted the challenge of compiling and using learner corpora. This is especially true for languages other than English as a Foreign Language. Inge Bartning and her team started developing the L2 French oral corpus *InterFra* almost 20 years ago within the framework of a research project on second language acquisition (*Interlangue Française – développement, variation et interaction* <http://www.fraita.su.se/interfra/>). Florence Myles and Rosamond Mitchell aim to exploit the growing database of *French Learner Language Oral Corpora* (oral and longitudinal corpora) to "produce a full account of the development of the language system of instructed learners of L2 French" (<http://www.flloc.soton.ac.uk/index.html>). Similarly, Tenfjord et al (2006) describe the design and interface of the ASK Corpus, a learner corpus of Norwegian as a second language, specifically intended to "function as a tool for doing research on second language acquisition" (Tenfjord et al 2006:1821).

3.3.1.3. Amount of data

As explained in section 3.3.1.1, data has traditionally had a rather low status in SLA studies, and more specifically in transfer studies, which have tended to focus on theory formulation and discussion to the detriment of data analysis and exploitation. As a result, researchers have generally offered very little evidence to substantiate their claims. This limited amount of data also stems from SLA researchers' interest in the **individual**. As Gass and Selinker (1992) argue,

> ... there are constraints on language transfer which go well beyond mere similarity and dissimilarity of the two languages involved. These constraints ultimately involve *the learner as an active participant in the learning process*[91], one who makes 'decisions' about what can and cannot be transferred. (Gass and Selinker 1992:235)

[91] My emphasis.

This explains the general predilection of transfer studies for **case studies** which focus on one learner participant (e.g. Schwartz and Sprouse 1996; Singleton 1987b; Yorio 1989; Jarvis 2003). It is noteworthy that Schwartz and Sprouse (1996) proposed the 'Full Access / Full Transfer Model' on the basis of data from a single Turkish learner of L2 German.

A marked preference for case studies has sometimes been accompanied by a feeling that group studies can only be carried out after the individual learner's IL behaviour has been properly investigated. Two decades ago, Kohn wrote:

> The emphasis is on the individual learner, because he is the one who is engaged in transfer in the first place. The question of whether transfer occurs in the interlanguage of learner X or not and if so, which type of transfer is involved, cannot be answered by studying a group of ten or even a hundred learners. The answer can only be found in the interlanguage of learner X, i.e. in what he knows and in how he performs on the basis of his subjective knowledge (...). Group studies on factors determining the emergence of transfer and on the effects of transfer on language learning in general are, of course, the ultimate goal. However, such issues can only be successfully investigated if we are prepared to deal adequately with the individual learner's transfer behaviour in the first place, i.e. with transfer affecting the development of the learner's interlanguage knowledge and his retrieval of this knowledge for use in production. (Kohn 1986:23)

This view has been challenged by researchers such as Selinker (1992), Gass and Selinker (2001) and Myles (2005). Selinker argues that an individual learner's IL behaviour is "never as predictable as that of a group of learners of a particular NL [native language] attempting to learner a particular TL [target language]" (Selinker 1992:208). On the other hand, Myles insists that "we need very large cross-sectional datasets, so that the number of learners in each well-defined stage is big enough for us to be confident that the results of the analysis are generalizable (or to capture what is variable in language development across learners for that matter)" (Myles 2005:375).

One important corollary of an extensive use of case studies is that results are hardly generalizable. Arabski (1979)[92] has been criticized by Selinker (1992) for generalizing results based on a classroom written task and describing them as if they represent the entire interlanguage. However, Arabski's data consisted in errors produced by three groups of Polish learners. Theoretical claims based on a single learner participant as found in Kohn (1986) or Schwartz and Sprouse (1996) arguably lend themselves to much more severe criticism. Their results only apply to the learner's IL behaviour investigated and any attempt at drawing more general theoretical conclusions (or pedagogical implications) will necessarily suffer from **over-generalization**. As Selinker (1992:240) explains, the only method of generalizing that

[92] Cf. note 71.

may be regarded appropriate is that of empirically based theorizing sustained by **statistical** reasoning. Selinker however cautions against equating theoretical significance with statistical significance. We will come back to the use of statistical measures to make generalizations from SLA data in section 7.3.2.2.2.

Next to case studies, there are transfer studies which are based on experimental data from more than one individual. In the transfer studies reviewed in section 3.2, the number of participants generally ranges from about ten (e.g. Irujo 1986; 1993) to around sixty (e.g. Biskup 1992; Bahns and Eldaw 1993; Laufer 200; Huang 2001). Sjöholm's (1998) study stands apart, with a number of participants amounting to 608 Swedish-speaking and 638 Finnish-speaking students in Finland. However, this was probably only made possible by the type of task used in this study, i.e. a multiple-choice test. As for learner corpus-based transfer studies, they usually rely on a sizeable amount of corpus data. However, size is generally described in terms of total number of words rather than number of participants (e.g. Granger 1998b; Borin and Prütz 2004[93]). Learner corpus research has emerged from corpus linguistics. It has inherited its methodologies, tools and theoretical precepts (cf. section 1.4). Unlike in experimental data, individual variation is usually lost in pooled data as the focus is on **repeated events and regularities**:

> There is no virtue in being small. Small is not beautiful; it is simply a limitation. (...) The main virtue of being large in a corpus is that the underlying regularities have a better chance of showing through the superficial variations (...). If similar events are repeated with variation, then the more often they are repeated, the more you are able to see the regularity, the repeated element of the event, rather than the individuality that accompanies every use of every word in a text (Sinclair 2004c:189)

In fact, the corpus research paradigm stands in sharp contrast to case study methodology. A comparison of the following quote by Sinclair and that of Kohn cited above shows that the two approaches move in opposite directions. Proponents of case studies investigate one individual's language behaviour and regard group studies as 'the ultimate goal'. By comparison, corpus linguists typically focus on large amounts of data produced by a sizeable number of individuals. The focus is not on individuals who are seldom compared or even described in corpus-based research. Rather, it is on linguistic events, and more precisely on **repeated linguistic events**. Individual instances can only be interpreted in the light of the framework provided by the repeated events:

[93] Borin and Prütz (2004) investigate L1 syntactic transfer in Swedish university students' written English.

> In gathering and organizing corpus evidence, the first focus is on repeated events rather than single occurrences. This initial stage does not mean that unique, one-off events are necessarily ignored, but rather that they cannot be evaluated in the absence of an interpretative framework provided by the repeated events. (...) When a reliable description of the regularities has been assembled, then individual texts can be read against it, and at that time the individual instance will make a balanced impact by comparison with the norms (Sinclair 2004 [1996]:28-29)

Nesselhauf (2003) proves an interesting exception by providing information about the number of essays that make up the learner corpora used and by giving the provenance (by means of an essay code) of each collocational mistake illustrated. However, the 'individual' variable is not used in the interpretation of results.

3.3.2. Selection of items to be investigated in transfer studies

SLA studies have often investigated "bits and pieces of learners' language chosen for analysis because they caught the researcher's eye, seemed to exhibit some systematicity, confirmed some intuition one had about SLA, or had been found interesting in L1 acquisition" (Lightbown 1984:245) as the following quotes clearly show:

> "Although I have been interested in conventionalized language for many years, the data for this paper came to me almost by accident. While involved in some preliminary studies in the area of fossilization, some obvious facts jumped out at me" (Yorio 1989:60)
>
> "As for many others working within the framework of classical CA, though the predicted IL data matched the structural facts they were not entirely hypothesized; I knew where to look. (...) before conducting the experimental research I had had occasion to listen to large amounts of Israeli talk, both in their IH [Israeli Hebrew] and in their English" (Selinker 1992:185).

Next to these types of data, items investigated in transfer studies have been largely selected on the basis of data from contrastive analysis and error analysis. Selinker (1992:7) traces back **Contrastive Analysis** (CA) to Fries and his much-quoted statement:

> The most efficient materials are those that are based upon a scientific description of the language to be learned, carefully compared with a parallel description of the native language of the learner. (Fries 1945:9)[94]

Fries's conception of contrastive analysis was deeply rooted in practical teaching concerns. In the 1950s and 1960s, researchers embarked on the arduous task of comparing languages so as to identify the differences and hence the difficulties that learners from different mother tongue

[94] Quoted in Selinker (1992:6): Fries U. (1945) *Teaching and learning English as a Foreign Language.* Michigan: University of Michigan Press.

backgrounds may experience (cf. Mitchell and Myles 2004:32). As Selinker (1992:6-7) further explains, the direct link between Fries's thought and CA – and thus interlanguage studies and second language acquisition theories – was, however, only established by Lado (1957) who stressed that Fries's statement necessarily implied the CA fundamental assumption that "the student who comes into contact with a foreign language will find some features of it quite easy and others extremely difficult. Those elements that are similar to his native language will be simple for him, and those elements that are different will be difficult" (Lado 1957:2).

The next decades witnessed a series of attacks against contrastive analysis (cf. Ellis 1994:306-309)[95]. Most severe criticisms were levelled at what has been referred to as the strong or *apriori* version of contrastive analysis which consists in a point by point comparison of two languages so as to "make predictions about what will be the points of difficulty for a speaker of language A, for example, who is attempting to learn language B, on the assumption that similarities will be easier to learn and differences harder" (Schachter 1974:205). However, in the early 1990s, interlanguage specialists still seemed to regard contrastive analysis as "the best place to begin language transfer studies since structural congruence (or at least, partial structural similarity) is most probably necessary, though not sufficient, for many of the claims regarding CLI [cross-linguistic influence]" (Selinker 1992:208-209). This may explain why studies of lexical transfer have often focused on **cognates and false friends** (cf. 3.2.1.2.3).

In the 1970s, the disillusionment with contrastive analysis drove researchers and teachers to turn their attention to the language actually produced by learners. Mitchell and Myles (2004:38) explain that it is Corder (1967)[96] who promoted **error analysis** by stressing the importance of systematically investigating learners' errors. A substantial amount of studies have sought to establish what processes result in learners' errors so as to distinguish between interlingual (or transfer) errors and intralingual errors, which "reflect the operation of learning strategies that are universal, i.e. evident in all learners irrespective of their L1" (Ellis and Barkhuizen 2005:65). In a review of a selected number of error analysis studies, Ellis (1994:302) notes that the percentage of errors reported ranges from 3% to 51%, a large variation which he attributes to the difficulty in determining whether an error is the result of transfer or intralingual processes.

[95] See section 3.3.3 for a discussion of the weak or a posteriori version of contrastive analysis.

[96] Corder P. (1967) The significance of learners' errors. *International Review of Applied Linguistics* 5:161-169.

The difficulty of assessing the extent of cross-linguistic influence is encountered in any type of transfer studies, irrespective of the type of data they use and of whether they focus on errors or other manifestations of transfer (e.g. avoidance, facilitation and overuse) (cf. Ellis 1994:301-306). In the next section, we thus focus on the criteria and methods that have been used to demonstrate transfer.

3.3.3. Types of evidence used to prove transfer

A fundamental issue in transfer studies is how the researcher demonstrates that language transfer has occurred. Singleton and Little (1991) refer to this question as the 'attribution problem'. Two types of evidence have generally been employed to prove L1 transfer. The first one involves using L1 in comparisons between learners' interlanguage and their mother tongue (IL <> L1). The second type of evidence is based on comparisons between the interlanguage of learners from several mother tongue backgrounds (IL <> IL). A third type of evidence has sometimes been used in transfer studies, i.e. comparing the interlanguage of learners sharing the same mother tongue background.

3.3.3.1. IL-L1 comparisons

There has often been a tendency in transfer studies to claim that any L2 error that shows a **similarity to an L1 feature** is the result of transfer (e.g. Swan and Smith 2001; Viberg 1998; Cieślicka 2006; Piasecka 2006). Thus, Faerch and Kasper (1986) state that the appearance of the erroneous form 'employedless' in Danish learner interlanguage most probably originates from Danish 'arbeijdsløs', i.e. the translation equivalent of En. 'unemployed' since the Danish suffix '-løs' generally corresponds to English '-less'. Kohn (1986) attributes German learners' use of simple past instead of past progressive in sentences such as 3.1 to negative transfer effects on the basis of IL-L1 similarity:

3.1. *While you *wrote* [were writing] *your letter, I left the house.*
Ge. 'Während du deinen Brief schriebst, verließ ich das Haus'.

Kohn provides the same explanation for German learners' non-use of the English auxiliary 'do' in sentence 3.2. On the other hand, he explains the erroneous structure illustrated in sentence 3.3 by overgeneralization of the DO-structure since there is no L1 equivalent.

3.2. **When came my mother?* < Ge. 'Wann kam meine Mutter?'

3.3. **Where did the train be?* < ? > Ge. 'Wo war der Zug?'

IL-L1 similarity is certainly the strongest type of evidence for L1 influence (cf. Jarvis 2000). However, the danger of systematically considering that equivalence means transfer and lack of equivalence proves absence of transfer is that other factors may pass unnoticed[97]. There is a case for revisiting Kohn's (1986) study in the light of developmental factors. Ellis (1994:59) describes learners' use of declarative word order in questions (**You like to swim*?) as an error stemming from an incomplete application of rules which involves a failure to fully develop a structure. Similarly, Faerch and Kasper's (1986) example of 'employedless' may be interpreted as an 'overgeneralization error', i.e. an error that arises "when the learner creates a deviant structure on the basis of other structures in the target language" (Ellis 1994:59).

The tendency of claiming that any L2 error that shows a similarity to an L1 feature is the result of transfer is also sometimes made apparent in the research questions formulated in transfer studies. For example, Piasecka (2006) makes use of a translation task to investigate L1 influence on Polish learners' use of 23 English idioms that are **formally identical** to Polish idioms. She formulates the following heavily biased research question:

> "Will they translate Polish idioms into English, **thus showing the effects of transfer**?" [98](Piasecka 2006:250)

This quote implies that if Polish learners produce a correct English idiom, it will be interpreted as the result of L1 influence while, in fact, learners may know that the L1 idioms investigated translate into English cognate idioms.

A limitation of most of the studies which make use of IL-L1 similarity is that they **hardly rely on L1 empirical data** to prove transfer. However, mere linguistic identity of IL and L1 does not prove the existence of the process of transfer. Lightbown describes researchers' tendency to make a case for transfer on the basis that the structure 'exists' in the L1 without further investigation as "'shot-in-the-dark' post hoc interpretive guesses which pass for explanations" (Lightbown 1984:245). Similarly, Jarvis talks about a "you-know-it-when-you-see-it phenomenon" (Jarvis 2000:246). Quite paradoxically, learner corpus-based studies have also often fallen into the trap of claiming L1 influence without relying on L1 empirical data (e.g. Chi et al. 1994; Granger 1998b; Nesselhauf 2003b). As Douglas explains, "the point here is not that these methods are faulty or that the interpretations are invalid, but

[97] On the other hand, Jarvis (2000) comments that transfer may also be obscured by other factors such as L2 influence, L2 proficiency and acquisitional universals.
[98] My emphasis.

only that little or no evidence is provided for either quality [reliability and validity]" (Douglas 2001:451).

A related problem is that of researchers' intuition about **translation equivalents** and similar structures across languages (cf. James 1980:175; Ellis 1994:3). For example, in a study entitled 'Verification of transfer', Ard and Homburg (1992) argue that "the devices used for measuring native language influence have been too subjective, too crude and not sufficiently verifiable" (Ard and Homburg 1992:63). They design a sophisticated matrix of form and meaning similarity between Spanish and English lexical items. However, their translations equivalents are open to criticism. A careful examination of their data reveals that there are several erroneous formal equivalents. For example, the Spanish translation equivalent of En 'emit' is 'emitir' (the form *emiter does not exist in Spanish) and En. 'divide' translates into Sp. 'dividir' (not *divider). Ard and Homburg propose *parentes instead of Sp. 'parientes' as a translation equivalent to En. 'parents'. There is an additional problem here: Sp 'parientes' is much more formal than En. 'parents' and it is most probable that Spanish speakers would use 'padres' instead. These inaccuracies cast serious doubt on the validity of Ard and Homburg's study[99].

There are, however, a number of transfer studies which have made use of **systematic empirically-based IL-L1 comparisons** (e.g. Tono 2004). These comparisons have sometimes been conducted within the broader framework of a triangular comparison between L1, IL and L2, which proves particularly helpful in highlighting **L1-related frequency effects** in interlanguage. For example, in a number of experimental word-order studies, Selinker compares Hebrew, English and IL English (cf. Selinker 1992:183-207). He argues that Hebrew-speaking learners' tendency to produce sentences in which adverbs are placed before the object (e.g. *I like very much movies*) is attributable to transfer of Hebrew's word order on the basis that the relative frequency of this structure in IL stands in sharp contrast with English but closely corresponds to its distribution in Hebrew. Selinker further argues that 'syntactic transfer' can be operationally defined as "a process which occurs whenever a statistically significant arrangement in the NL [native language] sentences reappears in IL behaviour" (Selinker 1992:201).

Other examples of studies which rely on triangular comparisons include Borin and Prütz (2004), Altenberg (2002) and Guillot (2005). Borin and Prütz (2004) compare part-of-speech

[99] Translation and comparable corpora clearly have a role to play in transfer studies by providing empirical evidence of translation equivalents (cf. Granger 2003b; Bowker 2003).

sequences (e.g. C NN V: sentence-initial conjunction – common noun – finite verb)[100] in (1) a corpus of L1 English texts, (2) a corpus of texts produced by Swedish-speaking learners of English and (3) a corpus of Swedish texts with the assumption that "significant common differences would reflect L1 interference in the IL, in the form of underuse or overuse of L2 constructions" (Borin and Prütz 2004:67). Altenberg (2002) carries out a comparison of Swedish and English corpora to test the hypothesis that overuse of causative *make* with adjective complements by Swedish-speaking learners is due to L1 transfer (cf. Altenberg and Granger 2001). Guillot (2005) compares the use of *il y a* (En. 'there is/there are'), *gens/personnes* (En. 'people') and the verb *dire* (En. 'say') in (1) a corpus of essays written by English-speaking learners of French, (2) a corpus of texts from the French newspaper *Le Monde* and (3) a corpus of texts from the English newspaper *The Guardian*. Results show that the relative frequencies of the French lexical items in the corpus of learner essays are closer to those of their English counterparts, thus suggesting L1 influence.

Odlin argues that "the comparison of the interlanguage with the native and target language has certain limitations, especially with regard to positive transfer. If the NL [native language] and TL [target language] show little or no difference in some structure common to both, any pattern of positive transfer should not differ much, and any actual difference in interlanguage patterns in such cases will not automatically say much about transfer" (Odlin 2003:447).

3.3.3.2. IL-IL comparisons

Proponents of IL-IL comparisons take the view that L1 influence is best identified through a comparison of the interlanguage of learners from at least two different mother tongues (cf. Kellerman 1984; Ringbom 1986; Odlin 1989; Ard and Homburg 1992). They argue that if interlanguage features are investigated for one L1 population only, it is impossible to prove that learners from different mother tongue background would not have performed identically. As Odlin explains,

> "Speakers of languages using these structures [definite and indefinite articles] might or might not have an advantage in using articles in a new language (e.g. a Spanish speaker learning English). Certainly, researchers sympathetic to contrastive analysis might take any success to indicate positive transfer, but skeptics might argue that any success results simply from acquisition strategies common to first and second language acquisition. Clearly, the way to resolve such

[100] See section 5.2.2 for a discussion of corpus formats.

> an impasse is to compare learners whose languages have articles with learners whose languages do not." (Odlin 2003:447)

Other L1 learner populations are thus needed as control groups to verify that L1 really is the major explanatory factor. As Kellerman puts it, "[w]hen differences in performance correspond to differences in language background, then it is reasonable to suppose, ceteris paribus, that the major reason for these differences is L1 influence" (Kellerman 1984:100).

A large proportion of the studies that have investigated lexical transfer by means of experimental techniques are nevertheless based on data from one L1 population (e.g. Irujo 1986; 1993; Bahns and Eldaw 1993; Laufer 2000; 2003). Notable exceptions are studies comparing Finnish-speaking and Swedish-speaking learners of English as initiated by Ringbom (e.g. Ringbom 1987; 2006; Sjöholm 1998). Similarly, comments on potential L1 influence are often found in corpus-based studies that investigate interlanguage features of one L1 population (e.g. Chi et al. 1994; Granger 1998b; Liu and Shaw 2001; Nesselhauf 2003a; Tono 2004; Flowerdew 2006). The number of corpus-based studies which have made use of at least two learner corpora specifically to verify transfer is more limited. Examples include Nesselhauf (2003b), Antenberg and Tapper (2001), Altenberg and Granger (2001) and Leńko-Szymańska (2007).

In 1986, Ringbom argued that "[t]he study of comparable groups of learners with different L1s is one of the many areas where more research is clearly needed" (Ringbom 1986:150). This statement appears equally valid today. However, the fact that transfer studies have not always privileged IL-IL comparisons is understandable. First, most of these studies have clear applied objectives and researchers most probably prefer to focus on the L1 learner population they are familiar with (cf. section 3.3.1.2). Second, making sense of data resulting from IL-IL comparisons requires at least some knowledge of other L1s. It is necessary to resort to data from other L1s to interpret cases where learners' use of an L2 feature is consistent with their L1 behaviour while being similar to its use by learners from other mother tongue backgrounds.

Consider, for example, Selinker's argument that Hebrew-speaking learners' tendency to produce sentences in which adverbs are placed before the object (e.g. *I like very much movies*) is attributable to transfer of Hebrew's word order (cf. section 3.3.3.1). It may at first seem to be challenged by Osborne's (2007) similar findings for French EFL learners. However, a comparison with interlanguages from a variety of L1 populations together with some linguistic knowledge about the different mother tongues helps explain these findings and provides a stronger type of evidence for L1 transfer. Osborne (to appear) compares adverb placement in the various interlanguages represented in the *International Corpus of Learner English* (cf. section 4.1.1) and explains that "[g]enerally, the results obtained from the learner corpora indicate that V-Adv-O order is most frequent in the productions of learners whose L1 has verb-raising (French, Italian and Spanish), and least frequent with speakers of V2 languages (Dutch, German and Swedish), with speakers of non-raising languages (Russian, Polish, Czech and Bulgarian) in between" (Osborne to appear). Apparently, Hebrew is a verb-raising language[101]: Selinker's findings are therefore supported by Osborne's study.

This example, however, shows that L1 influence may be obscured when the effects of the mother tongue of different L1 learner populations coincide to produce the same IL behaviour. Thus, researchers who rely exclusively on IL-IL comparisons face the danger of coming to the conclusion that there is no L1 influence in situations where it may play an important role. On the other hand, differences in interlanguage behaviour brought to light by IL-IL comparisons may not always be indicative of L1 influence. Other background-related factors such as culture and education may also be responsible for what Jarvis's (2000) refers to as 'inter-L1-group heterogeneity in learners' interlanguage performance'.

3.3.3.3. Comparing the interlanguage of learners sharing the same L1

A third type of evidence to prove L1 influence has recently been described by Jarvis (2000). It consists in verifying whether learners who share the same L1 background behave as a group with respect to a specific L2 feature. The use of this type of comparison in interlanguage studies has already been advocated by Faerch et al. (1984) who were particularly interested in investigating "how many individuals belonging to certain groups of learners use specific forms and functions" (Faerch et al. 1984:278). As Jarvis (2000) explains, however, the interlanguage behaviour of learners from a specific L1 population may lack homogeneity when several L2 options are available for serving the same communicative function or in areas of language use that are more likely to encourage individual variation. On the other hand, intra-L1-group homogeneity may be high as a result of other factors such as acquisitional universals or L2 influence and does not rule out the possibility that the IL feature under study is shared by learners from different mother tongue backgrounds. This third type of evidence is therefore clearly not sufficient to prove cross-linguistic influence.

[101] See http://ling.ucsc.edu/events/wccfl-21/abstracts/goldberg.pdf and Cole (1976)

3.3.3.4. A case for multiple sources of evidence

It is difficult to assess which of L1-IL and IL-IL comparisons provides a better picture of L1 influence. However, what is clear, as Jarvis stresses, is that "it is probably impossible to arrive at a truly complete picture of L1 influence without performing both types of comparisons" (Jarvis 2000:252). Jarvis further comments that "[o]ne might also argue that it is precisely the lack of the complete picture that has produced so much confusion concerning transfer" (ibid). As a result, he argues that transfer studies cannot make do with one single type of evidence but should systematically resort to the three types of comparisons described in this section (i.e. IL-L1 comparison, IL-IL comparison and a comparison of the interlanguage of learners sharing the same L1). In his 2000 article, he thus incorporates these three types of L1 observable effects into a unified framework for the study of L1 influence together with a theory-neutral definition of L1 influence and a list of variables that need to be controlled in transfer studies. We will come back to this framework in section 7.3.

3.3.4. Corpus linguistics versus experimental methods

> "[I]n the light of the enormity of our ignorance, we can perhaps draw a general lesson for L2 research – namely, that we cannot afford to ignore any avenue that holds the possibility of supplying information and insights about L2 processes. Purism in this matter is entirely out of place" (Singleton 1999:245)

The need for multiple types of data is now widely acknowledged by the SLA community (cf. section 3.3.1.2). As Ellis and Barkhuizen report, "there is an obvious need to employ multiple data collection methods on the grounds that no one method will provide an entirely valid picture of what a learner knows or thinks" (Ellis and Barkhuizen 2005:49). Evidence from experimentally elicited transfer studies arguably need to be checked against other types of learner language samples. One option is to make use of learner corpus data. The strengths of corpus linguistics techniques make up for the weaknesses of experimental methods (and vice versa). While studies based on experimentally elicited data have often been criticized for the limited number of learners they involve, they have also been praised for controlling variables that may influence learners' interlanguage, taking the individual learner into account, and interpreting results on the basis of a solid theoretical background. By contrast, the strength of learner corpus based research lies in the large amount of learner samples that can be investigated with the help of corpus linguistics tools and methodologies. Its major weaknesses today, however, are its tendency to conflate learners and analyze pooled data and its often deficient theoretical knowledge of SLA theories.[102]

The two fields also have a lot to learn from each other, especially regarding quantitative data analysis. The statistics used in SLA studies (e.g. correlation, analysis of variance, by-subjects and by-items analyses, etc.) remain largely unexploited in learner corpus based research[103] (cf. Gries 2006; 2007). We will come back to this issue in section 7.3.2.2. Conversely, SLA researchers rarely rely on textual data analysis (e.g. total number of words, relative frequencies, X^2, log-likelihood, etc.) as pointed out in section 3.3.1.1.

Myles argues that "the field of SLA needs to make use of corpora for addressing current theoretical issues" (Myles 2005:374)[104]. In addition, corpora could also help widen the scope of transfer studies as advocated by Ringbom (2006):

> "There is a fair amount of literature on transfer, but the scope of transfer studies needs to be widened. Transfer has mostly been discussed in connection with Error Analysis, where learners' L1 based deviations (especially syntactic ones) from the norm of the TL have been easy to spot, while the ways in which L1-knowledge has facilitated learning are much more difficult to notice." (Ringbom 2006:2)

By taking advantage of corpus linguistics techniques to handle quantitative data, learner corpus-based studies could help discover new types of evidence of transfer in interlanguage. Similarly, corpus linguistics tools could be used to uncover IL features worthy of investigation. The argument here is not that corpus data are better than experimentally elicited data. Rather, it is suggested that learner corpus-based studies can supplement the many transfer studies based on elicitation techniques by shedding new light on L1 influence and perhaps answering other questions. The important point is to recognize that data types "commit us to a proper recognition of the constraints on hypothesis building that each type of corpus[105] imposes on us" (Kellerman 1984:35). One of the objectives of this thesis will thus be **methodological**, viz. assess the usability of learner corpus data to lift some of the 'methodological fog' that covers transfer studies and investigate its potential to uncover new types of evidence of transfer.

[102] As Douglas suggests for language testing researchers, learner corpus researchers "would benefit by paying more attention to the insights about acquisition processes offered by SLA researchers as well as an emphasis in SLA on the examination of the specific language produced by research subjects, as opposed to [corpus linguists'] penchant for counting and measuring" (Douglas 2001:453).

[103] It should however be noted that SLA studies sometimes make use of very sophisticated statistical tests on very small amounts of data (e.g. Irujo 1993).

[104] Myles, however, argues that "the kind of corpora that are readily available are not necessarily those most suited to the investigation of SLA acquisition processes: they are nearly always written, crossectional and overwhelmingly from advanced learners of English." (Myles 2005:388)

[105] Kellerman uses 'corpus' in the sense of 'data source'.

3.4. Conclusion

Lack of shared terminology to refer to different types of phrasemes was already deplored in chapter 2. The problems it creates have appeared very clearly in this chapter. Studies are sometimes hardly comparable as they make use of the same term (e.g. idioms) to refer to different phrasemes and do not always provide a definition for their object of study. Transfer studies are so numerous that it is extremely difficult not to get lost in the labyrinth of theories, findings and claims that have appeared in the body of transfer-related literature of these last 50 years. It is also particularly challenging to make sense of contradictory findings. These difficulties have been shown to be compounded by methodological problems which cast doubt on what appear to be well-established research results in the literature.

Methodological issues in transfer studies have however been addressed by a number of SLA scholars such as Ellis (1994) and Ellis and Barkhuizen (2005). These researchers have highlighted the need for triangulation. Jarvis (2000) and Odlin (2003) have also argued for the use of more than one type of evidence to prove transfer. A leitmotif in this chapter has been the fact that learner corpus data remain largely unexploited in transfer studies, and indeed, in SLA studies more globally. It has been argued that learner corpus data could be used to supplement other types of learner language samples. In addition, they could be employed to overcome some of the limitations of transfer studies in terms of amount of data, type of data analysis, selection of items to be investigated, etc. However, these claims need empirical validation.

One objective of this thesis will thus be to investigate how (learner) corpora can be used to help inform the different steps of a transfer study, from item selection to data interpretation.

4. Data and methodology

As already stated in the introduction, this thesis has three major objectives:

- To examine words and phrasemes that native speakers and EFL learners use to serve typical organizational or rhetorical functions such as exemplifying and concluding;
- To investigate whether upper-intermediate to advanced EFL learners, irrespective of their mother tongue backgrounds, share a number of features that characterize the way they make use of lexical devices to serve specific rhetorical functions;
- To assess the role of the first language in EFL learners' specificities, by comparing French learners' use of EAP vocabulary with that of other learner populations.

Sections 4.1 and 4.2 describe the corpora and methodology used to pursue these three objectives. The method to analyse corpus data and compare the use of EAP words and phrasemes in native and learner corpora is based on the *Integrated Contrastive Model* described in section 4.2.1. The software used for the quantitative and qualitative description of EAP words and their co-occurrents is described in section 4.2.2. Section 4.2.3 reviews major statistical measures for co-occurrence extraction and compares them. It then presents the quantitative approach to phraseology adopted in this thesis.

4.1. Data

This thesis makes use of both learner and native speaker written data. It compares words and phrasemes that are used to serve rhetorical and organizational functions in EFL learner writing and expert academic prose. Researchers such as Lorenz (1999a) and Hyland and Milton (1997) criticize the use of professional writing in learner corpus research, arguing that it is "both unfair and descriptively inadequate" (Lorenz 1999a:14) and taking a stand against the "unrealistic standard of 'expert writer' models" (Hyland and Milton 1997:184). Native student writing is arguably a better type of comparable data to EFL learner writing if the objective of the comparison is to describe and evaluate interlanguage(s) as fairly as possible. It will be used in this thesis to identify features that are shared by EFL learners and native students and are therefore likely to be characteristic of novice writing. It is, however, highly questionable whether findings from such comparisons can make their way to the classroom. As Leech (1998, xix) puts it, "[n]ative-speaking students do not necessarily provide models that everyone would want to imitate". For example, native students have been shown to produce more dangling participles than EFL learners (cf. Granger 1997b) and different types of orthographic errors (cf. Cutting 2000).

The question of the norm can be settled by taking into account the aim of the comparison. Professional writing has a major role to play in learner corpus research as soon as pedagogical applications are considered. As put by Ädel (2006:206-207),

> On the one hand, it can be argued that in order to evaluate foreign learner writing by students justly, we need to use native-speaker writing that is also produced by students for comparison. On the other hand, it can also be argued that professional writing represents the norm that advanced foreign learner writers try to reach and their teachers try to promote. In this respect, a useful corpus for comparison is one which offers a collection of what Bazerman (1994: 131)[106] calls 'expert performances'.

We nevertheless agree with De Cock's (2003:196) comment that "argumentative essay writing has no exact equivalent in professional writing". Special care will thus be taken to interpret results in the light of genre analysis as differences between student essays and expert writing may simply reflect differences in their communicative goals and settings (cf. Neff et al. 2004).

Section 4.1.1 describes the learner corpus used. This thesis makes use of three types of native corpora, i.e. professional academic writing, student writing and speech, which are described in section 4.1.2. Section 4.1.3 addresses the contentious issue of comparing EFL learner writing to native models of English.

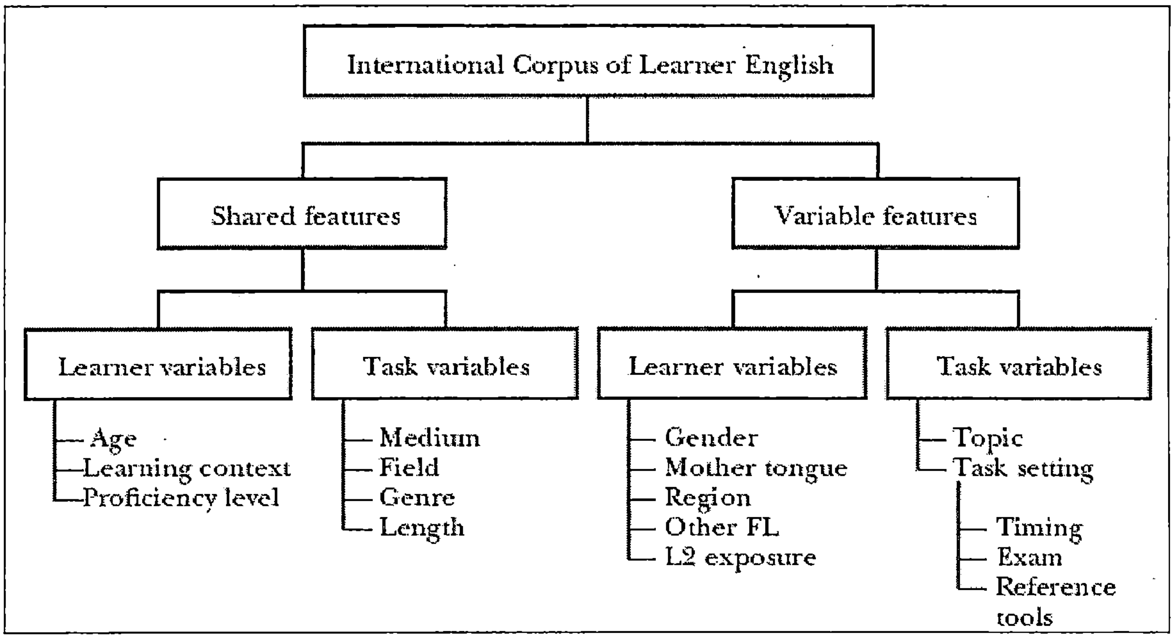
4.1.1. EAP learner writing

The learner data used in this thesis consist of ten sub-corpora of the *International Corpus of Learner English* (henceforth ICLE) compiled at the University of Louvain, Belgium, under the supervision of Sylviane Granger (cf. Granger et al 2002; Granger 2003). A computer learner corpus is an electronic collection of (near-)natural language learner texts assembled according to explicit design criteria (cf. Granger to appear) (see Ellis and Barkhuizen (2005) for a discussion of different types of learner data). Each learner text in the corpus is documented with a detailed profile questionnaire completed by the learner who wrote the essay. Profile questionnaires give two types of information about learner texts, i.e. information on the type of task and learner characteristics. Figure 4.1 shows that some of the task and learner variables (age, learning context and proficiency level of learners and medium,

[106] Bazerman C. (1984) Systems of genres and the enactment of social intentions. In Freedman A. and P. Medway (eds) Genre and the new rhetoric. London: Taylor and Francis, 79-101.

field, genre and length of the task) are shared by all learner texts: these variables were used as corpus design criteria. Other variables differ from text to text, e.g. gender and mother tongue of learners and topic and task setting of learner essay writing. Task and learner variables can be used to compile sub-corpora that allow for studies of the influence of each variable. Thus, a comparison of French- vs. Dutch-speaking learners can highlight the potential influence of the mother tongue while a comparison of time vs. untimed essays can give valuable insights into the potential influence of the time variable.

Figure 4.1: ICLE task and learner variables (Granger 2003: 539)



In ICLE, learners share a number of features: they are young adults who study English as a Foreign Language at university. They are all in their second, third or fourth year and their level is described as advanced although "individual learners and learner groups differ in proficiency" (Granger 2003:539). A number of texts written by learners from the 11 mother tongue backgrounds as available in the first version of the International Corpus of Learner English have recently been rated externally by a professional ESOL rater according to the descriptors for writing found in the Common European Framework of Reference for Languages. Results show that learner essays rate from B2 to C2, with a majority of C1 essays, and that the proportion of B2, C1 and C2 texts differs between the 11 mother tongue backgrounds (cf. Thewissen et al. 2006). Learners also differ in a number of features as illustrated in Figure 4.1. The learner variable on which emphasis will be placed in this thesis is the mother tongue variable. The effect of gender, region, other foreign language and L2 exposure will not be analyzed.

The ICLE data share a number of task attributes. Granger (2003:540) describes them as consisting "exclusively of written productions of a particular genre, namely, essay writing". A large majority of essay topics are argumentative and titles include *Crime does not pay; Most university degrees are theoretical and do not prepare students for the real word* and *Feminists have done more harm to the cause of women than good.* Essays differ in task settings: they may have been written in timed or untimed conditions, as part of an exam or not, with reference tools such as grammars and dictionaries or not. Most studies of ICLE data to date have not taken these task settings into consideration[107]. Studies in second and foreign language acquisition and teaching writing however insist on the influence of task types and conditions (cf. Shaw 2004, Kroll 1990). Learner essays were therefore carefully selected in an attempt to control external variables which may affect the written production of learners. Sub-corpora were compiled on the basis of two learner variables, i.e. the **mother tongue variable** and **the language at home variable**[108], and three task variables, i.e. all texts are **untimed argumentative** essays potentially written with the help of **reference tools.**

Although essays written without the help of reference tools would arguably be more representative of what advanced EFL learners can produce, we had to select untimed essays with reference tools as they represent the majority of essays in ICLE. Table 4.1 gives the number of essays per mother tongue that correspond to learner and task variables used to compile sub-corpora in this thesis and the number of essays that are left when we apply these criteria successively. Note that essays written by Bulgarian-speaking learners were not used in this thesis as, after applying the selection criteria, the resulting corpus was very small. The last column gives the total number of words per L1 sub-corpus. Word counts were made with the *WordList* option of *WordSmith Tools* (cf. 4.2.2.1) with the options 'hyphen does not break words' and 'no numbers in wordlist'.

[107] Ädel's (to appear) analysis of the time vs. untimed variable is an exception.

[108] The 'language at home' variable is used in addition to the 'mother tongue' variable to make sure that learners do not speak another language at home, thus controlling that learners are not bilinguals.

Table 4.1: Breakdown of ICLE essays

	Nr. of essays per L1	Nr. of essays per L1 = language at home	Nr. of argumentative essays	Conditions (based on argumentative essays)		Reference tools (based on untimed conditions)		Nr. of words	Average nr. of words per essay
				Nr. of timed essays	Nr. of untimed essays	Nr. of essays with no ref. tools	Nr. of essays with ref. tools		
Bulgarian	302	300	300	0	300	269	(31)[109]	-	-
Czech	244	225	182	1	180	29	147	130,768	890
Dutch	258	254	243	11	207	10	196	162,243	828
Finnish	259	254	229	33	193	26	167	125,292	750
French	311	303	256	2	244	5	228	136,343	598
German	433	425	410	176	205	23	179	109,556	612
Italian	391	364	122	36	83	4	79	47,739	604
Polish	363	361	357	116	238	12	221	140,521	636
Russian	274	271	270	10	224	27	194	165,937	855
Spanish	248	248	196	6	163	9	149	99,119	665
Swedish	352	337	283	155	118	5	81	48,060	593
TOTAL	**3435**	**3342**	**2848**	**546**	**2155**	**419**	**1641**	**1,165,524**	**697**

[109] Not added to the total number of texts.

4.1.2. EAP native writing

The native data used in this thesis consists of corpora of professional and student writing as well as a corpus of speech that was sometimes used to check whether specific word sequences are more frequent in speech or writing. Sections 4.1.2.1 and 4.1.2.2 describe the corpus of professional academic writing and that of student writing respectively. The corpus of academic speech is described in section 4.1.2.3.

4.1.2.1. Professional writing

The *British National Corpus* (BNC) is a synchronic general corpus that is supposed to represent contemporary British English as a whole. The BNC contains approximately 100 million words which reflect a wide variety of text types, genres and registers. The written component totals 90% of the corpus and includes samples of academic books, newspaper articles, popular fiction, letters, university essays and many other kinds of text. The spoken component represents 10% of the whole corpus and consists of monologues and dialogues in many different contexts, e.g. business, leisure and education. Aston and Burnard (1998) describe the text selection procedure as follows:

> In selecting texts for inclusion in the corpus, account was taken of both production, by sampling a wide variety of distinct types of material, and reception, by selecting instances of those types which have a wide distribution. Thus, having chosen to sample such things as popular novels, or technical writing, best-seller lists and library circulation statistics were consulted to select particular examples of them. (Aston and Burnard 1998:28)

The BNC is annotated according to the Text Encoding Initiative (TEI) mark-up guidelines (cf. Burnard 2002). Mark-ups include rich metadata on a variety of structural properties of texts (e.g. headings, sentences and paragraphs), file description, text profile, as well as linguistic information (see section 5.2.2.1 for more information on linguistic information).

Written texts in the BNC do not comprise more than 45,000 words: longer texts were sampled so as to allow for a wider coverage of text types and avoid over-representing idiosyncratic uses of the English language. However, such a decision presents problems for certain types of linguistic enquiries. A number of recent studies in the field of English for academic purposes have shown that words may behave differently and display different preferred phraseological uses in abstracts, introductions, methods and results sections as well as conclusions (cf. section 1.4.1). Quantitative comparisons between the BNC and other corpora should be made with caution, especially when the lexical items under study are closely linked to specific parts of texts (e.g. words and phrasemes used to introduce the main topic or a conclusion).

Three criteria were originally used to select written texts to design a balanced corpus: domain, time and medium. Domain refers to the subject field of the texts; time refers to the period when the text was written and medium refers to the type of publication, e.g. books, newspapers, periodicals, etc. Lee (2002) criticizes the domain categories for being "overtly broad and too inexplicit" and proposes "a proper navigational map for people wanting to deal with specific 'genres'". Table 4.2 gives the breakdown of genres in the BNC written corpus according to Lee's categorization and shows that genre labels are often hierarchically nested. Thus, if we want to analyse the language of natural sciences, we can select all texts classified under *W_ac_nat_science*. If we are not interested in sub-divisions into natural science, medicine, technology and engineering, humanities and arts, etc., we can select all texts whose categorizing labels begin with *W_ac*.

Table 4.2: Breakdown of written BNC genres (from Lee 2002)

BNC written	No. of words	%		No. of files
W_ac_humanities_arts	3,321,867	3.8%	Academic prose 17.7%	87
W_ac_medicine	1,421,933	1.6%		24
W_ac_nat_science	1,111,840	1.3%		43
W_ac_polit_law_edu	4,640,346	5.3%		186
W_ac_soc_science	4,247,592	4.9%		138
W_ac_tech_engin	686,004	0.8%		23
W_admin	219,946	0.3%		12
W_advert	558,133	0.6%		60
W_biography	3,528,564	4%		100
W_commerce	3,759,366	4.3%		112
W_email	213,045	0.2%		7
W_essay_sch	146,530	0.2%	Unpublished essays	7
W_essay_univ	65,388	0.1%		4
W_fict_drama	45,757	0.1%	Fiction 18.6%	2
W_fict_poetry	222,451	0.3%		30
W_fict_prose	15,926,677	18.2%		432
W_hansard	1,156,171	1.3%		4
W_institut_doc	546,261	0.6%		43
W_instructional	436,892	0.5%		15
W_letters_personal	52,480	0.1%	Letters 0.2%	6
W_letters_prof	66,031	0.1%		11
W_misc	9,140,957	10.5%		500
W_news_script	1,292,156	1.5%		32
W_news_brdsht_nat_arts	351,811	0.4%	Broadsheet national newspapers 3.5%	51
W_news_brdsht_nat_commerce	424,895	0.5%		44
W_news_brdsht_nat_editorial	101,742	0.1%		12
W_news_brdsht_nat_misc	1,032,943	1.2%		95

W_news_brdsht_nat_reportage	663,355	0.8%		49
W_news_brdsht_nat_science	65,293	0.1%		29
W_news_brdsht_nat_social	81,895	0.1%		36
W_news_brdsht_nat_sports	297,737	0.3%		24
W_news_other_arts	239,258	0.3%	Regional and local newspapers 6.4%	15
W_news_other_commerce	415,396	0.5%		17
W_news_other_report	2,717,444	3.1%		39
W_news_other_science	54,829	0.1%		23
W_news_other_social	1,143,024	1.3%		37
W_news_other_sports	1,027,843	1.2%		9
W_news_tabloid	728,413	0.8%		6
W_non_ac_humanities_arts	3,751,865	4.3%	Non-academic prose 19.1%	111
W_non_ac_medicine	498,679	0.6%		17
W_non_ac_nat_science	2,508,256	2.9%		62
W_non_ac_polit_law_edu	4,477,831	5.1%		93
W_non_ac_soc_science	4,187,649	4.8%		128
W_non_ac_tech_engin	1,209,796	1.4%		123
W_pop_lore	7,376,391	8.5%		211
W_religion	1,121,632	1.3%		35
TOTAL	87,284,364	100%		3144

This genre-based labelling of texts allows for a broad range of variation studies. The most general type of comparison that can be made is a comparison between writing and speech. Broad categories of writing can also be compared, e.g. academic prose versus newspaper articles or fiction samples. Third, sub-genres can be contrasted, e.g. the academic discourse of humanities and arts versus the academic prose of natural science.

The sub-corpus of academic prose in humanities and arts (W_AC_HUMANITIES_ARTS: 3,321,867 words) will be referred to as BNC-AC-HUM in this thesis and used as a comparable corpus to the ICLE sub-corpora. There are however major differences between the two corpora. First, ICLE is a corpus of unpublished university student essays while BNC-AC-HUM consists of samples of published articles and books. Second, student essays rarely total more than 1,000 words while samples in the BNC-AC-HUM are much larger (from 25,000 to 45,000 words). Third, topics in BNC-AC-HUM differ from those in ICLE (cf. section 4.1.1). They include, among others, *the people's peace; national liberation; the morality of freedom; Europe in the central middle ages; China's students; British literature since 1945; what is this thing called science?; Soviet relations with Latin America; Nietzsche on tragedy*, etc. Unlike in ICLE, topics in BNC-AC-HUM appear only once.

Despite these major differences, it is believed that BNC-AC-HUM is the most suited available corpus to be compared with student writing as found in the ICLE-sub-corpora for two main reasons. University students who wrote the ICLE essays are students of humanities:

texts in BNC-AC-HUM are therefore believed to be closer to the types of text these students might have come across during the first years at university. They also have the advantage of corresponding to the type of writing that learners will have to produce in their university curriculum. Comparisons will however be made with caution, especially quantitative ones.

The *British National Corpus* was accessed via two web-based tools that are described in section 4.2.2.2.

4.1.2.2. Student writing

The student writing corpus used in this thesis is a subpart of the *Louvain Corpus of Native Speaker Essays* (LOCNESS) (cf. Granger 1996; 1998a) and will be referred to as STUD-US-ARG. It consists of 88 essays written by American university students and totals 100,702 words. The average number of words per essay is 1,144. Texts were carefully selected in an attempt to control external variables which may affect the writing process and to be as comparable as possible to the learner data: they are all **untimed argumentative essays** potentially written with the help of **reference tools**. Essay titles include, among others, *death penalty, euthanasia, crime does not pay* and *money is the root of evil.* Note that the decision to use a corpus of essay writing by American students is largely based on a question of availability of corpora as the British English component of LOCNESS mainly contains literary essays.

4.1.2.3. Speech

The spoken part of the *British National Corpus* was regularly consulted to check whether words and co-occurrent items that are found in learner writing are more typical of native speech or academic writing. The BNC spoken corpus consists of 10,334,947 words and will be referred to as BNC-SP. As shown in Table 4.3, the corpus includes a wide variety of spoken registers, among others, broadcast documentary and news, interviews and lectures.

Table 4.3: Breakdown of spoken BNC genres (from Lee 2002)

BNC spoken	No. of words	%		No. of files
S_brdcast_discussn	757,317	7.3%	Broadcast 10.2%	53
S_brdcast_documentary	41,540	0.4%		10
S_brdcast_news	261,278	2.5%		12
S_classroom	429,970	4.2%		58
S_consult	138,011	1.3%		128
S_conv	4,206,058	40.7%		153

S_courtroom	127,474	1.2%		13
S_demonstratn	31,772	0.3%		6
S_interview	123,816	1.2%	Interviews 9.1%	13
S_interview_oral_history	815,540	7.9%		119
S_lect_commerce	15,105	0.1%	Lectures 2.9%	3
S_lect_humanities_arts	50,827	0.5%		4
S_lect_nat_science	22,681	0.2%		4
S_lect_polit_law_edu	50,881	0.5%		7
S_lect_soc_science	159,880	1.5%		13
S_meeting	1,377,520	13.3%		132
S_parliament	96,239	0.9%		6
S_pub_debate	283,507	2.7%		16
S_sermon	82,287	0.8%		16
S_speech_scripted	200,234	1.9%	Speeches 6.4%	26
S_speech_unscripted	464,937	4.5%		51
S_sportslive	33,320	0.3%		4
S_tutorial	143,199	1.4%		18
S_unclassified	421,554	4.1%		44
TOTAL	10,334,947	100%		909

4.1.3. Comparing learner writing to native writing

> "Could it be that the criticism of the native speaker and his/her unreliable intuitions is part of an abstract and politically correct mainstream position to which many linguists are most willing to pay lip service, but which is no longer valid as soon as one faces practical problems like the correction and proof-reading of non-native speakers' texts? Could it also be that even those critics of the native speaker concept and proponents of an English-as-a-lingua-franca norm who happen to be non-native speakers of English try to adhere to standards and norms set by native speakers when it comes to their own use of English? I wouldn't be surprised." (Mukherjee 2005:21)

In his preface to *Learner English on Computer* (Granger 1998), Leech describes the native control corpus as "a standard of comparison, or norm, against which to measure the characteristics of the learner corpora." This view has been challenged on two grounds which are discussed in this section (cf. Granger to appear). First, the general idea of *comparison* has been criticized on the basis that learner language is not a deviant form of the target language but a linguistic system in its own right. Interlanguage should thus be analysed in its own terms to characterize learners' linguistic competence in the L2. This criticism is not only directed at comparisons of learner and native corpora but also at other types of SLA data (e.g. grammaticality judgment tests) to the point that a large proportion of L2 research is described in SLA literature as having succumbed to what Bley-Vroman (1983) first referred to as the **'comparative fallacy'** (cf. Lakshmanan and Selinker 2001; Firth and Wagner 1997).

According to Bley-Vroman, L2 studies that are based on the 'target language as a norm' principle and employ concepts that are defined relative to the target language (e.g. obligatory context and error) will result in "incorrect or misleading assessments of the systematicity of the learner language" (1983:3) and are thus unlikely to understand the nature of interlanguage. He also warns that "[t]he comparative fallacy can have very serious effects on the validity of empirical studies in the way that it influences the interpretation and classification of data" (ibid: 15).

As Mukherjee has convincingly demonstrated, the target language in learner corpus research is **explicit** and **corpus-based.** It is an "abstraction based on the performance of many individuals in various communication situations" (Mukherjee 2005:14). By contrast, the target language in other types of learner data is often left implicit (cf. Sung Park 2004) and is most frequently based on a single researcher's intuition. Lakshmanan and Selinker (2001) address the issue of the comparative fallacy and warn against "judging language learner speech utterances as ungrammatical from the standpoint of the target grammar without first having compared the relevant interlanguage utterances with the related speech utterances in adult native-speaker spoken discourse" (Lakshmanan and Selinker 2001:401). Although they are not more explicit about it, their quote may be understood as a plea for more natural language data (i.e. corpus data) and a warning against hasty conclusions based on researchers' grammaticality judgments.

The second criticism directed at learner versus native language comparisons is concerned with the idea of **the 'native speaker' as embodying the target norm** (e.g. Piller 2001; Tan 2005). Mukherjee, however, argues that 'nativeness' remains a useful construct both for linguistics and for the ELT community, a 'useful myth' in Davies's (2003) terms, and proposes a usage-based definition of the native speaker based on three aspects that he regards as central to native-like performance, i.e. lexicogrammaticality, acceptability and idiomaticity (cf. Pawley and Syder 1983):

> The term 'native speaker' should be used for an *abstraction* of all language users (1) who have good intuitions about what is *lexicogrammatically* possible in a given language and speak/write accordingly, (2) who know to a large extent what is *acceptable* in a given communication situation and speak/write accordingly, (3) whose usage is largely *idiomatic* in terms of linguistic routines commonly used in a given speech community. If we refer to an individual speaker as a native speaker, this speaker is thus taken to exemplify the abstract native speaker model on grounds of his/her language use. (ibid:14)[110]

[110] Italics are mine.

He advocates a **corpus-approximation to the native speaker norm** and argues that corpus data can be used to describe this norm by "generalizing and abstracting from a vast amount of representative performance data" (ibid:15). In this thesis, the corpus-approximation to the native speaker norm is based on British and American English corpora. It should be noted, however, that the existence of a variety of norms is recognized in learner corpus research (cf. Granger to appear) and that other varieties of English could have been used. For example, it may make more sense to compare the Tswana component of the second edition of the *International Corpus of Learner English* to South African English rather than British or American English.

Authors such as Seidlhofer (2001) and Jenkins (2005, 2006) call for the development of a model based on proficient users of English as a Lingua Franca (ELF) on the basis that English is now used by many more non-native speakers than native speakers. However, recent studies have questioned the validity of such a model. Mollin (2006) analyses Euro-English, i.e. Continental Europe ELF, in terms of three processes that need to take place in the development of an endonormative standard, i.e. expansion in function, nativization in form and institutionalization of new norms. The author first investigates the expansion of English in Continental Europe, i.e. its use in domains as different as education, administration, media, literature and interpersonal communication and argues that "Europe has transcended the borders of mere EFL-status, but that it has not truly entered the realms of ESL" (Mollin 2006:87). To test nativization in form, Mollin selects a number of features commonly presented as characteristic of ELF and examines their use in a 400,000 word corpus of Euro-English. She investigates individual lexical items such as *actual* and *eventual*, complementation patterns, countability and articles and shows that what are commonly regarded as ELF features are not shared by a large proportion of speakers. Finally, she makes use of a questionnaire to test the institutionalization criterion and reports that a large proportion of European speakers of English clearly aim for a native-speaker standard. Mollin's results thus suggest that Euro-English cannot be described as a new variety of English in its own right and is better described as being situated in Kachru's (1985) Expanding Circle in which English is used as a Foreign Language only. In the author's words, Euro-English is "the Yeti of English varieties: a mythical entity surrounded by legend, but without any real-life evidence" (ibid: 197).

4.2. Methodology

The method used to investigate the phraseology of EAP vocabulary in native and learner corpora is based on the *Integrated Contrastive Model* developed by Granger (1996). This model is described in section 4.2.1 and the software programs used to analyze corpus data are presented in section 4.2.2. As noted by De Cock, corpus linguistic methods "have brought frequency and more specifically frequency of co-occurrence into the phraseological equation" (De Cock 2003: 41). Section 4.2.3 describes the quantitative approach to phraseology adopted in this thesis. It first discusses a number of parameters that can influence the results of a co-occurrence analysis and highlights two major limitations of raw frequency. It then reviews widely used association measures to extract significant co-occurrents and describes the extraction procedure used in this thesis.

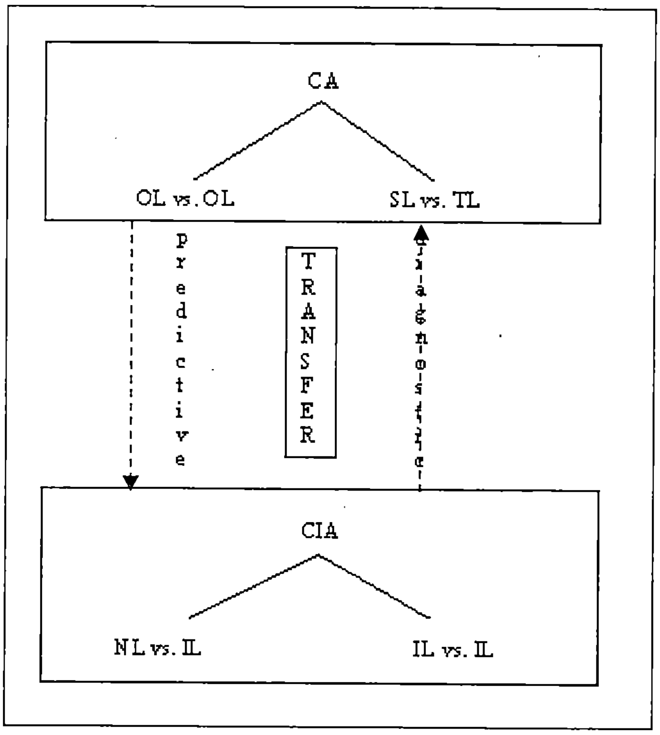
4.2.1. The Integrated Contrastive Analysis Model

The *Integrated Contrastive Model* (Granger 1996; Gilquin 2000/2001) provides a very useful framework to investigate the phraseology of EAP vocabulary that native speakers of English and advanced EFL learners use in academic discourse. The model combines *Contrastive Analysis* (CA) and *Contrastive Interlanguage Analysis* (CIA). CIA consists of two types of comparison. First, it involves a comparison of native and non-native production of the same language which aims to "shed light on non-native features of learner writing and speech through detailed comparisons of linguistic features in native and non-native corpora" (Granger 2002: 12). Such a comparison can "highlight a range of features of non-nativeness in learner writing and speech, i.e. not only errors but also instances of under- and overrepresentation of words, phrases and structures" (ibid). Second, CIA includes a comparison of different interlanguages of the same language, i.e. the English of French learners, Spanish learners, Dutch learners, etc. By comparing learners of different mother tongues, researchers should "gain a better insight into the nature of interlanguage" (Granger 1998a: 12) and "differentiate between features which are shared by several learner populations and are therefore more likely to be developmental and those which are peculiar to one national group and therefore possibly L1-dependent" (Granger 2002: 13).

The combination of CA and CIA proposed by Granger (1996) helps researchers link interlanguage behaviour to performance in the mother tongue. As illustrated in Figure 4.2, the *Integrated Contrastive Model* proposes two types of approaches to CA and CIA data. From CA to CIA, the approach is predictive and consists in formulating CA-based predictions about

L2 production which are then checked against CIA data. From CIA to CA, the approach is diagnostic: it aims to explain CIA findings, i.e. errors but also overuse and underuse, in the light of CA descriptions. The terms 'predictive' and 'diagnostic' are used to refer to working hypotheses which will be confirmed or refuted by corpus data (cf. Granger 1996:46). Together with Jarvis's framework, the latter approach is adopted in chapter 7 to assess the potential influence of the mother tongue but the more neutral term "explanatory" is preferred.

Figure 4.2: The Integrated Contrastive Model (Granger 1996) [111]



Contrastive Interlanguage Analysis has been very popular among researchers in the field of learner corpus research and highlighted an unprecedented number of features that characterize learner interlanguage(s). To date, however, most studies have failed to exploit the full potential of the CIA model, focusing on the comparison between a learner corpus and a native reference corpus, but neglecting the comparison of different learner corpora of the same target language (cf. section 1.4.2). The studies that have compared more than one IL have usually focused on learners from one mother tongue background and used data from one or two learner populations only to check whether the features they have highlighted in one corpus are common to other learners or are L1-specific, and thus possibly transfer-related. One of the objectives of this thesis is to make the most of CIA by comparing the use of EAP

[111] CA: Contrastive Analysis; OL: Original Language; SL: Source Language; TL: Target Language; CIA: Contrastive Interlanguage Analysis; NL: Native Language; IL: Interlanguage

vocabulary and its phraseology in ten learner corpora and two native corpora and to show that these types of comparison are indispensable if we want to identify the distinguishing features of learner language at a given stage of development (cf. Bartning 1997).

Differences between learner and native writing are highlighted by means of log-likelihood tests. The whole learner corpus is compared to BNC-AC-HUM (or STUD-US-ARG) but results are reported only if they are shared by learners from several mother-tongue backgrounds. The statistical tests used in chapter 7 to compare the French sub-corpus to the other learner populations so as to highlight transfer-induced factors are described in that chapter.

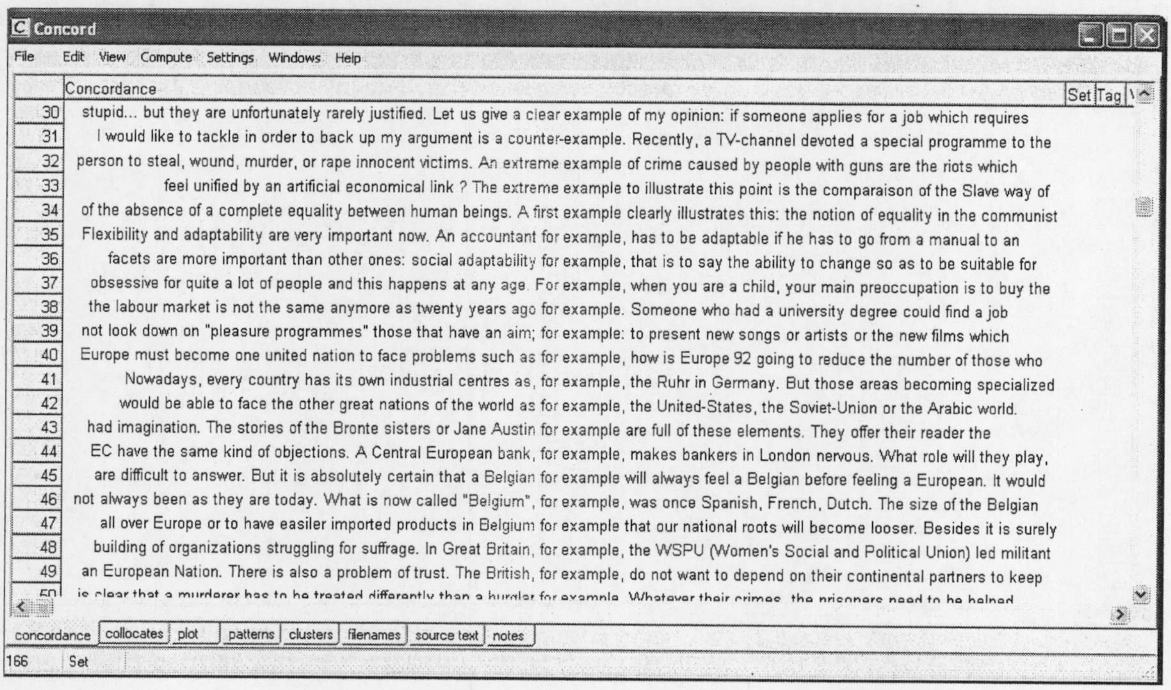
4.2.2. The software used

The study makes use of two types of software to analyse corpus data. It uses the software program *WordSmith Tools 4* (henceforth WST4) to analyse the ICLE sub-corpora and LOCNESS. The *British National Corpus* was accessed through two web-based concordancers described in section 4.2.2.2.

4.2.2.1. WordSmith Tools 4

A number of options available in *WordSmith Tools 4* (Scott 2004) were used to extract EAP vocabulary, most notably the wordlist option, the keyness option and (simple and detailed) consistency analysis. These options are described in Section 5.3 when discussing the quantitative criteria used to extract EAP lexical items. Once EAP words are selected, the next stage of the study is to analyse and contrast their preferred environment in native and learner corpora. The *Concord* option was used to examine the instances of a number of EAP words in context. Figure 4.3 illustrates the type of output given by the *Concord* option and shows the main features of the concordance program. Each occurrence of the word *example* in the corpus is displayed in the centre of each line. The words surrounding the search-word can be alphabetically sorted to reveal its preferred environments. In Figure 4.3, words are left-hand sorted alphabetically. The context size can be enlarged if more context is necessary and the source text can be accessed by double-clicking on a concordance line. In-built facilities in *Concord* also include, among others, *collocates, patterns* and *clusters*.

Figure 4.3: WordSmith Tools' Concord option

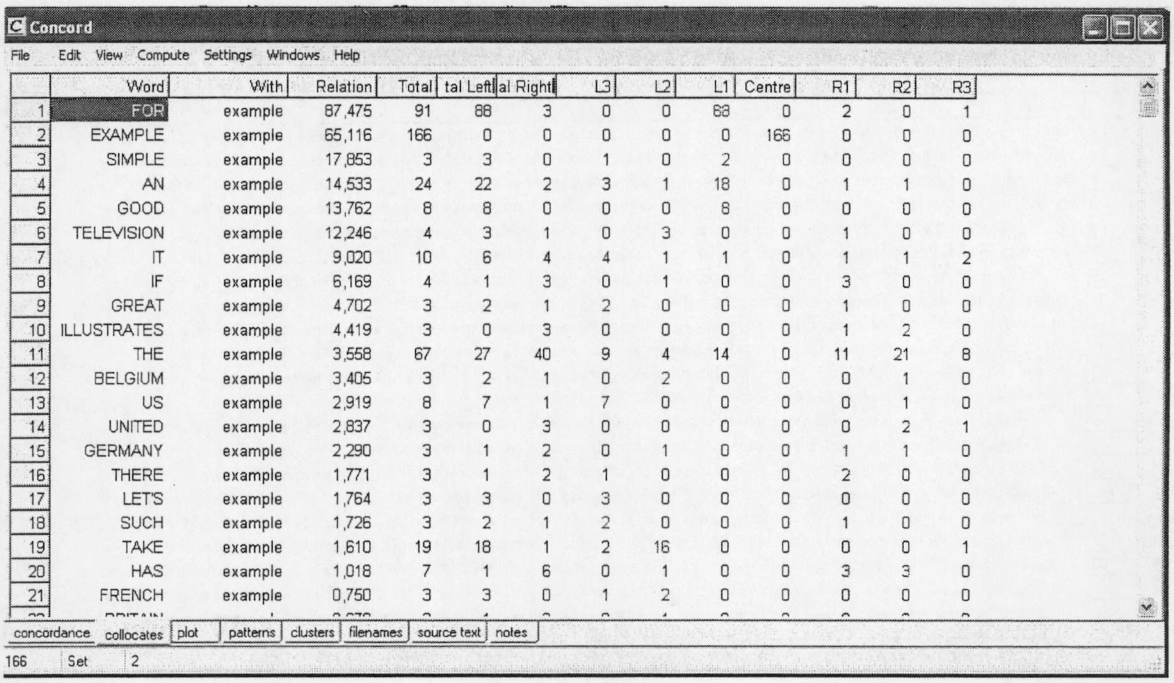


```
30  stupid... but they are unfortunately rarely justified. Let us give a clear example of my opinion: if someone applies for a job which requires
31  I would like to tackle in order to back up my argument is a counter-example. Recently, a TV-channel devoted a special programme to the
32  person to steal, wound, murder, or rape innocent victims. An extreme example of crime caused by people with guns are the riots which
33  feel unified by an artificial economical link ? The extreme example to illustrate this point is the comparaison of the Slave way of
34  of the absence of a complete equality between human beings. A first example clearly illustrates this: the notion of equality in the communist
35  Flexibility and adaptability are very important now. An accountant for example, has to be adaptable if he has to go from a manual to an
36  facets are more important than other ones: social adaptability for example, that is to say the ability to change so as to be suitable for
37  obsessive for quite a lot of people and this happens at any age. For example, when you are a child, your main preoccupation is to buy the
38  the labour market is not the same anymore as twenty years ago for example. Someone who had a university degree could find a job
39  not look down on "pleasure programmes" those that have an aim; for example: to present new songs or artists or the new films which
40  Europe must become one united nation to face problems such as for example, how is Europe 92 going to reduce the number of those who
41  Nowadays, every country has its own industrial centres as, for example, the Ruhr in Germany. But those areas becoming specialized
42  would be able to face the other great nations of the world as for example, the United-States, the Soviet-Union or the Arabic world.
43  had imagination. The stories of the Bronte sisters or Jane Austin for example are full of these elements. They offer their reader the
44  EC have the same kind of objections. A Central European bank, for example, makes bankers in London nervous. What role will they play,
45  are difficult to answer. But it is absolutely certain that a Belgian for example will always feel a Belgian before feeling a European. It would
46  not always been as they are today. What is now called "Belgium", for example, was once Spanish, French, Dutch. The size of the Belgian
47  all over Europe or to have easiler imported products in Belgium for example that our national roots will become looser. Besides it is surely
48  building of organizations struggling for suffrage. In Great Britain, for example, the WSPU (Women's Social and Political Union) led militant
49  an European Nation. There is also a problem of trust. The British, for example, do not want to depend on their continental partners to keep
50  is clear that a murderer has to be treated differently than a burglar for example. Whatever their crimes, the prisoners need to be helped
```

Collocates shows all the words that co-occur with the search-word (here, *example*) and gives their respective total frequencies as well as their frequencies in each position (see Figure 4.4). It also gives an association measure between the search-word and its co-occurrents. This feature was not implemented in the previous version of WST. In Figure 4.4, the most significant co-occurrent of *example* is the preposition *for*. The user can decide to use the mutual information, MI^3, the z-score or the log-likelihood to compute the association measure by selecting the appropriate statistical measure in the *Concord* settings (see section 4.2.3 for a discussion of association measures). This option will not be used in this thesis as association measures computed by WST4 were found not to correspond to results obtained with specialized statistical packages such as the *Ngram Statistics Package* (Banerjee and Pedersen 2003). Moreover, values are not reliable as they sometimes differ when the same analysis is repeated.

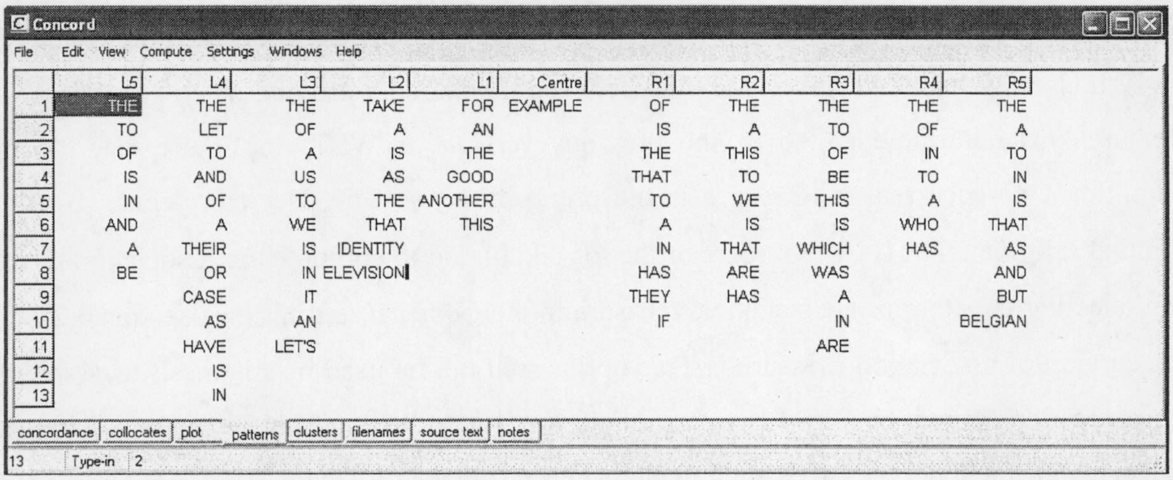
The *patterns* facility also shows co-occurrents of the search-word but does not rely on statistical measures. Each column represents a position within the selected window or horizon and gives a list of the most frequent co-occurrents sorted by decreasing frequency. Thus, in Figure 4.5, the most frequent word found in position L1 (i.e. one word to the left) is *for* and the second one is *an*.

Figure 4.4: Collocates



	Word	With	Relation	Total	tal Left	al Right	L3	L2	L1	Centre	R1	R2	R3
1	FOR	example	87,475	91	88	3	0	0	88	0	2	0	1
2	EXAMPLE	example	65,116	166	0	0	0	0	0	166	0	0	0
3	SIMPLE	example	17,853	3	3	0	1	0	2	0	0	0	0
4	AN	example	14,533	24	22	2	3	1	18	0	1	1	0
5	GOOD	example	13,762	8	8	0	0	0	8	0	0	0	0
6	TELEVISION	example	12,246	4	3	1	0	3	0	0	1	0	0
7	IT	example	9,020	10	6	4	4	1	1	0	1	1	2
8	IF	example	6,169	4	1	3	0	1	0	0	3	0	0
9	GREAT	example	4,702	3	2	1	2	0	0	0	1	0	0
10	ILLUSTRATES	example	4,419	3	0	3	0	0	0	0	1	2	0
11	THE	example	3,558	67	27	40	9	4	14	0	11	21	8
12	BELGIUM	example	3,405	3	2	1	0	2	0	0	0	1	0
13	US	example	2,919	8	7	1	7	0	0	0	0	1	0
14	UNITED	example	2,837	3	0	3	0	0	0	0	0	2	1
15	GERMANY	example	2,290	3	1	2	0	1	0	0	1	1	0
16	THERE	example	1,771	3	1	2	1	0	0	0	2	0	0
17	LET'S	example	1,764	3	3	0	3	0	0	0	0	0	0
18	SUCH	example	1,726	3	2	1	2	0	0	0	1	0	0
19	TAKE	example	1,610	19	18	1	2	16	0	0	0	0	1
20	HAS	example	1,018	7	1	6	0	1	0	0	3	3	0
21	FRENCH	example	0,750	3	3	0	1	2	0	0	0	0	0

Figure 4.5: Patterns

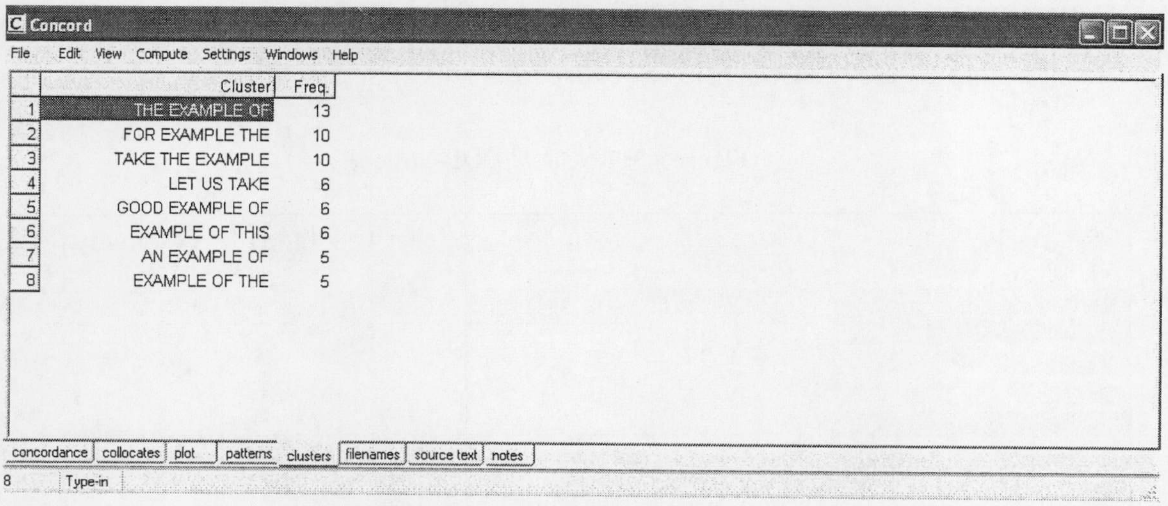


	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	THE	THE	THE	TAKE	FOR	EXAMPLE	OF	THE	THE	THE	THE
2	TO	LET	OF	A	AN		IS	A	TO	OF	A
3	OF	TO	A	IS	THE		THE	THIS	OF	IN	TO
4	IS	AND	US	AS	GOOD		THAT	TO	BE	TO	IN
5	IN	OF	TO	THE	ANOTHER		TO	WE	THIS	A	IS
6	AND	A	WE	THAT	THIS		A	IS	IS	WHO	THAT
7	A	THEIR	IS	IDENTITY			IN	THAT	WHICH	HAS	AS
8	BE	OR	IN	ELEVISION			HAS	ARE	WAS		AND
9		CASE	IT				THEY	HAS	A		BUT
10		AS	AN				IF		IN		BELGIAN
11		HAVE	LET'S						ARE		
12		IS									
13		IN									

Finally, *Clusters* gives a list of the most frequent n-grams or lexical bundles (cf. 1.4.1) that contain the search-word sorted by decreasing frequency. A number of settings can be adjusted:

- The length of the n-grams: Figure 4.6 shows a cluster analysis of 3-word sequences including the search-word *example* but 2-word sequences or larger clusters can also be extracted.
- The horizon within which clusters are extracted: it is possible to retrieve clusters that appear to the left or right of the search-word but which do not include it.
- Minimum frequency of the clusters.

Figure 4.6: Clusters



	Cluster	Freq.
1	THE EXAMPLE OF	13
2	FOR EXAMPLE THE	10
3	TAKE THE EXAMPLE	10
4	LET US TAKE	6
5	GOOD EXAMPLE OF	6
6	EXAMPLE OF THIS	6
7	AN EXAMPLE OF	5
8	EXAMPLE OF THE	5

Further information on WST4 and the types of analysis that can be performed with the software program can be found on Mike Scott's webpage http://www.lexically.net and in Scott and Tribble (2006).

4.2.2.2. BNC Web-based concordancers

The BNC was mainly accessed through the *BNCWeb* but occasional use was also made of the *Variation in English Words and Phrases* (VIEW) interface.

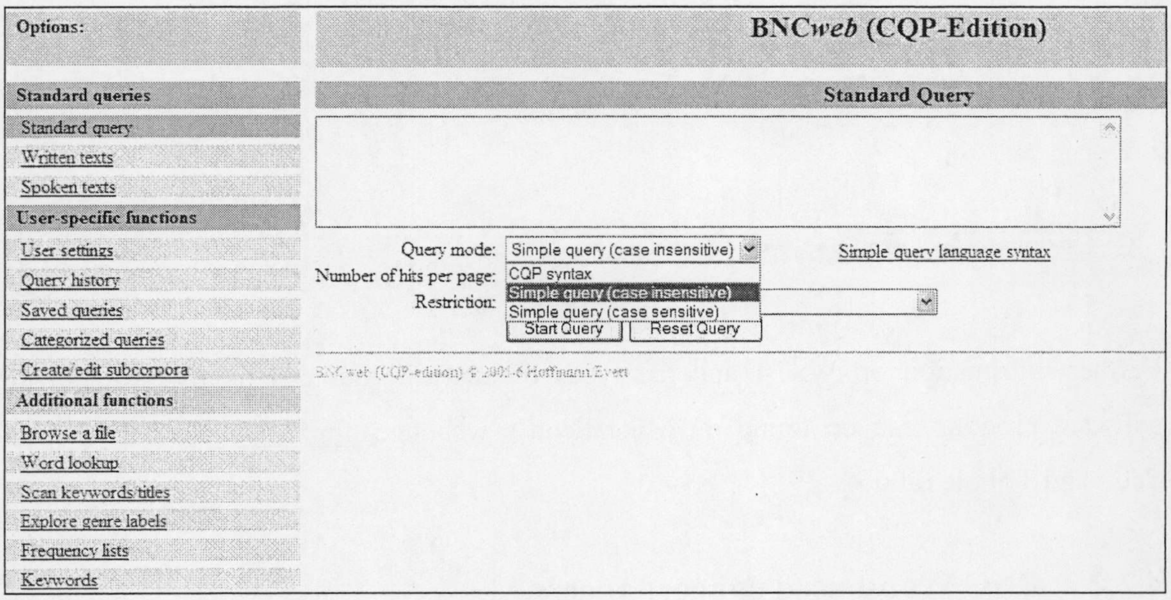
I) BNCweb (CQP edition)

The *BNCweb (CQP*[112] *edition)* is currently being developed by Stefan Evert and Sebastian Hoffmann. It is the result of a "marriage of two corpus tools" (Hoffmann and Evert 2006), i.e. the *BNCweb*, a web-based client developed at the University of Zurich which allows users to access the BNC by means of a Web browser (cf. Lehmann et al 2000) and the *IMS Corpus Workbench*, a generic query engine designed to process large corpora developed at the University of Stuttgart and whose main component is the corpus query processor CQP (see http://ims.uni-stuttgart.de/projekte/CorpusWorkbench for more information). The BNC*web* (CQP edition) combines the strengths of both software packages while overcoming their respective limitations (cf. Hoffmann and Evert 2006). It is a marriage between the efficiency and flexibility of CQP queries and the user-friendliness of BNC*web* and its wide range of query options and display facilities. Implementing the CQP in BNC*web* makes it possible to conduct more flexible and sophisticated searches. One of the major limitations of CQP, i.e. its

[112] CQP = Corpus Query Processor

lack of user-friendliness, was overcome by developing a simplified query language in parallel with the CQP language. Figure 4.7 shows that the user can decide to search the BNC by means of a simple query (case sensitive or case insensitive) or a CQP query.

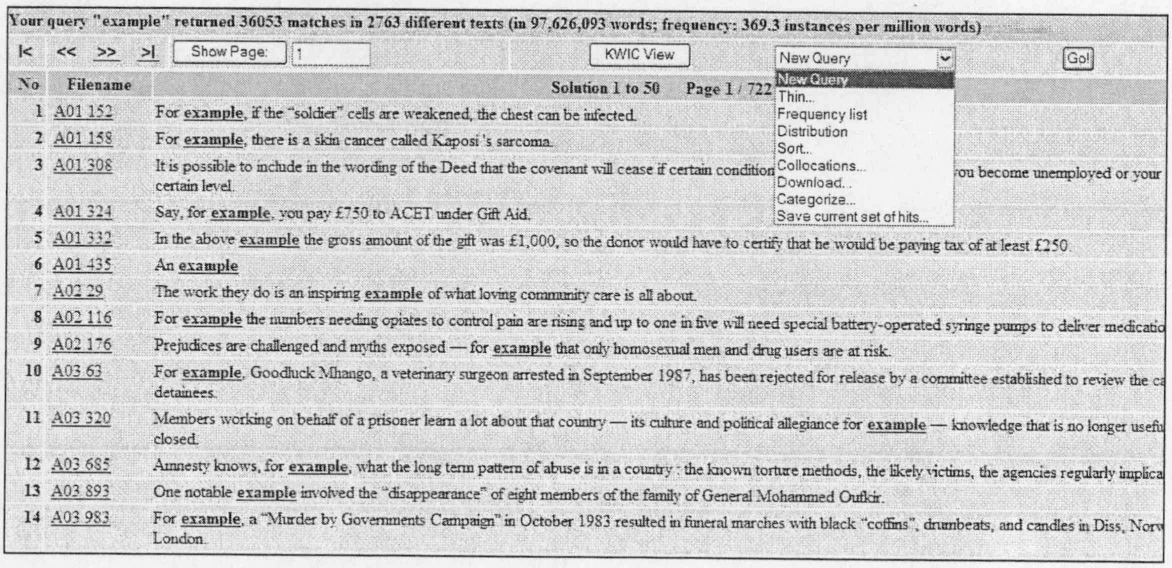
Figure 4.7: BNC*web* (CQP-edition)



Query mode: Simple query (case insensitive)

Figure 4.8 displays the first fourteen instances of the query result for the search word *example*. The title bar gives the total number of matches as well as the number of texts in which the search word appears (see the notion of 'range' in 5.3.1.2) and its relative frequency per million words. For each concordance line, the context can be enlarged by clicking on the search word and bibliographical information can be obtained by clicking on the filename (left column). A number of options are available from a query result page (see the pull-down menu on Figure 4.8). Concordance lines can be alphabetically sorted to the left and right of the node by means of the *sort* option. The *distribution* option gives the distribution of the search word over most text features, e.g. text type, text domain, age of author, sex of author, perceived level of difficulty, etc. Unfortunately, it does not give the distribution over David Lee's genre categories of the BNC texts (cf. 4.1.2.1) and we had to use the VIEW interface (see below) to get this type of information.

Figure 4.8: BNC*web* (CQP-edition) – search word: *example*

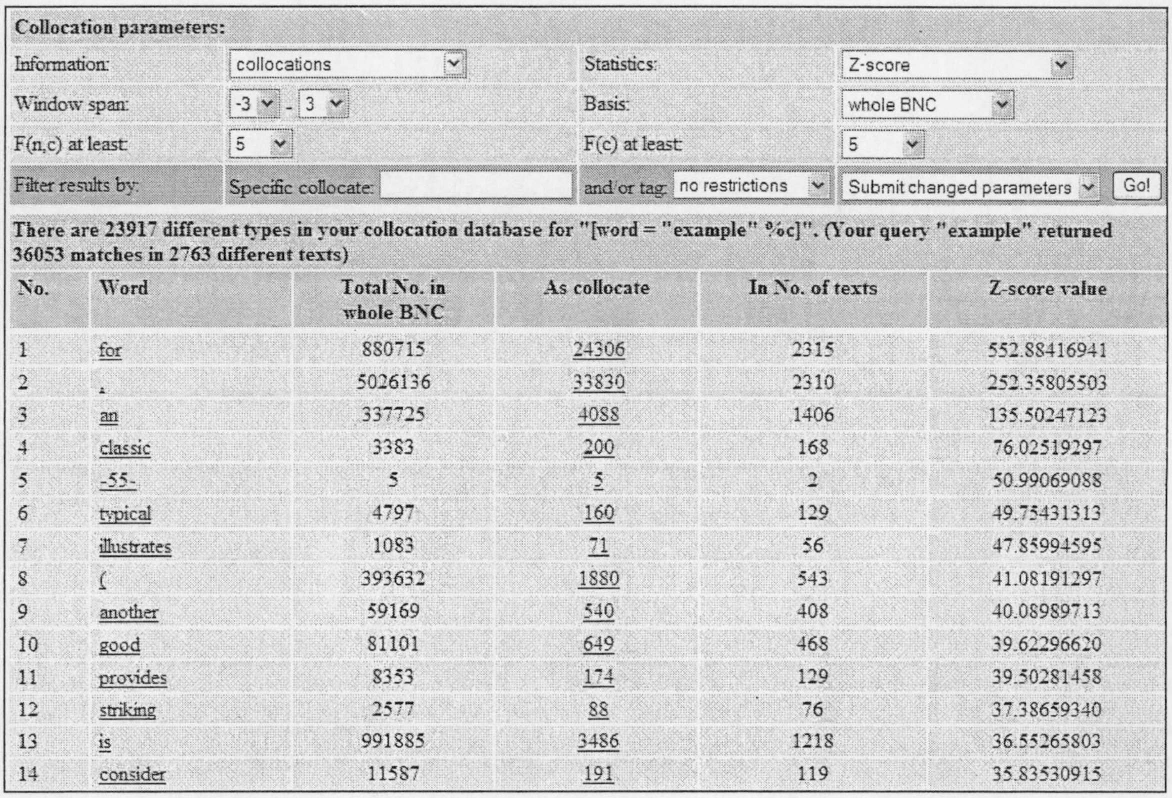


Your query "example" returned 36053 matches in 2763 different texts (in 97,626,093 words; frequency: 369.3 instances per million words)

No	Filename	Solution 1 to 50 Page 1 / 722
1	A01 152	For **example**, if the "soldier" cells are weakened, the chest can be infected.
2	A01 158	For **example**, there is a skin cancer called Kaposi 's sarcoma.
3	A01 308	It is possible to include in the wording of the Deed that the covenant will cease if certain condition ou become unemployed or your certain level.
4	A01 324	Say, for **example**, you pay £750 to ACET under Gift Aid.
5	A01 332	In the above **example** the gross amount of the gift was £1,000, so the donor would have to certify that he would be paying tax of at least £250.
6	A01 435	An **example**
7	A02 29	The work they do is an inspiring **example** of what loving community care is all about.
8	A02 116	For **example** the numbers needing opiates to control pain are rising and up to one in five will need special battery-operated syringe pumps to deliver medicatio
9	A02 176	Prejudices are challenged and myths exposed — for **example** that only homosexual men and drug users are at risk.
10	A03 63	For **example**, Goodluck Mhango, a veterinary surgeon arrested in September 1987, has been rejected for release by a committee established to review the ca detainees.
11	A03 320	Members working on behalf of a prisoner learn a lot about that country — its culture and political allegiance for **example** — knowledge that is no longer usefu closed.
12	A03 685	Amnesty knows, for **example**, what the long term pattern of abuse is in a country : the known torture methods, the likely victims, the agencies regularly implica
13	A03 893	One notable **example** involved the "disappearance" of eight members of the family of General Mohammed Oufkir.
14	A03 983	For **example**, a "Murder by Governments Campaign" in October 1983 resulted in funeral marches with black "coffins", drumbeats, and candles in Diss, Norw London.

The *Collocations* option gives significant co-occurrents of the search word on the basis of a number of association measures. Users can decide to use mutual information, MI^3, z-score, log-likelihood or log-log (see section 4.2.3). They can also sort co-occurrents by decreasing frequency. A number of other settings are customisable, e.g. maximum window span, minimum frequency of the co-occurrence f(n,c), minimum frequency of the co-occurrent f(c), inclusion of lemma and part-of-speech information, etc. Figure 4.9 displays a collocation query result. Significant co-occurrents are sorted by decreasing z-score values (right column). The frequency of the co-occurrence is given together with the number of texts in which it appears.

Figure 4.9: BNC*web* (CQP-edition) – *Collocations* option [search item: *example*]



Collocation parameters:

Information: collocations

Statistics: Z-score

Window span: -3 - 3

Basis: whole BNC

F(n,c) at least: 5

F(c) at least: 5

Filter results by: Specific collocate:

and/or tag: no restrictions

There are 23917 different types in your collocation database for "[word = "example" %c]". (Your query "example" returned 36053 matches in 2763 different texts)

No.	Word	Total No. in whole BNC	As collocate	In No. of texts	Z-score value
1	for	880715	24306	2315	552.88416941
2	,	5026136	33830	2310	252.35805503
3	an	337725	4088	1406	135.50247123
4	classic	3383	200	168	76.02519297
5	-55-	5	5	2	50.99069088
6	typical	4797	160	129	49.75431313
7	illustrates	1083	71	56	47.85994595
8	(	393632	1880	543	41.08191297
9	another	59169	540	408	40.08989713
10	good	81101	649	468	39.62296620
11	provides	8353	174	129	39.50281458
12	striking	2577	88	76	37.38659340
13	is	991885	3486	1218	36.55265803
14	consider	11587	191	119	35.83530915

Simple queries allow users to search for word forms (e.g. *issue*), lemmas (e.g. *{issue/N}*), word sequences with or without wildcards (e.g. *it was * of him to*) or co-occurrent items (e.g. *{such} <10> {etc.}*). CQP queries were designed to allow for more complex queries involving, among others, wild cards and CLAWS part-of-speech tags (cf. Appendix 5.1). Table 4.4 gives a few examples of CQP queries and their explanations.

Table 4.4: Examples of CQP queries

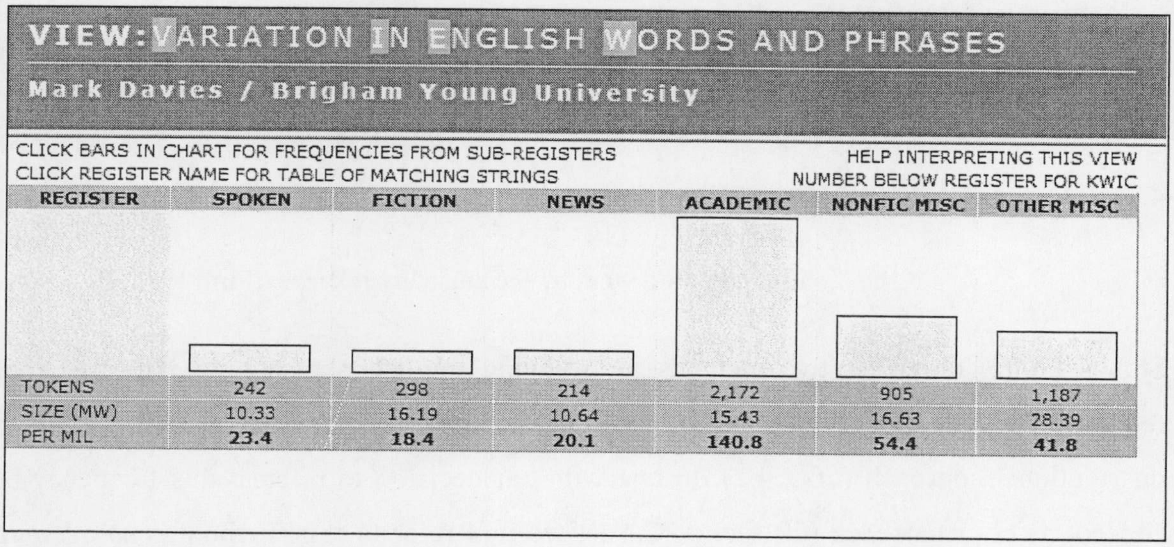
"a" [pos="N.*"] "of"	All occurrences of the collocational framework “a NOUN of”, e.g. *a number of, a result of, a variety of, a series of*
[pos="JJ"]	All occurrences of adjectives
[word="example.*" & pos!="MWU"]	All occurrences of the word *example* that do not appear in multi-word units
"as" [pos="NP.*"] []{0,2} [pos="V.*"]	All occurrences of the word *as* directly followed by a proper noun (NP) followed by a verb. The proper noun and the verb can be separated by maximum two words.
"look" [pos!=VB.*”] {0,10} "up|down"	All occurrences of *look* followed either by *up* or *down* with at most ten non verbal word forms in between

Features of BNC*web* (CQP edition) that were not used in this thesis but which may be of interest to the reader include frequency lists, keywords and user’s defined sub-corpora.

II) Variation in English Words and Phrases (VIEW)

Like other BNC web-interfaces, *Variation in English Words and Phrases* (henceforth VIEW) (http://view.byu.edu/) allows researchers to search for words and phrases as well as their co-occurrents within a ten-word window. Words can also be searched for by means of part-of-speech tags and wildcards. One of the unique features of the web-interface developed by Mark Davies at Brigham Young University is that it allows researchers to compare the frequency of words in specific genres and sub-genres as defined by David Lee (see section 4.1.2.1).[113] For example, Figure 4.10 gives the frequency distribution of the noun *conclusion* across the six main genres available in the BNC and shows that it is much more frequent in academic texts than in spoken English, newspaper articles or fiction texts.

Figure 4.10: VIEW – genre distribution of the word 'conclusion'



VIEW: VARIATION IN ENGLISH WORDS AND PHRASES

Mark Davies / Brigham Young University

CLICK BARS IN CHART FOR FREQUENCIES FROM SUB-REGISTERS
CLICK REGISTER NAME FOR TABLE OF MATCHING STRINGS

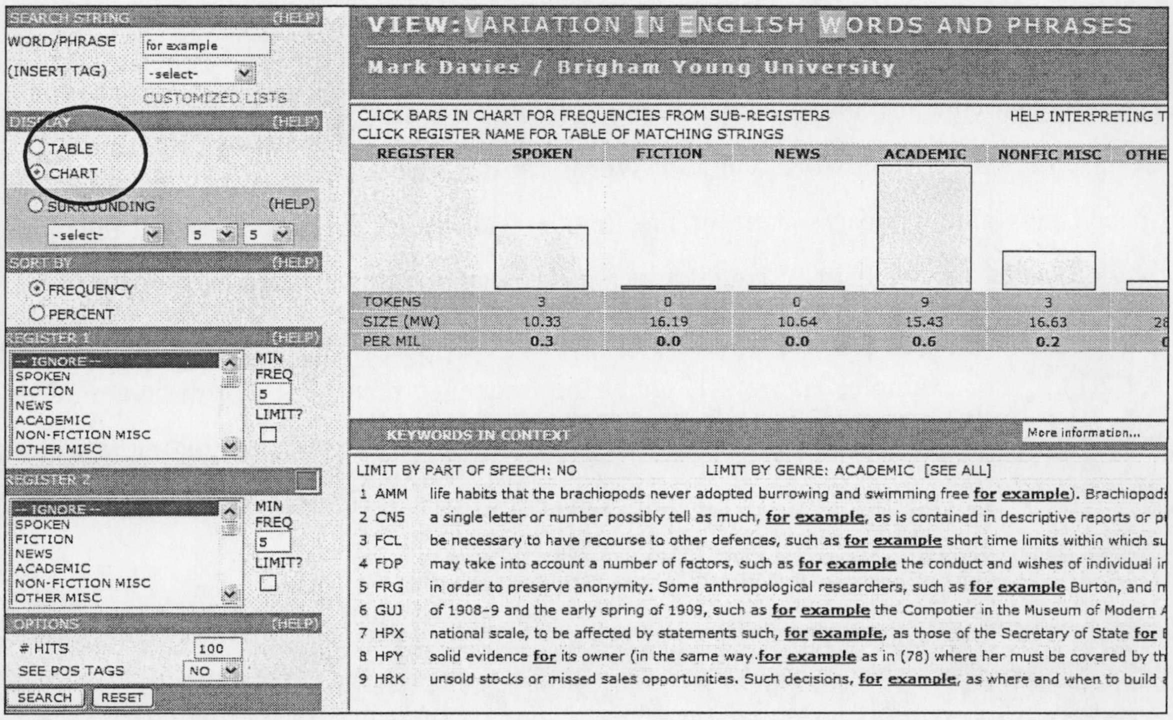
HELP INTERPRETING THIS VIEW
NUMBER BELOW REGISTER FOR KWIC

REGISTER	SPOKEN	FICTION	NEWS	ACADEMIC	NONFIC MISC	OTHER MISC
TOKENS	242	298	214	2,172	905	1,187
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28.39
PER MIL	**23.4**	**18.4**	**20.1**	**140.8**	**54.4**	**41.8**

It should be noted that the option does not function properly with word sequences that belong to CLAWS's dictionary of multi-word units, e.g. *for example* and *except for* (cf. section 5.2.2.2.1). When the chart display option is used, only the occurrences of the multi-word units that were not recognized by CLAWS and were therefore not accurately tagged are taken into account for the comparison. Thus, in Figure 4.11, *for example* is reported to occur 9 times in academic prose[114].

[113] Note that Mark Davies uses the term 'register' to refer to David Lee's genre classification.
[114] The bug was reported to the author.

Figure 4.11: VIEW – genre distribution of phrasemes: a bug?



WORD/PHRASE: for example
(INSERT TAG): -select-
○ TABLE
● CHART
○ SURROUNDING (HELP): -select- 5 5
● FREQUENCY
○ PERCENT
REGISTER 1: -- IGNORE -- SPOKEN FICTION NEWS ACADEMIC NON-FICTION MISC OTHER MISC; MIN FREQ 5; LIMIT? □
REGISTER 2: -- IGNORE -- SPOKEN FICTION NEWS ACADEMIC NON-FICTION MISC OTHER MISC; MIN FREQ 5; LIMIT? □
HITS: 100
SEE POS TAGS: NO

CLICK BARS IN CHART FOR FREQUENCIES FROM SUB-REGISTERS HELP INTERPRETING T
CLICK REGISTER NAME FOR TABLE OF MATCHING STRINGS

REGISTER	SPOKEN	FICTION	NEWS	ACADEMIC	NONFIC MISC	OTHE
TOKENS	3	0	0	9	3	
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28
PER MIL	**0.3**	**0.0**	**0.0**	**0.6**	**0.2**	0

KEYWORDS IN CONTEXT

LIMIT BY PART OF SPEECH: NO LIMIT BY GENRE: ACADEMIC [SEE ALL]

1 AMM life habits that the brachiopods never adopted burrowing and swimming free **for example**). Brachiopods
2 CN5 a single letter or number possibly tell as much, **for example**, as is contained in descriptive reports or p
3 FCL be necessary to have recourse to other defences, such as **for example** short time limits within which su
4 FDP may take into account a number of factors, such as **for example** the conduct and wishes of individual ir
5 FRG in order to preserve anonymity. Some anthropological researchers, such as **for example** Burton, and n
6 GUJ of 1908-9 and the early spring of 1909, such as **for example** the Compotier in the Museum of Modern A
7 HPX national scale, to be affected by statements such, **for example**, as those of the Secretary of State **for**
8 HPY solid evidence **for** its owner (in the same way **for example** as in (78) where her must be covered by th
9 HRK unsold stocks or missed sales opportunities. Such decisions, **for example**, as where and when to build

4.2.3. A quantitative approach to phraseology

> You shall know a word by the company it keeps. (Firth 1957)

Stubbs (2002) describes two methods for studying English phraseology. The first method relates to a purely frequency-based or statistical concept of collocation as defined within the distributional approach (cf. 2.4.2). In line with our decision to refer to this frequency-based concept as co-occurrence (cf. section 2.4.3), we will refer to this method as ***co-occurrence analysis***. The second method consists in extracting the most frequent phrases, in the sense of strings of uninterrupted word-forms, i.e. clusters or lexical bundles, from a corpus. It is most suited for the extraction of fixed phrasemes or collocational frameworks if wildcards are used. The reader is referred to Altenberg (1998) and De Cock (2003, 2004) for examples of this type of analysis as well as a discussion of its strengths and limitations. In this section, we focus on co-occurrence analysis.

Co-occurrence analysis basically involves three main stages (cf. Evert and Kermes 2003). First, the corpus is **(optionally) pre-processed and annotated**. Then co-occurrents are extracted within a given **span** and their joint frequencies are computed. **Filters** can be used to improve accuracy or limit the number of resulting word pairs. Typical filters include:

- A **minimum threshold** for the co-occurrence frequency under which word pairs are not analyzed.
- A **list of stopwords** that removes all word pairs that comprise a stopword as found in that list. Stopwords are typically grammatical words such as *a, the, their, us, is,* and *be*.
- A **list of patterns**, e.g. noun + verb, adjective + noun, to be analysed if a relational model of co-occurrence is used.

The third step is to choose an **association measure** to rank co-occurrence data.

In this section, we shall first discuss parameters that can have an important influence on the type of results obtained by the co-occurrence extraction procedure. We will then highlight two major shortcomings of raw frequency data for the extraction of co-occurrence data before reviewing a number of association measures that are frequently used to highlight significant word pairs. Co-occurrence data for the noun *evidence* in the BNC will be used to describe these measures and compare them. Finally, we will describe the parameters and association measures used in this thesis.

4.2.3.1. Parameters

A number of parameters may influence the outcome of a co-occurrence analysis. They include degrees of corpus pre-processing and annotation, the size of the co-occurrence window or span used and the use of filters such as a minimum frequency threshold or a stopword list.

4.2.3.1.1. Corpus pre-processing and annotation

Co-occurrence analysis can be carried out on raw corpus data or annotated data (i.e. lemmatized or POS-tagged data)[115]. Although a large proportion of collocations have been found to exhibit syntagmatic relations at the level of lexemes, other types of phrasemes exhibit syntagmatic relations between word forms (cf. section 2.3.1). Clear (1993:277) claims that "collocations are very rarely lost because of the lack of lemmatisation: one of the inflected forms will appear as a significant collocate, and the potential for the other forms in the paradigm to collocate will be apparent to the human analyst." Similarly, Bartsch (2004:111) states that "[f]rom a statistical point of view it appears that significant collocations are still highlighted even if they are distributed widely over the different word forms of a

[115] Section 5.2.2.1 discusses the issue of corpus annotation, its pros and cons, etc.

lexeme." On the other hand, it may also be argued that co-occurrence data based on word forms provide less information on preferred word pairs as they include a lot of repeated data, i.e. different word forms of the same lemma. This disadvantage is most severe for word pairs including verbs. Table 4.5 compares the first 20 verb forms and verb lemmas that co-occur most significantly with the noun *evidence* within a window of three words to the left and to the right of the node (3L-3R). The association measure used is the log-likelihood. Verb form co-occurrents (left column) only comprise forms of nine different lemmas which are repeated from two (e.g. *suggest/suggests, show/shows*) to four times (*give/giving/gave/given*). These nine verbs are also extracted by a co-occurrence analysis based on lemmatized data. As shown in the right part of Table 4.5, they are first ranked in the list of verb lemma co-occurrents of the noun *evidence* (in bold). A co-occurrence analysis based on lemmatized data also retrieves verbs such as *produce, present, adduce,* and *obtain*. There is no instance of a verb co-occurrent that is retrieved by a co-occurrence analysis based on raw data and not by a similar analysis on lemmatized data.

Table 4.5: The noun *evidence* and its verb form and verb lemma co-occurrents

'*evidence*' + verb forms			'*evidence*' + verb lemmas		
Verb form	**f(c)**	**f(n,c)**	**Verb lemma**	**f(c)**	**f(n,c)**
is	986618	3908	***be***	3244400	5686
was	883602	1499	***give***	125302	1051
suggests	6667	390	***suggest***	28246	739
give	43973	429	***provide***	51472	647
suggest	8776	282	***was***[116]	883602	1499
provide	22171	332	***support***	18597	398
support	15723	274	***have***	1319155	1507
be	651292	806	***show***	58832	473
show	20575	211	***find***	95790	371
have	461408	579	produce	30045	211
giving	12234	180	present	14079	170
gave	21975	202	adduce	185	75
found	47221	236	obtain	12689	142
provides	8353	143	indicate	12198	128
has	256861	387	hear	34747	161
given	37235	195	cite	2555	67
shows	10428	132	gather	4971	75
had	421199	438	base	19201	97
supporting	2420	86	require	28257	109
provided	12851	123	offer	28780	109

[116] Note that the occurrences of the word form '*was*' are erroneously counted separately from other word forms of the lemma 'be' in the BNC.

Partly related to lemmatisation is the issue of part-of-speech and syntactic annotation. **Syntactic annotation** allows for an analysis of co-occurrent words which appear in a specific (syntactic) structure, e.g. pre-modifying adjective + noun, verb + object noun, subject noun + verb. Kilgarriff and his colleagues recently developed the *Sketch Engine* to provide lexicographers with "corpus-based summaries of a word's grammatical and collocational behaviour" (Kilgarriff et al. 2004:105). Table 4.6 gives a sample of the word sketch for the noun *evidence* based on the BNC. Words are lemmatised. The association measure implemented in the *Sketch Engine* is the log-log. This measure differs from the log-likelihood in that it gives less prominence to highly frequent co-occurrents (cf. section 4.2.3.3.5).

Table 4.6: A sample of the word sketch for the noun *evidence*

object of	**5522**	**2.7**	**subject of**	**1982**	**1.9**	**adj. modifier**	**6173**	**2.4**
adduce	64	46.31	suggest	412	51.95	circumstantial	83	54.25
provide	622	40.61	support	117	33.69	conclusive	94	51.82
give	941	39.04	indicate	82	32.04	empirical	163	50.61
obtain	130	29.55	point	59	29.77	anecdotal	67	50.26
gather	68	28.89	show	146	28.6	ample	91	45.51
produce	187	28.44	exist	43	26.81	archaeological	75	41.35
find	334	27.62	emerge	40	26.76	forensic	57	40.86
present	120	27.49	accumulate	20	26.33	further	283	40.76
hear	144	26.96	implicate	16	25.85	sufficient	148	39.12
collect	62	24.57	relate	52	24.82	supporting	67	38.98
n. modifier	**820**	**0.4**	**pp_of-p**	**3614**	**3.3**	**pp_on-p**	**282**	**1.4**
documentary	115	59.59	senses	24	23.66	oath	9	24.54
hearsay	30	47.97	efficacy	13	20.82	behalf	9	22.15
expert	62	36.45	infection	25	19.99	issue	10	13.54
affidavit	21	35.9	abuse	26	19.89	matter	7	11.52
dating	19	32.79	damage	31	18.96	subject	7	11.46
research	72	30.47	ischaemium	6	18.7	point	9	11.27
fossil	20	29.48	witness	20	18.6	ground	5	10.15
confession	14	26.24	nephropathy	5	18.08	nature	5	9.32
parol	5	25.92	competence	15	17.86	effect	6	9.1
video	21	22.34	disease	34	17.08	side	5	7.31
pp_in-p	**393**	**0.8**	**pp_obj_to-p**	**187**	**0.7**	**pp_obj_by-p**	**248**	**1.6**
case	49	26.39	relate	15	21.81	support	66	38.83
court	41	25.69	point	12	21.49	unsupported	10	34.21
trial	21	24.61	regard	7	20.73	substantiate	5	20.74
proceedings	14	22.43	listen	9	20.12	contradict	5	18.55
prosecution	7	17.04	refer	8	16.25	convince	6	17.31
favour	6	15.93	reference	6	13.86	justify	5	13.31
chief	6	13.41	apply	6	12.33	prove	6	12.27
form	9	10.16	add	6	11.43	confirm	5	11.99
action	6	8.57	give	8	8.15	establish	5	9.83
area	7	7.05	make	5	4.11	suggest	5	9.42

Although syntactic parsing is arguably the most relevant type of annotation for co-occurrence analysis, co-occurrence studies are still largely based on **part-of-speech tagged corpora**, i.e.

corpora that provide information about the grammatical nature of a word (e.g. Bouma and Villada 2002; Bartsch 2004). Very few parsed corpora are readily available and parsers are rarely accompanied by user-friendly search tools such as the *Sketch Engine*. As the ICLE sub-corpora and LOCNESS are not available in parsed formats, co-occurrence data presented in chapter 6 are based on **lemmatised** and **part-of-speech** tagged corpora. The corpora were lemmatized and part-of-speech tagged with the help of *Wmatrix* (cf. section 5.2.2.2). Co-occurrence data based on part-of-speech tagged corpora is regarded as **an approximation to the relational model of co-occurrences** where the co-occurrent words appear in a specific (syntactic) structure, e.g. adjective + noun, verb + object noun (cf. section 4.2.3.3.1). Results of co-occurrence analyses within a given span, i.e. **positional co-occurrences**, were sorted by part-of-speech pairs (e.g. adjective + noun, noun + verb, adverb + adjective, etc.) and manually analysed to keep only those co-occurrent words that stand in a grammatical relation to the node, i.e. **relational co-occurrences**.

4.2.3.1.2. Span

The window size or **span** represents the maximum distance that can separate a word and its co-occurrent words in a co-occurrence analysis. The word under investigation is often referred to as the **node**. The words around the node are described in terms of their distance from it in number of words to the left or right of it. Thus, in the following sentence, the word *theoretical* is situated in position 3L (third word to the left of the node):

There	*has*	*been*	*much*	*theoretical*	*and*	*applied*	*research*	*on*	*collocations.*
		5L	4L	3L	2L	1L	node	1R	2R

Jones and Sinclair (1974) showed that 95% of the collocational influence of the node takes place within a 4L-4R span. Following Jones and Sinclair's very influential paper, numerous co-occurrence studies use a 4L-4R (or 4:4) span (e.g. Sinclair 1987; Sinclair 1991; Stubbs 2002; Xiao and McEnery 2006).

The selected span can have a significant impact on the outcome of a co-occurrence analysis. Clear (1993) compares co-occurrence data for the word *order* obtained within a span of 2:2 vs. 6:6 and concludes that "[t]he items which are lost from the 2:2 listing in the wider span (*tall* and *working*) seem to be worth more than the extra items which are gained" (Clear 1993:290). Moreover, collocations identified within a span of 2:2 are "lost amongst the increased volume of data gathered by the wider span setting" (ibid) and "fall much further down the significance ranking" (ibid). Similarly, Bartsch first selected a span of 5:5 as a starting point for her co-occurrence analysis but she later reduced it to a 3:3 span as a wider setting introduces "too much noise (i.e. irrelevant data) that obscures the statistical significance of other word combinations" (Bartsch 2004:93).

Most studies have used the same window size for all types of nodes, regardless of their part of speech. A notable exception is Berry-Roghe (1973) who proposes a span of 4:4 for all words except for adjectives for which a 2:2 span seems more appropriate. Similarly, most studies have used a symmetrical span, i.e. a span of the same size in either direction. Mason (2000:269) is strongly critical of these choices as "they don't consider that different words may have a different degree of influence on their lexical environment". He proposes "a variable span which is individual to each word" based on lexical gravity, i.e. the restriction a word places on the variability of its context (cf. also Cantos and Sánchez 2001). Unfortunately, no tool is currently available to carry out analyses of lexical gravity.

Left and right co-occurrents are often merged together in the literature. For example, left and right verb co-occurrents of a noun are often added up so as to retrieve both active and passive structures of the same collocation (e.g. *he **made** an important **decision**; a **decision** was **made***). Table 4.7 compares the first twenty verb co-occurrents of the noun *decision* for the following spans: 3L-1L, 3L-3R and 1R-3R (association measure: log-likelihood). Results for 3L-1L and 1R-3R differ significantly:

- Only 6 verbs, i.e. *make, take, reach, affect, follow* and *announce*, are significant co-occurrents of the noun *decision* in both left and right position.
- Left-hand only significant co-occurrents (3L-1L) include *influence, challenge, delay, reconsider, overturn, welcome* and *defer*.
- Right-hand only co-occurrents (1R-3R) mainly consist of high-frequency verbs and modals, e.g. *be, have, will, should, can* and *must*.

There is a strong case here for analysing left and right significant co-occurrents separately, thus distinguishing the phraseological environment to the left vs. right of a word.

Table 4.7: A comparison of span sizes: verb co-occurrents of 'decision'

3L-1L	f(n,c)	3L-3R	f(n,c)	1R-3R	f(n,c)
make	2455	make	4249	make	1794
take	601	be	4216	be	3436
reach	241	take	1548	take	947
influence	134	was	1339	was	1027
reverse	106	reach	395	have	1014
follow	152	have	1392	reach	154
announce	100	affect	214	affect	133
challenge	61	influence	162	will	400
affect	81	reverse	132	base	133
delay	49	base	200	should	178
reconsider	35	will	506	may	178
accept	82	follow	239	would	255
overturn	36	announce	149	can	243
regret	43	come	267	withdraw	52
implement	52	may	225	must	121
quash	31	should	222	allow	84
welcome	49	can	323	close	60
defend	44	would	324	follow	87
defer	28	overturn	56	proceed	36
come	156	delay	72	announce	49

In this thesis, left and right co-occurrents will thus be analysed separately. Two spans will be used, i.e. **3L-1L and 1R-3R**, for most types of word pairs except for adjective-noun co-occurrences, for which the following spans will be used: **2L-1L and 1R-2R** (cf. Berry-Roghe 1973). A two-word span to the left makes it possible to retrieve adjective collocates such as *strong* and *considerable* in *strong historical evidence* and *considerable experimental evidence.* Although a two-word span to the right extracts many adjectives that do not stand in a grammatical relation to the noun under study, it allows for the extraction of adjectives that are more frequently used in predicative position as illustrated in the following sentences:

4.1. *There is a great deal of written* ***evidence available*** *on the subject for those who care to take a deeper look.* (BNC)

4.2. *In cases where there was more than one accused, it was not unusual to discover that some* ***evidence*** *was* ***admissible*** *against one accused only and not the other accused.* (BNC)

4.3. *The* ***evidence*** *is* ***overwhelming*** *that he did not at this stage contemplate a coalition government.* (BNC)

4.2.3.1.3. Filters

Two types of filters are often used in co-occurrence analyses: (1) stopword lists are used to improve accuracy and (2) minimum frequency thresholds are used to limit the amount of resulting word pairs. **Stopword lists** are mainly used in the field of information retrieval and automatic term extraction (cf. EAGLES's article on multi-word recognition and extraction at http://www.ilc.cnr.it/EAGLES96/rep2/node38.html) but they are also sometimes used in collocation extraction (e.g. Pazos Bretaña 2005; Piao et al. 2006). They are used to filter out the most common words in the language, i.e. function words (cf. section 2.3.1), from co-occurrent lists so that only semantically meaningful word pairs remain. Stopword lists may include determiners (*a, the, their*), pronouns (*us, I, we*), prepositions (*up, down, out, by*), high frequency verbs (*have, be, become*), conjunctions (*and, although, but*) and adverbs (*accordingly, however*) (cf. Cornell's stoplist for information retrieval at ftp://ftp.cs.cornell.edu/pub/smart/english.stop and the Collocate Finder website at http://ell.phil.tu-chemnitz.de/collCollect/user/nph-index.cgi).

Not all linguists agree on the use of stopword lists in co-occurrence analysis as they may hide preferred collocational or colligational patterns of words. Clear writes that "[a] manual analysis of the concordance for *order* shows that *in* is by far the most significant on the left and *to* on the right, from which I determine that the fixed phrase *in order to* is the massively dominant collocational pattern for this word" (Clear 1993:284). Another argument against the use of a stopword list is that depending on the association measure used, function words may be given the lowest scores (cf. Grefenstette et al. 1994). For these two reasons, **no stopword list will be used in this thesis.**

The second parameter which can exert significant influence on the result of a co-occurrence analysis is the setting of a **minimum frequency threshold** under which word pairs are not considered. Published research does not reveal any clear consensus on the subject. There is no established standard value for minimum frequency thresholds and values are often still determined by trial and error. Clear (1993) carries out a co-occurrence analysis of the word *taste* in a 25 million word corpus and discards all word pairs that are observed fewer than three times. Bouma and Villada (2002), by contrast, study collocational prepositional phrases in a 16 million word corpus and compare co-occurrence data obtained with two different minimum frequency threshold values, i.e. 10 and 40.

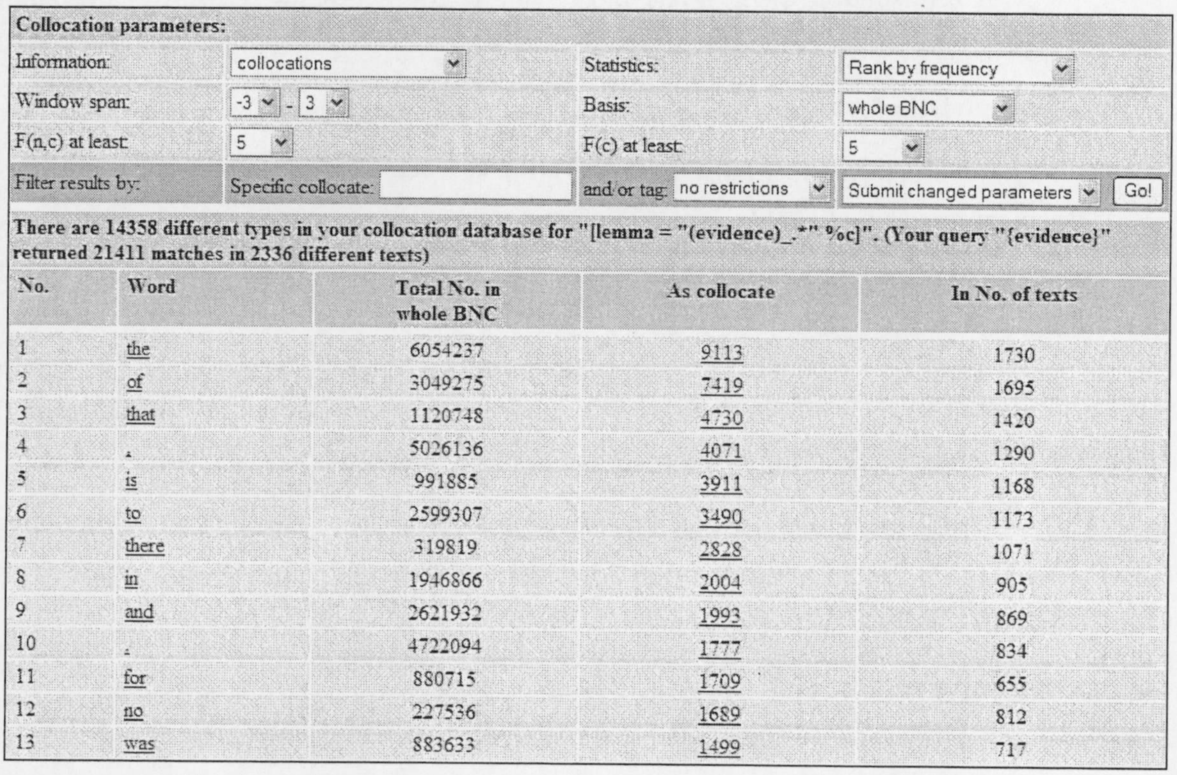
In his study of the statistics of word co-occurrences, Evert argues that "[d]ata with cooccurrence frequency f < 3, i.e. the hapax and dis-legomena, should always be excluded from the statistical analysis" (Evert 2004: 133) as expected frequencies (cf. 4.2.3.3.1) and p values (cf. 4.2.3.3.3) for low frequency words are distorted in unpredictable ways. The author advocates the use of a **minimum frequency threshold of 5**, which is adopted in this thesis to retrieve statistically significant word co-occurrences from the BNC.

4.2.3.2. Raw frequency

The simplest way of identifying collocations in corpora is to look at **raw frequencies** of co-occurrence. As shown in Figure 4.12, the word *the* ranks first when co-occurrents of the word *evidence* are sorted by decreasing frequency. The raw frequency is 9113 (column 'as collocate'), which means that *evidence* and *the* co-occur 9113 times in the whole BNC. It is directly followed by other high-frequency words such as *of, that, is, to, there* and *in* and by punctuation marks. Evert (2004:12) lists two major shortcomings of raw co-occurrence data. The first one relates to the fact that raw frequencies are often not meaningful as an indicator for the amount of 'glue' between two words. Raw frequency of co-occurrence foregrounds the most frequent words despite the fact that they may co-occur by sheer coincidence and may not provide conclusive evidence of significant collocation patterns. Barnbrook (1996:88) claims that, on the basis of frequency of co-occurrence alone, high-frequency words "would be found to collocate strongly with most keywords [i.e. *node words*]." A statistical interpretation of the frequency data is therefore necessary to determine the degree of association between the words.

The second shortcoming underlined by Evert (2004) relates to the fact that co-occurrence data "only provide information about the one particular corpus they are extracted from" (ibid). They cannot be used to make generalizations about the type of discourse (e.g. academic discourse, speech, fiction writing, news) represented by the corpus under study. This can be achieved by methods of **statistical inference** that "interpret the source corpus – and hence the co-occurrence data - as a random sample from the language or sub-language of interest" (ibid). These methods allow researchers to distinguish between observed co-occurrences that are merely due to **chance** (i.e. "the particular choice of the corpus" according to Evert) and those that are due to a **significant association** between the words that holds for the language represented by the corpus under study. The following section introduces several association measures that are often used for that purpose.

Figure 4.12: Co-occurrents of the lemma *evidence* (decreasing frequency)



Collocation parameters:

Information: collocations | Statistics: Rank by frequency

Window span: -3 - 3 | Basis: whole BNC

F(n,c) at least: 5 | F(c) at least: 5

Filter results by: Specific collocate: | and/or tag: no restrictions | Submit changed parameters

There are 14358 different types in your collocation database for "[lemma = "(evidence)_.*" %c]". (Your query "{evidence}" returned 21411 matches in 2336 different texts)

No.	Word	Total No. in whole BNC	As collocate	In No. of texts
1	the	6054237	9113	1730
2	of	3049275	7419	1695
3	that	1120748	4730	1420
4	,	5026136	4071	1290
5	is	991885	3911	1168
6	to	2599307	3490	1173
7	there	319819	2828	1071
8	in	1946866	2004	905
9	and	2621932	1993	869
10	.	4722094	1777	834
11	for	880715	1709	655
12	no	227536	1689	812
13	was	883633	1499	717

4.2.3.3. Association measures

Association measures are the most widely used method for distinguishing between casual and significant co-occurrences (cf. 2.4.3). Such measures compute an association score for each pair of words extracted from a corpus which is intended as an indicator of the **strength of their association correcting for random effects**. Evert (2004:76-77) distinguishes between four broad categories of association measures:

1. The **significance of association** group comprises association measures derived from statistical hypothesis tests which aim to quantify the amount of evidence that the observed sample provides against the null hypothesis. This category includes exact statistical hypothesis tests such as the **Fisher's exact test** and asymptotic statistical hypothesis tests such as the **z-score**, the **t-score**, the **chi-square** and the **log-likelihood**.
2. The **degree of association** group encompasses measures that are maximum-likelihood estimates for various coefficients of association strength (cf. Evert 2004:84-88), e.g. the **mu-value μ** and the **odds ratio θ**.

3. The **information theory** group consists of association measures borrowed from the field of information theory. The most widely used measure is **mutual information** (MI).
4. The **heuristic formulae** group contains association measures for which no theoretical basis can be given but which are nevertheless considered to be good indicators of association. Their formulae are often heuristic variants of measures from other groups, e.g. the **MI^3** and the **log-log** measure.

A review of all these measures is clearly beyond the scope of this thesis. The reader is referred to Stefan Evert's webpage for a comprehensive list of association measures and their mathematical interpretation: http://www.collocations.de/AM/index.html.

This section presents the association measures implemented in the collocation option of the BNC*web* (cf. 4.2.2.3.2), i.e. z-score, log-likelihood, mutual information, MI^3 and log-log[117], after introducing preliminary concepts necessary for a discussion of these measures and their major properties.

4.2.3.3.1. Preliminaries

Most association measures compare the observed frequencies of words in a corpus against their expected frequencies under the **null hypothesis**, i.e. the hypothesis according to which there is no difference between the two types of frequencies. The null hypothesis is assumed to be valid until statistical evidence indicates otherwise. **Observed frequencies** are co-occurrence frequency data for a word pair (a,b) as observed in the corpus under study and are often classified into a four-cell **contingency table**. Words *a* and *b* are two variables which determine the row and column categories as illustrated in Table 4.8. Cell O_{11} represents the **joint frequency** of the word pair (a,b); cell O_{12} stands for all word pairs containing *a* but not *b*; cell O_{21} represents all word pairs containing *b* but not *a* and cell O_{22} stands for all word pairs that do not comprise *a* nor *b*. The figures in the right-hand column (i.e. R1 and R2) and the bottom row (i.e. C1 and C2) are called **marginal totals**. Thus, R1 is the marginal frequency of *a*, i.e. the number of pair tokens whose first component is *a*. N is the **grand total** or the **sample size**, i.e. the addition of $O_{11} + O_{12} + O_{21} + O_{22}$.

[117] See http://homepage.mac.com/bncweb/manual/bncwebman-collocation.htm#formulae#formulae for more information on the formulae used to compute the association measures described in this section.

Table 4.8: A contingency table

	word 2 = b	word 2 ≠ b	row totals
word 1 = a	O_{11}	O_{12}	$= R_1$
word 1 ≠ a	O_{21}	O_{22}	$= R_2$
column totals	$= C_1$	$= C_2$	$= N$

The contingency table in Table 4.9 shows the observed frequencies of the adjective-noun pair (*conclusive, evidence*) in the BNC. The pair type has a joint frequency of 93. There are 442 pair tokens of the form (*conclusive,* *) (i.e. marginal frequency R1) and 21,411 pair tokens of the form (*, *evidence*) (marginal frequency C1). The full sample N totals 4,143,217 adjective-noun pair tokens extracted from the BNC.

Table 4.9: Contingency table for the adjacent word pair (*conclusive, evidence*)

	word 2 = evidence	word 2 ≠ evidence	
word 1 = conclusive	93	349	$R_1 = 442$
word 1 ≠ conclusive	21318	4,121,457	$R_2 = 4,142,775$
	$C_1 = 21,411$	$C_2 = 4,121,806$	N= 4,143,217

To calculate **expected frequencies** for the co-occurrent words, there must be a language model which predicts how those words would behave if there were no particular collocational attraction between them. The model most commonly used in corpus linguistics is **random distribution**. As Bartsch (2004:100) has pointed out, "the assumption of a random distribution, i.e. of a completely independent distribution of words in a language sample, is a mere methodological convenience, a myth that does not reflect faithfully the reality of linguistic structure." Words, due to their part of speech membership, can only occupy certain slots in a sentence which are constrained by the grammar of the language[118].

Random distribution is theoretically sounder when grammatical constraints are taken into account (cf. Barnbrook 1996:93). Evert and Krenn (2003) and Evert (2004) propose to use a **relational** model of co-occurrence, where the co-occurrent words appear in a specific (syntactic) structure, e.g. adjective + noun, verb + object noun, to address this methodological problem. Thus, by analysing adjective + noun co-occurrences, it is recognized that nouns often attract pre-modifying adjectives. The next step is to distinguish between adjectives that co-occur with the node word by chance and those that are attracted by the noun under study.

[118] See also Rietveld et al (2004:351) for a discussion of the complex issue of "sequential dependences", by which an observation can be predicted by the outcomes of preceding observations.

In other words, a distinction must be made between pair types that support the **null hypothesis of independence** (H_0) and those that provide clear evidence against it.

As stated above, most association measures compare the observed frequencies of words in a corpus against their expected frequencies under the null hypothesis. **Expected frequencies** are the frequencies that would be predicted in each cell of a contingency table if only the observed row and column totals were known and if the variables under comparison, i.e. two co-occurrent words *a* and *b*, were independent. They can easily be computed from the marginal totals and the sample size N as shown in Table 4.10. Table 4.11 shows the expected frequencies for the contingency table of the word pair (*conclusive, evidence*) given in Table 4.9. It shows that the observed frequency of *conclusive* + *evidence* is much more frequent than expected (observed frequency = 93; expected frequency = 2).

Table 4.10: Calculating expected frequencies

	word 2 = b	word 2 ≠ b	
word 1 = a	$E_{11}=(R_1C_1)/N$	$E_{12}=(R_1C_2)/N$	$= R_1$
word 1 ≠ a	$E_{21}=(R_2C_1)/N$	$E_{22}=(R_2C_2)/N$	$= R_2$
	$= C_1$	$= C_2$	= N

Table 4.11: Expected frequencies for the adjacent word pair (*conclusive, evidence*)

	word 2 = evidence	word 2 ≠ evidence	
word 1 = conclusive	(442*21,411)/4,143,217 = 2	(442*4,121,806)/4,143,217 = 440	R1= 442
word 1 ≠ conclusive	(4,142,775*21,411)/4,143,217 = 21,408	(4,142,775*4,121,806)/4,143,217 = 4,121,366	R2 = 4,142,775
	C_1 = 21,411	C_2 = 4,121,806	N= 4,143,217

4.2.3.3.2. z-score

The z-score is widely used and implemented in corpus tools such as SARA, Xaira, TACT, etc. (cf. McEnery et al. 2006:215). It is a measure which adjusts for the total frequencies of the co-occurrent words and indicates how far the observed frequencies deviate from what would be expected under the null hypothesis. In other words, it shows how much more frequent the co-occurrence is than one would expect from the respective frequencies of each co-occurrent word. A higher z-score indicates a greater degree of association between two words. The z-score assumes that data is normally distributed and has been criticized for artificially inflating the significance of infrequent words (see Dunning 1993). As can be seen from Figure 4.13, infrequent words such as *corroborative* (overall frequency of 28 in the whole BNC), *uncorroborated* (overall frequency of 18) and *substantiating* (overall frequency of 8) are given in the top list. By contrast, highly frequent words such as *clear, good* and *strong* appear much later in the list (*clear*: position 37; *strong*: position 42; *good*: position 173) although they occur much more frequently with the noun *evidence* than the top-ranked adjectives (*clear evidence* = 169; *strong evidence* = 114 and *good evidence* = 102 vs. *corroborative evidence* = 19 and *uncorroborated evidence* = 9).

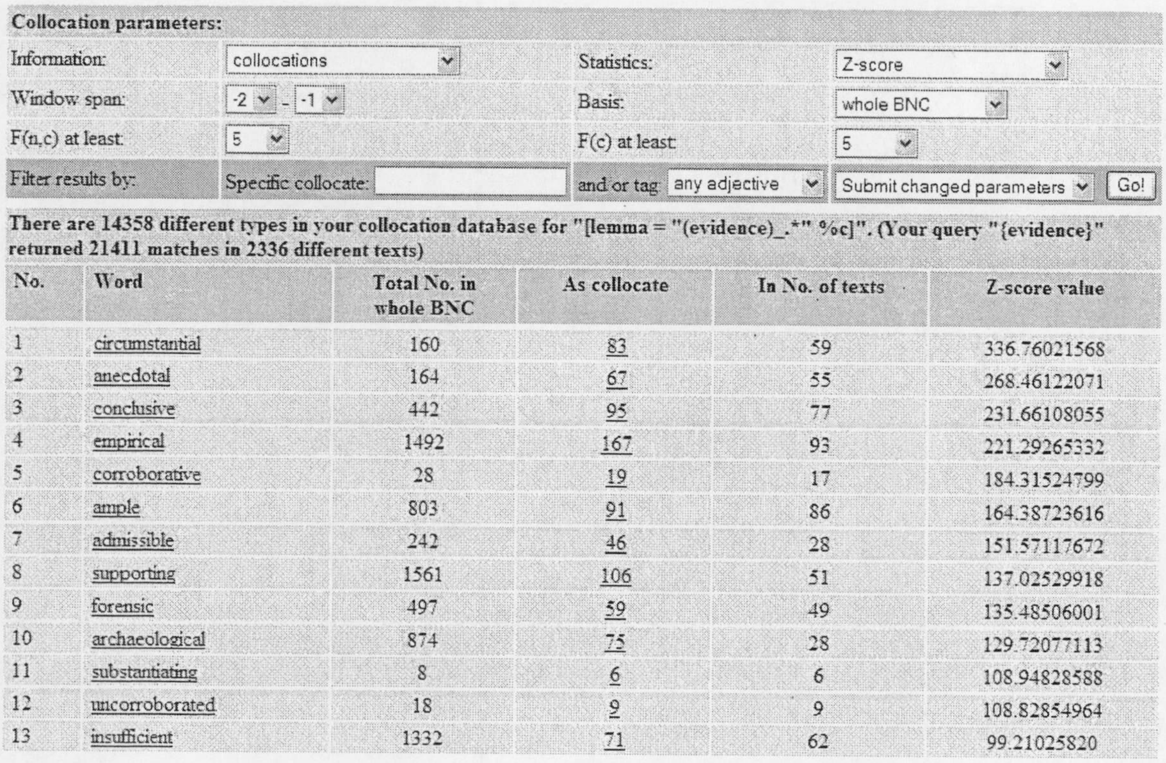


Figure 4.13: BNC – Co-occurrences of the lemma 'evidence' - z-score

Collocation parameters:

Information: collocations

Statistics: Z-score

Window span: -2 - -1

Basis: whole BNC

F(n,c) at least: 5

F(c) at least: 5

Filter results by: Specific collocate:

and or tag: any adjective

There are 14358 different types in your collocation database for "[lemma = "(evidence)_.*" %c]". (Your query "{evidence}" returned 21411 matches in 2336 different texts)

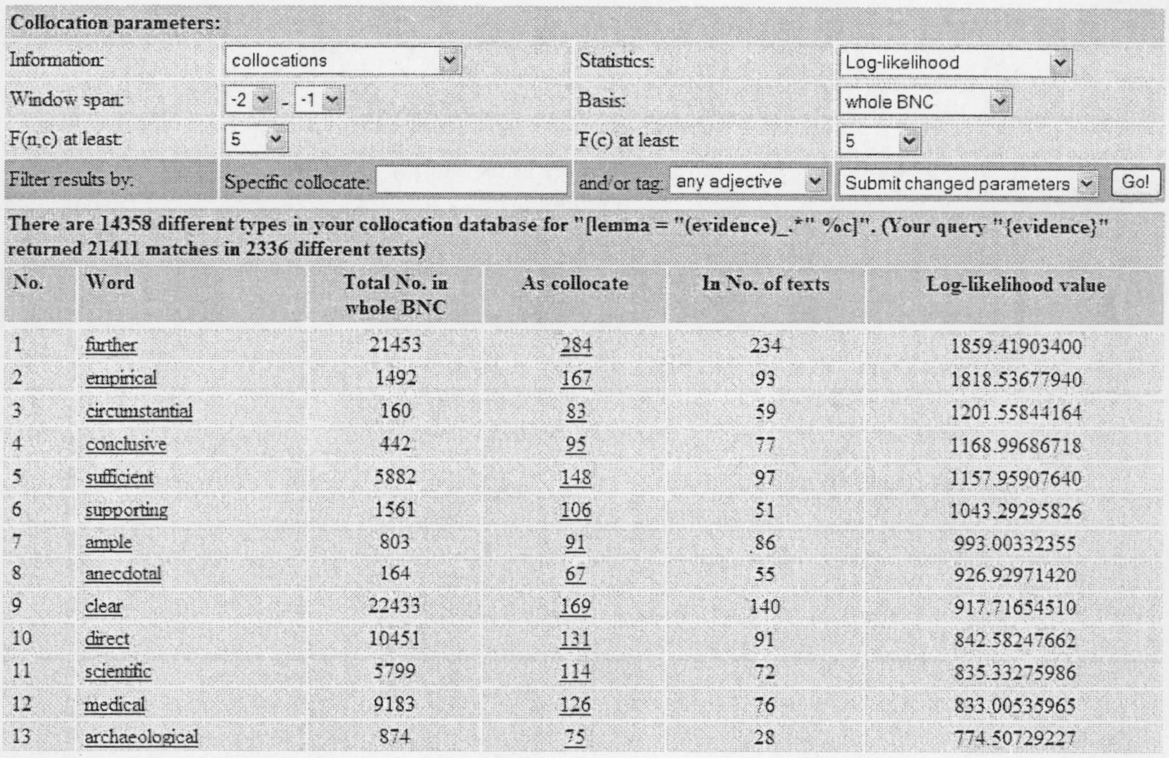
No.	Word	Total No. in whole BNC	As collocate	In No. of texts	Z-score value
1	circumstantial	160	83	59	336.76021568
2	anecdotal	164	67	55	268.46122071
3	conclusive	442	95	77	231.66108055
4	empirical	1492	167	93	221.29265332
5	corroborative	28	19	17	184.31524799
6	ample	803	91	86	164.38723616
7	admissible	242	46	28	151.57117672
8	supporting	1561	106	51	137.02529918
9	forensic	497	59	49	135.48506001
10	archaeological	874	75	28	129.72077113
11	substantiating	8	6	6	108.94828588
12	uncorroborated	18	9	9	108.82854964
13	insufficient	1332	71	62	99.21025820

The reader is referred to Berry-Rogghe (1973) and Barnbrook (1996:95-97) for more information on the z-score.

4.2.3.3.3. Log-likelihood

Dunning (1993) criticizes the z-score for assuming that data is normally distributed and proposes the log-likelihood (LogL) score, which, he argues, does not "depend so critically on assumptions of normality" and "works reasonably well with both large and small text samples and allows direct comparison of the significance of rare and common phenomena." Figure 4.14 shows that top-ranked collocates based on log-likelihood scores include both highly frequent (e.g. *further, available, clear*, with overall frequencies of 21,453 and more in the BNC) and less frequent words (e.g. *circumstantial, anecdotal*, with overall frequencies of 160 and 164 respectively). However, none of the less frequent words are as rare as adjectives extracted by the z-score such as *corroborative, uncorroborated* and *substantiating*.

Figure 4.14: BNC – Co-occurrences of the lemma 'evidence': Log-likelihood



Collocation parameters:

Information: collocations
Statistics: Log-likelihood
Window span: -2 - -1
Basis: whole BNC
F(n,c) at least: 5
F(c) at least: 5
Filter results by: Specific collocate: and/or tag: any adjective

There are 14358 different types in your collocation database for "[lemma = "(evidence)_.*" %c]". (Your query "{evidence}" returned 21411 matches in 2336 different texts)

No.	Word	Total No. in whole BNC	As collocate	In No. of texts	Log-likelihood value
1	further	21453	284	234	1859.41903400
2	empirical	1492	167	93	1818.53677940
3	circumstantial	160	83	59	1201.55844164
4	conclusive	442	95	77	1168.99686718
5	sufficient	5882	148	97	1157.95907640
6	supporting	1561	106	51	1043.29295826
7	ample	803	91	86	993.00332355
8	anecdotal	164	67	55	926.92971420
9	clear	22433	169	140	917.71654510
10	direct	10451	131	91	842.58247662
11	scientific	5799	114	72	835.33275986
12	medical	9183	126	76	833.00535965
13	archaeological	874	75	28	774.50729227

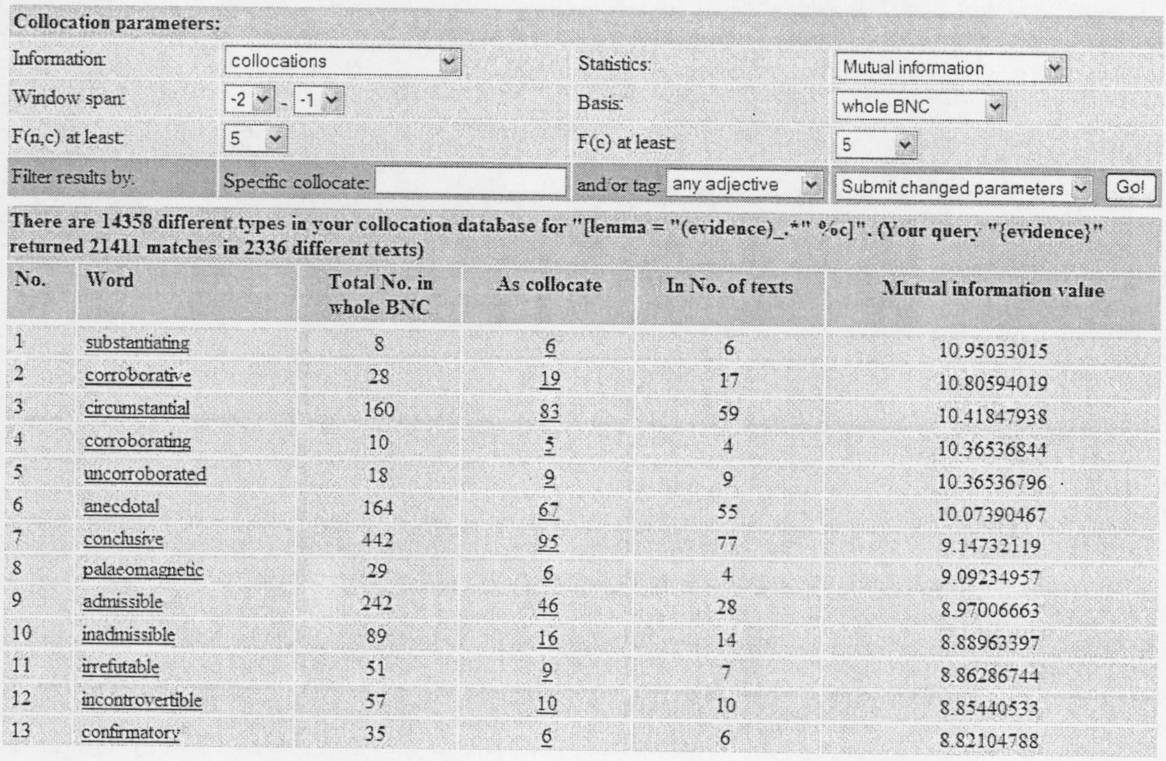
The log-likelihood scores can be directly compared with critical values of a chi-square distribution table (cf. Oakes 1998: 176). Rayson et al (2004), however, advise that a critical value of significance of 15.13 be used at p < 0.01 instead of 6.64.

4.2.3.3.4. Mutual information and MI^3

Mutual information is a statistical measure borrowed from information theory. It compares the probability of observing word *a* and word *b* together with the probabilities of observing *a* and *b* independently (cf. Church and Hanks 1990, Church et al. 1991). The MI score is defined by McEnery et al. (2006:56) as "a measure of collocational strength." Manning and Schütze (2000: 182), however, argue that MI is a good measure of independence but a bad indicator of dependence. Values close to 0 indicate independence between the two elements of a word pair. For dependence, the score is dependent upon the frequencies of word *a* and word *b*. Thus, word pairs composed of low-frequency or rare words will receive a higher score than word pairs composed of frequent words.

One solution that has been proposed to address the problem of sparseness is the use of a frequency threshold of at least 3 under which words are not taken into account in a collocational analysis (cf. Manning and Schütze 2000:182). Clear (1993:280) discards all word pairs having a frequency of co-occurrence of less than three in his collocational analysis of the word *taste* in a 30 million word corpus. Figure 4.15 nevertheless shows that a frequency threshold does not improve results significantly and that the MI score gives **too much weight to rare events**. Most of the top-ranked adjectives that co-occur with the noun *evidence* are infrequent words which occur less than 100 times in the BNC, e.g. *substantiating* (with an overall frequency of 8 in the BNC), *corroborating* (overall frequency of 10), and *uncorroborated* (overall frequency of 18).

Figure 4.15: BNC – Co-occurrences of the lemma 'evidence': Mutual Information



Collocation parameters:

Information: collocations
Statistics: Mutual information
Window span: -2 - -1
Basis: whole BNC
F(n,c) at least: 5
F(c) at least: 5
Filter results by: Specific collocate: and/or tag: any adjective

There are 14358 different types in your collocation database for "[lemma = "(evidence)_.*" %c]". (Your query "{evidence}" returned 21411 matches in 2336 different texts)

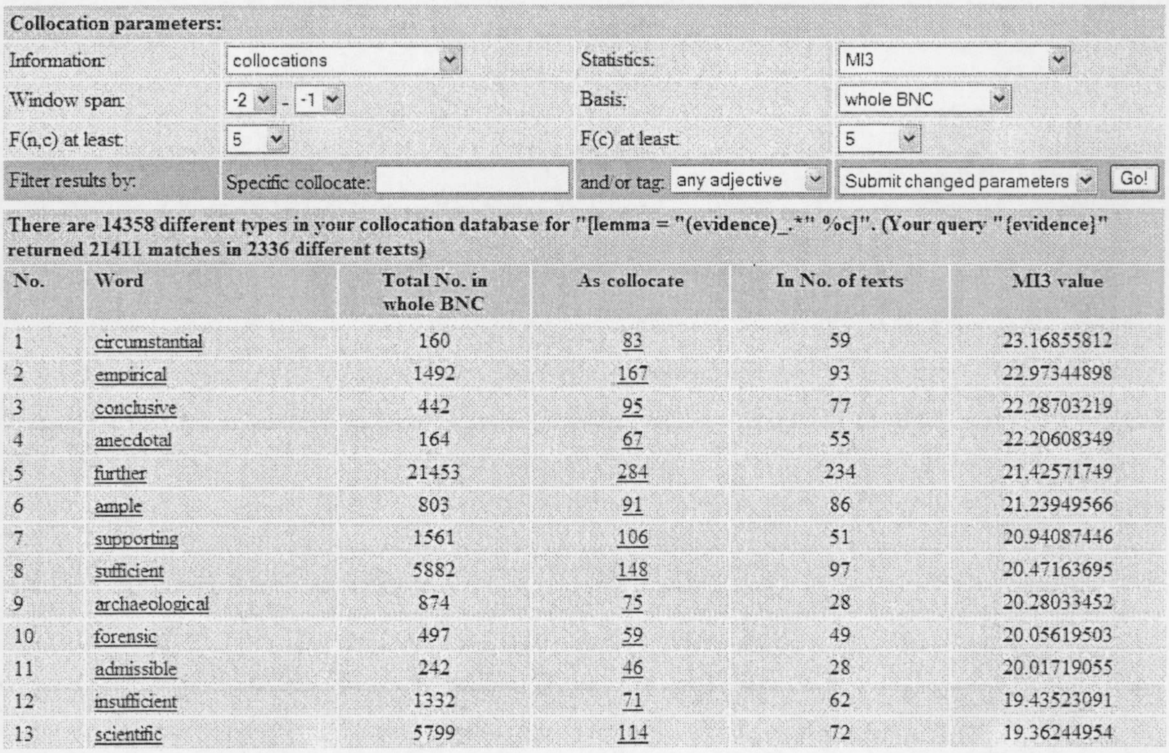
No.	Word	Total No. in whole BNC	As collocate	In No. of texts	Mutual information value
1	substantiating	8	6	6	10.95033015
2	corroborative	28	19	17	10.80594019
3	circumstantial	160	83	59	10.41847938
4	corroborating	10	5	4	10.36536844
5	uncorroborated	18	9	9	10.36536796
6	anecdotal	164	67	55	10.07390467
7	conclusive	442	95	77	9.14732119
8	palaeomagnetic	29	6	4	9.09234957
9	admissible	242	46	28	8.97006663
10	inadmissible	89	16	14	8.88963397
11	irrefutable	51	9	7	8.86286744
12	incontrovertible	57	10	10	8.85440533
13	confirmatory	35	6	6	8.82104788

Daille (1994) considered different ways of rebalancing the MI score to give less weight to rare words and more to frequent words by increasing the influence of the joint frequency (O_{11}) in the statistical formula. Daille tested versions of MI in which O_{11} was successively replaced by all powers of O_{11} from 2 to 10. The cube of the joint frequency was found to obtain the best results, yielding the following formula:

$$MI^3 = \log (O_{11})^3 / E_{11}$$

The MI^3 is thus a purely heuristic variant of MI and is not based on a sound theoretical model (see Daille 1994, Oakes 1998 or Evert 2004 for more information). As shown in Figure 4.16, it gives more weight to frequent words. Unlike with MI, top-ranked adjectives are rather frequent ones and occur more than 100 times in the BNC. Examples are *circumstantial, empirical, conclusive, anecdotal* and *admissible*. These examples show that MI^3 gives results that could be described as standing between the log-likelihood and the z-score (compare with Figures 4.13 and 4.14).

Figure 4.16: BNC – Co-occurrences of the lemma 'evidence': MI^3



Collocation parameters:

Information:	collocations	Statistics:	MI3
Window span:	-2 - -1	Basis:	whole BNC
F(n,c) at least:	5	F(c) at least:	5
Filter results by:	Specific collocate:	and/or tag: any adjective	Submit changed parameters

There are 14358 different types in your collocation database for "[lemma = "(evidence)_.*" %c]". (Your query "{evidence}" returned 21411 matches in 2336 different texts)

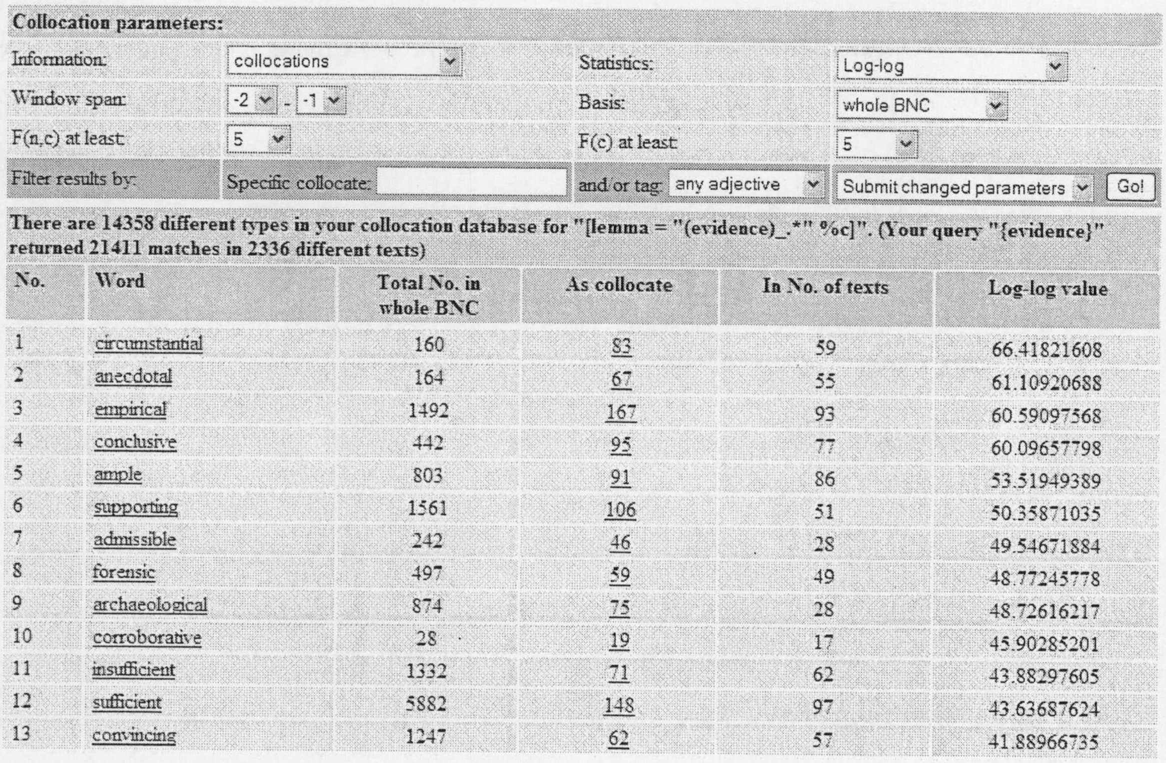
No.	Word	Total No. in whole BNC	As collocate	In No. of texts	MI3 value
1	circumstantial	160	83	59	23.16855812
2	empirical	1492	167	93	22.97344898
3	conclusive	442	95	77	22.28703219
4	anecdotal	164	67	55	22.20608349
5	further	21453	284	234	21.42571749
6	ample	803	91	86	21.23949566
7	supporting	1561	106	51	20.94087446
8	sufficient	5882	148	97	20.47163695
9	archaeological	874	75	28	20.28033452
10	forensic	497	59	49	20.05619503
11	admissible	242	46	28	20.01719055
12	insufficient	1332	71	62	19.43523091
13	scientific	5799	114	72	19.36244954

4.2.3.3.5. Log-log

According to the BNC manual, the log-log measure refers to the heuristic variant of MI used in the *Sketch Engine* developed by Adam Kilgarriff and his colleagues (see 4.2.3.1.1). It is the product of mutual information and the logarithm of the raw frequency of the co-occurrent word *b*. The log-log score gives less weight to low-frequency or rare word co-occurrents. Figure 4.17 shows that results differ significantly from top-ranked adjectives according to MI as illustrated in Figure 4.15. They also differ from results ranked by log-likelihood or MI^3 scores as top-ranked adjectives are much less frequent in the corpus (compare with Figures

4.14 and 4.16). High-frequency co-occurrent words are not top-ranked by the log-log measure.

Figure 4.17: BNC – Co-occurrences of the lemma ‘evidence’: Log-log



Collocation parameters:

Information: collocations

Statistics: Log-log

Window span: -2 - -1

Basis: whole BNC

F(n,c) at least: 5

F(c) at least: 5

Filter results by: Specific collocate: and/or tag: any adjective

There are 14358 different types in your collocation database for "[lemma = "(evidence)_.*" %c]". (Your query "{evidence}" returned 21411 matches in 2336 different texts)

No.	Word	Total No. in whole BNC	As collocate	In No. of texts	Log-log value
1	circumstantial	160	83	59	66.41821608
2	anecdotal	164	67	55	61.10920688
3	empirical	1492	167	93	60.59097568
4	conclusive	442	95	77	60.09657798
5	ample	803	91	86	53.51949389
6	supporting	1561	106	51	50.35871035
7	admissible	242	46	28	49.54671884
8	forensic	497	59	49	48.77245778
9	archaeological	874	75	28	48.72616217
10	corroborative	28	19	17	45.90285201
11	insufficient	1332	71	62	43.88297605
12	sufficient	5882	148	97	43.63687624
13	convincing	1247	62	57	41.88966735

4.2.3.4. Discussion and association measures adopted in this thesis

> “Frequency becomes interesting when it is interpreted as typicality” (Stubbs 2002:61)

Each of the association measures described in this section can be used to order lists of word pairs (*a,b*) extracted from corpus data and highlight those which appear to be most strongly collocationally attracted. Section 4.2.3.3 has offered evidence that each measure provides very different types of information. Table 4.12 shows the top twenty collocates of the node word *evidence* in the BNC for each association measure. The lists are sorted by alphabetical order so that a comparison is easier to make. Information is also given on the ranking of each word according to each measure. As explained by Barnbrook (1996:100), this type of display highlights two major types of difference between association measures. First, words which are included in one list but not in another: *ample, anecdotal, conclusive* and *forensic* belong to the top twenty co-occurrent adjectives of *evidence* as ranked by the five measures under study. By contrast, the adjectives *confirmatory, damning* and *serological* are highly ranked by MI only while *historical* and *strong* are only found in the top twenty list of the log-likelihood. Second, words whose ranks differ between the lists: while *ample* is top-ranked by the five measures, its position varies from 5 (LogL) to 17 (MI).

Table 4.13 presents a list of all adjectives found in the top twenty lists as described in Table 4.12, together with information on their overall frequency in the BNC and their joint frequency with the noun *evidence*. The right columns represent the association measure: when a word was found in the top twenty list of an association measure, a √ is printed in the corresponding cell. The table is sorted by increasing overall frequency of the adjectives. A number of observations can be made from this table. First, MI clearly stands apart from the other association measures. It systematically ranks best all the adjectives that have the **lowest overall frequencies** in the BNC. Second, MI^3 and log-log behave similarly: although ranks may differ slightly between the two measures (e.g. *further* – MI^3: rank 5; log-log: rank 14), they share 17 adjectives. The adjectives *clear, direct* and *medical* are extracted by MI^3 while *criminal, abundant* and *expert* are found in the log-log top-twenty list. Third, the log-likelihood is the only measure which does not have any adjective with an overall frequency of 150 or lower in its top twenty list. It ranks best adjectives with **high overall frequencies**. Finally, the z-score seems to rank best two types of adjectives, i.e. adjectives which are quite frequent in the BNC but, unlike the log-likelihood, not necessarily the most frequent ones (e.g. *anecdotal, empirical* and *supporting*) and infrequent adjectives but whose percentage of occurrences together with *evidence* is quite high. Thus, *substantiating* occurs 8 times in the BNC and 6 times together with *evidence* (75%). Similarly, *uncorroborated* occurs 18 times, among which 9 times as a co-occurrent of *evidence* (50%) and *corroborative* is found in 64% of its occurrences next to *evidence*. In section 4.2.3.3.2, it was said that the z-score is believed to artificially inflate the significance of infrequent words (cf. Dunning 1993). Our findings suggest that, unlike MI, the z-score does not inflate the significance of any type of infrequent word but ranks best **infrequent words whose percentage of occurrences together with the node word is very high**.

Table 4.12: A comparison of association measures

z-score		LogL		MI		MI^3		log-log	
word	**rank**	**word**	**rank**	**word**	**rank**	**word**	**rank**	**word**	**rank**
								abundant	19
admissible	7			admissible	9	admissible	11	admissible	7
ample	6	ample	7	ample	17	ample	6	ample	5
anecdotal	2	anecdotal	8	anecdotal	6	anecdotal	4	anecdotal	2
archaeological	10	archaeological	13			archaeological	9	archaeological	9
circumstantial	1	circumstantial	3	circumstantial	3	circumstantial	1	circumstantial	1
		clear	9			clear	17		
conclusive	3	conclusive	4	conclusive	7	conclusive	3	conclusive	4
				confirmatory	13				
convincing	16					convincing	19	convincing	13
corroborating	19			corroborating	4				
corroborative	5			corroborative	2	corroborative	14	corroborative	10
		criminal	17					criminal	20
				damning	15				
		direct	10			direct	18		
empirical	4	empirical	2	empirical	18	empirical	2	empirical	3
				epidemiological	19				
experimental	18	experimental	14			experimental	15	experimental	15
expert	20							expert	17
forensic	9	forensic	16	forensic	16	forensic	10	forensic	8
further	15	further	1			further	5	further	14
		historical	20						
inadmissible	17			inadmissible	10				
				incontrovertible	12				
				incriminating	14				
insufficient	13	insufficient	15			insufficient	12	insufficient	11
				irrefutable	11				
		medical	12			medical	16		
		oral	18			oral	20	oral	16
				palaeomagnetic	8				
		scientific	11			scientific	13	scientific	18
				serological	20				
		strong	19						
substantiating	11			substantiating	1				
sufficient	14	sufficient	5			sufficient	8	sufficient	12
supporting	8	supporting	6			supporting	7	supporting	6
uncorroborated	12			uncorroborated	5				

Table 4.13: Association measures and overall frequency

word	f(word)	f(word, *evidence*)	z-score	LogL	MI	MI^3	log-log
substantiating	8	6	√		√	√	√
corroborating	10	5	√		√		
uncorroborated	18	9	√		√		
corroborative	28	19	√		√	√	√
palaeomagnetic	29	6			√		
confirmatory	35	6			√		
irrefutable	51	9			√		
serological	54	5			√		
incontrovertible	57	10			√		
inadmissible	89	16			√		
incriminating	100	13			√		
damning	118	15			√		
circumstantial	160	83	√	√	√	√	√
anecdotal	164	67	√	√	√	√	√
epidemiological	179	20			√		
admissible	242	46			√		
conclusive	442	95	√	√	√	√	√
forensic	497	59	√	√	√	√	√
abundant	598	37					√
ample	803	91	√	√	√	√	√
archaeological	874	75	√	√		√	√
convincing	1247	62	√			√	√
insufficient	1332	71	√	√		√	√
empirical	1492	167	√	√	√	√	√
supporting	1561	106	√	√		√	√
expert	1599	63	√				√
oral	2272	75		√		√	√
experimental	2336	81	√	√		√	√
criminal	4645	88		√			√
historical	5492	86		√			
scientific	5799	114		√		√	√
sufficient	5882	148	√	√		√	√
medical	9183	126		√		√	
direct	10451	131		√		√	
strong	15703	114		√			
further	21453	284	√	√		√	√
clear	22433	169		√		√	

The choice of an association measure clearly depends on the **objectives** of a co-occurrence analysis. As McEnery et al (2006:217) have suggested, word pairs that are significant when MI is used are generally interesting for lexicographical purposes while they are of secondary importance for pedagogical purposes. By contrast, they argue that word pairs highlighted by MI^3 are probably "more useful for second language learners at beginning and intermediate levels." The objectives of the co-occurrence analysis conducted in this thesis are primarily **descriptive** and **applied**: it is intended to describe the phraseology of EAP vocabulary in academic texts written by English speakers and EFL learners. In chapter 1, we defined **phraseology** as the study of all syntagmatic relations between at least two word-forms or lexemes, contiguous or not, written separately or not, which are typically syntactically closely related constituents and constitute "'preferred' ways of saying things" for the language user since:

- They form a functional (referential, interactional, or structural) unit; and
- They display arbitrary lexical restrictions; and/ or
- They are characterized by a certain degree of semantic non-compositionality; and/or
- They display a certain degree of syntactic fixity.

The inclusion in our definition of the idea of "'preferred' ways of saying things" clearly favours measures such as the log-likelihood, MI^3 or the log-log which tend to give more prominence to frequent co-occurrences. By contrast, phrasemes that are characterized by a certain degree of semantic non-compositionality may be better ranked by MI as they are also often less frequent. For example, Moon (1998a) has shown that idioms such as *spill the beans* and *call the shorts* have frequencies of less than 1 per million words.

Barnbrook (1996:101) argues that it is "difficult, if not impossible, to select one measure which provides the best assessment of the collocates" and that it is "probably better to use as much information as possible in exploring collocation, and to take advantage of the different perspectives provided by the use of more than one measure." Similarly, Bartsch (2004) uses three association measures to ensure identification of relevant co-occurrence data. She uses the MI score as the prime statistic for filtering what she calls 'collocation candidates' from the BNC word pairs and the t-test and chi-square scores for cross-checking purposes as "these can support and sometimes supplement the data identified by MI" (ibid 112).

A similar strategy is adopted in this thesis when analysing BNC data but the **log-likelihood** is used as the prime statistic for filtering co-occurrence data. Following Rayson et al. (2004), we examine all co-occurrences that have a log likelihood higher than 15.13 (with $p \leq 0.01$). This means that there is a probability of 99.9% that the co-occurrence of two words is not due to chance. In other words, the probability of making an error by saying that the co-occurrence is not due to chance is only of 0.1%. There is no added value in using the log-likelihood together with MI^3 or log-log as these measures share a large proportion of best-ranked candidates.[119] As illustrated in Table 4.12, MI and log-likelihood rank co-occurrence data in widely different ways and can thus be regarded as complementary measures for descriptive purposes. The **MI** will thus be used for cross-checking purposes so that no infrequent phraseme remains unnoticed but results will not be systematically reported.

Association measures are applied on co-occurrences extracted from the BNC only. Log-likelihood measures are largely dependent on corpus size and word frequencies. Co-occurrence statistics are therefore not comparable across corpora of different sizes such as BNC-AC-HUM, the ICLE sub-corpora and STUD-US-ARG. Moreover, the ICLE sub-corpora and STUD-US-ARG are arguably too small for a statistical analysis of co-occurrences. The types of words analysed in this thesis are not high-frequency words such as *make, do* and *take* and co-occurrences in learner and native student writing often appear less than three times, a threshold under which association measures should be excluded from a statistical analysis. As a result, the following method is adopted: **word pairs in the ICLE sub-corpora and in STUD-US-ARG are classified into three groups according to their co-occurrence status in professional academic writing:**

- **Word pairs that do not appear in BNC-AC-HUM;**
- **Word pairs that appear in BNC-AC-HUM but are not statistically significant co-occurrences;**
- **Word pairs that are statistically significant co-occurrences in BNC-AC-HUM.**

4.3. Conclusion

> "Though some believe that the statistical methods have rendered linguistic analysis unnecessary, this is in fact not the case." (Sag et al. 2002)

This chapter described the data and methodology that are used in this thesis to investigate the phraseology of EAP vocabulary in native and EFL learner writing. Special care has been taken to select a set of learner essays from the *International Corpus of Learner English* that is as homogeneous as possible and to control a number of variables that have been found to influence EFL learner writing. Three types of native corpora are used as control corpora in the study reported in this thesis. The learner corpus is first compared to the academic sub-corpus of the *British National Corpus* (domain: humanities and arts) to identify **learner-specific**

[119] The log-likelihood is preferred over MI^3 and log-log as these measures are not based on a theoretically sound model of language (see section 4.2.3.3.4 and 4.2.3.3.5).

features in the way they use EAP vocabulary. It is then compared to STUD-US-ARG (i.e. a subset of the *Louvain Corpus of Native Speaker Essays* and similar essays) to identify features of **novice writing**. The spoken part of the *British National Corpus* is also sometimes used to check whether specific words and phrases that appear in learner corpora are **more typical of speech or writing**. The method used to investigate the phraseology of EAP vocabulary in native and learner corpora is based on the **Integrated Contrastive Model** and combines comparisons of learner vs. native writing corpora and comparisons of learner writing corpora.

The method for studying co-occurrences in the BNC combines automatic, statistical and manual procedures. First, repeated co-occurrences are automatically extracted on the basis of a number of parameters that have been carefully described in this chapter. Second, association measures are used to highlight co-occurrences that occur more often than predicted by chance and provide clear evidence against the null hypothesis of independence. The **log-likelihood** and the **MI** have been selected as they were found to possess complementary properties for descriptive purposes. Finally, co-occurrences are manually analysed and those that do not stand in a grammatical relation to the node are eliminated. Co-occurrences in ICLE and STUD-US-ARG are not statistically analysed but are classified **according to their status in BNC-AC-HUM.**

The data and methodology described here are used in chapters 6 and 7. Chapter 6 is devoted to a **comparison between native language and learner language**. The learner corpus is used as a whole when compared to native writing. However, **a comparison of interlanguages** is also made so as to highlight interlanguage features that are **shared across a majority of learner populations**. In chapter 7, the focus will be on the French learner sub-corpus which will be compared to the 9 other learner populations in order to investigate transfer effects on a number of overused EAP phrasemes in French learner writing. Before turning to the actual investigation of corpus data, chapter 5 discusses the issue of word selection for the purposes of this thesis and presents a methodology to extract EAP-specific words from corpora.

5. Selection of EAP vocabulary

> A great deal of research is still necessary to describe with any empirical rigour the lexis that is characteristic of particular purposes, genres and registers. (Milton 1999:223)

5.1. Introduction

Chapter 5 is dedicated to the selection of words that are typical of academic texts and which will provide a basis for the comparison of native speaker and EFL learner academic writing. In section 1.3.1.2, we examined vocabulary needs in EAP and described in detail the *Academic Word List* (Coxhead 2000). Although widely used today for receptive as well as productive purposes, the *AWL* was initially developed to complement the *General Service List of English Words* so as to approach the critical 95% coverage threshold needed for reading comprehension. This primary objective carries direct methodological implications that were shown to be inappropriate for the purposes of this thesis (cf. section 1.3.2).

In this chapter, we use a **corpus-driven** approach which "relies heavily on data and (largely) automatic procedures" (De Cock 2003:197) (cf. also Tognini-Bonelli 2001) and propose a new methodology based on the criteria of keyness, range and evenness of distribution to select EAP-specific words that would provide the basis for a **productively-oriented academic word list**. Unlike the *AWL*, our list is based on annotated corpora. Section 5.2 gives a description of the corpora used, discusses the pros and cons of corpus annotation and describes the annotation tools employed. Section 5.3 presents the methodology developed to select EAP words. Results for two corpus formats, i.e. word forms + POS-tags and lemmas + POS-tags, are compared in 5.4. Section 5.5 describes the procedure used to add to our list words that belong to well-represented semantic categories in EAP. The resulting *Academic Keyword List* is compared to the *Academic Word List* in section 5.6. Finally, section 5.7 discusses inherent limitations of the keyness approach adopted.

5.2. Data

In section 5.2.1, we describe the corpora of professional and student writing used for the extraction of EAP vocabulary. In section 5.2.2, we discuss the issue of corpus format, highlight the advantages of annotation for applied objectives and describe the annotation tools used in this study.

5.2.1. Corpora

Recent studies have drawn on corpora of academic texts to identify and examine generic features of academic writing, such as discourse structure, rhetorical strategies, citation practices, and lexical choice (see sections 1.2 and 1.4.1). The corpora used tend to consist solely of professional or expert writing, perhaps because this specific text type is in the public domain and therefore relatively easy to collect. Academic writing, however, includes other kinds of text types than professionally edited articles and books, notably student essays. As Nesi et al (2004:440) comment, "[n]ovice writers do not (...) begin by writing for publication, or for a readership of strangers. Their early attempts at academic writing are more likely to be assessed texts produced in the context of a course study". In this study, the selection of academic vocabulary is thus made on the basis of an analysis of both **professional** and **student writing**.

Although representativeness remains an 'act of faith' (Leech 1991: 127), efforts have also been made to use texts from different academic **disciplines** (e.g. arts, social science, applied science, etc.) and with different **purposes** (e.g. ESP texts vs. argumentative texts) to build up a picture of academic writing as complete as possible.

5.2.1.1. Professional writing

The professional academic corpora used are the *Micro-Concord corpus collection B* (henceforth MC) and the *Baby BNC academic corpus* (henceforth B-BNC), two 1,000,000-word corpora of published academic prose which consist of 33 and 30 texts respectively. The B-BNC consists entirely of texts written by native speakers of British English while MC also includes texts written by native speakers of American English. Table 5.1 shows that both corpora consist of five sub-corpora of about 200,000 words, each of which corresponds to a broad academic discipline. This division into academic disciplines is particularly well suited for our purposes as the study presented in this chapter seeks to extract words that are used by all members of the 'academic discourse community' (cf. Swales 1990) across disciplines.

Table 5.1: Professional academic writing

Corpus	Variety of English	Text type	Number of words
MC	mainly British English	books	**1,005,060**
Arts			180,496
Belief and religion			199,612
Science			219,596
Applied science			203,316
Social science			202,302
B-BNC	British English	books and periodicals	**1,021,007**
Humanities			262,476
Politics, education and law			196,322
Social science			132,678
Science			283,490
Technology and engineering			146,041
TOTAL			**2,026,067**

A number of studies have shown that academic writing conventions may differ markedly according to the discipline (see section 1.2.). 'Soft science' (e.g. arts, psychology, social science, etc.) and 'hard science' (e.g. applied science, chemistry, medicine, etc.) are often contrasted in the literature. Two corpora were thus compiled from the MC and the B-BNC, namely a corpus of professional 'soft science' (ProfSS) and a corpus of professional 'hard science' (ProfHS) as described in Table 5.2. This re-categorisation is intended to extract words that are typical of both soft and hard science academic texts.

Table 5.2: Recategorization of corpus data: Two corpora of professional academic writing

Corpus	Number of words	
ProfSS	**1,173,886**	
MC Arts		180,496
MC Belief and religion		199,612
MC Social science		202,302
BNC Humanities		262,476
BNC Politics, education and law		196,322
BNC Social science		132,678
ProfHS	**852,443**	
MC Science		219,596
MC Applied science		203,316
BNC Science		283,490
BNC Technology and engineering		146,041

5.2.1.2. Student writing

Two corpora of student writing were used for this study: a preliminary version of the *British Academic Written English* (BAWE) corpus, currently being developed at the Centre for English Language Teacher Education and the Warwick Writing Programme at the University of Warwick, England (Nesi et al. 2004) and the argumentative sub-part of the *Louvain Corpus of Native Speaker Essays* (LOCNESS) (cf. 4.1.2.2).

The BAWE is a corpus of assessed student essays between 1,000 and 5,000 words in length. The essays are grouped according to the disciplines, faculties and departments at Warwick University, e.g. arts, history, psychology and engineering. The preliminary version of the BAWE corpus contains 499 assignments from 18 departments. However, only a sample was used here for two main reasons. First, 27% of the contributors are not native speakers of English. Nesi et al. (2004: 444) argue that "the University of Warwick is a multicultural, multilingual environment, and in their departments students are assessed on merit, without regard for their language background", and therefore assume that "all contributors are proficient users of English, given that their assignments have been awarded high grades." As one of the main objectives of this thesis is to compare EFL learners' use of EAP words and phrasemes with that of English writers, we selected the essays written by native speakers of English only. Second, disciplines are not equally represented. The majority of essays come from the humanities while contributions from chemistry, computer science, mathematics, etc., are completely lacking. Texts were thus selected in four well-represented disciplines: arts, social science, psychology and history (cf. Table 5.3.).

Table 5.3: Student academic writing

Corpus	Variety of English	Text type	Number of words
BAWE	British English	essays	**845,344**
Arts			221,841
Social science			163,300
Psychology			201,946
History			258,257
LOCNESS	mainly American English	essays	**168,593**
TOTAL STUDCORP			**1,013,937**

Essay topics are very diverse in the BAWE corpus and seldom repeated (cf. Table 5.4. for examples). For more information on essay topics in LOCNESS, see section 4.1.2.2.

Table 5.4: Essay topics in the BAWE corpus

Arts [98 essays]	- Visual arts in Britain - Prince Arthur portrayed in books - Rise of aestheticism - Modes of writing essays
Social science [64 essays]	- Housing policy - Teachers as professionals - Would you agree that subordination was inscribed in the life of domestic servant?
Psychology [103 essays]	- Clinical depression - Psychology as a science - Expressing attitude - Is attention merely a matter of selection?
History [136 essays]	- Absolutism in early modern Europe - Why did America dominate the world film market by the 1920s? - Who was to blame for the Boxer rising?

5.2.2. Corpus format

The *Academic Word List* (Coxhead 2000) is a list of **word forms** that were manually classified into 570 word families (cf. section 1.3.1.2). Similarly, most studies of vocabulary in the field of English for Specific Purposes are based on raw corpora (e.g. James et al. 1994; Curado Fuentes 2001) and none of them discuss the issue of corpus format. However, annotated data have been shown to play an important role in lexicography (see section 4.2.2.1.1), where lemma, word class and meaning are often associated. The main objective of this chapter is to select nouns, verbs, adjectives, adverbs and other function words that are commonly used in academic texts. **Part-of-speech tagged** corpora were thus used to facilitate the extraction of specific word classes. The extraction procedure is tested on two corpus formats, i.e. **word form + morphosyntactic tag** and **lemma + morphosyntactic tag**, to assess the suitability of each format for selecting EAP vocabulary.

5.2.2.1. Annotating corpora

> Corpora are useful only if we can extract knowledge or information from them. The fact is that to extract information from a corpus, we often have to begin by building information in. (Leech 1997a:4)

Corpus annotation refers to the practice of adding linguistic information to an electronic corpus of language data. Various levels of annotation can be distinguished, starting from the addition of lemma information to each word in the corpus. A **lemma** is used to group together inflected forms of a word, such as the singular and the plural forms of a noun, or the different conjugated forms of a verb. A second type of annotation is the morphosyntactic level of annotation, which concerns the labelling of the Part-Of-Speech (POS) or grammatical category of each word in the corpus. **POS tagging** is the most popular kind of linguistic annotation applied to text: by providing information about the grammatical nature of a word, it makes it possible to extract information about the various meanings and uses of this word. Thus, it distinguishes between *left* as the past tense or past participle of *leave* ('I *left* early') and *left* as a word meaning the opposite of *right*, either as an adjective ('my *left* hand'), an adverb ('turn *left*') or a noun ('on your *left*') (cf. Leech 1997a: 4). Other levels of annotation are syntactic annotation or parsing (the analysis of sentences into their constituents), semantic annotation (the labelling of semantic fields) and discourse tagging (the annotation of discourse relations within the texts). For more information on the different levels of annotation, see Leech (1997a: 12).

Compelling reasons for annotating a corpus are numerous. Leech and Smith (1999: 31-36) examine the possible uses of annotation in detail and list applications in fields as diverse as information retrieval, word processing (e.g. spell-checkers), speech processing (e.g. distinguish homographs for speech synthesis), machine-aided translation, lexicography, etc. However, a number of criticisms have been directed at corpus annotation. One of the most widespread criticisms is that annotation reflects, at least to a certain extent, some theoretical perspective. Although the sets of categories and features used in annotating a corpus are generally chosen to be as uncontroversial classes as possible, the interpretative nature of corpus annotation has been understood as a way to impose pre-existing models of language on corpus data (cf. Tognini-Bonelli 2001: 73-74). These models of language date from a "pre-corpus" time and some of them derive from descriptions which ignore empirical evidence altogether (cf. Sinclair 2004: 52). The argument, though valid, is certainly not strong enough to ignore all the advantages of corpus annotation entirely but should be taken as a word of caution against the naive assumption that using annotating software is a neutral act. It is regrettable that "too many researchers nowadays expect, and accept, off-the-shelf tools that they do not examine too closely" (Sinclair 2004: 51).

These models of language are also limited by practical constraints such as the need for speed and accuracy in automatic tagging. As Leech (1997b:25) explains, an "armchair linguist" might conceive a new tagset based on sound linguistic principles only to discover that the software is incapable of assigning a specific tag with any degree of accuracy. A tagset should thus be considered as "a trade-off between what is linguistically most desirable and computationally feasible" (Leech 1997b:27) at a given point in time.

Another argument that is put forward against annotation is that it may introduce errors. While it is inevitable that annotation systems will sometimes get things wrong, the various levels of annotation distinguished above are performed with varying degrees of accuracy. POS tagging is a well-researched kind of linguistic annotation and taggers perform with very high levels of accuracy (see 5.2.2.2.1.) whereas discourse annotation systems, for example, are more recent and still need major improvement.

Although annotated data is often described as 'enriched' data (Leech and Smith 1999; Aarts 2002; Bowker and Pearson 2002), annotation has also sometimes been criticized for resulting in a loss of information (Sinclair 1992; 2004). The argument is summarised by Tognini-Bonelli as follows:

> It could be argued that in a tagged text no information is lost because the words of the text are still there and available, but the problem is that they are bypassed in the normal use of a tagged text. The actual loss of information takes place when, once the annotation of the corpus is completed and the tagsets are attached to the data, the linguist processes the tags rather than the raw data. By doing this the linguist will easily lose sight of the contextual features associated with a certain item and will accept single, uni-functional items – tags – as the primary data. What is lost, therefore, is the ability to analyse the inherent variability of language which is realised in the very tight interconnection between lexical and grammatical patterns. This is the price paid for simplification; a process that is so useful – but it is argued here that the interconnection between lexis and grammar is crucial in determining the meaning and function of a given unit: any processing that loses out on this is bound to lose out in accuracy. (Tognini-Bonelli 2001: 73-74)

The methodology adopted here is designed to give the best of both worlds, by first extracting EAP-specific words from annotated corpora and then returning to raw data to analyze their use in context.

Finally, it is worth underlining that this chapter does not intend to meet a theoretical but an **applied** objective and that even linguists that have directed the most severe criticisms against annotated data acknowledge that the "good point of annotation lies in its value in applications" (Tognini-Bonelli 2001: 73).

5.2.2.2. The software used

We make use of *Wmatrix*, a web-based corpus processing environment which gives researchers access to several corpus annotation and retrieval tools developed at the *University Centre for Computer Corpus Research on Language* (UCREL), Lancaster University. Tools available in *Wmatrix* include the *Constituent Likelihood Automatic Word-tagging System* (CLAWS) and the *UCREL Semantic Analysis System* (USAS) (cf. Rayson 2003).

5.2.2.2.1. The Constituent Likelihood Automatic Word-tagging System (CLAWS)

A corpus uploaded to the *Wmatrix* environment is first grammatically tagged with the *Constituent Likelihood Automatic Word-tagging System* (CLAWS) (cf. Garside and Smith 1997). The tagger makes use of a detailed tagset of 146 tags (cf. Appendix 5.1) as well as of two lexicons: (a) a lexicon of single words with all their possible parts-of-speech and associated lemmas and (b) a multi-word expression lexicon, which currently contains 18,710 patterns. Most multi-word expressions are phrasal verbs (*stub out*), noun phrases (*riding boots*), proper names (*United States of America*) or idioms (*living the life of Riley*). They are described as regular expressions or templates, i.e. sequences of words, parts of words and grammatical categories used to match similar patterns of text and extract them. Thus, the template *"ma[kd]*_V* {JJ, D*, AT*} sense_NN1"* will identify all occurrences of the verb *make* directly followed by an optional adjective (JJ), determiner (D*) or article (AT*) and the singular noun (NN1) *sense* and will consequently retrieve all instances of the MWU *make sense* and its variants *make no sense, little sense, more sense, common sense,* etc.

Part-of-speech tagging is essentially a **disambiguation task**. Many words are part-of-speech homographs, i.e. they are spelt the same but belong to different word classes. A tagger needs to determine which part-of-speech is the most probable given the immediate syntactic and semantic context of a homograph. Although close to 90% of the English types[120] have only one part-of-speech (e.g. *abound* can only be a verb and *kindness* is a noun), over 40% of the running words or tokens in a corpus are morpho-syntactically ambiguous (DeRose 1988:31). This is largely due to ambiguity for a number of high-frequency words such as '*that'*, which can be a determiner (*Do you remember **that** nice Mr. Hoskins who came to dinner?*)[121], a relative pronoun (*The people **that** live next door*), a conjunction (*I can't believe **that** he is only 17*) or even an adverb (*I hadn't realized the situation was **that** bad*!). Another very common source of ambiguity in English is homography between verbs and nouns, e.g. *use, issue, cause, abandon, craft,* etc. (cf. Ide 2005).

Most current part-of-speech taggers use, at least partly, a probabilistic approach to disambiguation: they rely on co-occurrence probabilities between neighbouring tags, indicating the relative likelihood of co-occurring tags. Co-occurrence probabilities are often automatically derived by training on manually disambiguated texts. For example, given that *x* is a determiner, the probability that the item to its immediate right is a noun or an adjective

[120] If a text is 75,000 words long, it has 75,000 *tokens*. But a lot of these words will be repeated, and there may be only 2,000 different words, i.e. *types*, in the text.

[121] Sentence examples are taken from the *Longman Dictionary of Contemporary English* (2005)

can be calculated. At the same time, non-probabilistic or rule-based taggers have been making a comeback with systems such as the one proposed by Brill (1992). Typical rule-based taggers use context frame rules to assign tags to unknown or ambiguous words. An example of a context frame rule is 'if an ambiguous or unknown word is preceded by a determiner and followed by a noun, tag it as an adjective.' See Voutilainen (1999) for a survey of the history of the different approaches to wordclass tagging.

CLAWS is a **hybrid tagger, combining both probabilistic and rule-based approaches**. This hybrid approach allows CLAWS to assign POS-tags with a very high degree of accuracy, the precise degree of accuracy varying according to text types (97-98% for written texts) and POS-tags (cf. Rayson 2003: 63). The tagger is commonly described as going through five major stages (cf. Garside 1987):

1. **A pre-editing or tokenization phase**: This stage prepares the text for the tagging process by segmenting it into words and sentence units, a task which is not trivial. A sentence is generally described as a string of words followed by a full stop. A full stop, however, does not necessarily signal the end of a sentence (e.g. in figures [5.8 or 14.28], title nouns [*Mr., Dr.*], and other types of abbreviations such as *i.e., viz., fig.*). Similarly, a word is generally considered as an orthographic word, i.e. a string of letters surrounded by white spaces. However, words are not always separated by a blank (e.g. in contractions such as *don't, it's, they're*).
2. **An initial part-of-speech assignment**: Once a text has been tokenized, the tagger assigns part-of-speech tags to all of the word tokens in the text in isolation. If a word is unambiguous, i.e. belongs to only one part-of-speech category or word class (e.g. *boat, person* and *belong*), it will be assigned a single tag. If a word is ambiguous, that is, if a word can belong to more than one word class (e.g. *use, cause, fire*, all of which can be categorized either as nouns or verbs), it is assigned several tags listed in decreasing likelihood. Thus, *fire* is first tagged as a noun and then as a verb as the probability of the word being a noun is higher than that of it being a verb. If a particular word is not found in the tagger's lexicon, it is assigned a tag based on various sets of rules, e.g. morphological rules, for tagging unknown items. Thus, a word ending in **ness* will be classified as a noun; a word ending in **ly* will be classified as an adverb, etc.
3. **A rule-based contextual part-of-speech assignment**: This stage assigns a single "ditto-tag" to two or more orthographic words which function as a single unit, e.g. *as*

well as is tagged as a conjunction and *in situ* as an adverb (see below for more details on ditto-tags and their advantages).

4. **The probabilistic tag-disambiguation program**: The task of the probabilistic tag-disambiguation program is to inspect all the cases where a word has been assigned two or more tags and choose a preferred tag by considering the context in which the word appears and assessing the probability of any particular sequence of tags. The probability of a tag sequence is typically a function of:
   - o the probability that one tag follows another one and
   - o the probability of a word being assigned a particular tag from the list of all its possible tags (Garside and Smith 1997: 104).

   If, for example, the word *run* has been assigned both a noun and a verb tag, it is less likely to be classified as a verb if it appears in the vicinity of another verb although *run* is more often a verb than a noun.
5. **Output**: The output data can be presented in intermediate format (vertical output for manual post-editing) or final format (horizontal and encoded in SGML[122]). Table 5.6 shows a typical CLAWS vertical output: each line represents a running word in the corpus and gives its POS-tag and lemma.

Table 5.6: CLAWS vertical output

	POS-tag	Word form	Lemma
0000005 730	AT	The	the
0000005 740	JJ	whole	whole
0000005 750	NN1	point	point
0000005 760	IO	of	of
0000005 770	AT	the	the
0000005 780	NN1	play	play
0000005 790	VVZ	seems	seem
0000005 800	TO	to	to
0000005 810	VBI	be	be
0000005 820	AT1	an	an
0000005 830	NN1	attack	attack
0000005 840	II	on	on
0000005 850	AT	the	the
0000005 860	NN1	Church	church
0000005 870	.	.	PUNC

The intermediate format has the advantage of allowing researchers to select the information needed. I wrote a Perl programme which takes this intermediate format as input and creates two horizontal corpora:

- a corpus with word forms followed by their POS-tags (cf. Table 5.7) and;

[122] Standard Generalized Markup Language

- a corpus with lemmas and POS-tags (cf. Table 5.8).

Table 5.7: CLAWS horizontal output [word form + POS]

```
The_AT whole_JJ point_NN1 of_IO the_AT play_NN1 seems_VVZ to_TO
be_VBI an_AT1 attack_NN1 on_II the_AT Church_NN1 ._PUNC

… with AT: article; JJ: adjective; NN1: singular common noun;
IO: of (as preposition); VVZ: -s form of lexical verb; TO:
infinitive marker 'to'; VBI: be, infinitive; AT1: singular
article; II: general preposition; PUNC: punctuation
```

Table 5.8: CLAWS horizontal output [lemma + POS]

```
the_AT whole_JJ point_NN1 of_IO the_AT play_NN1 seem_VVZ to_TO
be_VBI an_AT1 attack_NN1 on_II the_AT Church_NN1 ._PUNC
```

The problem with the format described in Table 5.8 is that word forms are replaced by their lemmas while POS-tags remain too specific, e.g. the information on number given by the tags NN1 (singular common noun) or DD1 (singular determiner) or the information on verbal forms given by the tags VVZ (*-s* form of a lexical verb) or VVG (*-ing* form of a lexical verb). As a result, frequency lists based on this corpus format generate different frequencies for '*example*_NN1' and '*example*_NN2'. POS-tags were thus automatically simplified by means of a Perl program to match the level of specificity of lemmas. Table 5.8b shows the same sentence as the one annotated in Table 5.8 after simplifying POS-tags. Simplification routines are described in Table 5.9.

Table 5.8b: CLAWS horizontal output [lemma + simplified POS tags]

```
the_AT whole_JJ point_NN of_IO the_AT play_NN seem_VV to_TO
be_VB an_AT attack_NN on_II the_AT Church_NN ._PUNC
```

Table 5.9: Simplification of CLAWS POS-tags

Simplified POS tags	CLAWS7 POS tags
Singular vs. plural forms	
MC (cardinal number)	MC1, MC2
NN (common nouns)	NN1, NN2
NNL (locative nouns, e.g. *island, street*)	NNL1, NNL2
NNO (numeral nouns, e.g. *hundred*)	NNO, NNO2
NNT (temporal nouns, e.g. *day, week*)	NNT1, NNT2
NNU (units of measurement, e.g. *inch*)	NNU1, NNU2
NP (proper nouns)	NP1, NP2
NPD (weekday noun)	NPD1, NPD2
NPM (month noun)	NPM1, NPM2
Comparative and superlative forms	
DA (after-determiners, e.g. *little, much, few*)	DAR (*more, less*), DAT (*most, fewest*)
JJ (adjective)	JJR, JJT
Verb forms	
VB (*be*)	VB0 (*be*, base form), VBDR (*were*), VBDZ (*was*), VBI (*be*, infinitive), VBM (*am*), VBN (*been*), VBR (*are*), VBZ (*is*)
VD (*do*)	VD0 (*do*, base form), VDD (*did*), VDG (*doing*), VDI (*do*, infinitive), VDN (*done*), VDZ (*does*)
VH (*have*)	VH0 (*have*, base form), VHD (*had*), VHG (*having*), VHI (*have*, infinitive), VHN (*had*), VHZ (*has*)
VV (lexical verbs)	VV0 (base form of lexical verb), VVD (pas tense), VVG (-ing participle), VVGK (-ing participle catenative, e.g. *be* ***going*** *to*), VVI (infinitive), VVN (past participle), VVNK (past participle catenative, e.g. *be* ***bound*** *to*), VVZ (-s form)

Finally, each CLAWS7 tag may be modified by the addition of a pair of digits to show that it occurs as part of a sequence of similar tags, representing a sequence of graphemic words which, for grammatical purposes, are best treated as a single unit. The expression *ahead of* is an example of a group of two graphemic words treated as a single preposition by receiving the following tags:

ahead_II21 of_II22.

II stands for a general preposition. The first of the two digits indicates the number of graphemic words in the sequence, and the second digit the position of each graphemic word within that sequence. Such "ditto tags" are not included in the lexicon but are assigned automatically by a rule-based component which is applied after initial part-of-speech assignment and before disambiguation and looks for a range of multi-word sequences included in a pre-established list.

Ditto tags are very useful for our purposes of developing a productively-oriented academic wordlist as they make it possible to extract typical EAP multi-word sequences as well as single words. However, the annotation format needs to be slightly modified to do so. Table 5.10 shows CLAWS vertical output for the complex preposition ‘*in terms of*’. Each graphemic word of the complex preposition is tagged and lemmatized independently. A wordlist based on CLAWS horizontal output would thus distinguish between the preposition ‘*in*’ (in_II) and the preposition ‘*in*’ used as the first word of a multi-word sequence (in_II31). It would not be able to retrieve the complex preposition ‘*in terms of*’. Any sequences of words with ditto tags (e.g. *in*_II31 *terms*_II32 *of*_II33) were thus automatically replaced by means of a Perl program by their component words separated by a hyphen and followed by their POS-tag (e.g. *in-terms-of*_II).

Table 5.10: CLAWS tagging of the complex preposition ‘in terms of’

```
0000006 040  II31        in          in
0000006 050  II32        terms       term
0000006 060  II33        of          of
```

5.2.2.2.2. The UCREL Semantic Analysis System (USAS)

A second layer of annotation is applied by the *UCREL Semantic Analysis System* (USAS). This tool assigns tags representing the general semantic field of words from a lexicon of single words and multi-word units. A semantic field is a theoretical construct which groups together “words that are related by virtue of their being connected – at some level of generality - with the same mental concept” (Wilson and Thomas 1997: 54). It will not only include synonyms and antonyms of a word but also its hypernyms and hyponyms, and any other words that are linked in other ways with the concept concerned. For example, the category ‘language and communication’ (Q) includes related words such as *answer, reply, response, question, query, statement, message, feedback, anecdote, explain,* and *explanation.*

The USAS tagset includes 21 major semantic fields (see Table 5.11), which, in turn, expand into 232 categories (see Appendix 5.2). Letters are used to denote the major semantic fields while numbers indicate field subdivisions. For example, the semantic tag A2.2 represents a word in the category ‘general and abstract words’ (A), the subcategory ‘affect’ (A2) and more precisely the sub-subcategory ‘cause / connected’ (A2.2). The semantic annotation does not apply to proper names and closed classes of words such as prepositions, conjunctions and pronouns. These categories are all marked with a Z-tag (for more information on the USAS tagset, see Archer et al. 2002).

Table 5.11: Semantic fields of the UCREL Semantic Analysis System

A	General and abstract terms
B	The body and the individual
C	Arts and crafts
E	Emotional actions, states and processes
F	Food and farming
G	Government and public
H	Architecture, house and the home
I	Money and commerce in industry
K	Entertainment, sports and games
L	Life and living things
M	Movement, location, travel and transport
N	Numbers and measurement
O	Substances, materials, objects and equipment
P	Education in general
Q	Language and communication
S	Social actions, states and processes
T	Time
W	World and environment
X	Psychological actions, states and processes
Y	Science and technology
Z	Names and grammar

Like part-of-speech tagging, semantic tagging subdivides broadly into a tag assignment phase and a tag disambiguation phase. First, a set of potential semantic tags are attached to each lexical unit. The next stage consists in the selection of the contextually appropriate semantic tag from the set of potential tags provided by the tag assignment algorithm. The program makes use of a number of sources of information in the disambiguation phase, notably POS-tags, domain of discourse and contextual rules (cf. Rayson 2003: 67-68). It assigns a semantic field tag to every word in the text with about 92% accuracy. Table 5.12 shows that in the sentence '*This chapter deals with the approach of the criminal law to behaviour which causes or risks causing death'*, the word *chapter* has been assigned the tags Q4.1 ('language and communication – media – books'), S5 ('social actions, states and processes - groups and affiliation'), S9 ('social actions, states and processes – religion and the supernatural') and T1.3. ('time – period').[123] The program ranked these semantic tags and chose Q4.1 as the semantic tag with the highest correctness probability which is displayed in the final output format (see Table 5.13).

[123] Note the erroneous annotation of 'risks' in Table 5.12 which should have received a VVZ tag (-s form of a lexical verb).

Table 5.12: USAS vertical output

```
            POS-tag  Word form                   Semantic tag
0000006 010  DD1      This                       M6 Z5 Z8
0000006 020  NN1      chapter                    Q4.1 T1.3 S9/S5 S5+
0000006 030  VVZ      deals      A1.1.1 I2.2 I2.1 A9- K5.2 F3/I2.2
0000006 040  IW       with                       Z5
0000006 050  AT       the                        Z5
0000006 060  NN1      approach                   X4.2 M1 E1 S1.1.1
0000006 070  IO       of                         Z5
0000006 080  AT       the                        Z5
0000006 090  JJ       criminal                G2.1[i1.2.1 G2.1- A5.1-
0000006 100  NN1      law                      G2.1[i1.2.2 G2.1 S6+ Y1
0000006 110  II       to                         Z5
0000006 120  NN1      behaviour                  S1.1.1 A1.1.1
0000006 130  DDQ      which                      Z8 Z5
0000006 140  VVZ      causes                     A2.2
0000006 150  CC       or                         Z5
0000006 160  NN2      risks                      A15-
0000006 170  VVG      causing                    A2.2
0000006 180  NN1      death                      L1-
0000006 181  .        .
```

Table 5.13: USAS horizontal output

```
 This_M6 chapter_Q4.1 deals_A1.1.1 with_Z5 the_Z5 approach_X4.2
of_Z5 the_Z5 criminal_G2.1[i1.2.1 law_G2.1[i1.2.2 to_Z5
behaviour_S1.1.1 which_Z8 causes_A2.2 or_Z5 risks_A15
causing_A2.2 death_L1- . PUNC
```

Words may signal simultaneously more than one semantic field for the same occurrence in a text. The word *chapter* in the sense of 'an ecclesiastical assembly of priests or monks' is a case in point. It belongs equally in the semantic fields of 'groups and affiliation' and 'religion'. The two semantic tags are thus assigned in the form of a single tag S5/S9 (see Table 5.12).

Provided that they are listed in the multi-word lexicon, multi-word units such as phrasal verbs (e.g. *break out, take off*), compounds (e.g. *academic year, advisory committee, bank account*), proper names (e.g. *Costa Rica, George Bush*) and idioms (e.g. *at the drop of a hat, to bark up the wrong tree, by the skin of one's teeth*) are also analysed as if they were single words, using ditto-tags similarly to part-of-speech tagging. For example, *criminal law* is tagged as follows: *criminal_G2.1[i1.2.1 law_G2.1[i1.2.2* (see Table 5.13).

5.3. Automatic extraction of EAP words

Methodological issues are particularly important for the selection of words that should be part and parcel of a productively-oriented academic word list. Section 5.3.1 provides a detailed description of the selection criteria used to extract potential EAP words, namely keyness, range and evenness of distribution. It also discusses the pros and cons of these criteria. Section 5.3.2 describes the manual adaptation of word spellings that was required to account for words that have two different spellings.

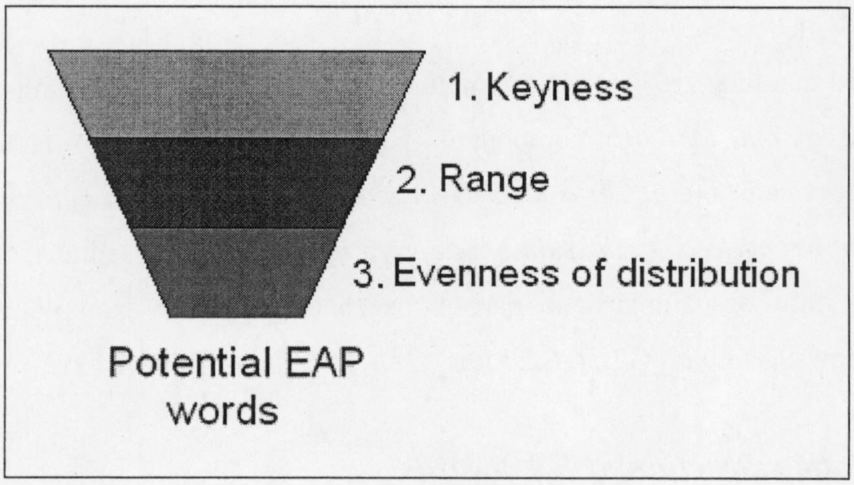
5.3.1. Selection criteria

Coxhead (2000) selected word families for the *Academic Word List* on the basis of three criteria (see section 1.3.1.2):

4. **Specialised occurrence**: a word family had not to be in the first 2,000 most frequent words of English as listed in West's (1953) *General Service List*.
5. **Range**: a word family had to occur in all 4 disciplines represented in the corpus with a frequency of at least 10 occurrences in each sub-corpus (about 875,000 words each) and in 15 or more of the 28 subject areas.
6. **Frequency**: a word family had to occur at least 100 times in the 3.5-million word Academic Corpus.

The method proposed here is primarily based on **keyness** (cf. Scott 2001), a criterion that has not been used by Coxhead (2000). Two quantitative filters, i.e. range and evenness of distribution, are subsequently used to overcome the limitations of the keyness criterion and narrow down the resulting list of EAP words (cf. Figure 5.1).

Figure 5.1: A three-layered sieve to extract EAP words



5.3.1.1. Keyness

The keyness method has been used in a variety of fields to extract distinctive words or keywords, e.g. words typical of speech vs. writing (Leech et al. 2001), business English words (Nelson 2000) and terminological items typical of specific sub-disciplines of English for Information Science and Technology (Curado Fuentes 2001). Keywords are words that "occur with unusual frequency in a given text" (Scott 1997:236), which does not mean high frequency but unusual frequency by comparison with a **reference corpus**.

For the purpose of this research, the two corpora of professional writing and the corpus of student academic writing described in section 5.2.1 were each compared with a large corpus of **fiction** on the basis of the hypothesis that typical EAP words would be particularly under-represented in this literary genre. Thus, our reference corpus was not compiled to represent all the varieties of the language[124] but to serve as a "strongly contrasting reference corpus" (Tribble 2001: 396). It consists of the categories K (general fiction), L (mystery and detective fiction), M (science fiction), N (adventure and western fiction) and P (romance and love story) of the LOB (Lancaster-Oslo/Bergen) Corpus, the FLOB (Freiburg Lancaster-Oslo/Bergen) Corpus, the BROWN corpus and the FROWN (Freiburg-Brown) Corpus[125] as well as of the Baby BNC fiction corpus (cf. Table 5.14.).

Table 5.14: The fiction corpus

Corpora	**Number of words**
LOB (categories K, L, M, N, P) FLOB (categories K, L, M, N, P) BROWN (categories K, L, M, N, P) FROWN (categories K, L, M, N, P)	946,337
Baby BNC fiction	999,688
TOTAL	**1,946,025**

The procedure for identifying keywords with *WordSmith Tools* (see section 4.2.1.3.1.) involves several stages (cf. Scott and Tribble 2006). First, wordlists are computed for each corpus, containing all the different types and their frequencies. Second, the wordlists of the three academic corpora are compared with the fiction corpus wordlist using the *Keywords* tool which produces a list of all the words that present statistically significant differences in

[124] See the definition of a reference corpus proposed by the Expert Advisory Group on Language Engineering Standards (EAGLES96) at http://www.ilc.cnr.it/EAGLES96/corpustyp/node18.html

[125] Each corpus consists of one-million words of British or American written English. The four corpora are equivalent in the sense that they were compiled using the same corpus design and sampling methods. For more information about these corpora, see http://khnt.hit.uib.no/icame/manuals

frequency between two wordlists. Keyness values can be calculated with chi-square or log-likelihood tests (cf. Dunning 1993). The latter was used in this thesis. The significance of the statistical test was set at 0.01 with a critical value of 15.13 (cf. Rayson *et al.* 2004), which means that there is less than 1% danger of mistakenly claiming a significant difference in frequency. Similarly, the minimum frequency of potential keywords was set at 10 occurrences to limit the extraction of rare words. Keywords were extracted for each of the corpora described in section 5.2.1, i.e. the Professional Soft Science (ProfSS) corpus, the Professional Hard Science (ProfHS) corpus and the native student writing corpus (STUDCORP), in word form + POS-tag format as well as lemma + POS-tag format. Tables 5.15 and 5.16 give the number of positive and negative keywords for each corpus in the two formats. Positive keywords are words that are statistically prominent in the three corpora of academic writing while negative keywords are words that have strikingly low frequency in this genre when it is compared with fiction writing. In other words, negative keywords are words that are statistically prominent in fiction texts.

Table 5.15: Word forms + POS-tags: Number of keywords

Corpus	Positive keywords	Negative keywords
Professional hard science	5,098	906
Professional soft science	5,623	1,343
Student writing	5,899	1,117

Table 5.16: Lemmas + POS-tags: Number of keywords

Corpus	Positive keywords	Negative keywords
Professional hard science	4,322	837
Professional soft science	4,656	1,201
Student writing	4,492	956

Positive keywords are more numerous than negative keywords for each academic corpus and for both corpus formats, which can be explained by the large amount of specialized vocabulary present in academic texts, e.g. *formula, cell* and *species* in hard science, *law, offence* and *policy* in soft science and *theory, factor* and *participant* in student writing. However, they do not all qualify as potential EAP or academic words in accordance with the definition of academic vocabulary provided in section 1.3.2.2. Keyness is a very powerful tool but it emphasizes words used with markedly high frequency only. The resulting list is therefore likely to include discipline or topic-related vocabulary, i.e. technical rather than semi-technical words that are not necessarily frequent in academic texts but are typically under-represented in fiction writing (e.g. *bacterium, methane, DNA, penicillin, chromosome,* *enzyme, jurisdiction, rape, archbishop, martyr,* etc.) (see section 1.3.1.3). Johannsson's (1981; quoted in Kennedy 1998) list of the most distinctive nouns in the academic subpart of the LOB corpus is a case in point. Table 5.17 clearly shows that the LOB sub-corpus of 'learned and scientific writings' (section J) contains more scientific texts than other discipline-related texts.

Table 5.17: The most distinctive nouns in the LOB sub-corpus of learned and scientific writing (Johannsson 1981)

constants, axis, equations, oxides, equation, theorem, coefficient, ions, correlation, electrons, impurities, oxidation, parameters, nickel, electron, impurity, diagram, ion, parameter, coefficients, oxygen, sodium, equilibrium

As a first step to overcome this inherent limitation of keyness analysis, I wrote a Perl program which automatically compares keywords for the three corpora and creates a list of positive keywords that are shared in the three corpora (cf. Scott's (1997) notion of 'key keywords'). Table 5.18 gives the number of shared positive keywords for each corpus format.

Table 5.18: Shared positive keywords

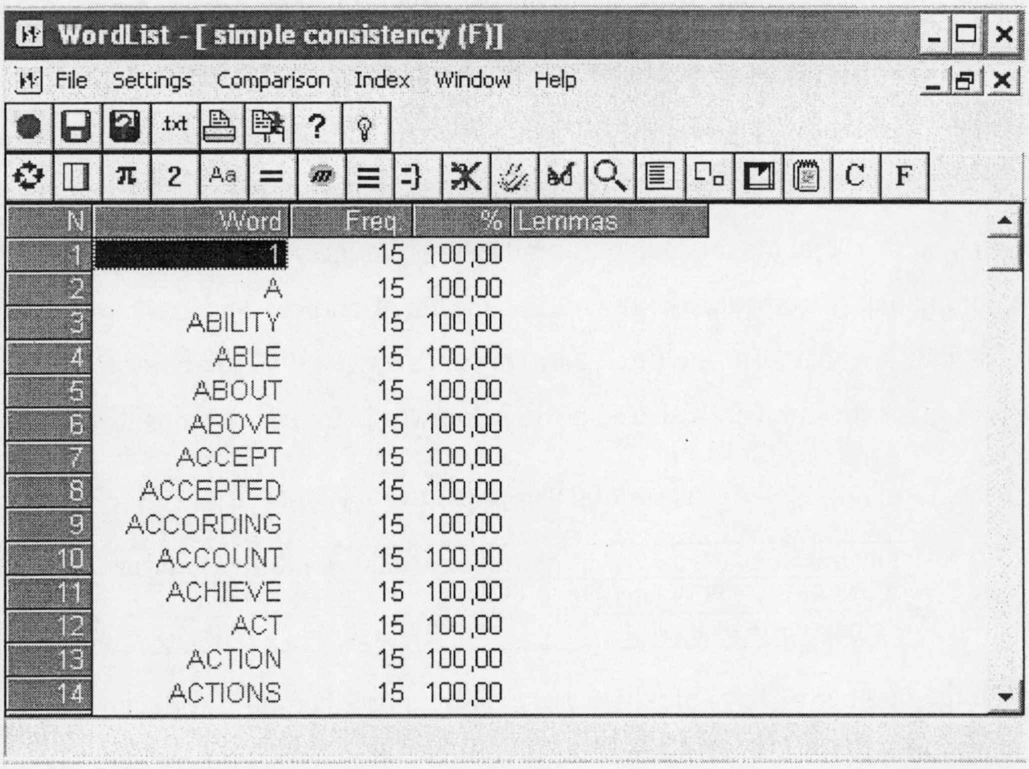
Format	**Number of shared positive keywords**
Word forms + POS-tags	2,048
Lemmas + POS-tags	1,642

Although the number of keywords was reduced by more than 60% for each corpus format, it remained quite high, i.e. 2,048 shared keywords for the word forms + POS-tags format and 1,642 shared keywords for the lemmas + POS-tags format. The criteria of range and evenness of distribution were subsequently used to refine the list of EAP keywords.

5.3.1.2. Range

To further distinguish keywords that are likely to be found in most academic texts from others that are restricted to a specific discipline (e.g. *property, psychological, treatment*, etc.), the criterion of **range** or **consistency**, i.e. frequency in terms of the number of texts, was also used (cf. Scott and Tribble 2006:29). The criterion of range, applied after a keyword analysis, helps determine whether a keyword is frequent because it occurs in most academic disciplines or whether it is frequent because of a very high usage in a limited subset of texts. It was calculated on the basis of the 15 sub-corpora described in sections 5.2.1.1 and 5.2.1.2 with the *Simple Consistency* facility of *WordSmith Tools*. This tool takes several wordlists as input, compares them and produces a wordlist which shows the 'frequency' of words in terms of the number of texts in which they appear, i.e. their range. Thus, in Figure 5.2, the words *ability*, *able* and *about*, for example, are shown to appear in the 15 sub-corpora, that is, in 100% of the corpora analyzed. For the purposes of this study, **only words appearing in the 15 academic sub-corpora were retained as potential EAP words**.

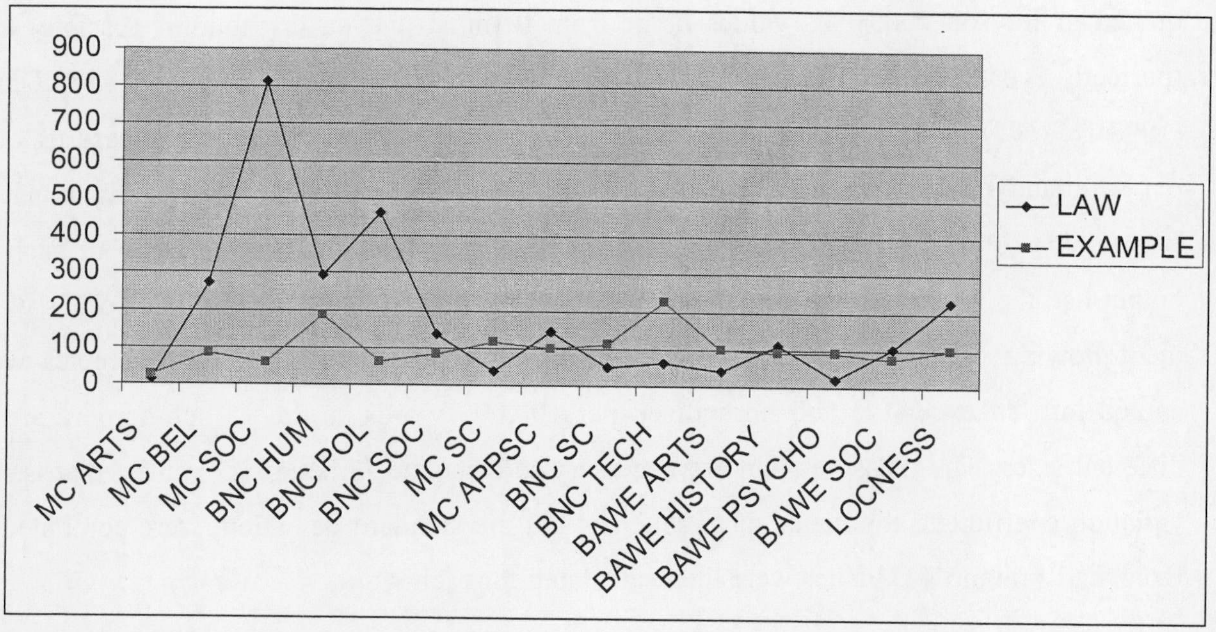
Figure 5.2: WordSmith Tools Simple Consistency Analysis



N	Word	Freq.	%	Lemmas
1	1	15	100,00	
2	A	15	100,00	
3	ABILITY	15	100,00	
4	ABLE	15	100,00	
5	ABOUT	15	100,00	
6	ABOVE	15	100,00	
7	ACCEPT	15	100,00	
8	ACCEPTED	15	100,00	
9	ACCORDING	15	100,00	
10	ACCOUNT	15	100,00	
11	ACHIEVE	15	100,00	
12	ACT	15	100,00	
13	ACTION	15	100,00	
14	ACTIONS	15	100,00	

Used alone, range also has an important limitation: it gives no information on the frequency of a word in the 15 sub-corpora. Thus, the criterion of range dismisses the words *sector, paradigm* and *variance* as they only appear in 11 sub-corpora but includes both the words *example*, which we intuitively regard as an EAP word, and *law,* the meaning of which is more discipline or topic-dependent (e.g. *the Canon Law, criminal law, the law of gravity*). Their respective frequencies are given in Figure 5.3 for each sub-corpus and reflect their difference. The frequency of the word *example* ranges from 26 to 226 in the 15 sub-corpora, while that of the word *law* ranges from 11 to 812. The wider frequency range of *law* can be explained by the peak frequency of occurrence of the noun in the professional soft science sub-corpora.

Figure 5.3: Distribution of the words *example* and *law* in the 15 sub-corpora



Differences such as these can be highlighted by a measure of the evenness of distribution of words in a corpus, the last criterion applied to further restrict our list of potential EAP words. For a similar technique, see Yang's (1986) measures of 'peakratio' and 'rangeratio'.

5.3.1.3. Evenness of distribution

The **evenness of distribution** or **dispersion** of a word is "a statistical coefficient of how evenly distributed a word is across successive sectors of the corpus" (Rayson 2003: 93). A number of studies (cf. Zhang et al. 2004) have used a measure of dispersion to define a core lexicon on the basis that "if a word is commonly used in a language, it will appear in different parts of the corpus. And if the word is used commonly enough, it will be well-distributed". One such measure is Juilland's D statistical coefficient. Juilland's D was first used in the *Frequency Dictionary of Spanish Words* (Juilland and Rodriguez 1964) and is calculated as follows:

$$\mathrm{D} = 1 - V / \sqrt{n}\text{-}1$$

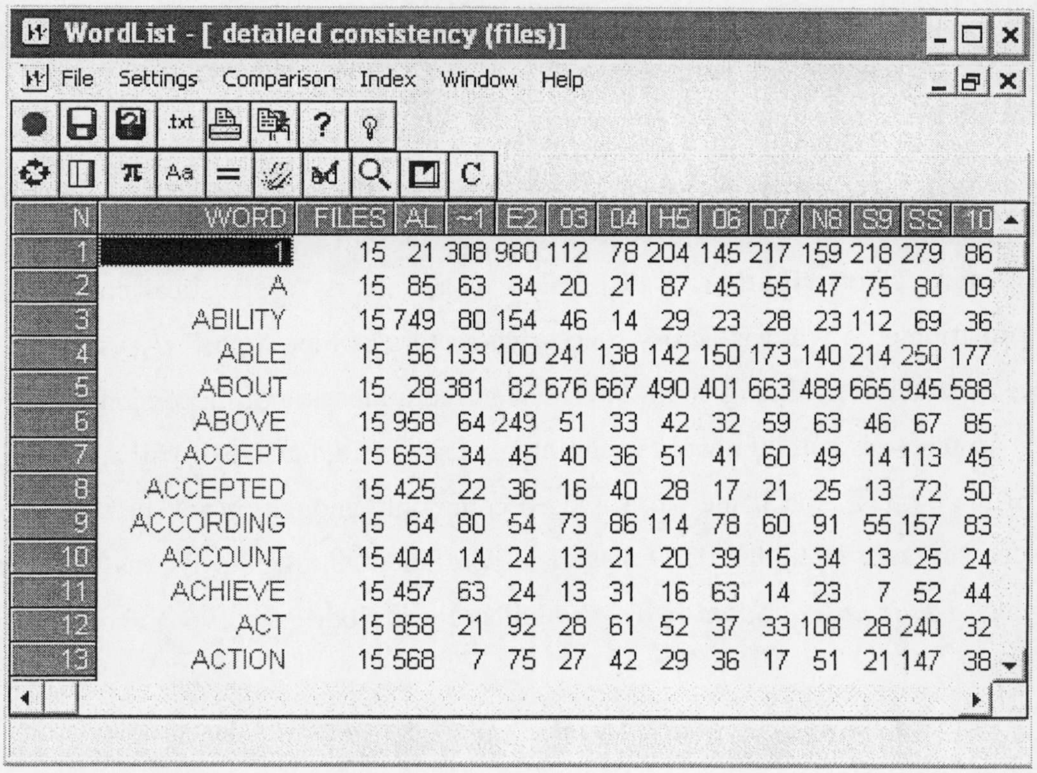
where n is the number of sectors, i.e. the number of sub-corpora or texts, in the corpus. The variation coefficient V is given by:

$$V = s / x$$

where x is the mean sub-frequency of the word in the corpus and s the standard deviation of these sub-frequencies. We have selected **Juilland's D value** as it has been shown to be the most reliable of the various dispersion coefficients that are available (Lynne 1985, 1986, quoted in Rayson 2003). Its values range from 0 (most uneven distribution possible) to 1 (perfectly even distribution across the sectors of the corpus). The reader is referred to Oakes (1998:189-192) for more information on evenness of distribution or dispersion measures.

Juilland's D values are calculated for each word on the basis of the output list of *WordSmith Tools Detailed Consistency Analysis*. Figure 5.4 shows an example of such an output list: the third column gives the total frequency of each word in the whole corpus while the following columns show its frequencies in each sub-corpus. These frequencies were copied into an excel file and normalized per 100,000 words as the 15 sub-corpora are of different sizes. The measures that are necessary to calculate Juilland's D values, that is, the variation coefficient, the mean sub-frequency and the standard deviation, were computed in Excel and Juilland's D values were then calculated for each word.[126]

Figure 5.4: WordSmith Tools Detailed Consistency Analysis



N	WORD	FILES	AL	~1	E2	03	04	H5	06	07	N8	S9	SS	10
1	1	15	21	308	980	112	78	204	145	217	159	218	279	86
2	A	15	85	63	34	20	21	87	45	55	47	75	80	09
3	ABILITY	15	749	80	154	46	14	29	23	28	23	112	69	36
4	ABLE	15	56	133	100	241	138	142	150	173	140	214	250	177
5	ABOUT	15	28	381	82	676	667	490	401	663	489	665	945	588
6	ABOVE	15	958	64	249	51	33	42	32	59	63	46	67	85
7	ACCEPT	15	653	34	45	40	36	51	41	60	49	14	113	45
8	ACCEPTED	15	425	22	36	16	40	28	17	21	25	13	72	50
9	ACCORDING	15	64	80	54	73	86	114	78	60	91	55	157	83
10	ACCOUNT	15	404	14	24	13	21	20	39	15	34	13	25	24
11	ACHIEVE	15	457	63	24	13	31	16	63	14	23	7	52	44
12	ACT	15	858	21	92	28	61	52	37	33	108	28	240	32
13	ACTION	15	568	7	75	27	42	29	36	17	51	21	147	38

For a word to be selected as an EAP word, it was decided that its Juilland's D value had to be higher than 0.8. A Juilland's D value of 0.8 is arguably quite low (cf. Rayson 2003:94); however increasing this value results in a sharp reduction of the number of potentially EAP

[126] Scott's (2004) *WordSmith Tools* 4 is now able to compute Juilland's D values but they are computed for words in a single file and are based on an arbitrary division of a text into 8 segments of equal size.

words retrieved by the method. Thus, the noun *example* is selected as a potential EAP word as its dispersion value equals 0.83 whereas the noun *law*, with a Juilland's D value of 0.69, is not. Dispersion values make it possible to avoid the wrong conclusion that these two words behave similarly in academic writing and confirm that only *example* is of widespread and general use in this particular genre, while the noun *law* is over-represented in the professional soft science corpus, and more specifically in the social science sub-corpus. Other examples of words that are selected as potential EAP words are the nouns *conclusion* (D= 0.88), *difference* (D = 0.9), *extent* (D = 0.87), *significance* (D = 0.86) and *consequence* (D = .85) and the verbs *prove* (D = 0.9), *appear* (D = 0.9), *provide* (D = 0.89), *discuss* (D = 0.88), *show* (D = 0.88), *result* (D = 0.87) and *illustrate* (D = 0.85). Examples of words that have D values lower than 0.8 and are therefore not selected include the nouns *health, employment* and *treatment* and the verbs *to label, to perceive* and *to yield.*

5.3.2. Manual adaptation of spelling differences

Results of the automatic extraction revealed that a number of words are not selected as potential EAP words only because they have two different spellings (e.g. *analyse* - *analyze*). We therefore manually adjusted all the frequencies of words that can be spelt differently and added their frequencies under a single lemma, that is, the most frequent lemma or the British spelling of the lemma when the difference was a British vs. American English spelling difference. Thus, the frequency of '*characterise*' was added to that of '*characterize*' (cf. Table 5.19), the frequency of '*center*' was added to that of '*centre*', the frequency of '*behavior*' was added to that of '*behaviour*', etc. Table 5.19 also shows that spelling differences are also found in a single variety of English and are not only due to the fact that we made use of corpora of British and American English.

Table 5.19: Distribution of the two spellings of 'characterize'

	Total	**Range**	**Distribution in the 15 sub-corpora**															**D**
characterise	110	12	1	1	0	35	2	5	1	0	7	10	9	8	21	10	0	0.68
characterize	148	12	17	19	9	19	15	13	17	8	23	2	5	0	0	0	1	-
Added frequencies	**258**	**15**	**18**	**20**	**9**	**54**	**17**	**18**	**18**	**8**	**30**	**12**	**14**	**8**	**21**	**10**	**1**	**0.85**

Adjusted frequencies allowed us to select as EAP words verbs such as *analyze, characterize* and *emphasize*.

5.4. Results of the automatic extraction

The automatic extraction procedure described in section 5.3 has been applied to the two corpus formats, i.e. word forms + POS-tags and lemmas + POS-tags. Results for both corpus formats are reported in section 5.4.1 and 5.4.2 respectively. Section 5.4.3 compares these results and discusses the pros and cons of each corpus format for the extraction of potential EAP words.

5.4.1. Corpus format: word form + POS-tag

In section 5.3.1.1, 2,048 key word forms were found to be shared across the three corpora of academic writing. These 2,048 key word forms include 798 nouns, 536 verbs, 363 adjectives and 128 adverbs. The high proportion of key **nouns** is consistent with Biber et al.'s (1999) finding that nouns are particularly frequent in academic prose. Table 5.20 shows that the application of the criteria of range and evenness of distribution reduces the total number of keywords by more than 65%. It also shows that the percentage of nouns decreases significantly (from 39% to 28% of the total number of keywords) when these criteria are applied. By contrast, the proportion of most other categories, and more specifically that of **verbs** and **adverbs**, increases. This seems to indicate that a larger proportion of key nouns are discipline-specific and have a technical meaning. Verbs represent the largest group with 32% of the keywords, which may suggest that they are more likely to have a sub-technical meaning (cf. section 1.3.2.2). Although usually disregarded in academic textbooks and teaching materials, **adjectives** amount to 19% of the potential EAP word forms. The category "others" mainly consists of cardinal numbers and proper names and is not further analyzed.

Table 5.20: EAP word forms

	Keywords		**In 15 texts**		**D > 0.8**	
Nouns	798	[39%]	425	[36.3%]	194	[28%]
Verbs	536	[26.2%]	328	[28%]	225	[32%]
Adjectives	363	[17.7%]	201	[17%]	134	[19%]
Adverbs	128	[6.25%]	89	[7.8%]	69	[10%]
Prepositions	45	[2.2%]	34	[2.9%]	28	[4%]
Conjunctions	16	[0.8%]	15	[1.3%]	14	[2%]
Determiners	16	[0.8%]	16	[1.4%]	16	[2.3%]
Pronouns and articles	6	[0.29%]	6	[0.6%]	6	[0.86%]
Ordinal numbers	3	[0.15%]	3	[0.3%]	3	[0.43%]
Others	137	[6.69%]	55	[4.69%]	10	[1.4%]
TOTAL	**2,048**	**[100%]**	**1,172**	**[100%]**	**699**	**[100%]**
	(100%)		(57%)		(34%)	

Table 5.21 gives the first 20 EAP word forms per POS-tag ordered in decreasing order of keyness. The complete list is given in Appendix 5.3. A high number of these words correspond to the restricted definition of **academic vocabulary** given in section 1.3.2.2, that is, words that "have in common a focus on research, analysis and evaluation – those activities which characterize academic word" (Martin 1976:92). Examples include *problems, evidence, approach, issue, result, shows, appears, considered, defined, significant, major, therefore, for example, consequently, especially, whereas, because, despite.* It is worthy of note that three word forms of the verb ***be*** – *is, are* and *be* - appear in the top five list of EAP verb forms. This finding is consistent with Biber et al (1999) who found that copular *be* is most frequent in academic prose (see Appendix 1.1). The first 20 EAP adverbs show that **linking adverbials** and **evaluative adverbs** are two important categories of adverbs in this particular genre (cf. Conrad 1999). Results also stress the added value of CLAWS **multi-word unit** lexicon as 40% of the prepositions are complex ones, e.g. *such as, according to, in terms of, because of.* The list of pronouns and articles is very short and is mainly composed of **articles and 3rd person pronouns**. The first two determiners are the **demonstratives** *this* and *these*, which are most probably used to introduce nouns functioning as retrospective labels, lexical cohesion being particularly frequent in academic prose (see section 1.3.2.2).

Table 5.21: First 20 EAP word forms per POS-tag

Nouns	*role, system, use, problems, concept, evidence, process, approach, individuals, effect, form, issue, values, environment, individual, result, effects, groups, influence, differences*
Verbs	*is, are, has, can, be, may, based, will, used, shows, using, becomes, seems, appears, provides, considered, defined, allows, provide, does*
Adjectives	*social, significant, human, important, individual, different, physical, greater, major, effective, specific, negative, similar, common, positive, central, single, general, complex, higher*
Adverbs	*also, however, therefore, for example, thus, more, most, largely, both, often, further, generally, clearly, consequently, effectively, especially, highly, previously, necessarily, in general*
Conjunctions	*that, although, as, whether, or, whereas, because, in that, provided, whether or not, given that, rather than, since, even though*
Prepositions	*of, as, in, by, such as, within, between, during, rather than, upon, according to, in terms of, despite, including, as well as, because of, for, to, subject to, as opposed to*
Pronouns and articles	*the, their, its, an, themselves, itself*
Determiners	*this, these, which, many, each, such, both, most, latter, same, former, less, little, fewer, those, several*
Ordinal numbers	*third, second, first*

5.4.2. Corpus format: lemma + POS-tag

In section 5.3.1.1, 1642 key lemmas were found to be shared by the two corpora of professional academic writing and the corpus of student writing. These 1,642 key lemmas include 624 nouns, 301 verbs, 371 adjectives and 130 adverbs. Table 5.22 shows that the application of the criteria of range and evenness of distribution reduces the total number by 63% and gives a breakdown per word class. **Nouns** constitute 27% of all potential EAP lemmas. The percentage of **verbs** is quite understandably smaller than for word forms as each verb form has been lemmatized. As a result, unlike in Table 5.21, verbs and nouns account for the same proportion of potential EAP lemmas. **Adjectives** represent 21% of all extracted lemmas, which suggests that more attention should be devoted to the teaching of adjectives and their specific role in academic writing. Like for the results based on word forms + POS-tags, the category "others" mainly consists of cardinal numbers and proper names. It is interesting to note that although the initial total number of key word forms was higher than the total number of key lemmas, the application of the criteria of range and evenness of distribution reduces the total number of key word forms more significantly than that of key lemmas. Except for nouns and verbs, the resulting list of potential EAP lemmas presents close similarities in number with our list of potential EAP word-forms. Thus, both lists include 69 adverbs, 28 prepositions, 6 pronouns or articles and 3 ordinal numbers.

Table 5.22: EAP lemmas

	Keywords		**In 15 texts**		**D > 0.8**	
Nouns	624	[38%]	370	[36.9%]	167	[27%]
Verbs	301	[18.3%]	213	[21.2%]	162	[26.5%]
Adjectives	371	[22.6%]	203	[20.2%]	132	[21%]
Adverbs	130	[7.9%]	9	[9%]	69	[11%]
Prepositions	17	[1%]	16	[1.6%]	15	[2.5%]
Conjunctions	46	[2.8%]	35	[3.5%]	28	[4.6%]
Determiners	6	[0.4%]	6	[0.6%]	6	[1%]
Pronouns and articles	17	[1%]	17	[1.7%]	17	[2.8%]
Ordinal numbers	3	[0.2%]	3	[0.3%]	3	[0.5%]
Others	127	[7.7%]	49	[4.9%]	10	[1.6%]
TOTAL	**1,642**	**[100%]**	**1,003**	**[100%]**	**609**	**[100%]**
	(100%)		**(61%)**		**(37%)**	

As it mainly consists of cardinal numbers and proper names, the category "others" will not be further analyzed. The total number of potential EAP lemmas is thus **599** (=609 – 10).

Table 5.23 gives the first 20 EAP lemmas per POS-tag. The reader is referred to Appendix 5.4 for the complete list ordered by decreasing keyness. A comparison of Table 5.21 and Table 5.23 shows that a large proportion of adjectives, adverbs, conjunctions, prepositions, pronouns and articles, determiners and ordinal number are shared by the two lists. The first 20 adverbs are the same irrespective of the corpus format used. By contrast, results are significantly different for **verbs**. The verb *to be* does not appear in the top 20 EAP lemmas while its forms *is, are* and *be* rank in the top 5 EAP word forms. It appears in 44th position, most probably because the high frequencies of *is, are* and *be* in academic writing are counterbalanced by the high frequency of *was*, and to a lesser extent, of *am* in fiction. Other highly frequent verb forms are *has* and *does*. Potential EAP verb lemmas mainly consist of specialized lexical verbs such as *argue, demonstrate* and *support.* Note, also, that pronouns, articles and determiners are not correctly lemmatized by CLAWS. Thus, *this, these* and *those* are analyzed as separate lemmas and are lemmatized as *this, these* and *those* respectively (cf. Table 5.23).

Table 5.23: First 20 EAP lemmas per POS-tag

Nouns	*role, development, system, result, problem, individual, effect, period, issue, value, concept, process, example, form, level, use, approach, group, relationship, evidence*
Verbs	*argue, provide, use, may, create, base, can, become, develop, support, define, suggest, increase, present, achieve, demonstrate, consider, represent, show, describe*
Adjectives	*social, significant, human, important, different, individual, physical, major, effective, specific, similar, common, negative, positive, central, single, new, general, complex, crucial*
Adverbs	*also, however, therefore, for example, thus, more, most, largely, both, often, further, generally, clearly, especially, effectively, consequently, highly, previously, necessarily, in general*
Conjunctions	*that, although, as, or, whether, whereas, because, in that, provided, whether or not, given that, rather than, since, even though, than*
Prepositions	*of, in, as, by, such as, within, between, during, rather than, upon, according to, in terms of, despite, to, for, including, as well as, because of, from, subject to*
Pronouns and articles	*the, their, its, an, themselves, itself*
Determiners	*this, these, which, many, such, each, both, most, same, latter, former, less, those, little, fewer, several, some*
Ordinal numbers	*third, second, first*

5.4.3. A comparison of corpus formats for the extraction of EAP vocabulary

The choice between word forms and lemmas for the selection procedure is an issue for **nouns** and, more importantly, **verbs**. In this section, results for the two formats are compared to decide on which format the selection of EAP words should be based. Nouns and verbs, both as key word forms and key lemmas, are compared and classified under three categories:

- Words that are only extracted by the automatic procedure when applied to a word-form + POS-tag corpus format (**'EAP word forms only'** in Table 5.24)
- Words that are only extracted by the automatic procedure when applied to a lemma + POS-tag corpus format (**'EAP lemmas only'** in Table 5.24)
- Words that are extracted from both corpus formats (**'Shared EAP words'** in Table 5.24)

Table 5.24 shows that while the percentage of shared nouns is relatively high (81%), the picture is quite different for shared verbs (58%).

Table 5.24: Word forms vs. lemmas

	EAP word forms only	EAP lemmas only	Shared EAP words	Total number of EAP words retrieved
Nouns	15	19	148 [81%]	182
Verbs	28	51 [27%]	111 [58%]	190

The selection procedure applied to word forms + POS-tags retrieves 111 verbs that are also extracted when lemmas + POS tags are used. It also retrieves 28 verbs that are only extracted when word forms + POS-tags are used. Among the 28 potential EAP verbs that are only retrieved by an analysis of the word form + POS-tag corpus format, many are word forms of **high frequency verbs** such as *have, make, need, seem, do, find, give* and *take* (cf. Table 5.25).

Table 5.25: EAP word forms only

VVZ (-s)	adds, continues, does, falls, has, lies, makes, means, needs, points, raises, seems, takes
VVN (past participle)	carried, completed, faced, found, given, held, ignored, made, needed, placed, published
VVO (base form)	continue, have, offer, seem, share, understood
VVI (infinitive)	assist, continue, ensure, satisfy
VVG (-ing)	understanding

When carried out on lemmas + POS tags, the same procedure extracts the 111 shared verbs as well as 51 more verbs. These additional verbs are mainly **specialized lexical verbs** that have clear rhetorical functions in EAP, e.g. *assign, clarify, classify, concentrate, conduct, consist, contrast, contribute, demonstrate, differentiate, effect, enhance, exclude, expand, experience, formulate, function, generate, initiate, neglect, overcome, reject, rely, resolve, strengthen,* and *stress.*

As it makes it possible to retrieve more lexical verbs, **the lemma + POS-tag corpus format is preferred to the word form + POS-tag corpus format to select words that are typical of academic texts.** However, the word form + POS-tag corpus format is also highly valuable as argued in Granger and Paquot (2005). An application of the selection procedure on this particular format reveals that **key verbs are not necessarily keywords in all their word forms in academic discourse.** Thus, the key verb lemmas *link* and *describe* are key word forms only as *–ed* forms.[127] The following sentences show typical uses of the key word forms *linked* and *described* in academic writing:

5.1. *These ideas are closely* ***linked*** *to Emerson's ideas within Self-Reliance and his beliefs in general.* (PSS)

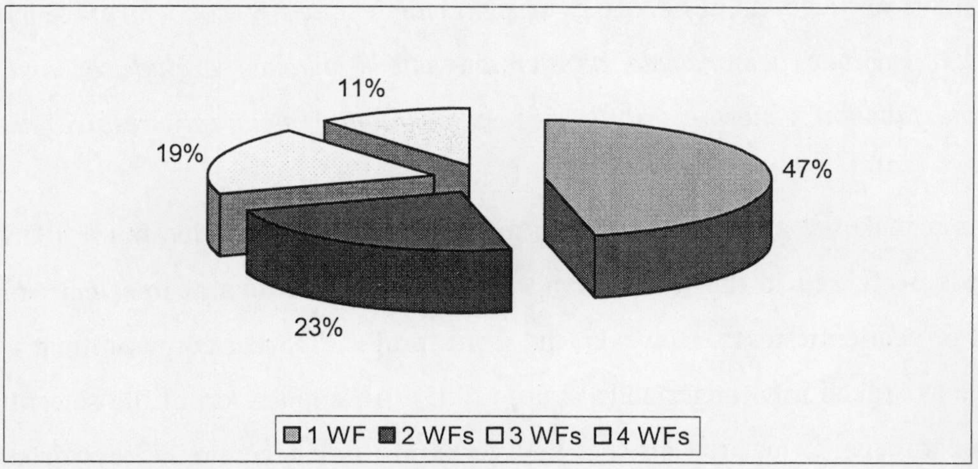
5.2. *The structure of these solutions will be* ***described*** *as they are derived in the following chapters.* (PHS)

It is important to note that almost **50% of the verbs that are selected as EAP key lemmas are keywords in one word form only** and that only 11% of these verbs are keywords in four word forms, e.g. *allow, attempt, lead* and *provide* (cf. Figure 5.5.) These findings support Hunston's (2002: 80-81) claim that "it cannot be assumed that all forms of a lemma behave in the same way". They also provide good reasons for teaching the most frequently used form in academic prose rather than the lemma in EAP courses. As argued by Sinclair,

> Traditionally, the 'base', or uninflected, form is used even when that form is hardly ever found on its own, or hardly ever found at all. But a case could be made for any of a number of alternatives, for example, that the most frequently encountered form should be used for the lemma (Sinclair 1991:42).

[127] Granger (2006) analyzes key verb forms and shows that *–ed* forms, and more particularly, past participle forms are the most distinctive verb forms in academic prose (see also Curado Fuentes 2001:111).

Figure 5.5: Number of key word forms per key lemma



5.5. Broadening the scope of well-represented semantic categories in EAP

The resulting list of the automatic extraction consists of 599 lemmas[128] that were retrieved on the basis of three criteria. First, they had to be **keywords** in the two corpora of professional hard and soft science academic writing and in the corpus of student writing. Second, they had to be characterized by **wide range**, that is, to appear in the 15 sub-corpora that represent different academic disciplines. Third, they had to be **well distributed** across the corpora and have a Juilland's D value higher than 0.8. Although results are already quite satisfactory, these criteria sometimes appear to be too restrictive. The sub-corpora are relatively small and frequencies may be skewed just because of a topic or an author's preferred turns of phrases. A semi-automatic procedure is thus proposed to complete the list of EAP lemmas and retrieve **semantically related words**.

In section 5.2.2.2, it was shown that a text uploaded to the web-based environment *Wmatrix* is morphosyntactically and semantically tagged. The semantic analysis is conducted with the *UCREL System* which classifies words and multi-word units into 21 major semantic categories. Table 5.26 shows the distribution of the 599 EAP lemmas across these semantic classes. Some words were automatically classified into more than one category but the figures given in Table 5.26 are based on the first semantic tag attributed to each word. Although there are errors, it is particularly interesting to note that 87% of the 599 EAP lemmas belong to only six of these categories. The category **'general and abstract terms'** includes almost half of

[128] As explained in section 5.4.2, the ten lemmas categorized as "others" are not further analyzed as they mainly consist of proper names and cardinal numbers.

the EAP lemmas. Examples include the nouns *activity, circumstance*, and *limitation* as well as the verbs *perform* and *cause*, the adjectives *detailed* and *particular* and the adverbs *similarly* and *conversely*. The category **'numbers and measurement'** accounts for more than 10 percent of the EAP lemmas and includes nouns (e.g. *degree, measure, amount, extent*), adjectives (e.g. *high, large, wide*), verbs (e.g. *extend, increase, reduce*), adverbs (e.g. *frequently, subsequently, also*) and prepositions (e.g. *in addition to*). The categories of **'language and communication'** (e.g. *argue, claim, define, suggest*), **'social actions, states and processes'** (e.g. *social, encourage, facilitate, impose*), **'psychological actions, states and processes'** (e.g. *assumption, analyse, interpretation, conclusion, attempt*) and **'names and grammar'** represent 5.6%, 7.7%, 9% and 8% of the EAP lemmas respectively. The category 'names and grammar' mainly consists of connective devices such as conjunctions (*or, whether*), prepositions (*such as, according to, since, during*) and adverbs (*moreover, thus, therefore*). The reader is referred to Appendix 5.5 for a categorization of all 599 potential EAP lemmas into the 232 semantic sub-categories of the USAS tagset.

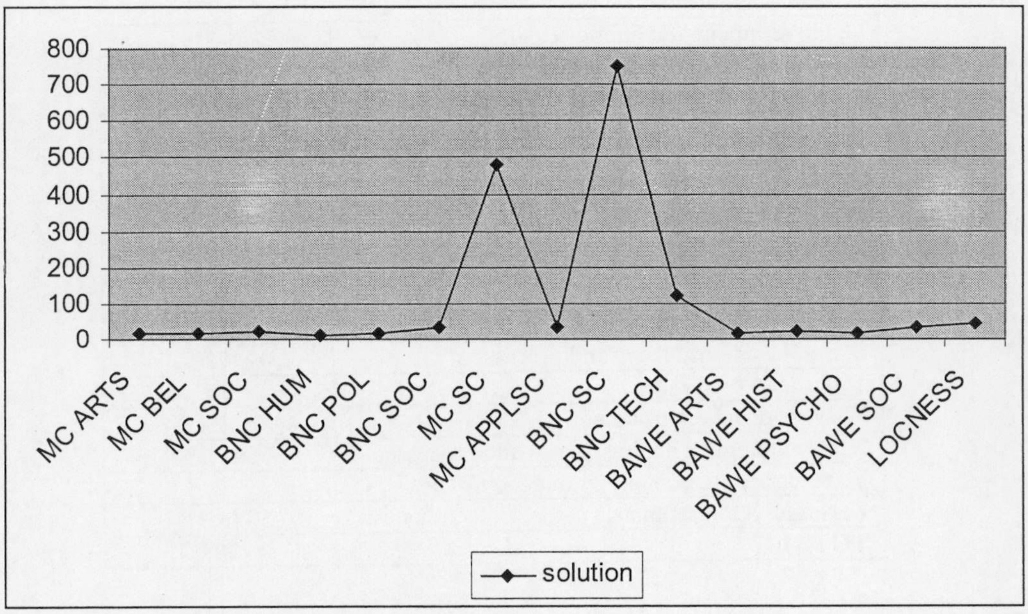
Table 5.26: Automatic semantic analysis of potential EAP words

Semantic categories	**Number of words**
A. General and abstract terms	267 [44.6%]
B. The body and the individual	2
C. Arts and crafts	2
E. Emotion	4
F. Food and farming	0
G. Government and public	4
H. Architecture, house and the home	2
I. Money and commerce in industry	7
K. Entertainment, sports and games	0
L. Live and living things	0
M. Movement, location, travel and transport	12
N. Numbers and measurement	74 [12.4%]
O. Substances, materials, objects and equipment	7
P. Education in general	4
Q. Language and communication	34 [5.7%]
S. Social actions, states and processes	47 [7.7%]
T. Time	26
W. World and environment	2
X. Psychological actions, states and processes	55 [9.2%]
Y. Science and technology in general	2
Z. Names and grammar	50 [8.3%]
TOTAL	**599**

On this basis, we added to our list of EAP lemmas keywords that do not have Juilland's D values higher than 0.8 but which belong to one of the six categories described above, namely 'general and abstract terms', 'numbers and measurement', 'language and communication', 'social actions, states and processes', 'psychological actions, states and processes' and 'names and grammar'. Among the words that are retrieved by this additional criterion, many are **morphologically related** to words that were automatically selected. For example, the noun *analysis* which is morphologically related to the EAP potential verb lemma *analyze* is retrieved by the semantic criterion although its Juilland's D is inferior to 0.8. It should be emphasized that the fact that many morphologically related words are retrieved by the semantic criterion is not an argument for using word families instead of lemmas. The criteria of minimum frequency and range still apply here and the noun *analysis* is retrieved only because it is very frequent in academic prose and appears in a wide range of academic texts. By contrast, other morphologically related words such as *analyst* or *analyzable* are not retrieved.

Other words have skewed Juilland's D values because of their **polysemy**. The noun *solution* is a case in point as it has both a sub-technical meaning and a technical meaning with different frequencies and distributional behaviours. Its sub-technical meaning is found in all academic sub-corpora while its technical meaning is particularly frequent in scientific writing and accounts for its much higher frequency in the two professional scientific sub-corpora (MC_SC and BNC-SC in Figure 5.6). These two peak frequencies of occurrence are responsible for the relatively low D value (54.6) of the noun *solution*.

Figure 5.6: Distribution of the noun 'solution'



The implementation of the semantic criterion, namely membership to the six prominent semantic categories in academic prose, results in an enlargement of our list of 331 words.

57% of these new words are **nouns** and 21% consist of verbs. Table 5.27 shows a breakdown of the final list by grammatical categories and the complete list is given in Appendix 5.5.

Table 5.27: Distribution of grammatical categories in the *Academic Keyword List*

	Automatic extraction[129]		New words		Total	
Nouns	167	[27.9%]	188	[57%]	355	[100%]
Verbs	162	[27%]	71	[21%]	233	[100%]
Adjectives	132	[22%]	48	[14%]	180	[100%]
Adverbs	69	[11.5%]	18	[5%]	87	[100%]
Others	69	[11.5%]	6	[1.8%]	75	[100%]
TOTAL	**599**	**[100%]**	**331**	**[100%]**	**930**	**[100%]**
	(64.4%)		**(35.6%)**		**(100%)**	

5.6. A comparison with the Academic Word List

The methodology used in this study to extract EAP words is quite different from that employed by Coxhead (2000) to design the *Academic Word List*. There are three major differences between the two approaches. First, Coxhead (2000) does not include General Service words in the *Academic Word List* but we do. Second, the author does not use the criterion of keyness, which is the first criterion employed here. Third, she applies the criteria of minimum frequency and range on word families rather than word forms or lemmas. It is therefore methodologically very interesting to compare our list with the *Academic Word List* and examine how different criteria influence the selection of EAP words. The comparison is made by uploading our *Academic Keyword List* to the *Web Vocab Profile* developed by Tom Cobb[130]. This web interface takes a text file as input, analyses its vocabulary and classifies words into four main categories: (1) words belonging to the first 1,000 most frequent words of English, (2) words belonging to the second 1,000 most frequent words of English, (3) words belonging to the *Academic Word List* and (4) off-list words.

Before uploading the list of EAP lemmas, it was necessary to remove multi-word units such as the adverbs *for example* and *for instance* or the complex prepositions *in addition to, due to, prior to, in the light of, in favour of* and *because of. Web Vocab Profile* cannot deal with multi-word units and would decompose them into their different parts. The comparison is thus based on single words only. Results show that only 40% of the EAP lemmas are shared with the *Academic Word List* while **57% belong to the first 2,000 most frequent words of**

[129] Excluding the 'others' category.
[130] Available at http://www.lextutor.ca/vp/eng/

English as described in West's (1953) *General Service List*. These results highlight the important role played by General Service words in academic prose and **justify their inclusion in a productively-oriented academic keyword list**. Table 5.28 gives the distribution of EAP keywords in the GSL and the AWL together with examples.

Table 5.28: Distribution of EAP keywords in the GSL and the AWL

Lists	%	Examples
GSL	57%	*aim, argue, argument, because, compare, comparison, differ, difference, discuss, example, exception, explain, explanation, importance, include, increasingly, likelihood, namely, point, reason, result, therefore, typically*
AWL	40%	*accurately, adequate, analysis, assess, comprise, conclude, conclusion, consequence, emphasize, hypothesis, inherent, method, proportion, relevance, scope, summary, survey, theory, validity, whereas*
Off-list	3%	*assertion, correlation, criticism, exemplify, proposition, reference, tackle, versus, viewpoint*

Web Vocab Profile also computes an index of Graeco-Latin and French cognates vs. Anglo-Saxon words. Our *Academic Keyword List* consists of **82.2% of Graeco-Latin and French cognates**, which gives empirical support to Corson's claim that "control of the Graeco-Latin academic vocabulary of English is essential to academic success" (Corson 1997:671). In addition, 74.61% of the EAP lemmas that belong to the GSL are also Graeco-Latin words. This provides yet another reason for including GSL words into a productively-oriented academic word list. As Corson explains,

> Graeco-Latin words in English tend to be opaque, even for most L1 language users. For ESL users, they tend to be opaque if the learners have had no experience with their etymology when learning English or came from a language background greatly removed structurally from Latin and Greek. These words also have a very low frequency of use in most people's everyday discourse. In summary, the attributes of Graeco-Latin word difficulty are as follows: They are usually non-concrete, low in imagery, low in frequency, and semantically opaque. When these features combine in words, they interfere with word use and with word learning (Corson 1997:696)

5.7. Corpus findings and pedagogical relevance

An important application of word frequency lists for course design is "in deriving intuitions about functional and notional areas which might be important for the syllabus" (Flowerdew 1993:237). The automatic semantic analysis of our *Academic Keyword List* (AKL) has shown that a large proportion of the words included belong to the categories of 'general and abstract terms', 'numbers and measurement', 'language and communication', 'social actions, states and processes', 'psychological actions, states and processes' and 'names and grammar'. Several sub-categories which could be described as essentially **'functional' or 'procedural'** (cf. section 1.3.2.2) in nature are particularly well-represented (e.g. 'A2.2: affect: cause – connected'; 'A5: evaluation', 'A6: comparing'; 'Q2.2: speech acts'; 'X2.1: thought and belief').[131]

The functional syllabus has a long tradition in English language teaching (cf. Wilkins 1976; Weissberg and Buker 1978). Jordan (1997:165) reports that most textbooks following a product approach to academic writing published in Britain in the 80s and 90s are organized according to language functions such as explanation, definition, exemplification, classification, cause and effect, and comparison and contrast (e.g. Jordan 1990). However, they have rarely been based on principled selection criteria, relying instead on materials writers' perceptions of good practice in academic writing.

Unlike textbooks adopting a functional approach, courses which use vocabulary as the unit of progression, introduce new vocabulary according to principles such as frequency and range of occurrence. Nation explains that "[s]uch courses generally combine a 'series' and a 'field' approach to selection and sequencing. In a series approach, the items in a course are ordered according to a principle such as frequency of occurrence, complexity or communicative need. In a field approach, a group of items is chosen and the course covers them in any order that is convenient, eventually checking that all the items are adequately covered. Courses which use vocabulary as the unit of progression tend to break vocabulary lists into manageable fields, (…), according to frequency, which are then covered in an opportunistic way" (Nation 2001:386). It is interesting to note that most pedagogical applications of the *Academic Word List* to date have adopted this particular approach, using the frequency-based AWL sub-lists as fields (e.g. Obenda 2004, Huntley 2006).

[131] By contrast, categories with very small frequencies often represent topic-dependent semantic classes such as 'the body and the individual' or 'world and environment'.

There is a need for teaching materials which would merge the two types of syllabus design, thus adopting a 'functional-product' approach (cf. Jordan 1997:165) to academic writing while introducing new vocabulary according to principled criteria such as frequency and range of occurrence. This is precisely where our *Academic Keyword List* has a role to play. As Swales argues, however, one of the most pressing issues related to English for Academic Purposes "concerns how to make effective and efficient use of a specialized corpus (...) in order to gain pedagogically utilisable insight into the discourses that have been collected" (Swales 2002:151). **The *Academic Keyword List* is not a final product and does not in itself carry any guarantee of pedagogical relevance**. It still needs 'pedagogic mediation' (cf. Widdowson 2003) and is thus better conceived of as a "*platform* from which to launch corpus-based pedagogical enterprises" (Swales 2002:151).

The *Academic Keyword List* needs validation and may require adjustment if it is to inform a functional syllabus for academic writing. Once well-represented rhetorical or organizational functions have been identified, it is necessary to check whether all relevant lexical items are included in the AKL. We made use of a number of teaching materials, and more specifically, textbooks, and looked for all the lexical items that are commonly listed as serving specific functions. For example, the AKL includes a number of words and phrasemes that are commonly used as exemplifiers: the word-like units *for example* and *for instance*, the noun *example*, the verbs *illustrate* and *exemplify*, the preposition *such as* and the adverbs *notably* and *e.g.* Other lexical items listed in textbooks and EAP/EFL materials but which are not found in the AKL are the expressions *by way of illustration* and *to name but a few*, the nouns *illustration* and *a case in point* and the preposition *like*. These lexical items will also be examined for two main reasons. First, their analysis will make it possible to assess whether they should be added to the AKL and described in a functional syllabus for academic writing. Second, their inclusion in the description of a specific function in academic writing will allow us to approximate as closely as possible to what Hoffmann (2004:190) referred to as **conceptual frequency** so that the frequency of each exemplificatory lexical item can be calculated as a proportion of the total number of exemplifiers. As Wray states in her book on formulaic language,

> To capture the extent to which a word string is the preferred way of expressing a given idea (for this is at the heart of how prefabrication is claimed to affect the selection of a message form), we need to know not only how often that form can be found in the sample, but also how often it *could* have occurred. In other words, we need a way to calculate the occurrences of a particular message form as a proportion of the total number of attempts to express that message (Wray 2002:30).

Such an approach may help us move towards "understanding the intersection of form and function" (Swales 2002:163) in academic prose.

Keywords were selected fully automatically. The fact that they are used to serve particular functions in academic discourse has "to be corroborated by concordancing" (Flowerdew 1993:237). Words such as the noun *illustration*, the verb *illustrate* and the preposition *like* are often employed as exemplifiers but can also be used to fulfil other functions or with a different meaning. The verb *illustrate*, for example, also means 'to put a picture in a book' and the preposition *like* is often used to compare:

5.3. *This is as unsatisfactory as reading about a picture which is not* ***illustrated.*** (BNC-AC-HUM)

5.4. *There is no shortage of writing about Pollock;* ***like*** *other star artists, he has an embarrassingly large number of apologists, but fortunately there are select bibliographies which can guide her to key publications.* (BNC-AC-HUM)

To achieve better results, homographs were isolated and only the 'functional' uses of the lexical items under study were further analyzed.

The *Academic Keyword List* is based on native corpora only, which has its limitations for an analysis of **learner writing**, especially if conceptual frequency is to be investigated. EFL learners may use other lexical devices than native writers to serve rhetorical functions. For example, they repeatedly use word-like units such as *in a nutshell, in brief* and *all in all* for summarizing and concluding, which are quite rare in academic prose. A corpus-driven method such as the one used by De Cock (2003) (see section 4.2.2) was therefore adopted to identify words and word sequences that EFL learners use to serve rhetorical functions. Examples of overused words with functional uses in learner writing which do not belong to the AKL include *like, thing, say, let, I, really, firstly, secondly, thirdly, opinion, maybe, say, sure, but, thanks, always, so* and *why*. Learner-specific word sequences are discussed in section 6.3.2.3.1.

5.8. Limitation of the method used

A limitation inherent to the keyness approach adopted in this thesis to design the *Academic Keyword List* is its use of a reference corpus. A reference corpus is characterized by a set of distinctive linguistic features, among which some that may be shared with the corpus under study and which will thus not be recognized as prominent characteristics. There is thus a strong case for using "strongly contrasting reference corpora" (cf. Tribble 2001:396) such as a corpus of academic writing vs. one of fiction. However, it is quite probable that a few prominent lexical items in academic writing may have passed unnoticed because they are also used in fiction, irrespective of differences in meaning or function.

5.9. Conclusion

Recent EAP materials designers have mistakenly taken for granted that Coxhead's (2000) *Academic Word List* can be used as a vocabulary syllabus in academic writing. Vocabulary courses such as Obenda (2004), Schmitt and Schmitt (2005) and Huntley (2006) focus on productive uses of words that belong to the AWL, and more specifically, on their collocations and patterns of use. In this chapter, we challenge this assumption and propose a new methodology for the identification of lexical items that should be part and parcel of a productively-oriented academic word list. The methodology makes use of the criteria of keyness, range and evenness of distribution and provides a good illustration of the usefulness of annotation for the development of practical applications such as our *Academic Keyword List.* It is shown that while the word form + POS-tag format offers invaluable information about how words are used in academic discourse and which word forms are preferred, the lemma + POS-tag format gives better results for the selection of EAP words.

One important feature of the methodology used is that it does not disregard the 2,000 most frequent words of English. As a result, it has made it possible to appreciate the paramount importance of general service words in academic prose. They clearly account for a sizeable proportion of the sub-technical, and more precisely, procedural vocabulary that EFL learners need to master in order to write appropriate and effective academic texts. They include some of the most frequent items used to serve organizational or rhetorical functions in academic prose, e.g. *also, although, argue, because, cause, clear, compare, comparison, describe, description, differ, difference, discuss, effect, example, explain, introduction, likelihood, likely, support, suggest* and *view.*

By questioning the well-established distinction between General Service words and EAP words, this chapter also challenges the underlying assumption that it is indispensable to learn the 2,000 most frequent words in English before studying EAP words. Some General Service words are arguably not very useful for reading and writing academic texts, e.g. *baggage, game, club, garage*, etc. Ward (1999:310) denounced the "inherent contradiction in using a general list for learners with specific purposes" and developed a list of vocabulary for

EAP engineering students on the basis of the criterion of frequency only. The 930 lemmas that constitute our *Academic Keyword List* could be used as a basic vocabulary syllabus **in an EAP course** instead of the addition of the 2,000 words of the *General Service List* and the 570 word families of the *Academic Word List*.

6. EAP vocabulary in native and learner writing

6.1. Introduction

It was shown in chapter 1 that learner corpus research has often focused on the writing of one learner population characterized by a shared mother tongue background. Several studies have compared an L1 learner population to one or two learner populations in order to distinguish features that are specific to the L1 learner population under study from those characteristics that are shared by learners from different mother tongue backgrounds and therefore more likely to be developmental or teaching induced. Very few studies have investigated the use of linguistic features in more than three L1 learner corpora. The few studies that have undertaken such a task have often been more quantitative than qualitative in nature, largely because of the amount of data examined.

In this chapter, we analyze words and phrasemes that EFL learners use to serve typical **organizational or rhetorical functions** and compare these lexical items with those found in professional academic writing. In chapter 1, it was shown that EAP students often fail to recognize and appropriately use the conventions and linguistic features of academic prose. The working hypothesis of this chapter will thus be that **upper-intermediate to advanced EFL learners, irrespective of their mother tongue backgrounds, share a number of linguistic features that characterize their academic writing**. The learner corpus used, henceforth ICLE, consists of essays written by EFL learners from 10 different mother tongue backgrounds – Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, Swedish (see section 4.1.1). Quantitative data and statistics will be given for ICLE as an **aggregate population**. Unless specified otherwise, however, only linguistic features **shared by at least half of the learner populations** will be reported and discussed in this chapter. Linguistic features that are specific to one L1 learner population will be the focus of chapter 7.

The initial aim of this thesis was to focus on EFL learners' use of EAP vocabulary. For lack of detailed descriptions of the phenomenon in native academic prose, however, it was necessary to start with thorough analyses in native academic writing. Section 6.2 deals with EAP vocabulary in academic writing and focuses on the types of lexical devices used by expert writers to serve rhetorical or organizational functions. Section 6.3 is devoted to EAP vocabulary in learner writing. It first presents a detailed comparison of exemplificatory devices in native and learner writing which aims to illustrate the type of data and results obtained when comparing the range of lexical strategies available to EFL learners to that of expert writers. I compared eight functions in native and learner writing within the framework of a collaboration between the Centre for English Corpus Linguistics and Macmillan Education on the second edition of the *Macmillan English Dictionary for Advanced Learners* (cf. Gilquin et al. 2007). The numerous analyses conducted are not presented in detail in this thesis as their reading would soon have become cumbersome. A synthesis of quantitative results for 5 functions – **exemplification**, **expressing cause and effect**, **comparing and contrasting**, **expressing a concession** and **reformulating** - is given in Appendices 6.1a to 6.1e. Instead, the focus is placed on the general interlanguage features that emerge from these analyses, which fall into six categories: aspects of overuse and underuse, register-awareness, phraseological patterns, semantic misuse, clusters of connectives and sentence position of connectors.

Our findings have important pedagogical implications which are examined in Section 6.4. Two key aspects are discussed: the influence of teaching practices on learner writing and the role of corpora, and more specifically, learner corpora in the development of EAP teaching materials. The section ends with an example of a writing section that I have developed to help EFL learners master the complex function of comparison and contrast.

6.2. EAP vocabulary in academic writing

> I am concerned to establish a methodology that concentrates in the first place on recurrent events rather than on unrepeated patterns. When the habitual usages of the majority of users are thoroughly described, we will have a sound base from which to approach the singularities, which may of course include much fine writing. (Sinclair 1999b:18)

This section deals with EAP vocabulary in academic writing. Section 6.2.1 presents a detailed analysis of exemplificatory devices in native writing which serves as an illustration of the type of data and results obtained when examining the whole range of lexical strategies available to expert writers when they establish cohesive links in their texts. It is impossible to describe in similar detail the other four functions that were analyzed in native writing so as to provide a basis for comparison to EFL learner writing. Section 6.2.2 briefly comments on the types of lexical devices used by expert writers to serve the functions of expressing cause and effect, comparing and contrasting, expressing a concession and reformulating in an attempt to characterize the phraseology of rhetorical functions in academic prose.

6.2.1. Exemplification in professional academic writing

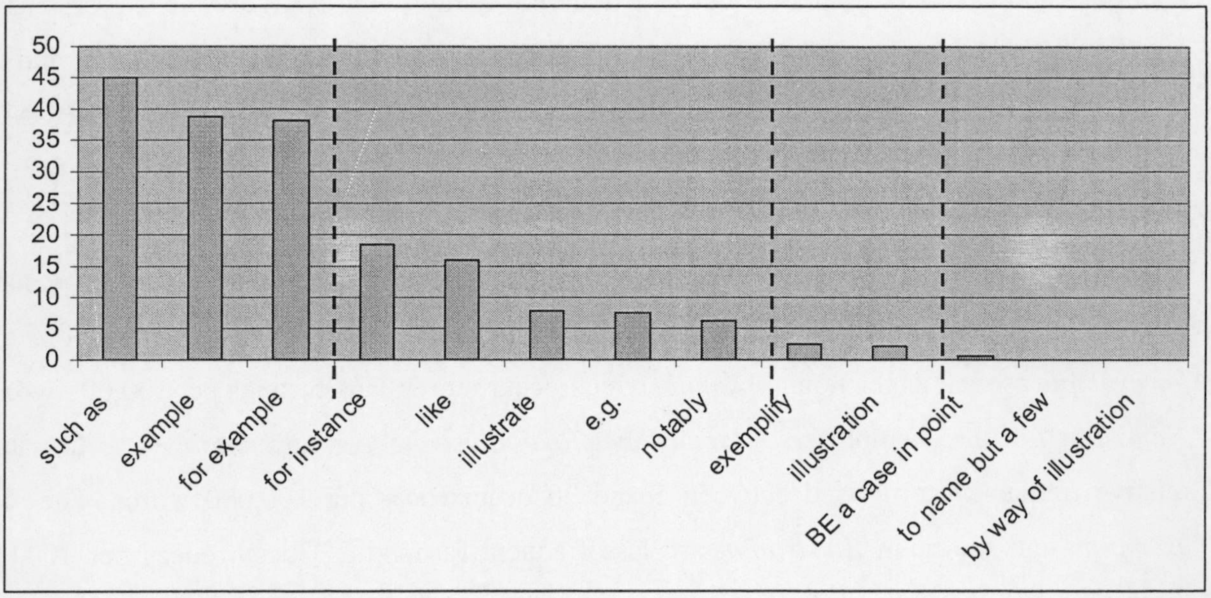
Siepmann (2005) shows that exemplificatory discourse markers occur in any kind of discursive prose and are particularly frequent in humanities texts. The author argues, however, that as an object of study, "exemplification continues to be the poor relation of other rhetorical devices" and that "[s]uch neglect has led to a commonly held view in both linguistic and the pedagogic literature that exemplification is a minor textual operation, subordinate to major discoursal stratagems such as 'inferring' and 'proving'" (ibid 111). Similarly, Coltier (1988) states that examples and exemplification merit close investigation at two levels: the exemplificatory strategies adopted (i.e. when and why are examples introduced in a text) and the wording of the example (i.e. the choice of the exemplifiers used). This section deals with the latter and focuses on the lexical items used by expert writers to give an example. For a rhetorical perspective on exemplifiers in native writing, see Siepmann (2005:112-118).

The *Academic Keyword List* (AKL) includes a number of words and phrasemes that are commonly used as exemplificatory discourse markers: the mono-lexemic or word-like units *for example* and *for instance*, the noun *example*, the verbs *illustrate* and *exemplify*, the preposition *such as* and the adverbs *notably* and *e.g.* Other lexical items commonly listed in textbooks and EAP/EFL materials but which are not found in the AKL are the expressions *by way of illustration* and *to name but a few*, the nouns *illustration* and *a case in point* and the preposition *like*. Table 6.1 gives the absolute frequencies of these words in BNC-AC-HUM as well as their relative frequencies per 100,000 words and the percentage of exemplificatory discourse markers they represent. In the bar chart (cf. Figure 6.1), the lexical items are ordered by decreasing relative frequency in the academic corpus. The most frequent exemplifiers in professional academic writing are the mono-lexemic phrasemes *such as* and *for example* as well as the noun *example*, which occur more than 35 times per 100,000 words. Almost half of the exemplifiers – *for instance, like, illustrate, e.g.* and *notably* - occur with a relative frequency comprised between 5 and 20 occurrences per 100,000 words. The verb *exemplify* and the noun *illustration* are less frequent (around 2.3 occurrences per 100,000 words) while the complex adverbs *to name but a few* and *by way of illustration* as well as the noun *case in point* are very rarely used in BNC-AC-HUM.

Table 6.1: Exemplification in BNC-AC-HUM

	Abs. freq.	**%**	**Rel. freq.**
nouns			
example	1285	21.6%	38.68
illustration	77	1.3%	2.3
(BE) a case in point	18	0.3%	0.5
TOTAL NOUNS	***1380***	***23.2%***	***41.5***
verbs			
illustrate	259	4.35%	7.8
exemplify	79	1.3%	2.38
TOTAL VERBS	***338***	***5.7%***	***10.2***
prepositions			
such as	1494	25%	45
like	532	8.9%	16
TOTAL PREP.	***2026***	***34%***	***61***
adverbs			
for example	1263	21.2%	38
for instance	609	10.2%	18.33
e.g.	259	4.35%	7.8
notably	77	1.3%	2.32
to name but a few	4	0.06%	0.12
by way of illustration	3	0.05%	0.09
TOTAL ADVERBS	***2215***	***37.2%***	***66.7***
TOTAL	**5959**	**100%**	**179.4**

Figure 6.1: Exemplification in BNC-AC-HUM



Section 6.2.1.1 discusses major findings about prepositions, adverbs and adverbial phrases which are used to serve an exemplificatory function and section 6.2.1.2 focuses on the exemplificatory use of nouns and verbs.

6.2.1.1. Using prepositions, adverbs and adverbial phrases

As shown in Figure 6.1, the complex preposition *such as* is the most frequent exemplifier in BNC-AC-HUM (example 6.1). Unlike in other genres such as speech and fiction (cf. section 6.2.2, Figure 6.7), it is much more frequent than the preposition *like* (example 6.2) in professional academic writing. Similarly, *for example* is twice as frequent as *for instance*. These two complex adverbs are commonly classified as 'code glosses' in metadiscourse theory (cf. section 2.5), as they are used to "supply additional information, by rephrasing, explaining, or elaborating what has been said, to ensure the reader is able to recover the writer's intended meaning" (Hyland 2005:52). Code glosses are 'interactive resources' in Hyland's typology, i.e. features used to "organize propositional information in ways that a projected target audience is likely to find coherent and convincing" (ibid 50). They are categorized as **textual phrasemes** in this thesis (see section 2.3.7) as they are recognized as mono-lexemic units with an organizational – exemplificatory - function.

6.1. *This is the arrangement in Holland whereby various institutions* ***such as*** *media, schools, cultural organizations, welfare services, and hospitals are duplicated, and run by the separate catholic and protestant communities.* (BNC-AC-HUM)

6.2. *Surrealist painting had publicity value, especially when executed by a showman* ***like*** *Salvador Dali, who married the former wife of the poet Paul Éluard.* (BNC-AC-HUM)

In the BNC-AC-HUM, *for example* and *for instance* are typically used within the sentence, enclosed by commas, especially after the subject of the sentence. They mainly function as advance labels (cf. section 1.3.2.2) to introduce a following example (cf. Table 6.2):

6.3. *A book on a single painter,* ***for example,*** *is a monograph.* (BNC-AC-HUM)

Table 6.2: 'for example' and 'for instance' in BNC-AC-HUM: advance vs. retrospective label

	Advance label	**Retrospective label**
for example	1185 (93.8%)	78 (6.2%)
for instance	588 (96.5%)	21 (3.5%)

Examples 6.4 and 6.5 show that *for example* and *for instance* can follow the subject of the exemplifying sentence while remaining essentially cataphoric in nature.

6.4. *Such associations of sexual deviance and political threat have a long history sedimented into our language and culture. The term "buggery",* ***for example,*** *derives from the religious as well as sexual nonconformity of an eleventh-century Bulgarian sect which* *practised the Manichaean heresy and refused to propagate the species; the OED tells us that it was later applied to other heretics, to whom abominable practices were also ascribed.* (BNC-AC-HUM)

6.5. *The small mammals living today in many different habitats and climatic zones have been described, so that the associations between faunal types and ecology are well documented [...]. Woodland faunas, **for instance,** are distinct from grassland faunas, and tropical faunas distinct from temperate faunas, and when these and more precise distinctions are made it is possible to correlate and even define ecological zones by their small mammal faunas.* (BNC-AC-HUM)

The complex adverbs *for example* and *for instance* can also function as retrospective labels and refer back to an example given before as illustrated in example 6.6. This use is, however, much less frequent.

6.6. *Thirdly, the debates over how far to forge a strategy either for winning power or for promoting economic development in a post-revolutionary society have not been satisfactorily resolved, and indeed perhaps cannot be, given that counter-revolutionary response to any successful formula will ensure that it will be that much more difficult to apply the same tactics in another situation. Such is the relation which Nicaragua bears to El Salvador, **for example.*** (BNC-AC-HUM)

In *Mieux écrire en anglais*, Laruelle (2004:96-97) writes that *for example* should be placed in sentence-initial position if the whole sentence has an exemplificatory function while the complex adverb should follow the subject, between commas, if the subject only is the example. This statement, however, is not confirmed by corpus data. Sentence 6.4 clearly shows that *for example* need not be placed in sentence-initial position to introduce an exemplificatory sentence.

Like nouns and verbs, mono-lexemic adverbial phrases can also have their own **phraseological patterns**. Three verbs, i.e. *consider* (f[n,c][132] = 13; log-likelihood = 92.5), *take* (f[n,c] = 7; log-likelihood = 19.1) and *see* (f[n,c] = 19; log-likelihood = 71.7) are significant left co-occurrents of *for example* in BNC-AC-HUM. They are used in the second person of the imperative. The verbs *consider* and *take* are typically used with *for example* to introduce an example that is discussed in further detail over several sentences:

[132] f[n,c] = f: frequency; n: node; c: co-occurrent

6.7. *It is worth pausing here momentarily to observe that such legally provided remedies can be morally justified even when applied to people who are not subject to the authority of the government and its laws.* ***Consider for example*** *the law of defamation. Assuming that it is what it should be, it does no more than incorporate into law a moral right existing independently of the law. The duty to compensate the defamed person is itself a moral duty. Enforcing such a duty against a person who refuses to pay damages is morally justified because it implements the moral rights of the defamed. One need not invoke the authority of the law over the defamer to justify such action. The law may not have authority over him.* (BNC-AC-HUM)

6.8. *But the concept of compresence is far from clear. If it implies that no time-lag is detectable between elements of an experienced" complex", then this is true only in a very limited sense.* ***Take, for example,*** *the perceptual experience that I have while looking at this bunch of carnations arranged in a vase on the table in the middle of the room. I see this" complex" as one whole. But while I am looking at it my eyes constantly wander from one flower to the next, pausing at some, ignoring others, picking out the details of their shapes and colours. Finally, without taking my eyes off the flowers, I may move the vase closer, or walk around the table and look at the flowers from different angles. The scene will keep constantly changing. As a result, I shall experience a succession of different" complexes of qualities" but I shall still be looking at the same bunch of flowers.* (BNC-AC-HUM)

Hyland (2002c:217) describes this type of imperatives as directives with a rhetorical purpose that "can steer readers to certain *cognitive acts*, where readers are initiated into a new domain of argument, led through a line of reasoning, or directed to understand a point in a certain way". He categorizes them as **interactional resources**, and more specifically as **engagement markers**, i.e. "devices that explicitly address readers, either to focus their attention or include them as discourse participants" (Hyland 2005:53).

The verb *see* is frequently used in professional academic writing as an **endophoric marker** to refer to tables, figures, or other sections of the article or to someone else's ideas or publications (cf. Hyland 1998, 2002c, 2005; Hyland and Tse 2006). The use of 2nd person imperative *see* "allow[s] academic writers to guide readers to some *textual act*, referring them to another part of the text [*internal reference*] or to another text [*external reference*]" (Hyland 2002c:217). In BNC-AC-HUM, 63% of the occurrences of the sequence *see for example* appear between brackets as in the following example:

6.9. *Afro-Caribbean and Asian children are indeed painfully aware that many teachers view them negatively and some studies have documented reports of routine racist remarks by teachers (**see for example** Wright in this volume).* (BNC-AC-HUM)

Swales et al. (1998) examined a corpus of research articles in 10 disciplines, namely art history, chemical engineering, communication studies, experimental geology, history, linguistics, literary criticism, philosophy, political science and statistics, and found that 2nd person imperative ***see*** was the most frequent imperative form across disciplines. Similarly, in his study of directives in academic writing, Hyland (2002c) analysed a corpus of 240 published research articles, 7 textbook chapters and 64 project reports written by final year Hong Kong undergraduates and found that 2nd person imperative ***see*** represented 45% of all imperatives in his corpus. Note that in both studies, variation in the use of 2nd person imperative ***see*** was found across disciplines.

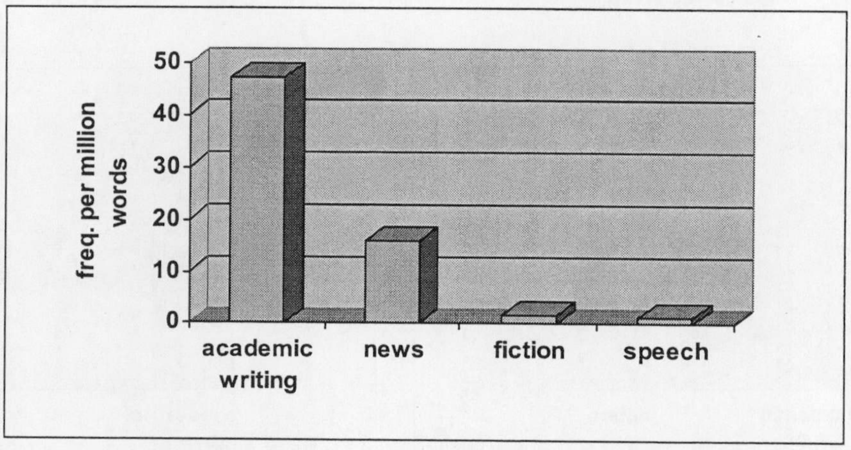
The sequences *take/consider/see for example* are regarded as **phrasemes** in this thesis as they form functional – textual – units and display arbitrary lexical restrictions. The **advantage of adopting a phraseological approach to metadiscourse resources** appears quite clearly here. The sequences *take/consider for example* consist of two metadiscourse resources in Hyland's categorisation scheme: the imperative forms *take* and *consider* are interactional resources and more specifically engagement markers while *for example* is a code gloss. Similarly, *see* is an endophoric marker in *see for example.* In our phraseological framework, the three sequences are categorized as **textual phrasemes**.

The adverb *notably* can be regarded as a typical EAP word: Figure 6.2 shows that it is much more frequent in academic writing than in any other genres. It is typically preceded by a comma (ex. 6.10) and is qualified by the adverb *most* (ex. 6.11) in 15.2% of its occurrences in BNC-AC-HUM.

6.10. *Some bishops, **notably** Jenkins of Durham, Sheppard of Liverpool, and Hapgood of York, have spoken out about deprivation in the inner cities, the miners ' strike, and the need for government to show a greater compassion for, and understanding of, the poor.* (BNC-AC-HUM)

6.11. *At leading public schools, **most notably** Eton, there is a tradition of providing MPs, government ministers, and prime ministers.* (BNC-AC-HUM)

Figure 6.2: Distribution of the adverb 'notably' across genres



The abbreviation *e.g.* (or less frequently *eg*) stands for the Latin form 'exempli gratia' and means the same as *for example*. It is quite common in the BNC-AC-HUM, in which 65.7% of its occurrences are used between brackets:

6.12. *Direct curative measures* ***(e.g.*** *flood protection) clearly are within the domain of a soil conservation policy.* (BNC-AC-HUM)

In contrast to *for example* and *for instance*, the great majority of occurrences of *e.g.* typically introduce one or more noun phrases rather than full clauses:

6.13. *It may help to refer the patient to other agencies* ***(e.g.*** *social services, a psychosexual problems clinic, self-help groups).*

When *e.g.* is used without brackets, it is preceded by a comma:

6.14. *Primary industries are those which produce things directly from the ground, the water, or the air,* ***e.g.*** *farming.*

As shown in Figure 6.1, the **textual phrasemes** *by way of illustration* and *to name but a few* are quite rare in BNC-AC-HUM. In fact, these expressions are generally very infrequent in all types of discourse. Figure 6.3 and 6.4 show the distribution of the two phrasemes in four main genres of the *British National Corpus* (BNC), namely academic writing, fiction, newspaper texts and speech. 36% (10 occurrences out of a total of 28) of the occurrences of *by way of illustration* in the BNC appear in academic texts and only one occurrence is found in speech. The expression *to name but a few* is more frequent than *by way of illustration* in the whole BNC but only 12.8% of its occurrences (10 out of 78 occurrences) appear in academic texts. No instance of *to name but a few* is found in speech.

Figure 6.3: Distribution of 'by way of illustration' across genres

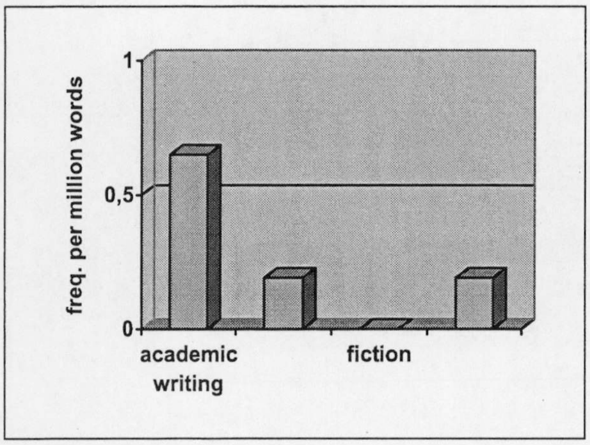
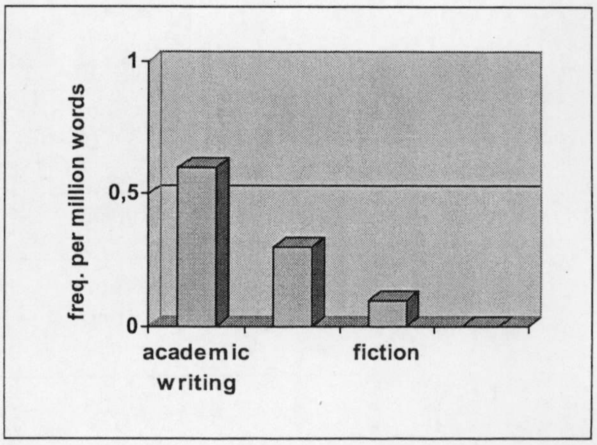


Figure 6.4: Distribution of 'to name but a few' across genres



6.2.1.2. Using nouns and verbs

Nouns and verbs are also used to give examples in specific phraseological patterns. The most frequent noun is *example*, which is much more common than *illustration* and *a case in point*. Table 6.3 shows that it is as frequently used as its connective counterpart, i.e. textual phraseme *for example*, in BNC-AC-HUM.

Table 6.3: 'example' and 'for example' in BNC-AC-HUM

	example		*for example*	
	Absolute freq.	**%**	**Absolute freq.**	**%**
BNC-AC-HUM	1285	50.43%	1263	49.57%

Table 6.4 shows the significant verb co-occurrents of the noun *example* in BNC-AC-HUM. The verb *be* is the most frequent verb co-occurrent of *example* both in a 3L-1L window and 1R-3R window. It is, however, twice as frequent in the left window. When *example* is preceded by the verb *be*, it mainly functions as a retrospective label, i.e. it refers back to the exemplifying element which is given in subject position. The noun *example* may directly refer back to a noun phrase as in example 6.15 or to the demonstrative pronoun *this* which further points to a previous exemplifying sentence (underlined in example 6.16).

6.15. *Vision* ***is a better example of*** *a modular processing system.* (BNC-AC-HUM)

6.16. *The designer at Olympia chose to represent the race by the moment before it started, as Polygnotos showed the sack of Troy in its aftermath.* ***This is the supreme surviving example of*** *the early classical taste for stillness and indirect narrative.* (BNC-AC-HUM)

By contrast, when introduced by *there* + *BE* (11%) or *here* + *BE* (15%), the noun *example* functions as an advance label which refers forward to a following example (underlined):

6.17. *In addition, of course, choices can result from lengthy weighing of odds.* ***Here is a simple example of*** *the complexity at issue.* *<u>I am driving along a narrow main road, used by fast-moving traffic, with my children in the back seat. A car some distance ahead strikes a large dog but does not stop, leaving the creature walking-wounded but in obvious distress. My children, seeing what occurred, cry out. I glance in the rear-view mirror to see other cars close behind; slowing down but then speeding up again. I do not stop</u>.* (BNC-AC-HUM)

When *example* is the subject of the verb *BE*, it always functions as an advance label. It is often qualified by an adjective (cf. examples 6.18 to 6.20) and the exemplified item is generally introduced by the preposition *of* (examples 6.19 and 6.20). In example 6.20, the exemplified item is the pronoun *this* which refers back to the previous sentences.

6.18. ***The prime example is*** *<u>the Dada movement, whose nihilistic work is now admired for qualities of imagination</u>.* (BNC-AC-HUM)

6.19. ***The clearest example of*** *emotive language* ***is*** *<u>poetry</u>, which is entirely concerned with the evocation of feelings or attitudes, and in which the writer's and reader's attention is not, or should not be, directed at any of the objective relationships between words and things.* (BNC-AC-HUM)

6.20. *Until the seventeenth century many, even most, European frontiers were very vague, zones in which the claims and jurisdictions of different rulers and their subjects overlapped and intersected in a complex and confusing way. This was especially true in eastern Europe, where many states were larger and central governments usually less effective at the peripheries of their territories than in the west.* ***The most striking example of this is*** *perhaps <u>the frontier in the Danubian plain between the Ottoman empire and the Habsburg territories in central Europe</u>.* (BNC-AC-HUM)

Copular clauses with the noun *example* consist of **sentence stems** (*An example of* Y *is* ...) and **rhemes** (... *is an example of* Y). They are classified as **textual phrasemes** in our typology as they fulfil a clear exemplificatory function.

Table 6.4: Significant verb co-occurrents of the noun 'example' in BNC-AC-HUM

Left co-occurrents		Right co-occurrents	
Verb	**freq.**	**Verb**	**freq.**
be	*139*	*be*	*84*
provide	*26*	*illustrate*	*14*
take	29	*show*	*21*
give	*12*	*give*	*15*
cite	5	suggest	12
consider	12	quote	6
illustrate	*7*	include	7
show	*9*	*provide*	*8*
see	10	concern	6
serve	5		
		will	16
		can	15
		would	13

Four other verbs, namely *provide, give, illustrate* and *show* (in italics in Table 6.4), are significant left and right co-occurrents of the noun *example*. The verbs *take, cite, consider, see* and *serve* are only significant left co-occurrents while the verbs *suggest, concern, quote* and *include* and the modals *will, can* and *would* are right co-occurrents. The verbs *provide, take, give, cite, consider, see, serve* and *include* often co-occur with the noun *example* to form **textual – exemplificatory - phrasemes**. The verb *provide* can be used in active or passive structures but active structures in which the subject is the example are more frequent (example 6.21). The verb *cite* is more often found in a passive structure in which *example* functions as a retrospective label (example 6.22). The two verbs often form **rhemes** with the noun *example*. .

6.21. *The Magdalen College affair, for example,* ***provides a classic example of*** *passive resistance.* (BNC-AC-HUM)

6.22. *A famous passage of art criticism* ***can be cited as one example*** *entirely beyond dispute.* (BNC-AC-HUM)

The verb *take* is mainly used in sentence-initial exemplificatory infinitive clauses with the noun *example* (68.9%; ex. 6.23). It also occurs in active structures with a personal pronoun subject (13.79%; ex. 6.24) and in imperative sentences (13.79%). When used in the imperative, it is generally used in the 2nd person (ex. 6.25) and only appears once in the 1st person plural (ex. 6.26). By contrast, the verb *consider* is mainly used in imperative sentences with the noun *example* (70%), i.e. 2nd person imperatives (ex. 6.27) and less frequently 1st person plural imperatives (ex. 6.28). The verb *see* is always used in the 2nd person of the imperative in co-occurrences with the noun *example* ex. 6.29). It is not used to introduce an example but as an endophoric marker to direct the reader's attention to an example situated elsewhere in the text (cf. Hyland's typology of metadiscourse markers in section 2.5).

6.23. ***To take one example,*** *at the beginning of the project seven committees were established, each consisting of about six people, to investigate one of a range of competing architectural possibilities.* (BNC-AC-HUM)

6.24. *In accordance with the theme of this chapter, I shall simply use "stylistics" as a convenient label (hence the inverted commas) for the branch of literary studies that concentrates on the linguistic form of texts, and I shall* ***take*** *four different* ***examples*** *of this kind of work as alternatives to the Prague School's and Jakobson's approach to the relationship between linguistics and literature.* (BNC-AC-HUM)

6.25. ***Take the example of*** *following an object by eye-movements (so-called "tracking").* (BNC-AC-HUM)

6.26. *By way of illustration,* ***let us take an example*** *from the development of Newton's theory that we have considered several times before, and consider the situation that confronted Leverrier and Adams when they addressed themselves to the troublesome orbit of the planet Uranus.* (BNC-AC-HUM)

6.27. ***Consider the following example.*** (BNC-AC-HUM)

6.28. *Some are evidently more equal than others in this system of justice;* ***let us consider another example.*** (BNC-AC-HUM)

6.29. *The most important vowel is set to two or more tied notes in a phrase designed to increase the lyrical expression (****see Example*** *47, above).* (BNC-AC-HUM)

The verb *include* is used with the plural form of the noun *example* in subject position to introduce an incomplete list of examples in object position:

6.30. *The floral* ***examples include*** *a large lotus calyx and two ivy leaves joined by a slight fillet.* (BNC-AC-HUM)

Among the verb co-occurrents of *example*, another set of verbs is used to discuss examples given in a text, e.g. *quote* (example 6.31); to talk about conclusions that can be drawn from these, e.g. *show* and *suggest* (examples 6.32 and 6.33) or to show what something is like or that something is true, e.g. *illustrate* (example 6.34).

6.31. *Thirdly, in all the* ***examples quoted here,*** *there is a sense in which all observers see the same thing.* (BNC-AC-HUM)

6.32. *The **example shows** that the objector's neat distinction between adjudicative and legislative authorities is mistaken.* (BNC-AC-HUM)

6.33. *As these **examples suggest**, the pontificate of Leo XIII (1879–1903) saw something of a recovery in the political fortunes of the Papacy.*

6.34. ***This example clearly illustrates** the theory dependence and hence fallibility of observation statements.* (BNC-AC-HUM)

The significant co-occurrences illustrated in sentences 6.31 to 6.34 do not qualify as restricted collocations as the meaning of the verbs is not restricted by the noun *example* and their combinations are fully explainable in semantic and syntactic terms. However, these co-occurrences are used in adverbial clauses (e.g. *as this example suggests* ...) and sentence stems (e.g. *this example* [adv.] *illustrates* ...) which serve to **describe** or **explain** examples previously given and **make suggestions** on their basis. These three actions appear in Pecman's (2004) ontology of General Scientific Language (cf. section 1.4.1) and are key EAP functions. The adverbial clauses and sentence stems used to fulfil these functions are thus classified as **textual phrasemes**.

The advantage of using the noun *example* rather than the complex adverbs *for example* or *for instance* to serve an exemplificatory function is that it allows the writer to evaluate the example in terms of its suitability, e.g. *good, outstanding, fine, excellent* (cf. example 6.35) or typicality, e.g. *classic, typical, prime* (cf. examples 6.36 to 6.38). The adjectives *above* and *following* are used to situate the example in the text (example 6.39): they are endophoric markers in metadiscourse terms (cf. section 2.5). Table 6.5 gives the 24 adjectives that significantly co-occur with the noun *example* in BNC-AC-HUM.

6.35. *An **outstanding example** of this type of narrative is Vargas Llosa 's Conversation in the Cathedral, which pivots around a four-hour conversation between two characters, the whole novel being made up of dialogue and narrative units generated in waves by the central conversation, as the two men's review of their past lives sparks off inner thoughts and recollections and conjures up other conversations and dramatized episodes.* (BNC-AC-HUM)

6.36. *The **prime example** is the Dada movement, whose nihilistic work is now admired for qualities of imagination.* (BNC-AC-HUM)

6.37. *Macmillan's personal meeting with Eisenhower in September is a **classic example** of how old friendships can have unfortunate consequences in diplomacy.* (BNC-AC-HUM)

6.38. ***Typical examples** are cases where one is given notice that everyone who enters a certain house, club, or park must abide by certain rules, obey a certain authority, or do so at his own risk.* (BNC-AC-HUM)

6.39. *Consider the **following example**.* (BNC-AC-HUM)

There is a case for considering the co-occurrence *classic example* as a nonce combination: the adjective *classic* is used with a meaning that is listed as its first sense in the LDOCE4 (*1.* TYPICAL: *having all the features that are typical or expected of a particular thing or situation*) and the Oxford English Dictionary Online[133] (*1. of the first class, of the highest rank or importance; approved as a model; standard, leading*). However, the adjective is only commonly used with a very limited number of nouns - *example, mistake and case*[134] - in this particular sense. This is clearly one more illustration of the difficulty of separating senses that a word has in isolation from those that it acquires in context (cf. section 2.3.3). In this thesis, co-occurrences of this type are classified as **collocations** on the basis that they display degrees of lexical restriction. The co-occurrence *prime* + *example* (see example 6.36 above) is a clearer example of a collocation: the adjective *prime* has two core meanings – 'most important' and 'of the very best quality or kin' – but a *prime example* is 'a very typical example of sth'. Collocations represent 8.3% of the types and 6.87% of the tokens of adjective + *example* co-occurrences.

Other adjectives form semantically and syntactically fully compositional sequences with the noun *example*. Thus, the meaning of an *outstanding example* is made up of the meaning of the adjective *outstanding* and the noun *example*. This does not mean, however, that they are uninteresting for pedagogical purposes. First, they constitute "preferred ways" of qualifying *example* as they are repeatedly used with the noun. Second, in her study of verb + noun combinations, Nesselhauf (2005) has shown that free or nonce combinations are also prone to erroneous or, at least, unidiomatic use in learner writing. Similarly, Lorenz (1998) has pointed out that German learners' use of adjectives, irrespective of their phraseological status, differs from that of native students (cf. section 1.4.2).

[133] http://www.oed.com/

[134] These three nouns are listed under the first sense of 'classic' in LDOCE4.

Table 6.5: Adjectives co-occurrents of the noun 'example' in BNC-AC-HUM

Adjective	freq.	Adjective	freq.
good	38	fine	9
above	15	notable	8
following	18	isolated	8
well-known	10	interesting	9
obvious	16	known	7
classic	11	excellent	6
typical	13	prime	7
outstanding	10	trivial	5
extreme	12	previous	6
clear	16	remarkable	5
simple	13	numerous	5
striking	9	single	6

The added value of using statistics, and more specifically association measures, to extract significant co-occurrents of a word in a large corpus is made clear by a comparison of the adjectives listed in Table 6.5 with those given in Siepmann (2005:137) (see section 1.4.1 for a discussion of Siepmann's methodology and the corpora used). In addition to most adjectives given in Table 6.5, Siepmann lists a number of adjectives that do not appear even once in the 87-million word written part of the BNC, e.g. *beguiling, consummate, eminent, apposite, anodyne, happy, alarming, crass, cautionary*, and adjectives which occur only once or twice in the corpus, e.g. *exquisite, well-worn, edifying, emotive, awe-inspiring, glittering, hideous*. To use Sinclair's (1999b) words, these co-occurrents are best described as "singularities" and do not represent "the habitual usages of the majority of users".

Apart from verbs and adjectives, other significant co-occurrents of the noun *example* are found in professional academic writing. Left co-occurrents mainly consist of determiners and the pronoun *this*. Indefinite determiners (*a, another* and *one*) are more frequent than the definite article *the* with *example*, which is mainly used when the noun is qualified by a superlative adjective (example 6.40) or preceded by ordinals such as *first, next* and *last* (example 6.41).

6.40. ***The best-known example of this**, and probably the most ambitious, was the political academy set up in Paris in 1712 by the Marquis de Torcy.*

6.41. ***The first two examples** discussed below illustrate different ways in which the linguistic model is used to develop a narrative model, and (...).*

The pronoun *this* is typically used in subject position of a sentence with the verb *be* to refer back to an example given in a previous sentence as illustrated in example 6.16 above. Right co-occurrents comprise the preposition *of* and the pronoun *this*. In 40% of its occurrences, the noun *example* is directly followed by the preposition *of* which introduces the idea, class or event exemplified. The idea, class or event exemplified is often determined by a demonstrative (ex. 6.35 above) or pronominalized to refer back to a previous sentence. These findings support Gledhill's (2000) view that **there may be a very specific phraseology and set of lexico-grammatical patterns for function words in academic discourse** (see section 1.4.1). Function words seem to display co-occurrence effects just as content words do (also see Renouf and Sinclair's (1991) notion of 'collocational framework' described in section 2.4.2.2.3). These findings also provide **strong evidence against the use of 'stop word' lists** when extracting co-occurrences from corpora as there is a serious danger of missing a whole set of phraseological patterns.

The verbs *illustrate* and *exemplify* can also be used as exemplifiers. The verb *illustrate* is used with the meaning of 'to be an example which shows that something is true or that a fact exists' (ex. 6.42) or 'to make the meaning of something clearer by giving examples' (ex. 6.43) (LDOCE4). The verb *exemplify* is used with the meanings of 'to be a very typical example of something' and 'to give an example of'.

> 6.42. *The narratives of the Passio Praeiecti and of the Vita Boniti both have their peculiarities, and it is possible that the appointment of Praeiectus and the retirement of Bonitus were less creditable than their hagiographers claim. Nevertheless they do* ***illustrate*** *the complexities of local ecclesiastical politics.* (BNC-AC-HUM)
>
> 6.43. *My aim will be to* ***illustrate*** *different ways of approaching literature through its linguistic form, ways involving the direct application of linguistic theory and linguistic methods of analysis in order to illuminate the specifically literary character of texts.* (BNC-AC-HUM)

Both verbs are more frequent in academic writing than in any other genre. Figure 6.5 compares the relative frequencies of the two verbs in academic writing with three main genres represented in the *British National Corpus*. The verb *illustrate* is not uncommon in news but a quick look at concordances shows that a significant proportion of its occurrences are not used to introduce an example but with the meaning of 'to put pictures in a book, article, etc' (see example 6.44). *Exemplify* is very rarely used in other genres.

6.44. *Also in the pipeline is an Australian children 's TV series based on Gumnut Factory Folk Tales (written,* ***illustrated*** *and published by Chris Trump, A$7.95).* (BNC-NEWS)

Figure 6.5: Distribution of the verbs 'illustrate' and 'exemplify' across genres

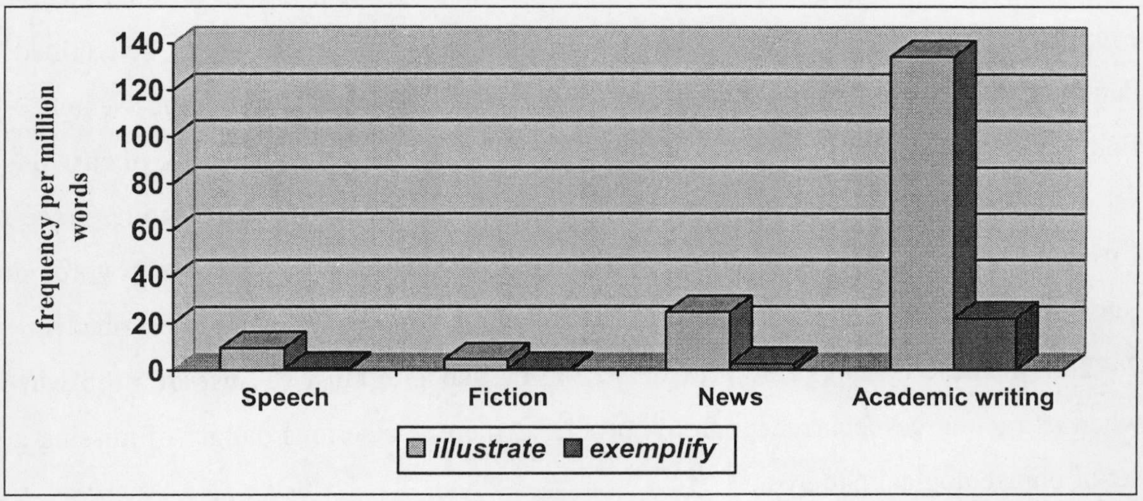


Figure 6.5 also shows that the verb *illustrate* is more frequent than *exemplify* in professional academic writing. Following Granger (2006), the frequencies of the two verb lemmas, their word forms and tenses are measured in BNC-AC-HUM[135]. Table 6.6 shows that there is no major difference in proportion between the verb forms *illustrate, illustrated* and *illustrates*. When used in active structures, the verb is often preceded by a non-human subject such as *example, figure, table, case* or *approach* (cf. example 6.45). Almost all occurrences of the past participle appear in the passive construction *BE illustrated by/in* (cf. example 6.46).

6.45. *This* ***example*** *clearly* ***illustrates*** *the theory dependence and hence fallibility of observation statements*. (BNC-AC-HUM)

6.46. *The contrast between the conditions on the coast and in the interior* ***is illustrated by*** *the climatic statistics for two stations less than 30 km (18.5 miles) apart.* (BNC-AC-HUM)

The sentence-initial adverbial clause *To illustrate this/the point/X, ...* (cf. example 6.47) represents 2.7% of the occurrences of the lemma *ILLUSTRATE* in BNC-AC-HUM.

6.47. *How many observations make up a large number? (...) Whatever the answer to such a question, examples can be produced that cast doubt on the invariable necessity for a large number of observations.* ***To illustrate this,*** *I refer to the strong public reaction against*

[135] Figures are based on disambiguated data. Thus, instances of *illustrate* used in the sense of 'to put pictures in a book, article, etc' are not added up.

nuclear warfare that followed the dropping of the first atomic bomb on Hiroshima towards the end of the Second World War. (BNC-AC-HUM)

In BNC-AC-HUM, *illustrate* significantly co-occurs with the noun *example* [LogL=112] in a 3L-1L window and with the nouns *point* [LogL=168.78], *example* [LogL=49.65] and *fig.* [LogL=45.08] in a 1R-3R window. The noun *point* is used as an object of *illustrate* which refers back to an idea put forward in a previous sentence:

6.48. *For most of this century it is those disorders gathered together under the heading of "schizophrenia" that have been used as the paradigm for trying to describe and understand psychosis. Yet even in this form, or forms -- for many would prefer to talk of "the schizophrenias" -- there is still no universally accepted set of criteria for diagnosis.* ***To illustrate the point,*** *one of the present authors was recently asked to review a paper submitted to a prominent psychiatric journal, proposing a new set of rules for diagnosing schizophrenia. In the course of their analysis the authors determined the extent to which their proposed criteria agreed with those contained in other existing diagnostic schemes -- some ten or twelve of them. Correlations varied over a very wide range.* (BNC-AC-HUM)

The noun *figure* (and the abbreviation *fig.*) is either used as the subject of the verb *illustrate* or in the passive structure *illustrated in figure x*. This co-occurrence is even more frequent in other academic genres such as social science, natural science and medicine which rely more on figures, tables and diagrams (see examples 6.49 and 6.50).

6.49. ***Figure 1 illustrates*** *the spread of results for the alcoholics and the controls.* (BNC-AC-MED)

6.50. *The advantages of the system are* ***illustrated in Fig.*** *8.2 and, like the Peruvian example discussed above, the fallow stage is contributing to crop productivity as well as providing protection against soil erosion.* (BNC-AC-SOC)

The adverbs *well, better, best,* and *clearly* are sometimes used with *illustrate* to evaluate the typicality or suitability of the example (example 6.51). The verb *illustrate* also co-occurs significantly with *how* to introduce a clause (example 6.52), with the verb *serve* (example 6.53) and the modals *will, can* and *may* (example 6.54).

6.51. *The history of the English monarchy* ***well illustrates*** *both the importance and the unimportance of war.* (BNC-AC-HUM)

6.52. *We recently did a simple experiment which happens to* ***illustrate how*** *children's knowledge of where an object is determines their behaviour.* (BNC-AC-HUM)

6.53. *While our discussion in this chapter is of the doctrine of neutrality as such, Rawls ' treatment of it will* ***serve to illustrate*** *the problems involved.* (BNC-AC-HUM)

6.54. *This prejudice against close involvement with " the secular government* ***may be illustrated by*** *an anecdote related in the about Molla Gurani.* (BNC-AC-HUM)

As shown in Table 6.7, the lexico-grammatical preferences of the verb *exemplify* differ widely from those of *illustrate*. A large proportion of the occurrences of the lemma *exemplify* are *–ed* forms, and more precisely past participle forms, of the verb. In BNC-AC-HUM, the verb significantly co-occurs with the verb *be* and the conjunction *as* in a 3L-1L window and with the prepositions *by* and *in* in a 1R-3R window. These significant co-occurrents highlight the preference of the verb for the passive structure *BE exemplified by/in* (cf. example 6.55) and the lexico-grammatical pattern *as exemplified by/in* (cf. example 6.56). *Exemplify* is also often used after a noun phrase, preceded by a comma (cf. ex. 6.57).

6.55. *The association of this material with the clerk* ***is clearly exemplified by*** *Chaucer's Wife of Bath's fifth husband, the clerk Jankyn, who, in the Wife of Bath's Prologue, reads antifeminist material to her from his book "Valerie and Theofraste".* (BNC-AC-HUM)

6.56. *He assumed, without argument, that science,* ***as exemplified by*** *physics, is superior to forms of knowledge that do not share its methodological characteristics.* (BNC-AC-HUM)

6.57. *Piaget's claim that thinking is a kind of internalized action,* ***exemplified*** *in the assimilation-accommodation theory of infant learning mentioned above, is really a global assumption in search of some refined, detailed and testable expression.* (BNC-AC-HUM)

Unlike *illustrate*, the verb *exemplify* does not co-occur significantly with nouns.

Table 6.6: 'illustrate' in BNC-AC-HUM

The lemma ILLUSTRATE	BNC-AC-HUM	
illustrate	**97**	**37.45%**
simple present	36	13.89%
infinitive	61	23.55%
illustrated	**84**	**32.43%**
simple past	7	2.7%
present/past perfect	0	0%
past participle	77	29.73%
illustrates	**63**	**24.32%**
illustrating	**15**	**5.79%**
continuous tense	2	0.77%
-ing clause	13	5%
Total	**259**	**100%**
Nr of words	**3,321,867**	
Relative freq. per 100,000 words	**7.8**	

Table 6.7: 'exemplify' in BNC-AC-HUM

The lemma EXEMPLIFY	BNC-AC-HUM	
exemplify	**9**	**11.4%**
simple present	5	6.33%
infinitive	4	5%
exemplified	**53**	**67%**
simple past	8	10%
present/past perfect	1	1.26%
past participle	44	55.7%
exemplifies	**15**	**19%**
exemplifying	**2**	**2.53%**
continuous tense	0	0%
-ing clause	2	2.53%
Total	**79**	**100%**
Nr of words	**3,321,867**	
Relative freq. per 100,000 words	**2.38**	

6.2.1.3. Discussion

The description of exemplifiers presented in this section does not aim at exhaustiveness but at **typicality** in professional academic writing. The corpus-based methodology adopted has highlighted a number of lexical items that are repeatedly used as exemplifiers in academic writing. The function of exemplification can be fulfilled by a whole spectrum of single words (the preposition *like*, the adverb *notably*, the abbreviation *e.g.*) and word combinations, i.e. word-like units or mono-lexemic phrasemes (the preposition *such as*, the complex adverbs *for example* and *for instance*), sentence stems (*An example of Y is X; Examples include ...*) and rhemes (... *is an example of ...; ... provides a classic example of ...*), imperative clauses (*Consider, for example ...*) and sentence-initial infinitive clauses (*To take one example, ...*). A large majority of these word combinations are **semantically and syntactically fully compositional** except for a few collocations such as *prime example* and *classic example.* They are, however, characterized by their **high frequency of use** and can be described as "preferred ways" (cf. Altenberg 1998) of giving an example in professional academic writing.

By contrast, Siepmann (2005) analyses a 9.5-million word corpus of academic writing but does not make use of statistical methods. He enumerates every single occurrence of word sequences used to give an example and lists rare events such as the infinitive clauses *to paint an extreme example* and *to pick just one example* (1 occurrence in his corpus), the co-occurrence *example + is afforded by* and the expression *for the sake of example.* It is argued here that privileging exhaustiveness over typicality is counter-productive in corpus linguistic research and that such an approach results in too much –unreliable - information. Siepmann, for example, writes that "English authors have ***a large range of exemplificatory imperatives***[136] at their disposal, using the direct second-person imperative VP ~ as well as the less imposing hortative *let us* + VP and the inclusive *let me* + VP. Of these last two, the former is around five times more frequent than the latter, showing a high degree of audience sensitivity among authors" (2005:120). A closer look at his frequency data (reprinted in Table 6.8), however, shows that the co-occurrences *see/take/consider* + *for example* account for 89% of the imperatives found by the author. First person plural imperatives are extremely rare and *let me* + VP only appears 3 times in the 9.5 million word corpus of professional academic writing used in his study. In addition, first person plural imperatives do not appear with the complex adverb *for example* but introduce a noun phrase headed by the noun *example.* In summary, although a large range of exemplificatory imperatives **may** be available to language users, only a very limited set of these are of widespread use in professional academic writing. The second person imperative is clearly more frequent than the hortative *let us* + VP and the inclusive *let me* + VP. However, it is commonly used with three verbs only, namely *consider, take* and *see*.

Table 6.8: 1st person plural imperatives in academic writing (based on Siepmann 2005:119)

(for example/for instance) see (for example/for instance) NP (200)
(for example) consider (for example) NP (54)
take, for (another) example, NP (16)
Consider a(n) (ADJ) example/instance (7)
take the example of (as examples of NP) (5)
consider (as an example) NP (3)
take, as an example, NP (1)
as an illustration (of this)/ by way of (brief) illustration, consider NP (2)
Take (even) NP (2)
Let us (now) take + (as) + DET + ADJ + example(s) (4)
Let us consider + DET + ADJ + example(s) (4)
Let me give (you) (but) one example (2)
Let me offer + DET (+ ADJ+) example (1)
Let us consider, for the sake of illustration, NP (1)

[136] My emphasis.

The analysis of exemplifiers presented in this section also validates the method used to design the *Academic Keyword List* as the exemplificatory lexical items which were extracted are of two types:

- the **most frequent** exemplifiers in academic writing (*such as, example, for example* and *for instance*) (cf. Figure 6.1 above);
- lexical items which are not as frequent as *such as, example, for example* and *for instance*, but which are **EAP-specific** in the sense that they are more frequent in academic prose than in any other genre (*illustrate, exemplify, e.g., notably*).

The preposition *like* can be used to fulfil an exemplificatory function in academic writing but it is much more common in other genres. The nouns *illustration* and *case in point* are quite characteristic of formal textual genres but are infrequent items. The expressions *to name but a few* and *by way of illustration* are rare events in all types of discourse.

In conclusion, while AKL items and their phraseological patterns should definitely be taught to EFL learners, other lexical items play an important role in EAP and therefore deserve to be included in an EAP syllabus. However, their pedagogical relevance will clearly depend on a range of factors such as EFL learners' proficiency level. Although the noun *case in point* and the expressions *to name but a few* and *by way of illustration* may be taught to advanced EFL learners with a view to increasing their lexical repertoire, these will certainly not be the first exemplificatory items taught to beginners. Crewe (1990) reviews the advantages of what he calls a 'reductionist approach' to the teaching of connectives, i.e. an approach by which learners are "simply forbidden from 'ringing the changes' on the connectives in a random manner and forced to come to terms with a small, relatively discrete, subset of the original long list" (1990:321) and argues that "a shortened list has the advantage of allowing the contrasts between the connectives to be more easily stressed" (1990:322). Another deciding factor is what learners actually do with these items: do they use the expressions *to name but a few* and *by way of illustration*? If so, do they use them correctly? And do they use them sparingly or do they make heavy use of these infrequent exemplifiers? These questions can only be answered by an analysis of learner corpus data as presented in section 6.3.1.

6.2.2. The phraseology of rhetorical functions in academic prose

It is impossible to describe in similar detail the other four functions that were analyzed in native writing so as to provide a basis for comparison to EFL learner writing. Instead, this section briefly comments on the types of lexical devices used by expert writers to serve the functions of expressing cause and effect, comparing and contrasting, expressing a concession and reformulating in an attempt to give a general overview of the way EAP vocabulary is used to serve specific rhetorical functions and to characterize the phraseology of these rhetorical functions in academic prose.

The organizational functions of **expressing a concession** and **reformulating: paraphrasing or clarifying** are mainly realized by single words and mono-lexemic phrasemes. As shown in Table 6.9, the lexical means available to express a concession consist of single word adverbs (e.g. *however, nevertheless, yet*), (complex) conjunctions (e.g. *although, even though*) and (complex) prepositions (e.g. *despite, in spite of*). Similarly, reformulation is most frequently achieved by means of the mono-lexemic units *that is* and *in other words*, the abbreviation *i.e.* and the adverb *namely* (cf. Table 6.10). Nouns, verbs and adjectives are not used to serve these two functions.

Table 6.9: 'Expressing a concession' in BNC-AC-HUM

	Abs. freq.	%	Rel. freq.
Adverbs			
however	3,353	28.59%	100.94
nevertheless	676	5.76%	20.35
nonetheless	66	0.56%	1.99
though ADV	144	1.23%	4.33
yet	1,817	15.49%	54.7
TOTAL ADVERBS	***6,056***	***51.64%***	***182.3***
Conjunctions			
although	2,292	19.54%	69
though CONJ	1,721	14.68%	51.8
even though	248	2.11%	7.47
(even if)	451	3.85%	13.57
albeit	80	0.68%	2.4
TOTAL CONJ.	***4,792***	***40.86%***	***144.26***
Prepositions			
despite	681	5.8%	20.5
in spite of	159	1.36%	4.79
notwithstanding	39	0.33%	1.17
TOTAL PREP.	***879***	***7.5%***	***26.46***
TOTAL	**11,727**	**100%**	**353**

Table 6.10: 'Reformulating: Paraphrasing or Clarifying' in BNC-AC-HUM

	Abs. freq.	**%**	**Rel. freq.**
i.e.	330	25.11%	9.93
that is	375	28.54%	11.29
that is to say	81	6.16%	2.44
in other words	210	15.98%	6.32
namely	187	14.23%	5.63
viz.	21	1.6%	0.63
or more precisely	12	0.91%	0.36
or more accurately	7	0.53%	0.21
or rather	91	6.93%	2.74
TOTAL	**1314**	**100%**	**39.56**

Adverbs, prepositions and conjunctions also represent a large proportion of the lexical devices used by expert writers to serve the functions of **comparing and contrasting** and **expressing cause and effect**. Unlike the functions of introducing a concession and reformulating, however, these functions can also be realized by means of **nouns, verbs and adjectives in specific phraseological or lexico-grammatical patterns**. Nouns account for 32.52% of the lexical means used to express a cause or an effect in academic writing, e.g. *cause, factor, source, effect, result, consequence, outcome* and *implication* (cf. Table 6.11). Verbs are also particularly common: *cause, bring about, contribute to, lead to, result in, derive, emerge,* and *stem*. Patterns involving nouns (e.g. *contrast, comparison, difference* and *distinction*) and verbs (e.g. *contrast, differ, distinguish* and *differentiate*) are often used to compare and contrast but adjectives play a more prominent role and account for 29.24% of the lexical means used by expert writers (cf. Table 6.12). Examples include *different, distinct, differing* and *distinctive*. These findings have important pedagogical implications and clearly point to EFL/EAP materials' need to give more prominence to nouns, verbs, adjectives and phrasemes when teaching cohesion. These materials tend to focus exclusively on monolexemic devices (see Section 6.4).

Table 6.11: 'Comparison and contrast' in BNC-AC-HUM

	Abs. freq.	**%**	**Rel. freq.**
nouns			
resemblance	116	0.4%	3.49
similarity	212	0.72%	6.38
parallel	147	0.5%	4.43
parallelism	19	0.06%	0.57
analogy	175	0.6%	5.27
contrast	522	1.78%	15.71
comparison	311	1.06%	9.36
difference	1,318	4.51%	39.68

differentiation	76	0.26%	2.29
distinction	595	2.03%	17.91
distinctiveness	10	0.03%	0.3
(the) same	559	1.91%	16.8
(the) contrary	28	0.1%	0.84
(the) opposite	85	0.29%	2.56
(the) reverse	56	0.19%	1.69
TOTAL NOUNS	***4,229***	**14.46%**	***127.3***
Adjectives			
same	2,580	0.88%	77.68
similar	1,027	3.51%	30.92
analogous	55	0.19%	1.66
common	1055	3.61%	31.76
comparable	223	0.76%	6.71
identical	137	0.47%	4.12
parallel	52	0.18%	1.57
alike	98	0.34%	2.95
contrasting	63	0.22%	1.90
different	2,496	8.53%	75.14
differing	72	0.25%	2.17
distinct	278	0.95%	8.37
distinctive	163	0.56%	4.91
distinguishable	33	0.11%	0.99
unlike	43	0.15%	1.29
contrary	27	0.09%	0.81
opposite	127	0.43%	3.8
reverse	23	0.08%	0.69
TOTAL ADJECTIVES	***8,552***	**29.24%**	***257.44***
Verbs			
resemble	138	0.47%	4.15
correspond	137	0.47%	4.12
look like	102	0.35%	3.07
compare	278	0.95%%	8.37
parallel	56	0.19%	1.69
contrast	137	0.47%	4.12
differ	242	0.83%	7.29
distinguish	404	1.38%	12.16
differentiate	74	0.25%	2.23
TOTAL VERBS	***1,568***	**5.36%**	***47.2***
Adverbs			
similarly	394	1.35%	11.86
analogously	2	0.01%	0.06
identically	2	0.01%	0.06
correspondingly	29	0.1%	0.87
parallely	0	0	0
likewise	118	0.4%	3.55
in the same way	56	0.19%	1.69
contrastingly	3	0.01%	0.09
differently	97	0.33%	2.92
by/in contrast by contrast in contrast	185 116 69	0.63%	5.57
by way of contrast	0	0%	0

by/in comparison	23	0.08%	0.69
by comparison	14		0.42
in comparison	9		0.27
comparatively	69	0.24%	2.08
comparatively			
*comparitively	-		
contrariwise	4	0.01%	0.12
distinctively	25	0.09%	0.75
on the other hand	372	1.27%	11.20
(on the one hand)	136	0.46%	4.09
on the contrary	95	0.32%	2.86
quite the contrary	2	0.005%	0.06
conversely	62	0.21%	1.87
TOTAL ADVERBS	***1,674***	**5.72%**	***50.39***
Prepositions			
like[137]	*2,812*	9.61%	*84.65*
unlike	244	0.83%	7.3
in parallel with	8	0.03%	0.24
as opposed to	121	0.41%	3.64
as against	46	0.16%	1.38
in contrast to/with	82	0.28%	2.47
in contrast to	73		2.2
in contrast with	9		0.27
versus	53	0.18%	1.60
contrary to	66	0.23%	1.99
by/in comparison with	52	0.18%	1.57
in comparison with	14		0.42
in comparison to	4		0.12
by comparison with	21		0.63
in comparison with	14		0.42
TOTAL PREP.	***3,484***	**11.91%**	***104.88***
Conjunctions			
as[1]	*5,045*	17.25%	*151.87*
while[1]	*1264*	4.32%	*38*
whereas	442	1.51%	13.31
whereas			
wheras			
TOTAL CONJ.	***6,751***	**23.08%**	***203.23***
Other expressions			
as … as	2,766	9.46%	83.26
in the same way as/that	38	0.13%	1.14
compared with/to	155	0.53%	4.67
compared with	113		3.4
compared to	42		1.26
CONJ compared to/with	32	0.1%	0.96
as compared to/with	11		0.33
when compared to/with	20		0.6
if compared to/with	1		0.03
TOTAL	**29,249**	**100%**	**880.5**

[137] Estimations based on an analysis of the first 200 occurrences of the conjunction in each corpus.

Table 6.12: 'Cause and effect' in BNC-AC-HUM

	Abs. freq.	%	Rel. freq.
nouns			
cause	755	2.85%	22.7
factor	550	2.08%	16.6
source	1,175	4.44%	35.4
origin	500	1.89%	15
root	183	0.69%	5.5
reason	1,802	6.8%	54.25
consequence	450	1.7%	13.6
effect	1,830	6.91%	55
result	813	3.07%	24.5
outcome	143	0.54%	4.3
implication	411	1.68%	12.37
TOTAL NOUNS	***8,612***	***32.52%***	***259.25***
Verbs			
cause	570	2.15%	17.2
bring about	125	0.47%	3.8
contribute to	276	1.04%	8.3
generate	227	0.85%	6.8
give rise to	101	0.38%	3
induce	67	0.25%	2
lead to	671	2.53%	20.2
prompt	115	0.43%	3.5
provoke	161	0.61%	4.9
result in	327	1.24%	9.8
yield	129	0.49%	3.9
make sb/sth do sth[138]	171	0.65%	5.16
arise from/out of	145	0.55%	4.4
derive	476	1.8%	14.3
emerge	466	1.76%	14
follow from	74	0.28%	2.2
trigger	56	0.21%	1.7
stem	95	0.36%	2.9
TOTAL VERBS	***4,252***	***16.06%***	***128***
Adjectives			
consequent	53	0.2%	1.6
responsible (for)	344	1.29%	10.4
TOTAL ADJ.	***397***	***1.49%***	***12***
Prepositions			
because of	599	2.26%	18
due to	195	0.74%	5.9
as a result of	196	0.75%	5.9
as a consequence of	22	0.08%	0.7
in consequence of	1	0%	0.03
in view of	66	0.25%	2
owing to	52	0.2%	1.6
in (the) light of	109	0.41%	3.3
thanks to	35	0.13%	1
on the grounds of	22	0.08%	0.7

[138] Estimations based on Gilquin (to appear)

on account of	24	0.09%	0.7
TOTAL PREP.	***1,321***	**4.99%**	***39.8***
Adverbs			
therefore	1,412	5.33%	42.5
accordingly	130	0.49%	3.9
consequently	143	0.54%	4.3
thus	1,767	6.67%	53.2
hence	283	1.07%	8.5
so	1,894	7.15%	57
thereby	182	0.69%	5.5
as a result	101	0.38%	3
as a consequence	20	0.08%	0.6
in consequence	14	0.05%	0.4
by implication	35	0.13%	1.05
TOTAL ADVERBS	***5,981***	**22.59%**	***180.04***
Conjunctions			
because	2,207	8.34%	66.4
since	*955*	3.6%	*28.74*
as[139]	*883*	3.34%	*26.58*
for	1,036	3.91%	31.2
so that	696	2.63%	21
PRO is why	52	0.19%	1.56
that is why	22	0.08%	0.66
this is why	18	0.07%	0.54
which is why	12	0.05%	0.36
on the grounds that	83	0.312%	2.5
seeing as	0	0%	0
TOTAL CONJ.	***5,912***	***22.33%***	***177.97***
TOTAL	**26,475**	**100%**	**796.99**

Table 6.13 shows a co-occurrence analysis (cf. section 4.2.3) of several nouns that are used to express a **cause or an effect** in academic prose: *reason, implication, effect, outcome, result* and *consequence*. Most of the listed co-occurrents form rather flexible and compositional sentence **stems** with their respective nominal node as illustrated in the following examples:

6.58. ***Another direct result of*** *conquest by force of arms* ***was*** *the development of slavery, which was widespread up to the beginning of the nineteenth century.* (BNC-AC-HUM)

6.59. ***This may be an effect of*** *the uncertainty around television 's textuality; but it is now an extremely limiting effect for the development of theory.* (BNC-AC-HUM)

6.60. *Health for women was held to be synonymous with healthy motherhood.* ***This had important implications for*** *the debate over access to birth control information and abortion -- rarely were demands for freer access to birth control information devoid of maternalist rhetoric.* (BNC-AC-HUM)

[139] Estimations based on an analysis of the first 200 occurrences of the conjunction in each corpus.

6.61. ***The reason is that** with Van Gogh art and life are not merely conditioned by each other to a greater degree than with any other artist, but actually merge with each other.* (BNC-AC-HUM)

6.62. *However **it is first necessary to consider another important consequence of** the view of psychosis being presented here.* (BNC-AC-HUM)

Like nonce combinations, these co-occurrences are entirely explainable by grammatical and semantic rules. Unlike nonce combinations, the co-occurrents involved are **preferred co-occurrents** of the nouns under study and, as such, are better described as part of the phraseological spectrum. More precisely, the phrasemes illustrated in sentences 6.58 to 6.62 are good examples of what Sinclair and his followers have called '**extended units of meaning**' where lexical and grammatical choices are "intertwined to build up a multi-word unit with a specific semantic preference, associating the formal patterning with a semantic field, and an identifiable semantic prosody, performing an attitudinal and pragmatic function in the discourse" (Tognini-Bonelli 2002:79) (cf. section 2.4.2.3). These extended units of meaning are categorized as **textual phrasemes** in our typology as they function as sentence stems to organize the propositional content at a metadiscoursal level.

A few co-occurrences are **restricted collocations**, e.g. *a knock-on effect* (example 6.63) and *carry implications* (example 6.64). The adjective *knock-on* is almost always used with the noun *effect* (88.8% of its occurrences in the whole BNC), with which it acquires a specific meaning metaphorically derived from the meaning of the verb *knock*. The verb *carry* is used in a delexical sense in the collocation *carry implications*, which basically means *have implications*.

6.63. *There seemed no end to the **knock-on effects** — the Magyar attempt to resist Vienna provoked the development of nationalism among Croats, Serbs, Roumanians and more.* (BNC-AC-HUM)

6.64. *We may certainly talk of animals, in the absence of speech, "consciously intending" or being compassionate, both of which **carry implications of** understanding to some degree.* (BNC-AC-HUM)

Table 6.13: Co-occurrents of nouns expressing cause and effect in BNC-AC-HUM

Table 6.13a: *reason*

Adjectives	**Object of ...**	**Other**
good	have	this (PRO & DET)
main	give	another
sufficient	see	**(no) reason to ...**
obvious	base on	believe
other	provide	suppose
different	find	doubt
alleged	examine	prefer
simple	**Complement of ...**	think
tactical	be	fear
political	seem	accept
major	**Subject of ...**	**reason(s) for ....**
additional	be	supposing
right	justify	believing
valid	**Prepositions (1R)...**	thinking
similar	for	accepting
fundamental	against	rejecting
real	**Preposition (2L)**	adopting
independent	for	**There + reason**
special	**Conjunctions**	There is (no) reason to
possible	why	There seems no reason
historical	which	There are (DET/ADJ) reasons
particular	that	

Table 6.13b: *implication*

Adjectives	**Complement of ...**
important	be
practical	**Prepositions (1R)...**
political	of
serious	for
social	**Prepositions (1L)**
Object of the verbs...	with
have	**Other**
carry	this (PRO & DET)
Subject of the verb...	that (CONJ)
be	

Table 6.13c: *effect*

Adjectives	**Object of ...**
adverse	have
overall	produce
good	achieve
profound	create
knock-on	cause
indirect	**Complement of ...**
far-reaching	be
damaging	**Subject of ...**
cumulative	be
dramatic	depend on
immediate	occur
excellent	**Prepositions (1R)...**
long-term	of
practical	on
particular	upon
powerful	**Other**
special	this (PRO & DET)
full	that (CONJ)
general	**Noun**
important	cause
other	

Table 6.13d: *outcome*

Adjectives	**Object of the verbs...**
logical	influence
eventual	determine
likely	represent
different	affect
inevitable	**Subject of the verb...**
final	be
Prepositions (1R)...	**Complement of ...**
of	be
Other	
this (PRO & DET)	

Table 6.13e: *result*

Adjectives	**Object of the verbs...**
inevitable	produce
direct	achieve
immediate	yield
beneficial	give
eventual	bring
interesting	lead to
practical	show
main	present
similar	interpret
Prepositions (1R)...	obtain
of	have
from	**Subject of the verb...**
Prepositions (3L)	be
with	**Complement of ...**
Other	be
this (PRO & DET)	

Table 6.13f: *consequence*

Adjectives	**Object of the verbs...**
inevitable	have
unintended	suffer (from)
unfortunate	avoid
direct	consider
important	outweigh
necessary	discuss
political	**Subject of the verb...**
natural	be
bad	follow
practical	ensue
social	**Complement of ...**
likely	be
major	**Prepositions (1R)...**
possible	of
Other	for
this (PRO & DET)	**Prepositions (3L)**
that (CONJ)	with
another	of

Like nouns, verbs that serve specific rhetorical or organizational functions in academic prose generally enter rather compositional and flexible sequences. Table 6.14 gives the most frequent lexical bundles which contain one of four verbs that serve to express **possibility or certainty**, namely *suggest, appear, prove* and *tend.* Most clusters are lexico-grammatical patterns of the verbs which function as sentence stems (e.g. *it has been suggested that, it appears that*), sentence-initial adverbial clauses (e.g. *as suggested above,* ...) or rhemes (e.g. ... *proved a complete failure*). It is worth noting that each verb form has its own phraseological patterns which constitute different **form/meaning pairings** and thus different complete units of meaning. For example, the verb form *suggested* is often used to report suggestions made by other people in impersonal structures introduced by *it* (e.g. *it has been suggested, it is sometimes suggested*), and in phrases introduced by the conjunction *as* (e.g. *as already suggested by*). It is also used in impersonal structures introduced by *it* followed by a modal verb (e.g. *it may/might be suggested that*) to make a tentative suggestion. The *–ed* form also appears in *as*-phrases including an endophoric marker (e.g. *as suggested above*) and/or the 1st person pronoun *I* (e.g. *as I have suggested*) to refer to a suggestion previously made. By contrast, the verb form *suggests* is typically used to make it clear that the suggestion offered is made on the basis of a particular thing which occupies the subject position of the sentence:

6.65. ***More recent evidence suggests**, however, that while it lives in woodland it actually hunts over nearby open areas (Glue & Hammond, 1974; Yalden, 1985).*

6.66. ***This suggests** a high degree of polarity between low income renters and high income owner-occupiers.*

It is also used to report a suggestion made by somebody else, as in the following sentence:

6.67. ***Sinclair Hood (1971) suggests that** woollen cloth and timber were sent to Egypt in exchange for linen or papyrus.*

There are very few noun + verb or verb + noun co-occurrences, with the exception of *evidence + suggest*.

Table 6.14: Co-occurrents of verbs expressing possibility and certainty in BNC-AC-HUM

Table 6.14a: *suggest*

suggested	suggest
- *it can / could / may be suggested that* - *it is (sometimes, commonly) suggested that* - *it was (first, also, even) suggested that* - *it has been suggested that* - *as (already) suggested by* - *as suggested above* - *this is suggested by* - *(as) I suggested / have (already) suggested*	- NP / *it / this might / may / would suggest* (*that*) - NP does suggest (that) - *there is evidence to suggest* - *I (would / want to) suggest* - NP / *it / this seems to suggest (that)*
suggests	suggesting
- NP / *it / this* (ADV: *strongly, also*) *suggests* (*that*) - *the evidence suggests (that)* - ..., *which suggests (that)* - *as* NP *suggests*	- ... , (ADV: *strongly*) *suggesting* (*that*) - *I am (not) suggesting that*

Table 6.14b: *prove*

proved	prove
- NP / *it / this proved to* - NP / *it / this proved* (ADV) ADJ (*to*) with ADJ: *difficult, unable, abortive, impossible, inadequate, successful, possible* - NP / *it / proved to be* (ADV) ADJ - NP *proved a (complete/dismal) failure*	- ADJ *(likely, difficult, easy, possible) to prove* ... *may / might / would prove* ADJ *to* - NP *was to prove* ADJ - *attempt to prove* - *seek to prove*
proves	proving
- NP *proves* ADJ (*impossible, necessary, inadequate, successful*) - NP *proves that*	- BE *proving* - ..., *proving that* - ... *by proving* - *hope/possibility/way of proving*

Table 6.14c: *appear*

appeared	appears
- *it appeared* (ADJ) *that* - *there appeared to be* - *this appeared to* V - ... *which appeared* ADJ/ *to* V	- NP / *it* / *this appears to* V - *which appears to* V - *what appears to* V - *there appears to* V - *it appears that* - *as appears from/in*
appear	appearing
NP *would/might/may appear to be/*V	/

Table 6.14d: *tend*

tended	tend
- NP tended to V (*be, favour, take, see*)	- NP tend to V (*be, see, look, regard*)
tends	tending
- NP *tends to* V - .. *which tends to* V - *it tends to* V V: *be, confirm, ignore, obscure, become, support, conclude*	/

Gläser (1998) has shown that idiomatic phrases, allusions to proverbs and quotations, and striking modifications of phrasemes, can be found in academic prose where they function as stylistic devices. Studies focusing on terminological terms in English for Specific Purposes have also revealed the pervasiveness of compounds (e.g. Bourigault et al. 2004) in specialized texts. Our findings indicate that the phraseology of rhetorical and organizational functions in academic prose does **not** consist of idioms, similes, phrasal verbs, idiomatic sentences, proverb fragments and the like (see also Pecman 2004 and Gledhill 2000). Figure 6.6 presents an adaptation of the typology of phrasemes as proposed in section 2.3.7 to account for the specific phraseology of rhetorical or organizational functions in academic prose. Types of phrasemes that belong to the phraseology of English for General Purposes but are not found in the phraseology of EAP rhetorical or organizational functions are crossed out (e.g. idioms in referential phrasemes and slogans in the category of communicative phrasemes). Categories of phrasemes that are added to account for the specific phraseology of EAP vocabulary, and more particularly, EAP functions, are in bold italics.

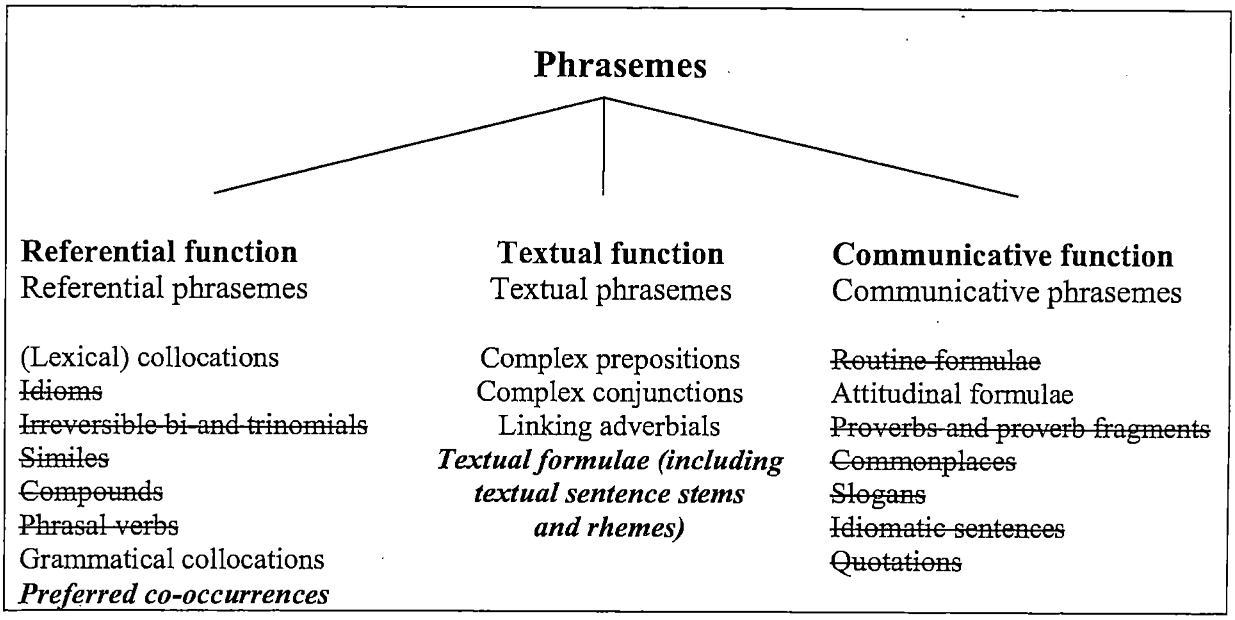
Referential phrasemes mainly consist of **lexical and grammatical collocations** (e.g. *carry implications, tend to*) and **preferred co-occurrences** (e.g. *direct result, evidence suggests, final outcome*). The latter category is added because it accounts for a large proportion of the phrasemes that are most typical of academic prose.

The category of **textual phrasemes** consists of three types of phrasemes in academic prose:

- **Complex prepositions** and **complex conjunctions** establish grammatical relations (cf. Burger's category of structural phrasemes).
- **Linking adverbials** are used to connect two stretches of discourse, e.g. *for example, to conclude, that is, as suggested above* (cf. Conrad 1999; section 1.4.1).
- I propose to use the general category of **textual formulae** to account for two types of phrasemes that are particularly prominent in academic prose. Textual formulae typically take the form of **sentence stems**, i.e. multiple clause elements involving a subject and a verb, which "form the springboard of utterances leading up to the communicatively most important – and lexically most variable – element" (Altenberg 1998:113). Examples include *It has been suggested, Another reason is ...,* and *It is argued that...* . They are also sometimes realized by means of **rhemes**, i.e. typically a verb and its post-verbal elements, which do not contain any thematic element (e.g. ... *is another issue*). Textual sentence stems are much more frequent than rhemes, which can be explained by the fact that rhemes are "usually tailored to expressing the particular new information the speakers want to convey to their listeners, and are therefore, as Altenberg (1998: 111) points out, 'composed of variable items drawn from an open set'" (De Cock 2003:269). Textual formulae display different degrees of flexibility, from rather flexible fragments such as 'DET (*a, another*) ADJ (*typical, classic, prime, good,* etc.) *example of* [NP] *is* ...' to rather inflexible phrasemes such as '... *is a case in point*'.

Attitudinal formulae make up a large proportion of communicative phrasemes in academic prose. They largely consist of (attitudinal and interactional) sentence stems such as *it is important/necessary that*, *it seems that* or *it is noteworthy.*

Figure 6.6: The phraseology of rhetorical/organizational functions in EAP



Phrasemes

Referential function
Referential phrasemes

(Lexical) collocations
~~Idioms~~
~~Irreversible bi-and trinomials~~
~~Similes~~
~~Compounds~~
~~Phrasal verbs~~
Grammatical collocations
Preferred co-occurrences

Textual function
Textual phrasemes

Complex prepositions
Complex conjunctions
Linking adverbials
Textual formulae (including textual sentence stems and rhemes)

Communicative function
Communicative phrasemes

~~Routine formulae~~
Attitudinal formulae
~~Proverbs and proverb fragments~~
~~Commonplaces~~
~~Slogans~~
~~Idiomatic sentences~~
~~Quotations~~

6.3. EAP vocabulary in learner writing

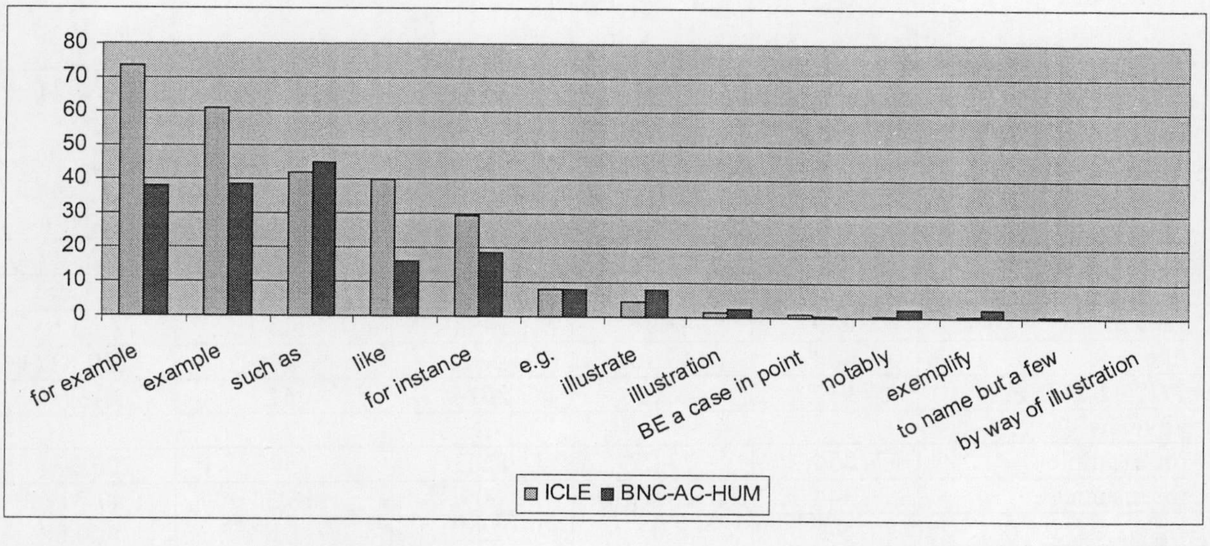
This section is devoted to EAP vocabulary in EFL learner writing. Section 6.3.1 first presents a detailed comparison of exemplificatory devices in native and learner writing which aims to illustrate the type of data and results obtained when comparing the range of lexical strategies available to EFL learners to that of expert writers. The focus of Section 6.3.2 is placed on the general interlanguage features that emerge when applying this methodology to learners' use of lexical items used to serve other rhetorical functions. These interlanguage features fall into six categories: aspects of overuse and underuse and the range of EFL learners' lexical repertoire, lack of register-awareness, phraseological infelicities, semantic misuse of connectors and abstract nouns, clusters of connectives and sentence position of connectors.

6.3.1. Exemplification in learner writing

A general finding of the comparison between ICLE and BNC-AC-HUM is that exemplificatory lexical items are **significantly more frequent in learner writing** than in professional academic prose. This finding highlights the importance of analysing several learner populations and comparing them so as to avoid faulty conclusions about EFL learner writing in general. Siepmann (2005) finds that the complex adverbs *for example* and *for instance* are less frequent in German learner writing than in native and non-native professional writing and argues that "[u]nder-use of exemplification as a rhetorical strategy in student writing may (…) bespeak a general lack of concern for comprehensibility" (Siepmann 2005:255). This explanation for German learners' underuse of exemplifiers is not entirely satisfactory[140] and does not apply to EFL learner writing in general: most L1 learner populations overuse exemplificatory discourse markers.

The bar chart in Figure 6.7 gives the relative frequencies per 100,000 words of exemplifiers in ICLE as compared with those in BNC-AC-HUM. Lexical items are ordered by decreasing relative frequency in ICLE (in blue). The bar chart shows that EFL learners' use of exemplifiers differs from that of professional writers in at least two ways. First, they do not show preference for the same exemplifiers. Thus, unlike in BNC-AC-HUM, the most frequent exemplifier in ICLE is the complex adverb *for example* (vs. *such as* in BNC-AC-HUM). Second, frequencies of individual items may differ widely. Figures and log-likelihood values for each corpus comparison are given in Table 6.15.

Figure 6.7: Exemplification in ICLE vs. BNC-AC-HUM



[140] Siepmann (2005) uses corpora of professional academic writing compiled from the internet as comparable corpora. Very few studies have investigated the linguistic features of web-based academic texts but it is most probable that the medium used influences the texts produced in terms of their level of formality, lexis, etc.

Table 6.15: Exemplification - comparisons based on total number of running words

	ICLE		BNC-AC-HUM		LogL
	Abs.	**Rel.**	**Abs.**	**Rel.**	
nouns					
example	713	61.17	1285	38.68	91.56 (++)
example	477	40.92	665	20.02	134.02 (++)
examples	230	19.73	620	18.66	0.52
*exemple[141]	4	0.34			
*exampl	1	0.09			
*examle	2	0.17			
*exaple	1	0.09			
illustration	17	1.46	77	2.3	3.29
illustration	16	1.37	63	1.9	
illustrations	1	0.09	14	0.42	
(BE) a case in point	10	0.86	18	0.5	1.29
TOTAL NOUNS	***740***	***63.5***	***1380***	***41.5***	***83.55 (++)***
verbs					
illustrate	51	4.38	259	7.8	16.1 (--)
illustrate	29	2.49	97	2.92	0.59
illustrates	14	1.2	63	1.9	2.62
illustrated	8	0.69	84	2.52	17.73 (--)
illustrating	0	0	15	0.45	9.02
exemplify	5	0.43	79	2.38	23.09 (--)
exemplify	2	0.17	9	0.27	0.37
exemplifies	2	0.17	15	0.45	2.1
exemplified	1	0.09	53	1.6	24.62 (--)
exemplifying	0	0	2	0.03	1.2
TOTAL VERBS	***56***	***4.8***	***338***	***10.2***	***32.14 (--)***
prepositions					
such as	489	41.96	1494	45	1.8
like	468	40.15	532	16	199.62 (++)
TOTAL PREP.	***957***	***82.1***	***2026***	***61***	***55.31 (++)***
adverbs					
for example	854	73.27	1263	38	206.97 (++)
for instance	344	29.51	609	18.33	47.33 (++)
e.g.	94	8.07	259	7.8	0.08
notably	5	0.43	77	2.32	22.13 (--)
to name but a few	3	0.26	4	0.12	0.93
by way of illustration	1	0.09	3	0.09	0
TOTAL ADVERBS	***1301***	***111.6***	***2215***	***66.7***	***206.19 (++)***
TOTAL	**3054**	**262**	**5959**	**179.4**	**277.04 (++)**

[141] The 'word list' option of WST4 (cf. section 4.4.2.1) was used to search for any misspelt form of the words under study in ICLE.

As shown in Table 6.15, EFL learners' **overuse** of the function of exemplification is largely explained by their massive overuse of the complex adverbs *for example*[142] and *for instance*[143], the noun *example*[144] and the preposition *like.* By contrast, learners tend to make little use of the verbs *illustrate* and *exemplify* and the adverb *notably*, which are **underused** in ICLE. There is no significant difference in use of the preposition *such as*, the abbreviation *e.g*, the nouns *illustration* and *case in point* and the expressions *to name but a few* and *by way of illustration* when comparisons are based on the total number of running words in each corpus. Except for the preposition *such as* and the abbreviation *e.g*, these lexical items are quite infrequent both in native and learner writing.

As explained in section 5.7, comparisons can also be based on the **total number of exemplifiers** - which will represent 100% - rather than on the total number of running words in each corpus. Corpus comparisons based on the total number of running words have shown that the function of exemplification is significantly more frequent in ICLE and that only four lexical items are responsible for this massive overuse. Comparisons based on the total number of exemplifiers, by contrast, allow us to answer different research questions. They give information on which lexical item(s) EFL learners prefer using any time they want to give an example and in which proportion. As shown in Table 6.16, *for example* is selected by EFL learners in 28% of the times they introduce an example.

[142] Overuse of *for example* is also found in other learner populations such as Japanese and Taiwanese learners (cf. Narita and Sugiura 2006; Chen 2006).

[143] The overuse of *for instance* has already been reported by Granger and Tyson (1996) for French learners and Altenberg and Tapper (1998) for Swedish learners.

[144] The relative frequencies of *for instance* and *example* are higher than in BNC-AC-HUM in most learner corpora. When learner corpora are analyzed separately, however, the differences in use are only significant for a few learner populations. Aggregated frequencies thus also help revealing general though moderate overuse in learner corpora.

Table 6.16: Exemplification - comparisons based on total number of exemplifiers

	ICLE		BNC-AC-HUM		LogL
	Abs.	**%**	**Abs.**	**%**	
nouns					
example	713	23.3%	1285	21.6%	2.87
illustration	17	0.56%	77	1.3%	11.65 (-)
(BE) a case in point	10	0.33%	18	0.3%	0.04
TOTAL NOUNS	***740***	***24.2%***	***1380***	***23.2%***	***0.98***
Verbs					
illustrate	51	1.67%	259	4.35%	47.52 (--)
exemplify	5	0.16%	79	1.3%	38.29 (--)
TOTAL VERBS	***56***	***1.8%***	***338***	***5.7%***	***78.77 (--)***
Prepositions					
such as	489	16%	1494	25%	79.47 (--)
like	468	15.3%	532	8.9%	71 (++)
TOTAL PREP.	***957***	***31.3%***	***2026***	***34%***	***4.37***
Adverbs					
for example	854	28%	1263	21.2%	38.33 (++)
for instance	344	11.3%	609	10.2%	2.06
e.g.	94	3%	259	4.35%	4.35
notably	5	0.16%	77	1.3%	36.88 (--)
to name but a few	3	0.09%	4	0.06%	0.24
by way of illustration	1	0.03%	3	0.05%	0.15
TOTAL ADVERBS	***1301***	***42.6***	***2215***	***37.2%***	***15.04 (++)***
TOTAL	**3054**	**100%**	**5959**	**100%**	

Both methodologies highlight a clear overuse of the preposition *like* and the complex adverb *for example,* which represent 15.3% and 28% respectively of all exemplifiers in ICLE, as opposed to 8.9% and 21.2% in BNC-AC-HUM (cf. Table 6.16). As shown in Table 6.17, results may also differ according to the methodology employed. The noun *example* has been shown to be overused in ICLE when comparisons are made on the total number of running words in each corpus. However, a comparison based on the total number of exemplifiers suggests that EFL learners do not select the noun *example* significantly more often than professional writers when they want to introduce an example (23.3% vs. 21.6%). Results also indicate that more lexical items are significantly underused when percentages are compared. In addition to *illustrate, exemplify* and *notably,* the noun *illustration* and the preposition *such as* are proportionally less often selected by EFL learners than by professional writers to give an example. This first broad picture of exemplifiers in learner and professional academic writing points to EFL learners' **limited repertoire** of lexical items used to serve a specific EAP function. This characteristic of learner writing is discussed in more detail in section 6.3.2.1.

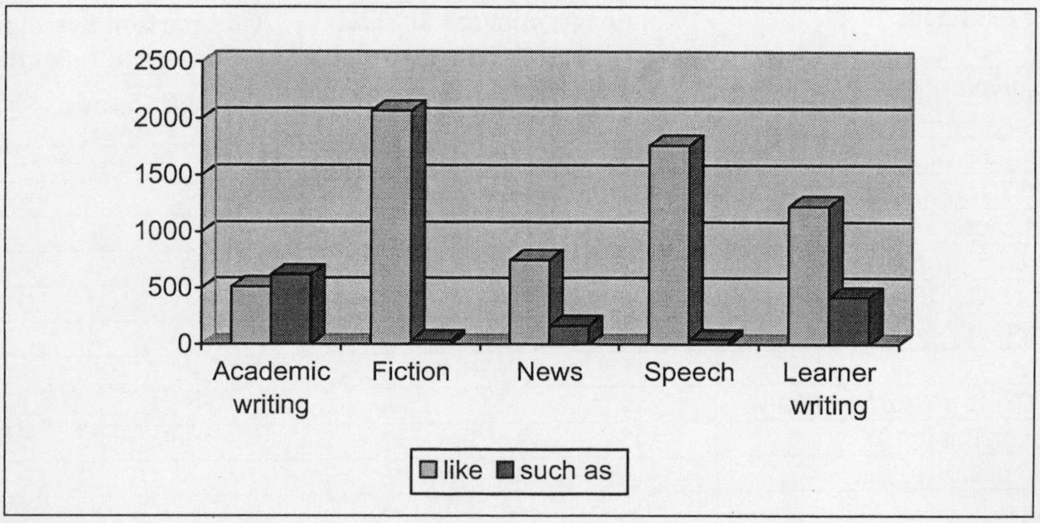
Table 6.17: Two methodologies for corpus comparisons

Lexical item	Comparison based on total number of running words	Comparison based on total number of exemplifiers
example	++	/
illustration	/	-
(BE) a case in point	/	/
TOTAL NOUNS	++	/
illustrate	--	--
exemplify	--	--
TOTAL VERBS	--	--
such as	/	--
like	++	++
TOTAL PREPOSITIONS	++	/
for example	++	++
for instance	++	/
e.g.	/	/
notably	--	--
to name but a few	/	/
by way of illustration	/	/
TOTAL ADVERBS	++	++

EFL learners' use of the exemplificatory prepositions *such as* and *like* reveals learners' **lack of register-awareness.** Learners overuse the preposition *like*, which is less typical of academic writing, irrespective of its various functions. By comparison, they underuse *such as*, which can be described as EAP-specific. Figure 6.8 shows the relative frequencies per 1,000,000 words of *like* and *such as* in four main genres of the *British National Corpus*, namely academic writing, fiction, newspaper texts and speech as well as in ICLE. The preposition *like* is much more frequent than *such as* in speech[145], fiction, news and learner writing but is less frequent in academic prose. By contrast, *such as* is more frequently used in academic prose. Learner writing can thus be paralleled with more informal genres such as conversation or fiction as far as the use of the prepositions *like* and *such as* is concerned.

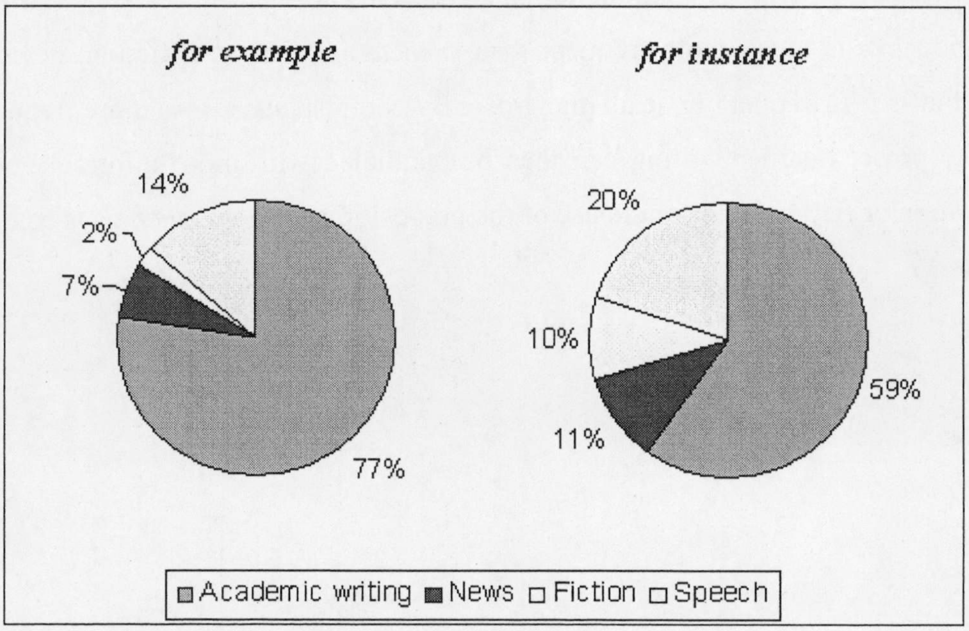
[145] See Miller and Weinert (1995), Siegel (2002) and Biber et al (1999:562) for specific functions of *like* in speech. See Müller (2005: 197-228) for an analysis of *like* as a discourse marker.

Figure 6.8: The prepositions 'like' and 'such as' across registers



Similarly, a large proportion of EFL learner populations make repeated use of the monolexemic unit *for instance*. The use of this textual phraseme in professional writing, however, differs significantly from that of *for example* both in terms of frequency and register. Figure 6.9 shows that 77% of all instances of *for example* in the BNC are found in the academic subcorpus. Comparatively, only 59% of the occurrences of *for instance* appear in academic prose while 30% are found in more informal genres such as speech and fiction.

Figure 6.9: Distribution of the conjuncts 'for example' and 'for instance' across registers (BNC)

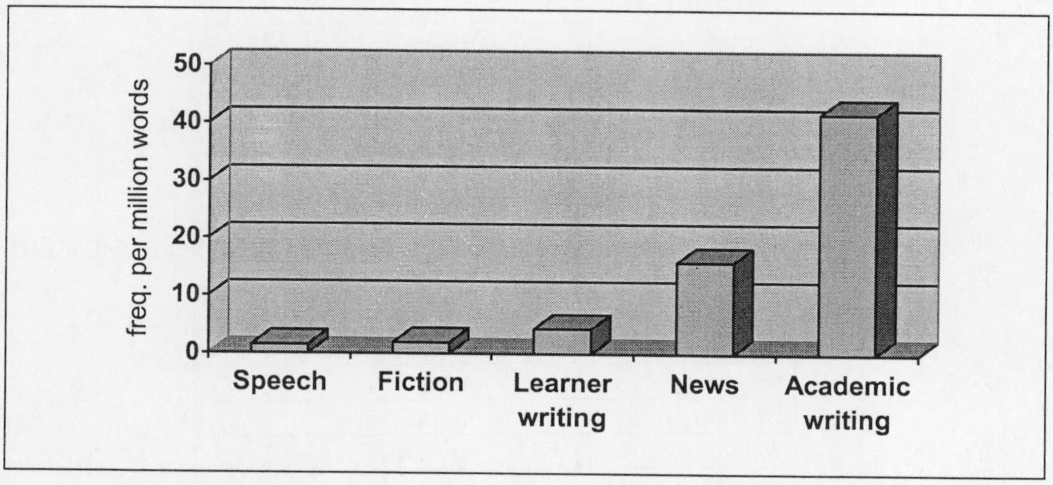


Lack of register-awareness manifests itself in a number of ways in learner academic writing. Examples include:

- Overuse of more informal linguistic features (e.g. the preposition *like*),
- Underuse of words and phrasemes that are typical of academic discourse (e.g. the adverb *notably* – cf. Figure 6.10 -; the verbs *illustrate* and *exemplify*),

Register-related learner specificities will be the focus of section 6.3.2.2.

Figure 6.10: Distribution of the adverb 'notably' across genres



The **phraseology** of EAP words can also be a primary source of learners' difficulties in academic writing. One of the main advantages of using a noun rather than the complex adverb *for example* and *for instance* to give an example is that the use of a noun allows the writer to qualify the example with an adjective (cf. section 6.2.1.2). Only 18% of the adjective co-occurrents (types) of the noun *example* in ICLE, however, are significant co-occurrents in BNC-AC-HUM (see Table 6.18). A quarter of the adjective co-occurrents of *example* in learner writing do not appear in the 100-million word BNC (cf. Table 6.19). A large proportion of these adjectives have been described by our native informant as forming awkward co-occurrences with *example* as illustrated in the following sentences:

6.68. *The story of Cinderelia is one more* ***impermissible example****. Cinderelia is a neglected child, and once again the step-family is the guilty party.* (ICLE-DU)

6.69. *For example a disliked politician will be shot through such a zoom as to expose his ugly bits. Which may most probably influence our feeling towards him. We all know thousands of such* ***manipulative examples****.* (ICLE-PO)

6.70. *This* ***mere example*** *proves that the ideal union people dream of is not yet a total reality : national conflicts are still at work, every nation defends its own interests before fighting for those of "the group" they joined.* (ICLE-FR)

6.71. *The* ***opposite example*** *is (the former?) USSR, where the union was imposed by a central power without real approbation of the states and against people's will.* (ICLE-FR)

6.72. *Of course, that was an* ***overstated example,*** *extreme, so to speak.* (ICLE-RU)

Table 6.18: Significant adjectives co-occurrents of the noun 'example' in ICLE

Adjective	freq.	Adjective	freq.
good	77	*excellent*	4
extreme	12	*typical*	3
above	8	*classic*	2
clear	8	*interesting*	2
striking	7	*numerous*	2
simple	6	*outstanding*	1
well-known	5		

Table 6.19: Adjectives co-occurrents of the noun 'example' in ICLE not found in BNC

Adjective	freq.	Adjective	freq.
big	2	*manipulative*	1
warning	2	*mere*	1
absolute	1	*model*	1
bright	1	*opposite*	1
cruel	1	*overstated*	1
present day	1	*polemic*	1
evident	1	*hair raising*	1
frightening	1	*striring*	1
impermissible	1	*upsetting*	1

Similarly, only 23% of the verbs (types) that are used with *example* in ICLE are significant co-occurrents of the noun in BNC-AC-HUM (see Table 6.20). 27% of the verb co-occurrents (types) of the noun *example* in ICLE do not appear in the whole BNC. They are listed in Table 6.21. Like adjective co-occurrents, several of these verbs form awkward co-occurrences with the noun *example*:

6.73. *In a*new society* made with less inequality, less poverty and more social justice we would not find the same quantity of crime that we find in our society.* ***I can make the example of*** *Naples: here there is everyday an incredible lot of crimes.* (ICLE-IT)

6.74. *Their understanding of the outside world differs. It originates in dissimilar climate, life-style, social organization, political and economical stability of the country.* ***To glide into an extreme example,*** *unequality appears even between people living in towns and villages.* (ICLE-CZ)

6.75. *The rules of the road you have to learn to pass your driving license* ***are plastered with examples of*** *children who cross the road unexpectedly, running after a ball.* (ICLE-GE)

Table 6.20: Significant verb co-occurrent types of the noun 'example' in ICLE

Left co-occurrents		Right co-occurrents	
Verb	**freq.**	**Verb**	**freq.**
be	162	*be*	119
take	36	*show*	31
give	28	*illustrate*	15
find	10	*concern*	2
show	10	*suggest*	1
serve	4	*suffice*	1
illustrate	3		
provide	2		
cite	2		
consider	1		
TOTAL	**258**	**TOTAL**	**169**

Table 6.21: Verb co-occurrent types of the noun 'example' in ICLE not found in BNC-AC-HUM

Left co-occurrents		Right co-occurrents	
Verb	**freq.**	**Verb**	**freq.**
culminate into	1	*say*	1
glide into	1	*reinforce*	1
state	1	*criticize*	1
plaster with	1	*point out*	1
derive	1	*express*	1
write	1		
help as	1		
appear	1		
TOTAL	**8**	**TOTAL**	**5**

The copular *be* is the most frequent left and right co-occurrent of the noun *example* in learner writing. Stems and rhemes with the verb *be* are significantly more frequent in learner writing than in professional academic writing (see Table 6.22). These results differ markedly from those reported in Paquot (to appear) in which French, Spanish, Italian and German learners were shown to underuse stems and rhemes with the verb *be*. Such a difference may be explained by a difference in comparable corpus type: unlike in this thesis, the comparable corpus used in Paquot (to appear) is a corpus of native student essays. We will discuss learner vs. native student writing differences in section 6.3.3.

Table 6.22: 'example' and 'be' in ICLE and BNC-AC-HUM

	be + example	***example + be***	**TOTAL**	**Rel. freq.**	**LogL**
ICLE	162 (57.65%)	119 (42.35%)	281	24.1	199.76 (++)
BNC-AC-HUM	139 (62.3%)	84 (37.7%)	223	6.71	

Table 6.23 shows that the structure *there* + *be* + *example* is more frequently used in learner writing. It is used in the 10 learner corpora as illustrated in the following sentence:

6.76. ***There is the example of*** *Great Britain where a professional army costs less than, for example, the French army based on conscription.* (ICLE-RU)

Table 6.23: '*There/Here* + BE + *example*' in ICLE

	There + BE + example		LogL
	Abs. freq.	Rel. freq.	34.52 (++)
ICLE	31 [11%]	2.66	
BNC-AC-HUM	15 [1.2%]	0.45	

In professional academic writing, the verb *take* is mainly used in sentence-initial exemplificatory infinitive clauses with the noun *example* (cf. example 6.77), a pattern which is very infrequent in ICLE. EFL learners prefer using the verb *take* in active structures introduced by the personal pronoun *I* (ex. 6.78) or in 1st person plural imperative sentences (ex. 6.79).

6.77. ***To take one example,*** *at the beginning of the project seven committees were established, each consisting of about six people, to investigate one of a range of competing architectural possibilities.* (BNC-AC-HUM)

6.78. ***I can take the example of*** *the "Société Générale de Belgique" which is directed by "Suez".* (ICLE-FR)

6.79. ***Let's take the example of*** *painting.* (ICLE-FR)

Learners use the verb *have* in the same structures as *take* to introduce an example as illustrated in sentences 6.80 and 6.81. The imperative sentence, however, was judged to be awkward by our native informant.

6.80. ***Let us have an example*** *- an extract out of the famous Figaro's soliloquy: There is a liberty of the press in Madrid now, so that I can write about anything I like, providing I will have it checked by two or three censors and an condition that I will not write against the government and religion.* (ICLE-CZ)

6.81. ***I have a good example*** *in my family.* (ICLE-PO)

Interestingly, the verb *have* and the 1st person plural imperative *let's* are not significant left co-occurrents of *example* in BNC-AC-HUM but they are in BNC-SPOKEN. The verb *have* is often used with an inclusive *we* as subject (ex. 6.82); *let's* is typically used with the verb *take* + *example* (cf. sentence 6.83).

6.82. *Er in relation to existing employment sites er and Mr Laycock referred to National Power, erm there* ***we have an example of*** *the attitude that the the council is taking towards the the re-use of employment sites.* (BNC-SPOKEN)

6.83. ***Let's take the example of*** *a cooker.* (BNC-SPOKEN)

The verb *give* is the most significant co-occurrent of the noun *example* in BNC-SPOKEN. It is used in questions and 1st person plural imperative sentences (examples 6.84 and 6.85), two patterns that are not found in BNC-AC-HUM despite the fact that the verb is also a significant co-occurrent of *example* in academic prose. By contrast, 1st person plural imperative sentences with the verb *give* appear in ICLE (example 86).

6.84. *Can you* ***give an example*** *when you say that the law is designed?* (BNC-SPOKEN)

6.85. ***Let me give you some examples.*** (BNC-SPOKEN)

6.86. ***Let me give you one example*** *– appaling shots from the war in ex-Yugoslavia that we can see nearly every day.* (ICLE-CZ)

In summary, verb co-occurrents of the noun *example* provide strong evidence for the **genre-bound nature of phrasemes**[146]. They also suggest that EFL learners sometimes select verb co-occurrents of the noun *example* that are more typical of speech and other more informal text types, which can be interpreted as further indication of their lack of register awareness.

Differences in phraseological or lexico-grammatical preferences can often be revealed by patterns of overuse and underuse of **word forms.** Thus, the different forms of the verbs *illustrate* and *exemplify* are not all underused in learner writing. Table 6.24 shows that the two verbs are underused in their *–ed* form only. This underuse corresponds to an underuse of the passive constructions *BE illustrated by/in* (example 6.87) and *BE exemplified by/in* (example 6.88), the past participle *exemplified* following a noun phrase (example 6.89) and the patterns *as illustrated/exemplified by/in* (example 6.90):

6.87. *The contrast between the conditions on the coast and in the interior* ***is illustrated by*** *the climatic statistics for two stations less than 30 km (18.5 miles) apart.* (BNC-AC-HUM)

[146] Other verb co-occurrents that are quite frequent in BNC-SPOKEN but not found in BNC-AC-HUM are the verbs *get* and *think*.
- *So* ***we've got some examples*** *here of some patterns that we want to learn using the N tuple method and tuple and tuple.* (BNC-SPOKEN)
- *Again* ***think of the example of*** *erm erm a social club you know, relationships between members, although they may be close and intimate and friendly and all that, are not the same as a relationship between members of a family.* (BNC-SPOKEN)

6.88. *The association of this material with the clerk* ***is clearly exemplified by*** *Chaucer's Wife of Bath's fifth husband, the clerk Jankyn, who, in the Wife of Bath's Prologue, reads antifeminist material to her from his book "Valerie and Theofraste".* (BNC-AC-HUM)

6.89. *Piaget's claim that thinking is a kind of internalized action,* ***exemplified*** *in the assimilation-accommodation theory of infant learning mentioned above, is really a global assumption in search of some refined, detailed and testable expression.* (BNC-AC-HUM)

6.90. *He assumed, without argument, that science,* ***as exemplified by*** *physics, is superior to forms of knowledge that do not share its methodological characteristics.* (BNC-AC-HUM)

Table 6.24: 'illustrate' and 'exemplify' in ICLE

	ICLE		BNC-AC-HUM		LogL
verbs	**Abs. freq.**	**Rel. freq.**	**Abs. freq.**	**Rel. freq.**	
illustrate	51	4.38	259	7.8	16.1 (--)
illustrate	29	2.49	97	2.92	0.59
illustrates	14	1.2	63	1.9	2.62
illustrated	8	0.69	84	2.52	17.73 (--)
illustrating	0	0	15	0.45	9.02
exemplify	5	0.43	79	2.38	23.09 (--)
exemplify	2	0.17	9	0.27	0.37
exemplifies	2	0.17	15	0.45	2.1
exemplified	1	0.09	53	1.6	24.62 (--)
exemplifying	0	0	2	0.03	1.2

The verb *illustrate* is more often used with human subjects (11.76%) in learner writing, and more specifically with the personal pronoun *I*:

6.91. ***I would like to illustrate that*** *by means of some examples which, as you will see, are very diverse;* ...(ICLE-DU)

6.92. *In the worst cases people decide to suicide.* ***I can illustrate*** *that by a real example.* (ICLE-CZ)

It is also frequently used in sentence-initial infinitive clauses (13.72%):

6.93. ***To illustrate the truth of this,*** *one has only to mention people's dissapointment* (sic) *when realizing how little value has the time spent at university.*(ICLE-SP)

6.94. ***To illustrate this point,*** *it would be interesting to compare our situation with the U.S.A.'s.* (ICLE-FR)

Like in professional academic writing, the phraseme *case in point* is very rarely used in learner writing. When used, however, the expression sometimes appears in lexico-grammatical patterns that are not found in native professional writing, e.g. in an infinitive clause with the verb *take* (sentence 6.95) or determined by a definite article and followed by a *that*-clause (sentence 6.96).

6.95. *However, wars always break out for economical reasons; For example, the first world war,* ***to take a case in point****, did not start because the murder or archduke Frank Ferdinand, heir of Autro-Hungary; that was only the straw that broke the camel's back.* (ICLE-SP)

6.96. *Professional observers see some even deeper danger in the emerging situation. A great number of children spend more and more time watching television. They take into consideration the behaviour patterns of film stars, they want to be like them.* ***The case in point is that*** *little children learn how to smoke how to drink how to be cunning and clever and get round the adults. Film stars are usually very attractive and it's not a surprise that children want to follow them.* (ICLE-RU)

EFL learners' phraseological and lexico-grammatical specificities or infelicities will be discussed in detail in section 6.3.2.3.

EFL learners may also experience difficulty with semantic features of single words and phrasemes. For example, they sometimes use the abbreviation *i.e.* instead of *e.g.* as an exemplificatory discourse marker (examples 6.97 to 6.100). The abbreviation *i.e.*, however, should only be used as a synonym of 'that is' to reformulate by paraphrasing or clarifying (cf. Appendix 6.1e).

6.97. *The states mostly tend to solve their politic problems in a peaceful way (*****i.e.*** **[e.g.]** *the split of Czech federation or the unification of Germany).* (ICLE-CZ)

6.98. *One of the examples that makes this point is related to children's toys, because nowadays children play with technological toys (*****i.e.:*** **[e.g.]** *video games), and these toys do not let the children develop their imagination and, in many cases, they are so inactive that playing with these toys does not permit physical exercise.* (ICLE-SP)

6.99. *In English every type of essay* ****i.e.*** **[e.g.]** *definition, cause and effect, comparison and contrast, argumentative etc. is ruled by its own conventions.* (ICLE-PO)

6.100. *It might seem absurd, but many progressive social changes (*****i.e.*** **[e.g.]** *an increase of individual liberty) may lead to further increase of crime.* (ICLE-RU)

Learners also sometimes use *as* in lieu of the complex preposition *such as* (examples 6.101 to 6.105). It should be noted, however, that this erroneous use is more frequently found in learner populations with Romance mother tongue backgrounds (see chapter 7 for a discussion of the potential influence of the mother tongue on the use of single words and phrasemes that serve specific rhetorical or organizational functions).

6.101. *Thus soldiers learned mostly bad habits* ****as*** **[such as]** *smoking, drinking (if possible) and being lazy in their leisure time.* (ICLE-CZ)

6.102. *In addition to the familiar subjects* ****as*** **[such as]** *reading, writing and mathematics, time should be reserved for making children conscious of the fact that there is more to life than the things we see.* (ICLE-DU)

6.103. *There should be particular institutions for those who are mentally alienated* ****as*** **[such as]** *the rapists, others for the young people, etc.* (ICLE-FR)

6.104. *In this essay I would like to show how, in my opinion, crime is caused by a predisposition of the individuals and how, of course, other factors* ****as*** **[such as]** *society, culture and politics can influence this natural inclination.* (ICLE-IT)

6.105. *Another proof will be the role that imagination plays in all the Arts* ****as*** **[such as]** *Literature, Music and Painting.* (ICLE-SP)

The adverb *namely* is also sometimes misused in EFL learner writing where it is used instead of *notably* or another exemplifier as illustrated in examples 6.106 and 6.107.

6.106. *This new wave of revolting trivial events is all the more worrying since it is linked to a rise of the small delinquance, implying a gerenalised climate of terror and a total mistrust of the citizens towards the police forces and the law, both accused of all vices and* ****namely*** **[(most) notably]** *of being too lax with those evils.* (ICLE-FR)

6.107. *No doubt they are important to us but it is also obvious that the progress and the amazing or, as it is sometimes said, appalling results do not produce those qualities that are characteristic only to humanbeings , and* ****namely-****: love , harmony with nature , kindness and soon and so forth .* **[most notably** love, harmony with nature and kindness] (ICLE-RU)

This confusion is relatively common and is even found on websites devoted to English connectors as illustrated in Figure 6.11.

Figure 6.11: The treatment of 'namely' on websites devoted to English connectors

Pour donner des exemples
For instance (par exemple).
For example (par exemple)
Such as (tel que)
Like (comme)
Namely (c'est-à-dire)
Above all (surtout)

http ://perso.orange.fr/frat.st.paul/Mots-outils.htm

un exemple : utilisez *for example, for instance, in other words, in particular, namely, to illustrate*, etc.
http://pages.usherbrooke.ca/notabene/chroniques/cleanglaise/cleanglaise2003-2004.htm

More generally, *namely* is very often misused in learner writing and it is not always clear what learners mean when they use this adverb:

6.108. *Because the campus consists of modern buildings, built closely together, it is no more than a ten minute's walk to get where you need to be for lectures and seminars. All the academic facilities are* ***?namely*** *located on the main campus.* (ICLE-DU)

6.109. *Why, then, so many people object to gay marriages and, at the same time, yearn for equality? It is* ***?namely*** *just equality what gay marriages are about, isn't it?* (ICLE-FI)

6.110. *We strive for a multiplex society where even the tiniest minority is allowed to live according to its belives and convictions. On the other hand, the freedom of speech is restricted for the same reasons.* ***?Namely****, it is difficult to think how the Finnish Jews could live a balanced and equal life, if the anti-Semitic circles were openly allowed to spread their truths. Censorship is freedom.* (ICLE-FI)

6.111. *The efforts made by the firms are obvious. They* ***?namely*** *create replacement products: they replace the gas in the aerosols and so we have ozone-friendly aerosols, ...* (ICLE-FR)

6.112. *Reluctance to eventually join The Common Market is* ***?namely*** *caused by fear, disbelieves, inferiority complex, short-sightedness or even nationalistic and xenophobic tendencies.*(ICLE-PO)

6.113. *The first possible answer is that theoretically all humans are equal, but in practice some are more equeal,* ***?namely****: some are more common, not so notable, than a few being on the top.*(ICLE-SP)

More examples of **semantic misuse** are illustrated and discussed in section 6.3.2.4.

Another explanation for the general overuse of the function of exemplification in learner writing may also be that exemplifiers are repeatedly used when they are superfluous, redundant or even when other rhetorical functions should be made explicit. Sentence 6.114 is an example of a superfluous exemplificatory discourse marker[147]:

6.114. *On the one hand, there is always a slight connection between sports and politics. At international sport events like world championships or Olympics athletes can only participate as representatives of their country. Thus* ***for example,*** *before all international football matches the national anthems of the teams are played. The same is true for victory ceremonies at Olympics.* (ICLE-GE)

In sentence 6.115, the logical relation between the two sentences is a causal link that is left implicit while an unnecessary exemplifier is used:

6.115. *I described there only some examples from the great number of criminal offences. After some years many of those criminals will be set free because of their relatively mild punishment. They had* ***for example*** *youthful age. (Youthful age - by the way in contrast to the punishment of 16 years old boys in our country, who got off with the light punishment, in England were recently sentenced two 10 years old boys for murder of a 3 years old boy to the lifelong punishment!)* (ICLE-CZ)

Section 6.3.2.5 will focus on the **redundant use** of lexical items that serve rhetorical or organizational functions as well as on learners' tendency to clutter up their texts with too many logical devices.

EFL learners' use of exemplifiers also differs from native professional writers with respect to their **sentence position**. Unlike in BNC-AC-HUM, sentence-initial position of the complex adverbs *for example* and *for instance* is clearly favoured in ICLE:

6.116. *But there are actually a number of things we all can do that make a difference.* ***For example,*** *there ought to be information about different ways to save electricity.* (ICLE-SW)

6.117. *There were a lot of wars due to the religion.* ***For instance,*** *England has always been divided according to the kind of religion in which a person believed.* (ICLE-SP)

The two complex adverbs are also repeatedly found at the end of a sentence in ICLE (7.14% of the occurrences of *for example* and 8.4% of the occurrences of *for instance*), a position which is rare in academic professional writing (*for example* = 1.6%; *for instance* = 1.3%):

[147] Note that this type of redundant use is not rare in speech.

6.118. *Let us have a good look at television* ***for example.*** (ICLE-POLISH)

6.119. *Furthermore the psyche surpassed the nature of cosmic allegorism, which involved the representation of thunderstorm as a performance of the god Zeus* ***for example.*** (ICLE-DUTCH)

6.120. *They only want an easy to operate camera, a Single Use Camera* ***for instance.*** (ICLE-DU)

6.121. *I find the pronunciation of English much more difficult than the pronunciation of Italian,* ***for instance.*** (ICLE-POLISH)

Aspects of **sentence position** are dealt with in section 6.3.2.6. Section 6.3.2.7 briefly comments on other learner-specific features such as spelling and punctuation errors.

In section 6.3.1, we argued that AKL lexical items and their phraseological patterns should definitely be taught to EFL learners. Learner corpus data reviewed here support this claim as each of these lexical items presents one or more learner-specific features. Examples include semantic misuse of the adverb *notably* and of the abbreviation *e.g.,* sentence position of the adverbs *for example* and *for instance* and learner-specific phraseological use of the noun *example* and the verbs *illustrate* and *exemplify*. It was also argued that, apart from learners' proficiency level, the pedagogical relevance of the other lexical items – the preposition *like*, the nouns *illustration* and *case in point* and the expressions *to name but a few* and *by way of illustration* - clearly depends on whether learners already use these exemplifiers and how they use them. Learner corpus data suggest that:

- A word of caution is needed against the heavy use of the preposition *like*;
- The noun *illustration* should be taught to advanced learner as it is underused in ICLE;
- The specific lexico-grammatical patterns of *case in point* should also be taught as this phraseme is repeatedly used in 'unidiomatic' patterns.

More **pedagogical implications** of learner corpus-based findings will be considered in section 6.4.

6.3.2. EAP vocabulary and general interlanguage features

Comparisons of the other four functions that were analyzed in native and learner writing cannot be provided in similar detail. Instead, this section focuses on the general interlanguage features that emerge from these analyses, which fall into six categories: aspects of overuse and underuse and the range of EFL learners' lexical repertoire, lack of register-awareness, phraseological infelicities, semantic misuse of connectors and abstract nouns, clusters of connectives and sentence position of connectors. A synthesis of quantitative results for 5 functions – **exemplification**, **expressing cause and effect**, **comparing and contrasting**, **expressing a concession** and **reformulating** - is given in Appendices 6.1a to 6.1e. The figures and examples presented in this section are based on these appendices, which the reader is encouraged to refer to for absolute and relative frequencies as well as percentages of each lexical item in both ICLE and BNC-AC-HUM and log-likelihood values for corpus comparisons.

6.3.2.1. Limited lexical repertoire

Several studies have argued that EFL learners are not equipped with the lexical repertoire necessary for writing academic texts (cf. section 1.4.2). An analysis of learners' use of the EAP-specific words that constitute the *Academic Keyword List* (AKL) further supports this view. Table 6.25 shows that almost 50% of AKL words are underused in ICLE, a percentage that rises up to 52.1% for nouns and 56.3% for adverbs. By contrast, the percentage of overused AKL words in learner academic writing is only 21.4%. The largest proportions of overused items are found in nouns and in the category 'other' which includes prepositions, conjunctions, determiners, etc.

Table 6.25: The *Academic Keyword List* in ICLE

	overused	**no statistical difference**	**underused**	
nouns	86 [24.2%]	84 [23.7%]	185 [52.1%]	[100%]
verbs	40 [17.2%]	93 [39.9%]	100 [42.9]	[100%]
adjectives	34 [18.9%]	59 [32.8%]	87 [48.3%]	[100%]
adverbs	16 [18.4%]	22 [25.3%]	49 [56.3%]	[100%]
other	21 [28%]	21 [28%]	33 [44%]	[100%]
TOTAL	**199 [21.4%]**	**277 [29.8%]**	**454 [48.8%]**	**[100%]**

Table 6.26 gives examples of overused and underused AKL words in ICLE. It might be accepted as a general tendency that "learner usage tends to amplify the high frequencies and

diminish the low ones" (Lorenz 1999b:59). For example, overused items such as the nouns *idea* and *problem*, the verbs *be* and *become* and the adjectives *difficult* and *important* are very frequent words in general English (relative frequencies higher than 200 occurrences per million words in the whole BNC). Conversely, underused items such as the nouns *hypothesis* and *validity*, the verbs *exemplify* and *advocate*, the adverbs *conversely* and *ultimately* and the prepositions *as opposed to* and *in the light of* are much less frequent in English (relative frequencies lower than 30 occurrences per million words in the whole BNC).

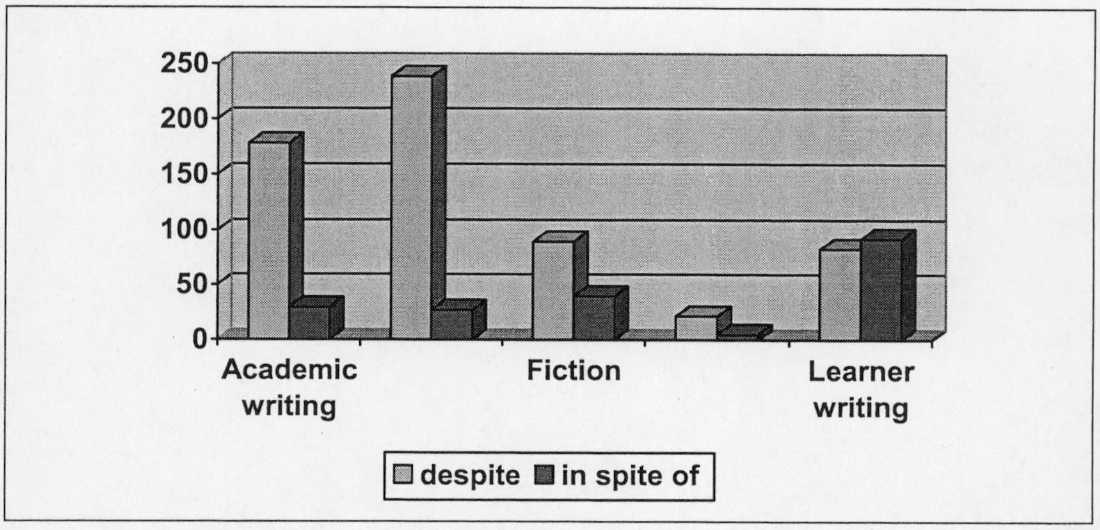
The picture, however, appears to be more complex than Lorenz's quote suggests. Not all high frequencies are amplified in EFL learner writing. Many frequent words[148] are underused in ICLE, e.g. the nouns *argument, difference* and *effect*, the verbs *argue* and *explain*, the adjectives *likely* and *significant* and the adverbs *generally* and *particularly* (in bold in Table 6.26). Key function words such as *between, in, by,* and *of* are quite representative of the nominal style of academic texts (cf. section 1.2) and of the fact that 60% of all noun phrases in academic prose have a modifier (cf. Biber 2006; Appendix 1.1). However, these highly frequent prepositions are underused in ICLE, a fact that can be related to EFL learners' tendency to avoid prepositional noun phrase postmodification (cf. Aarts and Granger 1998; Meunier 2000: 279). The preposition *despite* is underused while its much less frequent synonym, irrespective of genres, i.e. the complex preposition *in spite of*, is overused in learner writing (cf. Figure 6.12). In addition, words such as the noun *disadvantage*, the verbs *participate* and *solve* and the adverbs *consequently* and *moreover* (underlined in Table 6.26) are overused although they appear with frequencies lower than 50 occurrences per million words in the BNC.

[148] Words which appear with a relative frequency higher than 100 occurrences per million words in the whole BNC.

Table 6.26: Examples of overused and underused AKL words

	overused	**underused**
nouns	*advantage, aim, benefit, change, choice, conclusion, consequence, degree, disadvantage, example, fact, idea, influence, possibility, problem, reality, reason, risk, solution, stress*	*addition,* ***argument****, assumption,* ***basis****, bias, comparison, concept, contrast, criterion,* ***difference, effect,*** *emphasis,* ***evidence, extent, form,*** *hypothesis,* ***issue****, outcome, perspective,* ***position****, scope,* ***sense****, summary, theme,* ***theory****, validity*
verbs	*aim, allow, avoid, be, become, cause, choose, concern, consider, consist, contribute, create, deal, depend, develop, exist, improve, increase, influence, participate, prove, solve, study, treat, use*	*adopt, advocate,* ***argue****, assert, assess,* ***assume****, cite, comprise, conduct, contrast, define, derive,* ***describe****, emphasise, enhance,* ***ensure****, examine, exemplify,* ***explain****, highlight,* ***indicate, note,*** *propose,* ***reflect, reveal****, specify,* ***suggest****, view, yield*
adjectives	*common, different, difficult, important, interesting, main, necessary, obvious, possible, practical, real, special, true, useful*	*adequate,* ***appropriate****, comprehensive, critical, detailed, explicit, extensive, inherent,* ***likely****,* ***major****, misleading, parallel,* ***particular****,* ***prime****, relative, representative,* ***significant, similar,*** *subsequent, substantial, unlikely*
adverbs	*also, consequently, especially, extremely, however, mainly, more, moreover, often, only, secondly, successfully, therefore*	*adequately, conversely, effectively, essentially,* ***generally****, hence, increasingly, largely, notably, originally,* ***particularly****, potentially, previously, primarily, readily, relatively, similarly, specifically, subsequently, ultimately*
other	*according to, because, due to, during, each, for, less, many, or, same, several, some, than, this*	***although, an****, as opposed to,* ***between, by, despite, from,*** *given that,* ***in****, in relation to, in response to,* ***in terms of****, in the light of,* ***including, its****, latter,* ***of****, prior to,* ***provided, rather than,*** *subject to,* ***the, to,*** *unlike,* ***upon, which***

Figure 6.12: 'despite' and 'in spite of' across genres



The amplification of a restricted set of low frequencies in learner writing may be partly explained by teaching-induced factors as overused words such as *consequently*, *moreover* and *secondly* appear in the long and undifferentiated lists of connectors usually provided by EFL/EAP teaching materials (cf. section 6.4.1). This situation may be compounded by problems of semantic misuse, e.g. *moreover, on the contrary* (cf. section 6.3.2.4). Underuse of frequent, but semantically specialized, words probably stems from learners' tendency to rely heavily on all-purpose, general words where more precise vocabulary should be used (cf. Petch-Tyson 1999; Caldwell 2002). Another tentative explanation may be that EFL learners do not amplify any type of high frequencies but those of words that are highly frequent in **speech**. As argued by Baayen et al (2006), "the complexity of the frequency variable has been underestimated" and it may be that more emphasis should be placed on **the explanatory potential of spoken frequency counts**. Underused words such as *argument, issue, assume, indicate, appropriate*, and *particularly* are quite frequent in general English, i.e. the whole BNC, but their frequencies decrease significantly when the conversation component is analyzed separately.

In section 6.3.1, it was shown that, although they generally overuse exemplifiers, EFL learners make little use of a number of EAP-specific lexical means such as the verbs *illustrate* and *exemplify* or the adverb *notably*. They rely instead on a restricted lexical repertoire mainly composed of the complex adverbs *for example* and *for instance*, the noun *example* and the preposition *like*. The same conclusion holds for learners' use of **cause and effect** lexical items, which is compared with that of native professional writers in Appendix 6.1b. Broadly speaking, learners overuse logical links of cause and effect in their argumentative essays. This overuse, however, does not affect all grammatical categories. When corpus comparisons are based on the total number of running words in each corpus, it seems to be generally attributable to adverbs, prepositions and conjunctions. The categories of nouns, verbs and adjectives do not display significant patterns of overuse or underuse. When frequencies are measured against the total number of 'cause and effect' lexical items, by contrast, only prepositions and conjunctions are significantly overused while nouns and verbs are underused (cf. Table 6.27). This means that, compared to professional writers, EFL learners prefer using **prepositions, conjunctions and, to a lesser extent, adverbs, to the detriment of nouns and verbs** when they need to express a cause or an effect.

Table 6.27: Expressing a cause or an effect in ICLE: grammatical categories

	Absolute freq. / total freq. in corpus	% of 'cause and effect' lexical items
nouns	n[149]	--
verbs	n	--
adjectives	n	n
adverbs	++	n
prepositions	++	++
conjunctions	++	++

Table 6.28 shows that, even though EFL learners prefer using prepositions, conjunctions and adverbs to express a cause or an effect, not all individual connectors are overused in learner writing. Thus, the overuse of conjunctions largely stems from learners' marked preference for *because*, which represents 19.92% of all 'cause and effect' markers in ICLE. Lorenz (1999b) examines the use of causal links in essays written by 16-to-18-year-old German learners and German undergraduates and describes the marked overuse of the conjunction *because* as a 'wild-card use'. He argues that "[i]f a linguistic element is used as an all-purpose wild card, that usage is bound to include a number of instances of **over-extension**[150]. In other words, it can be expected that learners may disregard target-language restrictions which are not that obvious, or even accounted for in the standard grammars, but which are nevertheless observed by the native speakers. Such 'simplification' is one of the most frequently cited features of learner language" (Lorenz 1999b:60-61).

Several of the overused lexical items are **hugely overused** in learner writing. The adverb *so* represents 11.48% of the 'cause and effect' lexical items used by learners while it only accounts for 7.2% of those in professional writing. Other examples of 'lexical teddy bears' (cf. Hasselgren 1994) or 'pet' (cf. Tankó 2004) discourse markers are the prepositions *because of* and *due to*. In their study of expressions of doubt and certainty, Hyland and Milton (1997) have reported similar findings: Cantonese learners use a more limited range of epistemic modifiers, with the ten most frequently used items (*will, may, think, would, always, usually, know, in fact, actually,* and *probably*) accounting for 75% of the total.

[149] 'n' = no significant difference in use; '--' = underuse; '++' = overuse
[150] My emphasis.

Table 6.28: Cause and effect: overuse and underuse per grammatical categories (based on Appendix 6.1b)

	overuse	no statistical difference	underuse	TOTAL
nouns	2 [18.9%]	4 [36.7%]	5 [45.4%]	11 [100%]
	root[151], *consequence*	*cause, factor, reason, result*	*source, origin, effect, outcome, implication*	
verbs	1 [5.9%]	3 [17.6%]	13 [76.5%]	17 [100%]
	cause	*bring about, contribute to, lead to*	*generate, give rise to, induce, prompt, stem, provoke, result in, yield, arise, derive, emerge, follow, trigger*	
adjectives	0	1 [50%]	1 [50%]	2 [100%]
		responsible (for)	*consequent*	
adverbs	4 [40%]	2 [20%]	4 [40%]	10 [100%]
	consequently, as a result, as a consequence, so	*therefore, in consequence*	*accordingly, thus, hence, thereby*	
prepositions	3 [27.3%]	6 [54.5%]	2 [18.2%]	11 [100%]
	because of, due to, thanks to	*as a result of, owing to, as a consequence of, on the grounds of, in consequence of, on account of*	*in view of, in (the) light of*	
conjunctions	2 [40%]	0	3 [60%]	5 [100%]
	because, this/that is why		*for, so that, on the grounds that*	
TOTAL	**12 [21.4%]**	**16 [28.6%]**	**28 [50%]**	**56 [100%]**

On the other hand, 50% of the lexical means which serve to express a cause or an effect in native professional writing are underused in ICLE. While underuse is found in all grammatical categories, proportions vary significantly. **Nouns** and **verbs** constitute a large proportion of the lexical means available to express a cause or an effect in academic prose. However, 64.3% of them are underused in ICLE (e.g. the nouns *source, effect* and *implication* and the verbs *induce, result in, yield, arise, emerge* and *stem from*). This may be explained by teaching-induced factors as lexical cohesion has been largely neglected in teaching materials, i.e. textbooks and more particularly grammars, where the focus has generally been on adverbial connectors (cf. section 6.4.1).

An analysis of the lexical items which serve to express a **comparison or a contrast** in academic prose shows that the proportion of underuse is also quite high in this function. Table 6.29 reports that almost half of all 'comparison and contrast' items are underused. Like for

[151] The noun 'root' is most probably overused in ICLE because it appears in an essay prompt given to EFL learners, "In the words of the old song: "Money is the root of all evil", which is often re-used in learner texts.

'cause and effect' lexical items, the proportion of underuse varies significantly. **Nouns** and **adjectives** account for 59% of all underused 'comparison and contrast' lexical items, e.g. *resemblance, similarity, contrast, similar, distinct,* and *unlike*. The proportion of overuse is relatively low. In addition, overused items include words and phrasemes that are more frequent in speech (e.g. *look like, in the same way*) (cf. section 6.3.2.2) as well as commonly misused expressions such as *on the contrary* (cf. section 6.3.2.4). Unlike for 'cause and effect' lexical items, overused 'comparison and contrast' lexical items do not make up for the underused ones and the function is underused in learner writing.

In summary, it appears that EFL learners tend to rely heavily on **a restricted set of hugely overused adverbs, prepositions or conjunctions** to establish text cohesion. Logical links can also be provided by **nouns** (cf. the concept of 'labelling' explained in section 1.3.2.2), **verbs and adjectives**, which often account for a large proportion of the lexical strategies used to serve a specific rhetorical or organizational function in professional academic prose. These cohesive devices, however, do not seem to be readily accessible to upper-intermediate to advanced EFL learners[152], which is not particularly surprising as lexical cohesion has generally been neglected in teaching materials. Tables 6.28 and 6.29 provide useful information about learners' particular needs and have been used to inform academic writing sections (cf. section 6.4.3). They could also be used to develop vocabulary building exercises.

In this section, the breadth of EFL learners' lexical repertoire has been examined in terms of the proportion of overused and underused AKL single words and mono-lexemic units used to serve a specific rhetorical or organizational function. It should be stressed that the limited nature of EFL learners' lexical repertoire also stems from a restricted use of phrasemes and lexico-grammatical patterns typically found in professional academic prose (cf. Flowerdew 1998 and 2003), which will be discussed in section 6.3.2.3.

[152] It should be noted that these findings are not restricted to EFL learners and that English as a Second Language (ESL) speakers have been reported to experience the same difficulty (cf. Hinkel 2002 and section 1.4.2).

Table 6.29: Comparison and contrast: overuse and underuse per grammatical categories (based on Appendix 6.1c)

	overuse	no statistical difference	underuse	TOTAL
nouns	0	5 [33.33%]	10 [66.67%]	15 [100%]
		parallelism, difference, distinctiveness, the contrary, the opposite	*resemblance, similarity, parallel, analogy, contrast, comparison, differentiation, distinction, the same, the reverse*	
verbs	2 [22.22%]	5 [55.56%]	2 [22.22%]	9 [100%]
	look like, compare	*resemble, correspond, differ, distinguish, differentiate*	*parallel, contrast*	
adjectives	2 [11.11%]	4 [22.22%]	12 [66.67%]	18 [100%]
	same, different	*alike, contrary, opposite, reverse*	*similar, analogous, common, comparable, identical, parallel, contrasting, differing, distinct, distinctive, distinguishable, unlike*	
adverbs	4 [19%]	10 [47.62%]	7 [33.33%]	21 [100%]
	in the same way, on the other hand, on the one hand, on the contrary, + erroneous expressions	*analogously, differently, identically, parallely, reversely, contrariwise, by way of contrast, contrastingly, quite the contrary, comparatively*	*similarly, likewise, correspondingly, by/in comparison, conversely, by/in contrast, distinctively*	
prepositions	2 [22.22%]	3 [33.33%]	4 [44.44%]	9 [100%]
	like, by/in comparison with + erroneous expressions	*in parallel with, in contrast to/with, contrary to*	*unlike, as opposed to, as against, versus*	
conjunctions	0	1 [33.33%]	2 [66.67%]	3 [100%]
		whereas	*as, while*[153]	
other expressions	1 [25%]	3 [75%]	0	4 [100%]
	as ... as,	*in the same way as/ that, compared with/to, CONJ compared with/to*		
TOTAL	**11 [13.92%]**	**31 [39.24%]**	**37 [46.84%]**	**79**

[153] The underuse of the conjunctions *as* and *while* reported here must be taken with caution as it results from estimations based on an analysis of the first 100 occurrences of the conjunctions in each corpus.

6.3.2.2. Lack of register-awareness

Many learner corpus-based studies have reported on EFL learners' lack of register awareness (e.g. Granger and Rayson 1998; Lorenz 1999b; Altenberg and Tapper 1998; Meunier 2000; Ädel 2006). These studies, however, have often focused on the writing of learners sharing a single mother-tongue background (cf. section 1.4.2). The large-scale study undertaken in this thesis allows for a more systematic description of features of register-awareness in the way EFL learners use lexical means which serve a rhetorical function, irrespective of learners' mother-tongue backgrounds. In ICLE, most rhetorical or organizational functions are characterized by the **overuse of at least one lexical item that is more typical of speech** (or at least of more informal types of writing) than of professional academic prose, as shown in Table 6.30. Sentences 6.122 to 6.128 give examples of overused lexical items that are more frequent in the BNC spoken component than in BNC-AC-HUM: the adverb *so* to express an effect, the adverb *though* to introduce a concession, sentence-initial *and* and the adverb *besides* to add information, the complex adverb *of course* to express certainty, the stem *I am going to talk about* to introduce a new topic, the complex preposition *thanks to* to express a cause and the complex adverb *all in all* which is used to 'show that you are considering every part of a situation' (LDOCE4).

6.122. *Many people who are in this situation think that this is a waste of time: you lose an entire year.* ***So*** *they want to get rid of the military service.* (ICLE-DU)

6.123. *Spanish holds an important position in South America and increasingly so in the United States, too. According to Crystal it has little further potential ouside Spain,* ***though.*** (ICLE-FI)

6.124. *In summary, it can be stated that genetic engineering, given the present state of affairs in technology, is morally wrong.* ***And*** *it will continue to be so until we can exclude all risks that are still attached to genetic engineering.* (ICLE-DU)

6.125. *The high economic level adquired from the technology and industrialisation belongs to the developed countries but it is not extended or share with the underdeveloped countries.* ***Besides,*** *the latters are dependent of the formers.* (ICLE-SP)

6.126. *But practically everybody is able to dream.* ***Of course,*** *there are different people with different concepts of happiness, different thoughts and emotions.* (ICLE-RU)

6.127. *In this essay* ***I am going to talk about*** *the link between crime and politics; what I want to demostrate is that a good way of making politics can cut the roots to crime.* (ICLE-IT)

6.128. ***Thanks to*** *them anyone willing to broaden his\her general knowledge of the world has an easy access to useful information.* ***All in all,*** *there are many ways in which mass media affect our approach to reality and they are, by no means, all positive or good for us.* (ICLE-PO)

As shown in sentence 6.128, it is not infrequent that speech-like lexical items – *thanks to* and *all in all* - cluster together in learner writing.

Table 6.30: Speech-like overused lexical items per rhetorical function

Rhetorical function	Speech-like overused lexical item
Exemplification	*like*
Cause and effect	*thanks to* *so* *because* *that/this is why*
Comparison and contrast	*look like* *like*
Concession	the (sentence-final) adverb *though*
Adding information	sentence-initial *and* the adverb *besides*
Expressing personal opinion	*I think* *to my mind* *from my point of view* *it seems to me*
Expressing possibility and certainty	*really* *of course* *absolutely* *maybe*
Introducing topics and ideas	*I would like to/want/am going to talk about* *thing* *by the way*
Listing items	*first of all*
Reformulation: paraphrasing and clarifying	/
Quoting and reporting	*say*
Summarizing and drawing conclusions	*all in all*

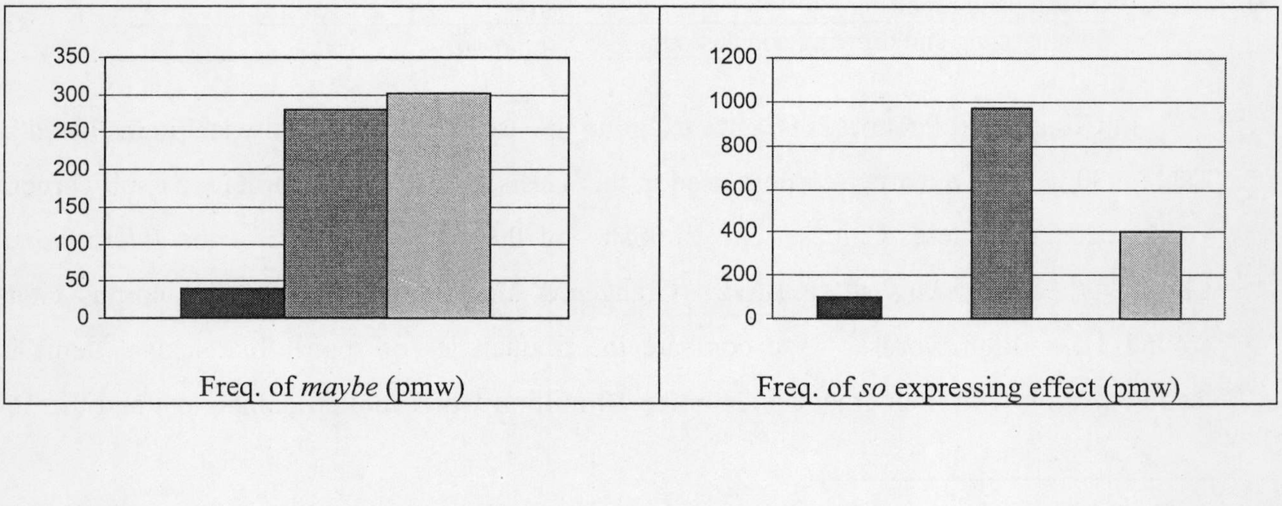
In Gilquin and Paquot (2006), we examine the use of some of the lexical items listed in Table 6.30 in the ten learner corpora used in this thesis as well as in four new L1 sub-corpora - Norwegian, Japanese, Chinese, and Turkish - of the second version of the *International Corpus of Learner English* (ICLEv2) (Granger et al. to appear 2008). The corpus totals around 1.5 million words.[154] We compare the frequencies of speech-like lexical items in learner writing with their frequencies in the 10-million word **spoken** component and the 15-

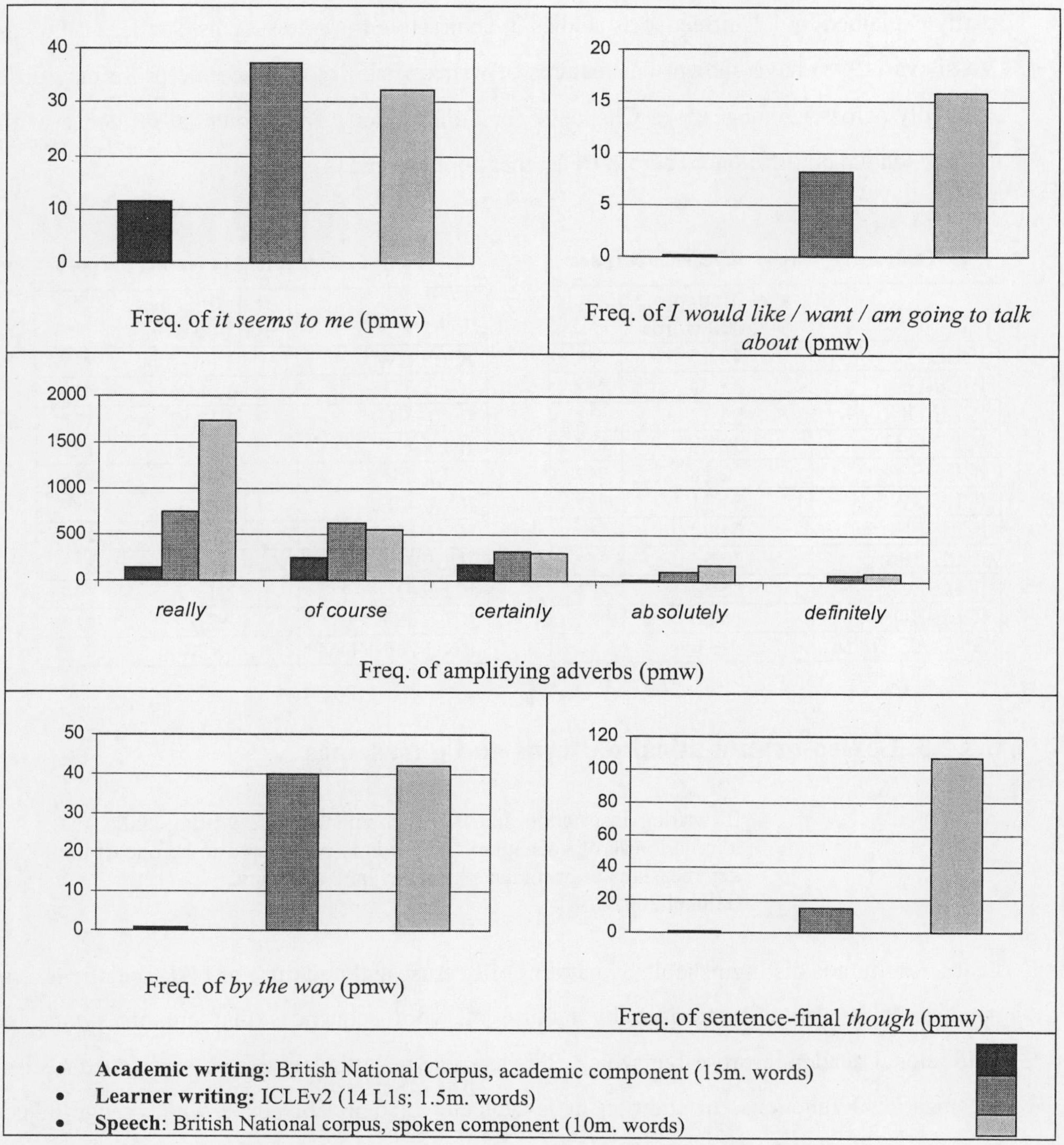
[154] The variables used to select the essays in the five new L1 sub-corpora are the same as for the corpora employed in this thesis: argumentative, untimed and with reference tools (cf. section 4.1.1).

million word **academic** sub-corpus of the *British National Corpus* (cf. 4.1.2). Our findings support Lorenz's (1999b:64) statement that there is "mounting evidence that text-type sensitivity does indeed lie at the heart of the NS/NNS numerical contrast." They show that the relative frequency of these speech-like lexical items in learner writing is often between their frequency in academic prose and that in speech (cf. graphs for *maybe, I would like/want/am going to talk about, really, absolutely, definitely, by the way* and *though* in Figure 6.13). Some of these items (e.g. *so* expressing effect, *it seems to me, of course* and *certainly*) are even more frequent in learner writing than in native speech.

The overuse of several of these speech-like lexical items has been highlighted in a number of studies focusing on specific L1 learner populations. For example, Chen (2006) reports on the overuse of *besides* in Taiwanese student writing; Lorenz (1999b) discusses the marked overuse of the conjunction *because* and the adverb *so* in German learner writing; French, Spanish and Swedish learners' heavy reliance on *I think* to express their personal opinion is reported by Granger (1998), Neff et al (2007) and Aijmer (2002) respectively; Japanese, French and Swedish learners' overuse of *of course* is highlighted by Narita and Sugiura (2006), Granger and Tyson (1996) and Altenberg and Tapper (1998) respectively. Our results suggest that these features are often not characteristic of one or two learner populations only. They are instead often **shared by a large proportion of the learner populations investigated** and are therefore **more likely to be developmental or teaching-induced**. It remains to be seen, however, whether lack of register awareness is a typical feature of EFL learner writing or whether it is a general characteristic of **novice writing**. This issue will be addressed in section 6.3.3.

Figure 6.13: Speech-like lexical items across registers (Gilquin and Paquot 2006)





Freq. of *it seems to me* (pmw)

Freq. of *I would like / want / am going to talk about* (pmw)

Freq. of amplifying adverbs (pmw)

Freq. of *by the way* (pmw)

Freq. of sentence-final *though* (pmw)

- **Academic writing**: British National Corpus, academic component (15m. words)
- **Learner writing:** ICLEv2 (14 L1s; 1.5m. words)
- **Speech**: British National corpus, spoken component (10m. words)

EFL learners, however, do not use speech-like lexical items similarly, irrespective of their mother-tongue background. Although all L1 learner populations overuse the adverb *maybe* when compared to BNC-AC-HUM, Table 6.31 shows that relative frequencies differ widely across L1 populations. Another example is EFL learners' use of *I think*, which is overused by all L1 learner populations while presenting marked differences in use across learner L1 sub-corpora. As shown in Table 6.32, relative frequencies range from 17.29 occurrences per 100,000 words in the Polish learner sub-corpus (ICLE-PO) to 143.57 occurrences per 100,000 words in the Swedish one (ICLE-SW). This huge difference may be partly explained by L1 influence as studies in contrastive rhetoric such as Connor (1996) and Vassileva (1998) have shown that features of writer visibility in academic prose may differ markedly across languages (see Chapter 7 for a discussion of the potential influence of the mother-tongue on differences across L1 learner populations).

Table 6.31: 'maybe' in learner corpora

	rel. freq. per 100,000 words	
ICLE-IT	48.18	***
ICLE-GE	38.34	***
ICLE-DU	35.13	***
ICLE-CZ	32.88	***
ICLE-SP	32.28	***
ICLE-SW	31.21	***
ICLE-FI	24.74	***
ICLE-FR	20.34	***
ICLE-PO	16.37	***
ICLE-RU	13.26	***
BNC-AC-HUM	1.93	

Table 6.32: 'I think' in learner corpora

	rel. freq. per 100,000 words	
ICLE-SW	143.57	***
ICLE-IT	134.06	***
ICLE-RU	121.13	***
ICLE-CZ	101.7	***
ICLE-FR	94.61	***
ICLE-GE	72.11	***
ICLE-SP	66.59	***
ICLE-FI	55.87	***
ICLE-DU	51.77	***
ICLE-PO	17.79	***
BNC-AC-HUM	6.14	

6.3.2.3. Lexico-grammatical patterns and phrasemes

> In writing instruction and the assessment of L2 writing skills, the idiomatic use of vocabulary and lexis is considered to be one of the key measures of proficiency, fluency, and accuracy.
> (Hinkel 2002:158)

Learner writing is distinguishable by patterns of overuse and underuse of EAP-specific lexical bundles. It is also characterized by its use of word sequences that are not typical of professional academic prose but which EFL learners nevertheless use to serve rhetorical or organizational functions. In addition, it is typically recognizable by a whole range of co-occurrences that differ from academic prose in quantitative and qualitative terms. Section 6.3.2.3.1 presents major findings of an analysis of word sequences in EFL learner writing. It focuses on aspects of overuse and underuse of word sequences that include AKL words before discussing learner-specific clusters that are not found in professional academic prose. Section 6.3.2.3.2 compares the co-occurrents of the noun *conclusion* in academic and learner writing and examines EFL learners' phraseological infelicities and lexico-grammatical errors.

6.3.2.3.1. An analysis of word sequences in EFL learner writing

Results presented in this section are based on an analysis of 2-to-5 word sequences that are overused or underused in learner writing (cf. Altenberg 1998; Stubbs 2002; De Cock 2003) and which were extracted with WST4 (cf. section 4.2.2.1). They show that learner writing is characterized by a **marked underuse** of a large proportion of 2-to-5 word sequences that include AKL words and that are typically used to serve specific rhetorical or organizational functions in academic prose. EFL learners rely instead on a restricted set of clusters which they massively overuse (e.g. *for example, main reason, another important, it depends*). Granger (1998b) suggests that the use of these sequences "could be viewed as instances of what Dechert (1984:227)[155] calls 'islands of reliability' or 'fixed anchorage points', i.e. prefabricated formulaic stretches of verbal behaviour whose linguistic and paralinguistic form and function need not be 'worked upon'" (Granger 1998b:156). This is also consistent with the author's statement that "while the foreign-soundingness of learners' productions has generally been related to a *lack* of prefabs, it can also be due to an excessive use of them" (Granger 1998b:155). The foreign-soundingness of EFL learner writing also stems from learners' overuse of AKL words in clusters that are not typical of the particular genre of academic prose but which are more frequently used in speech or more informal types of writing (e.g. *people* ***claim*** *that, I will* ***discuss****, from my* ***point*** *of* ***view****,* ***because*** *of the* ***fact***[156]).

Table 6.33 shows that EFL learners **overuse adjective + noun sequences with 'nuclear' adjectives** (cf. section 1.3.1.1) such as *main* (e.g. *main reason, main cause, main problem*), *real* (e.g. *real problem, real value*), *important* (e.g. *important role, important question, important factor*), *great* (e.g. *great number, great importance*), *different* (e.g. *different points, different problems, different reasons*) and *big* (e.g. *big problem*) to the detriment of more EAP-like phrasemes such as *extensive use, crucial importance, central issue, significant number, integral part, lesser extent* and *wide variety*. Similarly, they overuse adverb + adjective/adverb/conjunction sequences with highly frequent adverbs such as *mainly* (e.g. *mainly because*), *quite* (e.g. *quite clear*) and *very* (e.g. *very important*) but make little use of phrasemes such as *readily available, relatively few, significantly different, almost entirely, closely associated, particularly interesting, more generally, highly significant* and *precisely because.*

[155] Dechert, H. (1984). Second language production: Six hypotheses. In H. Dechert, D. Möhle and M. Raupach (eds), *Second language productions*. Tübingen: Gunter Narr, 211-230.

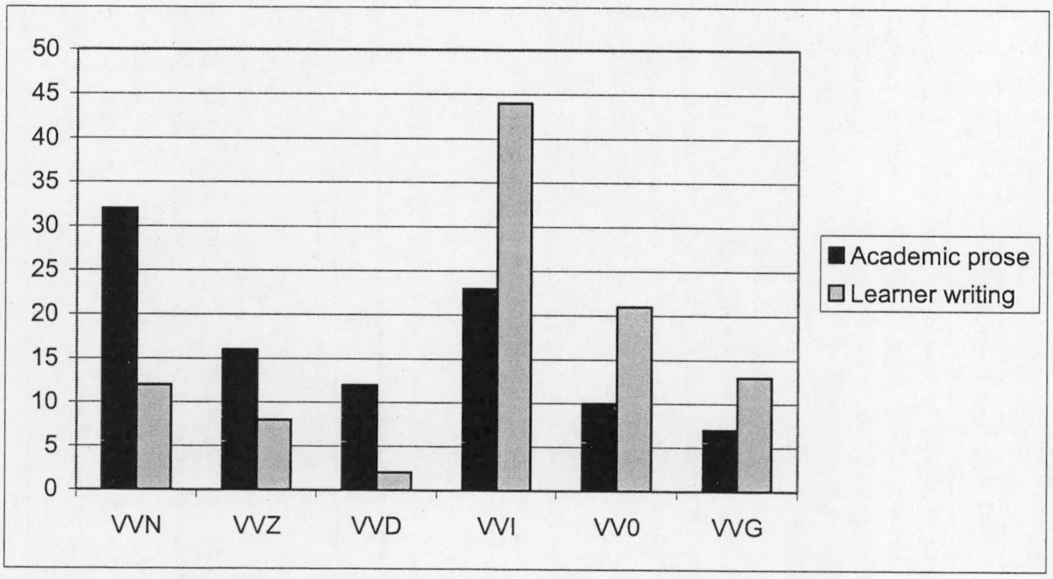
[156] AKL words are printed in bold in these examples.

Results also seem to support the widely held view that EFL learners' academic writing is characterized by "firmer assertions, more authoritative tone and stronger writer commitments when compared with native speaker discourse" (Hyland and Milton 1997:193) (see also Petch-Tyson 1998; Lorenz 1998; Neff et al 2004). EFL learners state propositions more forcefully and make a more overt persuasive effort: they **overuse communicative phrasemes** that serve as **attitude markers** (e.g. *it is very difficult to, it is very important to*) and **boosters** (e.g. *but it is true that, it is a fact that, it is obvious that*). By contrast, they underuse **hedges** such as *it is (more) likely that, it may be that, it seems likely that, it is possible that, it is unlikely that,* and *it would appear that.*

Word sequences used as **self mentions** are also much more frequent in learner writing than in professional academic prose (cf. Aijmer 2002; De Cock 2003; Ädel 2006). Examples include *therefore I, because I, I consider, we can, I can, in my view, I will discuss, provides us with*, and *from my point of view.* Conversely, academic writers use more clusters with 3rd person pronouns with an **evidential** function, e.g. *he remarks, he cites, his method, they suggest*, a difference which can be related to the more intertextual nature of professional academic texts (cf. Ädel 2006 and to appear, for a discussion of the influence of text type and intertextuality on academic writing).

It is worth noting that EFL learners underuse a whole set of word sequences involving the ***–ed* form of verbs**, and more precisely, their **past participle form**, which are typically used in professional academic prose. For example, they underuse the 2-word clusters *described as, suggested above, inferred from, listed above, discussed in* and *reported by*, the 3-word clusters *closely associated with, it was claimed, be ascribed to, when compared with, as noted above, is described in*; the 4-word clusters *can be related to, might have been expected, it is assumed that, may have been used*; and the 5-word clusters *it has been suggested that, it could be argued that, be defined in terms of,* and *be explained in terms of.* This is consistent with Granger's (2006) finding that past participle forms are the most frequent verb forms in academic prose but are highly underused in learner writing. As illustrated in Figure 6.14, Granger (2006) also shows that EFL learners massively overuse the base form (VV0) and the infinitive (VVI) form of lexical verbs, which can be related to our finding that learners use many more sequences involving the first person pronoun as subject followed by a lexical verb or an auxiliary (+ lexical verb), e.g. *I think, I can* and *I will discuss.*

Figure 6.14: Top 100 verb forms in academic prose and learner writing (Granger 2006)



As already underlined by Granger (2006), some verbs have similar frequencies as lemmas in learner writing and academic prose but display over- or underuse of some verb forms. Examples of AKL verbs following this pattern are *differ* and *discuss*. The lemmas do not display significant difference in use. However, *differ* is underused in its *–ing* form while *discuss* is overused in its unmarked form (*discuss*) and underused in its *–ed* form. Similarly, some verbs are underused or overused as lemmas but the general over- or underuse may not affect all verb forms. For example, *provide* is underused in learner writing when lemmas are compared but an analysis of word forms indicates that only *provided*[157] is underused while the other forms of the verb do not display significant difference in use. Table 6.34 shows that the picture can even be more complex: verb forms may be overused in specific lexical bundles while being underused in others. For example, the verb form *concerned* is overused in clusters such as *as I am concerned* and *concerned about* but underused in *been concerned with* or *we are concerned.* Similarly, EFL learners overuse the sequences *it allows* and *allows us to* and underuse the EAP sequence *allows for*. This shows that patterns of non-native usage are not limited to single lexical items which contravene the numerical native trend (cf. Lorenz 1999b:72).

[157] Note that the underuse of the form *provided* stands for an underuse of the *–ed* form of the verb *provide*; the conjunction *provided that* is not underused in ICLE.

Table 6.33: Examples of overused and underused clusters with AKL words

	Overused clusters	Underused clusters
2-word clusters	*for example, for instance, important to, main reason, opportunity of, therefore I, and therefore, have problems, are concerned, another important, mainly because, only because, quite clear, different reasons, totally different, different way, more difficult, great importance, very important, main cause, main problem, absolutely necessary, because I, negative consequences, real problem, great amount, good idea, I consider, great part, important part, big problem, best solution, allows us, conclusion I, different points, we can, can choose, it depends, good use, good example, real value, important question, important factor, I can, different problems*	*by contrast, in particular, was probably, a similar, the view, suggestion that, described as, suggested above, was effectively, still further, more generally, readily available, relatively few, more significantly, is ultimately, he concludes, on average, central issue, certain respects, radically different, consistent with, crucial importance, significantly different, extensive use, final analysis, they suggest, inferred from, listed above, general principles, inherent in, major source, particular attention, highly significant, by comparison, considerable degree, perhaps because, much emphasis, he cites, provide evidence, little evidence, central figure, in practice, reports that, allowing for, what appears, discussed in, may suggest, reported by, precisely because, crucial role, integral part, wide variety, they argued, partly because, somewhat different, almost entirely, he remarks, his method*
3-word clusters	*as a result, as a consequence, in my view, more and more, more or less, take into account, advantages and disadvantages, aim of this, pay attention to, as a conclusion, take into consideration, of great importance, it means that, affect our approach, people claim that, I will discuss, may say that, prevents us from, provides us with,*	*in terms of, the absence of, the view that, extent to which, the implications of, an account of, a theory of, in relation to, an attempt to, closely associated with, a considerable degree, as distinct from, high degree of, high proportion of, it seems likely, various forms of, a concern with, to this extent, despite the fact, the hypothesis that, the issue of, this need not, at any rate, by reference to, in certain respects, were subject to, in his view, in view of, it was claimed, it follows that, by showing that, this suggests that, be ascribed to, when compared with, as noted above, is described in*
4-word clusters	*the problem is that, it is very difficult, the fact is that, is the fact that, it is also true, there are also people, a great number of, it is high time, it is obvious that, as much as possible, it is true that, to a great extent, because of the fact, to answer this question, in order to achieve, it is necessary for,*	*it may be that, may well have been, to the effect that, are likely to be, would seem to be, to the extent that, with the exception of, it does not follow, it seems likely that, in the presence of, the edge of the, it was difficult to, the immediate aftermath of, it is possible that, can be related to, similar to that of, the total number of, it is unlikely that, a wide variety of, in the absence of, to the advantage of, on the assumption that, as an attempt to, on the basis of, in the belief that, might have been expected, with the exception of, the extent to which, was by no means, in the presence of, no reason to suppose, with the result that, it would appear that, it is assumed that, may have been used*

5-word	*from my point of view, far as I am concerned, there are more and more, it is very difficult to, but it is true that, this is not the case, as a matter of fact, it is very important to, it is a fact that, one of the most important*	*as in the case of, it has been suggested that, it could be argued that, in so far as they, it is more likely that, it is hardly surprising that, be defined in terms of, it is worth noting that, be explained in terms of*

Table 6.34: Overused and underused clusters with AKL verbs

Lemmas and over- or underused word forms	Overused clusters	Underused clusters
AFFECT (++) *affect* (++) *affects* (++)	*affect our, media affect, affect us, affects the, affect our approach, media affect our, mass media affect, affect our approach, media affect our approach, mass media affect our, affect our approach to reality, media affect our approach to*	*was affected, not affect the*
ALLOW (++) *allowed* (++)	*allowed to, not allowed, are allowed, not allow, be allowed, allow them, it allows, allows us, are not allowed, are allowed to, not allowed to, allow them to, be allowed to, allows us to, are not allowed to*	*allow them, allowed him, by allowing, allow it, allows for, to allow, allowing for, allow that, which allowed, allowed him, to allow for*
CONCERN (++) *concerning* (++)	*is concerned, are concerned, am concerned, concerned about, it concerns, concerning the, I am concerned, as I am concerned, far as I am concerned*	*was concerned, been concerned, concerned to, concerned with, we are concerned, been concerned with, was concerned with, concerned with the, is concerned with the*
DEPEND (++) *depends* (++) *depending* (++) *depended* (--)	*depends on, it depends, depending on, much depends, it depends on, depends on the, depending on the*	*depending upon, depended on, depended upon, will depend, depends upon the, depend upon the, will depend on*
DIFFER (//) *differing* (--)	-	*differed from, differs from the*
DISCUSS (//) *discuss* (++) *discussed* (--)	*will discuss, to discuss, I will discuss*	*was discussed, already discussed, and discussed, discussed below, in discussing, discussed in, discussed in chapter*
TEND (//) *tend* (++) *tended* (--)	*tend to, people tend, we tend, they tend, people tend to, we tend to, they tend to*	*they tended to, and tended to, has tended to, have tended to, tended to be*
PROVIDE (--) *provided* (--)	*provides us, provide us, provide them, can provide, provide us with, provides us with, provide them with*	*might provide, provide the, provides that, provide an, provide evidence, to provide, provide a, provides an, provides a, was to provide, to provide an, to provide a*

A **corpus-driven approach** to the phraseology of EFL learner writing is indispensable if we want to build up a full picture of all the possible lexical realizations of a rhetorical or organizational function in learner writing. It makes it possible to uncover a whole range of words and word sequences that are not recognized as typical of academic prose as they do not include an AKL word but which are nevertheless used by EFL learners to serve organizational or rhetorical functions (cf. section 5.7). Examples of **learner-specific sequences** that do not include an AKL word are given in Table 6.35. They include:

- Phrasemes that are more frequently used in **speech**, e.g. *of course, I think that, there are a lot of* (cf. section 6.3.2.2);
- Sequences that are **not used in English to establish the logical link** intended by EFL learners, e.g. *on the other side* (cf. section 6.3.2.4 on semantic misuse);
- Sequences that exist in English but are **very rare in all types of discourse**, e.g. the sequence *as far as I am concerned* which is repeatedly used to express one's opinion in ICLE.
- **'Unidiomatic' sequences** such as *according to me* used to express one's opinion in a restricted number of learner sub-corpora and *as a conclusion* used as a textual phraseme to introduce a conclusion (see section 6.3.2.3.2 for a co-occurrence analysis of the noun *conclusion* in ICLE).
- **Erroneous sequences** such as *in contrary, by the contrary, in the contrary, in contrary to* that are used to express a contrast in EFL learner writing.

EFL learners' overuse of rare native sequences such as *as far as I am concerned* or *last but not least* or 'unidiomatic' sequences such as *according to me* and *as a conclusion* may be partly explained by misguided teaching materials or/and L1 influence. Appendix 6.2.a shows a list of linking words published by Clairefontaine that French students are encouraged to use in the English test of the 'Baccalauréat', i.e. the final secondary school examination qualifying for university entrance, to 'enrich their essay and give more clarity to their argumentation'. It includes *as a conclusion* but does not list *in conclusion*[158]. The rare expression *as far as I am concerned* is also given as a key expression to voice one's opinion. Similarly, a web-page devoted to linking words and hosted by the 'Académie de Lille (Anglais BTS Informatique)' lists *according to me* as a direct translation equivalent of Fr. '*à*

[158] John Osborne kindly pointed out to me that the sequence *according to me* has also appeared in published textbooks such as *Ok*! (Lacoste and Marcelin, Nathan 1984), which was widely used in French colleges throughout the 80s and early 90s.

mon avis' as well as *as a conclusion* as a possible equivalent of Fr. 'pour conclure / pour résumer'. This web-page is reprinted in Appendix 6.2b.

Table 6.35: Examples of overused clusters in learner writing

	Examples
2-word clusters	*in sum, of course, in fact, is why, let us, I think, instead of, look at, we must, or maybe, really think, there are, my opinion, if you, but I, if we, there is, thanks to, we want, sure that, I believe, people say, people think, when I, said that, I agree, many things, no matter, means that, opinion is, I want, everybody knows, people often, let them, we look, I hope, at all, people believe, even worse, I really, so why, we think, people feel, we get, I guess, just imagine, think twice, quite sure, why we, I must, very serious, helps us*
3-word clusters	*in my opinion, in spite of, to sum up, first of all, I think that, in order to, I would like, that is why, on the contrary, I believe that, to my mind, we have to, all kinds of, I would say, we all know, people think that, if we want, it means that, by the way, a look at, on one hand, I am convinced, people believe that, I will try, I agree that, and of course, everybody knows that, many people think,*
4-word clusters	*on the one hand, last but not least, I would like to, some people say that, we can say that, in this essay I, are more and more, I am sure that, there are a lot, it is impossible to, I don't agree with, I want to say, but if we look, I am afraid that, it is easy to*
5-word clusters	*I do not think that, as a matter of fact, from my point of view, I would like to say, far as I am concerned, it seems to me that, I do not agree with, but at the same time, due to the fact that, I do not think so*

6.3.2.3.2. Preferred co-occurrences in EFL learner writing

In section 6.3.2.1, it was shown that EFL learners show a marked preference for a restricted set of single words and mono-lexemic phrasemes to express logical links. They also use **learner-specific functional equivalents** of these mono-lexemic items. As explained in section 6.3.2.3.1, they use the sequence *as a conclusion* instead of *in conclusion* as a textual phraseme to introduce a conclusion. This learner-specific phraseme represents 39.2% of the concluding textual phrasemes involving the noun *conclusion* in ICLE. In a longitudinal study of German learners' use of the noun *conclusion*, Mukherjee and Rohrback (2006) comment that the sequence *as a conclusion* is gaining grounds in learner writing to the extent that it is even more frequent than *in conclusion* in the more recent corpus they use:

> Interestingly, the most frequent phrase is no longer *in conclusion*, but *as a conclusion*. This certainly is a problematical development because *in conclusion* is much more frequent and idiomatic than *as a conclusion*, the latter being notoriously overused by German learners of English at university level as well. (Mukherjee and Rohrback 2006:224)

This development may be related to the increasing use of the internet for study purposes and of the type of teaching materials available on this channel (cf. section 6.3.2.3.1). Another example of a learner-specific functional equivalent is the use of *on the other side* instead of *on the other hand* to compare and contrast (cf. section 6.3.2.4 for more details on learners' use of *on the other side*).

In section 6.2.1.1, it was shown that mono-lexemic phrasemes such as *for example* have their own phraseological patterns in academic prose. However, these do not seem readily available to EFL learners who tend to **produce their own phraseological 'cascades'**, "collocational patterns which extend from a node to a collocate and on again to another node (in other words, chains of shared collocates)" (Gledhill 2000:212)[159]. Figure 6.15 shows that the textual phraseme *in conclusion* (or one of its learner-specific functional equivalents) is very often directly followed by the personal pronoun *I* in ICLE. This is consistent with Ädel's (2006) finding that personal metadiscourse, i.e. metadiscourse items that refer explicitly to the writer and/or reader, serves a wide range of rhetorical functions in Swedish learner writing, e.g. exemplifying, arguing, anticipating the reader's reaction, and concluding. The sequence *in conclusion, I* is then generally followed by the modal *would* to produce the word sequence *in conclusion, I would,* which, in turn, very often introduces the sequence *like to.* The sequence *in conclusion, I would like to* either introduces the verb *say* or another verb of saying such as *tell* or *mention.*

Figure 6.15: Phraseological cascades involving 'in conclusion' and learner-specific equivalent sequences

								say	6
								emphasize	2
In conclusion	59	*I*	37	*would*	21	*like to*	11	*tell*	1
As a conclusion	40		(36%)		(56.76%)		(52%)	*mention*	1
As conclusion	3							*speak about*	1
								reiterate	1
								quote	1

EFL learners prove to use AKL nouns and verbs **in different lexico-grammatical or phraseological patterns** than professional writers. This has already been illustrated by learners' use of the noun *example* and the verbs *illustrate* and *exemplify* in section 6.3.1. Another example is learners' use of the noun *conclusion.* Table 6.36 gives the verb co-occurrents of the noun *conclusion.* The percentage of verb co-occurrent types that are

[159] Gledhill (2000) uses the term 'collocational cascade' but we shall avoid using the adjective 'collocational' to refer to sequences of co-occurrents (cf. section 2.4.3)

significant co-occurrents of the noun *conclusion* in BNC-AC is 30.76%. Almost half of the verb co-occurrent types (46.2%) used in ICLE do not appear in BNC-AC. When tokens are analysed, the percentage of verb co-occurrents that are significant co-occurrents of the noun *conclusion* in BNC-AC rises to 75.8% as several of the verbs are repeatedly used in learner writing (e.g. *come to* and *draw*). Conversely, the percentage of verb co-occurrents that are not found in BNC-AC falls to 12% as 'non-native' co-occurrences are rarely repeated.

Table 6.36: Verb co-occurrents of the noun *conclusion* in ICLE

Verb + *conclusion* as object	**Freq in ICLE**	**Statistically significant co-occurrent in BNC-AC**[160]	**BNC-AC**[161]	***conclusion* as subject + verb**	**Freq in ICLE**	**Significant co-occurrent in BNC-AC**	**BNC-AC**
add up to	1	-	-	*emerge*	1	**	√
apply	1	-	-	*arise*	1	-	√
approach	1	-	-	*contain*	1	-	-
arrive at	5	**	√	*be*	23	**	√
bring	1	-	-	*come*	1	-	-
bring sb to	2	-	√	*need*	1	-	√
come to	52	**	√	*bring sb to*	1	-	-
**come into*	1	-	-				
confirm	1	**	√				
contain	1	-	-				
draw	25	**	√				
**draw up*	1	-	-				
end with	1	-	-				
escape	1	**	√				
express	1	**	√				
find	2	-	-				
gather	1	-	-				
get	1	-	√				
give	1	-	√				
have	1	-	√				
influence	1	-	√				
jump to	2	**	√				
lead to	4	**	√				
leave sb with	1	-	-				
look for	1	-	-				
make	11	-	√				
point to	1	*	√				
put	1	-	-				
put forward	1	-	-				
reach	3	**	√				
write as	1	-	-				
TOTAL	**128 tokens (32 types)**			**TOTAL**	**29 tokens (7 types)**		

[160] ** Significant co-occurrents in BNC-AC; - not significant co-occurrents in BNC-AC.

[161] √ : the co-occurrent appears in BNC-AC; - the co-occurrent is not found in BNC-AC.

Several of the verbs that are significant co-occurrents in BNC-AC form **collocations** with the noun *conclusion*. The verb *draw* has a weakened or delexical meaning and 'grammaticalizes' the agentive noun *conclusion*. The meaning of the support verb construction corresponds to that of the verb *conclude* (cf. section 2.4.1.2.2). Collocations involving the verbs *arrive at, come to, reach, jump to* or *lead to* and the noun *conclusion* can also be broadly described as equivalents of the verb *conclude* but these verbs have developed an abstract or figurative meaning in combination with *conclusion*, as is also the case for the verb *escape*.

EFL learners use the collocations *arrive at + conclusion, come to + conclusion, draw + conclusion, lead + conclusion* and *reach + conclusion*. However, they **do not always use them in native-like lexico-grammatical patterns** as illustrated in the following sentences. In sentences 6.129 and 6.130, the indefinite article *a* is used instead of the definite article *the*, which is always used in BNC-AC when the conclusion (underlined in the examples) is introduced by a *that*-clause. In sentence 131, the frequent phraseme *lead to the conclusion that* is used with the personal pronoun *us*, a pattern which is very rarely found in academic prose. In the context of EFL teaching/learning, these findings support Nesselhauf's (2005:25) argument that collocations should not be viewed as involving two lexemes only but should also include the other elements closely associated with them.

6.129. *However, when we consider all the pros and cons of fast food* ***we will certainly arrive at a conclusion that*** *it is not an ideal way of eating*. (ICLE-PO)

6.130. *And taking into consideration that Marx was a materialist* ***we can come to a conclusion that*** *he himself would be attracted by the advantages of television, and religion for him would remain the opium of the masses*. (ICLE-RU)

6.131. *To sums up, all I have mentioned before* ***lead us to the conclusion that*** *if our lifes were a little "easier" and we wouldn't be dominated by a world that is constantly changing, due to new techniques and industrialization, we could enjoy doing things as dream and imagine more frecuently.* (ICLE-SP)

The collocation *escape + conclusion* appears in two phraseological patterns in academic prose: '*it is difficult to escape the conclusion that*' and '*we cannot escape the conclusion that*'. The single occurrence of the collocation that appears in ICLE is used in the native-like lexico-grammatical pattern '*cannot escape the conclusion that*' but its subject is a nominal phrase headed by the noun *evaluation* (sentence 6.132).

6.132. *However, a more objective evaluation of the problem* ***cannot escape the conclusion that****, drug use and abuse have occurred in all civilisations all over the world, and that it is the criminalization of drugs that has created a much heavier burden on society.* (ICLE-DU)

In the native-like collocation *express + conclusion*, the verb *express* has acquired a semi-technical sense and means 'make something public'. It is mainly used in legal discourse and thus conveys a rather formal tone as illustrated in sentence 6.133. Its single occurrence in ICLE can be qualified as 'non-native like' as it appears with the first person singular pronoun *I* as subject and the possessive determiner *my* (example 6.134). It may be hypothesized that the learner who wrote this sentence has been influenced by the native-like co-occurrence '*express one's opinion/view*'.

6.133. *The Divisional Court* ***expressed its conclusion*** *in the following terms: ...* (BNC-AC-HUM)

6.134. *Finally, I wanted* ***to express my conclusions****.* (ICLE-SP)

All these examples also show that, like the textual phraseme *in conclusion*, the collocations involving the noun *conclusion* are often used with personal metadiscourse.

Other sentences are examples of EFL learners' attempts at using native-like collocations, which result in crude **approximations**. Thus, in sentence 6.135, the phrasal verb *draw up* is used in place of *draw* and in sentence 6.136, the preposition *into* replaces *to* and no article is used in an attempt at producing the native sequence '*came to the conclusion that*'.

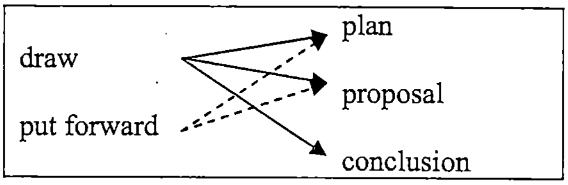
6.135. *Finally,* ***a conclusion can be drawn up*** *enphasizing our first statement, that is: technology, science and industrialisation have not killed dreams and imagination.* (ICLE-SP)

6.136. *The woman started to think about the price of progress and* ***came into conclusion that*** *automation causes more problems than it solves.* (ICLE-PO)

In sentence 6.137, the verb *put forward* is used with the noun *conclusion*. While this verb is commonly used with the nouns *plan* and *proposal*, two nouns that also combine with the verb *draw* to form collocations, the verb *put forward* is not used with the noun *conclusion* in native English (cf. Figure 6.16). This phenomenon is referred to as a **collocational overlap**, i.e. a set of nouns which have partially shared collocates, by Howarth (1996; 1998) (see also Lennon 1996).

6.137. ***Without putting forward premature conclusions,*** *we can nevertheless notice that a certain importance is granted to them.* (ICLE-FR)

Figure 6.16: A collocational overlap



The **semantic incongruity** of the co-occurrence 'put forward a conclusion' is made apparent by contrasting the definitions of *put forward* and *conclusion*. The verb *put forward* means 'to suggest an idea, explanation etc, especially one that other people later consider and discuss' (LDOCE4) while a *conclusion* is 'something you decide after considering all the information you have' (LDOCE4). Thus, a conclusion can hardly be put forward as it is supposed to be more than a suggestion and the result of serious consideration and discussion.

As already underlined by Nesselhauf (2005), EFL learners also produce deviant verb + noun nonce combinations. The noun *conclusion* enters in learner-specific V + N combinations that are not found in academic prose and which can be regarded as awkward nonce combinations on semantic grounds:

6.138. ***Looking for the conclusion*** *I would like to say that every person is individual and each has his or her own character.* (ICLE-RU)

6.139. *Having considered the various aspects of capitalism* ***a conclusion must be gathered:*** *the system cannot provide for the basic needs of the population; consequently it needs to take steps in order to prevent combativity which will endangered their interests.* (ICLE-SP)

The same remark can be made on several adjective + *conclusion* co-occurrences (cf. example 6.140). However, the distinctive feature of adjective + *conclusion* combinations in ICLE is that they are not the most typical combinations in academic prose even though a large proportion of them occur in BNC-AC (cf. Table 6.37). The first ten most significant adjective co-occurrents of the noun *conclusion* in BNC-AC are *general, logical, tentative, similar, foregone, main, firm, different, opposite*, and *definite*. None of these appear in learner writing except for *logical*. This reveals **learners' weak sense of what are native speakers' 'preferred ways of saying things'.**

6.140. *Looking at this idea from the Polish point of view, also brings* ***double standard conclusions****.* (ICLE-PO)

Table 6.37: Adjective co-occurrents of the noun *conclusion* in ICLE

Adjectives	Frequency	Significant co-occurrents of *conclusion* in BNC-AC[162]	BNC-AC
absolute	1	-	√
awful	1	-	-
certain	3	**	√
clear	1	**	√
clever	1	-	-
concrete	1	-	√
depressing	1	**	√
double standard	1	-	-
fair	1	-	√
false	1	-	√
final	5	**	√
frightening	1	-	-
interesting	1	-	√
liberal	1	-	-
logical	4	**	√
long-searched for	1	-	-
obvious	1	**	√
overall	1	**	√
only	1	-	√
own	4	**	√
personal	1	-	√
premature	1	-	√
private	1	-	-
radical	1	-	√
right	2	-	√
sad	1	-	-
same	2	**	√
satisfactory	2	**	√
satisfying	1	-	-
sensible	2	-	√
successful	1	**	√
terrifying	1	-	-
understated	1	-	-
unequivocal	1	-	√
wrong	1	-	√
TOTAL	**51 tokens (35 types)**		

The phraseology of EFL learner writing is also characterized by a number of **lexico-grammatical infelicities or errors**. Sentences 6.141 to 6.143 give examples of erroneous grammatical collocations with the AKL nouns *account, demand* and *possibility.*

[162] ** Significant co-occurrents in BNC-AC; - not significant co-occurrents in BNC-AC.

6.141. *Fifteen or twenty years ago an* **account *about** [of] *a murder of two girls by their mother would have called a scud of public resentment, but nowadays it's just a usual thing.* (ICLE-RU)

6.142. *And the stores are reacting to the* **demand *of** [for] *raw material ranging from flower, seeds, oil to different spices and exotic fruits.* (ICLE-GE)

6.143. *They should be given the* **possibility *to learn** **[of learning]** *a good job according to their natural abilities and to their preference and to acquire a certain amount of experience.* (ICLE-FR)

Sentences 6.144 and 6.145 illustrate learners' confusion between the prepositions *despite* and *in spite of*, which results in the blend **despite of* (cf. Dechert and Lennon 1989).

6.144. ***Despite of** [Despite] *the absence of such professionalism our nation overcame fascists.* (ICLE-RU)

6.145. *Therefore I also had to acknowledg that butterflies do not necessarily have to be better than caterpillars,* ***despite of** [despite] *the fact that they look nicer.* (ICLE-GE)

The following examples show that learners sometimes make use of the impersonal pronoun *it* in subject position after *as*:

6.146. *It is a matter of fact that these "things" cannot be bought and sold like shares on the stockmarket. Luckily, I would say because otherwise only the rich would be able to posses them* ***as it is** [as is] *unfortunately the case with many products in other areas of living.* (ICLE-GE)

6.147. *Because of the ambition for the power, their rivalry made them hold continuous battles,* ***as it was** [as was] *the case of Catholics and Protestants.* (ICLE-SP)

6.148. *All this proved that it is no good teaching people how to kill people (basically what the military service is about) and even worse is puting weapons at their disposal, by delivering permits for carrying firearms or by authorizing the legal sale of guns,* ***as it is** [as is] *the case in America (this point will not be discussed here, since it is outside the scope of this essay).* (ICLE-FR)

Another source of error is the adjective *same* which is sometimes preceded by the indefinite article in ICLE:

6.149. *Even within* ***a** [the] ***same** *ethnic group the crisis of the family determines substantial differences: between the Blacks, for instance, has stayed in relief that between those* *emigrants in America by now a normal family nucleus constitutes an exception, while the Etiopians or the West Indians live better for their strong family structure.* (ICLE-SP)

6.150. *The negative image of feminism makes it twice as hard for women to rise above it than it would be if men were facing* ***a*** **[the]** ***same*** *kind of dilemma.* (ICLE-FI)

6.151. *When different people read* ***a*** **[the]** ***same*** *book they have probably various imaginations while reading.* (ICLE-CZ)

Other examples of lexicogrammatical errors include *suggest *to* [V-*ing*], *related *with* [*to*], *attempt *of* [*to*], and *discuss *about* [ø]. It should be noted that very few of these errors are widespread in learner writing and that some of these may be partly L1-induced (cf. chapter 7).

6.3.2.4. Semantic misuse

> The misuse of logical connectives is an almost universal feature of ESL students' writing. (Crewe 1990:317)

In section 6.3.2.1, the function of **comparing and contrasting** has been shown to be generally underused in learner writing. An analysis of individual lexical items, however, reveals that the adverbials *on the contrary* and *on the other hand* are overused in ICLE. As Lorenz (1999b:72) has demonstrated, overuse is often accompanied by patterns of non-native usage. EFL learners' semantic misuse of the phraseme *on the contrary* has already been reported in the literature focusing on learners sharing the same mother tongue background:

> In Hong Kong, we are all familiar with students who use 'on the contrary' for 'however / on the other hand', thus adding an unintended 'corrective' force to the merely 'contrastive' function sought. (Crewe 1990:317)

Granger and Tyson (1996) report the same conceptual problems for French learners. Lake (2004) states that a large proportion of EAP non-native speakers who use *on the contrary* do so inappropriately, which is confirmed by our corpus-based analysis of EFL learners from different mother tongue backgrounds (see also Celce-Murcia and Larsen-Freeman 1999:534-535). The following sentences are examples of semantic misuse of *on the contrary*. EFL learners typically erroneously use *on the contrary* instead of a contrastive discourse marker such as *on the other hand* or *by contrast* to contrast qualities of two different subjects (underlined in the following examples). Thus, in example 6.152, the fact that Onasis had everything is contrasted with the fact that Raskolnikov had nothing and the phraseme *by contrast* would have been more appropriate.

6.152. *Raskolnikov differs from Onasis, of course. Onasis had everything but he wanted to have more. Raskolnikov,* ***on the contrary** [**by contrast**], *had nothing.* (ICLE-RU)

6.153. *The materialism, the base of this doctrine, denies the existence of spiritual substances and considers that the material is the only existing reality.* ***On the contrary** [**By contrast**], *the religion has been a link between the man and some superior powers during a long time.* (ICLE-SP)

6.154. *The young like crazy driving, overtaking and leading on the roads. Sports cars are created for this use and this may be the reason why their price is so high and use is expensive.* ***On the contrary** [**By contrast**], *station wagons are not expensive in maintenance. The main users of this kind of vehicles are families.* (ICLE-PO)

6.155. *For instance, most Americans have moved to the USA from different countries as immigrants.* ***On the contrary** [**By contrast**], *Europeans have lived in their countries for hundreds of years.* (ICLE-FI)

6.156. *As a result of all the burials, huge graveyards come into being and therefore, the Dutch government has decided that common graves will be cleared out ten years after the burials. It is easy to understand that if people are dug up after ten years, many people choose to be cremated.* ***On the contrary** [**By contrast**], *it is possible to be buried forever in a private grave, as long as someone pays for it, but a private grave is very expensive; the locals decide how much such a grave costs, resulting in increased prices.* (ICLE-DU)

The semantic inappropriacy of *on the contrary* in EFL learner writing has often been attributed to teaching practices. Teaching materials often provide lists of connectors in which the complex adverb *on the contrary* is described as a phrase of contrast, that is, as an equivalent alternative to *on the other hand, by contrast*, etc (cf. Crewe 1990). For pedagogical purposes, Lake (2004) proposes to use a checklist of contextual features that should be present in order to use *on the contrary*:

> As for the implications for learners, it now becomes possible to consult a checklist of contextual features that should be present in order for on the contrary to be appropriate:
>
> one subject;
> two contrasting qualities;
> one positive statement and one negative statement open to similar interpretations;
> an argument, either genuinely present or implied, to which the two statements, adjacent to the phrase both form a refutation.
>
> Such a checklist may be simplistic in that it does not cover all the possible lexico-syntactical environments in which the phrase might be encountered; but as a guideline for production, it ought to prove a useful starting point from which EAP teachers can devise their own practice materials. (Lake 2004:142)

Lake (2004) rules out the possibility of L1 influence on EFL learners' semantic misuse of *on the contrary* on the basis that over 70 per cent of international students from widely different mother tongue backgrounds produced two distinctly separate L1 equivalent items in a cloze test in which they were required to insert *on the contrary* or *on the other hand* and provide an equivalent phrase for both adverbials. It is, however, most probable that both misguided teaching practices and L1 influence interact here. The L1 equivalent forms to *on the contrary* and *on the other hand* may be characterized by different patterns of usage and thus be the source of negative transfer. Granger and Tyson (1996), for example, argue that French learners' overuse and misuse of *on the contrary* is probably due to an over-extension of the semantic properties of Fr. 'au contraire', which can be used to express both a concessive and antithetic link (cf. section 3.2.3.2 and section 7 for a discussion of L1 potential influence on French learners' use of *on the contrary*).

Lake (2004) also regards EFL learners' misuse of *on the contrary* as "something of an exception" and writes that "[i]n the EAP context, such functional phrases [connectives] are usually familiar to learners from an early stage, and do not pose great problems of usage" (Lake 2004:137). This view, however, is over-optimistic and is clearly not reflected in our corpus-based learner data. In section 6.3.1, EFL learners' inappropriate use of the abbreviation *i.e.* (in lieu of *e.g.*), the preposition *as* (instead of *such as*) and the adverb *namely* was discussed. Other examples of semantically misused lexical items include *on the other hand, on the other side, moreover, besides,* and *even if.*

Field and Yip (1992:25) report that *on the other hand* is frequently used by Cantonese speakers to make an additional point, with no implied contrast. They suggest that this semantic misuse may come from the same inappropriate use of an equivalent form in Chinese which is often misused by poor writers, taking it to mean 'another side or aspect' (cf. section 3.2.3.2). Although L1 influence may play a part in Hong Kong Chinese students' inappropriate use of the complex adverb, erroneous uses of *on the other hand* are found in most learner corpora, thus suggesting that there are other contributing factors to its semantic misuse. The following sentences are examples of inappropriate use of *on the other hand* in ICLE in which it would have been more appropriate to use no connector or to use an additive one:

6.157. *In this pragmatic society, material and practical things have limitted the development of trascendental things. We are too tide to reality, but this is so since we start facing life within society. Children are taught from the school according to fixed patterns and guidelines that every time are progressively more oriented to the Technology than to*

Humanistics. So, subjects such as Philosophy have been eliminatedfrom the plannings as compulsory , and have been replaced by others related to computers. [P][163] *Apart form the education children received, this overwheiming phenomenum, is aiso extended to their games and the way they play. For example, what do the children play with? Even toys are mechanized, the big majority of children have a computer or a video-game, with which spend (waste, in my opinion) a great number of hours.* [P] ***On the other hand*** **[In addition?],** *concerning industialization, there is the question of working. How much labour force has been omitted and substituted by machines? Definitely, rather than say that in our modern world there is no place for dreaming and imagination, we can rather say that we are deshumanizing our world progressively.* (ICLE-SP)

6.158. *However, when we think about the role and effects of massmedia upon the society, we shouldn't underestimate them, though,* ***on the other hand*** **[ø],** *people are mostly unconscious of the subtle influence TV and press have on them.* (ICLE-PO)

6.159. *I strongly believe that there is still a place for dreaming and imagination in our modern society.* [P] *Firstly, where there is a child, there are always dreams and imagination. Everybody knows that children like inventing funny stories and amusing plays by using their wide fantasy. This is one reason why children always bring happiness and awake the adults' childish part.* ***On the other hand,*** *fantasy is* **[also]** *a useful mean used by teachers in primary schools to teach school subjects to their little students. So, it is children who keep dreams and imagination alive! Therefore, as soon as there is a child in the world, dreaming and imagination will be possible too.* (ICLE-IT)

6.160. *The re-introduction of the death penalty may have positive sides, too. Criminality would be limited, because criminals would be afraid of the severe punishment.* [P] *This might be an illusion, because* ***on the other hand*** **[ø]** *the death penalty develops violence and is incompatible with the basic laws of humanity. What else will be the answer to violence but violence again?* (ICLE-GE)

6.161. *The function of punishment is to show that crimes are not acceptable or that they can solve any problems.* ***On the other hand*** *the aim of punishments is* **[also]** *to make the criminals obey the laws and show example to other's so that they will not follow the bad example and commit the same crime.* (ICLE-FI)

Unlike in native academic prose, the word combination *on the other side* sometimes appears in ICLE in places where a contrast seems to be the logical link intended by EFL learners, as illustrated in the following examples:

[163] [P] = new paragraph in learner writing

6.162. *The present attitude of the Twelve may support their fears and discourage them (the recent exclusion from the Community's conference on Polish issues). However, Poland cannot reply with isolation as the unification still remains the best solution to its problems.* ***On the other side,*** *all countries should understand that history and its consequences cannot divide the continent. The successful process of unification should be carried out with respect to nations' rights and without special privileges given to the powerful. Otherwise, Europe will never become a continent of success.* (ICLE-PO)

6.163. *In dramatic cases, for examples in families with very serious problems such as crimes, violence, intolerable relationships etc., it seems impossible not to conceive divorce as a good situation. In less problematic cases,* ***on the other side,*** *there is no point in continuing living with a no longer loved person (or totally indifferent), because it can create only more difficulties: in the past it was women's role to accept submittedly this situation, while nowadays this philosophy of self-sacrifice for the family's sake is no longer conceived, the right to happiness and to self-fulfilment has taken its place.* (ICLE-IT)

6.164. *Another big problem is our environment. There is pollution wherever you look. We can no longer enjoy the sun in summer because of the hole in the ozone layer. This hole is caused by technical impovements in the last decades. But* ***on the other side*** *it is sometimes hard to live without car or aerosols.* (ICLE-GE)

6.165. *Europe 92 means well a loss of identity since we'll be no longer Belgians, Italians, English ... but Europeans. But* ***on the other side*** *we will form a new nation with new hopes, new ideas ...* (ICLE-FR)

6.166. *The Iraqi leader Saddam Hussein is one the clearest example. If the world tolerate their crimes, another followers will appear.* [P] *In this sense I support the Gulf War.* ***On the other side*** *I can not familiarize myself with the idea, that it is glorious to fight for one's country.* (ICLE-CZ)

6.167. *I am not an enemy of alpha sciences. We need chemistry, for instance, to provide medication. I do object to the search for infinity. Let us all dream about paradise instead of looking for it. A perfect world with perfect people must be a very boring place. And* ***besides,*** *true knowledge is not to be found on your computer-screen; it is to be found in the heart.* (ICLE-DU)

There is also some confusion between the conjunctions *even if* and *even though* in EFL learner writing. Learners often use *even if* in lieu of *even though* to introduce a concession:

6.168. *However,* ****even if*** **[even though]** *I agree that the American public school system is defective, home schooling to me is no real alternative, as I feel that parents are not the best teachers for their own children.* (ICLE-GE)

6.169. *Although university degree is theoretical and does not really prepare for real life and world, they have their specific value. They give you the starting impetus for developing yourself and you decide if you want to continue or not.* ****Even if*** **[Even though]** *university knowledge is of very little use in real life and world and you will soon find out that you know almost nothing or just the basics for your work, it will motivate you to go on and you feel that you have at least very little, which is better then nothing.* (ICLE-CZ)

6.170. *As he writes: <*>. He was so opposed to industrializing that he preferred shipping the raw materials from America to Europe and then import them back as ready products* ****even if*** **[even though]** *he knew this would cost some extra money.* (ICLE-FI)

6.171. *The members of a community share a personal history, have a long tradition of customs which is still very vivid in their mind* ****even if*** **[even though]** *they do not always know the national anthem of their own country!* (ICLE-FR)

6.172. *In conclusion,* ****even if*** **[even though]** *our modern society is dominated by technology, computers and scientific experiments, we should always keep a place for dreams and imagination.* (ICLE-IT)

6.173. *We must forget about refrigerators containing CFC-11 and CFC-12,* ****even if*** **[even though]** *they are cheaper.* (ICLE-PO)

6.174. *We are as much a part of Europe as any other country here,* ****even if*** **[even though]** *we are not in the European Union.* (ICLE-SW)

Even if is used to introduce a condition, not a concession. Compare sentences 6.175 and 6.176:

6.175. ***Even if*** *these descriptions are valid they still leave open a number of questions, particularly why the same mechanisms do not operate with girls.* (BNC-AC)

6.176. ***Even though*** *these descriptions are valid they still leave open a number of questions, particularly why the same mechanisms do not operate with girls.* (BNC-AC)

In the second sentence, the writer knows and accepts that the descriptions are valid. In the first sentence, he or she does not.

Semantic misuse has often been discussed in the literature in relation to logical connectives. However, EFL learners also experience difficulty with the semantic properties of other types of cohesive devices, and more specifically, **labels**, i.e. abstract nouns such as *issue, argument,* and *claim* that are inherently unspecific and require lexical realization in their co-text, either beforehand or afterwards (cf. section 1.3.2.2). In reception, these nouns are likely to be problematic to EFL learners as they often refer to abstract ideas and processes and introduce additional propositional density to a text (cf. Corson 1997). Very few studies have investigated EFL learners' productive use of these nouns (cf. Flowerdew 2006 for an exception). In addition to phraseological and lexico-grammatical infelicities, EFL learners' use of labels is characterized by **semantic infelicity** or **lack of semantic precision**. Learners use the noun *problem* as an 'all purpose wild card' (cf. Lorenz 1999b) in lieu of more specific nouns such as *issue* or *question* as illustrated in the following sentences:

6.177. *This short discussion of the main points linked to the* ***problem*** **[issue]** *of capital punishment leads to the final question.* (ICLE-PO)

6.178. *The most important question conserning genetic engeneering is* ***the problem*** **[that]** *of gen manipulation with humans.* (ICLE-GE)

6.179. *If we are aware of the fact that such time-tables are very common for people living in a modern society like ours, the* ***problem*** **[question]** *of the place of imagination and dreaming is not even worth examining. Industrialisation has transformed dreaming into a waste of time which is now "cleverly" linked to money.* (ICLE-FR)

The noun *argument* also seems to cause difficulty to EFL learners. In sentence 6.180, the rather unidiomatic expression 'familiar arguments about' should be rephrased as 'widespread or popular beliefs about'. In sentence 6.181, the sentences that follow the label 'argument' are better described as 'reasons' why Big Tobacco did not depart from prepared statements.

6.180. *Female participation in making decisions concerning war and peace, economy and environmental protection would be to the benefit of all. However it will not be possible until males re-think and, hopefully, reject* ***familiar arguments*** **[widespread/popular beliefs]** ***about*** *women being unreliable, irrational and dependent on instincts. (ICLE-PO)*

6.181. *There are two main* ***arguments*** **[?reasons]** *that help us understand why Big Tobacco stuck to their statements for so long.* [P] *First, the companies feared the consequences that would follow a confession. They feared that there was going to be even more legislation and regulation if they would ever admit to lying. <*> and may be a reason for Congress to <*> says Martin Meehan Republican and cohair of the House tobacco task force <R>.* .... (ICLE-DU)

Other problematic labels include, among many others, *aspect* and *issue*. In sentence 6.182, 'another aspect' introduces a second example of the fact that "you are judged by what you do rather than by what you are", contrasting it with the first example about physicists and mathematicians. In sentence 6.183, 'in certain aspects' stands for 'in some respects' and 'the aspect of money' probably refers to the 'money issue' or the 'money question' in sentence 6.184.

6.182. *Our modern western society puts a lot of pressure on people as far as work is concerned. Your job is your "trademark". Or, in other words, you are judged by what you do rather than by what you are. Sad, but true. For example, according to popular opinion you must be very intelligent if you are a physicist or a mathematician.* ***And another aspect is that*** **[?by contrast,]** *the unemployed or housewives are sometimes treated as social outcasts.* (ICLE-GE)

6.183. *Actually, bits of information from the remotest parts of the globe reach us in an instant. Human beings can eventually feel as one great family, but only* ****in certain aspects*** **[in some respects]**, *for as far as real good relations among countries are concerned, it is still a matter of distant future.* (ICLE-PO)

6.184. *A legend exists that money was invented by the devil to tempt the mankind. The* ***aspect*** **[?issue/question]** *of money includes the problem of equality. There were and there are different ideas about making all people equal, because it was considered that this would lead to common happiness.* (ICLE-RU)

In sentence 6.185, it is not quite clear what ‘her issues’ refer to and in sentence 6.186, ‘issue’ most probably stands for ‘product’.

6.185. *Uta Ranke-Heinemann, the most famous woman in the field of Catholic theology, tries to provide answers to them. Her* ***issues*** **[?]** *lies on the verge of theology, philosophy and first of all, religion. She is employed in defining the relation between faith and the mind.* (ICLE-PO)

6.186. *The picture I draw from my dear old houseman admittedly is nothing but a mere cliché, a hyperbolic* ***issue*** **[product]** *of my vivid imagination.* (ICLE-GE)

6.3.2.5. Chains of connective devices

EFL learners’ texts are sometimes characterized by the use of too many connective devices (cf. Crewe 1990; Chen 2006; Narita and Sugiura 2006). The following text is an excerpt of an essay written by a French-speaking EFL learner. Each sentence contains at least one connective device – typically an adverbial connector or a sentence stem -, which is often found in sentence-initial position (cf. section 6.3.6).

<ICLE-FR-UCL-0092.3>
[1] **But what about** these prestigious institutions today? [2] To caricature them rapidly **one could say that** universities consist of courses given by professors (competent in their fields) in front of a silent audience who is conscientiously taking notes. [3] **So one can wonder if** a university degree really prepare students for real world and what his value is nowadays.

> [4] **I think it is true that** lectures in themselves are theoretical. [5] **Firstly because** students spend most of their time sitting in big classrooms which do not allow practical exercises **but only** ex cathedra lectures. [6] **Secondly because** the subjects of the lectures are theoretical. [7] **For example:** during a general methodology course (which, **we think,** could be more practical) different theories as Krashen's, Lado's or ? studied in detail **but** practical points are hardly ever considered.
>
> [8] **However is it true that** this formation does not prepare students for real world? [9] **I am of the opinion that** the answer is no. [10] **First I think that** university degrees are theoretical on purpose (**as opposed to** high schools which are more practical.). [11] **The reason is that, thanks to** the theoretical background they have learned, university students are able to build up their own way to achieve their aim.
>
> [12] **Moreover** they are also able to adapt or to modify their method according to the situation. [13] **To take the example of** a teacher again, **I could say that** a teacher in front of a classroom do not think about particular methodological theories again but that he has created his own methodology. [14] **Secondly, I think that** academic studies develop a critical mind. [15] The students are **indeed** trained to analyse pieces of information coming from different horizons from a critical point of view, **which means that** they have to dissect them, to confront them and then to be able to pass judgment on them. [16] **That is** the way they should create a personal opinion for themselves.
>
> [17] **Nevertheless, I do not want to** go too far. [18] **I really think that** theory is essential **but I am convinced that** practice should also be present. [19] **Let's take the example of** a student in economics who has his certificate in his pocket and proudly goes working in a big firm for the first time. [20] **I would compare** this business man to a gentleman who perfectly knows the highway code and who knows how to start and how to run through the gears but who finds himself in the center of Paris at the peak hours the first time he really drives! [21] **By this example, I want to show that** theory must always be accompagnied by practical applications, which is not often the case at university. [22] **I think that** this is a fully justified criticism against this institution.

Some EFL learners use many logical connectives between sentences simply to indicate to the reader that they are adding another point (e.g. *firstly, secondly, for example, first, moreover, to take the example of*). Several of these connectors are **superfluous** and sometimes **wrongly used** (e.g. *moreover* in sentence [12], *indeed* in sentence [15]). Crewe (1990) attributes EFL learners' massive overuse of connective devices to their attempt at imposing "surface logicality on a piece of writing where no deep logicality exists" (Crewe 1990:320). He adds that "[o]ver-use at best clutters up the text unnecessarily, and at worst causes the thread of the argument to zigzag about, as each connective points it in a different direction" (ibid 324). The following excerpt of an EFL learners' essay is a good example of EFL learners' use of logical connectors as 'stylistic enhancers', i.e. "words or expressions that may be sprinkled over a text in order to give it an 'educated' or 'academic' look" (Crewe 1990:316) but whose presence will not make the text coherent.

> 6.187. ***Furthermore***, *Hobbes is a stern determinist. He regards man, like nature, as subject to the chain of cause and effect.* ***Therefore*** *a concept like "free will" is impossible. Hobbes*

even *considers people as artificial creatures, not belonging to nature,* ***because*** *they are not able to live together in harmony, something which animals like bees and ants are capable of,* ***because*** *they are natural.* ***Of course,*** *these ideas were as much an insult to man's estimation of himself as Darwin's allegation, two hundred years later, that our ancestors used to live in trees.* ***As a consequence,*** *Hobbes was accused of being an atheist and forbidden to publish any more books.* (ICLE-DU)

As shown by Aijmer (2001) in a study of EFL Swedish student writing, learners use *I think* to make their claims more persuasive rather than to express a tentative degree of commitment. They often use *I think* or an equivalent expression (e.g. *I am of the opinion that, I am convinced that*) when it is communicatively unnecessary in the flow of argumentation. For example, sentence [18] in excerpt 6.186 could be rephrased as "Theory is essential but practice should also be present". The sequence *I think it is true* in sentence [4] corresponds to what Aijmer (2001) described as a 'rhetorical overstatement', which the author regards as typical of non-native speaker argumentative essays. The clusters *To me I think* and *as far as I am concerned* in sentences 6.188 and 6.189 respectively are other two examples of rhetorical overstatement.

6.188. ***To me I think*** *technology and imagination are very much interrelated,* ***and then on the other hand I understand that*** *they* ***also*** *can be seen as separate.* (ICLE-SW)

6.189. *I agree with George Orwell, because* ***as far as I am concerned I think that*** *in every country there are few people which are rich and many people which are poor.* (ICLE-IT)

The pedagogical implication of these findings is that, "important as these links are, learning when not to use them is as important as learning when to do so. In other words, students need to be taught that excessive use of linking devices, one for almost every sentence, can lead to prose that sounds both artificial and mechanical" (Zamel 1983:27).

6.3.2.6. Sentence position

Linking adverbials can occur in different sentence positions. They often occur initially, as do *however* and *in conclusion* in examples 6.190 and 6.191. They can also occur in medial position, i.e. within the sentence, often immediately after the subject, as shown in example 6.192. Final position is also possible as illustrated in sentence 6.193.

6.190. *In practice, the Red Army units did nothing to conciliate the Ukrainian Left or the peasants. Agriculture was brutally collectivized and no concessions were made in the use* *of the Ukrainian language and culture.* ***However,*** *Denikin's White armies counter-attacked and after seven months the Red Army was obliged to withdraw.* (BNC-AC-HUM)

6.191. <P>[164] ***In conclusion,*** *the population of England remained fairly stable for much of the fifteenth century, at a far lower level than in the first half of the fourteenth.* (BNC-AC-HUM)

6.192. *Coysevox's bust of Lebrun repeats -- again with a certain restraint -- the general outlines of Bernini's bust of Louis XIV. The face,* ***however,*** *shows a realism and subtlety of characterization that are Coysevox's own.* (BNC-AC-HUM)

6.193. *It 'd be worth asking him first,* ***though.*** (BNC-SP)

EFL learners' marked preference for sentence-initial position has been reported in various studies focusing on one L1 learner population (e.g. Field and Yip 1992; Lorenz 1999b; Zhang 2000; Narita and Sugiura 2006). Granger and Tyson (1996) comment that "[i]t is likely that this tendency for learners to place connectors in initial position is not language-specific" (Granger and Tyson 1996:24). Our analysis of connectors in ICLE supports this hypothesis. Table 6.38 shows that the total proportion of sentence-initial connectors in learner writing is much higher than that found in academic prose (13.17% vs. 6%). Examples include the preposition *despite* which appears in sentence-initial position in 52% of its occurrences in ICLE vs. 34.5% in BNC-AC-HUM (cf. example 6.194) and sentence-initial *due to* which is repeatedly used in learner writing but hardly ever occurs in academic prose (example 6.195).

6.194. ***Despite*** *its commercial character Christmas still means a lot to me.* (ICLE-FI)

6.195. ***Due to*** *these developments the production expanded enormously, which meant that a greater number of people could be fed.* (ICLE-DU)

Another example is the adverb *therefore* which often appears in sentence-initial position in ICLE but is not often used in that position in BNC-AC-HUM:

6.196. *Scientific research as well as individual observations prove that eating habits have a great impact on the condition of the human body and soul and, consequently, on rest, sleeping and even dreams.* ***Therefore*** *people should pay more attention to what they consume.* (ICLE-PO)

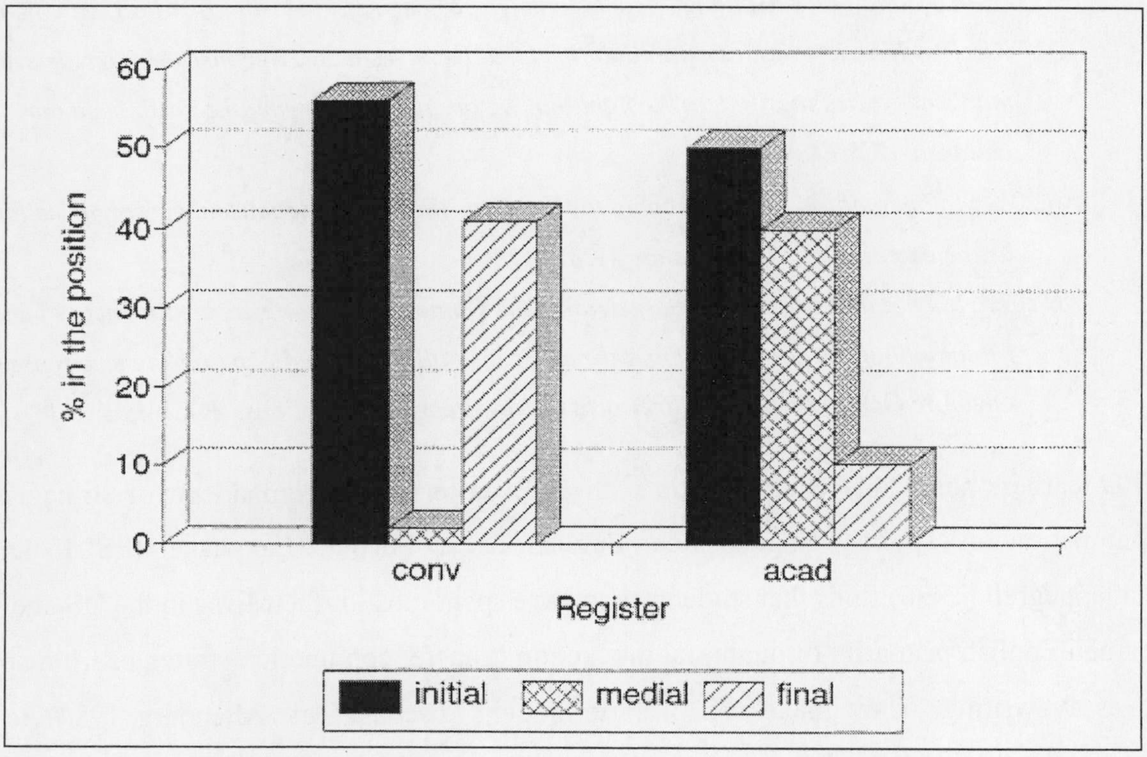
These findings provide evidence for EFL learners' lack of knowledge of the preferred syntactic positioning of connectors in English, which has often been attributed to **L2 writing instruction**. Flowerdew (1993) argues that teaching materials do not provide students with

[164] <P>: new paragraph

authentic descriptions of syntactic patterns of words. He shows that, contrary to what is often taught in course books, the adverbial connector *then* rarely occurs in sentence-initial position, but is more usually found in medial position. Similarly, Milton (1999:225) discusses the problematic aspects of teaching connectors by means of lists of undifferentiated items and suggests that one way instruction may skew EFL learners' style is "by the presentation of these expressions as if they occurred in **only**[165] sentence-initial position" (cf. also Narita and Sugiura 2006). Thus, EFL learners' tendency to place connectors in unmarked sentence-initial position seems to be **reinforced by teaching** (cf. Granger 2004:135).

Unmarkedness provides another possible explanation for EFL learners' massive overuse of sentence-initial connectors. Conrad (1999) studies variation in the use of linking adverbials across registers. She shows that, in both conversation and academic prose, the highest percentage of linking adverbials appears in sentence-initial position (cf. Figure 6.17 copied from Conrad 1999:13) and concludes that "[i]nitial position seems the **unmarked position**[166] for linking adverbials" (Conrad 1999:13) (cf. also Biber et al. 1999 and Quirk et al. 1985). EFL learners seem to use unmarked sentence-initial position as a safe bet.

Figure 6.17: Positions of linking adverbials in conversation and academic prose (Conrad 1999:13)



[165] My emphasis.
[166] My emphasis.

Contrary to our expectations, the proportion of sentence-initial *because* is lower in learner writing than in professional writing. However, sentence-initial *because* is significantly more frequent (relative frequency of 9.18 in learner writing vs. 4.54 in academic prose). It is also used to **serve different functions** in learner writing. In academic prose, sentence-initial *because*-clauses are attached to a main clause. As shown in the following examples, they introduce the cause of something that is described in the main clause:

6.197. ***Because*** *these changes were worldwide, Europe's history is inseparable from world history between 1880 and 1945.* (BNC-AC-HUM)

6.198. ***Because*** *the death-rate was high, marriages were usually short-term.* (BNC-AC-HUM)

Unlike expert writers, EFL learners sometimes use sentence-initial *because* to introduce new information in independent segments and give the cause of something that was referred to in the previous sentence:

6.199. *The crime rate would also strongly reduce and this is of course the main objective of all this measures.* ***Because*** *everybody wants to live in a safe society.* (ICLE-DU)

6.200. *To directly try to change people with "experience of life" would, at best, only be to win Pyrrhic-victories, compared to this effective investment.* ***Because*** *deep inside every man's heart lies the "Indian"-insight that we are only borrowing the earth from our children.* (ICLE-SW)

6.201. *In my opinion it is useful only for them, for their trial.* ***Because*** *their sorrow is found as the extenuating circumstance.* (ICLE-CZ)

6.202. *It's not that I am completely against the Games, but there is a certain uneasy feeling if I think about the costs and about the sense of it all.* ***Because*** *I can't make more sense of Olympic Games than of the fireworks at New Years Celebrations.* (ICLE-GE)

EFL learners share this characteristic with ESL writers. In a comparison of strategies for conjunction in spoken English and English as a Second Language (ESL) writing, Schleppegrell (1996) finds that students who have spent most of their lives in the US and have learned English primarily through oral interaction, transfer conjunction strategies from speech to essay writing. They make use of 'afterthought' *because* (cf. Altenberg 1984) to add information in independent segments as well as of other types of speech-like clause combining strategies.

Conrad (1999) reports that, in academic prose, most linking adverbials are placed in sentence-initial or medial position. Three types of **medial position** are particularly frequent (cf. Conrad 1999:14-15):

1. Linking adverbials occur immediately after the subject as illustrated in sentence 6.192 above.
2. Linking adverbials occur between an auxiliary and the main verb:
   All estimates of population size must ***therefore*** *allow for a large measure of conjecture, a fact stressed by all reputable modern historians who have worked on this intractable subject.* (BNC-AC-HUM)
3. Linking adverbials occur between the main verb and its complement:
   It is difficult to believe ***therefore*** *that one of these mosaics was not influenced by the other.* (BNC-AC-HUM)

Medial position of connectors is quite typical of academic prose. However, it is clearly less favoured by EFL learners. As underlined above, teaching materials focus on sentence-initial position and EFL learners most probably feel unsafe about the syntactic positioning of connectors.

Table 6.38: Sentence-initial position of connectors in BNC-AC-HUM and ICLE

	ICLE				BNC-AC-HUM			
	S-I	Total freq.	%	Rel. freq. pmw	S-I	Total freq.	%	Rel. freq. pmw
although	263	522	50.38%	225.6	676	2,276	29.7%	203.5
and	1456	32,236	4.5%	1249	1374	91,306	1.5%	413.6
as a result	71	103	68.9%	60.9	65	102	63.7%	19.6
as a result of	24	79	30.38%	20.6	22	194	11.34%	6.6
as far as X is concerned	96	167	57.48%	82.4	31	59	52.5%	9
because	107	2,493	4.29%	91.8	151	2,207	6.79%	45.4
because of	62	530	11.7%	53.2	46	599	7.67%	13.8
consequently	103	179	57.54%	88.4	60	143	41.96%	18
despite	50	96	52%	42.9	235	681	34.5%	70.7
due to	29	246	11.78%	24.9	3	195	1.54%	0.9
even if	83	274	30.29%	71.2	94	451	20.84%	28.2
even though	46	127	36.22%	39.5	28	248	11.29%	8.4
for example	235	854	27.52%	201.6	233	1263	18.45%	70
for instance	93	344	27%	79.8	86	609	14.12%	25.9
furthermore	113	127	96.58%	96.9	176	217	81.1%	53
however	673	1,128	59.66%	577.4	882	3,353	26.3%	265.5
in spite of	47	106	44.34%	40	42	159	26.4%	12.6
moreover	255	292	87.33%	218.8	365	495	73.73%	109.9
nevertheless	170	250	68%	145.8	392	676	57.99%	118
on the contrary	92	164	56.1%	78.9	48	95	50.53%	14.4
on the other hand	228	418	54.54%	195.6	155	372	41.67%	46.7
so	805	1,436	56%	690	675	1,894	35.64%	203.2
thanks to	68	199	34.17%	58..3	5	35	14.28%	1.5
therefore	340	689	49.35%	291.7	75	1,412	5.31%	22.5
thus	221	446	49.55%	189.6	756	1,767	42.78%	227.6
TOTAL	**5,730**	**43,505**	**13.17%**	**4,916.24**	**6,675**	**110,808**	**6%**	**2009**

Table 6.39 shows that, in ICLE, several connectors are repeatedly used in sentence-final position, which is quite uncommon in BNC-AC-HUM. As shown in Figure 6.17, final position is frequent in **conversation** but rare in academic prose. Conrad (1999), however, shows that only three highly-frequent items – *then, anyway* and *though* - account for the relatively high proportion of sentence-final linking adverbials in native conversation. She argues that these linking adverbials are commonly found in sentence-final position as they serve important interpersonal functions:

> [A]dverbials in conversation, in addition to showing a link with previous discourse, can also play important roles in the interpersonal interaction that takes place. These roles are often particularly noticeable for the common adverbials in final position. (…), final *though* often occurs when speakers are disagreeing or giving negative responses, final *anyway* is often associated with expressions of doubt or confusion, and (…) *then* typically indicates that a speaker is making an inteference (sic) based on another speaker's utterance. The placement of these adverbials in final position is consistent with previous corpus analysis of conversation that has found that elements with particular interpersonal importance are often placed at the end of a clause (…). It may be, then, that **in some cases in conversation there is a tension between placing the linking adverbial at the beginning of the clause, due to its linking function, and at the end of the clause, due to its interactional function.**[167] (Conrad 1999:14)

The type of interpersonal interaction that takes place in conversation is not typical of academic prose. Thus, none of the linking adverbials commonly associated with final position in conversation are common in formal writing. These findings suggest that positions of linking adverbials in native discourse are directly influenced by the **register** in which they appear, and the **textual and/or interpersonal functions** they serve.

Table 6.39: Sentence-final position of connectors in ICLE and BNC-AC-HUM

	ICLE				BNC-AC-HUM			
	S-F	**Tot. freq.**	**%**	**Rel. freq.**	**S-F**	**Tot. freq.**	**%**	**Rel. freq.**
anyway	25	132	18.94%	2.15	20	71	28.2%	0.6
for example	63	854	7.38%	5.4	20	1263	1.58%	0.6
for instance	31	344	9.01%	2.69	8	609	1.31%	0.24
indeed	15	257	5.84%	1.29	18	1413	1.27%	0.54
of course	34	750	4.5%	2.92	14	863	1.62%	0.42
then	35	1054	3.32%	3	17	3062	0.5%	0.5
though	11	256	4.3%	0.9	7	178	0.9%	0.2

[167] My emphasis.

6.3.2.7. Other features

Other learner-specific features include spelling errors and punctuation errors. Tables in Appendix 6.1 show that AKL words are sometimes **misspelt** in ICLE. For example, the word form *consequences* is spelt *consequenses, consecuences, consecuenses, consequencies* and *consecvencies* and *difference* appears in ICLE spelt as *diference, differance, differece, differency, differene* and *diffrence*. Note, however, that misspelt academic words constitute rare events and that a misspelt form is generally not repeated. Mispelt AKL words appear in all learner corpora, which suggests that academic words are not only difficult to master at the semantic and morphological level (cf. Corson 1997; Schmitt and Zimmerman 2002) but also at the orthographic level. On the other hand, proportions of misspelt AKL words vary significantly across learner populations with the highest proportion being found in the Spanish sub-corpus. This finding is supported by Lefer and Thewissen's (2007) comparison of spelling errors in the Dutch, French and Spanish corpora in which the highest proportion of spelling errors is also found in the Spanish sub-corpus.

Neff et al. (2007:208) report that **punctuation** accounts for 12% of the Spanish EFL errors and 10% of the other L1 populations. They comment that punctuation is, "most probably, grossly under-taught" in the EFL classroom. As a result, EFL learners produce run-on sentences, i.e. two or more complete sentences which are joined with a comma or without any punctuation marks or conjunctions (cf. example 6.203). They also forget commas after sentence-initial subordinate clauses or connectors (cf. example 6.204) or before and after appositives such as *that is* and *that is to say* (cf. example 6.205). By contrast, they sometimes erroneously use a comma after the conjunctions *although* or *(even) though* (cf. examples 6.206 and 6.207).

6.203. *Some time ago we used to enjoy various kinds of entertainment, we would have hobbies, go outside to theatres, cinemas, restaurants and sport events.* (ICLE-PO)

6.204. ***However** their advertising boasts that their products are of the best quality.* (ICLE-SW)

6.205. *According to von Mayer, however, what matters is *relative poverty* **that is to say** the sudden decrease of wealth.* (ICLE-IT)

6.206. *When I compare these languages I do not consider English as an easy language, **although,** I do admit that I have noticed some things that are easier about English than about the other languages that I had the chance to learn.* (ICLE-PO)

6.207. ***Even though,*** *the number of female and male births are roughly the same (slightly more females), more males die in every age group, form foetus to ninety years old.* (ICLE-SP)

6.3.3. Discussion

This section has shown that **academic, and more precisely, argumentative, essays written by upper-intermediate to advanced EFL learners share a number of linguistic features irrespective of learners' mother tongue backgrounds or language families**. The focus of our analysis has been on the lexical means available to learners to serve specific rhetorical and organizational functions. This textual dimension is particularly difficult to master and has been described by Perdue (1993) as the last developmental stage before bilingualism in second language acquisition. Results show that the expression of rhetorical and organizational functions in EFL writing is characterized by:

- **Limited lexical repertoire**: EFL learners tend to massively overuse a restricted set of words and phrasemes to serve a particular rhetorical function and to underuse a large proportion of all the lexical means available to expert writers. They also prove to prefer using conjunctions, adverbs and prepositions to the detriment of phraseological patterns with nouns, verbs and adjectives.
- **Lack of register awareness**: Texts produced by EFL learners often "give confusing signals of register" (Field and Yip 1992:26) as they display mixed patterns of formality and informality. The frequency of informal words and phrases in learner writing is often closer to their frequency in speech than in academic prose.
- **Lexico-grammatical and phraseological specificities:** EFL learning writing is distinguishable by a whole range of lexico-grammatical patterns and co-occurrences that differ from academic prose in both quantitative and qualitative terms. Preferred co-occurrences in ICLE are often not the same as in academic prose, which reveals learners' weak sense of what are native speakers' 'preferred ways of saying things'. Learners' attempts at using collocations are not always successful and sometimes result in crude approximations and lexico-grammatical infelicities. Results also support Lorenz's (1999b) remark that "advanced learners' deficits are most resilient in the area of lexico-grammar, where lexical items are employed to signal grammatical and textual relations" and that "a lack of coherence in advanced learners' writing must at least partly be attributable to lexico-grammatical deficits" (Lorenz 1999b: 56).

- **Semantic misuse**: EFL learners not only experience difficulty with the semantics of connectors but also with other types of cohesive devices, and more specifically, with labels, i.e. abstract nouns that are inherently unspecific and require lexical realization in their co-text, either beforehand or afterwards.
- **Chains of connective devices**: EFL learners' texts are sometimes characterized by the use of superfluous (and sometimes semantically inconsistent) connective devices.
- **Marked preference for sentence-initial position of connectors:** Connectors are often used in unmarked sentence-initial position in learner writing. Medial position is not favoured by EFL learners although it is typical of academic prose.
- **Misspellings of academic words:** A few misspelt academic words were found in ICLE but their frequencies are very low, which suggests that only a few learners still experience difficulties with spelling. In addition, misspellings are very unevenly distributed across L1 sub-corpora.
- **Punctuation errors and infelicities**

The methodology used in this chapter allows researchers to draw a general picture of the writing of upper-intermediate to advanced EFL learners from different mother tongue backgrounds, thus **avoiding hasty interpretations in terms of L1 influence**. Consider the following quotations by Zhang (2000), who attributes a number of features to the influence of the learners' mother tongue, that is, Chinese:

> The overuse of this expression [*more and more*] was most probably due to language transfer since a familiar expression in the Chinese language *ye lai yue* was popularly used. (Zhang 2000:77)
>
> The reason for the initial positioning of conjunctions was again due to the transfer of the Chinese language where conjunction devices with similar meaning are mostly used at the beginning of a sentence. (Zhang 2000:83)

As already mentioned, the sentence-initial positioning of conjunctions is common to most learner populations. The mother tongue may reinforce learners' preference for sentence-initial position but cannot be regarded as a unique explanation for this learner-specific feature. In section 6.3.2.6, teaching-induced factors have been proposed as possible explanations for learners' preference for sentence-initial position. Syntactic positioning of connectors is rarely taught and EFL learners use sentence-initial position as a safe position. As for the overuse of the expression *more and more*, although it is indeed very significant in the Chinese component of ICLE, this feature is actually common to all learners in ICLE. This suggests that, while transfer may be at work in the case of Chinese learners, it cannot be the only explanation and other factors, such as developmental factors, or teaching-induced effects, have to be taken into account too.

Another advantage of the method used is that, once linguistic features of upper-intermediate to advanced EFL learner writing have been highlighted, we can check to what extent they are specific to EFL learners or just typical of **novice writing**. This is precisely where a corpus of essays written by native university students such as STUD-US-ARG (cf. section 4.1.2.2) has a role to play, in tripartite comparisons between professional writing, foreign learner writing and native student writing, which make it possible to distinguish between learner-specific and developmental features. Whether a feature is learner-specific or developmental varies from lexical item to lexical item, but as a general rule, findings suggest that what is most likely to be shared by native and non-native novice writers is a **lack of register-awareness**. Figure 6.18 shows that a whole range of lexical items that Gilquin and Paquot (2006) found to be overused in learner writing - *maybe, so* expressing effect, *it seems to me, really,* sentence-final *though, this/that is why, I think* and *first of all* - are also often more frequently used in native novice writing than in academic prose (cf. examples 6.208 to 6.210). The overuse of *I think* in both learner and native student writing has already been reported by Neff et al. (2004) who describe it as a general "novice-writer characteristic of excessive visibility" (Neff et al 2004:152).

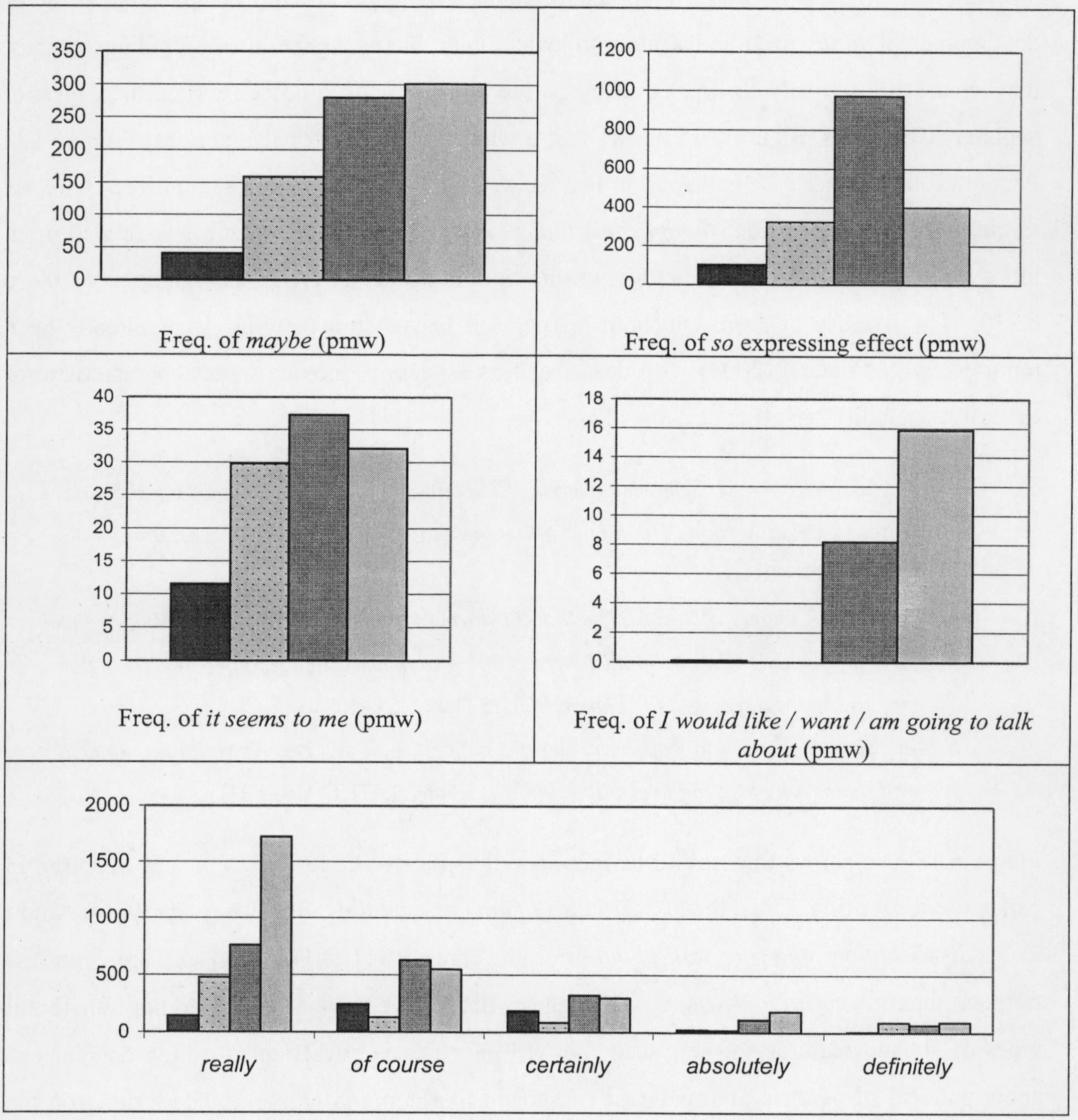
6.208. *Judge Robert H. Schnacke overruled the ban and allowed for reporters to be in the witness execution room. He did not allow them to bring in video cameras,* ***though.*** (STUD-US-ARG)

6.209. ***I think*** *society should realize that this form of punishment is being applied to those for doing exactly what the punishment is doing to them, which totally contradicts the morals involved (respect for human life).* (STUD-US-ARG)

6.210. *Shriver goes on to state that religious majorities do not respect religious minorities* ***and this is why*** *praying in public schools is wrong.* (STUD-US-ARG)

Figure 6.18 also shows that not all learner-specific spoken-like lexical items are overused in native novice writing. Thus, the lexical items *of course, certainly, absolutely, by the way* and *I would like/want/am going to talk about* are quite rare in STUD-US-ARG and are even less frequent than in academic prose, which suggests that native novice writers do not transfer all types of spoken features to their academic writing. It is very difficult to draw conclusions about patterns of overuse and underuse of spoken-like items in native novice writing on the basis of Figure 6.18 but it may be hypothesized that lexical items which are not so frequent in speech and rare in academic prose (e.g. *I want/would like/am going to talk about*) are less likely to be overused by native novice writers. By contrast, lexical items that are very frequent in speech and acceptable in academic prose are most likely to be overused (e.g. *maybe, so* expressing effect). These findings must be taken with caution as the native student corpus used, i.e. STUD-US-ARG, is arguably too small to be compared to the whole ICLE. Learner-corpus research is clearly in need of a larger comparable corpus of native novice writing (cf. Nesi et al. 2004).

Figure 6.18: Features of novice writing



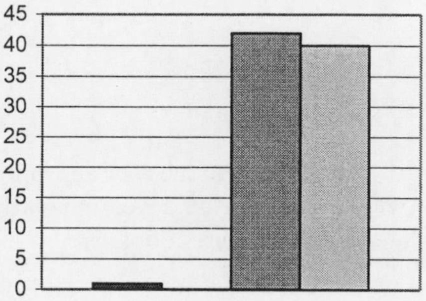
Freq. of *maybe* (pmw)

Freq. of *so* expressing effect (pmw)

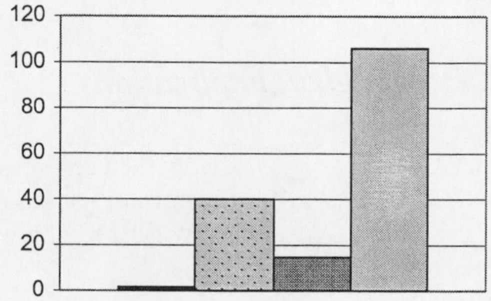
Freq. of *it seems to me* (pmw)

Freq. of *I would like / want / am going to talk about* (pmw)

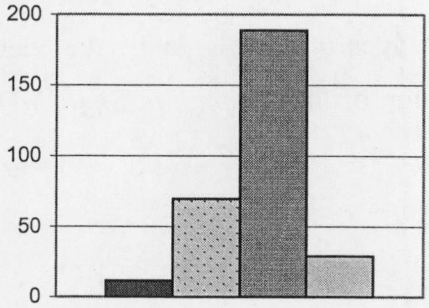
Freq. of amplifying adverbs (pmw)



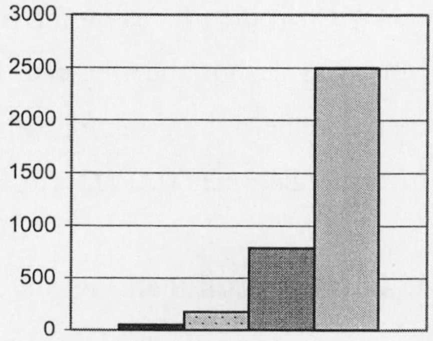
Freq. of *by the way* (pmw)



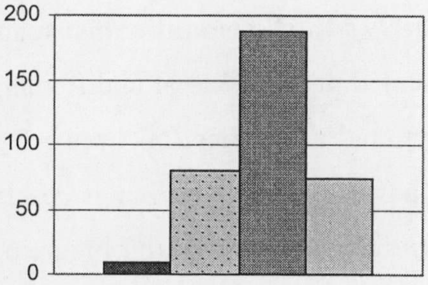
Freq. of sentence-final *though* (pmw)



Freq. of PRO (*this, that, which*) *is why* (pmw)



Freq. of *I think* (pmw)



Freq. of *first of all* (pmw)

- **Academic writing**: British National Corpus, academic component (15m. words) ■
- **Novice writing** : Sub-corpus of LOCNESS (100,702 words) ▨
- **Learner writing:** ICLEv2 (14 L1s; around 1.5m. words) ▩
- **Speech**: British National corpus, spoken component (10m. words) ▒

Other linguistic features are limited to non-native learners. Learner-specific features include **lexico-grammatical errors** (**a same; possibility *to; despite *of; discuss *about*); the use of **non-native like sequences** (e.g. *according to me* and *as a conclusion*); and the **overuse of relatively rare expressions** such as *in a nutshell.* As Gilquin, Granger and Paquot (in press) have argued, the issue of the degree of overlap between novice native writers and non-native writers has far-reaching methodological and pedagogical implications and is clearly in need of empirical studies.

6.4. Pedagogical implications

> How then could we raise our EFL learners' awareness of appropriate connector usage? One possible way is through the development of new EFL teaching materials. (Narita and Sugiura 2006:35)

The findings presented in this chapter have major pedagogical implications which are discussed in detail here. We will first place emphasis on a number of teaching-induced factors. We will then focus on the role of corpora, and more particularly, learner corpora, in EAP material design and illustrate how these two types of corpus data have been used to inform academic writing sections in the second edition of the *Macmillan English Dictionary for Advanced Learners* (MED2) (Rundell 2007).

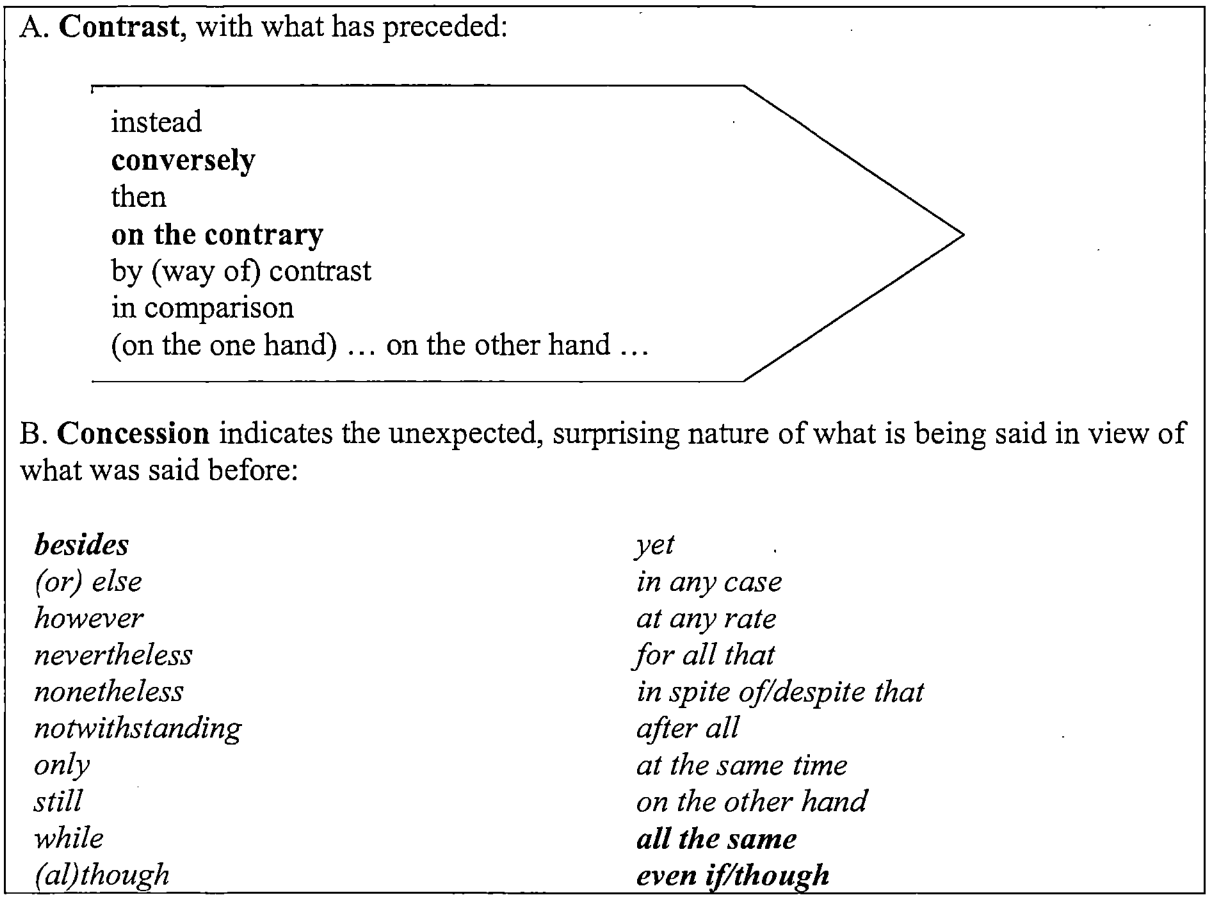
6.4.1. Teaching-induced factors

Teaching-induced factors have repeatedly been denounced in the literature as being responsible for a number of learners' inappropriate uses of connectors (cf. Zamel 1983; Hyland and Milton 1997; Flowerdew 1998; Milton 1999). First, **semantic misuse** may result from pedagogic practice as connectors are often presented in long lists of undifferentiated and supposedly equivalent items classified in broad functional categories (cf. Crewe 1990; Lake 2004). For example, in Jordan (1999), the complex adverb *on the contrary* is described as a phrase of contrast, thus equivalent to *on the other hand* and *by contrast* (cf. Figure 6.19). The same is true for *conversely*, which, however, should only be used for indicating that one situation is the exact opposite of another as in the following sentence:

6.211. *American consumers prefer white eggs;* ***conversely,*** *British buyers like brown eggs.* (LDOCE4)

Also problematic are the categorization of *besides* as a marker of concession and the misleading presentation of the conjunctions *even if* and *even though* as synonyms.

Figure 6.19: Connectives: contrast and concession (Jordan 1999:136)



A. **Contrast**, with what has preceded:

instead
conversely
then
on the contrary
by (way of) contrast
in comparison
(on the one hand) ... on the other hand ...

B. **Concession** indicates the unexpected, surprising nature of what is being said in view of what was said before:

besides
(or) else
however
nevertheless
nonetheless
notwithstanding
only
still
while
(al)though

yet
in any case
at any rate
for all that
in spite of/despite that
after all
at the same time
on the other hand
all the same
even if/though

Overuse of connectors such as *nevertheless, in a nutshell, as far as I am concerned, on the one hand,* and *on the other hand* can also be attributed to the long lists of connectors found in most textbooks (cf. Granger 2004:135) as no information is given about their frequency and semantic properties. Milton (1998) has shown that there is a strong correlation between the words and phrases overused by Hong Kong students and the functional lists of expressions distributed by Hong Kong tutorial schools, i.e. private institutions which prepare most high school students for English examinations. The **selection** of connectors to be taught may also lend itself to criticism. It was shown in section 6.3.2.3 that non-native like sequences such as *according to me* and *as a conclusion* are sometimes found in teaching materials, especially in lists of connectors freely available on the Internet[168]. In addition, the most frequent connectors to serve a rhetorical function are not always given and less

[168] The quality of teaching materials focusing on connectors that are freely available on the Internet is generally quite alarming, especially considering the fact that students increasingly use the Internet for study purposes.

idiomatic sequences are provided instead. Thus, in the *Fiche Essentielle du Baccalauréat*, connectors such as *first, second, in conclusion* and *in summary* are not provided and learners are encouraged to use rather unidiomatic sequences such as *as an introduction, as a conclusion* and expressions such as *in a word* which are less frequent (cf. Appendix 6.2a).

Another direct consequence of these lists is EFL learners' **stylistic inappropriateness** as Milton (1998) explains:

> Students are drilled in the categorical use of a short list of expressions – often those functioning as connectives or alternatively those which are colourful and complicated (and therefore impressive) – **regardless of whether they are used primarily in spoken or written language**[169] (if indeed at all), or to which text types they are appropriate (Milton 1998:190).

EFL learners generally do not have the means to distinguish between spoken-like linguistic features and academic writing conventions (cf. Milton 1999:228). Teachers should heighten learners' awareness of the stylistic restriction of individual connectors. However, connectors are very **rarely taught as register- or genre-specific**. Figure 6.19 shows that the spoken-like expression *all the same* is given as an equivalent alternative to more formal connectors such as *on the other hand* or *notwithstanding* in Jordan (1999). This example also illustrates the fact that **no information about grammatical category and syntactic properties** is made available to the learners. The preposition *notwithstanding* is listed together with adverbs and adverbial phrases (e.g. *however, yet*) as well as conjunctions (e.g. *although, while*). Learners' **marked preference for sentence-initial position** of connectors has also been related to L2 instruction (cf. Flowerdew 1998; Milton 1999; Narita and Sugiura 2006). Positional variation of connectors is usually not taught and learners use sentence-initial position as a safe bet. When preferred sentence positions of individual connectors are taught, they are often neither corpus-based nor confirmed by corpus data (cf. section 6.2.1.1).

Another problem of teaching practices which has not often been documented is that too much emphasis tends to be placed on connectors, that is, on grammatical cohesion (cf. Halliday and Hasan 1976), **to the detriment of lexical cohesion**[170]. In this chapter, however, nouns, verbs and adjectives have been shown to serve prominent rhetorical functions in academic prose. Labels, i.e. abstract nouns that are inherently unspecific and require lexical realization in their co-text, either beforehand or afterwards, have also been found to fulfil

[169] My emphasis.

[170] Cohesion has often been dealt with in grammars where the focus has always been on connectors. It is noteworthy that, in the new corpus-based *Cambridge Grammar of English* (Carter and McCarthy 2006), no attention is given to lexical cohesion although it has a chapter on textual cohesion ('Grammar across turns and sentences', pp. 242-262) as well as a full chapter on 'Grammar and Academic English' (pp. 266-294).

prominent cohesive roles in this particular genre. It is most probable that lexical cohesion has been neglected in EFL teaching because "there have been no good descriptions of the forms and functions of this phenomenon" (Flowerdew 2006:345). Hinkel (2002:255), however, argues that "the teaching of L2 lexical and syntactic features of text may not or should not be separated from the teaching of discourse and organizational skills."

6.4.2. The role of corpora in EAP materials design

Another important pedagogical implication of our findings is that the teaching of connectors and lexical cohesion by means of phrasemes involving nouns, adjectives and adverbs, necessarily needs to be based on **contextualized** data, and more specifically, on **authentic** texts. While teaching materials designed to help undergraduate students improve their academic writing skills are legion (e.g. Bailey 2006; Hamp-Lyons and Heasley 2006), few make use of authentic texts and very few are corpus-informed[171]. When they are corpus-informed, EAP resources tend to be based on native corpora only. Thus, Thurstun and Candlin's (1997) *Exploring Academic English*, which uses concordance lines to introduce new words in context and familiarise learners with phraseology patterns, rely exclusively on data from a native academic corpus. Although this tool is one of the most innovative EAP textbooks to date, it is arguably less useful for non-native learners, despite Thurstun and Candlin's (1998) claim that it is equally appropriate for native and non-native writers. As shown in section 6.3.3, learner writing is characterized by a number of linguistic features that differ from novice native writing.

The value of such pedagogical tools for non-native speakers of English would be greatly increased if findings from learner corpus data were also used to select what to teach and how to teach it. As stated by Flowerdew (1998), "when choosing which markers to teach, decisions made should also be based on findings from a parallel student corpus to ascertain where students' main deficiencies lie. If not, there is a danger that the emphasis on teaching the most frequent markers may focus on ones already familiar to and correctly used by students, or in this case, exacerbate the problem with their overuse" (Flowerdew 1998:338). By showing in context the types of infelicities learners produce and the types of errors they make, as well as the items they tend to underuse or overuse, learner corpora are the most valuable type of resources to design EAP materials addressing the specific problems

[171] See Gilquin, Granger and Paquot (in press) for a more detailed discussion of the role of corpora, and more specifically, learner corpora in EAP materials design and for possible explanations for the relatively modest role that corpora have played so far.

that EFL learners encounter as non-native writers. Yet, learner corpora have very rarely been used systematically to inform EAP materials (see Milton 1998 and Tseng and Liou 2006 for two exceptions in Computer-Assisted Language Learning).

The only type of resource in which learner corpus data have been relatively successfully implemented up to now is the **monolingual learners' dictionary** (MLD). The latest edition of the *Longman Dictionary of Contemporary English* (Summers 2003) and the *Cambridge Advanced Learner's Dictionary* (Gillard 2003) include a number of learner corpus-informed statements which warn against common learner errors (e.g. the confusion between the adjectives *actual* and *current*, the countable use of the noun *information*)[172]. MLDs are currently conceived of as fully comprehensive writing tools in that they include productively oriented information in areas such as syntactic behaviour, prevention of errors, phraseology and collocation. Yet, if MLDs are to take further "proactive steps to help learners negotiate known areas of difficulty" (Rundell 1999:47), learner corpora should not only be exploited to compile error notes but also to improve other aspects of the dictionary. In the next section, such an enterprise is described, in which learner corpus insights were used to inform a 30-page writing section in the second edition of the *Macmillan English Dictionary for Advanced Learners* (Rundell 2007).

6.4.3. An example of learner corpus-informed materials

In the previous two sections, it was shown that many learner-specific features in the use of connectors and lexical cohesive devices can be attributed to teaching-induced factors as well as to the complete lack of detailed corpus-based descriptions of the semantic, syntactic and phraseological properties of the lexical items which are available in English academic writing to serve specific organizational functions. It has also been argued that learner corpora have a role to play in teaching materials specifically designed for EFL learners. Yet, learner corpora may reveal variability, and more particularly, L1-specific variability. The method used in this thesis, however, has made it possible to distinguish shared features across learner populations from different mother tongue backgrounds from L1-speficic characteristics. These shared features can be used to inform generic tools such as the writing sections I designed in close collaboration with Gaëtanelle Gilquin and Sylviane Granger for the second edition of the *Macmillan English Dictionary for Advanced Learners*.

[172] De Cock and Granger (2004), however, have shown that there is still much room for improvement in the selection and presentation of learners' errors in MLDs.

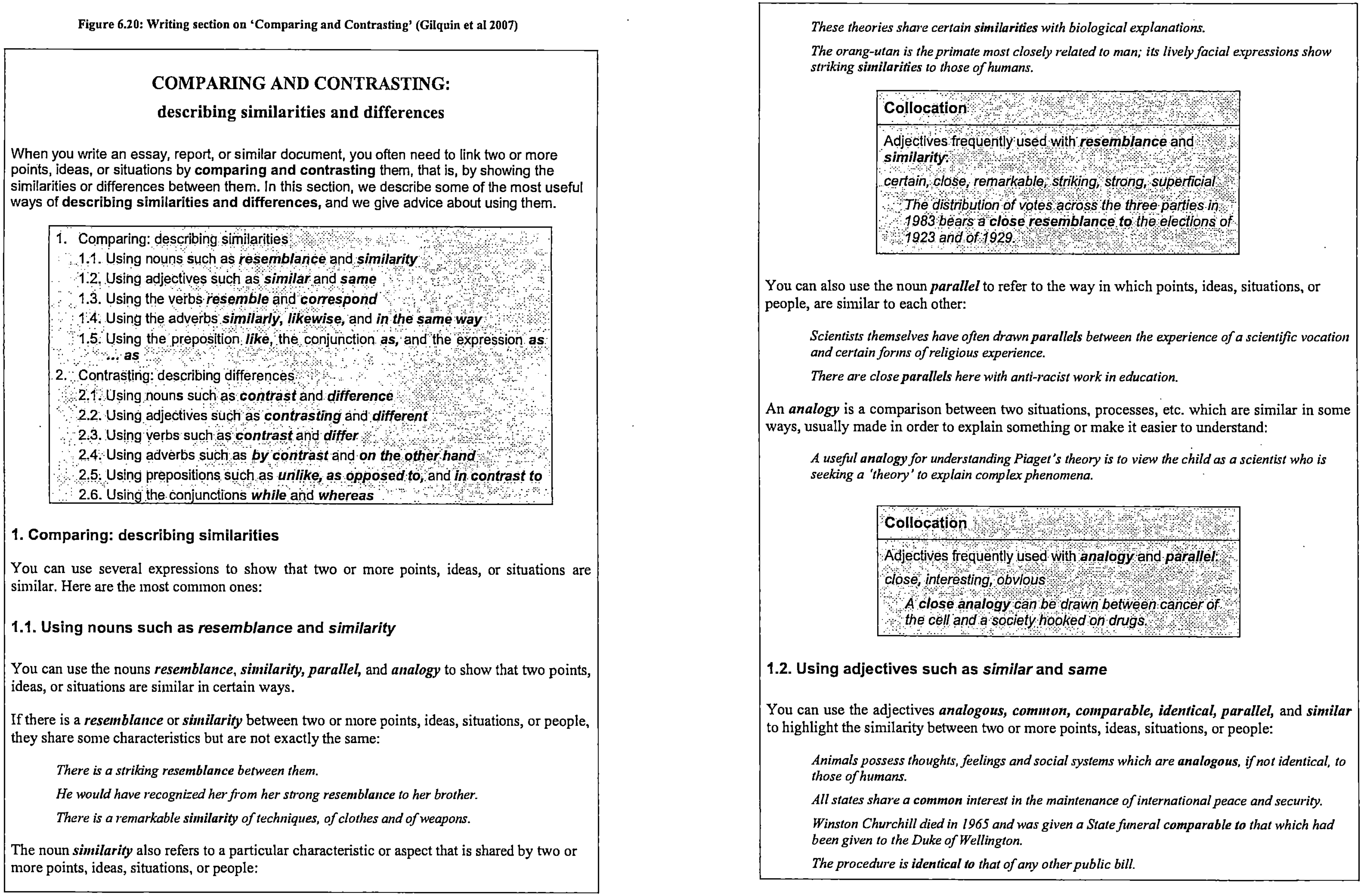
The writing section includes 12 functions that EFL learners need to master in order to write well-structured academic texts: (1) adding information; **(2) comparing and contrasting: describing similarities and differences; (3) exemplification: introducing examples;** (4) expressing cause and effect; **(5) expressing personal opinions**; (6) expressing possibility and certainty; **(7) introducing a concession; (8) introducing topics and related ideas;** (9) listing items; **(10) reformulation: paraphrasing or clarifying; (11) reporting and quoting; (12) summarizing and drawing conclusions** (Gilquin et al. 2007: IW1-IW29). I have written 8 of these 12 functions (in bold). For the reader's convenience, the function of **'comparing and contrasting'** is reprinted in Figure 6.20 while the other seven functions are in Appendix 6.3.

Each writing section includes a detailed corpus-based description of the many lexical means that are available to expert writers to perform a specific function. The words described were selected according to the corpus-driven method described in chapter 5. Special emphasis has been placed on nouns, adjectives and verbs as well as on their phraseological patterns. The sections provide information about how to use these words appropriately by focusing on their:

- Semantic properties
- Syntactic positioning
- Collocations
- Frequency
- Style and register differences

All examples come from the academic component of the *British National Corpus*. Evidence from learner corpora has been used in order to inform the writing sections in several ways. The writing sections specifically address the types of learners' problems discussed in this chapter, namely, restricted lexical repertoire, overuse and underuse, lack of register awareness, phraseological infelicities, semantic misuse, syntactic positioning, etc. Our treatment of these problems is mainly explicit, in that we draw learners' attention to error-prone items and we provide them with negative feedback in the form of "Be careful!" notes and "Get it right" boxes. The latter are intended to give guidance on how to avoid common errors while the former focus on problems of frequency (overuse and underuse), register confusion and atypical positioning. They are typically supported by frequency data, in the form of graphs which help the reader visualise the differences between learners' behaviour and that of native writers. Thus, in the section on 'Comparing and Contrasting', a graph is used to show that learners have a strong tendency to use the expression *look like,* which is relatively rare in academic prose (see Figure 6.20). Numerous authentic examples are provided to illustrate all the points we make. The reader is referred to Gilquin, Granger and Paquot (in press) for more detailed information on the principles that guided the design of these writing sections.

Figure 6.20: Writing section on 'Comparing and Contrasting' (Gilquin et al 2007)



COMPARING AND CONTRASTING:

describing similarities and differences

When you write an essay, report, or similar document, you often need to link two or more points, ideas, or situations by **comparing and contrasting** them, that is, by showing the similarities or differences between them. In this section, we describe some of the most useful ways of **describing similarities and differences,** and we give advice about using them.

1. Comparing: describing similarities
    - 1.1. Using nouns such as ***resemblance*** and ***similarity***
    - 1.2. Using adjectives such as ***similar*** and ***same***
    - 1.3. Using the verbs ***resemble*** and ***correspond***
    - 1.4. Using the adverbs ***similarly, likewise,*** and ***in the same way***
    - 1.5. Using the preposition ***like,*** the conjunction ***as,*** and the expression ***as ... as***
2. Contrasting: describing differences
    - 2.1. Using nouns such as ***contrast*** and ***difference***
    - 2.2. Using adjectives such as ***contrasting*** and ***different***
    - 2.3. Using verbs such as ***contrast*** and ***differ***
    - 2.4. Using adverbs such as ***by contrast*** and ***on the other hand***
    - 2.5. Using prepositions such as ***unlike, as opposed to,*** and ***in contrast to***
    - 2.6. Using the conjunctions ***while*** and ***whereas***

1. Comparing: describing similarities

You can use several expressions to show that two or more points, ideas, or situations are similar. Here are the most common ones:

1.1. Using nouns such as *resemblance* and *similarity*

You can use the nouns ***resemblance***, ***similarity, parallel,*** and ***analogy*** to show that two points, ideas, or situations are similar in certain ways.

If there is a ***resemblance*** or ***similarity*** between two or more points, ideas, situations, or people, they share some characteristics but are not exactly the same:

> *There is a striking **resemblance** between them.*
>
> *He would have recognized her from her strong **resemblance** to her brother.*
>
> *There is a remarkable **similarity** of techniques, of clothes and of weapons.*

The noun ***similarity*** also refers to a particular characteristic or aspect that is shared by two or more points, ideas, situations, or people:

> *These theories share certain **similarities** with biological explanations.*
>
> *The orang-utan is the primate most closely related to man; its lively facial expressions show striking **similarities** to those of humans.*

Collocation

Adjectives frequently used with ***resemblance*** and ***similarity:***
certain, close, remarkable, striking, strong, superficial

> *The distribution of votes across the three parties in 1983 bears a **close resemblance to** the elections of 1923 and of 1929.*

You can also use the noun ***parallel*** to refer to the way in which points, ideas, situations, or people, are similar to each other:

> *Scientists themselves have often drawn **parallels** between the experience of a scientific vocation and certain forms of religious experience.*
>
> *There are close **parallels** here with anti-racist work in education.*

An ***analogy*** is a comparison between two situations, processes, etc. which are similar in some ways, usually made in order to explain something or make it easier to understand:

> *A useful **analogy** for understanding Piaget's theory is to view the child as a scientist who is seeking a 'theory' to explain complex phenomena.*

Collocation

Adjectives frequently used with ***analogy*** and ***parallel:***
close, interesting, obvious

> *A **close analogy** can be drawn between cancer of the cell and a society hooked on drugs.*

1.2. Using adjectives such as *similar* and *same*

You can use the adjectives ***analogous, common, comparable, identical, parallel,*** and ***similar*** to highlight the similarity between two or more points, ideas, situations, or people:

> *Animals possess thoughts, feelings and social systems which are **analogous,** if not identical, to those of humans.*
>
> *All states share a **common** interest in the maintenance of international peace and security.*
>
> *Winston Churchill died in 1965 and was given a State funeral **comparable to** that which had been given to the Duke of Wellington.*
>
> *The procedure is **identical to** that of any other public bill.*

*A **parallel** but not identical distinction is between short-term and long-term memory.*

*The pattern of mortality is broadly **similar** for men and women.*

The adjective ***same*** is <u>always</u> used before the noun:

*The **same** pattern is also to be found in the discourse of parliamentary debates about apartheid.*

Note that ***same*** can also be used as a pronoun:

*The rules are almost the **same** as for domestic operations.*

*In all but a few minor respects, the privileges of the two Houses are the **same**.*

Get it right!

Same never comes after ***a***:

✗ *Women still have to work twice as hard as men for ~~a same~~ salary.*

✓ *Women still have to work twice as hard as men for <u>the same</u> salary.*

The adjective ***alike*** is <u>never</u> used before a noun. It is typically used after the verbs ***be*** and ***look***:

*On other issues such as education, health and social welfare the two mainstream parties **are** remarkably **alike.***

*Thus two individuals of different species from the same place **look** more **alike** than two individuals of the same species from different places.*

Collocation

Adverbs frequently used with:

- **comparable**: *broadly, directly, roughly*

  *The Scottish figures are not **directly comparable**.*

- **similar**: *broadly, fairly, quite, remarkably, roughly, somewhat, strikingly*

  ***Remarkably similar** results have been obtained by studies in the United Kingdom and other countries.*

- **same**: *essentially, exactly, much, precisely, quite, roughly*

  *They both contain **exactly the same** information.*

1.3. Using the verbs *resemble* and *correspond*

You can also use the verbs ***resemble*** and ***correspond*** to show that two or more points, ideas, or situations are similar:

*It is possible to suggest that the two poets **resemble** one another.*

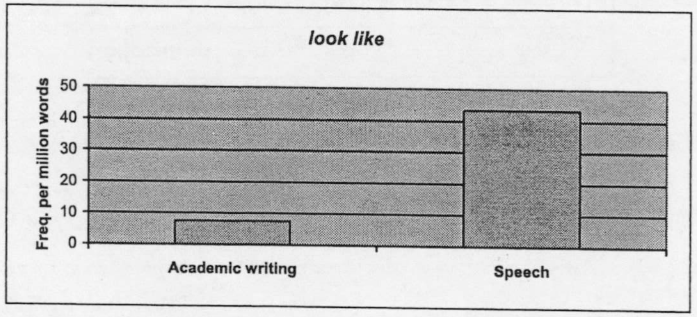
*Her views on capital punishment, immigration, and the trade unions **resemble** those of the right-wing tabloid press.*

*The techniques used with normal subjects give estimates that closely **correspond** to those derived from the clinical literature.*

*The political weakness of these states **corresponded to** their economic weakness.*

Be careful!

Many learners use the verb ***look like*** to show that two or more points, ideas, situations, or people, are similar. However, this verb is more frequent in speech and informal writing.



1.4. Using the adverbs *similarly, likewise,* and *in the same way*

You can use the adverbs ***similarly, likewise,*** and ***in the same way*** to show that the points, ideas, or situations you are comparing are alike. They are often used to modify the whole sentence and, in that case, are used at the beginning of a sentence, followed by a comma:

*One parent families may come about because of death, divorce or separation in a two parent family. **Similarly,** a one parent family may become a two parent family through marriage or remarriage.*

*Media theories must make the absence of state control their cornerstone. **Likewise,** proposals for the reform of the media must pay due attention to it.*

*Infants as young as 6 weeks consistently show preferences for familiar faces. **In the same way,** infants respond preferentially to their mother's voice compared to the voice of a stranger.*

When it is used inside the sentence, ***in the same way*** is normally followed by ***as***:

*Planning controls operate in rural areas **in the same way as** in urban areas.*

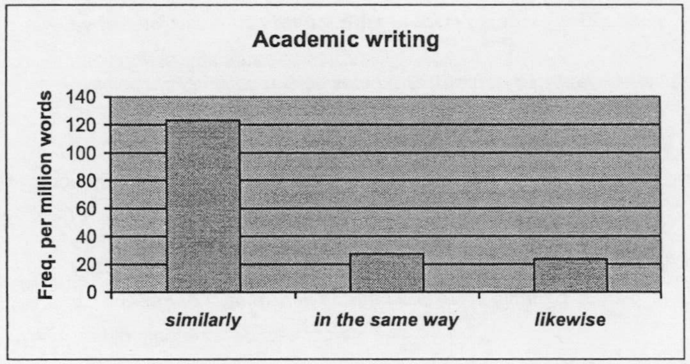
However, it can also be followed by ***that*** to introduce a clause:

Adverbs describe verbs ***in the same way that*** *adjectives describe nouns.*

The adverb ***similarly*** can also be used to modify an adjective:

A ***similarly*** *complex picture emerges from the results in the metropolitan authorities.*

Note that the adverb ***similarly*** is much more frequent than ***in the same way*** and ***likewise***.



1.5. Using the preposition *like*, the conjunction *as*, and the expression *as … as*

You can also use the preposition ***like***, the conjunction ***as***, and the expression ***as*** + ADJECTIVE / ADVERB + ***as …*** to describe similarities:

The preposition ***like*** is used before noun phrases:

Like *many others, Berkeley objected to the complete materialism of Hobbes.*

The police, ***like*** *most people, have stereotypical views as to the "typical" criminal or delinquent.*

The conjunction ***as*** introduces clauses. It is often preceded by a comma:

The "Celtic belt" was heavily forested in those days, ***as*** *was Italy in pre-Roman times.*

The contexts in which they work vary, ***as*** *do their personal and professional backgrounds.*

She had left him, just ***as*** *she so often threatened to do.*

He does not want opinion polls banned, ***as*** *is the case in Australia and some European countries.*

Get it right!

Don't use the impersonal pronoun ***it*** in subject position after ***as***:

✗ *Children are not playing marbles any more,* ~~***as it was the case***~~ *about thirty years ago.*

✓ *Children are not playing marbles any more,* ***as was the case*** *about thirty years ago.*

As can also introduce reduced verbless clauses, i.e. clauses that do not contain any verb:

Adaptation is the key principle of modern biology, ***as*** *described by Charles Darwin in the mid-nineteenth century.*

As + ADJECTIVE / ADVERB + ***as*** can be followed by a noun phrase or by a full clause:

Art is at least ***as important as*** *politics.*

There is the intriguing finding that some amnesiacs can learn certain new skills ***as quickly as*** *normal subjects.*

This task is by no means ***as simple as*** *it sounds.*

Get it right!

Like is not used to introduce full clauses in academic writing and professional reports. This use is very informal and some people consider it to be incorrect. Use ***as*** or ***in the same way as*** instead:

✗ ~~***Like***~~ *Marx said, religion is a falsification of life.*

✓ ***As*** *Marx said, religion is a falsification of life.*

✗ *You get addicted to television* ~~*like*~~ *you get addicted to opium.*

✓ *You get addicted to television* ***in the same way as*** *you get addicted to opium.*

Frequent phrases with *as*

In academic writing and professional reports, the conjunction **as** is used in a number of frequent phrases. These phrases have three main functions:

1. Referring to what you said before or to what you are going to say:

	defined	
	described	
	discussed	***above***
as	***explained***	***before***
	mentioned	***earlier***
	noted	***in chapter X***
	outlined	
	shown	
	stated	

This is not to say that Castro's actions ***as described above*** *can be explained solely in terms of Cuba's relations with Moscow.*

as already described / indicated / mentioned / noted / pointed out / stated / suggested

As already mentioned*, an increasing number of research reports shows that a frequent complaint made by patients is lack of information.*

as will be described / discussed / seen / shown

The comment was made in the British context, but ***as will be seen in*** *later chapters, it is of general applicability.*

2. Referring to diagrams, figures, tables, etc:

as described / illustrated / seen / shown in *[figure 3, table 6, etc]*

From the end of World War II to the early 1970s there was virtually full employment in developed countries, but in the next ten years the position changed considerably ***as shown in fig. 1.11.***

3. Reporting the ideas or findings of other writers:

as defined / described / shown / suggested by [Name]

Settlement was denser in some areas than in others; although such a division may only be recognised, ***as suggested by*** *Peter Fowler, from late Bronze Age times onwards.*

as *[Name]* ***explains / points out / reports / states / suggests***

Finally, ***as*** *Ehrlichman and Weinberger* ***point out****, differences in the distance between subject and experimenter may well account for certain of the inconsistencies in the results reported by different investigators.*

See the section on "Quoting and Reporting" for more information on these structures.

2. Contrasting: describing differences

You can use several expressions to show that two or more points, ideas, or situations are different. Here are the most common ones.

2.1. Using nouns such as *contrast* and *difference*

You can use the nouns ***contrast, difference,*** and ***distinction*** to express contrast:

However, there was an important ***contrast*** *between rural and urban settings.*

The tone of the report and its recommendations were ***in marked contrast with*** *those of earlier enquiries into child care scandals.*

Table 1 shows significant ***differences in*** *marital status.*

The ***difference*** *lies in the time the animal spends resting between meals.*

There is a sharp ***distinction between*** *domestic politics and international politics.*

Collocation

Adjectives frequently used with:

- ***contrast***: *direct, marked, sharp, stark, striking*

  There is a ***sharp contrast between*** *the US and Europe.*

- ***difference***: *considerable, crucial, essential, fundamental, main, major, marked, significant, striking, substantial*

  There are ***substantial differences in*** *mortality rates among older people in developed countries.*

- ***distinction***: *clear, crucial, fundamental, sharp*

  There appears to be a ***clear distinction between*** *the causes of uplift on the western and eastern sides of the central Andes.*

Verbs frequently used when these nouns are the object:

- ***contrast***: *draw, provide*

  Sussex ***provides*** *a great* ***contrast to*** *many other parts of England.*

- ***difference***: *find, notice, observe, show*

  We did ***find*** *a clear* ***difference between*** *the three- and the four-year-olds.*

- ***distinction***: *draw, make*

  It is simple to ***draw*** *a* ***distinction between*** *pluralist, elitist and Marxist approaches to the distribution of political power.*

You can also use the nouns ***the contrary, the opposite***, and ***the reverse*** to show that two or more points, ideas, or situations are so different that there is a strong opposition or conflict between them:

> *Scientists at first believed it was impossible for birds to directly infect humans with the virus. But an outbreak in Hong Kong **in** 1997 proved **the contrary**.*
>
> *Whatever characterises men, in their own view, women are defined as **the opposite**.*
>
> *Women from the higher classes have a greater tendency to stay single. **The reverse** is true for men, however, with a smaller proportion of working-class men marrying than their upper-class contemporaries.*

2.2. Using adjectives such as *contrasting* and *different*

You can use the adjectives ***contrasting, different, differing,*** and ***unlike*** to express a contrast:

> *It is, therefore, with two **contrasting** views of the discipline that I wish to begin.*
>
> *A rather **different** approach is taken by Turner, an architect who, for some time, worked in Lima.*
>
> *The essays in this book do not present a monolithic view, and they reflect **widely differing** perspectives.*

The adjective ***unlike*** is never used before a noun. It is typically used after the verb ***be***:

> *The Indian situation **is** quite **unlike** that of the Soviet Union following the October Revolution.*

You can use the adjectives ***contrary, opposite,*** and ***reverse*** to highlight a strong opposition:

> *In the absence of **contrary** agreement, the buyer need not make payment until delivery is made.*
>
> *Some stimuli in the environment will help you diet, and some will have just the **opposite** effect.*
>
> *To voters, who hope that a clear result will speed the end of the recession, an unclear result could have precisely the **reverse** effect.*

2.3. Using verbs such as *contrast* and *differ*

You can also use the verbs ***contrast*** and ***differ*** to show that two or more points, ideas, or situations are different:

> *This view **contrasts** sharply **with** the disillusioned, almost pessimistic, reflections of George Orwell, who wrote that there was no true internationalism amongst the British working class.*
>
> *This **contrasts with** the situation in the USA, where the clean-up of abandoned sites is a priority.*
>
> *Housing in rural areas **differs from** housing in the urban areas.*
>
> *The Arctic and Southern oceans **differ** in many respects.*

Collocation

Adverbs frequently used with:

- ***contrast***: *sharply, strongly*

  *This **contrasts strongly with** the simplicity of word processing, data base or even spreadsheet software.*
- ***differ***: *considerably, greatly, markedly, significantly, slightly, widely*

  *It did not **differ significantly** between groups.*

You can also show a contrast by using the verb ***compare*** in the expressions ***(as) compared with*** and ***(as) compared to***:

> *Britain's economic fortunes generally had fallen sharply **compared with** other western European countries since 1979.*
>
> *At each age, mortality rates are lower for women **as compared with** men.*
>
> *Furthermore, 12 per cent of males **compared to** 2 per cent of females are found guilty of, or cautioned for, criminal offences by the age of 17.*

You can use the verbs ***distinguish*** and ***differentiate*** to introduce a feature that makes someone or something clearly different from similar people or things. The distinguishing feature is shown in <u>subject</u> position:

> *This bold personal style **distinguishes** her from other Prime Ministers.*
>
> *This feature **differentiates** at least four groups of predator.*

The two verbs can also be used with the meaning of 'to recognize the differences between two or more points, ideas, or situations':

> *The first step towards understanding the crisis of 1931 is to **distinguish** between different types of coalition government.*
>
> *The introduction in our paper clearly **differentiates** the three levels of prevention on the basis of recent recommendations.*

Be careful!
All the verbs described above can be followed by a preposition. Acceptable prepositions vary from one verb to another:

> *The success is even more remarkable when **compared to** her predecessors.*
>
> *Another approach is to group together the readers of various right-wing papers and **contrast** them **with** readers of left-wing papers.*

Savannah animals ***differ from*** *those of tropical rainforests and these again* ***from*** *inhabitants of the tundra at high latitudes.*

For more information on which preposition to use with each verb, see their dictionary entries.

2.4. Using adverbs such as *by contrast* and *on the other hand*

You can use the adverbs ***by/in contrast*** and ***by/in comparison*** to show that two or more points, ideas, or situations are different. They are used:

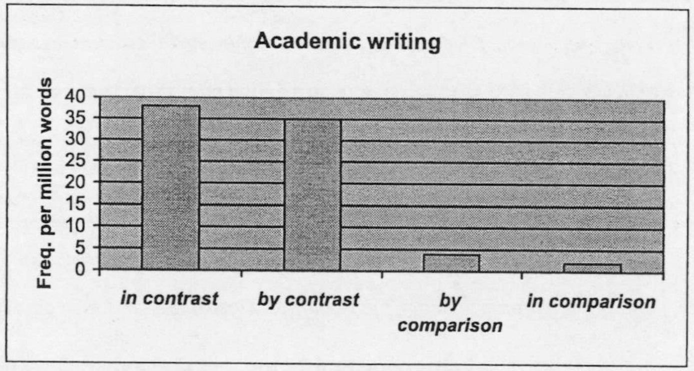
- at the beginning of the sentence, followed by a comma:

> *The rules of a factory may be written down and serve as strict regulators of behaviour.* ***In contrast,*** *the rules of dress or of how we eat are unwritten guides to behaviour.*
>
> *Gallstone disease is the most common cause of pancreatitis in Scotland.* ***By comparison,*** *alcohol is the most common cause of pancreatitis in patients in Finland.*

- inside the sentence, enclosed by commas:

> *For his final years in office, Warwick was known to be out of favour with the king. Gloucester,* ***by contrast,*** *kept the king's full confidence.*

Note that the adverbs ***by contrast*** and ***in contrast*** are much more frequent than ***by comparison*** and ***in comparison***:



You can also use the adverb ***on the other hand*** to talk about differences. It is often found after the subject, enclosed by commas:

> *Apollo is in various ways a god of higher civilization: he is, for instance, the god of medicine. Dionysus,* ***on the other hand,*** *is a god of nature and natural fertility.*
>
> *Consent is always given to the actions of other persons. Promises,* ***on the other hand,*** *cannot, except in very special circumstances, ever be made concerning the actions of another person.*

The adverb ***on the other hand*** is also used to describe two contrasting qualities of a single subject or topic. It is then more commonly found at the beginning of a sentence:

> *Working in groups can help to speed up the work where an investigation involves the results from several different cases.* ***On the other hand,*** *it can isolate quiet pupils from the action, and very quick pupils can ignore slower ones without intentional cruelty.*

Be careful!

Learners often use ***on the other hand*** in the expression ***on the one hand, … on the other hand, …*** In fact, the adverb ***on the other hand*** is more frequently used on its own:

> ***? On the one hand,*** *mobile phones are very useful if something unexpected happens.* ***On the other hand,*** *they are such a nuisance on the streets and everywhere!*
>
> ✓*Mobiles phones are very useful if something unexpected happens.* ***On the other hand,*** *they are such a nuisance on the streets and everywhere!*

Get it right!

The expression ***on the other side*** is not used for describing differences between two or more points, ideas, situations, people, etc:

✗ *Men have always been responsible for growing food.* ***~~On the other side~~,*** *women have always taken care of children.*

✓ *Men have always been responsible for growing food. Women,* ***on the other hand,*** *have always taken care of children.*

Get it right!

The expression ***on the contrary*** is not used for describing differences between two or more points, ideas, situations, people, etc:

✗ *Onasis had everything but he wanted to have more. Raskolnikov,* ***~~on the contrary,~~*** *had nothing.*

✓ *Onasis had everything but he wanted to have more. Raskolnikov,* ***by contrast,*** *had nothing.*

On the contrary expresses a direct denial of what has been asserted before, and means that the opposite is true. It is often used at the beginning of a sentence, followed by a comma, and usually comes after a negative sentence:

> *There is no single and all-important cause of attempted suicide.* ***On the contrary,*** *a variety of interpersonal, social, and psychological factors may contribute to it.*

You can use the adverb ***conversely*** for indicating that one situation is the exact opposite of another. It is typically used at the beginning of a sentence, followed by a comma:

> *The author clearly identifies an acceptance of poor health as an attribute of normal ageing.* ***Conversely,*** *youth is depicted as a time of vitality and good health.*

2.5. Using prepositions such as *unlike, as opposed to,* and *in contrast to*

You can also use prepositions to show a contrast. The prepositions ***unlike, in contrast to/with,*** and ***by/in comparison with*** are mainly used inside the sentence, but they are sometimes found at the beginning of the sentence:

> *A public company,* ***unlike*** *a private company, must be limited by shares.*
>
> ***Unlike*** *other prices (and wages), interest rates are to be politically determined.*
>
> *Government decisions in Britain are made in the name of the Cabinet,* ***in contrast to*** *the United States where they are made in the name of the President.*
>
> *Judges might find it easier to sympathise with and understand business law-breakers* ***in comparison with*** *other types of criminal.*

The prepositions ***as against*** and ***as opposed to*** are almost always used inside the sentence:

> *The rate of increase of convicted rapists is 39%,* ***as against*** *an increase in rape reports of 143%.*
>
> *Moving from insects to reptiles to mammals, the importance of learned,* ***as opposed to*** *genetically determined, behaviour gradually increases.*

The preposition ***versus*** (or sometimes its abbreviation ***vs.***) is used for stating that two or more points, ideas, or situations are being compared in order to show that they are different:

> *He brought the argument about Christian unity up against the fundamental question of Catholicism* ***versus*** *Protestantism.*
>
> *The American* ***versus*** *British female stereotype is a trans-Atlantic difference of opinion that has been raging for decades.*

It is commonly found in titles and sub-titles:

> *Microsoft* ***versus*** *Unix — collision course or co-existence?*

You can use the preposition ***contrary to*** to introduce a strong opposition. It is mainly used within the sentence, preceded by a comma, but can also be found at the beginning of the sentence.

> *It might well appeal to poets, who,* ***contrary to*** *popular myths about inspiration, are usually interested in the technical aspects of composition.*
>
> ***Contrary to*** *expectations, studies show that most people continue to regard themselves positively as they grow older.*

2.6. Using the conjunctions *while* and *whereas*

You can use the conjunctions ***while*** and ***whereas*** to balance contrasting points, ideas, or situations. They are used:

- inside the sentence, preceded by a comma:

> *Biographies will be included in the present chapter,* ***while*** *catalogues will be treated in the next.*
>
> *The majority of students of English are women,* ***whereas*** *the majority of academics teaching it are male.*

- less frequently, at the beginning of a sentence:

> ***While*** *teachers without full qualifications are poorly paid, qualified teachers are now relatively well off.*
>
> ***Whereas*** *the inexperienced pilot needs to fly at least once a month to be safe, more experienced pilots can go much longer without always becoming badly out of practice.*

Finally, you can also use the conjunction ***but*** to highlight a difference:

> *Inflation was the scourge of the mid-1970s* ***but*** *not the mid-1980s.*
>
> *He doesn't know anybody,* ***but*** *everybody knows him.*

6.5. Conclusion

> ... they have a shortfall of syntactic and lexical tools to enable them to produce competent written academic text (Hinkel 2002:160).

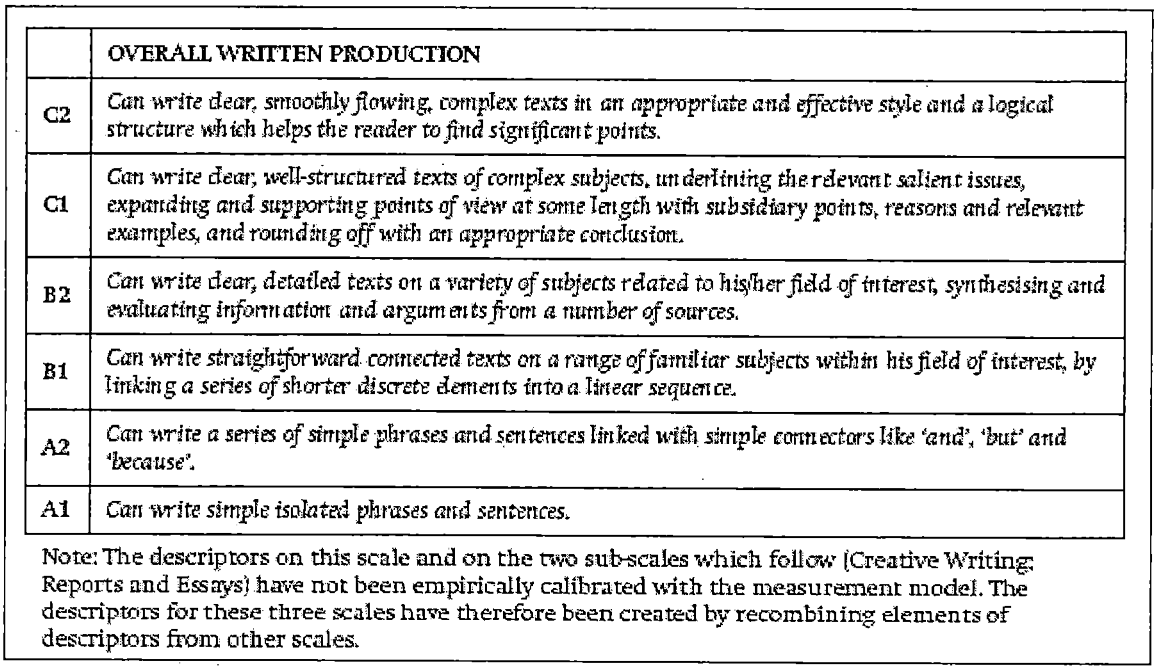
This chapter has established the value of the *Academic Keyword List* for theoretical, descriptive and pedagogical purposes. Detailed corpus-based descriptions of a selected list of AKL words in native professional writing have offered valuable insights into the distinctive nature of the phraseology of rhetorical functions in academic prose. Systematic use of both types of comparison involved in the model of *Contrastive Interlanguage Analysis*, that is, the comparison of interlanguage with native language, but also, of several interlanguages together, has made it possible to identify a whole range of learner-specific uses of the words common to a majority of L1 populations. These shared features can be regarded as a common core which characterizes the writing of upper-intermediate to advanced learners in institutional settings, irrespective of their L1 backgrounds.

It has been argued that a systematic analysis of several interlanguages is necessary for analyzing the potential influence of developmental, teaching-induced and transfer-related factors on EFL learner writing. By focusing on shared features across L1 learner populations, we have highlighted the important role played by developmental and teaching-induced factors in learner written production. We have also shown that it is not always possible to attribute learner-specific features to a single factor as developmental, teaching-induced and transfer-related effects can reinforce each other (cf. Granger 2004:135-136). Several examples of assumed L1 influence in the literature have been questioned and other explanations have been proposed in the light of data provided by our comparisons of several interlanguages. Transfer-related factors have only lightly been touched upon as they are the focus of the following chapter.

Shared learner-specific features have provided useful information on what EFL learners need in order to improve their academic writing skills. Together with data from native corpora, they have been used to inform academic writing sections in the second edition of the *Macmillan English Dictionary for Advanced Learners*. They could also be employed to help design other types of generic tools such as a corpus-informed EAP textbook for EFL learners or an electronic writing-aid tool. Detailed descriptions of AKL words in native and learner corpora are arguably the type of data needed to build an EAP dictionary (cf. Kosem and Krishnamurthy 2007).

The method used in this chapter has rarely been used in learner corpus research. However, the present study has brought to light its potential contribution to a number of theoretical discussions and applied projects. One avenue for future research would be to use this method to analyze a corpus stratified for proficiency levels rather than L1 backgrounds. Such a study could help improve the *Common European Framework* (CEF) descriptors for overall writing competence (cf. Figure 6.21). Unlike for reading, listening and speaking, the descriptors for writing are largely intuitive and were not empirically calibrated (cf. North 2002). Proficiency levels are very impressionistically described and the descriptors do not provide useful guidelines in order to distinguish between, for example, B2 and C1 texts.[173]

Figure 6.21: The Common European Framework: overall written production (Council of Europe 2001:61)



	OVERALL WRITTEN PRODUCTION
C2	*Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.*
C1	*Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.*
B2	*Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources.*
B1	*Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence.*
A2	*Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.*
A1	*Can write simple isolated phrases and sentences.*

Note: The descriptors on this scale and on the two sub-scales which follow (Creative Writing; Reports and Essays) have not been empirically calibrated with the measurement model. The descriptors for these three scales have therefore been created by recombining elements of descriptors from other scales.

More generally, the method could also be used to help identify other aspects of interlanguage – aspects of grammar, morphology, etc. - which are shared across learners from several mother tongue backgrounds and distinguish them from L1-specific features.

[173] Thewissen et al. (2005), Thewissen (2006) and Granger and Thewissen (2006) have examined the possible contribution of **error-tagged learner corpora** to the CEF specifications regarding the linguistic competences relevant for writing, e.g. lexical competence, semantic competence, grammatical competence. They have shown that looking at error domains can help to better distinguish between proficiency levels.

7. L1-specific variability in learners' use of EAP vocabulary and the issue of transfer

> "Learners clearly cannot be regarded as 'phraseologically virgin territory': they have a whole stock of prefabs in their mother tongue which will inevitably play a role – both positive and negative – in the acquisition of prefabs in the L2." (Granger 1998: 158)

7.1. Introduction

Chapter 6 has highlighted a number of linguistic features that are shared by most learner populations when compared to native academic writing. EFL learners, however, do not use all lexical items similarly, irrespective of their mother-tongue background. In section 6.3.2.2, it was shown that although all L1 learner populations overuse the adverb *maybe* and the expression *I think*, relative frequencies differ widely across L1 populations. These differences may be explained by a number of factors such as first language, essay prompt, vocabulary knowledge and proficiency level. The focus of Chapter 7 is on the potential influence of the first language on EFL learners' academic vocabulary. The objective is not to assess the extent to which L1-specific variability in learners' use of EAP words and phrasemes may be accounted for by L1 influence nor to determine how significant it is in L2 acquisition (cf. Ellis 1994:341), two questions which are clearly beyond the scope of this dissertation. Rather, the primary objectives of this chapter are **methodological**. This being said, preliminary results on a number of transfer effects will also highlight the theoretical significance of the methods adopted.

Methodological issues in transfer studies have been addressed in chapter 3 where we have argued that learner corpus data could be used to overcome some of the limitations of the field in terms of amount of data, type of data analysis, selection of items to be investigated, etc. This chapter thus aims to investigate how learner corpus data and corpus linguistics approaches can be used to help inform the different steps of transfer studies, and more particularly, item selection and transfer assessment.

It first proposes to make use of a **corpus-driven approach** to identify EAP clusters that are more frequently used by a specific L1 learner population in comparison with EFL learners from different mother tongue backgrounds (cf. section 7.2). It is hypothesized that this "raw discovery procedure" (De Cock 2003: 199) will bring to light a wide range of L1-specific word sequences which may serve as a new type of material for transfer studies.

An important outcome of chapter 3 is also that transfer studies should ideally rely on more than one source of evidence to prove L1 transfer. Section 7.3 thus describes Jarvis's (2000) unified framework for the study of L1 influence in which IL-L1 comparisons, IL-IL comparisons and comparisons of IL productions of learners sharing the same language are used. I propose one procedure to operationalize Jarvis's framework on learner corpus data which relies on corpus linguistics methods and several statistical measures. This procedure is repeated in four case studies which are described successively. In section 7.4, pedagogical implications of our results are briefly discussed.

7.2. Making use of learner corpora to select items to be investigated in transfer studies

This section describes one possible way of exploiting learner corpora to select items worthy of investigation in transfer studies. A **corpus-driven approach** similar to the one adopted in Chapter 5 to extract EAP words is first used to select word sequences that display L1-specific variability. The focus is then placed on word sequences including words that have been found to serve specific rhetorical or organizational functions in learner and/or native writing.

7.2.1. A corpus-driven approach

A corpus-driven approach to item selection seems particularly well suited to overcome the limitations of transfer studies which have been observed in section 3.3.2. As De Cock (2003) puts it, data in corpus-driven methods "constitute the starting point of a path-finding expedition that will allow linguists to uncover new grounds, new categories and formulate new hypotheses on the basis of the patterns that were observed" (De Cock 2003:197). Corpus-driven methods have been used to make IL-L1 comparisons and have already brought to light a number of findings related to L1-transfer (cf. De Cock 2003; 2004 – see sections 1.4.2 and 3.2.2.3). More studies of this type are clearly needed, especially for learner populations from other mother tongue backgrounds than French. The method used here, however, is different.

Corpus-driven methods have never been used to compare the interlanguage of learners from at least two different mother tongue backgrounds so as to verify that the first language really is the major explanatory factor (cf. section 3.3.3.2). In this section, we therefore intend to start filling this gap and demonstrate how promising the method is. We adopt a corpus-driven approach to L1-specific variability in learner corpora with the aim of highlighting 2-to-5-word sequences that are more (or less) frequently used by French, Spanish or Dutch learners[174] when compared to the other nine learner populations represented in the first version of the *International Corpus of Learner English*. The procedure used is as follows. We first make a list of 2-word sequences for ICLE-FR and a second one for the 9 other learner sub-corpora as a group with *WordSmith Tools* (cf. section 4.2.2.1). These two lists are then compared with the *Keywords* tool (cf. section 5.3.1.1). The same procedure is repeated for sequences of 3 words, 4 words and 5 words. The same method is used to compare 2-to-5 word sequences in ICLE-SP and ICLE-DU vs. the 9 other learner sub-corpora.

Table 7.1 gives the number of positive and negative key 2-to-5- word clusters[175] in ICLE-DU, ICLE-FR and ICLE-SP. It shows that the number of positive key clusters at least partly depends on **corpus size**. A higher number of positive key clusters is found in ICLE-DU, which totals 163,243 words, than in ICLE-SP, which only counts 99,119 words. This finding is easily explained by the fact that the bigger the corpus is, the more often a sequence may be repeated. Corpus size, however, does not help interpret the high number of positive key clusters found in ICLE-FR. A likely explanation is provided by the number of different **topics** which were used as essay prompts across the three corpora. Unlike in ICLE-DU and ICLE-SP, essays in ICLE-FR address a very restricted range of topics. The fact that there are many essays on the same topic provides more opportunities for students to make use of the same word sequences. Thus, 92 essays (42% of essays in ICLE-FR) deal with the essay prompt "Europe 92: loss of sovereignty or birth of a nation", which results in the following topic-related sequences being positive key clusters: *the birth of a nation, loss of national identity or, the united states of Europe, lead to a loss of, each country has its own, the creation of the European, the European countries, the unification of, all the European, the twelve countries, its own culture, a new nation, unification of Europe, the single market, the free movement, a great nation, andpolitical union, say that Europe, the new Europe, a united Europe,* etc.

[174] ICLE-FR, ICLE-SP and ICLE-DU were selected so that I was able to compare learners' use of lexical items in L2 with native usage of potential translation equivalents.
[175] Positive keywords have been defined as words that are statistically prominent in a corpus while negative keywords are words that have strikingly low frequency (cf. section 5.3.1.1).

Table 7.1: Positive and negative key 2-to-5-word clusters in ICLE-DU, ICLE-FR and ICLE-SP

		ICLE-DU 162,243 words	ICLE-FR 136,343 words	ICLE-SP 99,119 words
2-word clusters	+	795	937	541
	-	693	642	395
3-word clusters	+	344	520	248
	-	282	204	102
4-word clusters	+	82	156	75
	-	58	34	14
5-word clusters	+	19	43	16
	-	12	5	3

A sizeable proportion of negative keywords in ICLE-DU, ICLE-FR and ICLE-SP are also topic-related. Examples in ICLE-FR include clusters directly related to the essay topic "Most University degrees are theoretical and do not prepare us for the real life. Do you agree or disagree?" which is less frequently used as a prompt in other learner corpora: *students for the real world, prepare students for the real, students for the real, not prepare students for, prepare students for the, most university degrees are, university degrees are theoretical, do not prepare students, university degrees are*, etc. These examples stress the importance of taking **topic** into account when interpreting interlanguage data, something which is not often done in learner corpus research.

Next to topic-related word sequences, positive and negative key clusters also include a number of clusters which are particularly interesting to study as they involve words and phrasemes that have been shown to serve specific rhetorical or organizational functions in learner and/or native writing (cf. chapter 6). The next section is therefore devoted to this specific type of potential EAP key clusters, which will henceforth be referred to as **EAP key clusters** for simplicity's sake.

7.2.2. Focusing on 'EAP key clusters'

In Chapter 6, learners' use of EAP words and phrasemes has been shown to share similarities when compared to native writing. However, it is not because all L1 learner populations overuse or underuse a lexical item when compared to native writing that there is no L1-specific variability. An EAP word or phraseme may be frequently used by EFL learners in general and, even more heavily relied upon by one specific L1 learner population. Data resulting from IL-IL comparisons as described in section 7.2.1 suggests that a comparison of positive and negative EAP key clusters across learner corpora may help identify word sequences that only serve specific rhetorical or organizational functions in essays written by learners who share a single mother tongue background. These L1-specific clusters are arguably good candidates for transfer studies.

Ideally, an analysis of EAP key clusters in learner corpora should at least consist in the following steps:

- Operationalization of the concept of 'EAP key cluster': a possible solution is to define 'EAP key clusters' as clusters which involves a word or phraseme that belongs to the AKL list (e.g. *claim, example, conclusion*). The analysis should, however, also include learner-specific clusters such as *according to me* or *as a conclusion* (cf. section 6.3.2.3).
- Manual extraction of all EAP key clusters from the raw lists of positive and negative 2-to-5-word clusters
- Quantitative analysis: compare the frequency and proportion of positive and negative 2-to-5 word EAP key clusters across learner corpora to investigate whether there are learner populations who rely more heavily on EAP key clusters or rather, make less frequent use of this type of word sequences.
- Qualitative analysis: analyse positive and negative EAP key clusters for each learner corpus (e.g. structure, whether they rather involve nouns, verbs or adjectives; whether they are native-like, etc.) and compare them
- Interpretation: link findings with variables such as first language and proficiency level

Such an analysis clearly lies beyond the scope of this thesis. However, preliminary results are so promising that I would like to follow the procedure just described in future research.

In what follows, I focus primarily on **positive EAP key clusters in French learner writing**. The analysis does not aim to be exhaustive. Rather, it seeks to identify general trends in French learners' use of EAP clusters by contrasting them with word sequences particularly salient in Spanish and Dutch learner writing. Selected lists of positive EAP key clusters in ICLE-FR, ICLE-SP and ICLE-DU are provided in Table 7.2 to 7.4.

As Table 7.2 shows, when compared to other L1 learner populations, French learners tend to make significantly more use of 5-word sequences such as *from an economic point of view, as a matter of fact, is no longer the case*; 4-word sequences such as *is far from being, I would say that, on the one hand, that is to say, have the opportunity to, we can see that, we can say that*; 3-word sequences such as *will have to, in order to, on the contrary, it will be, as* *far as, as a conclusion, take the example, according to me, in other words, will certainly be, more and more, we have to, it will be*; and 2-word sequences such as *let us, considered as, we will, we may, even if, for instance, the case, thanks to, such as, first step, to conclude,* and *one could.* Some of these sequences are also frequently used by Spanish learners (cf. Table 7.3). Examples include *that is to say, in order to, we will* and *as a conclusion*. However, most of them are specific to French learners and Spanish learners tend to use other sequences such as *there are a lot of, I would like to say, as a result of this, as we can see, it is sure that, in the case of, to sum up, as we can see, we can,* and *if we*.

Table 7.2: A selected list of positive EAP key clusters in ICLE-FR

5-word sequences	from an economic point of (8), as a matter of fact (21), it is not always easy (5), is no longer the case (5), it is no longer the (5), we must not forget that (5), I do not think so (5), we can say that the (5)
4-word sequences	take the example of (10), as far as the (26), is far from being (14), I would say that (19), be considered as a (11), will be able to (24), it will be possible (6), will have to be (13), people will have to (5), no longer the case (5), on the one hand (25), I think that we (6), that is to say (16), the aim is to (5), it would be difficult (5), but I am not (5), it is obvious that (17), for the sake of (13), must not forget that (5), we must not forget (7), to keep in mind (6), should not forget that (6), we can say that (14), I think that the (15), it is not always (7), it is true that (20), it is no longer (7), have the opportunity to (5), do not think so (5), we can see that (7), we will have to (7), I do not think (20), I think that this (8)
3-word sequences	will have to (61), in order to (151), as far as (72), it will be (40), considered as a (18), as a conclusion (22), on the contrary (50), take the example (10), not forget that (19), according to me (13), in other words (40), will be able (24), would say that (19), people will have (9), I would rather (6), we may wonder (6), point of view (65), far from being (16), I would say (20), be considered as (18), longer the case (5), we can wonder (5), was considered as (7), have the impression (7), matter of fact (21), we can say (22), the opportunity to (22), with regard to (12), will certainly be (8), will be easier (6), how could we (7), we must not (12), a kind of (33), will be possible (7), more and more (66), the example of (13), would it be (7), question that arises (5), it be possible (5), not always easy (5), a first step (5), will always remain (6), let us not (6), the one hand (25), we will be (12), that is to (18), there wont be (6), could argue that (5), it is obvious (20), we can notice (6), is obvious that (17), in that respect (5), the aim is (5), from all this (5), lets take the (5), is to say (16), as I am (8), is not always (13), to show that (10), be possible to (8), look at it (5), because we have (5), it may also (5), look at what (5), I have just (5), a means to (5), idea of a (9), in a particular (7), let us take (7), we are now (6), is true that (21), one has to (13), we have to (48), I mean that (8), the cause of (24), keep in mind (9), I think that (66), even if they (13), it would be (47), is not yet (5), existence of a (5), to be considered (8), must not forget (7), less and less (6), problems in the (6), this is also (6), and more important (6), aim of the (6), is true for (5), present in the (5), but is this (5), not think so (5), this lack of (5), think that this (10), some people think (7), is probably the (7), may not be (7), could say that (9)

2-word sequences	will be (287), let us (67), it will (89), order to (151), will also (33), in order (156), as far (72), far as (73), considered as (39), speak of (28), will certainly (26), the contrary (58), is concerned (45), will perhaps (10), we will (64), we may (34), people will (38), appear as (8), will not (79), is thus (13), other words (42), project is (7) point of (77), lead to (48), even if (63), this will (22), will speak (6), favour the (6), may wonder (6), perhaps be (6), imply the (6), concerned by (6), also see (6), forget that (24), for instance (73), can wonder (8), aims at (10), a conclusion (22), ideal solution (5), indeed we (5), it implies (5), of fact (21), more easily (13), notice that (17), a first (13), the opportunity (27), the case (60), a bit (31), with regard (12), be more (44), an important (47), this assertion (7), be difficult (11), must keep (10), wonder whether (10), how could (15), it mean (11), we could (35), are concerned (20), true for (12), thanks to (44), be possible (19), may also (14), most striking (8), not imply (7), such as (88), first step (11), far from (30), could we (10), imply a (6), here again (6), added to (9), to conclude (17), the example (14), can notice (8), would it (11), can say (27), into account (22), is precisely (7), now turn (5), opportunity to (29), think that (126), to assert (6), first part (6), be considered (28), certainly not (15), but rather (12), no more (19), consider the (23), a matter (30), always remain (7), intend to (7), would rather (10), regard to (12), can also (32), will lead (9), must not (21), and particularly (6), would say (21), and yet (12), one hand (30), but also (68), are thus (5), see why (5), implies the (5), will always (26),linked with (7), know whether (7), a means (9), but this (33), to show (33), so that (51), not yet (15), if necessary (6), will remain (9), to illustrate (8), implies that (8), we notice (5), illustrate this (5), viewed as (5), an aim (5), but is (18), above all (19), in fact (60), is certainly (17), this point (22), is obvious (23), could argue (5), examine the (5), can now (5), these examples (7), are indeed (6), not speak (6), it may (36), mean that (30), problems of (17), this means (18), consider that (9), this question (19), obvious that (21), a particular (19), say that (97), we might (11), certainly be (8), what would (13), one must (10), I would (67), may not (16), but we (38), ’s take (8), what does (12), are therefore (6), is interesting (6), we always (5), cannot deny (5), indeed the (5), will provide (5), we wont (5), one could (15), the same (188), refer to (7), also means (7), put forward (7), great part (7), we shall (9), contrary to (8), lets take (8)

Table 7.3: A selected list of positive EAP key clusters in ICLE-SP

5-word sequences	there are a lot of (16), as a result of this (5), I would like to say (6)
4-word sequences	that is to say (15), as we can see (6), as an example of (5), in the case of (17), it is sure that (5), there are a lot (17), we only have to (5), is the case of (5), we have to do (5), in order to get (10), in order to be (12), the case of the (7), there have always been (5), a result of this (5), my point of view (9), would like to say (6)
3-word sequences	in order to (101), to sum up (17), we can see (23), the case of (26), this is the (33), depending on the (9), of this paper (6), the base of (7), a lot of (107), is related to (6), is sure that (6), is to say (16), because of this (10), that is to (16), the fact of (8), can talk about (5), the aim of (12), but we are (8), the most important (37), at the end (22), be said that (11), it is sure (5), an example of (11), in the case (17), we think that (6), as we can (6), in this sense (9), lead us to (5), as for example (5), we only have (5), can be considered (7), as a conclusion (10), the idea that (11), important thing is (10), as we have (8), as if they (8), in this way (19), all these facts (4), in relation to (6), talk about the (7), there have always (5), in spite of (18), we talk about (6), at this point (7), because of the (25), but you can (5), it would be (36), the possibility of (11), because I think (5), why do we (5), and so on (22)

2-word sequences	capacity of (13), related to (24), we can (130), in order (107), order to (103), this situation (23), case of (40), *imposible to (5), *necesary to (5), the *posibility (5), sump up (17), this paper (10), the capacity (8), depending on (14), said that (37), *lets think (6), to sum (17), lot of (111), thought that (17), this fact (16), related with (5), talk about (23), if we (80), base of (7), people is (14), comparison between (5), because of (71), in relation (9), because we (20), would be (116), this is (119), in conclusion (14), offers us (6), capacity to (7), because there (16), a conclusion (14), as we (28), most important (39), the cases (8), can see (26), which we (22), is very (61), talking about (15), besides the (7), the importance (16), the aim (14), idea that (13), respect to (7), consider a (5), necessity of (9), a lot (118), better than (16), because it (44), must be (59), so as (12), is sure (7), essay is (7), way of (63), solution for (8), lack of (41), if they (50), so they (20), such as (61), can talk (5), ideas in (5), an important (31), problems that (8), be considered (20), this point (17), but you (11), we only (8), and probably (7), because that (5), you are (48), the possible (9), consequence of (11), we consider (10), due to (34), we talk (7), another important (9), you have (41), the essay (5), opinion about (5), can observe (5), cases are (5), we will (33), another point (7), this subject (7), relation to (7), it must (19), this case (18), we must (39), why do (12), to say (47), we think (14), topic of (5), that although (5), we just (6), things that (21)

EAP key clusters in ICLE-DU are rarely statistically significant in ICLE-SP and/or ICLE-FR. Most are specific to Dutch learners: *is to be found in, seem to be able, there is a difference, argument in favour of, the question is whether, take for instance, the first argument, closer look at, the principle of, and perhaps even, will be discussed, this means that, be held, likely that, was introduced* and *proved that* (cf. Table 3.4).

Unlike in ICLE-SP and ICLE-FR, preferred EAP clusters in ICLE-DU are generally not highly frequent sequences. Rather, they show a statistically significant difference when compared to other learner sub-corpora because they display very small frequencies in other learner corpora or are not used at all by other L1 learner populations. Thus, the sequences *the first argument* and *there is a difference* appear 5 times in ICLE-DU but are not found in other learner corpora; there are 11 occurrences of *will be discussed* in ICLE-DU but no occurrence of this cluster in other learner corpora; *argument in favour of* and *the question is* appear 8 [relative frequency per 100,000 words = 4.9] and 20 [relative frequency per 100,000 words = 12.33] times respectively in ICLE-DU vs. 4 [relative frequency per 100,000 words = 0.4] and 48 [relative frequency per 100,000 words = 4.78] times in the 9 other corpora. In addition, preferred EAP clusters in ICLE-DU differ from those found in ICLE-FR and ICLE-DU in a significant way: they often correspond to native-like EAP phrasemes and lexico-grammatical patterns that have been found to be underused by EFL learners when compared to native writing (cf. chapter 6). Dutch learners' use of EAP clusters thus seems to better approximate native speakers' EAP phraseology, which may be at least partly explained by their higher level of proficiency (cf. section 4.1.1). Another possible explanation is that Dutch academic writing shares more discourse conventions, and thus more lexically congruent academic-like sequences, with EAP than do French and Spanish academic discourse (cf. section 7.3.4).

A combined examination of negative key clusters in ICLE-FR, ICLE-SP and ICLE-DU shows that French, Spanish and Dutch learners exhibit totally different patterns of use. For example, the sequence *as far as* is a positive key cluster in the French learner sub-corpus but is a negative key cluster in the Spanish and Dutch learner corpora; the sequence *that is to say* is more frequently used by French and Spanish learners than by Dutch-learners; there is a marked preference for *a lot of* in ICLE-DU and ICLE-SP but this sequence is less often used by French learners. More interestingly, positive and negative EAP key clusters seem to point to major differences between French and Spanish learners vs. Dutch learners in their use of clusters involving the first person plural pronouns *we* and *us*. Unlike French and Spanish learners, Dutch learners do not make heavy use of a wide range of sequences involving *we* and *us*. The sequences *let us, because we, when we, what we, if we, and we, we can, we cannot, we did, we may, we must, we need, we shall, we should, we would, we think, we can not, we can see, we must be, we tend to*, and *we will see* are negative EAP key clusters in ICLE-DU. Even though French and Spanish learners behave similarly in relying heavily on sequences involving *we* and *us*, these sequences are different in ICLE-SP and ICLE-FR. Examples of positive clusters in French learner writing include *let us, we may, we might, we (can) notice, we can say, we have to,* and *we (can/may) wonder*. By contrast, the following sequences are positive clusters in Spanish learner writing: *we can, if we, because we, as we, we consider, we can get, we can observe, we can see, we can talk,* and *we could say*.

Each positive or negative EAP key cluster in ICLE-FR, ICLE-SP and ICLE-DU is certainly worthy of in-depth investigation to shed light on the respective influence of factors such as L1-influence, transfer of training and proficiency level. This clearly shows that a **corpus-driven approach can help identify lexical items which could serve as a new type of raw data for SLA studies, and more particularly, transfer studies**. It is however clearly outside the scope of this thesis to conduct all these analyses. In the next section, we will therefore show how learner corpus data can be used to investigate transfer effects on the basis of **five positive EAP key clusters in ICLE-FR**. We will make use of Jarvis's (2000) methodological framework to assess transfer and examine potential L1 influence on French learners' massive use of the following words and phrases:

Table 7.4: A selected list of positive EAP key clusters in ICLE-DU

5-word sequences	is to be found in (7), it is a fact that (10), seem to be able to (5), on the other side of (8), as a result of the (6)
4-word sequences	is whether or not (5), there is a difference (5), argument in favour of (8), a lot of people (34), the question is whether (8), is a fact that (10), as a result of (23), this is of course (7), on the other side (16), it is a fact (10), seem to be able (5), not solve the problem (5), to come up with (6), is to be found (8), can be found in (6), a closer look at (7), so that they can (6), in this essay I (11)
3-word sequences	will be discussed (11), a lot of (177), be held responsible (8), this essay will (10), the inequality of (7), whether or not (17), has many advantages (6), the actions of (6), as said before (6), in favour of (28), it was clear (5), is a difference (5), take for instance (5), the first argument (5), is whether or (5), be discussed in (5), become clear that (5), this means that (19), to deal with (24), argument in favour (8), the benefits of (10), fact that the (20), held responsible for (6), come up with (12), be found in (16), for instance the (7), in this essay (22), which means that (11), is a fact (12), so that they (20), there are still (21), be clear that (5), that the latter (5), results in a (5), the chances of (5), the main differences (5), the position of (17), is that you (8), question is whether (8), in that way (12), a fact that (11), the question is (20), the fact that (91), show that the (7), this is of (7), it might be (14); were said to (5), a very difficult (5), the principle of (7), people will be (7), a good thing (11), mean that the (9), the story of (6), a difference between (6), some people will (5), to be answered (5), and perhaps even (5), the preservation of (5), needs to be (11), a result of (24), the advantages of (9), then there is (6), a change of (6); the reasons why (6), the things that (10), closer look at (7), said to have (7), another way to (5), this of course (5), if you were (5), can conclude that (5), could have been (9), due to a (6)
2-word sequences	a lot (230), stability in (11), this is (209), lot of (180), essay will (10), whether or (18), clear that (34), deal with (39), this essay (44), as said (6), favour of (28), conclude that (15), in favour (29), found in (25), meant that (9), this was (25), major changes (5), advantage for (5), operated on (5), will probably (21), to handle (16), becomes clear (8), because they (104), argument in (9), fact that (109), is whether (12), another example (18), inequality of (7), to deal (24), an argument (8), maybe even (9), a possible (13), be discussed (13), the chances (6), that consists (6), first argument (6), characterised by (6), be given (21), this means (24), against the (53), we saw (7), people just (8), perhaps even (8), be held (8), consists of (20), the latter (26), become too (10), showed that (10), a change (18), benefits of (11), a fact (16), you could (18), the disadvantages (10), look at (54), a major (18), end up (15), come up (15), the change (14), exactly what (12), which means (17), explain what (5), was clear (5), is designed (5), is that (160), it became (10), you were (10), was shown (7), much influence (6), possible for (18), likely that (10), was introduced (7), people will (29), as possible (34), main differences (5), become clear (5), the first (108), it was (107), the current (18), be clear (8), policy of (8), the principle (10), proved that (13), it might (23), reasons why (12), the case (59), now that (19), a difference (10), people are (108), way to (62), will improve (6), this view (6), very clear (5), elements in (5), problems will (5), difference in (5), finally there (5), to conclude (16), means that (40), the advantages (17), be admitted (7), as is (7), discussed in (7), many advantages (7), can still (10), all sorts (14), closer look (9), said before (8), a series (8), and perhaps (12), a solution (15), be said (22), sort of (34), ways to (17), he was (49), cent of (13), usually not (6), provided by (6), another advantage (6), and secondly (6), stand for (6), possible that (10), an advantage (7), whether this (7), argument against (8), talking to (8), change of (11), certainly not (14), position of (20), when you (41), just like (15), were said (5), established to (5), can conclude (5), will even (5), will also (17), because he (24), seem to (58), so that (54), I found (11), which causes (6), why would (6), is seen (6), an alternative (7), if they (72), can also (32), can even (13), sorts of (17), show that (21)

- The textual phraseme *on the contrary* about which contradictory findings have been reported in the literature (cf. Granger and Tyson's (1996) argument in favour of L1 influence vs. Lake's (2004) claim against transfer effects in section 6.3.2.4).
- The learner-specific cluster *according to me* which has been found in several learner sub-corpora but is even more frequent in French learner writing
- Lexico-grammatical patterns with the verb *illustrate*: this verb is used by French learners in sequences such as *illustrate this* and *to illustrate* (cf. Table 7.2) while it is very infrequent in other learner corpora.
- The sequences *let us* and *for instance* which have been shown to be overused by EFL learners from different mother tongue backgrounds when compared to native writing but are particularly frequent in French learner writing.

7.3. Making use of learner corpora to prove L1 transfer

In section 3.3.3.4, a case was made for relying on multiple sources of evidence when assessing potential L1 influence. Jarvis (2000) proposes a unified framework for assessing transfer which incorporates the three types of comparisons described in section 3.3.3 (i.e. IL-L1 comparison, IL-IL comparison and a comparison of the interlanguage of learners sharing the same first language). This section thus aims to examine how Jarvis's framework can be operationalized and applied to learner corpus data. It also seeks to investigate which type of transfer effects can be identified by relying on corpus data.

Jarvis's (2000) framework is described in section 7.3.1. The corpus linguistics methods and the statistical measures used to operationalize Jarvis's framework on learner corpus data are described in section 7.3.2. Section 7.3.3 presents four case studies in which the framework is applied to assess L1 influence on positive EAP key clusters in ICLE-FR. Findings resulting from these case studies point to an additional dimension which may play a prominent role in transfer effects but is not addressed in Jarvis's unified framework, i.e. L1 frequency. This is the focus of section 7.3.4. Section 7.3.5 discusses possible limitations of the approach.

7.3.1. Jarvis's (2000) unified methodological framework

As already pointed out in section 3.3.3.4, Jarvis (2000) argues that much of the confusion concerning the nature of L1 influence could be eliminated if a minimal set of methodological standards were adopted by researchers who work in this area. He proposes a unified framework for the study of L1 influence that consists of three components: (a) a theory-neutral definition of L1 influence, (b) a statement of the types of evidence that must be examined to verify the phenomenon, and (c) a list of external and internal variables to be controlled.

7.3.1.1. A theory-neutral definition of L1 influence

The first component of Jarvis's (2000) unified framework is a theory-neutral definition of L1 influence that serves as a methodological heuristic for empirical studies. Jarvis considers that L1 influence is "underlyingly a unitary phenomenon (or a conglomeration of interconnected processes, constraints, and possibilities) whose essence lies beyond the reach of the researcher" (Jarvis 2000:253-254). However, he recognizes the need for a working definition that would be "a statement of the general empirical evidence that is needed to establish convincingly that an IL behavior exhibits L1-related effects" (Jarvis 2000:252-253). Jarvis further argues that a working definition of L1 influence should reflect Odlin's (1989) and Selinker's (1992) recognition of the need for statistical probabilities and should be broad enough to subsume the two types of evidence for L1 influence privileged by these two researchers (cf. section 3.3.3). He thus proposes the following working definition of L1 influence:

> L1 influence refers to any instance of learner data where a statistically significant correlation (or probability-based relation) is shown to exist between some features of learners' IL performance and their L1 background. (Jarvis 2000:252)

This definition is clearly intended as a **methodological heuristic** to be used by transfer researchers. It specifies that, to establish the presence of L1 influence, transfer studies must verify that there is a **statistically significant** relationship between IL performance and L1 background. Both types of comparison – an IL-IL comparison or an IL-L1 comparison - could in principle be used to show statistical significance and attest to the presence of L1 influence. As explained by Jarvis, "what is additionally needed is a specification of the types of statistical evidence that are necessary and sufficient to achieve methodological rigor in an investigation of L1 influence" (Jarvis 2000:252).

7.3.1.2. Types of evidence

Jarvis translates his working definition of L1 influence into a list of the specific types of L1 observable effects that must be examined when investigating transfer. He argues that transfer studies should minimally consider at least three potential effects of L1 influence when presenting a case for or against L1 influence:

(1) **Intra-L1-group homogeneity in learners' IL performance** is found when learners who share the same L1 background behave as a group with respect to a specific L2 feature. To illustrate this first L1 effect, Jarvis uses Selinker's (1992) finding that Hebrew-speaking learners of English as a group tend to produce sentences in which adverbs are placed before the object (e.g. *I like very much movies*) (cf. section 3.3.3.1). Intra-L1-group homogeneity is verified by comparing the interlanguage of learners sharing the same first language (cf. section 3.3.3.3).

(2) **Inter-L1-group heterogeneity in learners' IL performance** is found when "comparable learners of a common L2 who speak different L1s diverge in their IL performance" (Jarvis 2000:254). To illustrate this second L1 effect, Jarvis refers to a number of studies reported in Ringbom (1987) that have shown that Finnish-speaking learners are more likely to omit English articles and prepositions than Swedish-speaking learners are. As stated by Jarvis, "this type of evidence strengthens the argument for L1 influence because it essentially rules out developmental and universal factors as the cause of the observed IL behavior. In other words, it shows that the IL behaviour in question (omission of function words) is not something that every learner does (to the same degree or in the same way) regardless of L1 background" (Jarvis 2000:254-255). Inter-L1-group heterogeneity is examined by comparing the interlanguage of learners from different mother tongue backgrounds (cf. section 3.3.3.2).

(3) **Intra-L1-group congruity between learners' L1 and IL performance** is found where "learners' use of some L2 feature can be shown to parallel their use of a corresponding L1 feature" (Jarvis 2000:255). Selinker (1992) uses this type of evidence to show that Hebrew-speaking learners' positioning of English adverbs parallels their use of adverbs in the L1. The added value of this third L1 effect is that it also has explanatory power by showing "what it is in the L1 that motivates the IL behavior" (ibid). Intra-L1-group congruity is confirmed by a IL-L1 comparison (cf. section 3.3.3.1).

The three effects described above can emerge in circumstances in which transfer is not at play and can thus be misleading when considered in isolation (cf. section 3.3.3). As shown in

Table 7.5, Jarvis concludes that despite differences in degrees of reliability, none of the three effects is sufficient evidence by itself to verify or characterize L1 influence.

Table 7.5: Jarvis's (2000) three effects of potential L1 influence

L1 effect	reliability	sufficient criterion
Intra-L1-group homogeneity in learners' IL performance:	poor	not sufficient
Inter-L1-group heterogeneity in learners' IL performance:	strong	not sufficient
Intra-L1-group congruity between learners' L1 and IL performance:	strongest	not sufficient

According to Jarvis, the identification of two simultaneous L1 effects is necessary to present a convincing case for L1 influence. The researcher argues that identifying the three L1 effects would be even more convincing. However, he acknowledges that "the ubiquity of conditions that can obscure L1 effects renders the three-effect requirement unrealistic in many cases" (ibid).

Jarvis also stresses the need for transfer studies to consider L1 potential influence in relation to other factors that may influence learners' IL performance:

> "Of course, even when sufficient evidence does emerge to indicate a presence for L1 influence, this should not be construed as ruling out the existence of other potential factors that affect a learner's use of the L2 (...). Instead, it should be recognized that multiple factors may combine to influence a learner's use of the L2 at any given moment and at all stages of development". (Jarvis 2000:259)

7.3.1.3. External and internal variables

The third component of Jarvis's (2000) unified framework is a list of variables that ideally should be controlled, i.e. either held constant or actively investigated, in any rigorous transfer study. As Table 7.6 shows, the list includes learner-external variables, i.e. variables that "relate to the environment in which learning takes place" (Ellis 1994:24) (e.g. social, educational and cultural factors), and learner-internal variables (e.g. age, personality, language aptitude). Jarvis argues that it is only by controlling variables that researchers can verify whether other variables may conceal L1 influence or whether learners' IL performance is the result of transfer or is associated with another variable.

Table 7.6: Jarvis's (2000:260-261) list of variables (based on Ellis 1994)

- age
- personality, motivation, and language aptitude
- social, educational and cultural background
- language background (all previous L1s and L2s)
- type and amount of target language exposure
- target language proficiency
- language distance between the L1 and target language
- task type and area of language use, and
- prototypicality and markedness of the linguistic feature

7.3.2. Applying Jarvis's (2000) framework to learner corpus data: methodological issues

Jarvis (2000) applies his methodological framework to data from three elicitation tasks in order to investigate whether Finnish and Swedish learners differ in their choice of L2 words for referring to objects and events and, if so, whether the difference can be attributed to L1 influence. The purpose of this section is to examine how this framework can be operationalized and applied to learner corpus data, and more specifically to the *International Corpus of Learner English*. It may be argued that the texts produced by learners in Jarvis's task one, i.e. a written narrative or film retell, come quite close to learner corpus data. However, the way Jarvis makes use of learner texts is more experimental in nature. The objective of the film retell task is to elicit words used to refer to a controlled set of objects and events that appear in the 8-minute "Alone and Hungry" segment of Chaplin's silent film *Modern Times*. Learner texts are not further exploited and no corpus linguistics techniques are used.

Applying Jarvis's unified framework to learner corpus data poses a number of practical problems and raises a number of complex issues that are addressed here. Section 7.3.2.1 focuses on the feasibility of controlling the many variables listed by Jarvis that may interact with L1 influence or obscure its effects. Section 7.3.2.2 discusses how the three types of evidence that must be examined to establish L1 influence can be investigated in learner corpus data.

7.3.2.1. Controlling variables

The learner data used in this thesis consist of ten sub-corpora of the *International Corpus of Learner English*. As explained in section 4.1.1, each learner text in the corpus is documented with detailed information about task and learner variables (cf. Figure 4.1). A number of these variables can be described as 'constant' as they were used as corpus design criteria. All learners are young adults who study English as a Foreign Language at university. They are all in their second, third or fourth year and their level is described as advanced (but see section 4.1.1 for a note of caution about defining proficiency levels on the basis of external criteria). The sub-corpora used in this thesis were compiled on the basis of two learner variables, i.e. the **mother tongue variable** and **the language at home variable**, and three task variables, i.e. all texts are **untimed argumentative** essays potentially written with the help of **reference tools**. Thus, a large proportion of the variables listed by Jarvis (cf. section 7.3.1.3) are controlled in our learner data, e.g. age, target language proficiency, task type and area of language use. Like in Jarvis's study, however, the variable 'personality, motivation and language aptitude' is not controlled. Although they most probably play an important role in EFL learners' use of words and phrasemes, the following variables are not controlled: (1) type and amount of target language exposure and (2) previous knowledge of other L2s. Information about these variables are available in ICLE. However, controlling these variables could only have been done at the expense of learner corpus size (cf. Table 4.1).

Variables that are held constant among learner texts are assumed not to account for any possible interlanguage variation between L1 populations or sub-corpora. On the other hand, variables that are not controlled may interact with L1 influence so that caution will be needed when interpreting data.

7.3.2.2. Investigating the three potential L1 effects

The *International Corpus of Learner English* appears to be ideally suited to analyzing the three potential effects of L1 influence described by Jarvis (2000). As shown in Table 7.7, **intra-L1-group homogeneity** in learners' performance can be investigated by comparing all the essays written by learners who share the same mother tongue background to verify whether they behave as a group with respect to a specific L2 feature. **Inter-L1-group heterogeneity** in learners' IL performance can be highlighted by a comparison of all the essays produced by different L1 populations. To establish **intra-L1-group congruity**

between learners' L1 and IL performance, an L1 sub-corpus (e.g. all the essays written by the Spanish learners) is compared to a comparable corpus in the first language (e.g. Spanish).

Table 7.7: L1 effects and ICLE

L1 effect	**Corpus comparisons**
Intra-L1-group homogeneity in learners' performance	A comparison of all the essays included in a L1 sub-corpus (e.g. all the essays written by French learners)
Inter-L1-group heterogeneity in learners' IL performance	A comparison of several learner sub-corpora
Intra-L1-group congruity between learners' L1 and IL performance	A comparison of a L1 sub-corpus to a L1 comparable corpus

The following sections discuss how these corpus comparisons are made in order to assess the extent of the three potential L1 effects and which statistical measures are used to operationalize Jarvis's (2000) working definition of transfer (cf. section 7.3.1.1).

As illustrated in Table 7.8, in Jarvis's case study, L2 learners are grouped into six categories according to their mother tongue (Finnish or Swedish), age, grade, number of years of English instruction and number of years of L2 Swedish or Finnish instruction.

Table 7.8: Jarvis's (2000) L2 participant groups

Group	*n*	L1	Ages	Grade	English instruction	Swedish instruction
F5	35	Finnish	11-12	5	3rd year	NA
F7	35	Finnish	13-14	7	5th year	1st year
F9A	35	Finnish	15-16	9	7th year	3rd year
F9B	35	Finnish	15-16	9	3rd year	7th year
						Finnish instruction
S7	35	Swedish	13-14	7	3rd year	5th year
S9	35	Swedish	15-16	9	5th year	7th year

This grouping allows Jarvis to investigate L1 effects as well as the effects of outside variables on learners' choice of L2 words for referring to objects and events. Jarvis uses two types of statistical procedures to test the three effects: Pearson's bivariate correlation coefficient and Cronbach's alpha. The **Pearson correlation** procedure "tests for internal consistency between the word choice frequencies of only two groups at a time" while **Cronbach's alpha** procedure "produces a measure of internal consistency (based on the average interim correlation) between any two or more groups simultaneously" (Jarvis 2000:278). Thus, Pearson's bivariate correlation coefficients identify pairs of groups that behave similarly in terms of word choices, which helps correlate word choices with external variables. Cronbach's alpha procedures are used to compare overall levels of intra-L1-group homogeneity and inter-L1-group heterogeneity where all groups are considered simultaneously. It is important to note that **groups are compared and correlated on the basis of all lexical choices and that no lexical item is analyzed individually.**

Jarvis examines all three types of evidence in the English lexical reference of Finnish-speaking and Swedish-speaking Finns in three elicitation tasks which constitute relatively controlled types of data. Learners have to refer to objects and events in a picture and the number of lexical choices to do so is necessarily limited. Because his objective is to examine L1 influence in terms of group tendencies, Jarvis **omits from further analysis all lexical options chosen by fewer than 10% of all groups**.

The data used in this thesis is very different. First, essay writing is much less controlled: learners were not asked to exemplify, express a cause or contrast. As a result, the proportion of lexical items used by more than 10% of all groups may be quite small and it may not be appropriate to omit the other items from further analysis. This means, however, that our data will presumably include many more zero values. **Zero values** are very problematic for correlation tests and cause artificially high correlation coefficients as groups sharing a large proportion of zero values are quite understandably interpreted as being highly correlated. Second, our objective is to investigate the effects of L1 influence on a number of lexical items **individually** and not to relate the use of one item to that of another. For these two reasons, it is necessary to use other statistical procedures (e.g. statistical measures based on comparisons of means) for testing L1 effects on our data. In the following sections, the three L1 effects are discussed in detail on the basis of the example of French learners' use of the textual phraseme *on the contrary*. Statistical tests proposed to verify each L1 effect are also described.

7.3.2.2.1. Intra-L1-group homogeneity in learners' IL performance

Investigating intra-L1-group homogeneity in learners' IL performance amounts to answering the question of whether the very frequent use of a lexical item in the French learner corpus is due to **a few individuals** (i.e. a few essays) or whether it is found in a **significant proportion of French learners' texts** (cf. the notion of range in section 5.3.1.2). There are statistical measures which test for homogeneity or variability within a given population (e.g. standard deviation, variance). However, intra-L1-group homogeneity is likely to be quite poor as the type of data used in this thesis typically allows for variation: learners often have several L2 options available for serving an organizational or rhetorical function, e.g. *for example* and *for* *instance* to exemplify. In addition, essay writing is a less 'oriented' type of data (cf. Granger and Monfort 1994:69) and may encourage individual variation.

The problem here is to determine what is meant by 'a significant proportion'. Thus, the French learner corpus consists of 228 essays, out of which 42 (that is 18.4%) contain one or more occurrences of the phraseme *on the contrary*. It is very difficult to assess whether 18.4% is a significant proportion without comparing this figure to other percentages. A first method is to compare the percentage of occurrences of the phraseme to that of other words in the French learner corpus. Table 7.9 shows that the phraseme *on the contrary* appears in as many texts as relatively frequent words such as *quite, speak, here, while* and *ago*. However, this does not help verify whether French learners behave as a group with respect to the overuse of *on the contrary*.

Table 7.9: 'on the contrary' in the French learner corpus

Lexical item	Number of texts in the French learner corpus	%
on the contrary	42/228	18.4%
is, a, the, of	228/228	100%
important, new, now, each	114/228	50%
difficult, feel, look, idea	57/228	25%
quite, speak, here, while, ago	42/228	18.4%

Intra-L1-group homogeneity is necessarily a **relative concept** when less controlled types of data are used. It can only be established by **comparing intra-L1-group homogeneity across several L1 populations**. In section 6.3.2.1, it was shown that the phraseme *on the contrary* is generally overused by L2 learners, irrespective of their mother tongue backgrounds. Table 7.10 however indicates that the percentage of texts in which the phraseme *on the contrary* is used is much higher in the French learner corpus than in other learner corpora. Thus, the proportion of 18.4% is revealing when compared to the (much) smaller percentages found in other learner sub-corpora.

Table 7.10: 'on the contrary' in learner essays

	Number of texts including on the contrary	Number of texts	%
Czech	10	147	6.8%
Dutch	10	196	5.1%
Finnish	8	167	4.8%
French	***42***	***228***	***18.4%***
German	8	179	4.5%
Italian	11	79	13.9%
Polish	16	221	7.2%
Russian	22	194	11.3%
Spanish	12	149	8%
Swedish	6	81	7.4%
TOTAL	**145**	**1641**	**8.8%**

7.3.2.2.2. Inter-L1-group heterogeneity in learners' IL performance

> "We need to constantly refine our methods and develop new ones with an eye to what is happening in disciplines with similarly quantitative foci: computational linguistics, psycholinguistics, psychology, etc." (Gries 2006:191)

As explained in section 7.3.1.2, inter-L1-group heterogeneity in learners' IL performance is found when L2 learners who speak different mother tongues diverge in their IL performance. To investigate this L1 effect on the basis of corpus data, it is thus necessary to compare learner sub-corpora. Corpus research has long been concerned with issues related to corpus comparison. The objective of numerous corpus-based studies is to answer the following two questions (cf. Kilgarriff 2001:98):

- How similar are corpora?
- In what ways do corpora differ?

Statistical methods based on **word frequencies** are used to find the words that are more characteristic of one corpus as against another. Two widely used statistical measures in learner corpus research are the **chi-square** (X^2) (e.g. Granger 1998b; Lorenz 1999b; Cobb and Horst 2004; Neff et al. 2004) and the **log-likelihood** (LogL) (e.g. Leech et al 2001; Rayson and Gardside 2000; Rayson 2003; Paquot 2007[176]). Linguists such as Kilgarriff (1996; 2001) and Gries (2007), however, disapprove of corpus linguists' heavy reliance on this kind of statistics for a number of reasons, two of which are of particular relevance here.

[176] See also the methodology used in chapter 5 to extract EAP words and in chapter 6 to compare frequencies of lexical items in learner vs. native corpora.

First, chi-squares and similar tests are not appropriate measures for **multiple corpus comparisons**. For each comparison between two corpora, a statistical test such as the X^2 estimates the probability of making a Type I error, i.e. reject the null hypothesis when it is in fact true. This probability is referred to as the **error rate per comparison**. When multiple comparisons between two corpora are made, the probability of making a Type I error increases by the number of comparisons made. The probability that the family of comparisons will contain at least one Type I error is called the **familywise error rate**. The more comparisons are made, the more the Type I error rate is inflated (cf. Howell 1997: 98-101 for more information on Type I errors). Multiple comparisons should thus ideally be made with statistical tests that are corrected to control for familywise error rate.

Second, these statistical measures are usually applied on corpus data without considering what corpora are made of and without measuring corpus homogeneity. **Internal variables** such as number of texts per corpus, length of texts and number of texts per writer/speaker are rarely taken into account in statistical analyses of corpora. As underlined by Rietveld et al. (2004:350), "[t]he speaker or writer level normally does not appear in the analysis, and the data obtained from the different speakers or writers are pooled" (cf. also section 3.3.1.3). Gries (2007), however, regards **variability** as an essential issue in corpus linguistics as "corpora are inherently variable internally" and deplores the fact that "there is not much work that systematically explores the issues of variability within corpora (i.e., corpus homogeneity) and between corpora". In an earlier paper, Gries (2006) describes two types of analyses, i.e. by-subjects statistics and by-items statistics, that are common currency in psycholinguistics and psychology and whose main objective is "to determine to what degree the observed effects hold across subjects and items different from those actually investigated in the experiment" (ibid 192). He expresses surprise at the fact that "these methodological issues have barely found their way into corpus-linguistic studies" (ibid). Corpus-based studies most of the time do not show "whether the overall results in fact mask speaker/file- dependent results (i.e., what would correspond loosely to by-subjects statistics) and/or lemma-dependent results (i.e., what would correspond to by-items statistics)" (ibid 193).

For these two reasons, other approaches have been proposed to compare corpora, e.g. comparisons of means (Neff et al. 2004b; Thewissen et al. 2006; Gries 2007; Neff van Aertselaer to appear), log odds ratio (Rietveld et al 2004), Mann-Whitney ranks (Kilgarrif

1996; 2001) and high ratio pairs (Oakes 2003)[177]. **Comparison of means** tests take into account the distribution of searched items across the different texts (hence the different learners) that compose the learner corpora under study. They are used in this thesis[178] to establish inter-L1-group heterogeneity in learners' IL performance. These tests are conducted on the basis of data files of the type illustrated in Figure 7.1. Each line represents a text file from one of the learner corpora described in section 4.1. The two-letter code is an identification of the L1 sub-corpus (e.g. CZ = the Czech sub-corpus of ICLE, FR = the French sub-corpus of ICLE) from which the text file comes. The figure corresponds to the relative frequency of the searched item per 100 words. Given the differences in essay length within and between learner corpora, I thought it preferable to calculate relative frequencies per 100 words for each essay rather than performing statistical tests on absolute frequencies. The data file used thus consists of 1641 lines which represent the 1641 texts that compose the 10 learner corpora (cf. Table 4.1). I wrote a Perl program which takes the 10 learner corpora as input and generates data files for each searched item automatically.

Figure 7.1: Data files for comparisons of means tests

```
CZ  0
CZ  0
CZ  0
CZ  0
CZ  0
CZ  0.23
CZ  0
CZ  0
CZ  0
CZ  0
CZ  0
CZ  0
CZ  0
CZ  1.2
CZ  0
CZ  0
CZ  0
CZ  0.1972
CZ  0
CZ  0
CZ  0
CZ  0
CZ  0
CZ  0
```

[177] See Baroni and Evert (to appear) for more information on statistical methods for corpus exploitation.

[178] I am greatly indebted to Yves Bestgen who kindly made all comparison of means tests with SAS software for me (see http://www.sas.com/ or http://wwwsas.stat.ucl.ac.be/sasdiscute.html for more information on SAS software).

Comparisons of means are performed with the **one-way analysis of variance (ANOVA)** as more than two learner populations are compared[179]. The ANOVA procedure examines two sources of variance: the variance **between** the groups, i.e. corpora, and the variance between individuals or texts **within** each group, i.e. corpus. The two types of variance are then compared with one another. If the variance between the corpora is significantly higher than the variance within each corpus, the interpretation is that the two corpora are not taken from the same population (cf. Oakes 1998:22)[180]. The result of an ANOVA is an F ratio which tells us whether at least one corpus in the set is different from the other corpora. The F ratio for the phraseme *on the contrary* in learner corpora is 7.34 ($p < 0.0001$), which means that at least one learner population behaves differently from the other ones when using the phraseme.

A **post-hoc test** must then be conducted to identify the corpus or corpora responsible for the significant difference. Many post-hoc tests have been proposed in the body of literature devoted to statistics, e.g. Bonferroni, Newman-Keuls, Tukey, Scheffé test and the Ryan procedure. The **Ryan procedure** (REGWQ) is occasionally used in this thesis to compare the 10 learner sub-corpora with each other. The output of a Ryan test is a set of L1 corpus groups within which differences between means are not statistically significant. In Figure 7.2, learner corpora are classified by decreasing mean of occurrence of *on the contrary* per essay (calculated on the basis of relative frequencies per 100 words). The third column (grouping with letter A) shows that, eight learner populations behave as a group when it comes to their use of *on the contrary*. The difference between their respective means is not statistically significant. Two learner populations differ from this group. The fourth column (grouping with letter B) shows that including the Italian learner population in the group is done at the expense of the Czech, Finnish and German learner populations. Italian learners' use of *on the contrary* is similar to that of Russian (RU), Swedish (SW), Polish (PO), Spanish (SP) and Dutch (DU) learners but, unlike these learner populations, it is statistically different from that of Czech (CZ), Finnish (FI) and German (GE) learners. The third column reveals that French learners' use of *on the contrary* differs significantly from all learner populations expect for Italian learners.

[179] When only two corpora are compared, a Student's t-test for independent groups should be used.

[180] See Howell (1997:299-347) for more information on ANOVA.

Figure 7.2: 'on the contrary': Ryan test

Corpus	Mean		Grouping	
FR	0.037400			C
IT	0.026173		B	C
RU	0.013538	A	B	
SW	0.012732	A	B	
PO	0.012049	A	B	
SP	0.011244	A	B	
DU	0.009103	A	B	
CZ	0.007698	A		
FI	0.006813	A		
GE	0.005951	A		

It is important to note that in order to compare the 10 learner corpora used in this thesis, a Ryan procedure makes 45 comparisons of means. As explained above, the problem of multiple comparisons can be described as the potential increase of Type I error (and familywise error rate) that occurs when statistical tests are used repeatedly. A simple example is provided by Salkind (2005):

> Imagine this scenario. You're a high-powered researcher at an advertising company, and you want to see if color makes a difference in sales. And you'll test this at the .05 level. So you put together a brochure that is all black and white, one that is 25% color, the next 50%, then 75%, and finally, 100% color, for five different levels. But since ANOVA is an omnibus test[181], you don't know where the source of the significant difference lies. So you take two groups at a time (such as 25% color and 75% color) and test them against each other. In fact, you test every combination of 2 against each other. Kosher? No way. This is called performing multiple t tests, and it is actually against the law in some juridictions. When you do this, the Type I error rate (which you set at .05) balloons depending on the number of tests you want to conduct. There are 10 possible comparisons (no color vs. 25%, no color vs. 50%, no color vs. 75%, etc.), and the real Type I error rate is $1 – (1 - \alpha)^k$, where
>
> α is the Type I error rate, which is .05 in this example
> k is the number of comparisons
>
> So, instead of .05, the actual error rate that each comparison is being tested at is .22, or
>
> 1 – (1 - .05)10 = .40 (!!!!!)
>
> Surely not .05. Quite a difference, no? (Salkind 2005:204-205)

In order to control for familywise error rate, post-hoc tests such as the Ryan procedure adjust the level of significance so that each comparison is controlled at the same significance level

[181] An 'omnibus test' tests for an overall difference between means.

(often .05), which means that the more comparisons are made, the more conservative the test becomes. A general principle in statistics is thus to make no more comparisons than are actually needed.

As the objective of this chapter is to investigate L1 effects in French learners' use of lexical items, the comparisons we are mainly interested in are those between the French learner corpus (i.e. the control corpus) and the other learner corpora. We thus make use of **Dunnett's test** as it is considered to be the most powerful[182] post-hoc test whenever one group is compared with each of the other groups (cf. Howell 1997:380-381). Table 7.11 shows the output of Dunnett's test for the phraseme *on the contrary*. Results offer direct evidence of (at least partial) inter-L1-group heterogeneity. The French learner corpus is compared to each corpus, from which it is shown to differ significantly, except for the Italian learner corpus.

Table 7.11: 'on the contrary': Dunnett's test

Corpus comparison	Simultaneous lower confidence limit	Difference between means	Simultaneous upper confidence limit	Significance[183]
IT-FR	-0.029202	-0.011226	0.006749	
RU-FR	-0.037311	-0.023862	-0.010413	***
SW-FR	-0.042477	-0.024667	-0.006857	***
PO-FR	-0.038348	-0.025351	-0.012353	***
SP-FR	-0.040660	-0.026156	-0.011651	***
DU-FR	-0.041708	-0.028296	-0.014885	***
CZ-FR	-0.044266	-0.029701	-0.015137	***
FI-FR	-0.044610	-0.030586	-0.016562	***
GE-FR	-0.045199	-0.031449	-0.017699	***

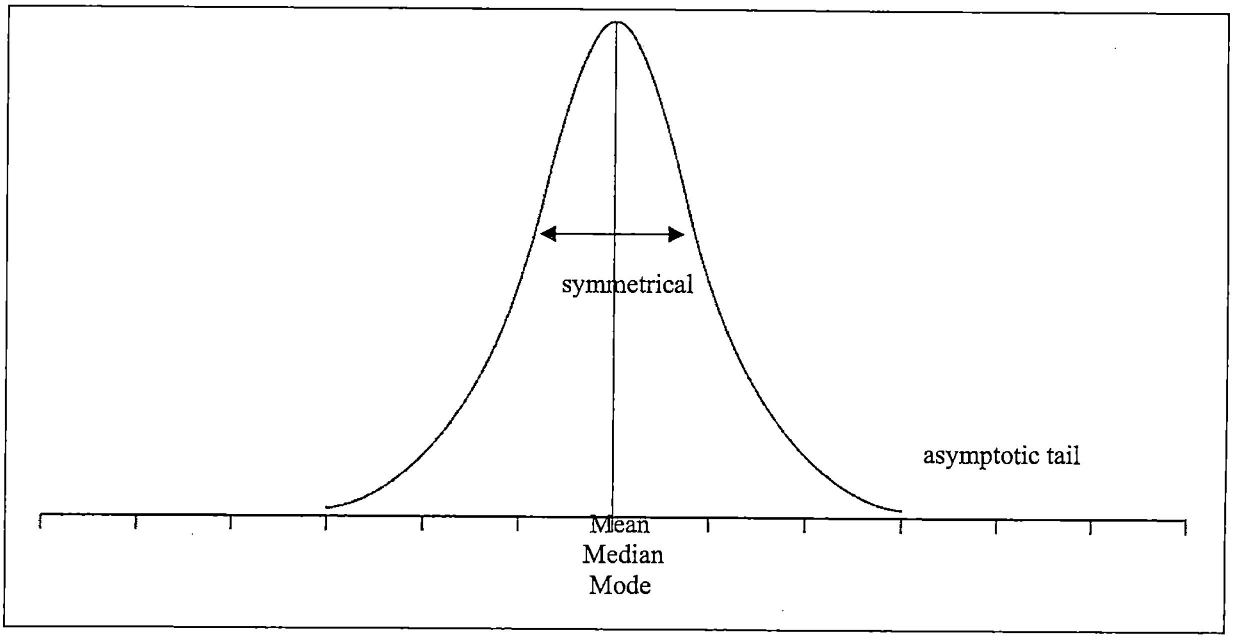
The ANOVA test, the Ryan procedure and the Dunnett's test are all **parametric** tests. They are based on the assumption that the populations from which the samples (i.e. the corpora) are drawn are normal. A normal distribution is represented by a bell-shaped curve (cf. Figure 7.3) which has the following three characteristics:

a. The mean, median and mode are equal to one another.
b. The bell-shaped (or normal) curve is perfectly symmetrical
c. The tails of the bell-shaped curve are asymptotic, i.e. they come closer and closer to the horizontal axis but never touch (cf. Salkind 2005:119-120).

[182] As explained by Salkind, "[p]ower is a construct that has to do with how well a statistical test can detect and reject a null hypothesis when it is false" (2005:148).

[183] Comparisons significant at the 0.05 level are indicated by '***'.

Figure 7.3: The normal distribution



Our data, however, are not normally distributed as they include a large proportion of zeros. The remaining values are mainly comprised between 0 and 1. Statistically-minded people could thus criticize our choice of parametric tests and argue that we should have used non-parametric tests such as the Wilcoxon's rank sum test or the Kruskal-Wallis One-way Analysis of Variance. According to Howell (1997:646), those who argue in favor of using parametric tests "argue, however, that the assumptions normally cited as being required of parametric tests are overly restrictive in practice and that the parametric tests are remarkably unaffected by violations of distribution assumptions." Moreover, parametric tests are said to be more powerful than non-parametric tests: they require fewer observations than do non-parametric tests and are more likely "to lead to rejection of a false null hypothesis" (ibid) than are their corresponding non-parametric tests. This advantage seems to be maintained "even when the distribution assumptions are violated to a moderate degree" (ibid).

7.3.2.2.3. Intra-L1-group congruity between learners' L1 and IL performance

The first question that needs to be addressed to investigate intra-L1-group congruity between French learners' L1 and IL performance is whether there is a French lexical item that is congruent with the English phraseme *on the contrary*. The phraseme *au contraire* appears as a straightforward translation equivalent to *on the contrary*. A related question is whether French learners' very frequent use of *on the contrary* corresponds to a very frequent use of its equivalent in French (cf. section 3.3.3.1; Selinker 1992; Guillot 2005). To answer this question, I compare the essays written in English by French-speaking students comprised in ICLE-FR to the *Corpus de Dissertations Françaises* (CODIF), a 225,174-word comparable corpus of essays written by French-speaking students collected at the University of Louvain. As shown in Table 7.12, the relative frequency of occurrences of *on the contrary* in the interlanguage of French learners is much closer to the relative frequency of *au contraire* in the essays written in French by French students than to the relative frequency of *on the contrary* in the essays written by English students.

Table 7.12: L1-interlanguage (IL) performance similarities – En. 'on the contrary' vs. Fr. 'au contraire'

	Relative freq. per 100,000 words
French EFL learners (ICLE-FR)	36.67
French students writing in French (CODIF corpus)	27
English students writing in English (STUD-US-ARG)	1.2

As the three L1 effects are found, it seems reasonable to conclude that conceptual problems and misguided teaching practices interact with L1 influence in French learners' use of *on the contrary*. This strongly supports Granger and Tyson's (1996) suggestion that French learners' overuse and misuse of *on the contrary* is probably due to an over-extension of the semantic properties of Fr. *au contraire*, which can be used to express both a concessive and antithetic link (cf. section 6.3.2.4).

7.3.3. Jarvis's (2000) framework applied

In this section, we follow the procedure described in section 7.3.2 and apply Jarvis's (2000) framework on four positive EAP key clusters in ICLE-FR: the learner-specific sequence *according to me*, the verb *illustrate*, the first person plural imperative form *let us* and the monolexemic phraseme *for instance*. Table 7.13 provides a synthesis of the methodology that I propose to use to operationalize Jarvis's framework on learner corpus data. It represents the different steps that need to be followed to investigate transfer effects on French learners' use of a specific interlanguage feature.

Table 7.13: Jarvis's (2000) framework applied to learner corpus data

SOURCE OF EVIDENCE	TYPE OF COMPARISON	CORPUS	STATISTICS
Intra-L1-group homogeneity in learners' IL performance	IL_a-IL_a	(ICLE-FR vs. ICLE-FR) vs. essays in other learner sub-corpora	% of essays in a learner corpus
Inter-L1-group heterogeneity in learners' IL performance	IL_a-IL_b	ICLE-FR vs. other learner sub-corpora	ANOVA Dunnett's test Ryan test
Intra-L1-group between learners' L1 and IL performance	IL-L1	ICLE-FR vs. CODIF	relative frequencies

As Jarvis explains, the hypotheses formulated to address the three types of evidence for L1 influence required by his methodological framework predict that "learners will produce (statistically demonstrable) high levels of intra-L1-group homogeneity (Hypothesis 1), low levels of inter-L1-group homogeneity (hypothesis 2) and high levels of L1-IL congruence (hypothesis 3)" (Jarvis 2000:289) in their use of specific lexical items.

7.3.3.1. French learners' use of 'according to me'

De Cock (2003) compares 2-to-5-word sequences in ICLE-FR and LOCNESS and finds that French learners overuse the 'idiosyncratic'[184] combination *according to me* (cf. section 1.4.2). In section 7.2, we compare 2-to-5-word sequences in ICLE-FR vs. other learner corpora and find that the use of *according to me* seems to be specific to French learners. An analysis of the 1641 learner essays reveals that this sequence only appears in 12 French learners' texts (5.26%), 5 Dutch learners' texts (2.55%) and 1 text written by a Swedish learner (1.23%) (cf. Table 7.14). It is difficult to interpret the small proportion of occurrences of *according to me* in French learners' essays (5.26%) as showing intra-L1 group homogeneity. This figure is, however, revealing when compared to the zero values found in most learner corpora.

[184] Note that the acceptability of this sequence is not quite clear. It occurs very rarely in the BNC (0.06 per million words) but seems to be gaining ground. However, monolingual learners' dictionaries such as the *Longman Dictionary of Contemporary English* advise against the use of *according to me*. What seems clear, however, is that it is not used in academic writing.

Table 7.14: 'according to me' in learner essays

	Rel. freq. of according to me per 100,000 words	Number of texts including according to me	Number of texts	%
French	***9.53***	***12***	***228***	***5.26%***
Czech	0	0	147	0%
Dutch	3.08	5	196	2.55%
Finnish	0	0	167	0%
German	0	0	179	0%
Italian	0	0	79	0%
Polish	0	0	221	0%
Russian	0	0	194	0%
Spanish	0	0	149	0%
Swedish	2.08	1	81	1.23%
TOTAL	**1.63**	**18**	**1641**	**1.10%**

Comparisons of means tests are very sensitive to zero values, which often make their results unreliable. However, a Dunnett's test shows that even though *according to me* appears in the Swedish learner corpus, Swedish learners' use of this expression differs significantly from that of French and Dutch learners (cf. Table 7.15). Thus, we can quite safely state that effect n°2, i.e. inter-L1 group heterogeneity in learners' IL performance, is found in French learners' use of *according to me*: ICLE-FR differs from all learner corpora except for ICLE-DU. We will come back to Dutch learners' use of *according to me* below in this section.

Table 7.15: 'according to me': Dunnett's test

Corpus comparison	Simultaneous lower confidence limit	Difference between means	Simultaneous upper confidence limit	Significance[185]
DU-FR	-0.009394	-0.004355	0.000684	
SW-FR	-0.014021	-0.007330	-0.000638	***
PO-FR	-0.013782	-0.008899	-0.004016	***
CZ-FR	-0.014790	-0.009318	-0.003846	***
GE-FR	-0.014484	-0.009318	-0.004151	***
FI-FR	-0.014587	-0.009318	-0.004049	***
RU-FR	-0.014371	-0.009318	-0.004265	***
SP-FR	-0.014767	-0.009318	-0.003868	***
IT-FR	-0.016071	-0.009318	-0.002564	***

Investigating Jarvis's (2000) third effect of L1 influence here amounts to checking whether there is congruity between French learners' use of the idiosyncratic expression *according to me* in L2 and a specific form in L1. In the *Corpus de Dissertations Françaises* (CODIF), the phraseme *selon moi* is used to express opinion, a sequence which can quite

[185] Comparisons significant at the 0.05 level are indicated by '***'.

safely be assumed to be responsible for French learners' use of *according to me* in English. English *according to* and French *selon* are most probably regarded as direct translation equivalents. They share a number of semantic features which are well explained in LDOCE4 for *according to*:

> 1. as shown by something or stated by someone:
> *According to the police, his attackers beat him with a blunt instrument.*
> *There is now widespread support for these proposals, according to a recent public opinion poll.*
> ! Do not say 'according to me' or 'according to my opinion/point of view'. Say 'in my opinion'
> *In my opinion his first book is much better.*
> 2. in a way that depends on differences in situations or amounts:
> *You will be paid according to the amount of work you do.*
> 3. in a way that agrees with a system or plan, or obeys a set of rules:
> *The game will be played according to rules laid down for the 1992 Cup.*
> *Everything went according to plan, and we arrived on time.* (LODCE4)

However, they differ in one significant way: while *according to me* is usually not accepted as a correct English phraseme (cf. sentence that warns against the use of *according to me* under the entry *according to* in LDOCE4 as reprinted above), *selon moi* is definitely a French phraseme. In addition, it is quite frequent in French student writing (relative frequency of 13.77 occurrences per 100,000 words), which may explain why French EFL learners are willing to use what they regard as a direct translation of a perfectly correct French lexical item. The following sentences illustrate French students' use of *selon moi* and French EFL learners' use of *according to me*:

7.1. ***Selon moi,*** *la chanson est un vecteur de culture parce qu'elle est un art qui impose l' engagement des différents acteurs.* (CODIF)

7.2. ***Selon moi,*** *tout le monde pense ce qu'il veut et comme il veut, agit comme il l'entend en respectant la loi et les codes établis.* (CODIF)

7.3. ***Selon moi,*** *ne connaître qu'une seule langue représente un certain handicap, c'est se condamner à l' isolement et à l' ignorance.* (CODIF)

7.4. ***According to me,*** *there is some danger behind the notion of europeanisation : monotony, boredom and loss of identity may arise if all the different cultures are reduced to one culture.* (ICLE-FR)

7.5. ***According to me,*** *the real problem now is not that man refuses to pay heed but that man refuses to make some sacrifices for the sake of ecology and to understand that the values that we have chosen are the wrong ones.* (ICLE-FR)

7.6. ***According to me,*** *the prison system is not outdated: it has never been a solution per se.* (ICLE-FR)

Figure 7.4 tries to represent schematically how the misleading translation equivalent is created by French EFL learners.

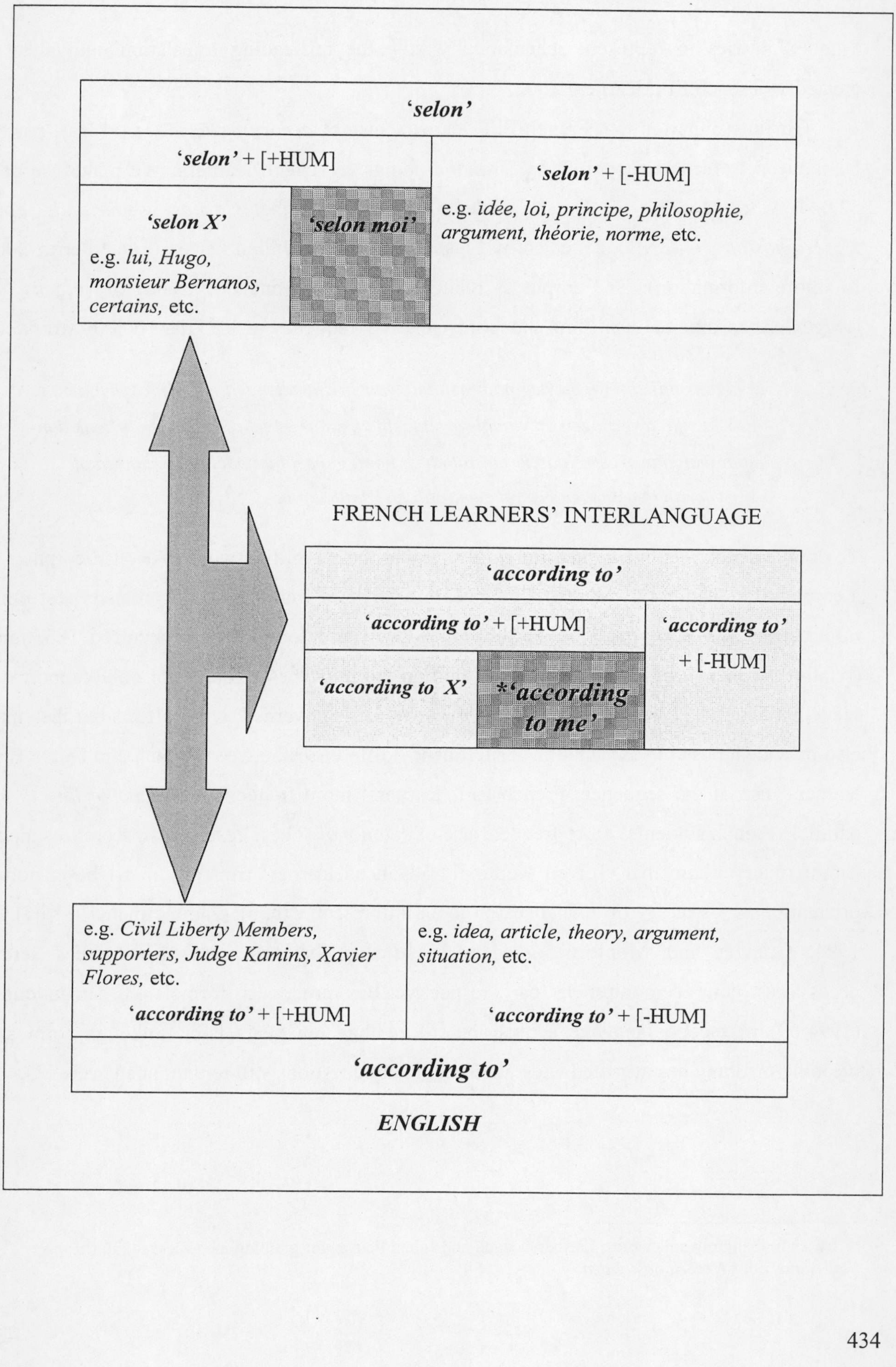
Our interpretation that French EFL learners' use of *according to me* is (at least partly) L1-induced is further supported by similar findings for Dutch learners. We make use of a 52,044 word corpus of Dutch student writing, i.e. the *Corpus Nederlands door Nederlandstaligen* (CNN), collected by Liesbet Degand and Julien Perrez[186] (see Perrez 2006 for more information). The corpus is relatively small but one occurrence of *volgens mij* (*volgens* 'according to' + *mij* 'me') is found (relative frequency of 1.85 per 100,000 words):

7.7. ***Volgens mij*** *zijn we op weg naar een tijd waarin dialecten stilaan hun communicatieve kracht zullen verhogen op vlaams grondgebied, maar in vele streken, en ik denk dan spontaan aan West-Vlaanderen, zullen zij nooit geheel hun poëtische klanken of verwoordingen moeten bestoft zien; althans, dat hoop ik.* (CNN)

To double-check this, we looked for *volgens mij* in the 20-million word PAROLE corpus, i.e. a corpus of contemporary written Dutch searchable via a web-based concordancer (for more information, see http://parole.inl.nl/html/index.html) and found that it occurred 163 times (relative frequency of 8 occurrences per million words). These results not only support our view that Dutch EFL learners' use of *according to me* is governed by L1 effects but they may also provide a possible explanation for frequency differences between French and Dutch EFL learners' use of the sequence. French EFL learners' more frequent use of *according to me* parallels French students' more frequent use of *selon moi*. These results also seem to support Zimmerman's claim that "[e]ven seemingly advanced learners can rely on L1-based form-orientation as a strategy of lexical search to an unforeseen extent" (Zimmerman 1987:65). In 1994, Granger and Monfort argued that "[d]es recherches plus approfondies seront nécessaires pour démontrer la part respective des processus formels et sémantiques" (1994:69) in second language acquisition. More than ten years later, although form and meaning mapping has attracted wide attention, many questions still remain unanswered.

[186] I wish to thank my colleagues, Liesbet Degand and Julien Perrez, for granting me access to the *Corpus Nederlands door Nederlandstaligen*.

Figure 7.4: Fr. 'according to me' in French learners' interlanguage


'selon'
'selon' + [+HUM]
'selon' + [-HUM]
'selon X'
e.g. lui, Hugo, monsieur Bernanos, certains, etc.
'selon moi'
e.g. idée, loi, principe, philosophie, argument, théorie, norme, etc.
FRENCH LEARNERS' INTERLANGUAGE
'according to'
'according to' + [+HUM]
'according to' + [-HUM]
'according to X'
*'according to me'
e.g. Civil Liberty Members, supporters, Judge Kamins, Xavier Flores, etc.
'according to' + [+HUM]
e.g. idea, article, theory, argument, situation, etc.
'according to' + [-HUM]
'according to'
ENGLISH

7.3.3.2. French learners' use of patterns with the verb 'illustrate'

In section 7.2, French EFL learners have been shown to overuse the 2-word sequences *to illustrate* and *illustrate this* compared with other learner populations. An investigation of intra-L1 group homogeneity shows that the verb *illustrate* only appears in 8.33% of the texts written by French learners. This proportion, however, is significantly higher than that found in other learner corpora. As shown in Table 7.16, the percentage of texts including *illustrate* in other learner populations ranges from 0% in ICLE-SW to 3.57% in ICLE-DU.

Table 7.16: 'illustrate' in learner essays

	Rel. freq. of the verb 'illustrate' per 100,000 words	Number of texts including the verb 'illustrate'	Number of texts	%
French	***14.67***	***19***	***228***	***8.33%***
Czech	3	3	147	2.04%
Dutch	4.31	7	196	3.57%
Finnish	1.60	2	167	1.20%
German	1.83	2	179	1.12%
Italian	2.09	1	79	1.27%
Polish	2.85	4	221	1.81%
Russian	3.62	6	194	3.09%
Spanish	6.05	5	149	3.36%
Swedish	0	0	81	0%
TOTAL	**4.46**	**49**	**1641**	**2.9%**

A Dunnett's test shows that French learners' use of the verb *illustrate* differs from that of most other learner populations except for Spanish learners (cf. Table 7.17). It can thus quite safely be stated that there is inter-L1 group heterogeneity.

Table 7.17: the verb 'illustrate': Dunnett's test

Corpus comparison	Simultaneous lower confidence limit	Difference between means	Simultaneous upper confidence limit	Significance[187]
SP-FR	-0.015073	-0.007018	0.001037	
DU-FR	-0.014929	-0.007481	-0.000033	***
RU-FR	-0.016140	-0.008672	-0.001203	***
CZ-FR	-0.018343	-0.010255	-0.002167	***
PO-FR	-0.017791	-0.010573	-0.003356	***
GE-FR	-0.018475	-0.010839	-0.003204	***
FI-FR	-0.020007	-0.012219	-0.004432	***
IT-FR	-0.022360	-0.012378	-0.002396	***
SW-FR	-0.023294	-0.013404	-0.003513	***

[187] Comparisons significant at the 0.05 level are indicated by '***'.

The English verb *illustrate* has a formal equivalent in French, i.e. *illustrer*, which has a relative frequency of 11.55 occurrences per 100,000 in CODIF. Table 7.18 shows that the verb is very frequently used in its infinitive form (example 7.8) or in simple present (example 7.9) in CODIF.

7.8. ***Pour illustrer cela,*** *prenons l'exemple des pâtes alimentaires italiennes.* (CODIF)

7.9. *Françoise Dolto qualifie d'ailleurs cet état de « complexe du homard »* ***afin d' illustrer*** *la position de faiblesse dans laquelle le jeune se trouve.* (CODIF)

Table 7.18: En. ‘illustrate’ and Fr. ‘illustrer’

	En. ‘illustrate’ in ICLE-FR		**Fr. ‘illustrer’ in CODIF**	
simple present	10	50%	8	30.77%
infinitive	***8***	***40%***	***13***	***50%***
past participle	***2***	***10%***	***3***	***11.54%***
imperative	0	0%	1	3.85%
past	0	0%	1	3.85%
TOTAL	20	100%	26	100%
Rel. freq. per 100,000 words	14.67		11.55	

The percentage of infinitive forms of the French verb *illustrer* thus differs significantly from the 23.55% of infinitive forms of the English verb *illustrate* that were found in BNC-AC-HUM (see Table 6.6; section 6.2.1.2). Similarly, past participle forms represent 11.54% of the occurrences of the verb (example 7.10), which differs significantly from the 29.73% of past participle forms of the verb *illustrate* found in academic writing.

7.10. *Voilà deux exemples d'identification aux personnages mais l'identification à l'intrigue, qu'en est-il ? Celle-ci peut être* ***illustrée*** *par l'adaptation en pièce de théâtre du roman de George Orwell, « 1984 », où la société est dirigée par un big brother qui a l'œil sur tout.* (CODIF)

It may thus be tentatively put forward that EFL learners' use of En. *illustrate* resembles French students' use of Fr. *illustrer* in two significant ways. First, the **frequency** of En. *illustrate* in ICLE-FR is closer to that of Fr. *illustrer* in CODIF than its frequency in BNC-AC-HUM (or STUD-US-ARG). Second, French EFL learners repeatedly use En. *illustrate* in infinitive structures, i.e. the most frequent **lexico-grammatical pattern** of its French counterpart (examples 7.11 and 7.12).

7.11. ***To illustrate this,*** *we can mention the notion of culture and language in the north of Belgium.* (ICLE-FR)

7.12. ***To illustrate this*** *point, it would be interesting to compare our situation with the U.S.A's.* (ICLE-FR)

By contrast, French learners do not make use of the past participle of the verb *illustrate* as often as professional native writers (BNC-AC-HUM). The frequency of the past participle of *illustrate* in ICLE-FR is again closer to the frequency of its French equivalent form in CODIF.

French learners' use of the verb *illustrate* is a clear example of what was referred to in section 3.2.4 as **transfer of use**. French learners' knowledge of the French verb *illustrer* most probably influences their knowledge of the English verb *illustrate* by transferring the lexico-grammatical preferences of the French verb to its English counterpart.

7.3.3.3. French learners' use of 'let us'

In section 6.3.2.3.1, the first personal plural imperative form *let us* has been shown to be overused by all L1 learner populations when compared to professional academic writing. In section 7.2, *let us* was also found to be much more frequent in ICLE-FR than in other L1 learner sub-corpora.

The percentage of French learners' essays which include one or more occurrences of *let us* (or its contracted form *let's*) is 25.88%, a proportion which is higher than that found for *on the contrary* (cf. section 7.3.2.2.1). Table 7.19 shows that this proportion is higher than that of all other learner populations. The difference is quite marked except for Russian learners. These results suggest that we can quite safely conclude that Jarvis's (2000) first L1 effect, i.e. intra-L1 group homogeneity in learners' IL performance, is found in French learners' use of English first person plural imperative.

Table 7.19: 'let us' in learner essays

	Rel. freq. of let us and 'let's' per 100,000 words	Number of texts including let us or 'let's'	Number of texts	%
French	***71.88***	***59***	***228***	***25.88%***
Czech	25.24	19	147	12.92%
Dutch	12.33	19	196	9.69%
Finnish	8.78	10	167	5.99%
German	13.69	14	179	7.82%
Italian	20.95	10	79	12.66%
Polish	19.21	23	221	10.40%
Russian	38.57	47	194	24.23%
Spanish	26.23	14	149	9.39%
Swedish	18.73	9	81	11.11%
TOTAL	**26.85**	**224**	**1641**	**13.65%**

A Dunnett's test confirms that French learners' use of *let us* differs from that of all other learner populations. Table 7.20 shows that there is a significant difference between ICLE-FR and all other learner corpora. This also holds true for the Russian learner sub-corpus in which there is a similar proportion of texts which include the first person plural imperative form *let us* or its contracted form *let's*. A plausible explanation for this difference is that in ICLE-FR, *let us* often appears more than once in a text while this is not true for ICLE-RU.

Table 7.20: 'let us': Dunnett's test

Corpus comparison	Simultaneous lower confidence limit	Difference between means	Simultaneous upper confidence limit	Significance[188]
RU-FR	-0.052969	-0.030214	-0.007459	***
CZ-FR	-0.068352	-0.043710	-0.019067	***
SP-FR	-0.069735	-0.045193	-0.020652	***
IT-FR	-0.080742	-0.050327	-0.019912	***
PO-FR	-0.072695	-0.050704	-0.028712	***
SW-FR	-0.083752	-0.053617	-0.023483	***
DU-FR	-0.079522	-0.056830	-0.034137	***
GE-FR	-0.081028	-0.057763	-0.034498	***
FI-FR	-0.084753	-0.061024	-0.037296	***

A Ryan test shows that there is no difference between L1 learner populations as far as their use of *let us* is concerned except for French learners. As illustrated in Figure 7.5, all learner populations form a group and the French learner sub-corpus stands apart. These findings clearly point to the second L1 effect described by Jarvis (2000), i.e. inter-L1 group heterogeneity in learners' IL performance.

[188] Comparisons significant at the 0.05 level are indicated by '***'.

Figure 7.5: 'let us': Ryan test

Corpus	Mean	Grouping	
FR	0.07141	A	
RU	0.04120		B
CZ	0.02770		B
SP	0.02622		B
IT	0.02108		B
PO	0.02071		B
SW	0.01779		B
DU	0.01458		B
GE	0.01365		B
FI	0.01039		B

There is no lexically equivalent form to En. *let us* in French. Equivalence is however found at the morphological level as French makes use of an inflectional suffix to mark the first imperative plural form. Thus, to investigate the third L1 effect, i.e. intra-L1 group congruity between learners' L1 and IL performance, we compare the frequency of *let us* in ICLE-FR with that of first person plural imperative verbs in CODIF. Table 7.21 shows that French EFL learners' use of *let us* is much closer to the frequency of first imperative plural verbs in French than in English. This suggests that the **frequency** of a lexical item in EFL learners' first languages may be reflected in their interlanguage[189].

Table 7.21: Imperative plural verbs in French L1, English L2 and English L1

	Total number of words	Number of 1st imperative plural verbs	Relative frequency per 100,000 words
French EFL learners (ICLE-FR)	136,343	98	71.88
French L1 students (CODIF)	225,174	215	95.48
English students (STUD-US-ARG)	100,702	3	2.98

If the frequencies of imperative plural verbs in L1 and IL are similar, the next question we need to address is whether the **function** of *let us* in French EFL learner writing also bears similarity to that of imperative plural verbs in French. An analysis of concordance lines for *let us* shows that this sequence is repeatedly used to serve a number of interactive and interpersonal metadiscourse functions in French learner writing (cf. section 2.5). For example,

[189] Only corpora of novice writing are compared here. In section 7.4, we add a further dimension to this analysis and examine the frequency of first imperative plural verbs in English and French professional writing so as to determine whether the 'novice writer' factor has a role to play in French students and EFL learners' heavy use of imperative plural forms.

it is used as a code gloss (example 7.13), a transition marker (examples 7.14 and 7.15), an attitude marker (example 7.16) and an engagement marker (example 7.17).

7.13. *To illustrate the truth of this,* ***let us take the example of*** *Britain which was already fighting its corner alone after Mrs Thatcher found herself totally isolated over the decision that Europe would have a single currency.* (ICLE-FR)

7.14. ***Let us then focus on*** *the new Europe as a giant whose parts are striving for unity.* (ICLE-FR)

7.15. ***Let us now turn our attention to*** *the students who want to apply for a job in the private sector.* (ICLE-FR)

7.16. ***Let us be clear that*** *we cannot let countries tear one another to pieces and if we closed our eyes to such an atrocity, our behaviour would be cowardly.* (ICLE-FR)

7.17. ***Let's first consider*** *the children.* (ICLE-FR)

The functions of *let us* in French EFL learner writing can be paralleled with the very frequent use of first person plural imperative verbs to organise discourse and interact with the reader in French student writing:

7.18. ***Prenons l'exemple*** *des sorciers ou des magiciens au Moyen Age.* (CODIF)

7.19. ***Ajoutons*** *qu'une partie plus spécifique de la population est touchée.* (CODIF)

7.20. ***Comparons*** *cela à la visite de la cathédrale d'Amiens.* (CODIF)

7.21. ***Citons comme exemple*** *le jugement difficile des autorités françaises sur les activités du régime de Vichy dans le cadre du procès de Touvier.* (CODIF)

7.22. ***Envisageons*** *tout d'abord la question économique.* (CODIF)

7.23. ***Examinons*** *successivement le problème de l'abolition des frontières d'un point de vue économique, juridique et enfin culturel.* (CODIF)

7.24. ***Imaginons*** *un monde ou règne une pensée unique.* (CODIF)

7.25. *Et* ***notons*** *que ces réalisations nous inspirent des rêves insoupçonnés jusque là, ...*

7.26. ***Considérons*** *un instant le cinéma actuel.* (CODIF)

7.27. ***Interrogeons-nous,*** *dans un second temps, sur la notion d'identité.* (CODIF)

7.28. ***Pensons, par exemple,*** *à l'Espagne, qui, pendant quatre à huit siècles, a appris à côtoyer les peuples arabes.* (CODIF)

These examples suggest that French learners transfer the discourse or organizational function of first person plural imperative verbs in French to equivalent imperative forms in English.

As explained in section 6.2.1.3, the first person plural imperative form *let us* can be found in professional academic writing but it is not frequent (relative frequency of 5.46

occurrences per 100,000 words). In addition, it only appears with a limited set of verbs (cf. Swales et al. 1998; Hyland 2002c). Table 7.22 gives the significant verb co-occurrents of *let us* in BNC-AC-HUM (horizon: 1R-2R) which are illustrated in examples 7.29 to 7.36.

Table 7.22: Significant verb co-occurrents of 'let us' in BNC-AC-HUM

Word	As collocate
consider	20
say	18
suppose	10
return	9
begin	9
look	10
take	10
have	6

7.29. ***Let us consider*** *more closely what was implied in the Greek refusal to look at the Bible.*

7.30. *To ground the many provisions of,* ***let us say,*** *the UN Declaration of 1948 in the mere possibility of their being defended by moral argument is to consign them to a very combative arena indeed, the vagaries of which we have explored in this chapter and are precisely those exploited by Hare in his gloomy quotation.* (BNC-AC-HUM)

7.31. ***Let us suppose,*** *now, that the child has mastered the correct us of like and dislike.*

7.32. ***Let us now return*** *to the topic of "existence predicates"..* (BNC-AC-HUM)

7.33. ***Let us begin with*** *the proposition that our visual experience does somehow involve a judgement.* (BNC-AC-HUM)

7.34. ***Let us look at*** *the second half of the twentieth century.* (BNC-AC-HUM)

7.35. ***Let us take,*** *as an illustration, the boundaries of bureaucracy.* (BNC-AC-HUM)

7.36. ***Let us have a look at*** *the passage to see how this might be.* (BNC-AC-HUM)

There is much more **lexical variety** in the verbs used in first person plural imperative forms in French EFL learner writing. This variety may again be interpreted as L1-related as French writers use a wide range of first person plural imperative verbs to organise discourse and interact with readers[190]. Table 7.23 shows that, for a large proportion of the imperative forms used in ICLE-FR, a direct translation equivalent can be found in French student writing (i.e. CODIF) and/or French editorials (see section 7.3.4 for a description of the corpus).

[190] Siepmann (2005: 119-121) has shown that, compared to English, French shows much greater reliance on the hortative to introduce an example, e.g. *considérons par exemple, citons l'exemple de, reprenons l'exemple de,* etc.

Table 7.23: Verb co-occurrents of 'let us' in ICLE-FR

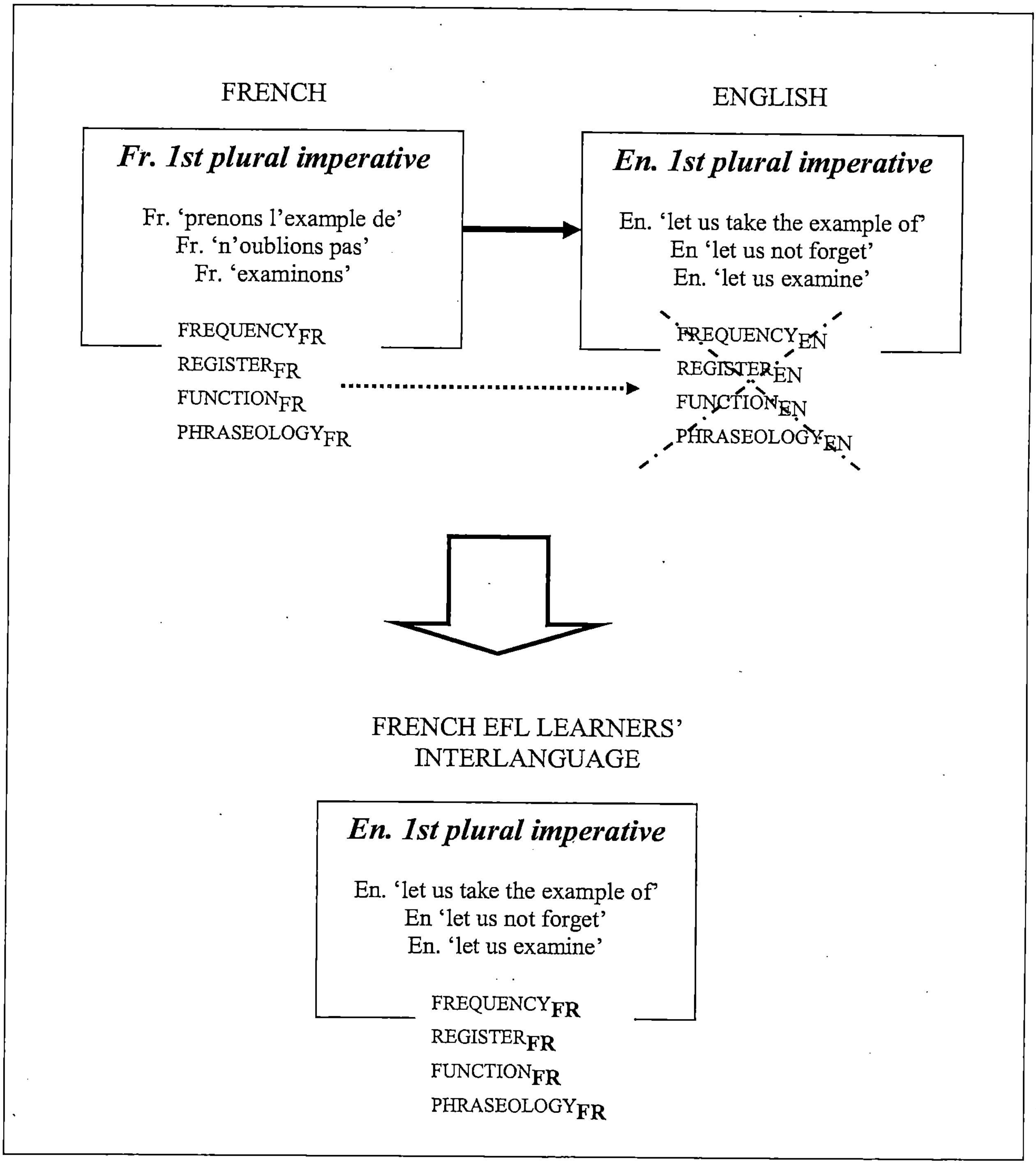
	ICLE-FR	CODIF	TRILLED-FR
let us agree	1	'acceptons'	--
let us analyse	1	'analysons'	--
let us be + ADJ/PP	4	'soyons + ADJ'	'soyons + ADJ'
let us comment	1	--	--
let us consider	4	'considérons'	'considérons'
let us create	1	--	'créons'
let us discuss	1	--	--
let us examine	5	'examinons'	'examinons'
let us explain	2	--	--
let us feel + ADJ	1	--	--
let us focus on	1	--	--
let us not fool ourselves	1	--	'ne soyons pas dupes'
let us (not/never) forget	8	'n'oublions pas'	'oublions'
let us give + example	2	'donnons un exemple'	--
let us give sb. a chance	1	--	--
let us help	1	--	'aidons'
let us have a look at	2	'jetons un regard'	--
let us hope	5	'espérons'	'espérons'
let us imagine	2	'imaginons'	'imaginons'
let us look at	2	'regardons'	'regardons'
let us (not) look down	1	--	--
let us look forward	1	--	--
let us make sb/sth ADJ	1	--	--
let us mention	1	'mentionnons'	--
let us move	1	--	--
let us prepare ourselves	1	--	--
let us prevent sb. from doing sth	1	--	--
let us put ourselves in sb's place	1	--	--
let us put it in another way	1	--	--
let us refer to	1	--	--
let us remember	3	'rappelons'	'rappelons'
let us say	1	'disons'	'disons'
let us see	1	'voyons'	'voyons'
let us solve	1	--	--
let us speak	1	'parlons'	'parlons'
let us start	1	'commençons'	'commençons'
let us state an example	1	--	--
let us take *+ example*	16 8	'prenons' 'prenons l'exemple de'	'prenons' 'prenons l'exemple de'
let us talk	2	'parlons'	'parlons'
let us think	4	'pensons'	--
let us try	2	--	'essayons'
let us (now) turn to	4	--	--
let us view	1	'voyons'	'voyons'
let us wait	1	--	'attendons'
let us wish	2	'espérons'	'espérons'
TOTAL [45 types]	**75 tokens**	**27/45 types (60%) ; 53/75 tokens (70.67%)**	

In addition, imperative forms that are repeated in ICLE-FR often have formal equivalents which are frequently used in French (e.g. *let us take the example of* 'prenons l'exemple de'; *let us hope* 'espérons'; *let us take* 'prenons'; *let us (not) forget* 'oublions/n'oublions pas que'). These similarities further support the hypothesis that French learners' use of *let us* is influenced by their L1. They are also evidence of **transfer of use** and more specifically **transfer of L1 patterns and collocations**.

The influence of the first language on French EFL learners' use of *let us* can also be described as **transfer of register** and **transfer of rhetorical conventions**. First person plural imperative verbs serve specific discourse strategies in French formal types of writing, and more specifically in academic writing. French EFL learners seem to transfer their knowledge of French academic writing conventions (cf. Connor 1996) and make use of imperatives in English academic writing in the same way as in French academic writing. However, *let us* (and more precisely its contracted form *let's*) is much more typical of speech in English (relative frequency of 42.5 occurrences per 100,000 words in BNC-SP vs. 5.3 in BNC-AC). As a result, the speech-like nature of *let us* in French EFL learner writing leads to an overall impression of stylistic inappropriateness.

In summary, the three L1 effects described by Jarvis (2000) are found in French EFL learners' use of *let us*, providing conclusive evidence in favour of L1 influence. French EFL learners use English first person plural imperatives in academic writing with the **frequency** of French imperative verbs in the corresponding **register**, in French-like **phraseological patterns** and to serve the same organizational and interactional **functions.** As illustrated in Figure 7.6, we can interpret our findings not only as evidence of transfer of form and meaning but also as evidence of transfer of frequency, register, function and phraseology. Thus, L1 phrasemes such as *let's take the example of, let's examine* or *let us not forget* mirror the stylistic profile of the French sequences *prenons l'exemple de .., examinons* et *n'oublions pas* in EFL French learner formal writing. This generalized overuse of the first person plural imperative in EFL French learner writing as a rhetorical strategy does not conform to English academic writing conventions but rather to French academic style.

Figure 7.6: Transfer of use



7.3.3.4. French learners' use of 'for instance'

The fourth case study investigates L1 effects on French EFL learners' overuse of *for instance*. This mono-lexemic phraseme appears in almost a quarter of all the essays produced by French learners, which suggests intra-L1 group homogeneity in French learners' IL performance. Table 7.24 shows that the phraseme appears in similar proportions in ICLE-DU and ICLE-SP.

Table 7.24: 'for instance' in learner essays

	Rel. freq. of for instance per 100,000 words	Number of texts including for instance	Number of texts	%
Czech	9.94	12	147	8.16%
Dutch	*36.37*	*42*	*196*	*21.43%*
Finnish	31.93	28	167	16.77%
French	***53.54***	***53***	***228***	***23.25%***
German	12.78	11	179	6.15%
Italian	41.89	14	79	17.72%
Polish	22.06	26	221	11.76%
Russian	25.91	34	194	17.53%
Spanish	*41.36*	*35*	*149*	*23.49%*
Swedish	20.80	10	81	12.35%
TOTAL	**29.51**	**265**	**1641**	**16.15%**

As illustrated in Table 7.25, a Dunnett's test confirms that there is no significant difference between French EFL learners' use of the phraseme *for instance* and that of Dutch and Spanish learners. Neither is there a difference between French and Italian learners. By contrast, learners' use of *for instance* in ICLE-FR differs from ICLE-FI, ICLE-RU, ICLE-SW, ICLE-PO, ICLE-GE and ICLE-PO.

Table 7.25: 'for instance': Dunnett's test

Corpus comparison	Simultaneous lower confidence limit	Difference between means	Simultaneous upper confidence limit	Significance[191]
SP-FR	-0.033561	-0.010080	0.013400	
IT-FR	-0.043621	-0.014522	0.014578	
DU-FR	-0.041994	-0.020282	0.001429	
FI-FR	-0.053387	-0.030685	-0.007983	***
RU-FR	-0.052795	-0.031023	-0.009252	***
SW-FR	-0.064739	-0.035907	-0.007076	***
PO-FR	-0.057243	-0.036202	-0.015162	***
GE-FR	-0.065757	-0.043498	-0.021239	***
CZ-FR	-0.072179	-0.048602	-0.025025	***

[191] Comparisons significant at the 0.05 level are indicated by '***'.

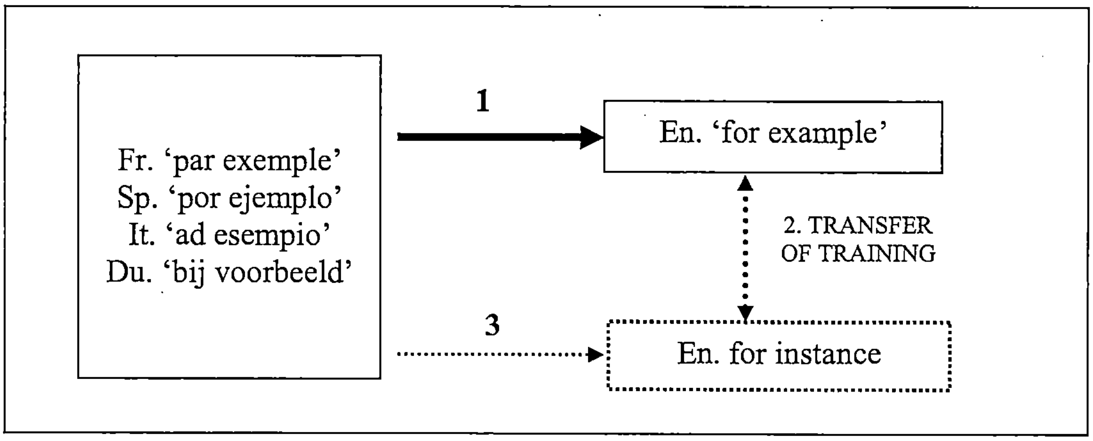
A Ryan test brings out a high degree of heterogeneity between all L1 learner populations and breaks done the learner corpora into several groups. No clear pattern of use arises from these results. L1-inter group heterogeneity is found but the French learner population does not stand apart (compare with Figure 7.2 for *on the contrary* and Figure 7.5 for *let us*).

Figure 7.7: 'for instance': Ryan test

Corpus	Mean	Grouping			
FR	0.057348			A	
SP	0.047268	B		A	
IT	0.042826	B		A	C
DU	0.037066	B	D	A	C
FI	0.026663	B	D		C
RU	0.026325	B	D		C
SW	0.021441	B	D		C
PO	0.021146	B	D		C
GE	0.013850		D		C
CZ	0.008746		D		

The mono-lexemic phraseme *for instance* does not have a direct formal equivalent in French, nor does it have one in Spanish, Italian and Dutch. By contrast, these four languages have a formal equivalent of *for example*. Our results suggest, however, that learners, especially from French, Spanish, Italian or Dutch mother tongue backgrounds, not only establish equivalence between Fr. *par exemple*, Sp. *por ejemplo*, It. *ad esempio* or Du. *bij voorbeeld* and its prototypical equivalent *for example* but also extend the equivalence to *for instance* (cf. Figure 7.8). One possible explanation for this rough equivalence is the influence of instruction or 'transfer of training' (cf. Selinker 1972): although they differ in terms of frequency and register, *for example* and *for instance* are often taught as functionally equivalent forms (see also sections 6.3.1 and 6.4.1). However, this hypothesis does not help explain why learner populations differ widely in their use of *for instance*. Only a careful analysis of ELT materials used in the different countries represented in ICLE could prove or reject the transfer of training hypothesis.

Figure 7.8: 'for instance': transfer of training



In summary, an investigation of Jarvis's (2000) three L1 effects suggests that French EFL learners' overuse of *for instance* is not L1-induced and may be better interpreted as evidence of transfer of training.

7.3.4. A fourth dimension to Jarvis's (2000) unified framework: L1 frequency

What immediately comes to mind when comparing French learners' use of EAP words and phrases with that of EFL learners from nine different mother tongue backgrounds is that French is often not the only language which has lexical items that are congruent with the English word sequences under study. Thus, there is also a direct translation equivalent to *on the contrary* in Spanish: *al contrario* or *por el contrario* (example 7.37). Similarly, the verb *illustreren* exists in Dutch (example 7.38).

7.37. *Ello no significa que sea prudente abrazar, sin más, el librecambismo cultural.* ***Al contrario,*** *mientras haya sectores en los que las industrias culturales europeas sean incapaces de competir, los esfuerzos de las políticas públicas deben ir encaminados a incrementar su competitividad, no su dependencia de las ayudas públicas.* (EDITO-SP)

7.38. *Dit voorbeeld* ***illustreert*** *dat er meer is dan de hoogte van de beloning die maakt of iemand een ministerspost ambieert.* (TRILLED-DU)
[This example illustrates that ..]

First person plural imperative verbs are also used in Spanish to organize discourse and interact with readers (examples 7.39 and 7.40).

7.39. ***Esperemos** que los californianos mantengan abiertas las puertas a quienes llegan hasta aquí en busca de las oportunidades que sus países -- que los gobiernos de sus países -- son incapaces de darles.* (EDITO-SP)
[Let us hope that ...]

7.40. ***Volvamos** al ejemplo de Richard Nixon.* (EDITO-SP)
[Let us return to Richard Nixon's example.]

Unlike French and Spanish, Dutch even has a formal equivalent to En. *let us*:

7.41. *En **laten we** onszelf niets wijs maken. Als we over fascisme of neo-fascisme praten, moeten we niet alleen maar denken aan groepjes fanatieke heethoofden.* (TRILLED-DU)
[And let us not fool ourselves....]

7.42. *We hebben hier weer de handen vol aan aan de Heertjes. De één zorgt voor onrust door wat hij schrijft, de ander door wat hij niet schrijft. **Laten we** met Raoul beginnen, die na de troonrede over Beatrix sprak als een volgevreten pseudo-vedette. "Deze overschatte miljonaire die met het laatste model gouden koets naar haar werk komt".* (TRILLED-DU)
[Let us begin with Raoul, who ...]

As already stated in section 3.3.3.1, Jarvis argues that intra-L1-group congruity between learners' L1 and IL performance is probably the strongest type of evidence for L1 influence. It is therefore quite surprising that French, Italian and Spanish learners differ in their use of *on the contrary*; that French and Spanish learners do not use *let us* similarly and that Dutch learners do not rely on *let us* (cf. section 7.2).

A combination of Contrastive Analysis (CA) and Contrastive Interlanguage Analysis (CIA) as described in Granger's (1996) *Integrated Contrastive Model* (cf. section 4.2.1) is necessary to explain those differences. It is used here to assess the potential influence of the first language on the fact that Spanish learners do not display statistically significant difference in their use of first person plural imperatives when compared to other learner populations while Dutch learners make little use of the sequence. Ideally, learners' use of *let us* would be compared to the use of first person plural imperative verbs in comparable corpora of student essay writting in L1. Unfortunately, we do not have access to such a corpus for Spanish. Nor do we have comparable corpora of Spanish, French and Dutch academic writing. The contrastive analysis presented here is thus based on corpora of editorials.

So (2005) argues that "in general, editorials and school argumentative essays indicate more overlaps than tensions as they share argumentation as their main rhetorical function and generic value" (So 2005:74). In addition, editorials are more easily available than student essays. Most press groups today have an online version of their newspapers on the Internet. Moreover, online newspapers are available in many different languages. We make use of the *TRILingual Louvain corpus of EDitorials* (TRILLED). The corpus is currently being compiled at the Centre for English Corpus Linguistics and is composed of editorials that are manually downloaded from English, French and Dutch newspapers' websites (but see Paquot and Fairon 2006 for the description of a RSS-based technique that can be used to build multilingual corpora from newspaper websites). The corpus of Spanish editorials used was kindly made available by JoAnne Neff van Aertselaer. We refer to this corpus as EDITO-SP. As shown in Table 7.26, it is smaller than the other corpora and mostly includes editorials from one newspaper only. Despite this limitation, it has proved very useful for our exploratory study.

Table 7.26: Comparable corpora of editorials

Corpus	language	newspapers	nb. of words	total
TRILLED-EN	English	The Guardian	421,647	1,011,430
		The Independent	61,479	
		The Times	72,629	
		The Observer	49,259	
		The Sunday Telegraph	50,812	
		The Daily Telegraph	296,340	
		The Economist	59,264	
TRILLED-FR	French	*Le Monde*	212,796	737,174
		Le Figaro	254,906	
		Libération	269,472	
TRILLED-DU	Dutch	*De NRC Handelsblad*	296,175	490,415
		Trouw	95,209	
		Het Parool	39,196	
		Utrechts Nieuwsblad	46,100	
		Haagsche Courant	13,735	
EDITO-SP	Spanish	mostly *El Pais*	-	151,011

As we are now examining the use of first person plural imperative verbs in editorials, we first need to replicate our analysis of imperative forms in TRILLED-FR (cf. section 7.3.3.3) before analyzing them in TRILLED-DU and EDITO-SP. Results are given in Table 7.25. The sequence *laten we* is relatively rare in Dutch editorials: first person imperative verbs do not seem to be widely used as signals of specific discourse strategies in Dutch. By contrast, they are repeatedly used in Spanish. Table 7.27 gives two relative frequencies of first person imperative plural verbs in EDITO-SP because it was found that, in one editorial (<SOp050s>), the writer used imperatives as a leitmotiv throughout his text: the editorial's title is a first person imperative plural verb ('Comparemos' *let compare*) and eight imperatives are used in the body of the text. Relative frequencies were recounted without considering this particular text so as to limit its influence on results.

Table 7.27: First person plural imperatives in French, Spanish, Dutch and English editorials

	Total number of words	**Number of 1st person imperative plural verbs**	**Relative frequency per 100,000 words**
French editorials (TRILLED-FR)	737,174	326	44.36
Spanish editorials (EDITO-SP)	151,011 (150,037)	44 (36)	29.14 (23.99)
Dutch editorials (TRILLED-DU)	490,415	8	1.63
English editorials (TRILLED-EN)	1,011,430	67	6.62

These findings support our hypothesis that transfer of use, and more particularly, **transfer of frequency**, may play a major part in EFL learners' use of specific lexical items. As described in section 7.3.3.3, the sequence *let us* (or *let's*) appears in the Spanish learner corpus but is not as frequent as in French learner writing. Similarly, Spanish first person plural imperative verbs are quite frequent in formal writing (i.e. editorials) though less frequent than their French equivalents. In addition, the infrequent use of *laten we* in Dutch formal writing is also reflected in Dutch learner writing and provides a plausible explanation for the low frequency of use of *let us* in ICLE-DU.

There are still a number of unanswered questions. First, it is not yet quite clear how the fact that first person plural imperative verbs are generally more frequent in learner writing than professional writing should be interpreted. A number of factors may be interacting here, i.e. genre differences, proficiency in L2 and novice writing factors. Neff and Bunce (to appear) show that Spanish EFL learners' preferred discourse strategies may change. They find that Spanish undergraduates as represented in ICLE use more structures introduced by *we* or *we can* (examples 7.43 and 7.44) than graduate students who attempt to make use of more sophisticated structures.

7.43. *As a consequence, **we can say that** objectivity is not very common on T.V.* (ICLE-SP)

7.44. *And this is where our capacity of thinking have to activate because, **if we have a look** at the Spanish channels' audience rates, **we can see that** the ones which have the highest picks in prime-time are the worst programmes in the world.* (ICLE-SP)

To draw a complete picture of L1 influence on discourse conventions in L2, EFL learners' use of *let us* should be examined in parallel with other over- or under-used word sequences that may serve metadiscourse functions such as *we can, we must* and *if we* (cf. section 7.2). For example, the introductory *we must* is overused in ICLE-SP. Neff et al (2004) suggest that Spanish learners' heavy reliance on *we must* as a metadiscourse marker may be attributed to L1-influence as its Spanish translational equivalent *deber* can mean either *must* or *should.* They comment that "*debemos* (we must/should) + reporting verb does not have this face-threatening connotation, but is used rather as a way of adding a further proposition to those already proposed" (Neff et al 2004:154) and further add that "[b]ecause of these differences in meaning (...), Spanish EFL writers might believe that, with the use of *must* + reporting verb, they are merely presenting the reader another proposition to be considered" (ibid. 155). The following sentences illustrate Spanish learners' use of *we must.* Their difficulty in differentiating between *we must* and *we should* is made apparent in example 7.46 in which both types of modality are used.

7.45. *So we must find the middle point,* ***we must remember that*** *we are looking for equality.* (ICLE-SP)

7.46. *As a result, what* ***we must admit*** *is that we all belong to a pyramid called society in which each one must fight to find his/her place within it.* ***We should also accept*** *the hunger to handle power as something inherent to the human race.* (ICLE-SP)

The potential influence of the first language should also be verified on other types of metadiscourse markers such as second person imperatives (examples 7.47 and 7.48) and impersonal structures (examples 7.49 and 7.50), which seem to be favoured by Dutch learners.

7.47. ***Take*** *The Cosby Show for example. It was quite all right in the beginning.* (ICLE-DU)

7.48. ***Take*** *an army that consists entirely of conscripts and an army that consists entirely of professionals and compare the results or their actions.* (ICLE-DU)

7.49. *Concerning the 'pretence of freedom' mentioned above,* ***it can be argued that*** *advertising is a means of making people buy the things that the "economic manipulators' want them to.* (ICLE-DU)

7.50. *Therefore,* ***it can be said that*** *fairy tales have been awfully misjudged in the past centuries and that they are not appropriate for young ears.* (ICLE-DU)

7.3.5. Conclusion

Results of the four case studies presented in this section already make it possible to draw methodological conclusions and state a number of preliminary theoretical implications with regard to transfer. First, because of the effects of individual variation, L1 influence is "most likely to occur in the form of general tendencies and probabilities, and not so much in the form of invariant patterns" (Jarvis 2000:251) (see also Odlin 1989: 251). This is especially true in less controlled data types such as learner corpora. As a result, unlike in experimentally elicited data, it appears to be very difficult to investigate the first transfer effect, i.e. intra-L1-group homogeneity in learners' IL performance, independently of the second effect, i.e. inter-L2-group heterogeneity, when analyzing learner corpus data. We have proposed to establish intra-L1-group homogeneity by comparing the proportion of essays in which a lexical item is found in the learner corpus under study with that of other learner sub-corpora.

Second, learner corpus data clearly dispute Ringbom's claim that "the only tangible signs of cross-linguistic influence are negative ones, errors" (Ringbom 1986:160). L1 transfer has been shown to manifest itself in patterns of overuse and underuse of EAP phrasemes, in collocational and lexico-grammatical preferences and in register differences. They also show that Ellis's (1994:305) conception of overuse due to L1 influence is too restricted. Ellis considers that overuse is often the direct consequence of the avoidance or underproduction of some difficult structure. However, overuse may also result from a preference for lexical items in the L1. These various transfer effects were first touched upon by Levenston (1971) in an article entitled 'Over-indulgence and under-representation – aspects of mother-tongue interference' but have not received the careful attention they deserve in SLA.

Third, results suggest that **transfer of form** not only often goes together with transfer of meaning but also with **transfer of function**. In second language writing research, L1 influence has been shown to manifest itself in idea-generating and idea-organizing activities (e.g. Wang and Wen 2002). It may be hypothesized that EAP multi-word units are most potentially transferable not only because they are relatively semantically and syntactically compositional, i.e. typically unmarked word combinations, but also because they are directly anchored to an organizational or rhetorical function. In addition, findings also point to a third type of transfer, i.e. **transfer of register and discourse conventions.**

Fourth, one of the most important outcomes of the case studies described here is certainly the extent of **transfer of L1 frequency**. Findings have shown that learners who share first languages that have congruent forms of an EAP phraseme do not necessarily display the same patterns of use of that multi-word unit. One plausible explanation lies in differences in L1 frequency, a dimension which seems to play a prominent role in transfer effects but is not addressed in Jarvis's unified framework. Comparisons of the interlanguage, the first language and the foreign language have shown that the frequency of a phraseological pattern in the interlanguage is often closer to the frequency of its L1 congruent form than to its frequency in the L2. The role of frequency is a key issue in second language acquisition. However, it has generally been conceived of in terms of L2 frequency[192]. *Studies in Second Language Acquisition* has devoted a special issue to frequency effects and their implications for all aspects of second language acquisition (volume 24, issue 02). However, no article deals with L1 frequency. They all largely focus on input frequency and its relation with language processing, intake[193], and implicit vs. explicit learning. Similarly, in a state-of-the-art article on SLA theory, Gregg (2003) only addresses the issue of frequency in relation with the role of input, thus restricting his discussion to the question of "how often does input of X need to be provided in order for X to be acquired?" (Gregg 2003:846)

Fifth, the manifestations of L1 influence that have been described here can all be subsumed under the general heading of **transfer of use** (cf. section 3.2.4). In accordance with Hoey's (2005) theory of lexical priming, we can also refer to these different types of transfer effects as **transfer of lexical priming** (cf. section 2.4.2.3 for more detail on lexical priming and section 3.2.4 for the SLA implications of Hoey's theory of lexical priming). EFL learners' knowledge of L1 words and phrasemes includes the fact that these lexical items occur "with certain other words in certain kinds of context" (Hoey 2005:8). Primings for collocational and contextual use of (at least a restricted set of frequent or core) L1 words and phrasemes are particularly strong in the mental lexicon of an adult EFL learner. They are the result of the many encounters with these words and word sequences in L1 speech and writing. Mental primings for the L1 lexicon most probably influence EFL learners' knowledge of English words and phrasemes by priming the lexico-grammatical preferences of an L1 item to its English counterpart.

[192] The major role of L1 frequency has already been identified in a few transfer studies focusing on phonology and syntax (cf. Selinker 1992:211; Kamimoto et al 1992).
[193] De Bot et al (2005) distinguish between input and intake as follows: "'Input' is everything around us we may perceive with our senses, and 'uptake' or 'intake' is what we pay attention to and notice" (De Bot et al 2005:8).

7.4. Pedagogical implications

> "At one time it was considered essential to avoid the mother-tongue in foreign-language teaching, and teachers would go through contortions to explain or demonstrate the meaning of words without translating. What often happened, of course, was that, after the teacher had spent ten minutes miming, say, *curtain* to a class of baffled French students, one of them would break into a relieved smile and say 'Ah, *rideau*'." (Swan 1997:166)

Our data carry at least two important pedagogical implications relating to the role of the first language in the classroom and pedagogical material development respectively. Transfer of lexical priming means that words or phrasemes in the foreign language may be primed for L1 use in terms of collocational and lexico-grammatical preferences, register and frequency. One of the many roles of teaching should thus be to counter these 'default' and sometimes misled primings in EFL learners' mental lexicon. For this purpose, awareness-raising activities focusing on similarities and differences between the mother tongue and the foreign language are clearly needed. These activities should not be restricted to "helping learners focus on errors typically committed by learners from a particular L1" (Hegelheimer and Fisher 2006:259). They should also raise learners' attention to more subtle differences such as register differences and collocational preferences of otherwise similar words in the two languages. This stands in sharp contrast to Bahns's (1993:56) claim that collocations that are direct translation equivalents do not have to be taught. Next to the fact that learners have no way of knowing when collocations are congruent in the mother tongue and the foreign language, differences between L1-L2 collocations may lie with aspects of use rather than form and meaning.

However, as Odlin (1989) explains, it is not always possible to make use of the first language in classroom and to rely on contrastive data:

> "Whatever the merits of contrastive materials in some contexts, it is clear that such materials are not always feasible. For example, when an ESL class consists of speakers of Chinese, Persian, Spanish, Tamil, and Yoruba, there is not likely to be any textbook that contrasts English verb phrases with verb phrases in all of those languages – and even if there were, teachers could not profitably spend the class time necessary to illuminate so many contrasts. Yet even in such classes, one type of contrastive information is frequently available: bilingual dictionaries. Although the comparisons are sometimes restricted to words in the native and target languages, the most carefully prepared dictionaries often provide some comparisons of pronunciation and grammar as well. If the class size allows it, teachers can help individual

> students in using any contrastive information that their dictionaries provide." (Odlin 1989:162)

Bilingual dictionaries should ideally facilitate the teacher's task in multilingual as well as in monolingual classrooms. However, it is questionable whether the type of contrastive information they provide is fully adequate. For example, the *Robert et Collins CD-Rom* (version 1.1)[194] includes an essay writing section in which first person plural imperatives in French are systematically translated by *let us* in English despite the fact that first person plural imperatives are clearly not the most favoured way of organizing discourse and interacting with the reader in English academic writing. The full section is reprinted in Appendix 7.2 but cases of infelicitous translation equivalence between first person plural imperatives in French and English are given in Table 7.28 for the reader's convenience. These findings can be interpreted more globally as being quite representative of a general lack of good contrastive studies on which pedagogical materials can be based. Multilingual corpora clearly have an important role to play here by providing an empirically based source of translation equivalents (cf. Bowker 2003; King 2003).

[194] The CD-Rom includes the contents of the *Robert et Collins Senior* and the *Robert et Collins Super Senior*.

Table 7.28: Collins et Cobuild CD-Rom (2003-2004): Essay writing

Essay writing : function	French sentence	Proposed English equivalence
Developing the argument	*Prenons comme point de départ le rôle que le gouvernement a joué dans l'élaboration de ces programmes*	= 'let us take ... as a starting point'
	En premier lieu, examinons ce qui fait obstacle à la paix	= 'firstly, let us examine'
The other side of the argument	*Après avoir étudié la progression de l'action, considérons maintenant le style*	= 'after studying ... let us now consider'
	Venons-en maintenant à l'analyse des retombées politiques	= 'now let us come to'
Assessing an idea	*Examinons les origines du problème ainsi que certaines des solutions suggérées*	= 'let us examine ... as well as'
	Sans nous appesantir or *nous attarder sur les détails, notons toutefois que le rôle du Conseil de l'ordre a été déterminant*	= 'without dwelling on the details, let us note, however, that'
	Nous reviendrons plus loin sur cette question, mais signalons déjà l'absence totale d'émotion dans ce passage	= 'we shall come back to this question later, but let us point out at this stage'
	Avant d'aborder la question du style, mentionnons brièvement le choix des métaphores	= 'before tackling ... let us mention briefly'
Adding or detailing	*Ajoutons à cela* or *Il faut ajouter à cela* or *À cela s'ajoute un sens remarquable du détail*	= 'let us add to this or added to this'
Introducing an example	*Prenons le cas de Louis dans «le Nœud de vipères»*	= '(let us) take the case of'
Stating facts	*Rappelons les faits. Victoria l'Américaine débarque à Londres en 1970 et réussit rapidement à s'imposer sur la scène musicale*	= 'let's recall the facts'
Emphasizing particular points	*N'oublions pas que, sur Terre, la gravité pilote absolument tous les phénomènes*	= 'let us not forget that'

7.5. Conclusion

The method proposed in this chapter to conduct a transfer study is a combination of a corpus-driven approach to item selection and a combined use of Jarvis's (2000) framework and Granger's (1996) Integrated Contrastive Model to assess L1 influence. Its reliance on programming and statistical analysis may make it sound complex, and perhaps also unnecessarily complicated. For example, Granger (2004) did not need such a sophisticated development to reach the conclusion that French learners' heavy reliance on the phraseme *on the contrary* is probably the result of a complex interplay between L1-influence and transfer of training.

In addition, the corpus-based methods used have their own limitations. The power of a corpus-driven approach to uncover new types of words and phrases that are worthy of investigation would definitely be an asset to transfer studies. However, this benefit can also prove to be a serious drawback. As De Cock (2003) puts it, "[a] corpus-driven linguist may end up with a large number of patterns and frequency counts and may be unable to see the wood from the trees and to interpret the results for lack of a guiding research question (De Cock 2003:197). As for transfer assessment, a major limitation of the corpus-based approach described in this chapter is that it is very difficult to collect the many types of corpora that are necessary, e.g. corpora of academic writing in French, Spanish, Dutch; corpora of student writing in different languages, etc. International collaboration of the type witnessed in the ICLE project is clearly needed to build comparable corpora of texts written in many different languages.

However, this (learner) corpus-based approach to L1 influence has brought to light interesting findings relating to L1 influence. First, a comparison of EAP key clusters across learner corpora has shown **indirect manifestations of preferred rhetorical strategies** in the first language. Second, it has helped identify a number of transfer effects that are largely undocumented in the SLA literature. These transfer effects make up what has been referred to as **'transfer of lexical priming'** and include L1 influence on collocational use and lexico-grammatical patterns of words and phrasemes, register preferences, and frequency of use. Learner corpora are probably the best type of samples of learner interlanguage to investigate these transfer effects. In addition, they arguably give a better account of the complexity and versatility of L1 influence. With its focus on frequency, register differences, and phraseology, to name just a few, corpus linguistics clearly has numerous resources and specific tools to offer to SLA researchers if they want to further investigate manifestations of L1 influence on learner interlanguage.

Our results also support Kellerman's claim that the "hoary old chestnut" according to which transfer does not afflict the more advanced learner "should finally be squashed underfoot as an unwarranted overgeneralization based on very limited evidence" (Kellerman 1984:121). They suggest, however, that transfer effects are more subtle at higher levels of proficiency.

The (learner) corpus-based method proposed in this chapter still has a lot to offer and I really hope that other researchers will take up the challenge of making use of Jarvis's framework and learner corpora to investigate L1 influence on other types of interlanguage features, e.g. learners' use of high-frequency verbs, modals, discourse conventions. The approach could also be used to revisit, in the light of learner corpus evidence, a number of studies in which transfer claims are built on shaky foundations or suffer from what Jarvis referred to as a "you-know-it-when-you-see" syndrome.

8. General conclusion

This thesis lies at the intersection of four research areas, namely (1) Phraseology, (2) English for Academic Purposes, (3) Second Language Acquisition and (4) (Learner) Corpus Research. In this section I take stock of the main findings of the present study and bring out its major contributions to each of the four areas in turn. The section concludes with some avenues for future research in the fields of both EAP and learner corpus research.

Contribution to the theory of phraseology

Phraseologists differ in their delimitation of the scope of phraseology and the types of word combinations they include in the field. These two observations have made it necessary to define the phraseological spectrum much more precisely. A selected review of typologies taken from fields as diverse as lexicology and lexicography, English language teaching and psycholinguistics, has made it possible to identify the most prominent **defining criteria** of phrasemes: internal structure, syntactic function, degrees of compositionality, function in the language and degrees of syntactic flexibility and collocability. Differences in categorization reflect the importance that the different fields, and more exactly individual researchers within these fields, attach to these criteria.

The need for an **extended view of phraseology** has also been emphasized. The field should account for the pervasiveness of a wide range of largely compositional and non-salient multi-word units – 'preferred ways of saying things' – which have so far largely gone unnoticed but which **corpus linguistics** techniques have helped to uncover. These word combinations have traditionally been considered as falling outside the limits of phraseology but recent corpus-based research has highlighted their pervasive nature in language.

It is essential to integrate the new insights derived from corpus-based approaches into a theory of phraseology. Acknowledging the existence of other types of word combinations next to idioms, collocations, proverbs, routine formulae, etc., requires adapting 'traditional' typologies to include preferred co-occurrences and sentence stems. Following De Cock (2003), I have adopted a structural and a functional model to categorize phrasemes as there is no one-to-one correspondence between form and function. The originality of the functional model lies in its integration of a third category, namely that of **textual phrasemes**, next to the already well-established categories of referential and communicative phrasemes. This new category has been added to encompass a whole set of word combinations which stand at the phraseology-discourse interface and which are typically used to serve rhetorical or organizational functions.

More importantly perhaps, the benefits of combining two approaches to the study of phraseology have been highlighted: the heuristic value of the frequency-based (and corpus-based) approach and the fine-grained linguistic analysis of traditional phraseological theory. In this thesis, statistically-defined word combinations extracted from corpora have been used as raw material which has subsequently been refined and submitted to a linguistic analysis. However, the fact that the same terms are regularly used in the literature to describe different types of word combinations in the two approaches leads to confusion. For example, the term 'collocation' has been shown to refer to arbitrarily lexically restricted word combinations in the phraseological approach and to recurrent or statistically prominent word combinations in the distributional approach. It has been suggested that the term **'collocation'** should be reserved for linguistically defined word combinations and that **'co-occurrence'** should be used to refer to the probabilistic phenomenon described within the distributional approach to collocation. More generally, I have argued for a clear distinction between terms used to describe (1) a **method of extraction** of word combinations and (2) a **linguistic analysis** of the results.

Contribution to English for Academic Purposes

On a theoretical level, this thesis supports and somehow substantiates the concept of 'English for (General) Academic Purposes' as a macro-genre which subsumes a wide range of text types in academic settings. Large scale corpus-based studies such as those conducted by Biber have shed light on distinctive linguistic features of academic discourse: they have shown that academic texts typically have an informational and non-narrative focus; require highly explicit, text-internal reference and deal with abstract, conceptual or technical subject matter. My own contribution to legitimizing EAP has been to demonstrate – on the basis of corpus data - that there is a **common core of words and phrasemes** irrespective of differences across genres and disciplines.

This common core has made it possible to provide a more precise description of the construct of **academic vocabulary**. Next to discipline-specific vocabulary, there is a wide range of words and phrasemes that serve to refer to activities which are characteristic of academic discourse, and more generally, of scientific knowledge. They also contribute to discourse organization and cohesion, from topic introduction to concluding statements. A large proportion of what has been referred to as academic words in this thesis consists of **core words**, a category which has so far largely been neglected in EAP .courses. The present analyses have thus questioned the fuzzy but well-established frequency-based distinction between high-frequency words and academic words.

Several studies have pointed to the highly conventionalized nature of academic discourse. My own work has demonstrated that EAP phraseology largely consists of 'lexical extensions' of a set of academic words (e.g. *conclusion, issue, claim, argue*). These words acquire their organizational or rhetorical function in specific word combinations that are essentially semantically and syntactically compositional and which therefore belong to phraseology in its wider sense (e.g. *as discussed below, an example of ... is ..., the aim of this study, the next section aims at ..., it has been suggested*).

The present study has also served to dethrone adverbs from their dominant position as default cohesive markers. Adverbs do not have a monopoly on lexical cohesion and discourse organization in academic writing. My research results have provided ample evidence for the prominent discursive role of **nouns, verbs and adjectives** and their phraseological patterns, a role which is hardly ever mentioned in EFL teaching. These part-of-speech categories, however, serve organisational functions as diverse as exemplification, comparing and contrasting, expressing cause and effect, etc and are therefore as worthy of inclusion in EFL tools as adverbs. .

On a pedagogical level, the results of my study demonstrate that teachers should not assume that EAP students know the first 2,000 words of English. Numerous so-called general service words are not fully mastered by L2 learners, even at high-intermediate to advanced proficiency levels. However, these words serve important discourse-organizing functions in academic writing, which suggests that they should still be the target of productive activities. These findings call into question the systematic use of Coxhead's *Academic Word List* as the exclusive vocabulary syllabus in a number of recent productively-oriented vocabulary textbooks. Another fact that stands out is that a clear distinction should be made between vocabulary needs for academic reading and writing.

As a result, I have developed a new rigorous and empirically-based procedure to select lexical items that should be part and parcel of a **productively-oriented academic word list**. The methodology makes use of the criteria of keyness, range and evenness of distribution and provides a good illustration of the usefulness of making use of POS-tagged corpora for applied purposes. One important feature of the methodology adopted here is that it includes the 2,000 most frequent words in English, thus making it possible to appreciate the paramount importance of core English words in academic prose.

The outcome of this procedure is the *Academic Keyword List*. This list, however, should **not** be regarded as an end product. In its current form (cf. appendix 5.5), the list is the raw result of the application of purely quantitative criteria on native corpus data. As such, it does not yet fully deserve to be referred to as a 'productively oriented' academic word list. Each word still needs 'pedagogical validation' and missing words may be further added to the list. The next step will be to describe each word's lexico-grammatical patterning and phraseology in academic prose and to make use of learner corpus data to inform these descriptions. The procedure has already been applied to the study of words that serve a rhetorical or organizational function in academic discourse and has demonstrated that a phraseological approach to the description of these words provides a mine of valuable information for pedagogical tools.

Contribution to Second Language Acquisition

It has further been shown that academic, and more precisely argumentative, essays written by upper-intermediate to advanced EFL learners share a number of linguistic features irrespective of learners' mother tongue backgrounds or language families. This **common core of interlanguage features** that characterize the expression of rhetorical and organizational functions in EFL writing includes a limited lexical repertoire, a lack of register awareness as well as lexico-grammatical and phraseological specificities, the semantic misuse of connectors and labels, the extensive use of chains of connective devices and a marked preference for sentence-initial position of connectors.

Several of these linguistic features, and more specifically, the lack of register awareness, may perhaps be better interpreted as **developmental** rather than learner-specific characteristics as they have also been found in novice native writing. However, other features such as lexico-grammatical errors, the use of non-native-like sequences and the overuse of relatively rare expressions seem to be largely learner-specific.

Next to a common core of learner-specific and developmental characteristics, I have also identified a number of **L1-specific** features that characterize the writing of EFL learners who share the same mother tongue background. Within the scope of this thesis, it has not been possible to measure the extent of transfer on these L1-specific features. Rather, my study has paved the way for such an analysis by testing the suitability of a number of corpus linguistics techniques in conducting transfer studies. A literature review has pointed to a whole range of methodological problems that need to be addressed before undertaking a large scale study of transfer effects on EFL learners' phraseology. My contribution to transfer studies has thus consisted in taking up the challenge of doing this fundamental methodological spadework. The method proposed is a combination of a corpus-driven approach to item selection and a combined use of Jarvis's (2000) unified framework and Granger's (1996) Integrated Contrastive Model to assess L1 influence.

Despite its primary methodological objective, our investigation into L1-specific interlanguage features has already pointed to a number of interesting findings on a theoretical level. It has identified a number of transfer effects that remain largely undocumented in the SLA literature. Lexical transfer has too often been narrowed down to transfer of form/meaning mappings and the third aspect of word knowledge, i.e. use, has rarely been investigated. However, this thesis has shed light on several aspects of word use which seem to play a prominent role in potential L1 influence: collocational use and lexico-grammatical patterning, function in the language, register preferences and frequency. I have proposed to subsume these different manifestations of L1 influence under the general heading of '**transfer of lexical priming**'.

The valuable theoretical insights provided by a learner-corpus based approach to the study of L1 influence bring to the fore the potential contribution of learner corpora for SLA studies. Learner corpora are probably the best – if not the sole - type of samples of learner interlanguage to investigate transfer of lexical priming. There are however many other variables that interact in learners' interlanguage and these variables are also in need of careful operationalization. Learner corpora could clearly act as a testbed for more studies that aim to provide empirical evidence for second language acquisition theories.

Contribution to (Learner) Corpus Research

The present thesis has relied almost entirely on corpus data to fulfil its theoretical, descriptive and pedagogical objectives. It has made use of a combination of corpus-based and corpus-driven approaches to select EAP vocabulary, analyze its phraseology in native and learner writing, identify L1-specific interlanguage features and investigate the potential influence of the first language on these features.

One important methodological aspect of this study is that it has sought to exploit the full potential of Granger's (1996) **Contrastive Interlanguage Analysis** by comparing (1) learner corpora to corpora of professional academic writing and novice student writing, (2) different learner corpora of the same target language with each other and (3) learner corpora to corpora of L1 writing. The present study has also demonstrated the usefulness of making use of several corpus types and of relying on a large corpus such as the British National Corpus as a control corpus against which results from smaller corpora can be checked.

Except in a few studies (e.g. De Cock 2003), learner corpus data have generally been used to confirm hypotheses and have hardly ever been allowed to speak for themselves. Learner corpus research has also made more extensive use of comparisons of learner vs. native corpora than of learner corpora. I have adopted a **corpus-driven approach** to compare L1 learner corpora with a view to identifying word sequences that are prominent in one L1 learner population but are less frequently used in other learner sub-corpora. While corpus-driven approaches yield an unmanageable amount of data, the procedure I have used restricts the output to what I refer to as EAP key clusters. The method has only been partially exploited here but it has already helped identify a number of indirect manifestations of preferred rhetorical strategies in the first language. In addition, it has stressed the potential influence of topic on learners' phraseology, a variable whose influence has largely been downplayed in corpus studies and more particularly in corpus studies which focus on phraseology.

This thesis has also taken on board certain limitations of current learner corpus research. First, the learner corpora used are more homogeneous than in many learner corpus-based studies. The learner essays that make up each L1 learner corpus were selected so as to **control for several learner and task variables** that have repeatedly been shown to influence interlanguage production. Two influential variables that have rarely been considered in the literature but which are controlled for in the present study are the time allotted to learners to write an essay and whether or not they are given access to reference tools. Second, I have complemented the analysis of pooled data sets with an investigation of individual variability. For this purpose, I have addressed yet another limitation of learner corpus research, i.e. its over-reliance on a single type of statistical measures (chi-squares and log-likelihood tests). I have made use of statistical techniques based on **comparisons of means** which rely on frequency information for each individual text and thus take variability between learner essays into consideration. I can only hope that these aspects of my thesis will help, if only modestly, to convince SLA specialists of the many advantages that learner corpora have to offer to interlanguage studies. Learner corpora are not the exclusive preserve of learner corpus researchers. They should feature prominently in the battery of data types used by SLA specialists.

This thesis has also provided a vivid illustration of the interdisciplinary nature of learner corpus research:

- It has resorted to corpus linguistics techniques and computer programming to process corpora;
- It has made use of several statistical tests, including measures that are not well established in corpus linguistic research, to compare corpora;
- It has interpreted learner interlanguage features in the light of a number of SLA concepts and theories;
- It has discussed implications of the findings for foreign language teaching.

This does not mean however that all learner corpus researchers should acquire programming and statistical skills. I profoundly disagree with Geoffrey Sampson's comment that:

> Nowadays, the ability to write computer code is one of the necessary skills in large areas of linguistics, and people who want to engage in the subject but lack this skill ought to realize that it is up to them to acquire it. (The fact that someone may have studied predominantly humanities subjects does not give him a kind of divine right never to get his hands dirty hacking a bit of code to solve a problem he encounters, like Ancien Regime aristocrats who by birth were entitled not to pay taxes!) Nobody is obliged to devote himself to research on a given subject, but someone who does must take the rough with the smooth. (Geoffrey Sampson, Wed, 29 Jul 1998, Corpora List < http://torvald.aksis.uib.no/corpora/1998-3/0037.html>)

I would rather argue that learner corpus researchers first need to be **good linguists**. Obviously, they must be good **corpus** linguists so as be able to make use of corpus-handling tools to analyze (learner) corpora. Learner corpus researchers also need to have an additional string to their bow in the form of background **SLA knowledge** which is arguably a must. In addition, SLA knowledge needs to be supplemented with **EFL experience** when learner corpora are used for applied purposes.

It would be unreasonable to expect all learner corpus researchers to become experienced statisticians and computer scientists as well. By trying their hand at more disciplines, they run the risk of becoming 'jacks-of-all-trades' in the most pejorative sense of the term. Learner corpus researchers should not be overly concerned about their inability to program or select the most appropriate statistical test. However, they should seek expert advice and work in close collaboration with statisticians and computer scientists. This being said, I would like to add a more personal note to the debate and say that I have always found it extremely empowering to have programming skills and a basic knowledge of statistics, if only to know what can be done with these tools and techniques or to interact with computer scientists and statisticians. It is also extremely gratifying to be able to write basic programs to meet one's needs and help colleagues with these issues.

Avenues for future research

Data from the Academic Keyword List have already been used to inform original academic writing sections in the second edition of the Macmillan English Dictionary. Even though these writing sections can be described as a proactive step to help learners write efficient academic texts, monolingual learner dictionaries still need major improvements to earn the title of 'comprehensive writing tools'. The writing sections largely function as self-contained materials in the middle of the dictionary and are not ideally integrated into the dictionary. The next step is arguably to incorporate data from the writing sections into each word's individual entry. Other possible applications for a fully-fledged productively oriented academic wordlist include helping design a corpus-based EAP textbook for EFL learners or providing source data for an electronic EAP dictionary that would include a writing-aid component. These are research avenues I wish to follow in the near future.

The **electronic medium** is probably the best suited for at least two reasons. First, the wide understanding of phraseology that I have adopted is hardly compatible with a word-based paper dictionary. Size and access to information would clearly be problematic in such a dictionary while they are no longer an issue for electronic tools. Second, the electronic medium also makes it possible to incorporate L1-specific information and contrastive data. It is clear to me that the next generation of electronic dictionaries for EFL learners will consist of adaptive tools which will take into account EFL learners' profile information, notably their mother tongue background as well as their specific needs and requirements.

The last few years have witnessed a burgeoning of studies which focus on genre-specific vocabulary. However, these studies have tended to be largely word-based. Until very recently, the study of genre-specific phrasemes was deeply rooted in terminology research, which explains why compounds have been the object of much attention. Following studies such as Gledhill (2000), I have adopted a **genre-based approach to phraseology**. More corpus-based studies of this type are clearly needed to describe phrasemes. Comparisons of phrasemes in speech vs. writing, academic prose vs. non-academic prose, or news vs. fiction are also called for to help pinpoint the register and genre specificities of phrasemes.

New avenues of research can now be explored by SLA specialists, corpus learner researchers and teachers alike. Not only have a number of largely unrecognized transfer effects been brought to light but the potential influence of **L1 frequency** on learner interlanguage has also been highlighted. This area of research is particularly stimulating and can be expected to be the object of much attention in the next few years.

Another promising area of research which has only been touched upon here is the systematic comparison of learner writing and novice native writing which is essential to distinguish between **learner-specific and developmental features**. Hoey (2005) insists that primings are constrained by register and genre. He gives the example of the word research which is **primed in the mind of academic language users** to occur with recent in academic discourse and news reports on research. The collocation is not primed to occur in other text types or other contexts. A direct implication of Hoey's theory of lexical priming is that academic phraseology cannot be assumed to be primed in the mental lexicon of novice native writers who have had little contact with academic disciplines. Further research is clearly needed to shed some light on the similarities and differences between EFL learners' use of EAP phraseology and that of novice native writers.

All in all, I have shown that the research paradigm of **corpus linguistics** is ideally suited to studying the lexical specificities of academic discourse in native and learner writing. The many corpora available already make it possible to examine a wide range of genres and text types. However, much more could be achieved in the field if other types of corpora were collected, notably longitudinal corpora which are sorely lacking. L1 writing skills would also need to figure more prominently in future research. It does not really make sense to expect learners to write properly in English and produce coherent and cohesive texts if they are not able to perform such a complex task in their mother tongue in the first place. Learner corpus research would greatly benefit from the design of comparable corpora of L1 and L2 writing produced by the same learners. There is also an urgent need for learner corpora which represent academic text types other than argumentative essays. Projects such as the British Academic Written Corpus and the Michigan Corpus of Upper-level Student Papers are thus particularly welcome. If we want to systematize the use of empirical data to perform IL-L1 comparisons in transfer studies, comparable L1 corpora are also highly desirable.

My journey into EAP vocabulary has led me to explore a large number of fascinating fields of research and experiment with a wide range of innovative tools and methods. Navigating my way through the complexity of each of these research areas, I have sought to unify several aspects of English for Academic Purposes, Phraseology, Learner Corpus Research, and Second Language Acquisition into a coherent whole. The challenges presented by such a cross-disciplinary position have quickly been brushed aside by the fresh light the approach has shed on key issues such as the nature of academic vocabulary, the scope of the phraseological spectrum, the respective influence of developmental features vs. transfer effects and the methodological aspects of interlanguage studies. I hope that my thesis will serve as a prompt for further research into the many issues raised. There is still so much to explore.

References[195]

Aarts J. (2002) Does corpus linguistic exist? Some old and new issues. In Breivik L. and A. Hasselgren (eds) *From the COLT's mouth ... and others*. Amsterdam: Rodopi, 1-17.

Aarts J. and S. Granger (1998) Tag sequences in learner corpora: a key to interlanguage grammar and discourse. In Granger S. (ed.), 132-141.

Abdullah K. and H. Jackson (1998) Idioms and the language learner: contrasting English and Syrian Arabic. *Languages in Contrast* 1:83-107.

Ädel A. (2005) Involvement and detachment in writing: The effects of task setting and intertextuality. Paper presented at *ICAME 26 (International Computer Archive of Modern and Medieval English) - AAACL 6 (American Association of Applied Corpus Linguistics)*, University of Michigan, 12-15 May 2005.

Ädel A. (2006) *The use of metadiscourse in argumentative texts by advanced learners and native speakers of English*. Amsterdam: Benjamins.

Ädel A. (to appear) Involvement features in writing: do time and interaction trump register awareness. In Diez-Bedmar B.M., G. Gilquin and S. Papp (eds) *Linking up contrastive and learner corpus research*. Rodopi.

Agerström J. (2000) Hedges in argumentative writing: a comparison of native and non-native speakers of English. In Virtanen T. and J. Agerström (eds), 5-42.

Agustín Llach M.P., A. Fernández Fontecha and S. Moreno Espinosa (2006) Differences in the written production of young Spanish and German learners: evidence from lexical errors in a composition. *Barcelona English Language and Literature Studies 14*. Available from <http://www.publicacions.ub.es/revistes/bells14/PDF/sec_lan_02.pdf>

Aijmer K. (1996) *Conversational Routines in English: Convention and Creativity*. London and New York: Longman.

Aijmer K. (2001) *I think* as a marker of discourse style in argumentative Swedish student writing. In Aijmer K. (ed.) *A Wealth of English. Studies in Honour of Göran Kjellmer*. Göteborg: Acta Universitatis Gothoburgensis, 247-257.

Aijmer K. (2002) Modality in advanced Swedish learners' written interlanguage. In Granger et al. (2002), 55-76.

Aisenstadt E. (1979) Collocability restrictions in dictionaries. In Hartmann R.R.K. (ed.) *Dictionaries and their users*. Leuven: Katholieke Universiteit, 71-74.

Alexander R.J. (1984) Fixed expressions in English: reference books and the teacher. *ELT Journal* 38(2):127-134.

Allerton D.J. (1984) Three (or four) levels of word cooccurrence restriction. *Lingua* 63: 17-40.

Allerton D.J. (1994) Valency and valency grammar. In Asher R.E. (ed.) *The Encyclopaedia of Language and Linguistics* 9. Oxford: Pergamon, 4878-4887.

Allerton D.J., N. Nesselhauf and P. Skandera (eds) (2004) *Phraseological units: basic concepts and their application*. Basel: Schwabe.

Allerton D.J. and J. Wieser (2005) Cross-language homonymy and polysemy: a semantic view of "false friends". In Allerton D.J., C. Tschichold and J. Wieser (eds) *Linguistics, Language Learning and Language Teaching* (ICSELL 10). Basel: Schwabe, 57-83.

[195] All cited internet sources were correct as of January 15th 2007.

Altenberg B. (1984) Causal linking in spoken and written English. *Studia Linguistica* 38:20-69.

Altenberg B. (1998) On the phraseology of spoken English: the evidence of recurrent word-Combinations. In Cowie A.P. (ed.), 101-122.

Altenberg B. (2002) Using bilingual corpus evidence in learner corpus research. In Granger et al. (eds), 37-53.

Altenberg B. and S. Granger-(2001) The grammatical and lexical patterning of *make* in native and non-native student writing. *Applied Linguistics* 22(2): 173-194.

Altenberg B. and M. Tapper (1998) The use of adverbial connectors in advanced Swedish learners' written English. In Granger S. (ed.), 80-93.

Andersen R. (1983) Transfer to somewhere. In Gass S. and L. Selinker (eds) *Language Transfer in Language Learning*. Rowley, MA: Newbury House, 177-201.

Anderson W.J. (2006) *The phraseology of administrative French: a corpus-based study*. Amsterdam and New York: Rodopi.

Arabski J. (ed.) (2006) *Cross-linguistic influences in the second language lexicon*. Clevedon: Multilingual Matters

Archer D., A. Wilson and P. Rayson (2002) *Introduction to the USAS category system*. Available from <http://www.comp.lancs.ac.uk/computing/research/ucrel/usas/usas%20guide.pdf.>

Ard J. and T. Homburg (1992) Verification of language transfer. In Gass S. and L. Selinker (eds), 47-70.

Aston G. and L. Burnard (1998) *The BNC Handbook*. Edinburgh: Edinburgh University Press.

Baayen R. H., L.F. Feldman and R. Schreuder (2006) Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language* 53: 496-512. Available from< http://www.mpi.nl/world/persons/private/baayen/publications.html>

Bailey S. (2006). *Academic Writing: A Handbook for International Students* (2nd edition). London and New York: Routledge.

Bahns J. (1993) Lexical collocations: A contrastive view. *ELT Journal* 47(1): 56-63.

Bahns J. and M. Eldaw (1993) Should we teach EFL students collocations? *System* 21(1):101-114.

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* 4:91-105.

Bally C. (1909) *Traité de Stylistique Française*. Paris: Klincksiek.

Banerjee S. and T. Pedersen (2003) The design, implementation and use of the Ngram Statistics Package. Proceedings of the *Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, February 17-21, 2003, Mexico City. Available from < http://www.d.umn.edu/~tpederse/Pubs/cicling2003-2.pdf>

Barkema H. (1993) Idiomaticity in English NPs. In Aarts J., P. de Haan and N. Oostdijk (eds) *English Language Corpora: Design, Analysis and Exploitation. Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Nijmegen*. Amsterdam: Rodopi, 257-278.

Barkema H. (1996a) The effect of inherent and contextual factors on the grammatical flexibility of idioms. In Percy C.E., C.F. Meyer and I. Lancashire (eds) *Synchronic corpus linguistics*. Amsterdam: Rodopi, 69-83.

Barkema H. (1996b) Idiomaticity and terminology: a multi-dimensional descriptive model. *Studia Linguistica* 50(2): 125-160.

Barkema H. (1997) Lexicalised noun phrases: the relation between collocability, compositionality, syntactic structure and flexibility. In Aarts J., I. de Mönnink and H. Wekker (eds) *Studies in English language and teaching in honour of Flor Aarts.* Amsterdam: Rodopi, 23-41.

Barnbrook G. (1996) *Language and Computers*. Edinburgh: Edinburgh University Press.

Baroni M. and S. Evert (to appear) Statistical methods for corpus exploitation. In Lüdeling A. and M. Kytö (eds) *Corpus Linguistics. An International Handbook*, chapter 38. Berlin: Mouton de Gruyter.

Bartning I. (1997) L'apprenant dit avancé et son acquisition d'une langue étrangère: tour d'horizon et esquisse d'une caractérisation de la variété avancée. *AILE* 9: 9-50.

Bartsch S. (2004) *Structural and functional properties of collocations in English.* Tübingen: Gunter Narr Verlag Tübingen.

Batoux D. (2003) Les verbes supports. In La Grammaticalisation. La Terminologie. *Travaux - Cercle linguistique d'Aix-en-Provence* 18 : 83-98.

Bauer L. (1983) *English Word Formation.* Cambridge: Cambridge University Press.

Bauer L. (2001) *Morphological productivity.* Cambridge: Cambridge University Press.

Bauer L. and I.S.P. Nation (1993) Word families. *International Journal of Lexicography* 6(4):253-279.

Becker J. (1975) The phrasal lexicon. Available from <http://acl.ldc.upenn.edu/T/T75/T75-2013.pdf>

Beheydt L. (2005) The development of an academic vocabulary. In Battaner P. and J. DeCesaris (eds) *De lexicografia. Actes del I Symposium Internacional de Lexicografia.* Série actvitats 15. Barcelona: Institut universitari de linguistica applicada, 241-250.

Benson M., E. Benson and R. Ilson (1997) *The BBI Dictionary of English word combinations.* Amsterdam: Benjamins.

Berry-Roghe G.L. (1973) The computation of collocations and their relevance in lexical studies. In Aitken A.J., R. Bailey and N. Hamilton-Smith (eds) *The computer and literary studies*. Edinburgh: Edinburgh University Press. Available from <http://www.chilton-computing.org.uk/acl/applications/cocoa/p010.htm>

Bertram R., R. H. Baayen and R. Schreuder (2000) Effects of family size for complex words. *Journal of Memory and Language* 42(3): 390–405.

Bhatia V. (2002) A generic view of academic discourse. In Flowerdew J. (ed.) *Academic discourse*. Harlow: Longman, 21-39.

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

Biber D. (2003) Variation among university spoken and written registers: a new multi-dimensional analysis. In Leistyna P. and C.F. Meyer (eds) *Corpus Analysis: Language Structure and Language Use.* Amsterdam and New York: Rodopi, 47-70.

Biber D. (2004) Lexical bundles in academic speech and writing. In Lewandowska-Tomaszczyk B. (ed.) *Practical Applications in Language and Computers (PALC 2003).* Frankfurt am Main: Peter Lang, 165-178.

Biber D. (2006) *University Language: A corpus-based study of spoken and written registers.* Amsterdam and Philadelphia: Benjamins.

Biber D. and S. Conrad (1999) Lexical bundles in conversation and academic prose. In Hasselgård H. and S. Oksefjell (eds) *Out of Corpora: Studies in Honour of Stig Johansson.* Amsterdam: Rodopi, 181-190.

Biber D., S. Conrad and V. Cortes (2003) Lexical bundles in speech and writing: an initial taxonomy. In Wilson A., P. Rayson and T. McEnery (eds) *Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech.* Frankfurt: Peter Lang, 71-92.

Biber D., S. Conrad and V. Cortes (2004) *If you look at ....*: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371- 405.

Biber D., S. Conrad, R. Reppen, P. Byrd and M. Helt (2002) Speaking and writing in the university: a multidimensional comparison. *TESOL Quarterly* 36(1): 9-48.

Biber D. and E. Finegan (1994) Intra-textual variation within medical research articles. In Oostdijk N. and P. de Haan (eds) *Corpus-based research into language*, Language and computers 12. Amsterdam: Rodopi, 201-222.

Biber D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *Longman Grammar of Spoken and Written English.* Harlow: Longman.

Biber D. and R. Reppen (1998) Comparing native and learner perspectives on English grammar: A study of complement clauses. In Granger S. (ed.),145-158.

Biskup D. (1992) L1 influence on learners' renderings of English collocations: a Polish/ German empirical study. In Arnaud P. and H. Béjoint (eds) *Vocabulary and Applied Linguistics*. Londres: Macmillan, 85-93.

Blasco Mateo E. (2002) La lexicalización y las colocaciones. *LEA* 14: 35-61.

Bley-Vroman R. (1983) The comparative fallacy in interlanguage studies: the case of systematicity. *Language Learning* 33: 1–17.

Blonski Hardin V. (2001) Transfer and variation in cognitive reading strategies of Latino fourth-grade students in a late-exit bilingual program. *Bilingual Research Journal* 25(4). Available from < http://brj.asu.edu/content/vol25_no4/html/art7.htm>

Bolton K., G. Nelson and J. Hung (2003) A corpus-based study of connectors in student writing. *International Journal of Corpus Linguistics* 7(2):165-182.

Bondi M. (2004) The discourse function of contrastive connectors in academic abstracts. In Aijmer K. and A-B. Stenström (eds) *Discourse patterns in spoken and written corpora.* Amsterdam and Philadelphia: Benjamins, 139-156.

Borin L. and K. Prütz (2004) New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In Aston G., S. Bernardini and D. Stewart (eds) *Corpora and Language Learners*. Amsterdam: Benjamins, 67-87.

Bosque I. (2001) Sobre el concepto de 'colocación' y sus límites. *LEA* 23(1):9-40.

Boström Aronsson M. (2005) Themes in Swedish advanced learners' written English. Unpublished PhD thesis. Göteborg: University of Göteborg.

Bou Franch P. (1998) On pragmatic transfer. *Studies in Language and Linguistics* 0:5-20.

Bouma G. and B. Villada (2002) Corpus-based acquisition of collocational prepositional phrases. *Computational Linguistics in the Netherlands* (CLIN) 2001, Twente University. Available from <http://odur.let.rug.nl/~begona/>

Bourigault D., N. Aussenac-Gilles, J. Charlet (2004) Construction de ressources terminologiques ou ontologiques à partir de textes: un cadre unificateur pour trois études de cas. Revue d'Intelligence Artificielle 18(1):87-110. Available from <http://w3.univ-tlse2.fr/erss/textes/pagespersos/bourigault/RIA-bourigault-aussenac-charlet.doc>

Bowker L. (2003) Corpus-based applications for translator training: exploring the possibilities. In Granger et al (eds), 169-183.

Brill E. (1992) A simple rule-based part of speech tagger. *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing.* Available from <http://citeseer.ist.psu.edu/brill92simple.html>

Bulut T. (2004) Idiom processing in L2: through rose-colored glasses. *The Reading Matrix* 4(2):105-116.

Burger H. (1998) *Phraseologie. Eine Einführung am Beispiel des Deutschen.* Berlin: Erich Schmidt VerčČlag.

Bussmann H. (1996) *Routledge Dictionary of Language and Linguistics.* London and New-York: Routledge.

Butler C.S. (1998) Multi-word lexical phenomena in functional grammar. *Revista Canaria de Estudios Ingleses* 36: 13-36.

Butler C.S. (2003) Multi-word sequences and their relevance for recent models of functional grammar. *Functions of language* 10(2): 179-208.

Butler F.A., A.L. Bailey, R. Stevens, B. Huang and C. Lord (2004) Academic English in Fifth-grade mathematics, science, and social studies textbooks. Centre for the Study of Evaluation, Report 642. Available from <http://www.cse.ucla.edu/reports/r642.pdf>

Burnard L. (2002) Validation manual for written language resources. Available from <http://www.oucs.ox.ac.uk/rts/elra/D1.xml>

Caldwell C. (2002) *Lexical vagueness in Student Writing.* Cambridge University: Unpublished PhD thesis.

Campbell L. (2001) What's wrong with grammaticalization. *Language Sciences* 23:113-161.

Campbell L. and R. Janda (2001) Introduction: conceptions of grammaticalization and their problems. *Language Sciences* 23:93-112

Campion M.E. and W.B. Elley (1971) *An academic vocabulary list.* Wellington: NZCER.

Cantos P. and A. Sánchez (2001) Lexical constellations: what collocates fail to tell. *International Journal of Corpus Linguistics* 6(2):199-228.

Carroll S. (1992) On cognates. *Second Language Research* 8(2):93-119.

Carter R. (1998 [1987]) *Vocabulary: Applied linguistic perspectives* (2nd edition). London: Routledge.

Carter R. and M. McCarthy (1988) Lexis and discourse: vocabulary in use. In Carter R. and M. McCarthy (eds) *Vocabulary and language teaching.* New York: Longman, 201-220.

Carter R. and M. McCarthy (2006) *Cambridge Grammar of English: A Comprehensive Guide. Spoken and Written English Grammar and Usage.* Cambridge: Cambridge University Press.

Casares J. (1992 [1950]) *Introducción a la lexicografía moderna.* Madrid: C.S.I.C.

Celce-Murcia M. and D. Larsen-Freeman (1999) *The Grammar Book: an ESL/EFL Teacher's course* (2nd edition). Boston: Heinle and Heinle.

Cenoz J., B. Hufeisen and U. Jessner (2001) (eds) *Cross-linguistic influence in third language acquisition: psycholinguistic perspectives.* Clevedon: Multilingual Matters.

Chang Y-Y. and J. Swales (1999) Informal elements in English academic writing: threats or opportunities for advanced non-native speakers. In C. Candlin and K. Hyland (eds) *Writing texts, processes and practices.* London: Longman, 145-167.

Channell J. (1994) *Vague Language.* Oxford: Oxford University Press.

Charlebois J. (2004) Pragmatics: the heart and soul of linguistic proficiency. *The Language Teacher* 28(4):3-8.

Charles M. (2003) 'This mystery...': a corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes* 2(4): 313–326.

Charles M. (2006) Phraseological patterns in reporting clauses used in citation: a corpus-based study of theses in two disciplines. *English for Specific Purposes* 25(3):310-331.

Chen C.W. (2006) The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics* 11(1):113-130.

Chi Man-Lai A., K. XWong Pui-Yiu and M. Wong Chau-Ping (1994) Collocational problems amongst learners: a corpus based study. In Flowerdew L. and A. K. Tong (eds) *Entering Text.* Hong Kong: Language Centre, HKUST, 157-165.

Chujo K. and M. Utiyama (2006) Selecting level-specific specialized vocabulary using statistical measures. *System* 34: 255-269.

Chung T. and P. Nation (2003) Technical vocabulary in specialised texts. *Reading in a Foreign Language* 15(2). Available from <http://nflrc.hawai.edu/rfl>

Chung T.M. and P. Nation (2004) Identifying technical vocabulary. *System* 32(2): 251-263.

Church K., W. Gale, P. Hanks and D. Hindle (1991) Using statistics in lexical analysis. In Zernik U. (ed.) *Lexical acquisition: exploiting on-line resources to build a lexicon.* Lawrence Erlbaum, 116-164. Available from <http://www.patrickhanks.com/papers/usingStats.pdf>

Church K. and P. Hanks (1990) Word association norms, mutual information and lexicography. *Computational Linguistics* 16(1): 22-29. Available from <http://acl.ldc.upenn.edu/P/P89/P89-1010.pdf>

Cieślicka A. (2006) On building castles on the sand, or exploring the issue of transfer in the interpretation and production of L2 fixed expressions. In Arabski (ed.), 226-245.

Clas A. (1994) Collocations et langues de spécialité. *Meta* 39(4):576-580.

Clear J. (1993) From Firth Principles: Computational tools for the study of collocation. In Baker M., G. Francis and E. Tognini-Bonelli (eds) *Text and technology: in honour of John Sinclair.* Amsterdam: Benjamins, 271-292.

Cobb T. (2003) Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review* 59(3):393-423.

Cobb T. and M. Horst (2001) Reading academic English: carrying learners across the lexical threshold. In J. Flowerdew and M. Peacock (eds) *Research perspectives on English for Academic Purposes.* London: Cambridge University Press, 315-329.

Cobb T. and M. Horst (2004) Is there room for an academic word list in French? In Laufer B. and P. Bogaards (eds) *Vocabulary in a second language: selection, acquisition, and testing.* Amsterdam and Philadelphia: Benjamins, 15-38.

Cohen A.D., H. Glasman, P.R. Rosenbaum-Cohen, J. Ferrera and J. Fine (1988) Reading English for specialised purposes: discourse analysis and the use of student informants. In Carrell P., J. Devine and D.E. Eskey (eds) *Interactive approaches to second language reading.* Cambridge: Cambridge University Press, 152-167.

Cole P. (1976) The interface of theory and description: notes on Modern Hebrew relativization. *Language* 52(3): 563-583

Coltier D. (1988) Introduction et gestion des exemples dans les textes à thèse. *Pratiques* 58:23-41.

Connor U. (1996) *Contrastive Rhetoric: Cross-Cultural Aspects of Second-Language Writing.* New York: Cambridge University Press.

Connor U. (2002) New directions in contrastive rhetoric. *TESOL Quarterly* 36(4): 493-510.

Conrad S. (1996) Investigating academic texts with corpus-based techniques: an example from biology. *Linguistics and Education* 8: 299-326.

Conrad S. (1999) The importance of corpus-based research for language teachers. *System* 27:1–18.

Conrad S. and D. Biber (2000) Adverbial marking of stance in speech and writing. In Hunston S. and G. Thompson (eds) *Evaluation in text: Authorial stance and the construction of discourse.* Oxford: Oxford University Press, 56-73.

Corder S. P. (1973) The elicitation of interlanguage. In Svartvik J. (ed.) *Errata: Papers in Error Analysis.* Lund: Gleerup, 36-47.

Corpas Pastor G. (1996) *Manual de fraseología española.* Madrid: Gredos.

Corson D. (1997) The learning and use of academic English words. *Language Learning* 47(4): 671-718.

Cortes V. (2002) Lexical bundles in Freshman composition. In Reppen R., S.M. Fitzmaurice and D. Biber (eds) *Using corpora to explore linguistic variation.* Amsterdam and Philadelphia: Benjamins, 131-145.

Cortes V. (2004) Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes* 23 (4): 397-423.

Coseriu E. (1964) Las solidaridades lèxicas. In Coseriu (ed.) (1981 [1977]) *Principios de semántica estructural.* Madrid: Gredos, 143- 161.

Cosme C., C. Gouverneur, F. Meunier and M. Paquot (eds) *Phraseology 2005. The Many Faces of Phraseology: An interdisciplinary conference.* Louvain-la-Neuve: Centre for English Corpus Linguistics.

Costa A. (2005) Lexical access in bilingual production. In Kroll J.F. and A.M.B. de Groot (eds) *Handbook of Bilingualism: Psycholinguistic Approaches.* Oxford: Oxford University Press, 308-325.

Coulmas F. (1979) On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics* 3: 239-266.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Cowan J. R. (1974) Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly* 8(4): 389-400.

Cowie A. P. (1991) Multi-word Units in Newspaper Language. In Granger S (ed.) *Perspectives on the English Lexicon.* Louvain-la-Neuve: Cahiers de l'Institut de Linguistique de Louvain, 101-116.

Cowie A.P. (1992) Multi-word lexical units and communicative language teaching. In Arnaud P. and H. Béjoint (eds) *Vocabulary and Applied Linguistics.* London: MacMillan, 1-12.

Cowie A.P. (1994) Phraseology. In Asher R.E. (ed.) *The Encyclopaedia of Language and Linguistics.* Oxford: Pergamon, 3168-3171.

Cowie A.P. (1997) Phraseology in formal academic prose. In Aarts J., I. de Mönnink and H. Wekker (eds) *Studies in English language and teaching in honour of Flor Aarts.* Amsterdam and Atlanta: Rodopi, 43-56.

Cowie A.P. (ed.) (1998) *Phraseology: theory, analysis and applications.* Oxford: Oxford University Press.

Cowie A.P. (1998a) Introduction. In Cowie A.P. (ed.), 1-20.

Cowie A.P. (1998b) Phraseological dictionaries: some east-west comparisons. In Cowie A.P. (ed.), 209-228.

Cowie A.P. (2001) Exploring native-speaker knowledge of phraseology: informant testing or corpus research? In Burger H., A.H. Buhofer and G. Greciano (eds) *Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zur Methodologie und Kulturspezifik der Phraseologie.* Essen: Schneider Verlag Hogengehren GmbH, 73-81.

Cowie A., R. Mackin and I.R. McCaig (1983) *Oxford Dictionary of Current Idiomatic English.* Oxford: Oxford University Press.

Coxhead A. (2000). A new Academic Word List. *TESOL Quarterly* 34 (2): 213-238.

Coxhead A. and P. Nation (2001) The specialised vocabulary of English for Academic Purposes. In Flowerdew J. and M. Peacock (eds) *Research perspectives on English for Academic Purposes.* Cambridge: Cambridge University Press, 252-267.

Crewe W. (1990) The illogic of logical connectors. *ELT Journal* 44(4):316-325.

Crismore A., R. Markkanen and M. Steffensen (1993) Metadiscourse in persuasive writing: a study of texts written by American and Finnish university students. *Written Communication* 5: 184-202.

Cruse D. A. (1986) *Lexical semantics*. Cambridge: Cambridge University Press.

Cruse D. A. (2000) *Meaning in language: an introduction to semantics and pragmatics.* Oxford: Oxford University Press.

Curado Fuentes A. (2001) Lexical behaviour in academic and technical corpora: implications for ESP development. *Language Learning and Technology* 5(3):106-129.

Cutting J. (2000) Written errors of international students and English native speaker students. In Blue G.M., J. Milton and J. Saville (eds) *Assessing English for Academic Purposes.* Frankfurt am Main: Peter Lang, 97-113.

Dagut M. and B. Laufer (1985) Avoidance of phrasal verbs – A case for contrastive analysis. *Studies in Second Language Acquisition* 7:73-79.

Dahl T. (2004) Textual metadiscourse in research articles: a marker of national culture or of academic discipline. *Journal of Pragmatics* 36:1807-1825.

Daille B. (1994) *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis. Université Paris 7.

Damascelli A.T. (2004) The use of connectors in argumentative essays by Italian EFL learners. In Prat Zagrebelsky M. T. (ed.) *Computer Learner Corpora: Theoretical issues and empirical case studies of Italian advanced EFL learner's interlanguage.* Alessandria: Edizioni dell'Orso, 138-160.

Davies A. (2003) *The native speaker: myth and reality*. Clevedon: Multilingual Matters.

Davis B. and R. Lunsford (2005) Metonymy in Alzeihmer's speech: when the whole is more than the sum of its parts. In Cosme et al (eds), 81-82.

De Bot K. (1992) A Bilingual Production Model: Levelt's 'Speaking' Model Adapted. *Applied Linguistics* 13(1):1-24.

De Bot K. (2004) The multilingual lexicon: modelling selection and control. *The International Journal of Multilingualism* 1(1):17-32.

De Bot K., W. Lowie and M. Verspoor (2005) *Second Language Acquisition: an advanced resource book.* London and New York: Routledge.

De Cock S. (2003) *Recurrent sequences of words in native speaker and advanced learner spoken and written English: a corpus-driven approach.* Unpublished PhD thesis. Louvain-la-Neuve: Université catholique de Louvain.

De Cock S. (2004) Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL), New Series 2*: 225-246.

De Cock S., S. Granger, G. Leech and T. McEnery (1998) An automated approach to the phrasicon of EFL learners. In Granger S. (ed.), 67-79.

Dechert H. and P. Lennon (1989) Collocational blends of advanced language learners: a preliminary analysis. In Olesky W. (ed.) *Contrastive Pragmatics.* Amsterdam: John Benjamins, 131-168.

DeRose S. (1988) Grammatical category disambiguation by statistical optimization. *Computational Linguistics* 14: 31-39.

Douglas D. (2001) Performance consistency in second language acquisition and language testing. *Second Language Research* 17(4): 442-456.

Downing A. and P. Locke (2002) *A University Course in English Grammar.* London: Routledge.

Doyle P. (2005) Replication and corpus linguistics: lexical networks in texts. *The Corpus Linguistics Conference Series 1(1), Corpus Linguistics 2005.* Available from <www.corpus.bham.ac.uk/PCLC>

Dudley-Evans T. and M.J. St John (1998) *Developments in English for Specific Purposes.* Cambridge: Cambridge University Press.

Dulay H.C. and M.K. Burt (1972) Goofing: an indicator of children's second language learning strategies. *Language Learning* 22:235-52.

Dunning T. (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61-74.

Duyck W. (2004) *Lexical and semantic organization in bilinguals.* Unpublished PhD thesis. Gent: Universiteit Gent. Available from <https://archive.ugent.be/handle/1854/7084>

Engels L.K. (1968) The fallacy of word counts. *International Review of Applied Linguistics* 10: 213-231.

Ellis R. (1994) *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Ellis R. and G. Barkhuizen (2005) *Analysing Learner Language.* Oxford: Oxford University Press.

Evans S. and C. Green (2006) Why EAP is necessary: a survey of Hong Kong tertiary students. *Journal of English for Academic Purposes* 6:3-17.

Evert S. (2004) *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Available from <http://www.collocations.de/phd.html>

Evert S. and H. Kermes (2003). Experiments on candidate data for collocation extraction. In *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics,* 83 - 86. Available from <http://purl.org/stefan.evert/PUB/EvertKermes2003a.pdf>

Evert S. and B. Krenn (2003). Computational approaches to collocations. Introductory course at the *European Summer School on Logic, Language, and Information (ESSLLI 2003),* Vienna. Available from <www.collocations.de>

Faerch C., K. Haastrup and R. Phillipson (1984) *Learner Language and Language Learning.* Copenhagen: Gyldendalske Boghandel.

Faerch C. and G. Kasper (1986) Cognitive dimensions of language transfer. In Kellerman E. and M. Sharwood-Smith (eds), 49-65.

Farrell P. (1990) Vocabulary in ESP: a lexical analysis of the English of electronics and a study of semi-technical vocabulary. *CLCS Occasional Paper* 25:1-83.

Fernando C. (1996) *Idioms and idiomaticity*. Oxford: Oxford University Press.

Field Y. and L.M.O. Yip (1992) A comparison of internal conjunctive cohesion in the English essay writing of Cantonese and native speakers of English. *RELC Journal* 23(1):15-28.

Firth A. and J. Wagner (1997) On discourse, communication and (some) fundamental concepts in SLA research. *Modern Language Journal* 81:285-300.

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

Fløttum K., T. Dahl and T. Kinn (2006a) *Academic Voices – across languages and disciplines*. Amsterdam: Benjamins.

Fløttum K., T. Kinn and T. Dahl. (2006b) "We now report on ..." versus "Let us now see how ...". Author roles and interaction with readers in research articles. In Hyland K. and M. Bondi (eds) *Academic discourse across disciplines*. Bern: Peter Lang, 203-224.

Flowerdew J. (1993) Concordancing as a tool in course design. *System* 21(2):231-244.

Flowerdew J. (ed.) (2002) *Academic Discourse*. Harlow: Longman.

Flowerdew J. (2003) Signalling nouns in discourse. *English for Specific Purposes* 22: 329-346.

Flowerdew J. (2006) Signalling nouns in a learner corpus. *International Journal of Corpus Linguistics* 11(3):345-362.

Flowerdew L. (1998) Integrating 'expert' and 'interlanguage' computer corpora findings on causality: discoveries for teachers and students. *English for Specific Purposes* 17(4): 329-345.

Flowerdew L. (2002) Corpus-based analyses in EAP. In Flowerdew J. (ed.) *Academic Discourse*. Harlow: Longman, 95-114.

Flowerdew L. (2003) A combined corpus and systemic-functional analysis of the problem-solution pattern in a student and professional corpus of technical writing. *TESOL Quarterly* 37(3): 489-511.

Flowerdew L. (2004) The problem-solution pattern in apprentice vs. professional technical writing: an application of appraisal theory. In Aston G., S. Bernardini and D. Stewart (eds) *Corpora and language learners*. Amsterdam and Philadelphia: Benjamins, 125-135.

Francis G. (1994) Labelling discourse: an aspect of nominal-group lexical cohesion. In Coulthard M. (ed.) *Advances in written text analysis*. London and New-York: Routledge, 82-101.

Fraser B. (1970) Idioms within a transformational grammar. *Foundations of Language* 6: 22-42.

Freddi M. (2005) Arguing linguistics: corpus investigation of one functional variety of academic discourse. *Journal of English for Academic Purposes* 4: 5-26.

Gabryś-Barker D. (2006) The interaction of languages in the lexical search of multilingual language users. In Arabski (ed), 144-165.

Garside R. (1987) The CLAWS word-tagging system. In Garside R., G. Leech and G. Sampson (eds) *The Computational Analysis of English*. London and New York: Longman.

Garside R., G. Leech and A. McEnery (eds) (1997) *Corpus annotation: linguistic information from computer text corpora*. New York: Addison Wesley Longman

Garside R. and N. Smith (1997) A Hybrid Grammatical Tagger: CLAWS4. In Garside et al (eds), 102-121.

Gass S. and L. Selinker (1992) (eds) *Language Transfer in Language Learning*. Amsterdam and Philadelphia: Benjamins.

Gass S. and L. Selinker (1992) Afterword. In Gass S. and L. Selinker (eds), 233-236.

Gass S. and L. Selinker (2001) *Second Language Acquisition. An Introductory Course.* Mahwah, NJ: Lawrence Erlbaum.

Ghadessy M. (1979) Frequency counts, word lists and materials preparation: a new approach. *English Teaching Forum* 17:24-27.

Gibbs R.W. (1990) Psycholinguistic studies on the conceptual basis of idiomaticity. *Cognitive Linguistics* 1(4): 417-451.

Giegerich H.Z. (2004) Compound or phrase? English noun-plus-noun constructions and the stress criterion. *English Language and Linguistics* 8(1):1-24.

Giegerich H.Z. (2005) Associative adjectives and the lexicon-syntax interface. *Journal of Linguistics* 41:571-591.

Giegerich H.Z. (2006) Attribution in English and the distinction between phrases and compounds. In Rösel P. (ed.) *Englisch in Zeit und Raum - English in Time and Space: Forschungsbericht für Klaus Faiss*. Trier: Wissenschaftlicher Verlag Trier. Available from <http://www.englang.ed.ac.uk/people/heinz.html>

Gilquin G. (2005) Putting prototypicality to the test. Paper presented at *New Directions in Cognitive Linguistics. First UK Cognitive Linguistics Conference*. University of Sussex, Brighton (United Kingdom), 23-25 October 2005.

Gilquin G. (2000/2001) The Integrated Contrastive Model. Spicing up your data. *Languages in Contrast* 3(1): 95-123.

Gilquin G. (to appear) Combining contrastive and interlanguage analysis to apprehend transfer. In Diez-Bedmar B.M., G. Gilquin and S. Papp (eds) *Linking up contrastive and learner corpus research*. Rodopi.

Gilquin G., S. Granger and M. Paquot (2007) Improve your writing skills: writing sections. In Rundell M. (editor in chief) *Macmillan English Dictionary for Advanced Learners* (second edition). Oxford: Macmillan Education.

Gilquin G., S. Granger and M. Paquot (in press) Learner Corpora: the missing link in EAP pedagogy. In Thompson P. (ed.) *Corpus-based EAP pedagogy*. Special issue of the *Journal of English for Academic Purposes*.

Gilquin G. and M. Paquot (2006) Finding one's voice in losing it. Learner academic writing and medium variation. Paper presented at the third *BAAHE (British Association of Anglicists in Higher Education) International Conference*, 7-9 December, Leuven, Belgique.

Gillard P. and A. Gadsby (1998) Using a learners' corpus in compiling ELT dictionaries. In Granger, S. (ed.), 159-171.

Gläser R. (1986) *Phraseologie der englischen Sprache*. Tübingen: Max Niemeyer Verlag.

Gläser R. (1998) The stylistic potential of phraseological units in the light of genre analysis. In Cowie A.P. (ed.), 125-143.

Gledhill C. (2000) *Collocations in Science Writing*. Language in Performance 22. Tuebingen: Gunter Narr Verlag.

Goldberg A.E. (1985) *Constructions: a construction grammar approach to argument structure*. Chicago: Chicago University Press.

González Rey I. (2002) *La Phraséologie du Français*. Toulouse: Presses Universitaires du Mirail.

Goodman A. and E. Payne (1981) A taxonomic approach to the lexis of science. In Selinker L., E. Tarone and V. Hnazeli (eds) *English for academic and technical purposes: studies in honour of Louis Trimble*. Rowley MA: Newbury House, 23-39.

Gouverneur C. (in preparation) *Phraseology in instructed Second Language Acquisition: A corpus of EFL textbooks under scrutiny*. PhD thesis.

Gramley S. and M. Pätzold (1992) *A survey of Modern English*. London: Routledge.

Granger S. (1993) Cognates: an aid or a barrier to successful L2 vocabulary development? *ITL Review of Applied Linguistics* 99/100: 43-56.

Granger S. (1996) From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In Aijmer K., B. Altenberg and M. Johansson (eds) *Languages in contrast: text-based cross-linguistic studies. Lund Studies in English 88.* Lund: Lund University Press, 37-51.

Granger S. (1996b) Romance words in English: from history to pedagogy. In Svartvik J. (ed.) *Words. Proceedings of an International Symposium*. Stockholm: Almqvist and Wiksell International, 105-121.

Granger S. (1997a) Automated retrieval of passives from native and learner corpora: precision and recall. *Journal of English Linguistics* 25(4): 365-374. .

Granger S. (1997b) On identifying the syntactic and discourse features of participle clauses in academic English: native and non-native writers compared. In Aarts J., I. de Mönnink and H. Wekker (eds) *Studies in English Language and Teaching*. Amsterdam and Atlanta: Rodopi, 185-198.

Granger S. (ed.) (1998) *Learner English on Computer*. London and New York: Addison Wesley Longman.

Granger S. (1998a) The computer learner corpus: a versatile new source of data for SLA research. In Granger S. (ed.), 3-18.

Granger S. (1998b) Prefabricated patterns in advanced EFL writing: collocations and formulae. In Cowie A.P. (ed.), 145-160.

Granger S. (2002) A bird's-eye view of learner corpus research. In Granger et al (eds), 3-33.

Granger S. (2003) The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly* 37(3): 538-546.

Granger S. (2003b) The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies. In Granger S., J. Lerot and S. Petch-Tyson (eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam and Atlanta: Rodopi, 17-29.

Granger S. (2004) Computer learner corpus research: current status and future prospects. In Connor U. and T. Upton (eds) *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam and Atlanta: Rodopi, 123-145.

Granger S. (2005) Pushing back the limits of phraseology: how far can we go? In Cosme et al (eds), 165-168.

Granger S. (2006) Lexico-grammatical patterns of EAP verbs: how do learners cope? Paper presented at *Exploring the Lexis-Grammar Interface*, 5-7 october 2006, University of Hanover, Germany.

Granger S. (to appear) The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In Aijmer K. (ed.) *Corpora and Language Teaching*. Benjamins

Granger S., E. Dagneaux and F. Meunier (eds) (2002b) *The International Corpus of Learner English*. CD-ROM and Handbook. Presses universitaires de Louvain: Louvain-la-Neuve.

Granger S., E. Dagneaux, F. Meunier and M. Paquot (to appear 2008) *The International Corpus of Learner English*. Version 2. Handbook & CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger S., J. Hung and S. Petch-Tyson (eds) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Language Learning and Language Teaching 6. Amsterdam and Philadelphia: Benjamins

Granger S., J. Lerot and S. Petch-Tyson (eds) (2003) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam and Atlanta: Rodopi.

Granger S. and F. Meunier (eds) (to appear) *Phraseology: An Interdisciplinary Perspective*. Amsterdam and Philadelphia: Benjamins.

Granger S., F. Meunier and S. Tyson (1994) *New insights into the learner lexicon: a preliminary report from the International Corpus of Learner English*. In Flowerdew L. and K.K. Tong (eds) *Entering Text*. Hong Kong: The Hong Kong University of Science and Technology, 102-113.

Granger S. and G. Monfort (1994) La description de la compétence lexicale en langue étrangère: perspectives méthodologiques. *Acquisition et Interaction en Langue Etrangère (AILE)* 3: 55-75.

Granger S. and M. Paquot (2005) The phraseology of EFL academic writing: Methodological issues and research findings. Paper presented at *ICAME 26 – AAAACL6 (International Computer Archive of Modern and Medieval English - American Association of Applied Corpus Linguistics*), 12-15 May 2005, University of Michigan, USA.

Granger S. and M. Paquot (to appear) Disentangling the phraseological web. In Granger S. and F. Meunier (eds).

Granger S., M. Paquot and P. Rayson (2006) Extraction of multi-word units from EFL and native English corpora. The phraseology of the verb 'make'. In Häcki Buhofer A. and H. Burger (eds) *Phraseology in Motion I: Methoden und Kritik. Akten der Internationalen Tagung zur Phraseologie (Basel, 2004)*. Baltmannsweiler: Schneider Verlag Hohengehren, 57-68.

Granger S. and S. Petch-Tyson (1996) Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15: 19-29.

Granger S. and P. Rayson (1998) Automatic profiling of learner texts. In Granger S. (ed.), 119-131.

Granger S. and H. Swallow (1988) False friends: a kaleidoscope of translation difficulties. *Langage et l'Homme* 23: 108-120.

Granger S. and J. Thewissen (2005) Towards a reconciliation of a Can Do and Can't Do approach to language assessment. Paper presented at the Second Annual Conference of EALTA (European Association of Language Testing and Assessment), Voss (Norvège), 2-5 June 2005.

Grant L. and L. Bauer (2004) Criteria for re-defining idioms: are we barking up the wrong tree? *Applied Linguistics* 25(1): 38-61.

Gréciano G. (1997) Collocations rythmologiques. *Meta* XLII(1):33-44.

Gréciano G. and A. Rothkegel (1997) *Phraseme in Kontext und Kontrast.* Bochum: Brochmeyer.

Greenbaum S. (1974) Some verb-intensifier collocations in American and British English. *American speech* 49(1-2): 79-89.

Grefenstette G., S. Teufel, J. Gaschler and B.M. Schulze (1994) *Designing and evaluating extraction tools for collocations in dictionaries and corpora. DECIDE Deliverable D-2b: Specifications for collocation extraction tools.* Grenoble: RXRC. Available from <http://www.ims.uni-stuttgart.de/ftp/pub/projekte/decide/>

Gregg K.R. (2003) SLA theory: Construction and assessment. In Doughty C. and M.H. Long (eds) *Handbook of Second Language Research.* London: Blackwell, 831-865.

Gries S. (2006) Some proposals towards more rigorous corpus linguistics. *ZAA* 54(2):191-202.

Gries S. (2007) Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2):109-151.

Gries S. (to appear a) Phraseology and linguistic theory: a brief survey. In Granger S. and F. Meunier (eds).

Gries S. (to appear b) Corpus-based methods in analyses of SLA data. In Robinson P. and N. Ellis (eds) *Handbook of Cognitive Linguistics and Second Language Acquisition.* Mahwah, NJ: Lawrence Erlbaum.

Groom N. (2005) Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4(3): 257-277.

Gross G. (1996) *Les expressions figées en Français: noms composés et autres locutions.* Paris: Ophrys.

Grossmann F. and A. Tutin (2003) Quelques pistes pour le traitement des collocations. In Grossmann F. and A. Tutin (eds), 5-22.

Grossmann F. and A. Tutin (eds) (2003) *Les collocations: analyse et traitement.* Travaux et Recherches en Linguistique Appliquée, Lexicologie et Lexicographie 1. Amsterdam: De Wereld.

Guenthner F. and X. Blanco (2004) Multi-lexemic expressions: an overview. In Leclère C., E. Laporte, M. Piot and M. Silberztein (eds) *Lexique, syntaxe et lexique-grammaire / Syntax, lexis and lexicon-grammar.* Amsterdam and New York: Benjamins, 239-252.

Guest M. (1998) Spoken grammar: easing the transitions. *The Language Teacher* 22(06). Available from < http://jalt-publications.org/tlt/files/98/jun/guest.html>

Guillot M-N. (2005) *Il y a des gens qui disent que …* 'there are people who say that…' Beyond grammatical accuracy in FL learners' writing: issues of non-nativeness. *International Review of Applied Linguistics* 43: 109-128.

Hägglund M. (2001) Do Swedish advanced learners use spoken language when they write in English? A quantitative study of some common phrasal verbs in the written language of Swedish advanced learners, native students and professional writers. *Moderna Språk* XCV(1):2-8.

Halliday M.A.K. (1961) Categories of the theory of grammar. *Word* 17.

Halliday M.A.K. (1966) Lexis as a linguistic level. In Bazell C., J.C. Catford and M.A.K. Halliday (eds) *In memory of J.R. Firth.* London: Longman, 148-162.

Halliday M.A.K. (1994) *An introduction to functional grammar.* London: Arnold.

Halliday M. and R. Hasan (1976) *Cohesion in English.* London: Longman.

Hamm A. (2004) What are proverbs? In Allerton et al (eds), 67-86.

Hamp-Lyons L. and B. Heasley (2006) *Study Writing: A Course in Writing Skills for Academic Purposes*. Cambridge: Cambridge University Press.

Harwood N. (2005) 'We do not seem to have a theory. The theory I present here attempts to fill this gap': Inclusive and exclusive pronouns in academic writing. *Applied Linguistics* 26: 343-375.

Hasselgård H. (to appear) Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In Aijmer K. (ed.) *Corpora and Language Teaching*. Amsterdam and Philadelphia: Benjamins.

Hasselgren A. (1994) Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4(2): 237-260.

Hausmann F.J. (1979) Un dictionnaire des collocations est-il possible? *Travaux de Linguistique et de Littérature* 17(1): 187-195.

Hausmann F.J. (1989) Le dictionnaire de collocations. In Hausmann F.J., H.E. Wiegand and L. Zgusta (eds) *Wörterbücher, dictionaries, dictionnaires. Ein internationales Handbuch zur Lexicographie*. Berlin: de Gruyter, 1010-1019.

Håkansson G., M. Pienemann and S. Sayehli (2002) Transfer and typological proximity in the context of second language processing. *Second Language Research* 18(3): 250-273.

Hegelheimer V. and D. Fisher (2006) Grammar, writing, and technology: A sample technology-supported approach to teaching grammar and improving writing for ESL learners. *CALICO Journal* 23(2): 257-279.

Heid U. (2002) Collocations in lexicography. Paper presented at *Colloc02*, workshop on computational approaches to collocations, 23 August 2002. Austria: Vienna. Available from <http://www.ofai.at/~brigitte.krenn/colloc02/workshop_prog.html>

Herbst T. (1996) What are collocations: sandy beaches or false teeth? *English Studies* 4: 379-393.

Herbst T. (1999) English valency structures – a first sketch. *Erfurter electronic studies in English* 6. Available from <http://webdoc.gwdg.de/edoc/ia/eese/artic99/herbst/6_99.html >

Hewings M. and A. Hewings (2002) "It is interesting to note that …": a comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes* 21 (4): 367-383.

Hiltunen T. (2006) Coming-to-know verbs in research articles in three academic disciplines. Proceedings of the 5th International AELFE (Asociación Europea de Lenguas para Fines Específicos) Conference, 246-251. Available from <http://www.unizar.es/aelfe2006/>

Hindmarsh R. (1980) *Cambridge English Lexicon*. Cambridge: Cambridge University Press.

Hinkel E. (2002) *Second language writers' text: linguistic and rhetorical features*. London: Lawrence Erlbaum Associates.

Hinkel E. (2003a) Adverbial markers and tone in L1 and L2 students' writing. *Journal of Pragmatics* 35(7):1049-1068.

Hinkel E. (2003b) Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly* 37: 275-301.

Hinkel E. (2004) *Teaching academic ESL writing: practical techniques in vocabulary and grammar*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hirsh D. and P. Nation (1992) What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language* 8: 689-696.

Hoey M. (1993) A common signal in discourse: how the word *reason* is used in texts. In Sinclair J.M., M. Hoey and G. Fox (eds) *Techniques of description*. London and New York: Routledge, 67-82.

Hoey M. (1994) Signalling in discourse: a functional analysis of a common discourse pattern in written and spoken English. In Coulthard M. (ed.) *Advances in written text analysis*. London: Routledge, 26-45.

Hoey M. (2004) Textual colligation: a special kind of lexical priming. In Aijmer K. and B. Altenberg (eds) *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora*, Göteborg, 22-26 May 2002. Amsterdam and New York: Rodopi, 171-194.

Hoey M. (2005) *Lexical priming: a new theory of words and language*. London and New-York: Routledge.

Hoffmann S. (2004) Are low-frequency complex prepositions grammaticalized? On the limits of corpus data – and the importance of intuition. In Lindquist H. and C. Mair (eds) *Corpus Approaches to Grammaticalization in English*. Amsterdam and Philadelphia: Benjamins, 171-210.

Hoffmann S. and S. Evert (2006) BNCweb (CQP-edition): The marriage of two corpus tools. In S. Braun, K. Kohn and J. Mukherjee (eds) *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, volume 3 of English Corpus Linguistics. Peter Lang, Frankfurt am Main, 177-195. Available from
<http://purl.org/stefan.evert/PUB/HoffmannEvert2006.pdf>

Housen A. (2002) A corpus-based study of the L2-acquisition of the English verb system. In Granger et al. (2002), 77-117.

Howarth P. (1996) *Phraseology in English Academic Writing: Some Implications for language learning and dictionary making*. Tübingen: Max Niemeyer Verlag.

Howarth P. (1998) The phraseology of learners' academic writing. In Cowie A.P. (ed.), 161-186.

Howarth P. (1999) Phraseological standards in EAP. In Bool H. and P. Luford (eds) *Academic standards and expectations: the role of EAP*. Nottingham: Nottingham University Press, 149-158.

Howell D. (1997) *Statistical Methods for Psychology*. Belmont: Wadsworth.

Hu M. and P. Nation (2000) Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language* 13(1):403-431.

Huang L-S (2001) Knowledge of English collocations: an analysis of Taiwanese EFL learners. In Luke C. and B. Rubrecht (eds) *Texas Papers in Foreign Language Education. Selected proceedings from the Texas Foreign Language Education Conference 2001*. ERIC Document ED465288. Available from < http://www.eric.ed.gov/home.html>

Hulstijn J. and E. Marchena (1989) Avoidance: Grammatical or semantic causes. *Studies in Second Language Acquisition* 11: 242-55

Hunston S. (2002) *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston S. and G. Francis (2000) *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia: Benjamins.

Huntley H. (2006) *Essential Academic Vocabulary: Mastering the Complete Academic Word List*. Boston: Houghton Mifflin Company.

Hyland K. (1998) Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics 30:* 437-455.

Hyland K. (1999) Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics* 20(3): 341-67.

Hyland K (2000) *Disciplinary discourses: social interactions in academic writing*. Harlow: Pearson Education Limited.

Hyland K. (2002a) Specificity revisited: how far should we go now? *English for Specific Purposes* 21:385-395.

Hyland (2002b) Activity and evaluation: reporting practices in academic discourse. In Flowerdew J. (ed.), 115-130.

Hyland K. (2002c) Directives: Argument and engagement in academic writing. *Applied Linguistics* 23: 215-239.

Hyland K (2002d) Options of identity in academic writing. *ELT Journal* 56(4):351-358..

Hyland K. (2003) Patterns of Engagement: Dialogic Features and L2 Student Writing. In Ravelli L. and R. Ellis (eds) *Academic Writing in Context: Social-functional Perspectives on Theory and Practice*. London: Continuum, 5-23.

Hyland K. (2005) *Metadiscourse*. London and New York: Continuum.

Hyland K. and L. Hamp-Lyons (2002) EAP: issues and directions. *Journal of English Academic Purposes* 1:1-12.

Hyland K. and J. Milton (1997) Qualifications and certainty in L1 and L2 students' writing. *Journal of Second Language Writing* 6(2): 183-205.

Hyland K. and P. Tse (2006) Metadiscourse in Academic Writing: a reappraisal. *Applied Linguistics* 25(2):156-177.

Ide N. (2005) Preparation and analysis of linguistic corpora. In Schreibman S., R. Siemens and J. Unsworth (eds) *A Companion to Digital Humanities*. Oxford: Blackwell, 289-306.

Irujo S. (1986) Don't put your leg in your mouth: transfer in the acquisition of idioms in a second language. *TESOL Quarterly* 20: 287-304.

Irujo S. (1993) Steering clear: avoidance in the production of idioms. *International Review of Applied Linguistics in Language Teaching* 31 (3): 205-219.

James C. (1980) *Contrastive Analysis*. Harlow: Longman.

James G., R. Davison, A. Cheung and S. Deerwester (1994) *English in Computer Science: a corpus-based lexical analysis*. Hong-Kong: Hong-Kong University of Science and Technology and Longman Asia.

Jansen L.M. (2000) Second language acquisition: from theory to data. *Second Language Research* 16(1): 27-43.

Jarvis S. (2000) Methodological rigor in the study of transfer: identifying L1 influence in the interlanguage lexicon. *Language Learning* 50(2): 245-309.

Jarvis S. (2003) Probing the effects of the L2 on the L1: a case study. In Cook V. (ed.) *Effects of the Second Language on the First*. Clevedon: Multilingual Matters, 81-102.

Jenkins J. (2005) ELF at the gate: the position of English as a Lingua Franca. *Humanising Language Teaching* 7(2). Available from <www.hltmag.co.uk/mar05/idea.htm>

Jenkins J. (2006) Points of view and blind spots: ELF and SLA. *International Journal of Applied Linguistics* 16(2):137-162.

Jiang N. (2002) Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition* 24: 617-637.

Jiang N. (2004a) Semantic transfer and development in adult L2 vocabulary acquisition. In Bogaards P. and B. Laufer B. (eds) *Vocabulary in a Second Language: selection, acquisition and testing*. Amsterdam and Philadelphia: Benjamins, 101-126.

Jiang N. (2004b) Semantic transfer and its implications for vocabulary teaching in a second language. *The Modern Language Journal* 88:416-432.

Jiménez R.T., E. García and D. Pearson (1996) The reading strategies of bilingual Latina/o students who are successful English readers: Opportunities and obstacles. *Reading Research Quarterly* 31: 90-112.

Johns A. (1997) *Text, role and context: developing academic literacies*. Cambridge: Cambridge University Press.

Jones S. and J.M. Sinclair (1974) English lexical collocations. *Cahiers de Lexicologie* 24: 15-61.

Jordan R.R. (1999) *English for Academic Purposes. A guide and resource book for teachers.* Cambridge: Cambridge University Press.

Juilland A. and E. C. Rodriguez (1964) *Frequency Dictionary of Spanish Words*. La Haye: Mouton.

Kachru B. (1985) Standards, codification and sociolinguistic realism: the English language in the outer circle. In Quirk R. and H. Widdowson (eds) *English in the World: teaching and learning the language and literatures*. Cambridge: Cambridge University Press, 11-30.

Källkvist M. (1999) *Form-class and task-type effects in learner English*. Lund: Lund University Press.

Källkvist M. (1998) Lexical infelicity in English: the case of nouns and verbs. In Haastrup K. and A. Viberg (eds) *Perspectives on Lexical Acquisition in a Second Language*. Lund: Lund University Press, 149-174.

Kamimoto T., A. Shimura and E. Kellerman (1992) A second language classic reconsidered – the case of Schachter's avoidance. *Second Language Research* 8(3): 251-277.

Kasper G. (1992) Pragmatic transfer. *Second Language Research* 8, 203-231.

Kasper G. and K.R. Rose (1999) Pragmatics and SLA. *Annual Review of Applied Linguistics* 19:81-104.

Kaszubski P. (2000) *Selected aspects of lexicon, phraseology and style in the writing of Polish advanced learners of English: a contrastive, corpus-based approach.* Unpublished PhD Thesis. Poznań: Adam Mickiewicz University

Kellerman E. (1977) Towards a characterization of the strategy of transfer in second language learning. *Interlanguage Studies Bulletin* 2(1): 58-145.

Kellerman E (1978) Giving learners a break: native language intuitions as a source of predictions about transferability. *Working Papers in Bilingualism* 15: 59-92.

Kellerman E. (1979) Transfer and non-transfer: where are we now? *Studies in Second Language Acquisition* 2: 37-57.

Kellerman E. (1984) The empirical evidence for the influence of the L1 in interlanguage. In Davies A., C. Criper and A. Howatt (eds) *Interlanguage*. Edinburgh: Edinburgh University Press, 98-122.

Kellerman E. (1986) An eye for an eye: cross-linguistic constraints on the development of the L2 lexicon. In Kellerman E. and M. Sharwood-Smith (eds), 35-48.

Kellerman E. (1995) Cross-linguistic influence: Transfer to nowhere? *Annual Review of Applied Linguistics* 15:125-150.

Kellerman E. (2000) Lo que la fruta puede decirnos acerca de la transferencia léxico-semántica : una dimensión no estructural de las percepciones que tiene el aprendiz sobre las relaciones lingüísticas. In Muñoz C. (ed.) *Segundas lenguas: adquisición en el aula.* Barcelona: Ariel, 21-37.

Kellerman E. and M. Sharwood-Smith (1986) (eds) *Cross-Linguistic Influence in Second Language Acquisition.* New York: Pergamon Press.

Kellerman E. and M. Sharwood-Smith (1986) Cross-linguistic influence in second language acquisition: an introduction. In Kellerman E. and M. Sharwood-Smith (eds), 1-9.

Kennedy G. (1998) *An Introduction to Corpus Linguistics.* London and New York: Longman.

Kilgarriff A. (1996) Why chi-square doesn't work, and an improved LOB-Brown comparison. *Proc. ALLC-ACH Conference.* Bergen, Norway: 169-172. Available from <http://www.kilgarriff.co.uk/publications.htm>

Kilgarriff A. (2001) Comparing corpora. *International Journal of Corpus Linguistics* 6(1):1-37.

Kilgarriff A., P. Rychly, P. Smrz and D. Tugwell (2004) The Sketch Engine. In Williams G. and S. Vessier (eds) *Proceedings of the Eleventh EURALEX International Congress.* Lorient: Université de Bretagne-Sud, 105-116.

King P. (1989) The uncommon core: Some discourse features of student writing. *System* 17(1):13-20.

King P. (2003) Parallel concordancing and its applications. In Granger et al. (eds), 157-168.

Kjellmer G. (1987) Aspects of English collocations. In Meis W. (ed.) *Corpus linguistics and beyond. Proceedings of the seventh international conference on English language research on computerised corpora.* Amsterdam: Rodopi, 133-140.

Kleinmann E. (1977) Avoidance behaviour in adult second language acquisition. *Language Learning* 27:93-107.

Kohn K. (1986) The analysis of transfer. In Kellerman E. and M. Sharwood-Smith (eds), 21-34.

Koike K. (2002) Comportamientos semánticos en las colocaciones léxicas. *LEA* 24(1): 5-23.

Kosem I. and R. Krishnamurthy (2007) The missing link between native speaker and learner dictionaries? The need for an EAP dictionary. Paper presented at the 2007 BALEAP Conference '*EAP in a globalising world: English as an academic lingua franca*', Durham University 15-17 April 2007.

Koya T. (2005) *The acquisition of basic collocations by Japanese learners of English.* Unpublished PhD thesis. Waseda, Japan: Waseda University. Available from < http://dspace.wul.waseda.ac.jp/dspace/handle/2065/5285 >

Krishnamurthy R. (ed.) (2004) *English Collocation Studies: The OSTI Report.* London and New York: Continuum.

Kroll B. (1990) What does time buy? ESL student performance on home vs. class compositions. In Kroll B. (ed.) *Second Language Writing.* Cambridge: Cambridge University Press, 140-154.

Kroll J. F. and A. Dijkstra (2002) The bilingual lexicon. In R. Kaplan (ed.) *Handbook of Applied Linguistics.* Oxford: Oxford University Press, 301-321.

Kroll J.F. and N. Tokowicz (2005) Models of bilingual representation and processing. In Kroll J.F. and A.M.B. de Groot (eds) *Handbook of Bilingualism: Psycholinguistic Approaches.* Oxford: Oxford University Press, 531-553.

Lado R. (1957) *Linguistics across Cultures: Applied Linguistics for Language Teachers*. Michigan: University of Michigan Press.

Lake J. (2004) Using 'on the contrary': the conceptual problems for EAP students. *ELT Journal* 58(2):137-144.

Lakoff G. (1987) *Women, fire, and dangerous things*. Chicago: University of Chicago Press.

Lakshmanan U. and L. Selinker (2001) Analysing interlanguage: how do we know what learners know? *Second Language Research* 17: 393-420.

Laruelle P. (2004) *Mieux écrire en anglais*. Paris: Presses Universitaires de France.

Latkowska J. (2006) On the use of translation in studies of language contact. In Arabski (ed.), 210-225.

Laufer B. (1992). How much lexis is necessary for reading comprehension? In Arnaud P.J. and H. Béjoint (eds) *Vocabulary and applied linguistics*. London: MacMillan, 126-132.

Laufer B. (1997) What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words. In Schmitt N. and M. McCarthy (eds) *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, 140-155.

Laufer B. (2000) Avoidance of idioms in a second language: the effect of L1-L2 degree of similarity. *Studia Linguistica* 54(2):186-196.

Laufer B. (2003) The Influence of L2 on L1 Collocational Knowledge and on L1 Lexical Diversity in Free Written Expression. In Cook V. (ed.) *Effects of the Second Language on the First*. Clevedon: Multilingual Matters, 19-31.

Laufer B. and S. Eliasson (1993) What Causes Avoidance in L2 Learning: L1 L2 difference, L1/L2 Similarity, or L2 Complexity? *Studies in Second Language Acquisition* 15(1): 35-48

Lee D. (2002) *Notes to accompany the BNC World Edition (Bibliographical) Index*. Available from <http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/home/BNCWIndexNotes.pdf>

Leech G. (1991) The state of the art in corpus linguistics. In Aijmer K. and B. Altenberg (eds) *English corpus linguistics: studies in honour of Jan Svartvik*. Harlow: Longman, 8-29.

Leech G. (1997a) Introducing corpus annotation. In Garside et al. (eds), 1-18.

Leech G. (1997b) Grammatical tagging. In Garside et al. (eds), 19-33.

Leech G. (1998) *Preface*. In S. Granger (ed).

Leech G., P. Rayson and A. Wilson (2001) *Word frequencies in written and spoken English*. London: Longman.

Leech G. and N. Smith (1999) The use of tagging. In van Halteren H. (ed.) *Syntactic wordclass tagging*. Dordrecht: Kluwer Academic Publishers, 23-36.

Lefer M-A. and J. Thewissen (2007) Orthographic and morphological errors in learner writing. Paper presented at the *28th ICAME Conference*, Startford-upon-Avon, 23-27 May 2007, England.

Lehmann C. (2002) New reflections on grammaticalization and lexicalization. In Wischer I. and G. Diewald (eds) *New reflections on grammaticalization*. Amsterdam and Philadephia: Benjamins, 1-18.

Lehmann H.-M., P. Schneider and S. Hoffmann (2000) BNCweb. In Kirk J. (ed.) *Corpora Galore: Analysis and techniques in describing English*. Amsterdam: Rodopi, 259-266.

Leńko-Szymańska A. (2007) The role of L1 influence and L2 instruction in the choice of rhetorical strategies by EFL learners. In Walinski J., K. Kredens and S. Gozdz-Roszkowski (eds) *Practical Applications in Language and Computers* 2005. Frankfurt & Main: Peter Lang, 357-372.

Lennon P. (1996) Getting 'easy' verbs wrong at the advanced level. *IRAL* 34(1):23-36.

Leńko-Szymańska A. (2002) How to Trace the Growth in Learners' Active Vocabulary. A Corpus-based Study. In Kettemann B. and G. Marko (eds) *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000.* Amsterdam and New York: Rodopi, 217-230.

Lewis M. (1993) *The lexical approach: the state of ELT and a way forward.* Hove: Language Teaching Publications.

Leśniewska J. (2006) Is cross-linguistic influence a factor in advanced EFL learners' use of collocations? In Arabski (ed.), 65-77.

Levenston E.A. (1971) Over-indulgence and under-representation aspects of mother-tongue interference. In Nickel G. (ed.) *Papers in Contrastive Linguistics*. Cambridge: Cambridge University Press, 115-121.

Lewis M. (2002) *Implementing the lexical approach: putting theory into practice.* Heinle: Boston.

Liao Y. and Y.J. Fuyuka (2004) Avoidance of phrasal verbs: the case of Chinese learners of English. *Language Learning* 54(2):193-226.

Lightbown P.M. (1984) The relationship between theory and method in second-language-acquisition research. In Davies A., C. Criper and A. Howatt (eds) *Interlanguage.* Edinburgh: Edinburgh University Press, 241-252.

Lipka L. (1994) Lexicalization and institutionalization. In Asher R.E. (ed.) *The Encyclopaedia of Language and Linguistics* 4. Oxford: Pergamon, 2164-2167.

Liu E. and P. Shaw (2001) Investigating learner vocabulary: a possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *IRAL* 39: 171-194.

Ljung M. (2002) What vocabulary tells us about genre differences: A study of lexis in five newspaper genres. In Breivik L.E. and A. Hasselgren (eds) *From the COLT's mouth ... and others'. Language Corpora Studies. In honour of Anna-Brita Stenström.* Amsterdam and New York: Rodopi, 181-196.

Lorenz G. (1998) Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. In Granger S. (ed.), 53-66.

Lorenz G. (1999a) Adjective intensification - learners versus native speakers. A corpus study of argumentative writing. *Language and Computers: Studies in Practical Linguistics 27.* Amsterdam and Atlanta: Rodopi.

Lorenz G. (1999b) Learning to cohere: causal links in native vs. non-native argumentative writing. In W. Bublitz, U. Lenk and E. Ventola (eds) *Coherence in Spoken and Written Discourse. How to create it and how to describe it.* Amsterdam and Philadelphia: Benjamins, 55-75.

Louw B. (1993) Irony in the text or insincerity in the writer? In Baker M., G. Francis and E. Tognini-Bonelli (eds) *Text and technology: in honour of John Sinclair.* Amsterdam: Benjamins, 157-176

Louw B. (2000) Contextual prosodic theory: bringing semantic prosodies to life. In Heffer C., H. Sauntson and G. Fox (eds) *Words in Context: a tribute to John Sinclair on his retirement.* Birmingham: University of Birmingham.

Luzón Marco M.J. (2000) Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes* 19(1): 63-86.

Lynn R.W. (1973) Preparing word lists: a suggested method. *RELC Journal* 4(1): 25-32.

Mackin R. (1978) On collocations: "words shall be known by the company they keep". In Strevens P. (ed.) *In Honour of A.S. Hornby*. Oxford: Oxford University Press, 149-165.

Maclagan M., B. Davis and G. Tillard (2005) Fixed phrases in the speech of patients with dementia. In Cosme et al. (eds), 247-248.

MacWhinney B. (1992) Transfer and competition in L2 learning. In R.J. Jackson (ed.) *Cognitive processing in bilinguals*. Amsterdam: North Holland, 371-390.

MacWhinney B. (1997) Second language acquisition and the competition model. In de Groot A.M.B. and J.F. Kroll (eds) *Tutorials in bilingualism: psycholinguistic perspectives*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 113-142.

MacWhinney B. (2001) The competition model: the input, the context, and the brain. In Robinson P. (ed.) *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press, 69-90.

MacWhinney B. (2004) A Unified Model of Language Acquisition. In Kroll J. and A. De Groot (eds) *Handbook of bilingualism: Psycholinguistic approaches*. Oxford: Oxford University Press.

Mahmoud A. (2002) Interlingual transfer of idioms by Arab learners of English. The Internet TESL Journal 8(12). Available from < http://iteslj.org/Articles/Mahmoud-Idioms.html>

Mahmoud A. (2005) Collocation errors made by Arab learners of English. Asian EFL Journal. Professional Teaching Articles 5. Available from < http://www.asian-efl-journal.com/pta_August_05_ma.php>

Makkai A. (1972) *Idiom structure in English*. The Hague and Paris: Mouton.

Manning C. and H. Schütze (2000) *Foundations of statistical natural language processing*. Cambridge and Massachusetts: MIT press.

Martin A. (1976) Teaching Academic Vocabulary to Foreign Graduate Students. *TESOL Quarterly* 10(1): 91-97.

Martínez I. A. (2003). Aspects of theme in the method and discussion sections of biology journal articles in English. *Journal of English for Academic Purposes* 2: 103–123.

Mason O. (2000) Parameters of collocation: the word in the centre of gravity. In Kirk J. (ed.) *Corpora Galore: Analyses and techniques in describing English. ICAME 19 proceedings*. Amsterdam and Atlanta: Rodopi, 267-280.

McCarthy M. (1991) *Discourse analysis for language teachers*. Cambridge: Cambridge University Press.

McCarthy M. and R. Carter (1994) *Language as discourse*. Harlow: Longman.

McEnery T. and N.A. Kifle (2002) Epistemic modality in argumentative essays of second-language writers. In Flowerdew J. (ed.), 182-215.

McEnery A., R. Xiao and Y. Tono (2006) *Corpus-based language studies: an advanced resource book*. London and New-York: Routledge.

McIntosh A. (1961) Patterns and range. *Language* 37: 325-337.

Meara P. (1993) The bilingual lexicon and the teaching of vocabulary. In Schreuder R. and B. Weltens (eds) *The Bilingual Lexicon*. Amsterdam: Benjamins, 279-295.

Meara P. (2002) The rediscovery of vocabulary. *Second Language Research* 18(4):393-407.

Mejri S. (2005) Introduction: polysémie et polylexicalité. In Mejri S. (ed.) *Polysémie et Polylexicalité. Syntaxe et Sémantique* 5 : 13-30.

Mel'čuk I. (1995) Phrasemes in language and phraseology in linguistics. In Everaert M., E.J. Van der Linden and A. Schenk (eds) *Idioms: structural and psychological perspectives*. Hillsdale: Lawrence Erlbaum Associates, 167-232.

Mel'čuk I. (1998). Collocations and lexical functions. In Cowie (ed.), 23-53.

Melka F. (1997) Receptive vs. productive aspects of vocabulary. In Schmitt N. and M. McCarthy (eds) *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, 84-102.

Meunier F. (2000) *A computer corpus linguistics approach to interlanguage grammar: noun phrase complexity in advanced learner writing*. Unpublished PhD thesis. Université catholique de Louvain: Louvain-la-Neuve.

Meunier F. (in preparation) *Using corpus linguistics to analyse native/non-native interaction and discourse strategies: the MOO experience.*

Meyer P.G. (1997) *Coming to know: studies in the lexical semantics and pragmatics of academic English*. Tübingen: Gunter Narr Verlag Tübingen.

Michiels A. (2002) Le traitement de la phraséologie dans DEFI. *Linguisticae Antverpiensia* 1:349-364.

Mieko M. (1997) An interview with Ulla Connor. *The Language Teacher Online* 21(4). Available from < http://www.jalt-publications.org/tlt/files/97/apr/connor.html>

Miller J. and R. Weinert (1995) The function of *like* in dialogue. *Journal of Pragmatics* 23:365-93.

Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (ed.), 186-198.

Milton J. (1999) Lexical thickets and electronic gateways: making text accessible by novice writers. In Candlin C.N. and K. Hyland (eds) *Writing: Texts, Processes and Practices*. London and New York: Longman, 221-243.

Milton J. and E. Tsang (1993). A corpus-based study of logical connectors in EFL students' writing: directions for future research. In R. Pemberton and E. Tsang (eds) *Studies in Lexis*. Hong Kong: The Hong Kong University of Science and Technology, 215-246.

Mitchell T.F. (1975) Linguistics 'goings-on': collocations and other lexical matters arising on the syntagmatic record. *Longman Linguistics Library* 19: 99-136.

Mitchell R. and F. Myles (2004) *Second Language Learning Theories* (second edition). London: Arnold.

Modiano M. (1999) International English in the global village. *English Today* 15: 22-28.

Mollin S. (2006) *Euro-English: Assessing Variety Status*. Tübingen: Narr.

Montoro del Arco E.T. (2006) *Teoría fraseológica de las locuciones particulares: las locuciones prepositivas, conjuntivas y marcadoras en español*. Frankfurt am main: Peter Lang.

Moon R. (1998a) *Fixed expressions and idioms in English*. Oxford: Clarendon Press.

Moon R. (1998b) Vocabulary connections: multi-word items in English. In McCarthy M. and N. Schmitt (eds) *Vocabulary: description, acquisition and pedagogy*. Cambridge: Cambridge University Press, 40-63.

Moon R. (1998c) Frequencies and forms of phrasal lexemes in English. In Cowie A.P. (ed.), 79-100.

Moreno A. (2004) Retrospective labelling in premise – conclusion metatext: an English-Spanish contrastive study of research articles on business and economics. *Journal of English for Academic Purposes* 3: 321-339.

Moss G. (1992) Cognate recognition: its importance in the teaching of ESP reading courses to Spanish learnes. *English for Specific Purposes* 11(2):141-158.

Mudraya O. (2006) Engineering English: A lexical frequency instructional model. *English for Specific Purposes* 25(2): 235-256.

Mudraya O., S. Piao, L. Löfberg, P. Rayson and D. Archer (2005) English-Russian-Finnish cross-language comparison of phrasal verb translation equivalents. In Cosme, C., C. Gouverneur, F. Meunier and M. Paquot (eds) *Proceedings of the Phraseology 2005 Conference*, Louvain-la-Neuve, 13-15 October 2005, 277-281.

Mukherjee J. (2005) The native speaker is alive and kicking – linguistic and language-pedagogical perspectives. *Anglistik* 16(2):7-23.

Mukherjee J. and J-M. Rohrback (2006) Rethinking applied corpus linguistics from a language-pedagogical perspective: new departures in learner corpus research. In Kettemann B. and G. Marko (eds) *Planning, gluing and painting corpora: inside the applied corpus linguist's workhop.* Frankfurt am Main: Peter Lang, 205-232. Available from <http://www.uni-giessen.de/anglistik/LING/Staff/mukherjee/>

Mulak J.I. (2000) The discourse functions of *because* clauses in L1 and L2 argumentative writing including a quantitative survey of reason/cause, concession and condition clauses. In Virtanen T. and J. Agerström (eds), 81-139.

Müller S. (2005) *Discourse markers in native and non-native English discourse.* Amsterdam and Philadelphia: Benjamins.

Myles F. (2005) Review Article: Interlanguage corpora and second language acquisition research. *Second Language Research* 21(4):373-391.

Nagy W., R.C. Anderson, M. Schommer, J.A. Scott and A.C. Stallman (1989) Morphological families in the internal lexicon. *Reading Research Quarterly* 24: 262–282.

Narita M. and M. Sugiura (2006) The use of adverbial connectors in argumentative essays by Japanese EFL college students. *English Corpus Studies* 13:23-42.

Nation P. (2001) *Learning Vocabulary in another Language.* Cambridge: Cambridge University Press.

Nation P. and K. Hwang (1995) Where would general service vocabulary stop and special purposes vocabulary begin? *System* 23(1):35-41.

Nation P. and R. Waring (1997) Vocabulary size, text coverage and word lists. In Schmitt N. and P. Nation (eds) *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press, 6-19.

Nattinger J.R. and J.S. Decarrico (1992) *Lexical phrases and language teaching.* Oxford: Oxford University Press.

Neff J., F. Ballesteros, E. Dafouz, F. Martínez and J.P. Rica (2004) The expression of writer stance in native and non-native argumentative texts. In Facchinetti R. and F. Palmer (eds) *English Modality in Perspective.* Frankfurt am Main: Peter Lang, pages.

Neff J., F. Ballesteros, E. Dafouz, F. Martínez and J-P. Rica (2007) A contrastive functional analysis of errors in Spanish EFL university writers' argumentative texts: corpus-based study. In Fitzpatrick E. (ed.) *Corpus Linguistics beyond the Word: Corpus Research from Phrase to Discourse* (Language and Computers 23). Amsterdam: Rodopi, 203-225.

Neff J., E. Dafouz, M. Díez, R. Prieto and C. Chaudron (2004b) Contrastive discourse analysis: argumentative text in English and Spanish. In Moder C. and A. Martinovic-Zic (eds) *Discourse across languages and cultures.* Amsterdam and Philadelphia: Benjamins, 267-283.

Neff J., E. Dafouz, H. Herrera, F. Martínez, J.P. Rica, M. Diez, R. Prieto and C. Sancho (2003) Contrasting learner corpora: the use of modal and reporting verbs in the expression of writer stance. In Granger S. and S. Petch-Tyson (eds) *Extending the scope of corpus-*

based research. New applications, new challenges. Amsterdam and New York: Rodopi, 211-230.

Neff van Aertselaer J. (2006) A rhetorical analysis approach to English for academic purposes. *Revista de Lingüística y Lenguas Aplicadas* 1:63-72.

Neff van Aertselaer J. (to appear) Contrasting English-Spanish interpersonal discourse phrases: a corpus study. In Granger S. and F. Meunier (eds) *Phraseology in Language Learning and Teaching*. Amsterdam: Benjamins.

Neff van Aertselaer J. and C. Bunce (to appear) Pragmatic word order errors and discourse-grammar interdependence. In Gómez González (ed.) *IV Conferencia de Lingüística Contrastiva*. Santiago de Compostela: Universidad de Santiago.

Nelson M. (2000) *A Corpus-Based Study of Business English and Business English Teaching Materials*. Unpublished PhD Thesis. Manchester: University of Manchester.

Nesi H. (2002) An English spoken academic wordlist. In Braash A. and C. Povlsen (eds) *Proceedings of the Tenth EURALEX International Congress*, EURALEX 2002, vol. 1, 351-357.

Nesi H. and H. Basturkmen (2006) Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics* 11(3):147-168.

Nesi H., G. Sharpling and L. Ganobcsik-Williams (2004) Student papers across the curriculum: designing and developing a corpus of British student writing. *Computers and Composition* 21: 439-450.

Nesselhauf (2003a) The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2): 223-242.

Nesselhauf (2003b) Transfer at the locutional level: an investigation of German-speaking and French-speaking learners of English. In Tschichold C. (ed.) *English Core Linguistics. Essays in Honour of D.J. Allerton*. Bern: Lang, 269-286.

Nesselhauf N (2004) What are collocations? In Allerton et al (eds), 1-22.

Nesselhauf (2004b) How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In Aston G., S. Bernardini and D. Stewart (eds) *Corpora and Language Learners*. Amsterdam and Philadelphia: Benjamins, 109-124.

Nesselhauf N. (2005) *Collocations in a learner corpus*. Amsterdam: Benjamins.

North B. (2002) Developing descriptor scales of language proficiency for the CEF common reference levels. In Alderson J.C. (ed.) *Case Studies in applying the Common European Framework*. Strasbourg: Council of Europe, 87-105.

Nunberg G., I. Sag and T. Wasow (1994) Idioms. *Language* 70(3): 491-538.

Oakes M.P. (1998) *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Oakes M. (2003) Text categorization: automatic discrimination between US and UK English using the chi-square and high-ratio pairs. Research in Language 1:143-156. Available from <http://www.cet.sunderland.ac.uk/IR/oakesRL2003.pdf>

Oakey D. (2002) Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In Reppen R., S.M. Fitzmaurice and D. Biber (eds) *Using corpora to explore linguistic variation*. Amsterdam and Philadelphia: Longman, 111–129.

Obenda D. (ed.) (2004) *Academic word power* (1- 4). Boston and New York: Houghton Mifflin Company.

Odlin T. (1989) *Language Transfer: Cross-linguistic influence in language learning*. Cambridge: Cambridge University Press.

Odlin T. (2003) Cross-linguistic influence. In Doughty C.J. & M.H. Long (eds) *The Handbook of Second Language Acquisition.* Oxford: Blackwell, 436-486.

Osborne J. (2007) Why do they keep making the same mistakes? Evidence for error motivation in a learner corpus. In Walinski, J., K. Kredens and S. Gozdz-Roszkowski (eds) *Practical Applications in Language and Computers 2005.* Frankfurt and Main: Peter Lang, 343-355.

Osborne J. (to appear) Phraseology effects as a trigger for errors in L2 English: the case of more advanced learners. In Meunier F. and S. Granger (eds) *Phraseology in Language Learning and Teaching.* Amsterdam and Philadelphia: Benjamins.

Ozturk I. (2007). The textual organization of research article introductions in applied linguistics: variability within a single discipline. *English for Specific Purposes* 26: 25-38.

Paltridge B. (2002) Thesis and dissertation writing: an examination of published advice and actual practice. *English for Specific Purposes* 21: 125-143.

Paquot M. (2007). Towards a productively-oriented academic word list. In J. Walinski, K. Kredens and S. Gozdz-Roszkowski (eds) *Corpora and ICT in Language Studies. PALC 2005. Lodz Studies in Language 13.* Frankfurt am main: Peter Lang, 127-140.

Paquot M. (to appear) Exemplification in learner writing: a cross-linguistic perspective. In Granger S. and F. Meunier (eds) *Phraseology in Language Learning and Teaching.* Amsterdam: Benjamins.

Paquot M. and C. Fairon (2006) Investigating L1-induced learner variability: using the Web as a source of L1 comparable data. Paper presented at *ICAME 27* (International Computer Archive of Modern and Medieval English), 24-28 May 2006, University of Helsinki, Finland.

Partington A. (1998) *Patterns and Meanings. Using Corpora for English Language Research and Teaching.* Amsterdam and Philadelphia: Benjamins.

Partington A. (2004) "Utterly content in each other's company": semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1): 131-156.

Pawley A. and F.H. Syder (1983) Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Richards J.C. and W. Schmidt (eds) *Language and communication.* London and New-York: Longman, 29-59.

Pazos Bretaña J.M. (2005) *Detección automatizada de fraseologismos.* Unpublished PhD thesis. University of Grenada. Available from <http://hera.ugr.es/tesisugr/15476935.pdf>

Peacock M. (2006) A cross-disciplinary comparison of boosting in research articles. *Corpora* 1: 61-84.

Pecman M. (2004) *Phraséologie contrastive anglais-français: analyse et traitement en vue de l'aide à la rédaction scientifique.* Unpublished PhD Thesis, Université de Nice, Sophia Antipolis.

Perdue C. (1993) Comment rendre compte de la "logique" de l'acquisition d'une langue étrangère par l'adulte? *Etudes de Linguistique Appliquée* 92:8-22.

Perrez J. (2006) *Connectieven, tekstbegrip en vreemdetaalvereving. De impact van causale en contrastieve connectieven op het begrijpen van teksten in het Nederlands als vreemde taal.* Unpublished PhD thesis. Louvain-la-Neuve: Université catholique de Louvain.

Petch-Tyson S. (1998) Reader/writer visibility in EFL persuasive writing. In Granger S. (ed.), 107-118.

Petch-Tyson S. (1999) Demonstrative expressions in argumentative discourse - A computer-based comparison of non-native and native English. In Botley S. P. and A. M. McEnery (eds) *Corpus-based and Computational Approaches to Discourse Anaphora.* Amsterdam and Philadelphia: Benjamins, 43-64.

Peters A.M. (1983) *Units of language acquisition.* Cambridge: Cambridge University Press.

Piao S.L., G. Sun, P. Rayson and Q. Yuan (2006) Automatic extraction of Chinese multi-word expressions with a statistical tool. In proceedings of the *Workshop on multi-word expressions in a multilingual context* held in conjunction with the *11th Conference of the European Chapter of the Association for Computational Linguistics* (EACL2006), Trento, Italy, April 3, 2006, 17-24.

Piasecka L. (2006) 'Don't lose your head' or how Polish learners of English cope with L2 idiomatic expressions. In Arabski (ed), 246-258.

Piller I. (2001) Who, if anyone, is a native speaker? *Anglistik* 12(2):109-121.

Poulisse N. (1997) Language production in bilinguals. In De Groot and Kroll (eds) *Tutorials in bilingualism: Psycholinguistic perspectives.* Mahwah, NJ: Lawrence Erlbaum Publishers, 201-224.

Poulisse N. and T. Bongaerts (1994) First language use in second language production. *Applied Linguistics* 15(1):36-57.

Poulsen S. (2005) *Collocations as a language resource: A functional and cognitive study in English phraseology.* University of Southern Denmark: Unpublished PhD thesis. Available from <http://www.humaniora.sdu.dk/phd/dokumenter/filer/Afhandlinger-48.pdf>

Praninskas J. (1972) *American University Word List.* London: Longman.

Pravec N. (2002) Survey of learner corpora. *ICAME Journal* 26: 81-114.

Pullum J.K. and R. Huddleston (2002) Adjectives and adverbs. In Pullum J.K. and R. Huddleston (eds) *The Cambridge Grammar of the English Language.* Cambridge: Cambridge University Press, 525-596.

Quirk R., S. Greenbaum, G. Leech and J. Svartvik (1972) *A Grammar of Contemporary English.* London: Longman.

Quirk R., S. Greenbaum, G. Leech and J. Svartvik (1985) *A Comprehensive Grammar of the English Language.* London: Longman.

Rayson P. (2003) *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison.* Unpublished PhD thesis, Lancaster University.

Rayson P., D. Berridge and B. Francis (2004) Extending the Cochran rule for the comparison of word frequencies between corpora. In Purnelle G., C. Fairon C. and A. Dister (eds) *Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004),* Louvain-la-Neuve, Belgium, March 10-12, 2004. Louvain-la-Neuve: Presses universitaires de Louvain, 926 - 936.

Rayson P. and R. Garside (2000) Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics*,. October 2000, Hong Kong, 1-6.

Renouf A. and J. Sinclair (1991) Collocational frameworks in English. In Aijmer K. and B. Altenberg (eds) *English corpus linguistics: studies in honour of Jan Svartvik.* London and New York: Longman, 128-143.

Reynolds D. (2005) Linguistic correlates of second language literacy development: Evidence from middle-grade learner essays. *Journal of Second Language Writing* 14(1): 19-45.

Rietveld T., R. Van Hout and M. Ernestus (2004) Pitfalls in Corpus Research. *Computers and the Humanities* 38:343-362.

Ringbom H. (1978) The influence of the mother tongue on the translation of lexical items. *Interlanguage Studies Bulletin* 3: 80-101.

Ringbom H. (1986) Cross-linguistic influence and the foreign language learning process. In Kellerman E. and M. Sharwood Smith (eds) *Cross-linguistic Influence in Second Language Acquisition*. New York: Pergamon Press, 150-162.

Ringbom H. (1987) *The Role of the First Language in Foreign Language Learning*. Clevedon and Philadelphia: Multilingual Matters.

Ringbom H. (1998) Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In Granger S. (ed.), 41-52.

Ringbom H. (1999) High-frequency verbs in the ICLE corpus. In Renouf A. (ed.) *Explorations in Corpus Linguistics*. Amsterdam and Atlanta: Rodopi, 191-200.

Ringbom H. (2001) Lexical transfer in L3 production. In Cenoz J., B. Hufeisen and U. Jessner (eds) *Cross-linguistic influence in third language acquisition: psycholinguistic perspectives*. Clevedon: Multilingual Matters, 59-68.

Ringbom H. (2006) The importance of different types of similarity in transfer studies. In Arabski J. (ed.), 35-45.

Ringbom H. (2007) *Cross-linguistic Similarity in Foreign Language Learning*. Clevedon: Multilingual Matters.

Ruiz Gurillo L. (1997) Aspectos de fraseología teórica española. Anejo XXIV de *Cuadernos de Filología*.

Rundell M. (1998) Recent trends in English pedagogical lexicography. *International Journal of Lexicography* 11(4): 315-342.

Rundell M. (1999) Dictionary use in production. *International Journal of Lexicography* 12(1):35-53.

Rundell M. (ed.) (2007) *Macmillan English Dictionary for Advanced Learners* (Second Edition). Oxford: Macmillan Education.

Sag I., T. Baldwin, F. Bond, A. Copestake and D. Flickinger (2002) Multi-word expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING 2002), Mexico City, 1-15.

Salkie R. (2002) Two types of translation equivalence. In Altenberg B. and S. Granger (eds) *Lexis in Contrast: corpus-based approaches*. Amsterdam and Philadelphia: Benjamins, 51-71.

Salkind N.J. (2005) *Statistics for people who (think they) hate statistics*. Thousand Oaks, California: Sage.

Saussure F. (de) (1982) *Cours de Linguistique Générale* (published by Ch. Bally et A. Sechehaye, with the collaboration of A. Riedlinger ; critical edition by T. de Mauro). Paris: Payot.

Scarcella R.C. and C.B. Zimmerman (2005) Cognates, cognition and writing: an investigation of the use of cognates by university second-language learners. In Tyler A., M. Takada, Y. Kim and D. Marinova (eds) *Language in Use: Cognitive and Discourse Perspectives on Language and Language Learning*. Washington: Georgetown University Press, 123-136.

Schachter J. (1974) An error in error analysis. *Language Learning* 24(2): 205-214.

Schenk A. (1995) The syntactic behavior of idioms. In Everaert M., E.J. Van der Linden and A. Schenk (eds) *Idioms: structural and psychological perspectives*. Hillsdale, Hove: Lawrence Erlbaum Associates, 253-271.

Schleppegrell M.J. (1996) Conjunction in Spoken English and ESL Writing. *Applied Linguistics* 17(3):271-285.

Schmid H-J. (2000) *English Abstract Nouns as conceptual shells: From corpus to cognition.* Berlin: Mouton de Gruyter.

Schmid H.-J. (2003) Collocation: hard to pin down, but bloody useful. *ZAA* 51(3): 235-258.

Schmitt N. (ed.) (2004) *Formulaic Sequences: Acquisition, Processing and Use.* Amsterdam and Philadelphia: Benjamins.

Schmitt N. and D. Schmitt (2005) *Focus on vocabulary: Mastering the Academic Word List.* London: Longman.

Schmitt N. and C.B. Zimmerman (2002) Derivative word forms: what do learners know? *TESOL Quarterly* 36(2):145-171.

Schwartz B.D. and R.A. Sprouse (1996) L2 cognitive states and the full transfer/full access model. *Second Language Research* 12: 40-72.

Scott M. (1997) PC analysis of keywords and key keywords. *System* 25(2): 233-245.

Scott M. (2001) Comparing corpora and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs. In Ghadessy M., A. Henry and L. Roseberry (eds) *Small Corpus Studies and ELT.* Amsterdam: Benjamins, 47-67.

Scott M. (2004) *WordSmith Tools 4*. Oxford: Oxford University Press.

Scott M. and C. Tribble (2006) *Textual patterns: Key words and corpus analysis in language education.* Amsterdam: Benjamins.

Seidlhofer B. (2001) Closing a conceptual gap: the case for a description of English as a lingua franca. *International Journal of Applied Linguistics* 11(2):133-58.

Selinker L. (1972) Interlanguage. *IRAL* X(3):209-231.

Selinker L. (1992) *Rediscovering interlanguage.* London and New York: Longman.

Shaw P. (2004) The development of Swedish university students' written English, appropriacy, scope and coherence. Proceedings of the *Ninth Nordic Conference for English Studies, Aarhus, Denmark, May 27-29.* Available from <http://www.hum.au.dk/engelsk/naes2004/papers.html>

Siegel M. (2002) *Like*: the discourse particle and semantic. *Journal of Semantics* 19(1):35-71.

Siepmann D. (2005) *Discourse markers across languages: a contrastive study of second-level discourse markers in native and non-native text with implications for general and pedagogic lexicography.* London and New York: Routledge.

Simpson R. and D. Mendis (2003) A corpus-based study of idioms in academic speech. *TESOL Quarterly* 37(3): 419-441.

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

Sinclair J. (1992) The automatic analysis of corpora. In Svartvik J. (ed.) *Directions in Corpus Linguistics.* Berlin and New York: Mouton de Gruyter, 378-397.

Sinclair J. (1996) The empty lexicon. *International Journal of Corpus Linguistics* 1(1): 99-119.

Sinclair J. (1999a) A way with common words. In Hasselgård H. and S. Oksefjell (eds) *Out of corpora: studies in honour of Stig Johansson.* Amsterdam and Atlanta: Rodopi, 157-179.

Sinclair J. (1999b) The lexical item. In Weigand E. (ed.) *Contrastive lexical semantics.* Current Issues in Linguistic Theory 17. Amsterdam and Philadelphia: Benjamins, 1-24.

Sinclair J. (2004 [1996]) The search for units of meaning. In Sinclair J. (ed.) *Trust the Text: language, corpus and discourse.* New York: Routledge, 24-48.

Sinclair J. (2004a) New evidence, new priorities, new attitudes. In Sinclair J. (ed.) *How to use corpora in language teaching.* Amsterdam and Philadelphia: Benjamins, 271-299.

Sinclair J. (2004b) Intuition and annotation – the discussion continues. In Aijmer K. and B. Altenberg (eds) *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora.* Amsterdam and New York: Rodopi, 39-59.

Sinclair J. (2004c) Current issues in corpus linguistics. In Sinclair J. (ed.) *Trust the Text: language, corpus and discourse.* New York: Routledge, 185-193.

Sinclair J. (2005) Corpus and Text - Basic Principles. In M. Wynne (ed.) *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxbow Books, 1-16. Available online from <http://ahds.ac.uk/linguistic-corpora/>

Singleton D. (1987a) The Fall and Rise of Language Transfer. In Coleman J. and R. Towell (eds) *The Advanced Language Learner.* London: Centre for information on language teaching and research, 27-53.

Singleton D. (1987b) Mother and other tongue influence on learner French: a case study. *Studies in Second Language Acquisition* 9: 32745.

Singleton D. (1999) *Exploring the second language mental lexicon.* Cambridge: Cambridge University Press.

Singleton D. and D. Little (1991) The second language lexicon: some evidence from university-level learners of French and German. *Second Language Research* 7(1):61-81.

Sjöholm K. (1998) A reappraisal of the role of cross-linguitic and environmental factors in lexical L2 acquisition. In Haastrup K. and A. Viberg (eds) *Perspectives on Lexical Acquisition in a Second Language.* Lund: Lund University Press, 209-236.

Skandera P. (2004) What are idioms? In Allerton et al. (eds), 23-36.

Smadja F. (1993) Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1): 143-177.

So B. (2005) From analysis to pedagogic applications: using newspaper genres to write school genres. *Journal of English for Academic Purposes* 4: 67-82.

Sonck-Mercier A., B. Couvreur-Loosen and J.P. Nyssen (1991) Towards a common core for ESP. In Granger S. (ed.) *Perspectives on the English Lexicon: A Tribute to Jacques Van Roey.* Cahiers de l'Institut de Linguistique de Louvain 17:281-289.

Söderberg Arnfast J. and J. Normann Jørgensen (2003) Code-switching as a communication, learning, and social negotiation strategy in first-year learners of Danish. *International Journal of Applied Linguistics* 13(1):23-53.

Stefanowitsch A. and S. Gries (2003) Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209-243.

Stotesbury H. (2003) Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes* 2(4):327-341.

Strevens P. (1973) Technical, technological, and Scientific English. *ELT Journal* 27(3): 223-234.

Stubbs M. (1986) Language development, lexical competence and nuclear vocabulary. In Stubbs M. (ed.) *Educational Linguistics.* Oxford and New York: Blackwell, 98-115.

Stubbs M. (1993) British traditions in text analysis: from Firth to Sinclair. In Baker M., G. Francis and E. Tognini-Bonelli (eds) *Text and technology: in honour of John Sinclair*. Philadelphia and Amsterdam: Benjamins, 1-33.

Stubbs M. (1995) Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language* 2(1): 23-55.

Stubbs M. (2001) *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

Stubbs M. (2002) Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7(2): 215-244.

Stubbs M. (2006) Review of Dirk Siepmann, Discourse Markers Across Languages. A Contrastive Study of Second-Level Discourse Markers in Native and Non-native Text with Implications for General and Pedagogic Lexicography. London and New-York: Routledge, 2005. *Languages in Contrast* 7(1):101-103.

Sung Park E. (2004) The comparative fallacy in UG studies. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics* 4(1). Available from <http://www.tc.columbia.edu/academic/tesol/WJFiles/pdf/EunSung2004.pdf>

Svensson M.H. (2002) Critères de figement et conditions nécessaires et suffisantes. *Romansk Forum* 16(2): 777-783.

Swales John M. (1990) *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales J.M. (2002) Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (ed.) *Academic Discourse*. Harlow: Pearson Education, 150-164.

Swales J., U. Ahmad, Y. Chang, D. Chavez, D. Dressen and R. Seymour (1998) "Consider this: the role of imperatives in scholarly writing". *Applied Linguistics* 19(1): 97-121.

Swan M. (1997) The influence of the mother tongue on second language vocabulary acquisition and use. In Schmitt N. and M. McCarthy (eds) *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, 156-180.

Swan M. and B. Smith (2001) *Learner English: A teacher's guide to interference and other problems* (second edition). Cambridge: Cambridge University Press.

Tan M. (2005) Authentic language or language errors? Lessons from a learner corpus. *ELT Journal* 59(2):126-134.

Tankó G. (2004) The use of adverbial connectors in Hungarian university students' argumentative essays. In Sinclair J. (ed.) *How to Use Corpora in Language Teaching*. Amsterdam and Philadelphia: Benjamins, 157-181.

Tapper M. (2005) Connectives in advanced Swedish EFL learners' written English – preliminary results. *The Department of English in Lund: Working Papers in Linguistics* 5:115-144. Available from <http://www.englund.lu.se/content/view/173/209/>

Tenfjord K., P. Meurer and K. Hofland (2006) The ASK Corpus – a language learner corpus of Norwegian as a second language. *LREC 2006 Proceedings Online*, 1821-1824. Available from <http://nl.ijs.si/sdjt/bib/lrec06/>

Thewissen J. (2006) *Using evidence from error-tagged learner corpora to inform written language assessment: The case of the CEF descriptors for essay writing*. Unpublished MA dissertation. Louvain-la-Neuve: Université catholique de Louvain.

Thewissen J., Y. Bestgen and S. Granger (2006) Using error-tagged learner corpora to create English-specific CEF descriptors. Paper presented at the *Third Annual Conference of EALTA*, 19-21 May 2006, Krakow, Poland.

Thompson G. (2001) Interaction in academic writing: learning to argue with the reader. *Applied Linguistics* 22(1): 58-78.

Thompson G. (2001) A pedagogically-motivated corpus-based examination of PhD theses: macrostructure, citation practices and uses of modal verbs. Unpublished PhD thesis, University of Reading, UK. Available from <http://paulslals.org.uk/>

Thompson P. (2006) Assessing the contribution of corpora to EAP practice. In Z. Kantaridou, I. Papadopoulou and I. Mahili (eds) *Motivation in Learning Language for Specific and Academic Purposes*. Macedonia: University of Macedonia [CDROM], no page numbers given. Available from <http://www.rdg.ac.uk/app_ling/thompson_pub.htm>

Thompson P. and C. Tribble (2001) Looking at citations: using corpora in English for Academic Purposes. *Language Learning and Technology* 5(3): 91-105.

Thurstun J. and C. Candlin (1997) *Exploring Academic English. A Workbook for Student Essay Writing*. Sydney: NCELTR Publications.

Thurstun J. and C. Candlin (1998) Concordancing and the teaching of the vocabulary of Academic English. *English for Specific Purposes*, 17(3): 267-280.

Tognini-Bonelli E. (2001) *Corpus Linguistics at Work*. Amsterdam and Philadelphia: Benjamins.

Tognini-Bonelli E. (2002) Functionally complete units of meaning across English and Italian: towards a corpus-driven approach. In Altenberg B. and S. Granger (eds) *Lexis in contrast: corpus-based approaches*. Amsterdam and Philadephia: Benjamins, 73-95.

Tono Y. (2004) Multiple comparisons of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In Aston G., S. Bernardini and D. Stewart (eds) *Corpora and Language Learners*. Amsterdam and Philadelphia: Benjamins, 45-66.

Trauth G. and K. Kazzazi (eds) (1996) *Routledge Dictionary of Language and Linguistics*. London and New York: Routledge.

Tribble C. (1998) *Writing Difficult Texts*. Unpublished PhD thesis, Lancaster University. Available from <http://www.ctribble.co.uk/text/phd.htm>

Tribble C. (2000) Genres, keywords, teaching: towards a pedagogic account of the language of project proposals. In Burnard L. and T. McEnery (eds) *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora*. Hamburg: Peter Lang. Available from <http://www.ctribble.co.uk/text/Genre.htm>

Tribble C. (2001) Small corpora and teaching writing: towards a corpus-informed pedagogy of writing. In Ghadessy M., A. Henry and R. Roseberry (eds) *Small corpus studies and ELT: theory and practice*. Amsterdam and Philadelphia: Benjamins, 381-408.

Tschichold C. (2000) *Multi-word Units in a Lexicon for Natural Language Processing*. Olms: Hildesheim.

Tseng, Y.-C. and H.-C. Liou (2006) The effects of online conjunction materials on college EFL students' writing. *System* 34: 270-283.

Tsohatzidis S.L. (1990) Introduction. In Tsohatzidis S.L. (ed.) *Meanings and Prototypes. Studies in Linguistic Categorization*. London and New York : Routledge, 1-13.

Tutin A. and F. Grossmann (2002) Collocations régulières et irrégulières: esquisse d'une typologie du phénomène collocatif. *Revue Française de Linguistique Appliquée* 7(1): 7-25. Available from <http://www.u-grenoble3.fr/grossmann/Articles%20en%20ligne/article_RFLA_tutin_grossmann.htm>

Vande Kopple (1985) Some exploratory discourse on metadiscourse. *College Composition and Communication* 36:82-93.

Van der Meer G. (1998) Collocations as one particular type of conventional word combinations: their definitions and character. In Fontenelle T., P. Hiligsmann, A. Michiels, A. Moulin and S. Theissen (eds) *Euralex'98* proceedings.

Van Roey J. (1990) *French-English contrastive lexicology: an introduction.* Louvain-la-Neuve: Peeters.

Vassileva I. (1998) Who am I/who are we in academic writing? A contrastive analysis of authorial presence in English, German, French, Russian and Bulgarian. *International Journal of Applied Linguistics* 8(2):163-190.

Viberg A. (1998) Cross-linguistic perspectives on lexical acquisition: the case of language-specific semantic differentiation. In Haastrup K. and A. Viberg (eds) *Perspectives on Lexical Acquisition in a Second Language.* Lund: Lund University Press, 175-208.

Viberg A. (2002) Basic verbs in lexical progression and regression. In Burmeister P., T. Piske and A. Rohde (eds) *An Integrated View of Language Development. Papers in Honor of Henning Wode.* Trier: Wissenshaftliche Verlag, 109-134.

Viberg A. (2004/2005) The lexical typological profile of Swedish mental verbs. *Languages in Contrast* 5(1):121-157.

Vinay J.P. and J. Darbelnet (1958) *Stylistique comparée de l'anglais et du français.* Montréal : Beauchemin.

Violi P. (2000) Prototypicality, typicality and context. In Albertazzi L. (ed.) *Meaning and Cognition.* Amsterdam and Philadelphia: Benjamins, 103-122.

Virtanen T. (1998) Direct questions in argumentative student writing. In Granger S. (ed.), 94-118.

Virtanen T. and J. Agerström (eds) (2000) *Three Studies of Learner Discourse: Evidence from the International Corpus of Learner English.* Rapporter Från Växjö Universitet, Humaniora 10, Växjö Universitet

Vold E.T. (2006) Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study. *International Journal of Applied Linguistics* 16 (1): 61–87.

Voutilainen A. (1999) A short history of tagging. In van Halteren H. (ed.) *Syntactic wordclass tagging.* Dordrecht: Kluwer Academic Publishers, 9-21.

Wang W. and Q. Wen (2002) L1 use in the L2 composing process: An exploratory study of 16 Chinese EFL writers. *Journal of Second Language Writing* 11(3): 225-246.

Ward J. W. (1999) How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language* 12(2): 309–324.

Weissberg R. and S. Buker (1978) Strategies for teaching the rhetoric of written English for Science and Technology. *TESOL Quarterly* 12(3):321-329.

Weisser M. (2001) A corpus-based methodology for comparing and evaluating native and non-native speaker accents. Unpublished PhD thesis. Lancaster University. Available from <http://www.tu-chemnitz.de/phil/english/chairs/linguist/documents/mw/publications/Thesis.pdf>

West M. (1937) The present position in vocabulary selection for foreign language teaching. *The Modern Language Journal* 21(6): 433-437.

West M. (1953) *A General Service List of English Words.* London: Longman.

Wiberg L. (2000) Involvement in argumentative writing: a comparison between Swedish and American argumentative essays and American editorials. In Virtanen T. and J. Agerström (eds), 43-80.

Widdowson H.G. (2003) *Defining issues in English language teaching*. Oxford: Oxford University Press.

Wiktorsson M. (2001) Register differences between prefabs in native and EFL English. The Department of English in Lund: *Working Papers in Linguistics* 1:85-94. Available from <http://ask.lub.lu.se/archive/00009277/01/Maria.pdf>

Wiktorsson M. (2003) *Learning idiomaticity: a corpus-based study of idiomatic expressions in learners' written production*. Lund: Lund Studies in English 105.

Wilkins D.A. (1976) *Notional Syllabuses*. Oxford : Oxford University Press.

Williams G. (2001) Sur les caractéristiques de la collocation. TALN2001, Tours, 2-5 July 2001, 9-16.

Williams G. (2003) Les collocations et l'école contextualiste britannique. In Grossman and Tutin (eds), 33-44.

Willis D. (2003) *Rules, Patterns and Words. Grammar and Lexis in English Language Teaching*. Cambridge: Cambridge University Press.

Wilson A. and J. Thomas (1997) Semantic annotation. In Garside et al (eds), 53-65.

Winter E. (1994) Clause relations as information structure: two basic text structures in English. In Coulthard M. (ed.) *Advances in written text analysis*. London: Routledge, 46-68.

Witt Jörg (2000) English as a global language: The case of the European Union. *Erfurt Electronic Studies in English* 11. Available from <http://webdoc.gwdg.de/-edoc/ia/eese/-artic20/witte/6_2000.html>

Wolter B. (2001) Comparing the L1 and L2 mental lexicon. *Studies in Second Language Acquisition* 23(1):41-69.

Wolter B. (2006) Lexical network structures and L2 vocabulary acquisition: the role of L1 lexical/conceptual knowledge. *Applied Linguistics* 27(4):741-747.

Woolard G. (2000) Collocation – encouraging learner independence. In Lewis M. (ed.) *Teaching Collocation: Further Developments in the Lexical Approach*. London: LTP, 28-46.

Wray A. (1999) Formulaic language in learners and native speakers. *Language Teaching* 32: 213-231.

Wray A. (2000) Formulaic sequences in second language teaching: theory and practice. *Applied Linguistics* 21(4): 463-489.

Wray A. (2002) *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray A. and M. Perkins (2000) The functions of formulaic language: an integrated model. *Language and Communication* 20: 1-28.

Xiao Z. and A. McEnery (2006) Collocation, semantic prosody and near synonymy: a cross-linguistic perspective. *Applied Linguistics* 27(1):103-12.

Xue G. and P. Nation (1984) A University Word List. *Language Learning and Communication* 3(2): 215-229.

Yang H. (1986) A new technique for identifying scientific / technical terms and describing science texts. *Literary and Linguistic Computing* 1(2): 93-103.

Yorio C. A. (1989) Idiomaticity as an indicator of second language proficiency. In Hyltenstam K. and L. Obler (eds) *Bilingualism across the lifespan*. Cambridge: Cambridge University Press, 55-72.

Zamel V. (1983) Teaching those missing links in writing. *ELT Journal* 37(1): 22-29.

Zhang M. (2000) Cohesive features in the expository writing of undergraduates in two Chinese universities. *RELC Journal* 31:61-95.

Zhang H., C. Huang and S. Yu (2004) Distributional consistency: as a general method for defining a core lexicon. Proceedings of the *Fourth International Conference on Language Resources and Evaluation,* Lisbon, Portugal, 26-28 may 2004. <http://data.cstr.ed.ac.uk/internal/library/proceedings/2004/lrec2004/>

Zgusta L. (1971) *Manual of lexicography.* Prague: Academia.

Zimmerman (1987) Form-oriented and content oriented lexical errors in L2 learners. *IRAL* 25(1): 55-67

Zuluaga A. (1980) *Introducción al estudio de las expresiones fijas*. Frankfurt aim Bern: Peter Lang.

Zuluaga A. (2002) Los “enlaces frecuentes” de María Moliner. Observaciones sobre las llamadas colocaciones. *LEA* 24(1): 97-114.