Title Page

A corpus-based investigation of the phraseology in various genres of written English with applications to the teaching of English for academic purposes

John Anthony McKenny

Submitted in accordance with the requirements for the degree of Ph.D.

The University of Leeds, School of English.

October 2006

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

In memory of Patrick John McKenny 1918-2005

Acknowledgements

I would like to thank my first supervisor, Professor Katie Wales, for safely guiding me in the writing and learning process of this thesis and for encouraging me to keep going. My second supervisor, Dr. Anthea Fraser Gupta gave me the searching criticism and sound advice I needed to complete my thesis.

My friend and colleague, Mark Garner, challenged me to think and write better by his discussion and criticism of my work.

Thanks are due to Peter Howarth, Göran Kjellmer, Alice Deignan, Eric Atwell, Breffni O'Rourke, David Hardisty, Loreto Todd, Peter Sercombe, Chris Reed, Peter Grundy, David Common, Kim Willcocks and Stephen Parkin for their inspiration and encouragement and to Hans Van Halteren, Attila Ajtai, Matthias Romppel and Paul Rayson for allowing me to benefit from their technical expertise.

Contents

Abstract

This thesis describes an investigation into the various kinds of fixed expressions or prefabricated language which occur in certain genres of written academic English. A basic premise is that language users, when they write, remember phrases as much as, or more than, they compose them. Although the grammar of a language licenses the use of a variety of forms to express any proposition only a small subset of these grammatically possible locutions are considered natural and native-like. It is demonstrated that prefabs, my preferred term for prefabricated language, serve many functions in written academic discourse. Their use helps the writer to generate idiomatic text which meets the expectations of the reader and shows clearly the writer's discourse community. Prefabs also enable the configuration of writer's stance more explicitly. At the same time, thanks to prefabs the reader is better able to navigate within the text.

The main focus of the investigation is the use by non-native speakers of English of prefabs in writing. Corpus linguistics is presented as the most appropriate methodology for this investigation. Two main kinds of corpora are constructed: an experimental corpus of argumentative essays produced by Portuguese 'apprentice' English for Academic Purposes (EAP) writers and three control corpora of texts of comparable length, by different categories of native speakers of English. The kinds and frequencies of prefabs in the corpora are measured and tabulated. Finally, recommendations are made on how to use the findings of the research to improve EAP teaching and learning programmes.

Abbreviations used in this thesis

ADJP	Adjective phrase
ADVP	Adverbial phrase
AI	Artificial intelligence
ALTE	Association of Language Testers in Europe and corresponding proficiency bandings 1-4
AP	Adjective phrase and/or adverbial phrase
ASCII	American Standard Code for Information Interchange. Text with no mark-up.
AWL	Academic Wordlist (Averil Coxhead 1997)
Bawe corpus	British Academic Written English corpus held at Warwick University
BNC corpus	British National Corpus (100 million words of 1990's English)
BNC-Baby corpus	Four million-word sub-corpora of BNC in XML
BNC World corpus	Version 2 of British National Corpus
BRICLE corpus	Brazilian sub-corpus of *ICLE*
Brown corpus	One-million word corpus of written American English from 1961 compiled at Brown University
CAE	Cambridge Advanced English Examination
CANCODE corpus	Cambridge Nottingham Corpus of Demotic English
CARS	Create a Research Space (Swales 1990)
CIA	Contrastive interlanguage analysis
CLAWS	The Constituent Likelihood Automatic Word-tagging System
CLC	Computer Learner Corpora
CPE	Cambridge Proficiency of English Examination
CofE corpus	Corpus of Expert writers
CORPORA	mailing list for scholars interested in all aspects of corpus and computational linguistics
DOTA	Dogmatism Text Analysis
DQ	Dogmatism quotient
EA	Error analysis
EAGLES	Expert Advisory Groups on Language Engineering Standards
EAP	English for academic purposes

EFL	English as a foreign language
EIL	English as an international language
ELF	English as a lingua franca
ELT	English language teaching
ESOL	English to speakers of other languages
ESP	English for specific purposes
FCE	First Certificate Examination of English
FLOB	The Freiburg-LOB Corpus of British English from 1991
FRICLE corpus	French sub-corpus of ICLE
HECTOR corpus	18-million word corpus of written English used by Moon (1998) in study of idiom
HKUST corpus	Hong Kong University of Science and Technology corpus
ICLE corpus	International Corpus of Learner English
IELTS	International English Language Testing Services
KFNgram	A program which generates lists of *n-grams* in text and HTML files
KWIC	Key word in context
L1	First language
L2	Second language
LL	Log likelihood
LINDSEI corpus	Louvain International Database of Spoken English Interlanguage
Locness corpus	Louvain Corpus of Native Essays
LOB	Lancaster Oslo Bergen corpus. One-million-word corpus of written British English designed to mirror the *Brown* corpus
LLC corpus	Longman Learner Corpus or London Lund Corpus
N-gram	A sequence of N words occurring in a text
NP	Noun phrase
NS(s)	Native speaker(s)
NNS(s)	Non-native speaker(s)
POS	Part of Speech (tagging)
PORTICLE corpus	Portuguese sub-corpus of ICLE
PP	Prepositional phrase
RP	CLAWS tag for prepositional adverb or particle
SLA	Second language acquisition
SPICLE corpus	Spanish sub-corpus of *ICLE*
STTR	Standardized type-token ratio
SWICLE corpus	Swedish sub-corpus of *ICLE*
TESOL	Teaching English to speakers of other languages
TOEFL	Test of English as a Foreign language

TOSCA	Tools for Syntactic Corpus Analysis
UCLES	University of Cambridge Local Examination Syndicate
USAS	The UCREL semantic analysis system is a software system for undertaking the automatic semantic analysis of text.
VP	Verb phrase
TTR	Type-token ratio
Wmatrix	A software tool for corpus analysis and comparison.
XLI	Cross-linguistic influence

List of tables and figures

List of tables and figures

Chapter 1 A phraseological approach to written academic English

1.0 Introduction

The pre-eminent role of English as the *lingua franca* of science, technology and other branches of learning has increased the importance of English for academic purposes (EAP) in university curricula throughout the world. Concomitantly, there has been an intensification of research in applied linguistics, the discipline which has traditionally sustained the teaching of EAP.

The educational purpose and praxis which informs this thesis is the teaching of EAP. The starting point and inspiration of the study is the classroom reality of non-native tertiary-level students learning and being taught to write English for academic purposes. The main objective of EAP teaching is to help students learn to listen to, speak, read and write English more effectively in the pursuit of their studies and in their subsequent professional lives. This thesis investigates aspects of phraseology in written academic English and applies the results to the teaching of EAP. When referring to English for academic purposes or when using the corresponding acronym, EAP, care is needed to avoid reification or the supposition that a body of knowledge already exists. Elbow (1998:148) warns against such hypostatization: 'the problem is that we can't teach academic discourse because there's no such thing to teach'.

Applied linguistics is an approach to language which has grown principally, but not exclusively, out of the context of language learning and teaching. Consequently, applied linguists have a responsibility to consider the criteria for an educationally relevant approach to language and to avoid 'the uncritical assumption that applied linguistics must necessarily be the application of linguistics' (Widdowson 1984:19). The educational purpose or the praxis comes first and the theory is developed for its relevance and utility. This applied linguistics is opposed to 'linguistics applied' where linguistics is applied uncritically to some practices such as language teaching, speech therapy or lexicography, resulting in a

practice which is 'essentially conformist... with the tendency...to dance attendance to whatever tune is currently in theoretical fashion' (Widdowson 1984).

In learning the conventions of EAP, the students need to write texts which read naturally and idiomatically. *The Oxford English Dictionary* (Vol. 7, 1989) defines 'idiomatic' in the following way:

> Peculiar to or characteristic of a particular language; pertaining to or exhibiting the expressions, constructions, or phraseology approved by the peculiar usage of a language, esp. as differing from a strictly grammatical or logical use of words; vernacular; colloquial.

It is clear from this definition, however, that the EAP teacher's main task is not simply a question of increasing students' vocabulary or making them more grammatically aware. There would appear to be a competence between the level of lexical choice and the level of syntactic decision where the language user knows or intuits which words go together and avoids other equally grammatical sequences. Howarth (1998:36) refers to this level as that of 'phraseological competence'.

The research for this thesis, which developed out of my work as an EAP teacher, investigates the extent to which writers choose their words in memorized chunks or clusters rather than word by word in linear sequence. The contribution that these chunks make to the idiomaticity of texts is also considered. In particular, a comparison is made of language learners' and native speakers' usage of prefabricated language in argumentative essays. The manner in which their level of expertise affects various groups of writers' use of prefabricated language is investigated by comparing a corpus of learner English and three native-speaker control corpora. A study is made of those sequences of words found in the corpora whose joint selection appears to be constrained by the idiomatic patternings of the English language. These combinations of words recur in texts more frequently than would be expected from the individual frequencies of their constituent words. A major presupposition of my thesis is that as well as providing the building blocks for idiomatic texts, these prefabricated chunks also provide the mortar in that they fulfil important textual and pragmatic functions.

Although chunks usually comply with the rules of syntax (with some notable non-canonical exceptions, e.g. *by and large*), it is unlikely that they, and only they, could be

generated by a set of phrase structure and transformation rules or, for that matter, by any other grammatical model. Indeed it should be observed that some thinkers (e.g. Garner 2004) suggest that the roles of syntax and prefabricated chunks should be reversed and that the syntax of a language is actually derived from the chunks or patterns which it exhibits. If the grammar cannot generate all and only the chunks which make up the patterning of the language, this means that they must be stored in memory or be in some way routinized. If language is learned and stored as chunks, then recall as chunks may be facilitated. Now, at this stage, it is not being suggested that chunks are held consciously in memory: merely that they can be summoned by the speaker/writer from his/her mental lexicon. In other words, these chunks or prefabs, as I shall call them, might be automatized or internalized in the native speaker. Some writers, notably Wray (2002) and Cook (1998), suggest that adult language learners do not usually exploit such formulaic sequences in their language acquisition but tend to have a more analytical word-centred approach to language learning. Might a way be found to make such patterns available to EAP learners?

The concept of chunking has relevance beyond explanations of how users produce and comprehend spoken language. In psycholinguistic modelling of speech production and comprehension, the time constraints in most conversational situations make plausible the supposition that participants utilize and recognize memorized chunks. However, in writing, time is not always such a pressing factor. Obviously, the writer might string together set phrases and insert the occasional personal touch when writing such routine texts as memoranda, business e-mails, notes to a schoolteacher excusing a child's absence, or writing character references. Some genres are more formulaic than others. What *is* a pressing factor is the set of expectations held by the reader and the constraints of the genre being instantiated. The writer has to go some way towards meeting the reader's expectations by conforming to the generic structure. This equal focus on the writer and the reader recalls Bakhtin's (1929) argument that all writing is dialogic.

The reader's expectations are for certain words and phrases (and themes) to appear as predicted by the genre and for the text to *feel* right. If a list of key lexical items is given in advance, a reader can often predict the content of the text (Phillips 1984). In many writing situations, the writer does not wish the language to obtrude but wants the message to be understood. Even in the case of the most creative kinds of writing, we see the interplay between the novel and the conventional: arguably, the world's greatest writers and

thinkers are also the greatest 'plagiarists'. Science and literature, even the most iconoclastic, build on what has gone before. But if most language is borrowed, then writers' originality lies in how they configure their communal phrases and create bridges between them. Kristeva (1980:66), writing on Bakhtin, describes this well: 'every text is constructed as a mosaic of quotations, every text absorbs and transforms other texts'. Pascal (1662), in the introduction to his *Pensées* was aware of the same problem:

> Let no one say that I have said nothing new; the arrangement of the subject is new. When we play tennis, both players use the same ball, but one of them has a better aim.
>
> (Pascal 1662:247)

During the twentieth century, the idea that some language is prefabricated, and that it is recalled rather than constructed, appears in the writings of various linguists. They conclude that the divide between lexis and grammar is difficult to sustain and that grammatical generalizations are insufficient to explain word co-occurrence and the existence of preferred ways of saying things. Scholars from various disciplines, particularly Palmer (1933), Becker (1975), Cowie and Mackin (1975), Bolinger (1977) and Pawley and Syder (1983) drew attention to word collocations and word combinations of varying degrees of fixedness.

1.1 Outline of the thesis

Embedded as it in the context of EAP teaching and learning, this thesis belongs to the field of applied linguistics. Having become aware of the phraseological dimension of language through my reading, teaching and language learning experience, I decided to make it the focus of my Ph.D. research. Gradually, the following two research questions were established for this thesis:

1. **Do non-native (in this case, Portuguese) writers of academic English use a similar quantity of prefabs in their essays in comparison with native speakers?**

2. **Do these non-native writers use prefabs to perform the same functions as native speakers?**

These two research questions are addressed in this thesis using a corpus-based methodology. In order to investigate the properties of learner and native-speaker writing, corpora were compiled which represent populations of writers with degrees of expertise ranging from non-native speaker (NNS) undergraduate writers of EAP essays, through native-speaker (NS) undergraduates to professional journalists writing editorials and essays. Chapter 1 traces the recent history of ideas relevant to this investigation and provides a review of the literature that sustains it. In Chapter 2, those features of corpus linguistics which are relevant to this thesis are elucidated. In particular, the methodology of corpus compilation or selection and the use of concordancing software to isolate prefabs are explained.

Chapter 3 provides a report of the computer-aided examination of the corpora where the main focus is the phraseology found in a corpus of learner writing contrasted with three control corpora of native writing. Given the wealth of recurrent sequences in these corpora, there is a need to focus on certain kinds of prefabs or certain of the functions they are used to fulfil. As the essays collected in the corpora contain mostly argumentative or persuasive writing, it was decided in advance to pay particular attention to those expressions which contribute to writer's stance, audience design, and the management of the interpersonal dimension of communication. Writer's stance and related concepts are discussed below in Section 1.14. Although all the prefabs found in the corpora are reported on, e.g. the phrasal verbs and noun compounds, particular attention is paid to those examples which contribute to the configuration of writer's stance. Chapter 4 undertakes an analysis and explanation of the patterns registered in Chapter 3. The final chapter of the thesis, Chapter 5, examines the implications of the research for the EAP curriculum and possible pedagogical responses.

The remainder of the present chapter is organized as follows. The discipline of phraseology (Cowie and Howarth 1996) provides a large part of the theoretical framework of this work and is the subject of Section 1.4 of the present chapter. Sections 1.4-1.6 provide an introduction to recent ideas about collocations, idioms, chunks and other kinds of routinized language which need to be considered in order to arrive at an understanding of

prefabs. The work presented in these sections comes from phraseology, lexicography, second language acquisition (SLA) theory, Artificial Intelligence (AI), and psycholinguistics, and a certain convergence in the thinking of the researchers in these different fields is noted. The role of prefabricated sequences in the writing process is examined in Sections 1.7, 1.8 and 1.10. In order to focus my work more on the prevalence of chunking in written language, the differences between written and spoken language are rehearsed in Section 1.11. This section goes on to map out a <formal> <informal> cline which might help in assessing the texts contained in the corpora. Research into writing from an EAP perspective is described in Section 1.12. The theoretical constructs of genre analysis are introduced in Section 1.13 and then applied to argumentative essays and other written genres. Section 1.14 presents a model for the analysis of writer's stance, and the ways in which evaluation is incorporated into written texts.

1.2 A definition of prefab

The terms 'chunk' and 'chunking' were developed in psychology in relation to a theory of memory. 'Chunk' is one of the many terms which have been used to describe prefabricated language within linguistics and related fields. Various writers refer to the proliferation of terms found in the literature (Mel'čuk 1998; Cowie 1998:4; Wray 2002:8-9). Other terms which have been used to talk about recurrent expressions from, for example, a lexicographic point of view include *multi-word item* (Moon 1998), *composite* (Cowie 1998), and *phraseme* (Mel'čuk 1998). The 60 different terms listed by Wray (2002:9) however do not all refer to the same phenomenon and for that reason are not interchangeable. There may be some overlap in the denotations of some terms while certain terms designate much more restricted aspects of phraseology than others. One writer might naively adopt a non-technical term to designate a specific meaning while another writer deliberately selects a term together with all its pre-established connotations and theoretical presuppositions. The lexical patterns under consideration can cover a large range of linguistic phenomena, such as phrasal verbs (e.g. *add up*), nominal compounds (e.g. *telephone box*), and institutionalized phrases (e.g. *fish and chips*), and they can be

syntactically idiosyncratic (e.g. *by and large*) and/or semantically idiosyncratic in nature (e.g. *white noise*). Such strings of words are used frequently in everyday language, usually to express precisely ideas and concepts that cannot be compressed into a single word. Drawing on the concept of *chunk,* I define my own technical term, *prefab,* which is intended to be applicable to the phraseological study of machine-readable corpora.

In the course of the elaboration of this thesis, one term, *formulaic sequence,* appeared to have established itself as the one that was 'here to stay'. The term seems to have gained wide acceptance in the literature (see Schmitt and Carter 2004:3) and its definition is not incompatible with my own definition of prefab given below. Wray defines *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)

This label reappeared in the title and throughout the chapters of *Formulaic Sequences* (Schmitt (ed.) 2004). This latter work used the term proposed by Wray but, in so doing, deviated from the original meaning (Wray, personal communication). As a result, either a new term is needed or scholars must agree to return to using *formulaic sequence* in Wray's original sense.

This unresolved terminological problem influenced the choice of the term 'prefab' in this work. Several morphological features of the word made it preferable to other candidate terms. Firstly, it shows its back derivation from *fabricate.* Secondly, the prefix *pre-* conveys the sense of an expression already existing in the common stock of lexicalised phrases. Thirdly, it is a countable noun and permits reference to the number of prefabs occurring per 100,000 words and so on. The term prefab is also used by Bolinger (1977) and Cowie (1998:1), pioneers in the field of phraseology, and by scholars at Lund and Aston universities. The concept of prefab will be shown to have psycholinguistic, sociolinguistic, phraseological and pedagogical dimensions in what follows.

In this chapter, a number of definitions of prefabricated or formulaic language are given. The authors (Sinclair 1991; Kjellmer 1994; Howarth 1998; Wray 2002) use different

terms for this phenomenon but I would suggest that there is a common core of meaning shared by these diversely named concepts. Each author seems to give greater emphasis to one particular feature of this complex phenomenon. For instance, Wiktorsson (2003) gives the following definition of prefab, adapted from Erman and Warren (2000:31), which relies heavily on the notion of conventionalization.

> a prefab is a combination of at least two words used by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization.
>
> (Wiktorsson 2003:1)

The identification of prefabs in accordance with this definition requires a close investigation of all possible word combinations in each text. To be counted as a prefab a combination of words must manifest some feature of conventionalization (i.e. the combination is selected by native speakers in preference to other expressions). A search for Wiktorssonian prefabs is of necessity a labour-intensive process in which the investigator continually asks of each word sequence whether or not it could be varied: in other words, if it is possible for one element of a prefab to be replaced by a synonymous word without causing change of meaning or function and/or idiomaticity. An alternative, more anthropomorphic, definition of prefab is given by Van Roey (1990:46):

> the linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its 'synonyms' because of constraints which are not on the level of syntax or conceptual meaning but on that of usage.

Kjellmer (1982: 25) observes that a defining attribute of collocations is that they are word combinations which co-occur 'more often than the frequencies in the corpus of the constituents of the combination would lead us to expect'. The problem with this definition, Kjellmer laments, is that it includes not only combinations such as *last year* but also non-grammatical combinations such as *although he* or *and the.* In a later piece of work, Kjellmer (1987:133) proposed that in a one-million-word corpus such as the *Brown* corpus, a sequence of words must occur more than once and be grammatical in order to be accepted as a collocation. Kjellmer's use of the word 'grammatical' is puzzling and seems to be

more of an aesthetic than a syntactic concept. One potential collocation from the *Brown* Corpus that he discusses is *but too.* It seems that when Kjellmer requires that the sequence be grammatical or well-formed, the selectional restrictions of the first word must enable, or make conceivable the appearance of the second word. In the case of *but too* this seems to be the case: e.g. *but too many.* The important point for Kjellmer (personal communication) is that the criteria be specified in advance and applied strictly to avoid charges of subjectivity. Kjellmer's concern with the avoidance of subjectivity and adherence to a principled non-*ad hoc* approach resulted in highly predictable and relatively uninteresting collocations such as *a night* or *of the night* being included in his three-volume *Dictionary of English Collocations* (1994).

Wray and Namba (2003) propose eleven criteria which, when applied consecutively, make identification of prefabs more certain and transparent. The application of these criteria reproduces the ratiocination involved in distinguishing between prefabs and recurrent non-formulaic sequences.

Eleven criteria for identifying formulaic sequences (Wray and Namba 2003:28)

A: By my judgement, there is something grammatically unusual about this wordstring.

B: By my judgement, part or all of the wordstring lacks semantic transparency.

C: By my judgement, this wordstring is associated with a specific situation and/or register.

D: By my judgement, the wordstring as a whole performs a function in communication or discourse other than, or in addition to, conveying the meaning of the words themselves.

E: By my judgement, this precise formulation is the one most commonly used by this speaker/writer when conveying this idea.

F: By my judgement, the speaker/writer has accompanied this wordstring with an action, use of punctuation, or phonological pattern that gives it special status as a unit, and/or is repeating something he has just heard or read.

G: By my judgement, the speaker/writer, or someone else, has marked this wordstring grammatically or lexically in a way that gives it special status as a unit

H: By my judgement, based on direct evidence or my intuition, there is a greater than chance-level probability that the speaker/writer will have encountered this precise formulation before in communication from other people.

I: By my judgement, although this wordstring is novel, it is a clear derivation, deliberate or otherwise, of something that can be demonstrated to be formulaic in its own right.

J: By my judgement, this wordstring is formulaic, but it has been unintentionally applied inappropriately.

K: By my judgement, this wordstring contains linguistic material that is too sophisticated, or not sophisticated enough, to match the speaker's general grammatical and lexical competence.

These criteria were developed by Wray and Namba (2003) for the analysis of child language and were intended as a checklist, which when applied together, would help in the identification of prefabs. They pay careful attention to the different aspects of 'prefabhood': whether the sequence displays syntactic or semantic idiosyncrasy, is situation-bound, is non-compositional, frequent, salient, memorized, intended as prefab, misapplied, or is too advanced or not advanced enough for the speaker's linguistic competence. Used together, they provide a useful set of analytical tools for a linguist who wishes to initiate a quest for prefabs by interrogating a corpus. This checklist has the added advantage of providing a degree of objectivity when used by a group of investigators.

In research conducted at Lund University, Hudson (1998) arrives at a classification of what she describes as 'fixed expressions. In this classification, fixed expressions are classified according to the grammatical function they serve in the context. This results in a wordclass-based typology: compound verbs, compound nouns, compound adjectives, compound adverbs, compound prepositions, compound connectives and compound quantifiers.

Hudson's model of fixed expressions is of particular relevance to this research as she allows for degrees of fixedness. In her study of fixedness she uses the following two variability criteria:

1. UNEXPECTED SYNTACTIC CONSTRAINTS ON THE CONSTITUENT PARTS

NUMBER	the other day	* the other days
	(cf. the other boy/the other boys)	
ARTICLE	strike a light	* strike the light
	(cf. strike a match/strike the match)	

WORD ORDER trials and tribulations *tribulations and trials

(cf. sorrow and pain/pain and sorrow)

2. UNEXPECTED COLLOCATIONAL RESTRICTIONS WITHIN THE EXPRESSION

-first of all *second of all (cf. first in line/second in line)

-above board * below board (cf. above standard/below standard)

-disaster area * catastrophe area (major disaster/major catastrophe)

-how do you do * how do they do (how do you do it?/how do they do it?)

Similarly, modification is abnormally constrained:

1 for good * for very good

2 kick the bucket * kick the plastic bucket

(Hudson 1998: 37)

Hudson (1998:9) applies these two variability criteria in the following way:

> An expression that fails on either of the two variability criteria I consider to be fixed to some degree. The term fixed expression will be used to refer exclusively to expressions that are fixed according to variability criteria.

An important implication of this analysis is that it allows for degrees of fixedness and thus admits such phenomena as collocational frames e.g. *a ___ of; too ____ to* (Renouf and Sinclair 1991) and variable idioms (e.g. *close shave/narrow shave*). It also provides a framework for analysing the arbitrary way in which the extension of collocability by analogy is often blocked: we can *capture* or *catch someone's imagination* but probably not *take* it, while we can *catch* or *take someone's fancy* but not usually *capture* it.

Hudson's approach using substitutions as a test of fixedness has much in common with that of Howarth, which is discussed below in Section 1.5. The challenge is to find a way to separate two kinds of language, the free and the restricted (to some degree). As examples of free combinations, Howarth (1998:27) cites Hausman's (1979:188) reference to run-of-the-mill combinations, 'épithètes aussi banales que *belle, grande, vielle* avec

valise' (everyday adjectives like *beautiful , large, old* with *bag*) that are predictable and generated by the language system with nothing distinctive in their semantics or communicative function to make them institutionalized or memorable e.g. *affect world trade.* It would appear that, *pace* Hausman, it is hard not to stumble on conventionalized meanings in the most banal language production, e.g. *vielle valise* translated into English.

There is an inevitable degree of subjectivity in any attempt to filter out the prefabs from the recurrent N-grams (computational linguistics term for a sequence of N words) which occur in all English texts. For example, Altenberg (1993) found that 70% of the words of running text in the half-million-word London-Lund Corpus belong to recurrent word sequences. Many of these N-grams will be examples of Hausman's predictable run-of-the-mill combinations generated by the language system. Some scholars have warned that this category of free combinations may not be as large as has been surmised (Bolinger 1975). Arnold (1973), when discussing substitution as a means of delineating restrictedness, feels the need to set up an intermediate category of 'semi-fixed combinations':

> In semi-fixed combinations we are not only able to say that such substitutes exist, but fix their boundaries by stating the semantic properties of words that can be used for substitution, or even listing them. ...For example, the pattern consisting of the verb 'go' followed by a preposition and a noun with no article before it ('go to school', 'go to market', etc) is used with nouns of places where definite actions or functions are performed.
>
> (Arnold 1986:167-8)

Having studied the various definitions of formulaicity and idiomaticity for a number of years and applied them to numerous written texts, I learned to recognize those features that most definitions shared. My concentration on these particular features probably results from my interest in the pragmatic interpretation of actual instances of use.

Shared characteristics of definitions of prefab:

–Prefabs are not constructed at the time of utterance by applying the grammar to the lexis.
–They are stored in memory and recalled.
–They are conventionalized: i.e. there is a form-function mapping so that the prefab is the socially sanctioned and preferred form of expression in a given situation.

–Prefabs can be continuous or discontinuous strings of words.
–Although they are usually fairly transparent, the meaning of prefabs is often not completely derivable from the meaning of the constituent words; i.e. there is an additional meaning probably resulting from repeated application of the term to a situation (e.g. *bed and breakfast*).

For the purposes of this thesis, I have developed the following definition of prefab.

A prefab is a recognizable cluster of words which is stored as a unitary whole in memory and recalled for use. This recurrent continuous or discontinuous string of words is the preferred form of expression in certain repeated situations in the social world.

Particular emphasis is placed on those aspects of prefabs which influence language learning and the production of idiomatic written academic prose. In the case of some of the terms in my definition, e.g. *cluster, recurrent, string,* it is not too difficult to see how the use of the computer could greatly facilitate the investigation. These characteristics of lexical items can be quantified fairly automatically by text retrieval software. Certain other terms contained in the definition, e.g. *recognizable, unitary, preferred*, could only be applied by speakers of the language using their intuition and introspection or, in other words, their minds.

The next section examines some of the functional theories of language which have been advanced. Halliday's (1989) theory of the three metafunctions of language, the i*deational*, the *interpersonal* and the *textual*, is presented.

1.3 Prefabs and macro-functions

One way to classify prefabs is to divide them according to the parts of speech they most resemble. Hudson (1998), as mentioned in Section 1.2, demonstrates that dividing prefabs according to the roles they play in sentence structure is often a useful taxonomic principle. She identifies prefabs by their syntactic 'function,' i.e. 'according to the wordclass roles they play in context' (Hudson 1998:37). Those prefabs which function as nouns she calls

NP (noun phrase) and so on (VP, ADJP, ADVP, PP). She complains, however, of the intractability of the adverb class. As Quirk et al. (1972) observe:

> the adverb is the least satisfactory of the traditional parts of speech. Indeed, it is tempting to say simply that the adverb is an item that does not fit the definitions for other parts of speech.
>
> (Quirk et al. 1972: 267)

An example of the difficulties presented by adverbs is the way in which prepositional phrases can function as adverbials, e.g. *in fact,* and *in consequence.* I found in the search for adverb modifiers of adjectives that a search using a wild-card * *+ly* captured many words which were not adverbs (e.g. *silly, fly*) and did not capture some of the most frequent adverbs (*quite, very, too*).

Function is more frequently used in linguistics to refer to the semantic, discourse, or communicative roles of utterances or sentences. Wilkins (1976) draws a distinction between what we *do* through language and what we report by means of language. It is now well-established in applied linguistics that recurring situations in the social world commonly elicit the same or very similar verbal responses from speakers or writers. Such responses have come to be designated *functions* or are sometimes referred to as *speech acts* (following Austin 1962 and Searle 1969). Many lists of functions have been drawn up. For example, Schmitt and Carter (2004:9) give the following list:

> apologizing
> making requests
> giving directions
> complaining
> offering sympathy
> complying with a request

The Council of Europe publications, *The Waystage Level, The Threshold Level,* and *The Vantage Level* (van Ek 1975; van Ek, Alexander and Fitzpatrick 1977; van Ek and Alexander 1980; van Ek and Trim 1996) contain much lengthier lists of communicative functions. One of the problems experienced by the communicative approach to language teaching has been that the list of functions is potentially endless. Nattinger and DeCarrico

(1992: 62-63) imply a solution when they observe that common functions typically have conventionalised language attached to them, such as *I'm very sorry to hear about* _____ used to express sympathy or *I'd be pleased/happy to* _____ in acceptance of responsibility.

There have been several bipartite and tripartite classifications of language functions. Austin (1962) and Searle's (1969) theory of speech acts with locutions and their illocutionary and perlocutionary force is one such tripartite system as are Bühler's (1934) representational, conative and expressive functions. Brown and Yule (1983) divided language utterances into those having a more *transactional* function and those with a more *interactional* function. The transactional function in Brown and Yule's theory corresponds to language used to represent factual reality and experience. When this function is invoked it is the propositional content of what is said or written which is being considered. The interactional function refers to the social relations and personal attitudes which are expressed through language. Hunston (2001) develops a distinction between the autonomous and the interactive plane. A writer, suggests Hunston, is simultaneously an informer and a text-constructor. In each sentence the reader is informed of the content of the text (autonomous plane) and is simultaneously informed of the structure of the text (interactive plane). 'Every sentence in a text operates on each plane simultaneously' (Hunston 2001:183). Each of these functional classifications emphasizes different aspects of written communication. In a similar way, Hyland (2002:80-83) proposes a tripartite framework of *writer*, *texts*, and *audience* to deal with the question 'How should we teach writing?'

One of the first book-length and comprehensive treatments of prefabs is Wray's (2002) *Formulaic Language and the Lexicon,* although, not surprisingly, her term *formulaic sequence* is favoured. She examines the language of native speakers, adult and child learners, and aphasics and discusses how they learn and use 'formulaic language'. In a discussion of the linguistic function of formulaic sequences, she lists four candidate functions:

they play a role in easing the speaker's effort

where there are no such pressures on the speaker, formulaic sequences might take pressure of the hearer's comprehension (e.g. in a scripted weather forecast or auction which has a lot of information but only a small part being relevant to each listener)

they signal the speaker's identity as an individual or member of a group

they manipulate a hearer into a desired action or perception

(Wray 2002:93)

Wray (2002:93) goes on to inquire about the 'motivation *behind* the desire to speak fluently, express identity, organize text, and help the hearer to understand what you say' and suggests that, instead of seeing formulaic sequences as solutions to linguistic problems, they could instead be viewed as linguistic solutions to a single non-linguistic problem: the promotion of the speaker's interests. Wray provides an egoistic interpretation of even apparently altruistic efforts to choose forms that facilitate the hearer's understanding:

> In all of these manipulative expressions, it is in the speaker's interests to ensure that the hearer understands, since the intended effect of the utterance is to create a situation beneficial to the speaker.
>
> (Wray 2002:95)

Thus, according to Wray, the four subsidiary functions listed above all serve one primary function which is to further the interests of the speaker. This focus on language serving a non-linguistic, interpersonal aim recalls the earlier work of Halliday and Hasan (1989) where three functions of language are defined and described. These functions are (1) the ideational, (2) the interpersonal (3) and the textual functions. In their discussion of the functions, Halliday and Hasan (1989: 29, 48) also describe a fourth function, the logical function, which is a subsidiary function subsumed under the ideational function but that complexity is ignored in this discussion and the more straightforward tri-partite theory is presented.

Halliday and Hasan (1989) describe these functions as together forming a conceptual framework for viewing language from the outside, in non-linguistic terms. Although these functions provide a grid for interpreting the different ways in which people use language, Halliday and Hasan stress that their theory goes deeper than other functional theories which simply equate function with use.

> Function will be interpreted not as the use of language but as a fundamental property of language itself, something that is basic to the evolution of the semantic system.
>
> (Halliday and Hasan 1989: 17)

The ideational function for Halliday and Hasan is the learning or thinking function through which language users share their experiences. An examination of a sentence from the ideational perspective asks

> what it is about – its meaning as the expression of some kind of process, some event, action, state, or other phenomenal aspect of the real world to which it bears some kind of symbolic relation.
>
> (Halliday and Hasan 1989: 18)

The interpersonal function is when the speaker or writer does something through language. 'Every utterance has both an interpersonal and ideational component to it. It does something, and it is about something' (Halliday and Hasan 1989: 17). When language is considered from the interpersonal point of view, its function in the process of social interaction is emphasized and it is interpreted as a mode of doing. The interpersonal function refers to how the roles of speaker and listener, or reader and writer are managed. Epistemic and attitudinal stance, which are discussed in Section 1.14 below, contribute to the realization of the interpersonal function, as does modality.

The textual function is not a way of using language but a means of ensuring that what is said is relevant and relates to its context. This function covers thematic structure, the use of cohesive devices, discourse markers, the metalingual references contained in the text (e.g. *see above, in the diagram below, in the previous chapter*) and references to other texts. Some typical examples of the realization of this function are *according to___, as far as ___ is concerned, with regard to___, as for___.*

Halliday and Hasan's three macrofunctions correspond neatly with the three components of the *context of situation*, namely field, tenor and mode, which are discussed in Chapter 1.13 below. Figure 1.1 illustrates how each of these features of the context corresponds closely with one of the three functions.

Figure 1.1 Correspondences within Halliday's functional theory of language
(from Halliday and Hasan 1989:26)

SITUATION Feature of the context	**(realized by)**	**TEXT** Functional component of semantic system
Field of discourse (what is going on)	→	Ideational meanings
Tenor of discourse (who are taking part)	→	Interpersonal meanings
Mode of discourse (role assigned to language)	→	Textual meanings

A recent computer-assisted analysis of prefabricated language is to be found in the *Longman Grammar of Spoken and Written English* (Biber et al. 1999). This grammar compares written academic English and conversation contained in two corpora each of c. 5 million words and uses supporting written and spoken reference corpora. Analyses are also made of two other registers, viz. fiction and press reportage, using similar sized corpora. Biber et al. (1999) use a purely empirical and statistical approach and count any sequences of three or more words which occur more than a certain number of times across a certain range of texts. They coin the term 'lexical bundles' to describe these recurrent phrases and these could be viewed as trigrams and N-grams (n>3) which recur across at least five different texts with a frequency of more than:

N=3 >10 times per million
n=4 >10 times per million
n=5 >5 times per million
n=5+ >5 times per million

Lexical bundles are selected purely on the basis of their frequency of occurrence in the corpus. These word sequences, e.g. *at the end of the,* are picked out by a computer search for frequently recurrent word clusters. Biber et al. (1999) provide lists of these bundles to users of their grammar. Although this work by Biber and his colleagues reveals a great deal of hitherto unsuspected information about the patternings of different genres of written and spoken English, I have some misgivings about how learners might be able to use bundles in their writing. Questions remain as to whether these bundles could be learned, internalized and recalled by language learners, given their typically truncated or fragmented appearance. For this reason, I will not classify recurrent sequences ending in *a, an* or *the* as prefabs.

The Longman Grammar of Spoken and Written English Grammar (Biber et al. 1999) having identified the recurrent word sequences through their statistical frequency, categorizes them using structural and grammatical features. The frequency of occurrence of lexical bundles in different registers is compared. Later, Cortes (2004) designed a taxonomy for classifying bundles based on their function. This taxonomy was developed further in Biber, Conrad and Cortes (2004) and is outlined in Figure 1.2.

Figure 1.2 Taxonomy classifying lexical bundles by function (based on Cortes 2004)

FUNCTION	DESCRIPTION	EXAMPLE
referential bundles	time, place, or text markers	*at the beginning of* *the end of the* *at the same time*
text organizers	word combinations used to express comparison, contrast, inference or focus.	*on the other hand,* *as a result of* *it is important that*
stance bundles	express attitudes that frame some other preposition	*I don't know why* *are more likely to*
interactional bundles	usually conversational word combinations to express politeness or to report	*thank you very much* *I said to him*

Similarities are apparent between the kinds of bundles in this taxonomy and the Halliday and Hasan model of language function. Referential bundles could be subsumed

under the ideational meanings. Stance and interactional bundles could be classified under interpersonal meanings and text organizers would realize textual meanings. This close correspondence provides a degree of reassurance that substantive features of the written language are being captured by both approaches.

1.4 Phraseology

This section describes the discipline which has in recent years adopted the name phraseology, thus aligning English-speaking scholars with those Russian and Eastern European scholars (e.g. Arnold 1973; Gläser 1988) who have used this term for many years. Gläser (1988:265) gives this definition of phraseology: 'the linguistic description of set expressions whose meaning cannot be derived from the meaning of their parts'.This definition is, however, unduly restrictive as it limits phraseology to the study of idioms and restricted collocations. Cowie (1998:19) claims that 'studies of collocations have pushed the boundary that roughly demarcates the "phraseological" more and more into the zone formerly thought of as "free"'.

Fox (1998) observes that, in language teaching and learning, undue emphasis on the meaning of words can lead to an unrealistic view of how meaning is created in text and may also produce writing by language learners which bears little resemblance to the writing produced by native speakers. Wray (2002:209) suggests that L2 learners, by concentrating on word-by-word output, may attempt to assemble too many meaningful words or may insert too much meaning in each slot on the syntagmatic chain. One of the facts reported by researchers on large corpora (Sinclair 1991; Fox 1998) is that in most kinds of written (and spoken) texts the most frequent words, and the most frequent senses of the less frequent words, tend to have less of a clear and independent meaning than less frequent words or senses. A good example is provided by the delexicalized verbs *have, make, give,* and *take,* which are often best explained in terms of the words they collocate with. As the following pairs of examples for each verb show, the verb could be removed with little or no loss of meaning: *have a look, have a holiday; make a report, make war; give a lecture; give a sigh; take shelter; take note.*

It is self-evident that frequent words and senses of words feature prominently in most kinds of texts. Willis (1990:47) observes that the 700 most frequent words in a corpus of written English account for over 70% of its size. The occurrence in texts of these (usually delexicalized) words creates redundancy, which allows the reader to concentrate on the content of writing rather than the manner in which it is written. According to Fox (1998):

> It is certainly true that there are times when you do want your audience to be impressed by your use of language...But most of the time you don't want that. You just want to get your message across. And you do that by being unremarkable in your language, by being conventional and predictable, and – dare I say it – boring!
>
> (Fox 1998:28)

Thus when the perspective changes from the level of individual word choice to the level where words collocate, there is evidence of specialization of meaning and arbitrary restrictions on co-occurrence. Phraseologists ascertain the amount of restriction on co-occurrence by measuring the 'commutability' of the expression. Commutability is determined by the degree to which the component words of a prefab can be replaced by synonyms without loss of meaning or idiomaticity.

Collocation

One of the earliest uses of the notion of 'collocation' was by Palmer (1933:1) in relation to the teaching of English as a Foreign Language (EFL): 'A collocation is a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts'. In the work of Firth (1957), collocation becomes a 'mode of meaning'. Just as the light of mixed wavelengths disperses into a spectrum, 'the lexical meaning of any given word is achieved by multiple statements of meaning at different levels', e.g. the orthographic level, the phonological level, the grammatical level, and the collocational level (Firth 1957:192). Firth adds formal and etymological meaning together with social indications of usage. He distinguishes contextual meaning from meaning by collocation and uses collocation as a technique for the stylistic criticism of literary works.

Firth's student, McIntosh (Halliday, McIntosh and Strevens 1964) saw collocation as a useful indicator of register. The collocation of *free* with *kick* in a text would suggest the

register of sports journalism. McIntosh thought that, sometimes, even the presence of one word can make us fairly sure of the field: *cleanse* he thought would confine the text to writing about cosmetics, detergents or possibly religion. (History has, ironically, produced a much more sinister use of the word to refer to events in South East Europe.)

McIntosh's student, Sinclair, defined the term collocation thus: 'collocation is the occurrence of two or more words within a short space of each other in a text' (Sinclair 1991:1). For Sinclair the term includes only the lexical co-occurrence of words but some scholars conjoin in their definition of collocation such lexical patterning with grammatical choice as well. Kjellmer (1984:163) gives the following definition of collocation: 'lexically determined and grammatically restricted sequence of words'. According to this definition, only recurring sequences that are grammatically well-formed can be considered as collocations. Kjellmer also attempts to elaborate a set of rules for assessing 'collocational distinctiveness'. Briefly, these rules classify a sequence as highly distinctive if it appears frequently in many and different categories of texts; if it is long (the minimum length is two words and the longer the more distinctive); and it is structurally complex.

There appears to be a reasonable degree of agreement among the writers discussed above (Palmer, Firth, McIntosh and Sinclair) about the definition of collocation. To describe a combination of words as a collocation refers to their physical proximity in texts. What is the relationship between collocations and prefabs? Kjellmer's definition of collocation moves towards the restrictedness and arbitrariness more characteristic of prefabs. Prefabs could be characterized as the subset of collocations which are recognizable, memorable and have some degree of idiomaticity.

1.5 The case of idioms

Although some writers, e.g. Gläser (1998), give prominence to idioms, they make up only a small part of the total number of phraseological units or prefabs in the language (cf. Fillmore 1978; Howarth 1998; Biber et al. 1999). Moon (1998) reports that 70% of the idioms in her corpus of 18 million words of written English, *HECTOR*, occurred with a frequency of less than once per million words of corpus text while 51% occurred with such

low frequencies that their presence or absence was no better than could be expected due to random chance and was, therefore, statistically insignificant. Another 37% of Moon's 4,000 or so idioms, although statistically significant, still had frequencies of less than one per million and only 11% had frequencies between 1 and 5 per million words. The remaining 1% of idioms occurred between 5 and 50 times per million.

Despite the fact that idioms occur most infrequently in written corpora (e.g. Moon 1998), and indeed are not used particularly frequently in spoken language production (cf. Biber et al.1999)—see below— they are nevertheless important in the present study for several reasons. Firstly, they form one end of a continuum stretching from pure idioms at one end (Figure 1.3) with free combinations at the other (Howarth 1998). (Note that the adjective 'free' is not used in Figure 1.3 in an absolute sense, as words are constrained by their word class, syntax, selectional restrictions and logical and cultural constraints).

Identifying idioms can therefore help isolate prefabs. Howarth's taxonomy could be used to produce an *extensive* definition of prefab, i.e. a definition that 'simply enumerates all the items that may be named by the word. Prefabs include word combinations that lie along the spectrum from pure idiom, through figurative idioms and literal idioms as far as restricted collocations but excluding free combinations.

Figure 1.3 Continuum of freedom –restrictedness in word combinations (after Howarth 1998)

Pure idioms	**Figurative idioms**	**Literal idioms**	**Restricted collocations**	**free combinations**
let the cat out of the bag	move the goalposts	on foot	explode a myth/theory/ notion/idea/belief	read a book
make off with	drop names	one day	wage freeze	
	catch fire	in the meantime		

Secondly, idioms are important because they have been very carefully studied and catalogued by many scholars (e.g. Makkai 1972, Fernando 1996), and are perhaps among the most closely analysed categories of prefabs. Examination of these studies of idioms provides guidance on methods of identifying and classifying prefabs. Thirdly, idioms are salient for advanced language learners, as witness the number of Internet sites which

provide practice in the comprehension and use of idioms. Perhaps learners see idioms as important repositories of information about the culture(s) of the target language.

Howarth (1998), however, in introducing phraseology, cautions that perhaps undue emphasis had been placed on idioms in phraseological studies and in pedagogical application of those studies. Howarth uses 'idiom' in a narrower sense than is often used by some writers in the British tradition. He agrees with Gläser's definition of 'idiom': 'an idiom is a lexicalised, reproducible word group which has semantic and syntactic stability and whose meaning cannot be derived from the meanings of its constituents' (Gläser 1988:266). Idioms in this sense (e.g. *armed to the teeth*, *bite the bullet*, *cut and dried*, *face the music* and *let the cat out of the bag*) are surprisingly infrequent in corpora of written and spoken English with average occurrences of less than once per million words (Moon 1998). In EAP texts such idioms will have even less prominence, as their function, it could be argued, is to establish a degree of intimacy between writer and reader untypical of most academic written genres. Howarth defines idioms as those phrases which are relatively frozen and semantically opaque (Howarth 1995:19, 24). He warns that it would be wrong to suggest that formulaic or prefabricated language is a single category which can be contrasted with compositional language or language generated by rules. Instead, he posits four properties of prefabricated language: (1) the formulaic or conventional nature of expressions; (2) memorization, a psycholinguistic feature; (3) lexicalisation: when a multi-word unit is stored and processed unanalysed as if it were a simple lexical item; and (4) fixedness, which refers to the degree of flexibility in relation to the substitution of synonyms or the word order.

Howarth suggests that the above four properties are gradable (1998:26) and advocates the adoption of a continuum model. Notwithstanding his criticism of the previous over-emphasis on idioms, he feels the need, in his close examination of fixed expressions and composites, to deal with idioms both for completeness and also because they form one end of a continuum which ranges from completely frozen to unrestricted free collocations.

Howarth (1998:26) examines the literature on phraseology and separates out two main kinds of idioms. Pure idioms he defines as those which have 'a unitary meaning that cannot be derived from the meanings of the components and are the most opaque and fixed category' (*blow the gaff, under the weather*); while figurative idioms have metaphorical meanings in terms of the whole and have a current literal interpretation (*blow your own*

trumpet, under the microscope). Howarth (1995 and 1998) makes a bold claim for an *a priori* programme of phraseological study, which he believes should precede any statistical measure of the distribution of composites in texts. As shown in Figure 1.4, he divides 'composites' (his favourite term for fixed expressions) into a collocational continuum ranging from free to frozen or fixed.

Figure 1.4 The free-restricted continuum for verb–noun and preposition–verb combinations (adapted from Howarth 1998:28)

	Free combinations	**Restricted collocations**	**Figurative idioms**	**Pure idioms**
Lexical composites verb+noun	*blow a trumpet*	*blow a fuse*	*blow your own trumpet*	*blow the gaff*
Grammatical composites preposition+noun	*under the table*	*under attack*	*under the microscope*	*under the weather*

In a previous work, Howarth (1995) developed just such a complex of features and used the criterion of commutability to subdivide the pivotal category of *restricted collocations* above into various levels of 'restrictedness.' He concentrated on verb + noun combinations. The result was another continuum, this time with five categories, ranging from free or unrestricted combinations through varying degrees of admissable substitution of verb or noun. The limit point at the restricted end of the spectrum, where no substitution of verb or noun is permitted (e.g. *curry favour*), is the category of idioms. In his research on academic texts, Howarth found that main verbs revealed some degree of restrictedness in 30-37% of all the occurrences of the verbs studied. This research reveals key characteristics of academic prose with implications for approaches to EAP writing. Howarth's work provides a robust system for the categorization of combinations. He meticulously maps out the area of phraseology, while his insistence on small-scale and sometimes, 'manual' text analysis contrasts with trends in corpus linguistics towards ever larger corpora.

Gläser's concentric model (1988) for English phraseology is useful because it sets up a core type of expression, which she calls a 'nominative' and that can be used as a basis for searching corpora. This core circle of nominatives is surrounded by two concentric rings, the first with fragments or reductions of propositions (e.g. stereotyped comparisons,

irreversible binomials), and the outer circle with propositions (e.g. commonplaces, slogans, proverbs). She defines phraseological units in nominative function (the inner circle) thus:

> Word-like units which designate phenomena, objects, events, processes, actions, states, qualities, relations, etc. in the outside world, and a few word groups which only function as operators in that they designate relations between phenomena or objects. Idiomatized prepositions (by virtue of, by dint of) and conjunctions (in order to, on condition that) may serve as examples.
>
> (Gläser 1988:273)

Gläser's nominatives include many expressions which are features of academic English, as illustrated by the idiomatized prepositions and conjunctions in this definition. This recalls Hudson's (1998) typology of fixed expressions (see Section 1.2), where she discusses compound prepositions and compound connectives. There is also a close correspondence between nominatives and exponents of the ideational function of language, which is discussed above in Section 1.3.

1.6 The recent history of chunks

The suggestion that the meaning of words might reside in the company that they are found in was put forward by Firth (1957) but a perusal of the history of language teaching pedagogy shows that similar theories had been in circulation in previous centuries. A notable example is to be found in the voluminous works of Comenius, especially his magnum opus *The Great Didactic* (1657), and the revolutionary language teaching textbook, *Orbis Sensualium Pictus* (1658), whose influence is still felt today by language teaching methodologists. In the former, Comenius insists that the teaching of words in mother tongue, foreign language and in classical language instruction should be through the presentation of illustrative sentences. In the latter, words to be taught are contextualized in exemplificatory texts.

The idea that words are stored in the mental lexicon in various arrangements and are retrievable as 'atoms' or as 'molecules', so to speak, has been circulating since at least the 1960s. The idea of chunking itself came from psychology. Miller (1956) observed that

many psychological phenomena pointed to the existence of a short-term memory with a limited capacity for holding information. Subjects seemed to be able to recall seven-digit telephone numbers but if they chunked the digits into higher order groupings (of, say, five), they could recall many more digits or binary number sequences (nearly forty). But the number of chunks remains roughly constant at seven items (plus or minus two).

This idea of chunking was applied to many complex cognitive and motor skills, such as learning to type, playing chess or playing a musical instrument. For example, what distinguishes a novice from a chess grandmaster is the latter's ability to 'chunk' the board into a few zones of influence or meaningful relationships between pieces, whereas the tyro sees all the pieces and pawns and their positions individually. Lerdahl and Jackendoff (1988) apply chunking to classical musicians or jazz virtuosi when they are playing fast. They suggest that when such musicians play intricate or difficult pieces

> larger and larger passages form simplex units from the point of view of awareness - to chunk the input and output. This suggests that processing speed is linked not so much to the gross measure of information processed as to the number of highest-level units that must be treated serially.
>
> (Lerdahl and Jackendoff 1988:125)

Lord (1960) suggests that oral epics as told by storytellers in cultures with oral traditions are not memorized, but recreated at each telling by concatenating formulaic expressions using a familiar plot as a structuring device. These studies of mnemonic techniques for recounting as accurately as possible tribally significant stories point to earlier strategies deployed by communities to preserve and transmit their culture. It was an attempt to make spoken language more permanent in the way that written language would later revolutionize the encapsulation of knowledge. The fact that this resource within oracy is still available to modern literate people suggests that it still has its utility. Although Lord's model is related to the way in which speakers produce longer narrative turns, support can be found for the application of a similar model to language acquisition (Skehan 1998) and language production (Pawley and Syder 1983).

Cohen, although referring to sentences, gives an early statement of a model of prefabrication in language production:

> The sentences of a language, not the words, are like tools in being either the stock means to certain frequently desired ends or the ad hoc means, specially constructed, to ends that may or may not be so frequently desired.
>
> (Cohen 1962:74)

Another early proponent of this chunk-driven model of language performance was Becker, a worker in Artificial Intelligence:

> Utterances are composed by the recitation, modification, concatenation and interdigitalization of previously known phrases consisting of more than one word. I suspect that we speak mostly by stitching together swatches of text that we have heard before.
>
> (Becker 1975:3081)

Becker estimates that we know 25,000 stock phrases and more than a hundred similes (e.g. *as pleased as punch*). Bolinger (1977) hypothesizes that memory plays an extremely important role in speech production. Indeed he suggests that our capacity to remember seems to be almost unlimited. Mel'čuk (1998) suggests that 'phrasemes' or set phrases outnumber words in the lexicon of any language by roughly ten to one. An interesting commentary on this question is found in Fillmore (1978:149):

> We have to face the inconvenient reality that a number of expressions in any language have to be viewed both as lexical items and as entities having grammatical structure on a level higher than that of word-formation. I have in mind, not only the much discussed 'idioms', which probably make up a small proportion of the total 'phrasicon,' but also the vast repertory of fixed phrases, clichés, speech formulas-in general, all conventionalized ways of saying things-that a speaker acquires independently of the process of learning the grammatical rules of the language.

Wong-Fillmore (1979) studies the phenomenon from the point of view of children's second language acquisition and describes chunks as 'formulaic frames with analysed slots'. Peters (1983), investigating first language acquisition, claims that ordinary conversation consists almost entirely of 'institutionalized clauses' which, unlike idioms, can be analysed according to the normal rules of syntax but which are stored in memory because of their usefulness or frequency in conversation. These clauses are stored and retrieved as single units. Peters suggests that, considering these institutionalized bits of language, any hard and fast division between lexis and syntax seems to be blurred; she

posits instead a dynamic and fluid continuum from the completely fixed to the completely original.

Nattinger (1988:89), following Bolinger, suggests that language in the form of memorized or routinized phrases is thought to be 'stored redundantly not only as morphemes, words, parts of phrases or even as longer memorized chunks of speech, and ... retrieved from memory as ... pre-assembled chunks'. Nattinger discusses the advantages of such language performance:

> This prefabricated speech has both the advantage of more efficient retrieval, and of permitting speakers to direct attention to the larger structure of the discourse, rather than keeping it focused narrowly on individual words as they are produced.
>
> (Nattinger 1988:90)

Nattinger regards language use as a 'compositional' process, as, basically a 'stitching together' of these pre-assembled phrases into discourse. In later publications he and his collaborators assess the implications for applied linguistics of this basic tenet. Nattinger and DeCarrico (1992) put forward a typology of 'lexical phrases' divided into six categories. This typology has been criticised (Howarth 1995) for using a combination of structural and pragmatic criteria for distinguishing between these six categories, and indeed some are a mixture of functional and structural types: e.g. 'polyword' which has vague distinguishing criteria such as 'short phrases' and 'no variability' (Nattinger and DeCarrico 1992:38). Their book is, nonetheless, influential among language teachers. This is partly because of their non-technical writing style, but also because its Firthian analysis of fixed expressions applied to the classroom gives teachers a challenging new perspective on conventionalized language.

A recurrent theme in discussions of prefabs is their 'memorizability'. This quality of being relatively easy to memorize underlies observations about prefabs being 'previously known phrases' (Becker 1975), being 'useful' (Peters 1983) and being conducive to 'more efficient retrieval' (Nattinger 1988). In order to conceptualize these processes of memorizing, retrieving and concatenating prefabs, it is necessary to examine some of the psycholinguistic models which have been proposed.

1.7 The role of habitually used expressions in language production

According to Bolinger (1975:7) there is a 'greater degree of unfreedom in every syntactic combination that is not random' than allowed for by generative linguists. There seems to be a level above the lexical, where the choice of one word constrains the words which might co-occur with it. These constraints seem to be arbitrary and conventionalised. One of the earliest and most influential accounts of this 'unfreedom' is found in the work of Pawley and Syder (1983) in their book chapter, 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency'. They call into question the Chomskyan notion of creativity in the generation of grammatical sentences. For Pawley and Syder, only a small subset of the innumerable or indefinitely large number of sentences which would be generated by a Chomskyan grammar are nativelike in form – in the sense of being acceptable to native informants as ordinary, natural forms of expression, in contrast to expressions that are grammatical but are judged to be 'unidiomatic', 'odd' or 'foreignisms.'

Native speakers, when communicating in real time, have access to a rich collection of prefabs which are 'units of clause length or longer whose grammatical form and lexical content is wholly or largely fixed; [their] fixed elements form a standard label for a culturally recognised concept' (Pawley and Syder 1983:191). These stems differ from idioms and, according to Pawley and Syder, number hundreds of thousands. Here are a few examples which they claim to have introspected after 'a few minutes' reflection' (pp. 206-7):

Clauses	**Sequences longer than a simple clause**
Is everything OK?	I'll believe it when I see it.
You would ask that question.	It just goes to show, you can't be too careful.
There's no pleasing some people.	There's nothing you can do about it now.

Pawley and Syder (1983:191) suggest these native-like locutions are stored in the long-term memory. A startling demonstration they give of the enormous number that there must be of such stored expressions is a list of habitually used expressions in which the verb

think occurs. Their list, which they describe as partial, has 47 expressions. Here are the first five from the list:

> *Come to think of it...*
> *What do you think?*
> *I thought better of it.*
> *Think nothing of it.*
> *Think it over.*
>
> (Pawley and Syder 1983:191)

Reading through the lists contained in this work occasionally reminded me of the taunts, gibes and other incantatory language of my primary school days. It would seem that a rich stock of these expressions is built up in childhood and many remain part of a speaker's active repertoire into adulthood. For Pawley and Syder, learning the habitually used expressions of a language is an integral part of acculturation, given that every culture has its preferred expressions or expressions upon which they confer idiomaticity. It should be noted that these lexicalized sentence stems are 'units of clause length or longer' (Pawley and Syder 1983:191). Other investigators, e.g. Fillmore et al. (1988:504) do not consider idiomaticity to be a feature of language that is only found at clause or sentence level. Many opaque idioms (i.e. idioms that cannot be comprehended by means of knowledge of vocabulary and grammar alone) occur as short phrases, e.g. *red tape* or *red herring.* Also transparent idiomatic expressions, such as *answer the door* (possible to comprehend but whose status as conventionalized is not given by the grammar and vocabulary) exist at the phrase level. In fact, the large majority of the prefabs investigated in this thesis exist at the phrase level and, when removed from their co-text, retain their idiomaticity.

A problematic feature of these habitually used expressions is that they are known twice: both as lexical units and as products of syntactic rules. This poses problems for those who, in evaluating a description or theory, choose according to principles of economy and generality. Pawley and Syder agree with their mentor, George Grace, in a footnote, that there has been an 'uncritical acceptance of Occam's Razor as providing a satisfactory way of choosing between competing theories, not only in linguistics but in Western science generally since Newton' (Pawley and Syder 1983: 224 fn. 21). They continue: 'The

question arises as to what parsimony has to do with the organisation of the speaker's knowledge' (1983:224).

There could be what Aitchison (1987) calls an 'embarrassment of riches' within the mental lexicon. Speakers may have multiple access through words and meanings and several paths to stored phrases including the ability to compose them syntactically as nonce forms. Gläser's (1988:275) observation about some idioms having a literal meaning e.g. *to roll out the red carpet* or the term *wet blanket* is apposite here.

Although Pawley and Syder's ideas were not entirely new, they were cogently and forcefully expounded and this publication has played a fulcral role in the literature of fixed expressions. They draw together work by Weinreich (1969), Becker (1975), Bolinger (1977) and Nattinger (1980) and posit a challenging hypothesis to confront the Chomskyan model of language production. According to the transformational generative model propounded by Chomsky and accepted widely, the internalized competence of a native speaker of a language can generate an infinite number of sentences most of which have never been said before, as demonstrated by young L1 speakers. Pawley and Syder's claim is that this vast potential might, in reality, be extremely circumscribed within the constraints of the phraseology of the language. They would not deny that we can produce an endless amount of novel utterances but, they suggested, many of these utterances would be questioned as unidiomatic. Their views are later echoed in work by various thinkers in applied linguistics. The editors of a festschrift to H. G. Widdowson (Cook and Seidlhofer 1995: vi) comment in their preface on the pivotal role of Pawley and Syder's work.

1.8 Chunking in written language

Most of the research carried out on chunking, and reported in previous sections, has viewed the phenomenon of chunking or pre-assembled phrases as primarily a characteristic of spoken language. It should be remembered that Pawley and Syder's findings concerned only spoken language. Until recent years, there have been fewer investigations of prefab use in written English. The present work focuses entirely on prefabs occurring in written language. A theoretical approach is needed, which would support an empirical investigation of the prefabs in use in academic texts.

In recent years, several important lexicographers have proposed models of language production which take into account the recurrence of lexicalised phrases and it is interesting that these models seem to be equally applicable to spoken and written language. These models view language production as the interplay between stable and creative aspects of language use (Cowie 1988) or between the choice principle and the idiom principle (Sinclair 1991), as discussed below in Section 1.10. Speakers and writers of a language draw on a very large repertoire of pre-assembled patterns of prefabricated language in most communicative situations and do not need to construct such phrases anew on each occasion (although the capacity to do so is always in place). One psycholinguistic explanation is that speakers and writers resort to memory first and only compute if that fails them. This allusion to memory affords us one possible distinguishing mark of prefabs: what some writers call *lexicalization* (Pawley and Syder 1983) and others call *conventionalization* (Howarth 1998). Both concepts suggest a socially shared stock of words and phrases belonging to *langue* or the 'storehouse filled by the members of a given community through their active use of speaking, a grammatical system that has a potential existence...in the brains of a group of individuals' (Saussure 1915:13-14). In keeping with my interest in the contents of the mental lexicon, I shall call this process *memorization*. As Pawley and Syder (1983) point out, not everyone memorizes the same chunks, so this concept relates to the individual and is open to psycholinguistic study. As my experimental data exemplifies the interlanguage of language learners, the individual variability implicit in the psycholinguistic concept makes it more suitable for my purposes.

It becomes clear on examining the data obtained from the written work of Portuguese advanced learners of English, which is introduced in Chapter 3, that individual students find different sequences memorable or salient. There are many more ways to deviate from written norms than ways of conforming. On the road to acquisition of a second language, the temptation to lapse into fossilization is great. Many of the Portuguese writers whose argumentative essays are examined in Chapter 3 have developed a dependence on certain prefabs which they tend to overuse to the detriment of other expressions which their linguistic competence would enable them to generate. I recognize this overuse of favourite prefabs from my own language learning experience, but recall the pleasure of discovering alternatives and using them. This thesis will provide a better description of the prefabs actually used by Portuguese advanced learners. As a result, EAP

teachers can assist their students in the elegant variation of overused prefabs and in appreciating the usefulness of underused prefabs.

Prefabs generally start off their diachronic life-cycle, which often spans centuries, as coinings which somehow catch the ear of a listener or the eye of a reader and are repeated and thus spread through the speech community. With the advent of the mass media, the potential for the spread of expressions has been greatly increased but the older technology, the printing press and the disseminative power of word of mouth should not be underestimated. Such a model does not preclude the same prefab being coined independently at different times and places. Nonetheless the resultant confluence of such separate mintings might expose a greater number of language users to the candidate prefab helping it to win out against competing forms to become the accepted or idiomatic way of saying something. This pre-eminence should then be revealed in any sizeable corpus of the language, although even very large corpora can fail to capture certain well-established prefabs. For example, Barkema (1993:271) notes that the following collocations commonly listed in dictionaries did not occur in the *Birmingham Corpus,* the 20-million-word predecessor of the *COBUILD* corpus: *baker's dozen, black frost, breach of promise, complementary colours, fortified wine, compassionate leave, false pride.*

Exactly how a prefab catches on might be explainable in terms of the concept of 'meme' invented by Dawkins (2000) and applied in psychology by Blackmore (1999). Finally, it is worth recalling Hymes's (1968) acute observation that formulaic language is not always intended for a wide circulation:

> Some sequences become idiomatic for a person or group because of a memorable novelty..., but more because sensed as appropriate or as needed. Most do not achieve generality or persistence, but some would lose value if they did, being intended or enjoyed as distinctive, or private to a few.
>
> (Hymes 1968:127-128)

Howarth (1998:28) observes that the significance of composites is regarded as psychological; their degree of restrictedness relates to mental storage and processing. Significance is therefore gradable and the result of a complex of features rather than simply a statistical measure. Prefabrication might be what makes languages learnable and transmittable to new generations. Those nonce forms that survive, that get transferred into

the common stock, might do so because they are more memorable than the other competing nonce forms which pass into oblivion. This line of argument might incur the criticism that it is *post hoc* reasoning. In the terminology of modern popular science, the phrase which catches on is a successful meme (Blackmore 1999). The rhyme, alliteration and assonance found in many proverbs come to mind (Gläser 1988:275). The transformation of some novel expressions into conventional ones may be much more complex than has been suggested here. Indeed, it could be the result of chance: one way of saying something prevails over other competing ways because it gets said and written by people and in circumstances more propitious for propagation.

Some writers take a fairly quantitative view of lexicalization and individual memorization and calculate how many times in a lifetime we might hear certain phrases according to their frequency in large reference corpora. Hoffman and Lehmann (2000) suggest that the BNC would take 4 years to read at 8 hours a day and in their estimation represents roughly 10 years of linguistic experience for the average speaker in terms of quantity. They wonder how native speakers and learners can memorize prefabs that they hear only 5 times a year. The interaction between memory and received input might be more complex. Perhaps native speakers are ready for certain collocations because they embody the spirit of the language and experience a kind of collective dejà vu with certain strings of words. Successful pop songs, jingles and slogans might succeed because they tap into this expectancy we have.

The huge difference in exposure to the target language between the typical L1 learner and the NNS learner places a great deal of pressure on the designers of language learning syllabuses. A way must be found to focus on those features of the language which are most typical and generalizable. Current orthodoxy in EAP recommends using the limited classroom contact hours to equip students with useful strategies for autonomous learning. The importance of 'noticing' for learning is strongly emphasized. Speakers/writers tend to produce utterances containing prefabs which may or may not be variable. For example, the noun in a prefab such as *in fact* can be modified to produce *in actual fact,* or *a bargain* can be *made, reached, arrived at* or *struck.* The first task for the language learner is to notice such prefabs and be able to recognize, or even produce them on later occasions. If all prefabs were completely fixed expressions, this learning task would be a sufficiently onerous one but, because many prefabs can be varied, the learner

often needs to use fuzzier categories when deciding on the fixedness of expressions. The most difficult types of prefabs to master are the numerous expressions which are established and partially variable. The challenge for teachers and students is the all-pervasive patterning in English and the large number of *almost* fixed expressions.

Kennedy (2003:478) remarks that 'learning to associate forms with forms, forms with semantic or pragmatic functions, and forms and functions with contexts requires huge amounts of exposure'. Work by Kirsner (1994) stresses the importance of implicit knowledge in L1 acquisition and suggests that the amount of exposure to and practice of lexical items which L1 learners undergo may have been greatly underestimated. Young L1 learners, according to Kirsner, need first to acquire a core lexicon of words and routines before they can acquire the L1. Similarly, in SLA, if fluency is to be achieved in speaking and writing, this depends on implicit learning and retrieval skills. If it is assumed that 'the L1 learner may receive 30,000-40,000 hours of exposure to and practice in using the language by the age of about 12 years' (Kennedy 2003:481), the question for EAP teachers is how to recreate similar implicit learning conditions for their students. Kennedy (2003) reminds us that

> Ensuring that language learners get frequent opportunities for internalizing prefabricated word groups is not the only task of the language teacher, but surely one of the most neglected... Because frequency of experience significantly affects learning, the provision of systematic, repeated exposure to collocations in meaningful contexts lies at the heart of the teaching experience.
>
> (Kennedy 2003:481)

1.9 Using corpora to model the mental lexicon

Caution is needed when a corpus is used to somehow represent the lexical experience of an individual language user. A large corpus like the BNC purports to sample spoken and written British English of the 1990s in all its breadth. Hoffman and Lehman (2000) have rendered an excellent service to the corpus linguistics community at the University of Zurich by making the BNC easier to consult online but when they use it as a measure of individual exposure to various words and phrases they are straining the analogy and using

the corpus for a purpose it cannot serve and for which it was not designed. As Wray (2002:27) observes:

> Only relatively few people regularly read both tabloid and broadsheet newspapers and listen to both pop quizzes and heavy current affairs programmes on the radio - the sort of data that are thrown together in a corpus.

A different angle on this question can be found in Hunston (2002), where she discusses the notion of 'mental concordance' first suggested in a conversation by Michael Hoey (cf. Hoey 2005:11-14). The following quotation from Hunston (2002:31) contains what appears to be a blend of Bakhtin's (1929) notion of intertextuality and Saussure's concept of *langue* as a collective storehouse in the brains of a group of individuals:

> The meanings in the text are dependent on the typical behaviours of sequences of words. This is, in a sense, a manifestation of intertextuality: the meaning of this one text depends on thousands of other texts, and the repeated patterns that are found in them. The instance depends on the whole, but the instance also influences the whole.
>
> (Hunston 2002:31)

According to the many versions of his theory, which Chomsky has developed over the years, a person competent in a language is capable of producing, and recognizing as grammatical, an infinite number of novel utterances by applying powerful rules to the lexicon. This notion of creativity was propounded by Chomsky as part of his attack on the behaviourist theory of language of Skinner. Chomsky (1965) pointed out the impossibility of young children's inductively discovering the rules of their mother tongue on the basis of the data they receive from their parents and in the short period of time they have to learn their mother tongue. A more rigorous version of this observation is stated in Gold's (1967) theorem. On closer examination, the creativity of all native speakers of all languages becomes less startling. It is usually only in songs, rhymes, poetry or experimental fiction that examples of syntactic innovation are found. Most novel utterances recycle the same structures with a rich variety of lexical and phraseological strings inserted in the same old phrase structures.

In a seminar, McEnery (1998) mentioned that he and his colleagues, Baker and Wilson, were investigating how people choose or store collocations. They postulated that if people consciously or systematically choose collocates for words they wish to use, then the underlying system should be regarded as an important part of linguistic description. Pawley and Syder (1983) had already suggested a separate component of language, a 'phrase book with grammatical notes' intermediate between lexis and grammar to deal with:

> the large body of institutionalised complex lexical forms and the semi-productive rules for generating new, native-like sequences by inflecting, expanding or transforming these forms...But any compartmentalization would not truly reflect the native speaker's grammatical knowledge if the facts are (as we believe) that lexicalization and productivity are each matters of degree.
>
> (Pawley and Syder 1983:219-20)

An important way to view prefabs and to classify them is to look at the function(s) they serve in speech and writing. This perspective on prefabs was examined in depth in Section 1.3 of the present chapter.

1.10 Prefabs in writing

Several of the theories which describe the use of prefabs in language production appear to be applicable to prefabs in both writing and speaking. For example, the wording of Widdowson's (1989) definition of communicative competence can be interpreted as referring to either written or spoken language:

> ...communicative competence is not a matter of knowing rules for the composition of sentences. It is much more a matter of knowing a stock of partially pre-assembled patterns, formulaic frameworks, a kit of rules, so to speak, and being able to apply the rules to make whatever adjustments are necessary according to contextual demands. Communicative competence in this view is essentially a matter of adaptation, and rules are not generative but regulative and subservient.
>
> (Widdowson 1989:135)

Sinclair (1991), in postulating his two models of interpretation of language, the free choice principle and the principle of idiom, is also careful to refer to *language users* and mentions *listening*, *speaking*, *reading and writing* in his explanation of the models, thus not ruling out the application of his two principles to writing. He postulated these principles after many years of research on text corpora of ever increasing size, which consisted mainly of written texts. According to Sinclair, the two principles are diametrically opposed but are both needed to explain language performance. This is how Sinclair defines the open-choice principle:

> It is often called a 'slot and filler' model, envisaging texts as a series of slots which have to be filled from a lexicon which satisfies local constraints. At each slot virtually any word can occur. Since language is believed to operate simultaneously on several levels, there is a very complex pattern of choices in progress at any one moment, but the underlying principle is simple enough.
>
> (Sinclair 1991:109)

In Sinclair's open-choice principle the only constraint is grammaticalness. In this way, it resembles the Saussurian model of language or *langue* as a system of arbitrary signs. The second and more radical principle Sinclair defines thus:

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments
>
> (Sinclair 1991:110)

Thus at a point in a text where the open-choice principle would admit a wide range of possible choices, the 'idiom principle' can limit this choice dramatically. A single choice in a slot can dictate what comes in the contiguous slot or slots, it can severely reduce the choice or it can even prevent the exercise of choice. For Sinclair the dominant or default principle is the idiom principle and the open-choice principle is only activated when the idiom principle is unable to deliver the flow of language perhaps because of lexical choices which are difficult in their environment. Sinclair suggests that the interpretative process switches to the open-choice principle in these or similar circumstances and then quickly switches back again (Sinclair 1991).

Sinclair's (1991) model of language production with its two alternating principles, the choice and the idiom principles, might be simplified to a unitary model where idioms

and words, which are stored together in the mental lexicon, are combined in language production. The question arises as to how the syntactic structure of the stored multi-word items is accessed. This is particularly important in the case of syntactically variable prefabs, e.g. *be in sb's nature, sth is not in sb's nature, it is not in the nature of sb to DO/BE STH,* which Sinclair recognizes as alternations of the same stored item. Such a unitary model might provide a better framework for understanding the way in which stored multi-word form-meaning pairings are integrated into language production and perception. Langacker (1987:46) describes just such a framework:

> The grammar lists the full set of particular statements representing a speaker's grasp of linguistic convention, including those subsumed by general statements. Rather than thinking them an embarrassment, cognitive grammarians regard statements as the matrix from which general statements (rules) are extracted. Speakers do not necessarily forget the forms they already know once the rules are extracted, nor does the rule preclude their learning additional forms as established units. Consequently, particular statements (specific forms) coexist with general statements (rules accounting for those forms) in speakers' representation of linguistic convention, which incorporates a huge inventory of specific forms learned as units (conventional expressions). Out of this sea of particularity speakers extract whatever generalisations they can. Most of these are of limited scope, and some forms cannot be assimilated to any general patterns at all.
>
> (Langacker 1987:46)

Kjellmer (1991) refers to these two models of interpretation of language, the free choice principle and the principle of idiom, in his book chapter 'A Mint of Phrases'. He compares Sinclair's image of switching from model to model and back again and the image put forward by Aitchison (1987) in an article on the mental lexicon. Aitchison draws an analogy between the processes involved when language users access their mental lexicon and two alternative forms of urban transport, namely buses and taxis. Buses represent those language users who use ready-made combinations of linguistic elements (in this case morphemes to form words). Taxis represent language users who use discrete elements capable of being combined into larger structures. Her conclusion is that people try to be buses and only turn into taxis if the bus-route proves to be unsuitable. Humans start out by using memory and routine possibilities. Only when this proves inadequate do they turn to computation. This is very similar to Sinclair's two principles. For normal texts the first

mode to be applied is the idiom principle. Whenever there is good reason, the interpretative process switches to the open choice principle and quickly back again.

Kjellmer suggests his own analogy when he compares the act of speaking or writing to driving a car:

> We normally have a goal in speaking or writing...We have to obey the rules laid down by the grammar of our language and we normally follow certain 'lexical stretches', i.e. well-established sequences of words ... Decisions will of course have to be taken, minor ones at the 'cross-roads', at breaks between lexical stretches, and major ones at the 'main junctions', where one train of thought succeeds another. Again, few personal deviations from the established pattern occur, such as choosing unexpected words or ungrammatical forms. So, just as in driving, we use semi-automated routines in speaking and writing; both traffic rules/ grammatical rules and a road network/ a set of lexical stretches are essential to ensure adequate communication.
>
> (Kjellmer 1991:122)

Kjellmer's text is quoted here at length for several reasons. Firstly, he is at pains to include both speaking and writing in his model of language production. The phrase 'in speaking and writing' occurs twice in these lines and a third time on the same page. Secondly, it provides a powerful set of metaphors for modelling some of the psycholinguistic processes of written text production.

The terms used by Widdowson, Sinclair and Kjellmer, (1. 'a stock of partially pre-assembled patterns'; 2. 'a large number of semi-preconstructed phrases' and 3. 'certain *lexical stretches*, i.e. well-established sequences of words' respectively) have a denotation broadly similar to prefab as this was provisionally defined above in Section 1.2 above.

Warren (1999) decided to test the two principles proposed by Sinclair to explain language production. Paraphrasing Sinclair, she writes 'sometimes we compose word for word, sometimes we retrieve more or less ready-made multi-word units.' Her research programme includes the questions:

> 1. What is the average proportion of prefabs in texts?
>
> 2. How do prefabs combine with each other and with words combined according to the open-choice principle?
>
> 3. What are the lengths of prefab and non-prefab strings?
>
> (Warren 1999: 3)

Warren's results show the proportion of prefabs in texts to be 58.6% for her spoken corpus and 52.3% for her written corpus. It should be pointed out that in the spoken corpus she counted verb contractions (e.g. *I'm, don't, isn't, let's*) or reducibles, as she called them, as prefabs. For consistency, she also considered the full written forms of these connections (*I am, do not, is not, let us*) as prefabs. Although this estimate of prefabrication was carried out 'manually' on a small sample from the LOB corpus, it suggests that a sizeable amount of written text is composed of prefabs. This work encouraged me to pursue a computerized investigation of prefabs using larger corpora.

1.11 Writing and speech compared

The prime difference between spoken and written language is the medium or channel of communication: phonic and graphic substance. Substance here refers to and is used in the Saussurian model of language to contrast with form. Spoken language is transmitted by sound and written language by visible marks. Spoken language is time-bound, dynamic and transient while written language is space-bound, static and permanent. These bifurcations, often deployed by linguists looking at this question, can be multiplied. Speech is many thousands of years older and it develops naturally in children, whereas writing has to be taught.

An examination of some of the technical differences between the potentialities of the phonic and graphic substance shows that typing is slower than speaking (about 60 words per minute (wpm) for a competent typist compared with 150 wpm for moderately paced speech). Reading is considerably faster than listening with competent readers achieving speeds greater than 360 wpm. In spoken language the mode of transmission is serial: written language is also serial but access can be random because the entire text is available. The reader can scan, skim, skip and go back to search the text for information as required. Despite these and other differences between spoken and written language, some scholars prefer to conceive of the two media or channels as forming a continuum with one set of characteristics, typical of spoken language, at one end and another set of

characteristics, typical of written language, at the other. In this model, written and spoken texts can occupy positions at most points of the spectrum.

One measure of lexical variation in texts which has been used since the nineteenth century in stylometric studies (Kenny 1982) is type-token ratio. The number of word types (i.e. different words) in a text is divided by the number of tokens (total words counting each and every repetition). The result of this calculation is the type-token ratio (henceforth TTR) of the text. One of the limitations of TTR as a measure of the lexical variation of a text is that it is much too sensitive to the size of the text or corpus. If a text is 1,000 words long, it is said to have 1,000 tokens. But a lot of these words will be repeated, and there may be only say 400 different words in the text. Types, therefore, are the different words. The ratio between types and tokens in this example expressed as a percentage would be 40%. But this ratio varies very widely in accordance with the length of the text – or corpus of texts – which is being studied. A 1,000-word article might have a type/token ratio of 40%; a shorter one might reach 70%; 4 million words will probably give a type/token ratio of about 2%, and so on. A further illustration of just how dependent TTR is on corpus size is provided by a comparison of a short corpus, say this present chapter of my thesis, with 1,896 types and 24,248 tokens and the *Brown* Corpus which has 44,000 types in its million tokens, which gives TTRs of 7.82% and 4.4% respectively. The elaboration of the standardised TTR, or STTR, was an attempt by the designer of Wordsmith Tools, Scott (1999) to counteract this deficiency of TTR. By averaging the TTRs of successive chunks of text (from 10-20,000 words), the standardized type/token ratios can be used to compare texts of differing lengths. Meunier (1998:32), however, shows that standardized type-token ratio is not a discriminating feature between NS and NNS writers and that 'a lexically rich essay seems not necessarily to be a good quality one'.

In view of the dependence of TTR on corpus size, a number of scholars have shown an interest in the method developed by Ure (1971) in her pioneering work on lexical density. Ure used the proportion of lexical to grammatical words in text to measure the lexical density of texts and situate them on an idealized spoken-written continuum (Ure 1971). In her innovative comparison of two small corpora (total 42,000 words), Ure (1971) ascertained the lexical or content words by using a stoplist of 100 function words (grammatical or closed class words) and then making a wordlist of types which did not include these 100 most frequent words. The lexical density of each corpus was calculated

as the ratio of the lexical tokens in relation to the total number of tokens, expressed as a percentage. Stubbs (1996) replicated and corroborated the experiment on a larger scale (1.5 million words). The results of Ure's and Stubbs' experiments show that writing and speaking are not really contraries or mutually exclusive polarities but are better conceptualized as lying on a continuum. Other linguists support this conception of continuum. Tannen (1982) believes that

> both oral and literate strategies can be seen in spoken discourse. Understanding this, let us not think of orality and literacy as an absolute split.
>
> (Tannen 1982:49)

The influence of the written form on the spoken form occurs perhaps because we are a highly literate society, and the two forms inevitably influence each other.

Halliday (1989) using sample written texts aligned with their spoken 'translations,' demonstrates that spoken and written language differ in terms of lexical density and syntactic complexity. Compared with spoken language, written language tends to have higher lexical density and has a much higher proportion of nouns and nominalizations. Although spoken language tends to be more intricate using more clauses, it has a sparser information load than written language. If higher lexical density is associated with written language, the question arises as to whether this dimension could be used to gauge authorial expertise. This question is examined in Chapter 3.9. Certain values of lexical density for a text would indicate whether it complies with the levels of information content and interpersonal involvement appropriate to the genre or text-type instantiated. McCarthy (1990) observes that lexical density is determined by text type and is largely independent of text length.

McCarthy and Carter (1994) suggest that we should view the two kinds of language production as lying on a continuum. They make a distinction between the 'medium' and the 'mode' of a message. By 'medium' they refer to the means by which a message is transmitted, i.e. phonic or graphic substance or speaking and writing. 'Mode' refers to the choices the sender makes as to whether features normally associated with speaking or writing shall be included. For example, a university lecture, although spoken, will have many of the features associated with the mode of a written article in an academic journal (carefully planned and structured language, impersonal grammatical forms, etc.), whereas

an advertisement, although written, may immediately evoke a spoken, conversational mode of language use with, for example,. direct address to the receiver, ellipted and contracted forms etc. (McCarthy and Carter 1994).

McCarthy and Carter are using *mode* in a different way from Halliday and Hasan (1989) who use it in their theory of register to refer to, among other things, what McCarthy and Carter call *medium*, the phonic or graphic channels. Halliday and Hasan, however, also include 'form' under mode. The cline between the spoken and written mode of McCarthy and Carter bears a strong resemblance to Tannen's (1982) continuum of oral and literate strategies mentioned above. In much of the work done in psycholinguistics, e.g. in the work of Pawley and Syder (1983), which is examined in Section 1.7, the main consideration is the spoken language. In the typical face-to-face spoken encounter the time factor plays an important role. With so many higher order activities to perform, it is suggested that a person involved in a conversation would just not have the time to produce the utterances by composing them using lexical access and grammatical rules.

This situation of time boundedness which usually exists in spoken language does not characterize the writing situation. The writer is typically removed geographically from the intended reader(s) and, therefore, has more time (and space) within which to compose the message. Also, in the writing situations of EAP students it seems fair to suggest that, no matter how hard-pressed for time they might be, they would normally allocate sufficient time for their writing assignments. If people tend to have more time when they write, the question arises as to why prefabrication should be found in different kinds of writing. For example, in work done at Lund University, Wiktorsson (1998) reports levels of prefabrication in a written corpus of 39.4% (using samples from a novel and from newspapers and magazines). Surprisingly, a corpus of poetry, in a follow-up study, was found to contain 21% prefabricated language (Erman and Warren 2000). I should add that Wiktorsson and Erman and Warren counted grammatical contractions (*he's, shouldn't* and so on) and proper names (famous ones such as Bill Clinton and not so famous ones) as prefabrications.

Corpus-based work being carried out by Carter and McCarthy (2006) is gradually changing the 'geography' of the spoken-written continuum referred to in this section. Until relatively recently, grammar was based on written English and there was a perception that spoken English adhered less strictly to the rules of this grammar. Through work on the

CANCODE corpus, a much more rounded picture of the structure of conversational English is emerging. Carter and McCarthy have revealed that spoken English has its own distinctive structures. In light of this recent vindication of the systematicity of spoken English, reference to writing which is more like spoken English, or spoken English which is more like written, might become confusing. In order to overcome the difficulties of these and similar counter-intuitive observations which result from the spoken-written continuum, an alternative way of clarifying this graduated difference among written and spoken texts might be along a <formal><informal> cline (Carter 2005, personal communication). The elements of such an approach are contained implicitly in two of the major grammars of the English language created in the late twentieth century (Quirk et al. (1972) and Leech and Svartvik (1975:28-30)) although the application of such a <formal><informal> cline is not fully worked out.

The terms *formality* or *informality* of texts are used here to refer to ' the way in which the style or tone of language will vary in appropriateness according to the social context, the situation and the relationship between addresser and addressee' (Wales 2002: 160). Points on the scale or continuum ranging from very formal to very informal are each correlated with specific linguistic features. Formal language tends to be public, relatively serious and nearly always written. It is used in speech, for example in formal lectures and public speeches (Leech and Svartvik 1985:24). The following are some of the linguistic features which are more correlated with formal English: words of French, Latin, and Greek origin; passive mood; nominalizations; abstract subjects; avoidance of the pronouns *I, you, we*; co-ordination and subordination of clauses; and the use of precise intratextual links.

At certain points in Chapters 3 and 4, the level of formality found in the writing of the corpora is discussed. At that point in the thesis it will become clear that achieving an appropriate level of formality is one of the major challenges faced by NS and NNS apprentice writers of argumentative academic prose.

1.12 Approaches to teaching EAP writing

The history of the teaching of writing reveals two broad trends. There have been teachers who advocate mimesis as the best form of initiation into the writing skill and those who stress individual creativity and self-expression as the path to learning. A notable advocate of imitating models is Ascham (1561), who, in his famous sixteenth century method of 'back translation' of the classics, translated Cicero into 'neutral' English and asked his students to translate these texts back into Latin. The students were told that the nearer their Latin was to Cicero's the better it was. This is not simple imitation as the students were encouraged to invoke from their own minds the Latin of Tully. Presumably they approximated ever nearer each time they checked their efforts with the Master. 'Learning teacheth more in one year than experience in twenty, and learning teacheth safely, when experience maketh more miserable than wise' (Ascham 1561: 8). Although today this might seem a slavish way to learn to write in a foreign language, for most of recorded history the orthodox view of the activities of science, philosophy, education and the arts has been that their practitioners copy, repeat or mimic reality.

Feyerband (1987) cites Homer, Aristotle, Leonardo and Bacon as proponents of an imitation theory of either the arts or the sciences or both. But he shows that such mimesis does not have to be passive:

> This brief survey already shows that imitation is not unambiguous but involves a series of choices. One is the choice of the material in which the copies are to be produced. The imitator must take the properties of the material into account. These properties may be due to laws of nature, they may be results of custom (the standard phrases, the grammar, the words of the language used in written reports; metre, musical modes, standard gestures in tragedy).
>
> (Feyerband 1987:129-30)

Bagnall (1985) smiles at this earlier predilection for copying in most European cultures but still writes wryly of the consequences of the cult of the new and the promotion of self-expression, which became prevalent in British education in the 1960s. Some of the written school compositions which he cites from this period show that the freedom from prescribed models and the invitation to 'commune with their souls' or to 'emote' which was

extended to schoolchildren often produced texts of the utmost banality. Bagnall suggests that stock phrases have an important role to play in self-expression. One of the consequences of the cult of originality is that:

> in our efforts to be different, to cast off the dead paraphernalia of the past – to stand as it were, in our own shoes – we find ourselves strangely naked. We have congratulated ourselves on our freedom, and have ended up tongue-tied.
>
> (Bagnall 1985: 71)

Moore and Carling (1988) agree with Bagnall when they warn that:

> what we value most in language-creativity, expressiveness... allows us to succeed less well in having others understand us than the largely prefabricated phrases we use to say almost the same thing over and over again...routines, set phrases, clichés are not invariable nor are they dispensable. We need them.
>
> (Moore and Carling 1988:71-2)

Bagnall (1985: 71) argues for the usefulness of clichés to give readers or listeners a sense of belonging. He laments the loss to writers resultant from Partridge's censoriousness in his *Dictionary of Clichés*. He wishes that Partridge (1940) could have labelled some expressions as 'overworked' 'used' or 'tired' with the suggestion that they not be overused instead of proscribing nearly all 3,000 of them. Howarth (1995), after investigating all the main approaches to cliché, decides that it is not a phraseological type and that the term is used as a judgement on the contextual appropriateness of a word combination which might belong to any of the phraseological categories discussed in Section 1.4.

Teachers of writing, then, need to be careful about encouraging originality in their students and should have clear ideas about what they mean by it. It may be desirable for students of, say, literature to produce written work which is, in some way, original but they still need to conform largely to the style prevalent in the discourse community of academic literary criticism. Perhaps what EAP students should be trying to do is to 'sound' like someone from the discourse community they wish to join. It is almost impossible to get an academic article accepted by a peer-reviewed journal if it does not follow the discourse patterns of the discipline. The need is still felt by people writing in their first language to base their writing on good models, although Flowerdew's (1993:309) reference to 'closet'

in the following quotation suggests a fear of being out of fashion (or out of step with the times):

> Many native speakers make use of others' writing or speech to model their own work in their native language when the genre is unfamiliar. It is time that this skill was brought out of the closet, and exploited as an aid for learning.

Obviously, EAP students will need to be made aware of the pitfalls of using generic stock phrases inappropriately. Nevertheless, because the journal articles, text books and handouts which students are required to read in their course of study contain domain-specific and domain-defining phraseologies, part of any language awareness activities involves bringing these sequences to their attention so that they produce essays, reports and dissertations with these phrasal elements comfortably incorporated.

In many EAP and ESP situations, the students need very little encouragement to turn to 'appropriate texts' within their disciplines and choose 'acknowledged sources' as their model for writing. These might have been written by co-nationals or other non-native speakers of English or generated by such writers and 'polished' by commissioned native speakers.

If much of our written language output consists of variously arranged chunks of pre-assembled words then perhaps the question of plagiarism needs to be reconsidered. In an internationalized educational milieu, the differing attitudes towards citation and borrowing within different cultures need to be confronted and, at the institutional or national level, decisions taken about what is to be considered legitimate and illegitimate in relation to such borrowings. Bloor and Bloor (1991) discuss, in the context of EAP, those ideas which belong to all in the academic community. They suggest that, to write about the square on the hypotenuse of a triangle being equal to the sum of the squares on the other two sides, it would still be necessary to cite Pythagoras. However, they suggest that in writing about linguistic competence, we would not have to acknowledge the formulation of this concept by Chomsky. The latter concept they call *free goods* after Goffman (1967:12): 'Free goods are those which, in a given situation, anyone can use without seeking permission'. In relation to this discussion of plagiarism, originality and prefabrication, the Bloors have an interesting observation to make about the academic world:

> Perhaps unfortunately, only the very great are permitted idiosyncratic, quaint or individual forms of written expression in this community.
>
> (Bloor and Bloor 1991:11)

A less pessimistic gloss of this quote from the Bloors could be 'you have to learn to walk before you can run'.

Students who are non-native speakers of English need to be initiated into, or discover for themselves, the 'house style' or acceptable norms for writing within their chosen profession or area of study. As well as achieving an acceptable level of lexical, grammatical and discoursal accuracy, they also have to use certain key phrases to show that they belong to their chosen discourse community. To put it another way: it is not enough to write correctly avoiding any infringement of the grammatical rules of the language. As Pawley and Syder (1983) forcefully point out, a much smaller set from the innumerable number of sentences generated by the grammar of a language is considered native-like or idiomatic. Each language chooses certain ways to say things and many other (grammatically) possible ways sound decidedly odd. Students of EAP writing need help with phraseology: if lexicalised sentence stems are as numerous as Pawley and Syder think they are, then this help might take the form of awareness-raising and guided self-instruction.

During the last 30 years of the twentieth century there were various swings in the approach to teaching EAP writing. These fashions are succinctly summed up in Ann Raimes's (1991) state-of-the-art article for the 25th anniversary number of *TESOL Quarterly*. Starting 25 years earlier, her synoptic history describes four main stages which are summarized here:

1) 1966–1975 Focus on Form

Under the audiolingual method, students only performed grammatical transformations, e.g. changing verbs from present into past. Grammar was not the only form that was focused on: contrastive rhetoric studies posited that other cultures had different ways of building and organising paragraphs. The pedagogical approach inspired by contrastive rhetoric trained students in the use of topic

sentences followed by examples and illustrations to make paragraphs. Students imitated outlines, did paragraph completion and re-ordered scrambled paragraphs.

2) 1976–present day Focus on the Writer

This approach moved away from sentence combining and controlled composition. Influenced by L1 writing research, it taught writing as a process and looked at what L2 writers actually do as they write. Concern with 'accuracy' and 'patterns' was replaced by concern with 'process,' 'making meaning,' 'invention,' 'multiple drafts' using journals, free writing, peer collaboration, and revisions.

3) 1986–present day Focus on Content

EAP courses are attached to a content course in the adjunct model. Learners are helped with the language of the thinking processes and the structure or shape of the content. A good deal of experimentation with team-teaching was undertaken and EAP teachers collaborated with subject teachers in teaching linked courses and topic-centred modules.

4) 1986–present day Focus on the Reader

Focus on the Reader is a return to an emphasis on form but on rhetorical rather than grammatical forms. Genre analysis (discussed in Section 1.13) becomes central and students are socialized into their target academic community. The reader, often conceived as 'omniscient' and 'all-powerful' becomes an 'abstract representation, a generalized construct... reified from an examination of academic assignments and texts' (Raimes 1991: 412). Raimes' reference to the reader as an abstraction although surprising at first is borne out by the experience of most EAP writers. Even in the case of Masters dissertations, there are usually only one or two readers involved (e.g. the tutor and a second reader).

This overview of the changes of emphasis within the EAP teaching community over the last decades shows the multifaceted nature of the writing process. There is little doubt that insights from all four approaches are needed to do justice to the complexity of writing. Form, be it grammatical or rhetorical, writers and their creative processes, content and the

concepts and structure of each subject area and readers and their expectations all have to be taken into account by teachers if they are to help apprentice writers develop written styles appropriate to their needs.

Tribble (1996), in a handbook for teachers of writing, presents a similar schema, from the student writer's point of view. He suggests that writers require four kinds of knowledge in order to be able to respond adequately to a specific writing task. They are rearranged here to bring out the parallels with Raimes's historical synopsis:

> ***language knowledge***
> knowledge of those aspects of the language system necessary for the completion of the task
> ***writing process knowledge***
> knowledge of the most appropriate way of preparing for a specific task
> ***content knowledge***
> knowledge of the concepts involved in the subject area
> ***context knowledge***
> knowledge of the social context in which the text will be read, and co-texts related to the writing task in hand
>
> (Tribble 1996:43)

As each new approach developed, it brought into existence a new curriculum, methodology, and a range of procedures and techniques but, interestingly, none of the four approaches ever really disappeared. In relation to the penultimate revolution mentioned by Haimes, with the focus on the reader, the concept of genre came to the fore. The next section discusses how this concept has been developed within EAP and the central role it plays in using corpus analysis.

1.13 A working model of genre

The focus on the writer mentioned in the previous section was influenced by research into L1 writing where, for example, writing instructors sought ways to help 18-year-old undergraduates who were not yet able to write effectively in their first language to succeed on their university courses. The resultant emphasis on the writing process is important and all writers need to know how to generate ideas, write recursively, loop back to previous drafts, edit, splice different sections or versions and proof-read. Nevertheless, most EAP students are already competent writers in their first language and might not need so much help with ways to prepare for the writing task. Perhaps what they do need is knowledge of the conventions and constraints related to their new and unfamiliar readership. This concern with the reader goes beyond the process of creating text to look at how the writer matches a text to a social purpose and provides EAP students with what Tribble (1996) designated 'context knowledge'. As was remarked above, a key concept which serves to mediate this knowledge is the concept of genre.

Genre, like many other concepts, is often easier to exemplify than to capture in a definition. It originally comes from literary studies and art history where it is used to classify different kinds of texts or paintings. For example, the genre 'poetry' numbers among its subgenres the sonnet, the epic poem, the lyric poem, the ballad, the pastoral elegy and so on. Formal linguistic criteria are often used to distinguish these literary subgenres. In the case of poetry, features such as length and number of lines, diction, rhythm, metre, versification and rhyme, among others, might be invoked to assign a work to one of the sub-genres.

With the spread of English for Specific Purposes in general and EAP in particular, many different kinds of texts began to be analysed with the same care and attention to detail as had only been lavished, in the past, on legal, literary, or religious texts. In fact it was in these last two domains that concordances were first laboriously developed long before the advent of the computer.

Late twentieth–century societies are peculiarly text-based. Bernstein (1990) discusses professionals who specialize in discourse such as priests, scientists, lawyers, teachers and administrators while Giddens (1991) points out the vast literature of guide books which deal with all aspects of modern life. Educators at different levels of the educational system from the primary to the tertiary level began to perceive that the ability

to process a wide variety of text types and to produce some of them, many of a prosaic and practical nature, enhances the social effectiveness of their pupils or students.

In the fields of language education and linguistics, the concept 'genre' is extended to classify uses of language to communicate in all areas of life. Thus, the term describes types of activities such as prayers, sermons, songs and poems 'which regularly occur in society' (Dudley-Evans 1989:77) and are recognized by the speech community as being of the same type. The concept of genre was introduced into EAP studies by Tarone et al. (1981) in an article investigating the reasons for, and effects of, choosing the active or passive voice in two journal articles on astrophysics. The conclusion reached was that it was the writer's communicative purpose which governed grammatical and lexical choices. This criterion, of writers' purpose, was to remain the defining feature of genre for most theorists in EAP and serves to distinguish it from the concept of literary genre as discussed above. In EAP, therefore, the emphasis is on the means by which a text realizes its communicative purpose rather than on the evolution of a system for the classification of genres.

The concept of genre was developed by Martin (1985) in the context of primary and secondary education:

> Genres are how things get done, when language is used to accomplish them. They range from literary to far from literary forms: poems, narratives, expositions, lectures, seminars, recipes, manuals, appointment marking, service encounters, news broadcasts and so on. The term genre is used here to enhance each of the linguistically realized activity types which comprise so much of our culture.
>
> (Martin 1985:250)

Martin in this passage refers to Australian culture. His concept of genre developed out of Halliday's (1978) systemic functional grammar and its concept of register. Register is a semantic concept which can be defined as a configuration of meanings that are typically associated with a *context of situation* which comprises the three components, *field, mode* and *tenor*. The first of these, *field,* includes setting or scene, topic and action; *tenor* is made up of participants, their purposes and role relations; and *mode* includes the channel used for the communication and its form.

Martin suggests removing purpose from its place within tenor and moving it upwards, out of context of situation, to become part of a new higher level called *context of culture*. He proposes *genre* as a suitable term to describe text types at this level of generalization. Martin gives the following definition of genre for the purposes of this discussion:

> For us genre is a staged, goal-oriented, purposeful activity in which speakers engage as members of our culture.
>
> (Martin 1984: 21)

For Martin, virtually anything a person does involves participating in one genre or another. Culture is thus 'a set of generically interpretable activities'. In answer to the question as to how genre is realized, Martin suggests the visual image of 'waves of field, mode, and tenor flowing through text giving it a distinctive purpose-oriented, staged structure' (Martin 1984:21). A text, therefore, can be viewed at the following levels:

CONTEXT OF CULTURE	Genre
CONTEXT OF SITUATION	Register (field, mode, tenor)
SEMANTIC SYSTEM	Linguistic realizations

(Martin 1984: 21-30)

Appropriating Sandberg's (1959) famous aphorism in relation to slang, it could be said that genre is 'language that rolls its sleeves up, spits on its hands and goes to work'. Swales (1990) in the context of EAP develops a very similar concept:

> A genre comprises a class of communicative events...which share some set of communicative purposes...recognized by the expert members of the parent discourse community and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of discourse and ...choice of content and style. Communicative purpose... operates to keep the scope of a genre ...focused on comparable rhetorical action... exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience.
>
> (Swales 1990:58)

Martin and Swales broadly concur in their definitions but Swales stresses the receptive aspect of the activity (the reader) while Martin gives more emphasis to production (by the speaker or writer).

Most of Swales' (1981) early generic research was done on the introductions to experimental research articles. When Swales looked at his corpus of such article introductions, he discerned the 'moves' contained in such writing (the concept 'move' he adopted from Sinclair and Coulthard [1975] and their model of discourse analysis). These moves are the vehicle of the action and its structure and purpose as referred to in the definition given above. Over the years, Swales has evolved his original four 'move' model into a three 'move' model which could be summarized as (1) reviewing the literature or establishing territory (2) indicating a gap in the literature and (3) outlining purposes. He describes this as his Create a Research Space (CARS) approach. The criticism that such moves are a superimposition of Swales' own perceptual predispositions has been partially met in an investigation by Crookes (1986) where he obtained substantial agreement among subjects using Swales older four 'move' model to analyse article introductions. Other models involving 'moves' have been developed by Dudley-Evans (1986) and Bhatia (1993) and more generalized patterns of text organization have been applied to academic writing such as that of the situation-problem-response-and-evaluation of response- model (Winter 1977, Hoey 1983). This last model is applicable to a wider variety of genres than the Swalesian model, e.g. advertisements, letters of complaint, scientific discourse, short stories and novels (Hoey 1983). However, some commentators prefer to categorize the situation, problem, response and evaluation of response framework as a text type schema (Paltridge 1996).

Biber (1988) points out that the term 'genre' categorizes texts on the basis of external criteria such as their communicative purpose and the discourse community to which they belong, in contrast with the notion of 'text type', which represents a grouping of texts which are similar in terms of the co-occurrence of linguistic patterns regardless of genre. Thus, the terms 'genre' and 'text type' provide different but complementary perspectives on texts. Biber found that the same genre can differ greatly in its linguistic realizations and that different genres can be quite similar linguistically.

The patterns of moves (Dudley-Evans 1989, Swales 1990, Bhatia 1993) have clear pedagogical utility in the teaching of writing. Students of EAP can be made aware of such

'moves' which can be translated into functional terms: 'at this stage of the text you are citing previous research as a support for your claim' or again: 'here you are generalizing from your results to make a knowledge claim either in the form of a hypotheses or deduction.' Such rhetorical features can be linked to grammatical choices. References can be made to the choices of verb tense and aspect, for example, where a writer chooses the present simple, present perfect or past simple tense for decreasing claims of generality for facts contained in a citation.

Genre, then, is a way to look at texts in action and identify the purposes and communities they serve. It provides an intuitive way to identify and assemble texts into collections according to their communicative purpose. Providing what Biber (1988) calls an external situational criterion for text selection, the concept of genre can be used by researchers to construct apprentice writer and expert writer corpora. These corpora can then be examined for patterns at the phraseological level to see if certain expressions are overused or underused by apprentice writers.

Genre is a metalinguistic term equally useful to the researcher and to the EAP teacher based, as it is, on the practicalities of written language performance. Genre is worth knowing at all levels of the educational system from primary school upwards. Knowledge of genres can be liberating and gain its possessor admission into certain worlds, academic and professional. As Thorne (1988) puts it:

> to know a language in the fullest sense is to know not just its grammar and (at least in part) its vocabulary but the very many ways in which it can be used. In the case of English this includes knowing what is a report, a narrative, a commentary, a synopsis...
>
> (Thorne 1988:142)

1.14 Modality, evaluation and stance in writing

An important dimension of writing which rewards closer investigation is the way in which writers express their attitudes, value judgments, assessments and feelings about the propositions in their writing. This area of study has been variously designated *modalization* (Halliday 1994), *evaluation* (Hunston and Thompson (eds.) 2001) and *epistemic* and

attitudinal stance (Conrad and Biber 2001). Several related concepts have been defined and discussed in the literature on academic literacy including *voice* (Elbow 1998), *persona* (Bizzell 1992), *identity* (Ivanič 1998) and *metadiscourse* (Hyland 2000).

This thesis, as stated in Section1.1, investigates the relationships between the phraseologies used by Portuguese undergraduate EAP writers and the discourse functions they wish to perform. By attending to the functions of prefabs in writing, their contribution to writer's *stance* can be appreciated. My experience as an EAP teacher made me aware that this dimension of discourse presents advanced learners of English with some of their greatest difficulties. The role that prefabs play in writers' self-affirmation is studied in this thesis. The contribution of prefabs to the expression of evaluation and stance are also elucidated in this work.

Hunston (2001) distinguishes two planes in written texts - the 'autonomous' and the 'interactive'. On the interactive plane, claims are evaluated and there are references to discourse-entities. Each clause shows how it is related to the previous clauses or the way the text is constructed and the relationships of averral, evaluation and corroboration between clauses are established. On the autonomous plane, world-entities are referred to and the content of the text is revealed. In both academic and non-academic argument, the writer normally avers a statement or attributes it to others. These other sources can be people with whom the writer agrees or disagrees.

> Status on the interactive plane, that is, status ascribed to the statements in a text, is largely, though not by definition, concerned with the evaluative parameter of certainty. In other words the statements differ from each other largely in terms of how certain or uncertain they are. For example 'a fact' is more certain than an 'assessment', and an averred assessment is treated by the text as more certain than an attributed statement.
>
> (Hunston 2000:185)

This aspect of evaluation, namely that of epistemic stance, has particular relevance for the writing of my student subjects. In pioneering work since the 1970s, this dimension of language was referred to as 'hedging' (Lakoff 1972; Skelton 1985). Language learners need to learn how to modulate what they are saying or writing so as not to sound dogmatic or arrogant on the one hand, nor too tentative on the other. There are many ways to perform such hedging, lexically, grammatically and morphologically.

Figure 1.5 Five ways to hedge

From notes taken at an M.Sc. lecture delivered by John Skelton, 29 October 1986, Language Studies Unit, Aston University

The world is flat	*It is said the world is flat*
God is great	*God may be great*
Frank Bruno is unstoppable	*Frank Bruno looks unstoppable*
Robin Cook has a forceful personality	*I'm not sure that Robin Cook has a forceful personality*
Sarah Ferguson has red hair	*Sarah Ferguson has reddish hair*

Two kinds of hedges are distinguished by Lakoff (1972) - the 'Shield' and the 'Approximator'. The Shield is used to protect the speaker or writer in case they are mistaken:

SHIELD: *I suspect the moon isn't made of green cheese after all*

APPROXIMATOR: *It's made of some sort of rock stuff.*

In the original formulation of the distinction, Lakoff (1972) suggested that the Shield hedged the speaker, giving their degree of commitment to the proposition, while the Approximator hedged the proposition and gave the degree of truth being claimed in it. On closer inspection we find that most Approximators are Shields. This distinction between the likelihood of the proposition being true and the degree of commitment to the assertion reappears in Conrad and Biber's (2001) distinction between epistemic and attitudinal stance. Skelton (1985) makes a distinction between a proposition and the comment about the proposition. A comment modulates what is said in the proposition. Modulations function by specifying as precisely as possible the relationship between speaker (or writer) and the proposition. The kind and degree of specificity appropriate in relation to a proposition is governed by the *contract of inexactitude* (1985).

Cf. It was a bit hot this afternoon
At 5.15 the temperature reached 33.2 ° C.

There appears to be a growing realization among scholars in different disciplinary areas that this phenomenon, which started out as an interesting feature of pragmatics and of importance to the teaching EAP writing, might be of much more far-reaching importance in the language sciences. Publications by Naess (1966), Lemke (1999), Salager-Meyer (2000) and Hunston and Thompson (2001) affirm the philosophical/scientific importance of imprecision and hedging. An earlier analysis of the role of precision/imprecision at each stage in discourse is found in the work of the philosopher, Naess (1966). His concept of 'precization' presaged Skelton's (1985) contract of inexactitude. According to Naess, statements at each stage of argumentation or discussion require a certain level of clarity and detail, no more and no less, if they are to be felicitous. In research on vague language, Channell (1994:21) found that learners of English had not been equipped with the 'vagueness category identifiers' they needed and did not always recognize when and to what degree their language should lack precision. Even advanced NNS writers of EAP can appear dogmatic in their texts through their inability to use downtoners and other hedges.

Work on modality by Aijmer (2001) is pertinent to this discussion of evaluation in text and also relates to considerations of writer's stance. Aijmer (2001) investigates the discourse function of *I think.* She points out that the modals *must, have got to,* and *should,* in their root or deontic uses, are characteristic of persuasive patterns of argumentation (Aijmer 2001). In argumentative essays they are invariably used rhetorically to persuade the reader to accept the writer's point of view. What she terms 'epistemic modal forms' (i.e. expressions of doubt and certainty) are used to reinforce or moderate claims and to indicate the type of evidence for a claim. On examining the Swedish sub-corpus of *ICLE*, *Swicle*, she discerns two distinct discourse functions of *I think (that)*:

(1) A deliberative or authoritative function: the speaker or writer is making a strong self-assertive pronouncement as in example (1a) below.

(2) A more modest admission of some degree of uncertainty or tentativeness. It often occurs non-initially as in example (2b) and weakens the claim made by the writer.

(Aijmer 2001: 248).

Aijmer (2001:48) supplies two exponents of these discourse functions both taken from the *Swicle* sub-corpus:

1a Therefore my conclusion is inevitably an over-all-one: I think that both assimilation and integration is needed.

2b Here, I think, the percentage of immigrants is around 70-80% and that has created problems.

(Aijmer 2001: 248)

Simon-Vanderbergen (2000) studies the use of the expression *I think* in televised political interviews and political debates and shows that it has different functions depending on register and mode. It is indexically linked not only to epistemic or affective stance but also with interpersonal relations of power and authority, social identity (learner, novice writer) and group identity (e.g. social class or gender).

Related to these notions of power and identity is the phenomenological reality of language use which is discussed by Verschueren (2000) in his study of metapragmatics. In this study he gives such emphasis to a speaker or writer's *reflexive awareness* that he affirms:

> all verbal communication is self-referential to a certain degree. In other words, there is no language use without a constant calibration between pragmatic and metapragmatic functioning.
>
> (Verschueren 2000:187).

Verschueren's idea about the universality of metapragmatic functioning could be translated into Halliday and Hasan's terms so that the realization of the ideational function would also

exercise a self-referential i.e. a textual function. A speaker or writer is stating a proposition in using a prefab or several together but is also using this self-same piece of language, which is known to be shared, to affirm membership of the speech community and to self-consciously resonate with other usages of the same sequence.

Adverbial marking of stance

Following on from their work on the *Longman Grammar of Spoken and Written English* (Biber et al. 1999), Conrad and Biber (2001) give a report of an analysis of stance in two five-million-word corpora of British English conversation and American academic prose. They identify three kinds of meanings associated with stance in speech and writing:

(1) epistemic stance, which indicates how certain the speaker or writer is, or where the information comes from (e.g. *perhaps*, *according to France Presse*);
(2) attitudinal stance, which indicates feelings or judgements about what is said or written (e.g. *amazingly, sadly*);
(3) style stance, which indicates how something is said or written (e.g. *frankly, in short*)

The three most common grammatical realizations of stance adverbials are

(1) single adverbs (Biber and Conrad include *sort of* and *of course* in this category, as fixed, multi-word expressions)
(2) prepositional phrases (included is *in fact* where the words retain their meaning and the noun can be modified, e.g. *in actual fact*)
(3) finite subordinate clauses

Conrad and Biber (2001) observe that stance adverbials are relatively rare in all registers compared to circumstance adverbials (with meanings such as time, place, and manner), which occur with ten times more frequency. Stance adverbials occur almost twice

as frequently in conversation as in academic prose. Conrad and Biber (2001) also used a five-million-word corpus of British news reportage for occasional comparisons and found, to their surprise (2001:64), that academic prose writers use stance markers almost twice as often as newspaper writers. Finally, Conrad and Biber, in this same study, found that epistemic stance adverbials are by far the most frequent of the three kinds of stance adverbial, and they found a fair number of author attitudes in manuals, textbooks and in technical reports.

These three grammatical realizations (single adverbs, prepositional phrases and finite subordinate clauses) account for over 90% of all stance adverbials in conversation, academic prose and news reportage. Prepositional phrases are most common in academic prose as realizations of stance adverbials (e.g. *on the whole, in most cases, from our perspective*). These prepositional phrases are of value in limiting the generalizability of claims or in stating plainly that it is the author's view that is being expressed. In Biber and Conrad's study, finite subordinate clauses are by far the most common grammatical realizations of stance adverbials in their conversational corpus. They observe that this would seem to belie the stereotyping of conversation as structurally simple. The majority of these finite subordinate clauses are comment clauses of the form *I guess, I think*. In Chapter 3, evidence is given of the advanced learner's predilection for using these finite subordinate clauses to express their authorial stance.

Stance and content analysis

A different experimental approach to hedging and epistemic levels of certainty is used in Chapter 4. It looks at one aspect of evaluation in text: the degree of 'dogmatism' shown by writers and how this relates to their lexical and phraseological expression of stance. The main source of ideas for the psychological constructs, dogmatism and open-mindedness is the work of Rokeach (1960), who defined dogmatism as 'the extent to which the total mind is an open or closed one'. Dogmatism in Rokeach's view is related to: (1) relatively closed cognitive organizations of beliefs and disbeliefs about reality, (2) organized around a central set of beliefs about absolute authority, (3) which provides a framework for intolerance toward others (Rokeach 1960:195).The concepts of 'dogmatism' and 'open-mindedness' were refined and developed in the work of the psychologist, Ertel (1985) and his followers (Berth and Romppel, 1999). A form of content analysis, Dogmatism Text

Analysis (henceforth DOTA), is used to measure the level of dogmatism in oral or written speech production. This is a fully automated analysis which can provide triangulation for a corpus linguistic approach. In Chapter 4, the essays in three corpora are submitted to the DOTA Content Analysis. The observation above that apprentice writers need to learn hedging skills is tested using this Dogmatism Quotient Test. Some of the research conducted with DOTA is reminiscent of work in genre analysis, which divides texts into moves, each with its own specific rhetorical strategies (Swales 1990) and work in academic literacy (Kelly and Bazerman 2003), which looks at the optimal use of levels of epistemic certainty at different stages in the development of academic essays.

1.15 Conclusion

This chapter has discussed the emerging conviction among linguists that a grammar, no matter how powerful, could not explain the degree of fixedness and predictability of lexical co-selection in the creation of text. It became clear that the constraints operating on the collocability of words were much more extensive than the selectional restrictions of a generative grammar. The discipline of phraseology emerged and its focus lay somewhere between the generality of grammar and the specificity of lexis. According to estimates by Bolinger (1976), Pawley and Syder (1983), Langacker (1987) and Jackendoff (1997), the prefabs stored in the mental lexicon of the average native speaker of a language are roughly as numerous as the single lexical items. One challenge for phraseology is the need to impose some kind of order on this large numbers of syntagms. The generalizations found in grammars, often expressed in rules with a manageable number of exceptions, may not be available in this kind of classification. One difficulty for phraseologists who are also EAP teachers is that the results of phraseological analysis tend to be more probabilistic and have a more localised application than the powerful rules of descriptive or pedagogic grammars.

In Chapter 1.0, the two research questions which inform this work are stated. The overall aim of this thesis is to study the phraseology of Portuguese EAP writers and the way in which they use prefabs in comparison with three groups of native speaker writers of English. Various definitions of prefab were examined and a set of criteria were defined for

identifying prefabs in texts. Phraseology, which provides a theoretical framework for the study of prefabs, was introduced and an outline of the recent history of the prefab was provided. It was noted how the focus in phraseology has gradually shifted from idioms, which are relatively frozen and opaque, to more or less transparent restricted collocations in 'ordinary language'. My own interest is in the prefabs which occur in argumentative essays in EAP and secondarily, in newspaper editorials. To date, there have been few investigations of prefabs in learner English, with the exception of Granger (1998b); DeCock et al. (1998) and Milton (1998) and none dealing with the learner English of Portuguese University students.

In order to assess the extent to which Portuguese students use prefabs and the way in which their prefab use differs from that of native speakers, a corpus methodology is used. Chapter 2 begins with a discussion of the way in which this new approach has been received in the disciplinary community of linguistics. An account is given of the initial resistance and the final coming of age of this new approach to language study. The Chapter goes on to describe the theory of corpus compilation and the methodology of corpus interrogation. The later development of Computer Learner Corpora (CLC) is also chronicled in Chapter 2.

Chapter 2 Corpus linguistics: From method to paradigm.

There was a time when members of the linguistics corps regarded a corpus as a corpse.

(Allén 1992)

2.0 Introduction

Practitioners of corpus linguistics and its opponents have made various claims about its academic and scientific standing. The first section of this chapter traces the emergence, in the sixties, of corpus linguistics and its rapid development, until the present, as a research methodology made possible by advances in computing. Chapter 2.2 looks at the 'hard' and 'soft' versions of the research programmes (or the corpus-driven vs. corpus-based approaches) which have been proposed by corpus linguists. 'Hard' and 'soft' here refers to the philosophy of science espoused by the corpus linguists when they use a corpus methodology. Practitioners of hard corpus linguistics try to be as free as possible from theoretical preconceptions so that they can follow where the data lead. Soft corpus linguists view corpora as one more tool for pursuing the answers to the traditional questions in linguistic study.

An assessment is made of the contribution this new approach has made to the study of language (Chapter 2.3) and such analytically useful concepts as semantic prosody, colligation and pattern grammar are elucidated. In Section 2.4, one of the more recent initiatives within corpus linguistics, the study of learner corpora, or Computer Learner Corpora (CLC), is described. This section looks at the contribution that CLC can make to a better understanding of second language acquisition and EAP student writing. Details are given of the research culture of the CLC scholars and the contrastive studies of learner language they have carried out. The contribution that corpus methodology can make to the study of prefabs is assessed in Section 2.5. In this section techniques and criteria which can be used for identifying and classifying prefabs are described and the efficacy of the different approaches is compared.

Section 2.6 takes further the discussion in Section 2.1 about the authenticity and representativeness of the data and their epistemological status. The various sections of this

chapter together with Chapter 1 provide the theoretical foundations for the methodological decisions taken in Chapter 3.

2.1 Corpus or Corpse Linguistics

A concise definition of corpus is provided by Atkins and Clear (1992:5): 'A corpus is a body of text assembled according to explicit design criteria for a specific purpose'. Although linguists and anthropologists used corpus techniques long before the advent of the microcomputer, the present-day term typically refers to a computerised corpus. Nowadays, although manual work continues to be done in corpus linguistics (Howarth 1998, Wiktorsson 1998, Warren 1999) scholars generally accept that computer use at some stage of the process is normal working procedure (for example, Howarth and Wiktorsson extract their small corpora from the Lancaster Oslo Bergen computer corpus). Corpus linguistics therefore means 'computer corpus linguistics' (Leech 1998: xvi).

There is a constellation of disciplines which study the lexicon and use computerized language corpora as sources of data. This investigation draws on a number of rapidly evolving areas of enquiry, namely computational linguistics, computational lexicology, computational lexicography and corpus linguistics. Each of these disciplines investigates the lexicon in different ways (Ooi 1998:1). The first, computational linguistics, is the study of computer systems for understanding and generating natural language (Ooi 1998:23). Computational lexicography and computational lexicology (Ooi 1998: 29) can be seen as complementary studies of the lexicon, the former engaged in the design and production of dictionaries and the latter supplying the theoretical knowledge of the lexicon which underpins such endeavours. The final discipline in the list above and the subject of the present chapter section, corpus linguistics, has been posited as the 'mother discipline' from which the others derive their methodological and theoretical frameworks.

Leech (1992:105) suggests that corpus linguistics refers 'not to a domain of study, but rather to a methodological basis for pursuing linguistic research'. In this view, corpus linguistics is not a branch of linguistics, as are syntax, semantics and phonetics, but a method, an approach to linguistic enquiry. Syntax, semantics, phonetics, pragmatics and sociolinguistics are just some of the areas of linguistic enquiry that can be investigated

with this methodology. Corpus linguistics can be used to differentiate between approaches to language study and so there is corpus-based syntax and non-corpus-based syntax, corpus-based semantics, non-corpus-based semantics, and so on, in most sub-disciplines of linguistics.

As mentioned above, corpus linguistics has been practised for longer than the short lifetime of the computer. In the earlier part of this century, descriptive linguistics, particularly in America, was predicated on the collection and analysis of corpora. For the American linguists of the period, the creation of corpora of Amerindian languages threatened with extinction was an urgent task. Boas (1911) and workers in the empiricist/behaviourist tradition insisted that a scientific theory of language should reject all data that are not directly observable or physically measurable (Bloomfield 1935). For such scholars the corpus was the *sine qua non* of scientific description (Leech 1991b).

The paradigm shift away from behaviourist linguistics was heralded by the publication of *Syntactic Structures* (Chomsky 1957). Even today, some 50 years later, corpus linguistics still takes Chomsky's critique of corpora into consideration (McEnery and Wilson 1996). One of the criticisms Chomsky made of a corpus-based approach was that it modelled the wrong aspect of language, namely performance, whereas he argued that the object of linguistic inquiry must be the knowledge or competence underlying such performance data. The key question for Chomsky was what the brain must be like to account for the learning and production of language. This critique of any linguistic theory based on empirical data is intimately linked to the theory which undergirds the whole Chomskian programme, his modern version of rationalism. Chomsky rejected behaviourism and its philosophical forebear, empiricism, and deliberately espoused the philosophy of rationalism. Instead of accounting for language observationally, he enjoined linguists to investigate language introspectively. A theory should be designed, he argued, which economically accounts for all or as many of the language data as possible. What counted as language data for Chomsky were the intuitions of the native speaker. Various accounts of the debate between rationalist and empiricist linguistics are provided in the literature (Leech 1991b; McEnery and Wilson 1996; Kennedy 1998).

Chomsky's (1957:159) second criticism of corpora was that 'any natural corpus will be skewed'. According to Chomsky, corpora are partial and in two ways. Firstly,

because, except in the case of a dead language, they are incomplete. They will contain some, but not all, of the valid sentences of a language. Secondly they are partial through being skewed because the frequency of a feature is an important determiner of its inclusion in the corpus. Many infrequent sentences would be excluded (Chomsky's often-quoted example was the sentence *I live in Dayton Ohio* being much less likely to occur than *I live in New York* for obvious demographic reasons). Those infrequent sentences which do enter the corpus might then be given undue significance.

> Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so widely skewed that the description would be no more than a mere list.
>
> (Chomsky 1957:159)

This criticism attacks the assumptions made by the users of corpora in traditional linguistic scholarship before the advent of Chomsky. The behaviourist linguists (e.g. Bloomfield 1933, Harris 1951) in their work on the transcription and analysis of extant, but previously unwritten or unstudied, languages had been assuming that language was finite or that the sentences of a language were enumerable.

Chomsky's third objection to corpora was the problem of deciding the grammaticality of a sentence which has not been said or written to date and has therefore not been captured in a corpus. Without recourse to introspection or to a native speaker, how can we distinguish between ungrammatical sentences and grammatical ones which merely have not yet occurred?

In response to Chomsky's criticism that a corpus models performance rather than competence, the corpus linguist today could point to the quality and quantity of language data which modern computer science enables us to examine. The scope and representativeness of the latest generation of mega-corpora could be adduced as defence against the charge of incompleteness and skew. A possible rejoinder to Chomsky's third criticism (the difficulty of deciding the grammaticality of a sentence hitherto unattested in a corpus) would be that a new sentence could be compared to patterns of sentences in the corpus. It is important to remember that novelty at the syntactic level is unusual. Thus, some of the points in Chomsky's (1957) critique of corpora could have been answered at

the time but such was his authority that an embargo was placed on corpora for most linguists for many years. Gradually, some members of the community of linguists have become aware of the prestidigitation involved in some of his arguments (Harris 1993; Hoey 2002).

Chomsky was successful in giving corpora a bad academic name for some years but even in the 1960s there were dissident voices. Kučera and Francis (1967) produced the first computerized corpus. The *Brown* corpus, as it came to be known, comprised one million words, made up of 500 written text samples of 2,000 words each, randomly sampled with fifteen text-type categories (or genres) from material, published in the United States in 1961 (Kučera and Francis 1967). The one-million-word Lancaster-Oslo-Bergen corpus (*LOB* 1980) sampled written British English from 1961 using the same methodology. It is interesting to note that the list of the 50 most common words in the *Brown* corpus is almost exactly replicated in the 50 most common words in the *LOB* corpus. In both corpora there were c. 50,000 word types of which c. 50 per cent occur only once in the corpus. Both these corpora are still of great utility for researchers today (and many of the studies referred to in this chapter draw on *LOB*). As Leech remarks, 'computer corpora have promising applications totally unforeseen by their early compilers' (Leech 1991b: 80).

To a considerable extent, the massive increase in the processing and storage capacity of computers, as predicted by Moore's Law (Cringely 1996), and the vast corpora that result, have answered Chomsky's objections to the partial and skewed nature of corpus studies. Moore's law is the empirical observation that, at our rate of technological development, the complexity of an integrated circuit with respect to minimum component cost will double in about 24 months. This increment in processing power has enabled corpus linguistics to expand rapidly as an academic discipline. There has been a steady increase in the versatility and sophistication of available corpora and purpose-built text retrieval software available to the corpus linguist. The *British National Corpus (BNC)*, completed in 1994, could be viewed as a direct descendant of the *LOB* corpus and purports to give a balanced picture of British English in the 1990s. If this 100,000,000-word corpus (with 90% written and 10% transcribed spoken text) could be viewed as belonging to the second generation of corpora, then *The Bank of English*, which stands at more than 500 million running words, might belong to a third generation. The creators of this last corpus

(Sinclair 1995) prefer to call it a collection of texts because, unlike the aforementioned *Brown*, *LOB* and *BNC* corpora, it is not considered complete, but is being continually added to and some texts which are considered in some way dated are being removed. *The Bank of English Collection* is growing inexorably to a new benchmark of 1,000,000,000 words.

It should be added that a recent trend towards smaller and more specialized corpora is evidenced in the presentations at the Corpus Linguistics 2003 conference held on March 28-31, at the University of Lancaster (Archer et al. 2003). Examples of these 'lighter', more focused projects are the various *International Corpus of English* corpora of regional varieties of English (e.g. the Irish *ICE*) with one million words, or what might be termed ESP or EAP corpora such as the *Michigan Corpus of Academic Spoken English (MICASE)* with 1.7 million words (for information about all the corpora mentioned in the text see Appendix 1). Another smaller corpus, which proved very useful in this research, is an offspring of the *BNC*, the *BNC Baby* (Berglund and Wynne 2005). This specialized corpus contains four one-million word samples from the larger *BNC* representing conversation, fiction, newsprint and academic writing. Words and phrases can be searched for and their relative frequency in the four styles of English is clearly shown in a pie or bar chart.

Despite the misgivings of some linguists about corpus linguistics, which have been examined in this section, this new computerized way of analysing large amounts of naturally-occurring language frees linguists 'from drudgery and empowers them to focus their creative energies on doing what machines cannot do' (Rundell and Stock 1992:14). The interplay between the intuition of the linguist and the heuristic power of automated language analysis promises significant advances in our understanding of language in use.

There is a cline of approaches available to the contemporary corpus linguist, ranging from fully automated to fully 'manual' ones. At each stage of the research, the choices made by the investigator may profoundly affect the research path and outcomes. Improvement in corpus linguistic software is such that, soon, corpus linguists may obtain exactly what they need to carry out their research. To this end, computer scientists are already building in features which might 'assist in finding research questions' (Rayson 2003) and are trying to establish exactly what kinds of operations linguists wish to carry out. A good example of this new generation of research tools is Wmatrix (Rayson 2003), a

multifunctional corpus analysis tool capable of annotating large corpora with part of speech (POS) and semantic tags and comparing the relative overuse and underuse of words, strings and tags in corpora. Some use of the Wmatrix software is made in the research for this thesis where appropriate.

The student of corpus linguistics, when consulting the literature, needs to be aware of the technological horizons prevailing at each stage in the historical development of the sub-discipline. Nowadays, even a supposedly manual approach uses the computer to select and print out the essays to be examined. (Interestingly, a word for 'not using computers in any way' does not yet seem to have been agreed upon by English speakers). Corpus linguists can put their hypotheses to the test very soon after they formulate them. The efficiency of computers in corpus linguistics allows initial hypotheses to be tested and then reformulated, thus allowing more real-time flexibility in the analysis carried out.

Corpus linguistics can inform theoretical linguistics. It has won wider acceptance in the academic world in the years during which this thesis was written. Some writers on corpus linguistics, (e.g. Hunston and Francis 2000; McEnery and Wilson 1996) anticipate a major breakthrough in language description thanks to the greater quantity and quality of linguistic information now available through advances in computing. Corpus linguistics has already provided exciting new descriptions of the English language, in particular Biber et al.'s (1999) *The Longman Grammar of Spoken and Written English* and Carter and McCarthy's (2006) *Cambridge Grammar of English.* Its techniques are available to teachers and students as well as to researchers.

Using a corpus approach to language study, specific genres and text types can be focused upon and new categorizations and distinctions tried out. With the ready availability of corpora and concordancing software in the classroom, studies of linguistic features can be carried out with ease, hypotheses can be formulated and tested and palpable progress made. A new kind of empirically based knowledge becomes attainable and students of language no longer have to rely on book knowledge. A possible criticism of corpus linguistics might be that taking naturally occurring language from its milieu and computerizing it destroys its vitality and communicative power. This is an extreme version of the Observer's Paradox. Does a node word or phrase in a concordance line with a span of four, five or ten words on either side of it retain enough of its original co-text to make it fully intelligible to the peruser of the concordance page or screen?

In response it could be argued that, in most central and typical cases, a concordance with a span of four words is enough to reveal the most important collocates of the node word. Moreover, the charge of decontextualization is overstated because current text retrieval software can supply the full co-text from which each concordance line is taken. In seconds, a great deal of sociolinguistic contextual information can be called up, via a header or from a separate file, for all texts in the corpus: e.g. the gender, age-group, social class, dialect and educational level of each speaker or writer is registered together with the size of the audience or readership. The advantage of corpus linguistics is not simply that without it all the counting of word clusters, adverbs, adjectives, verbs, nouns and prepositions would have to be done by hand. Probably, in the absence of corpus methodology, such comparative measurement is unlikely to have been undertaken. The researcher, teacher, or student (I embody all three) is enabled to ask different questions and swiftly obtain contrastive information along multiple dimensions. As more is learnt about particular language varieties, the potential for intervarietal study grows.

If corpus linguistics is more a methodology for doing linguistics than a substantive area of study, as Leech (1998) avers, then charting the way becomes at least as important as reaching destinations. Given that the process of discovery is of central importance, each step in corpus linguistics research needs to be well-documented.

2.2 The hard and soft approaches to corpus linguistics

The Bank of English Collection grew out of the *Collins Birmingham University International Language Database (COBUILD)* project. The director of both projects, Sinclair, proposed a distinction between 'corpus-based' and 'corpus-driven' linguistics (from Sinclair 1996 as reported in Ooi 1998). This section examines the contrast between these two different approaches to corpora. Basically, according to the terminological distinction, corpus-based linguistics is traditional linguistics carried on with corpora while corpus-driven linguistics is a non-theory-laden approach to language without preconceived ideas which starts from scratch and follows where the data lead. Figure 2.1 is taken from Ooi (1998: 51) and provides a more schematic analysis of the distinction between the two

approaches. Corpus-driven linguistics is depicted as a bold approach to the study of language which is deliberately iconoclastic and innovative.

Figure 2.1 Two approaches to corpus linguistics compared

CORPUS- 'BASED' LINGUISTICS	CORPUS- 'DRIVEN' LINGUISTICS
- a corpus is used to validate, check, and improve linguistic observations that have already been made; the corpus-based linguist does not feel 'threatened' by corpus at all.	- corpus is of prime importance in bringing out new ideas for the examination of data.
- the linguist does not question received theoretical positions or well-established descriptive categories; instead, his position on language structure is already well-formed.	- the linguist believes that the kind of evidence emerging from corpora may be difficult to reconcile with established positions in the discipline, and he leaves open the possibility of the need for a radical change in linguistic theory in order to cope with the evidence.
-the corpus is used to help extend and improve linguistic description.	-evidence from the corpus is paramount, therefore the linguist makes as few assumptions as possible about the nature of the theoretical and descriptive categories.
-An example of a relevant question: "Is *whom* still used in English, and if so how?"	-An example of a relevant question: "Is the distinction between grammar and lexis necessary?"

Ooi (1998:52) suggests that corpus-based and corpus-driven linguistics could be described respectively as 'top-down' and 'bottom-up' approaches to the analysis of lexical data. The corpus-driven approach allegedly espouses a radical philosophy of science. The mainframe and personal computer are seen as observation instruments capable of revolutionizing the science of linguistics as the telescope and the microscope revolutionized astronomy and medicine (Stubbs 1996: 231). This image of the computer reflects the pivotal role attributed to scientific tools in a recent encyclopaedia of the history of science: 'Scientific instruments have governed the rate at which science advances' (Bud and Warner 1999: 234). Sinclair and Renouf point out the relentless efficiency of the computer when used to observe language:

> Retrieval systems, unlike human beings, miss nothing if properly instructed: – no usage can be overlooked because it is too ordinary or too familiar...The human being, contrary to popular belief, is not well

> organized for isolating consciously what is central and typical in the language; anything unusual is sharply perceived, but the humdrum everyday events are appreciated subliminally.
>
> (Sinclair and Renouf (1988:151)

An interesting extension of this metaphor of linguistics as a new way of viewing language is employed by De Beaugrande:

> English… can… be said to be based upon extensive and delicate reserves of colligability and collocability. If … not, people would encounter harrowing difficulties in deciding what to say. Any word could be combined equally well with any other word, and the language would be completely flat and uniform, like a featureless desert of sand particles. But…the English language is far more similar to an astonishingly diverse and well-tended landscape; and the closer you look at real language, the more delicate details you will find, much like a landscape viewed from the window as the airliner descends from the higher altitudes down towards the ground.
>
> (De Beaugrande 2001:12)

The corpus-driven paradigm invites the linguist to look afresh at words and their patterns in concordances and texts in order to devise new *a posteriori* explanatory models which not only cover the general patterns and tendencies which can be discerned but also attempt to encompass every irregularity or localized idiosyncrasy of language that is uncovered. Of course, corpus-driven linguists do not have to jettison all their theories and models but those theoretical assumptions they are working with should be declared from the outset. Corpus-driven linguists maintain that they are prepared to reject or revise any theoretical assumption if it is falsified by the data. In this respect, Kennedy (1992) warns that the corpus linguist should strive to maintain an open mind because:

> Language description can be challenged by corpus research. It should not be assumed that the types identified in traditional descriptions should be the only things which are quantified.
>
> (Kennedy 1992: 367)

The linguist seems to be confronted by a stark choice between a 'soft' scientific approach, i.e. corpus-based linguistics, and a 'hard' scientific approach, namely corpus-driven linguistics. Perhaps the strongest statement of this stark choice is found in two

chapters of Tognini-Bonelli's (2001) influential book, *Corpus Linguistics at Work*. In Chapter 4 of this work, Tognini-Bonelli characterises corpus-based linguistics and in Chapter 5, she gives an account of corpus-driven linguistics. The two cogently written chapters contain highly persuasive writing. The author depicts corpus-based linguists as intellectual conservatives using corpus data to consolidate their preconceived ideas. On the corpus-driven side, there is courage, openness to the data, a preparedness to jettison concepts and theories which are falsified by the empirical data. Although this reading of the argumentation might be denounced as caricature, it is perhaps not going too far to suggest that she is accusing corpus-based linguists of wishful thinking or, worse still, 'cooking the books'. I use this personalised unacademic language advisedly because Chapter 4 and Chapter 5 of Tognini-Bonelli's work as a superbly constructed *argumentum ad hominem*. This drawing up of the two ways of doing corpus linguistics appears to be a subtle form of surreptitious legislation. Either corpus linguistics is done one way or the other. The argumentation is reminiscent of the famous 'armchair linguist' debate of Fillmore (1992), where he lambasts those who theorize about language from their armchairs at home instead of venturing outdoors in search of data. Although I devised the title of my thesis before the appearance of Tognini-Bonelli's book, I do not feel constrained to change it. Her corpus-based linguist could be viewed as a 'Straw Man' and her corpus-driven linguist as an unattainable idealisation.

Sinclair (2003) gives a more nuanced and conciliatory description of the utility of corpora in language studies in the introduction to his work *Reading Concordances:*

> The amount of variation in actual usage makes accurate generalisation rather difficult. The difference is often said to be between "top down" and "bottom-up" approaches; starting from the "top" it is extremely difficult to arrive at a description that fits the facts of usage, while starting from the "bottom" it is not easy to formulate sufficiently general statements. However, the experience and intuition of the researcher are available in both approaches, and so the so-called "bottom-up" approach, properly conducted, is really a two-pronged attack on the data from the top and bottom simultaneously.
>
> (Sinclair 2003)

The bottom-up approach starts with words and moves up to sequences of words and finally texts. In Chapter 4, this is supplemented by cautious speculation in order to attempt to

explain the phenomena observed. In the case of learner data there may be multiple causal factors – no single candidate explanation may be adequate. The researcher has to excogitate at times about what is going on inside the head of the learner or what learners need to know at a given stage in their L2 development to produce the linguistic forms they do.

For the rest of this thesis, 'corpus-based' is used as a superordinate term for investigative activities which are conducted with, alternately, a more corpus-based or a more corpus-driven methodology. The next section describes some of the more interesting findings of corpus linguists and discusses their relevance to research into prefabs.

2.3 Achievements due to the corpus revolution

COBUILD and its successor, *The Bank of English*, *The British National Corpus*, and several other large scale corpus projects have produced an impressive range of grammar and usage books for learners of English, the 'Big 5' UK English learner dictionaries, and some coursebooks for learning English: perhaps most notably to date *The COBUILD English Course* (Willis and Willis 1988) and *The COBUILD English Grammar* (2002). Some previously unsuspected aspects of language have already begun to emerge as a result of the computer analysis of large amounts of text. An example of this is the positive or negative semantic prosody of certain words. Words like *cause* (Stubbs 1995) or *happen* (Sinclair 1991) usually collocate with words associated with unpleasant or negative events or states of affairs. The remarks made by Benson (1986a) about the usual negative collocations of *cause* now seem prophetic:

> The verb to cause and some of its synonyms, combine freely with hundreds of nouns (which often have a negative meaning)...Let us cite a few examples showing nouns beginning with the letter d: to cause damage, danger, deafness, a death, a debacle, decay, decompression, defeat, a defect, a deficiency, deflation, a deformity, degeneracy, dehydration, dejection, delay, delinquency, delirium, delusion, etc.
>
> (Benson, Ilson and Benson 1986a:257)

Some other words like *provide* have a more positive semantic prosody (compare 'to cause work' and 'to provide work' (Stubbs 1996:174)). Tribble (1999) found that the words *experience* and *international* had specific local semantic prosodies in his corpus of English

for management consultants and collocated only with words with positive meanings, unlike the same two words in a general corpus. The study of semantic prosody using computer corpora, developed by Louw (1993) and extended by Hoey (1997), provides the means to discern patterns of a type which may have been suspected by literary critics but can now be identified in empirical detail. Hoey identifies four semantic prosodies for *consequence* and gives the following breakdown of the collocations he found in the *BNC* corpus:

> the logic of underlying processes - 56% (inevitable, inexorable, likely, probable)
> the badness of an outcome - 15% (dire, appalling, regrettable)
> the seriousness of an outcome -11% (important, decisive)
> the expectedness or otherwise of an outcome - 9% (unintended, odd
>
> (Hoey 1997:3)

More recently Hunston (2002) has shown that semantic prosody is much more widespread than even its discoverers originally suspected. For her, semantic prosody is a kind of collocation but involves relations between whole groups of words of a particular type. This insight builds on the work on pattern grammar produced by Hunston in collaboration with Francis (2000). The *COBUILD* project was fortunate in having not only excellent lexicographers and computer experts but also innovative grammarians. Hunston and Francis (2000) questioned Melčuk's (1998) notion of an area of language which is 'fixed' and a larger area which is free:

> Collections of lexical phrases are, ultimately, fairly random lists of phrases, organised either according to their relative fixedness, or to their function (discourse-organising, opinion-giving and so on) or to one of their keywords. They are an attempt to account for a portion only of the lexicon. Grammar patterns, on the other hand, constitute an attempt to describe the whole of the language (or most frequent items in the language).
>
> (Hunston and Francis 2000)

In suggesting that lexicographic work which concentrates on set phrases or phrasemes (Melčuk's term) lacks rigour and comprehensiveness, Hunston and Francis, in this passage, level strong criticism against a great deal of work in phraseology. They recommend the study of patterns and suggest that prefabs could be viewed as particularly

congealed patterns. Given that my research focuses on a small portion of the lexicon and a minor subset of patterns or component parts of language patterns, the question arises as to whether, in learning a foreign language, focusing on patterns might have a higher surrender value (i.e. yield greater utility for the time invested in learning them) than a prefab approach. Alternatively, using prefabs might provide a short cut to creating spoken and written texts. In my view, if it is found that EAP students are already using prefabs, their attention should be drawn to the overuse of certain terms and strategies provided for elegant variation. On the other hand, a longer-term pedagogical aim might be to situate teaching in the larger context of phraseology and verb complementation.

I have come to the conclusion that Bolinger's (1975) 'unfreedom' (referred to in Chapter 1.4 above) might be a larger area than hitherto suspected. Howarth's continuum from 'free' to 'restricted' (Section 1.5 above), and the fact that the freedom of many of his free examples could be questioned (*under the table*, *on the table*), raises the question whether there can be a continuum without two poles. Hunston (2002) draws attention to a similar problem in delineating the area of operation of Sinclair's (1991) 'open-choice' principle. The operation of this principle can only be established negatively: by the absence of prefabs. In the 'manual' search for prefabs, described in the next chapter, the truth of Hunston's observation was continually confirmed. Interpreted psycholinguistically, the open choice principle might prove to be superfluous if language is produced by combining chunks and single words which are stored together in the mental lexicon: one concatenative process which chooses now a single word now a prefab according to the processing, semantic, syntactic and stylistic requirements within the emergent discourse. Alternatively, the choice principle could be viewed as operating at a higher level, selecting and combining words and phrases into texts.

Hunston and Francis (2000), in their investigation of pattern grammar, postulated that a subset of graded adjectives which have a number of patterns of their own are candidates for 'class-hood'. They labelled this class 'evaluative adjectives'. One pattern involving evaluative adjectives that they studied is the pattern *it* followed by the verb *be* (or another link verb), followed by an adjective or adjective group and a *that*-clause. Typical exemplars of this class of adjectives which Hunston and Francis isolated from the *COBUILD* corpus are: *apparent, appropriate, arguable, awful, bad, clear, essential, evident, extraordinary, fair, funny, good, important, improbable, inevitable, interesting,*

likely, lucky, natural, necessary, obvious, plain, possible, probable, reassuring, sad, true, and *unlikely.*

These adjectives fall into groups according to their meaning but all the meanings involved are within evaluative scales such as *good/bad, easy/difficult, probable/impossible,* and so on. Another pattern within which evaluative adjectives are found is *it* + *be* (or other link verb), adjective or adjective group + to-infinitive clause:

> It's more expensive to live alone
> It is important to check the success of a university's graduates on the job market.

So powerful are these patterns that an adjective has only to appear within one of them to be identified as evaluative. Hunston and Francis provide an interesting example of the way in which the pattern carries its own meaning which makes the unusual choice, *shimmering,* interpretable in the following extract:

> In this climate of success-driven theatre, it is shimmering to find work that reflects such passion.
>
> (Hunston and Francis 2000:191)

The explanation which Hunston and Francis give of this phenomenon ends with a resonant aphorism:

> We have a mental stereotype of this *it* pattern and we know the common currency of evaluation from which this use of *shimmering* deviates. The pattern is basically a chunk of meaning, involving the co-selection of items in a predictable way. But this example shows the productivity of such patterns: we always have recourse to the paradigms of the possible as an alternative to relying on the syntagms of the typical.
>
> (Hunston and Francis 2000:191)

This particular section of *Pattern Grammar* (Hunston and Francis 2000) covers much the same ground as Lemke (1999). Section 3.6 of the next chapter will examine examples of the use, in the four corpora, of this pattern containing the evaluative adjectives.

If words and phrases in texts display certain patterns of behaviour, occurring in some contexts but not in others, this is a question of collocation i.e. a relationship between a word or phrase and other words. But the context can refer instead to the grammatical behaviour

of a word – the grammar patterns that a word prefers, e.g. tending to occur at the beginning or end of the sentence, functioning as the subject or object of the sentence and so on. Firth (1957) coined the term *colligation* to refer to these grammatical patterns. Hoey revived this important concept and defined it, using an interesting degree of anthropomorphism, 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 1998)

Whereas collocations occur within a text, colligations are intertextual. According to Hunston:

> The implied meanings and clause patterns are associated, not with individual lexical items alone, but with phraseologies, that is with a lexical word in conjunction with one or more grammatical words (as in *on the downside*) or with a sequence of grammatical words (as in *I may not be a*)...the meaning of this one text depends upon thousands of other texts, and the repeated patterns that are found in them
> (Hunston 2002:31)

The term 'implied meanings' in this quotation refers in particular to semantic prosody which, together with clause patterns or colligations, pervades the language. I have gone into detail with this aspect of Hunston and Francis's (2000) work not only because it illustrates a new way of describing language, but also because their ideas about evaluative adjectives are used in a case study which is reported in Chapter 4.

The written part of the *BNC* (90% of the total) is approximately 30% print journalism, while the Bank of English at mid-1996 was 68% journalism: 180 million words of print journalism and 40 million words of transcribed radio broadcasts (Rundell 1998). The preponderance of journalism in two of the major current corpora might suggest the continued validity of Chomsky's claim, that a language corpus can only represent a small sample of a large and potentially infinite population and is, therefore, skewed and unrepresentative. A similar criticism can be levelled, however, against any scientific endeavour which uses sampling techniques rather than investigating an entire and finite population – in other words, a great deal, if not most of the scientific and social scientific

research being carried out today. In favour of newsprint it should be added that, although newspapers have their own style, they are a good source of general information about language change as they incorporate changes more quickly than other kinds of discourse do (Hundt and Mair 1999). A corpus comprising only newsprint (e.g. the *Reuter* corpus, see Appendix 1):

> sacrifices 'what is desirable' (adding to a general corpus from all of its component registers every year) to 'what is feasible' (developing a corpus as restricted in terms of register but expansive in size and in currency).
>
> (Hunston 2002:31 based on personal communication with Wolfgang Teubert).

Recent advances in computer hardware and software and the concomitant increase in the size and availability of corpora have resulted in quantitative increases which are so great that they can be regarded as a qualitative shift (Clear 1993: 274). It is not simply that much more data is available, but it is increasingly of a different kind.

Chomsky's objection to performance data, that they are affected by 'memory limitations, distractions, shifts of attention and interest and errors' (Chomsky 1965:3), is not valid for many corpus studies because they look at recurrent collocations and co-selection of lexis and syntax. The data consists of the language performance of many different people and the resultant corpus is examinable by many scholars. Normally, the focus is on what is central and typical and the idiosyncratic or unique is automatically filtered out. For example, Kjellmer (1984) and Sinclair (1991) discount hapax legomena and collocations which occur only once, while Clear (1993:277) discards collocations which have been observed fewer than three times.

The basic issue is the nature of the data which should be used to inform linguistic theory. A major distinction between Chomskyan linguistics and corpus linguistics is that the former relies on artificially induced observations (such as grammaticality judgments) or introspections and the latter uses naturally occurring data. Moreover, Kjellmer (1991) points out that the corpus is not the only tool of the corpus linguist, who will also use elicitation and introspection techniques. McEnery and Wilson (1996) situate themselves at the midpoint between rationalism and empiricism when they state that the corpus linguist

relies on theoretical insight to explain empirical data. Hunston (2002: 23) observes 'the corpus simply offers the researcher plenty of examples; only intuition can interpret them'. Another defence of corpus linguistics is provided by Partington (1998) when he suggests a corpus is neither 'performance' nor 'competence' but supplants the differentiation between the two concepts. The concept of competence, the ideal speaker's knowledge of the language, raises certain philosophical problems because claims about 'ideal knowledge' are not falsifiable by any evidence. The individuality of performance needs to be transcended:

> Information about particular communicative events by itself is of limited, we might even say purely anecdotal value. By and large, we are not methodologically justified in interpreting the significance of a particular linguistic event unless we can compare it with other similar events. The corpus can provide 'background information' against which particular events can be seen.
>
> (Partington 1998:146)

This argument is similar to the rejoinder to Chomsky's criticism of the corpus's inability to deal with novel utterances (Section 2.1 above). At that point, it was suggested that most utterances are made up of a rich variety of lexis and word combinations inserted in a limited number of phrase structures with syntactic innovation being highly unusual. Nevertheless some theory is still needed to mediate between the particular linguistic event and the generality of patterns.

2.4 Computer Learner Corpora

After more than two decades of corpus-based/driven research on native speaker English and on an ever-increasing number of other languages, projects were conceived, apparently in several places at more or less the same time, of compiling corpora of learner English. This area of study, Computer Learner Corpora (CLC), after a slow start, is now flourishing. Granger (2003) posits a definition of CLC based on Sinclair's (1996a) definition of corpora (in general) for the EAGLES standard-setting committee:

> Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular

> SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance.
>
> (Granger 2003: 7)

Leech (1998:xvi) speculates that the delay in the rise of learner corpora investigation in the early 1990s following the earlier commencement of studies on mother tongue corpora in the 1960s may have been due to two principal factors, one practical and the other theoretical. In the practical sphere, Leech suggests that education tended to be the 'Cinderella' of the academic world, and consequently funding for computer equipment and research time was not generally available, especially for language teaching. The second, more theoretical reason he gives for the slowness to apply corpus methodology to student output was the unfavourable intellectual climate for investigating learner data. On the one hand, the dominant approach of the time, the so-called 'communicative approach', emphasized process rather than product. By their very nature, corpora contain only products, what has been said or written. On the other hand, Second Language Acquisition (SLA) scholars were interested in the acquisition of competence which takes place in the human mind and is only indirectly related to the language learner's observable performance. These investigators saw little value in studying performance data. Leech (1998: xvi) also suggests that the initial resistance to CLC might have been partially ascribable to the negative image which remained from the Error Analysis studies of the 1960s and 1970s.

Despite the methodological faults which were rightly attributed to Error Analysis, CLC researchers stress that examining learners' errors can tell us a great deal about what different learners find easy and difficult at different stages of their acquisition of a foreign language. Teachers and materials designers need to know what a learner could be expected to know by a certain stage in their interlanguage development. This knowledge can contribute greatly in evaluation, course design and in the production of suitable materials. The original Error Analysis (EA) considered errors to be a totally negative aspect of learner language and did not pay sufficient attention to what the learners were getting right or which structures and lexis they were avoiding. Obviously, avoidance does not lead to error but to under-representation and over-representation of words or structures in learners' use of the target language. This undue emphasis on error had an adverse effect on language

teaching, with teachers focusing excessively on learners' idiosyncrasies. The early EA practitioners extracted the errors from the surrounding text, which was stripped away and discarded, and then made collections of de-contextualized errors. CLC is able to show the full context of each error and call up the other instances of this error as well as the contexts in which the student uses the structure, rule or word correctly. Thus, we get a much fuller picture of each learner's performance. Several large learner corpus initiatives were undertaken in different parts of the world, starting in the late 1980s. The creation of the Longman Corpus of Learner English (*LLC*) began in 1987 and grew to 10 million words by 1997. This corpus contains samples of different genres of written English (examination answers, letters, reports, diaries and student essays) from learners with 8 different levels of proficiency from more than 160 different language backgrounds. It is designed for academic and lexicographical research and to support the preparation of language teaching materials. The Hong Kong University of Science and Technology (*HKUST*) Corpus (Milton and Tong 1991) contains 5 million words of English written by mainly Cantonese-speaking learners of English. This corpus, which is annotated with grammatical and discourse tags, is, to date, the largest collection of writings by a single learner group. One of the main uses the compilers of the *HKUST* corpus had in mind was the development of teaching materials and the production of more interactive and creative CALL programs.

All five of the major EFL dictionaries in use throughout the world today are based firmly on major corpora. *The Longman Dictionary of Contemporary English* is based on the purpose-built 30-million-word Longman Lancaster Corpus. The *COBUILD English Dictionary* is sustained by the 600-million+ *Bank of English* monitor corpus. The *Cambridge Advanced Learner's Dictionary* draws on the 16-million-word Cambridge Learner Corpus. *The Oxford Advanced Learners' Dictionary* and *Macmillan English Dictionary for Advanced Learners* (2002) have their empirical foundation in the *BNC*. Given that all five dictionaries are based on native corpora, CLC data appear less significant in comparison. Nonetheless, *The Longman Activator* (1993) was the first learner's production dictionary to incorporate CLC data, and the *Longman Essential Activator* (1997) was greatly influenced in its revolutionary lexicographical approach by the use of learner data from the Longman Learner Corpus. As studies continued, CLC data gradually received the attention it deserved and the various international projects could be said to have reached critical mass. Indeed, so rapid has been the adoption of CLC

approaches in applied linguistics and lexicography that one of the leading practitioners of learner corpus compilation and investigation, Granger, has argued that in the event of there being a conflict between insights gained from a NS and a learner corpus she would give precedence to the evidence yielded by the CLC corpus (Granger 2003: 24).

2.4.1 The ICLE project

A pioneering project in the collection of learners' writing and perhaps the one which has engendered most international collaboration is the International Corpus of Learner English (*ICLE*), based at the Catholic University of Louvain. This has over 2 million words written by advanced learners of English from 11 mother tongue groups (Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish and Swedish). The ICLE corpus is increasing as further sub-corpora are being compiled (Brazilian, Chinese, Japanese, Norwegian, and Portuguese sub-corpora are under construction). The decision by ICLE to include complete texts from each contributor is an important one. The final sub-corpus in this list, the Portuguese, is compiled by me and is the focus of my thesis investigation. This sub-corpus I have named *Porticle* and, when completed, it will contain 200,000 words. The *Porticle* sub-corpus is designed to represent the EAP writing of Portuguese undergraduates in their final two years of study and to allow comparison with writing by the other groups of learners represented in *ICLE*.

One of the main imbalances in corpus linguistics to date has been the under-representation of the spoken language. This is mainly due to the expense of recording and then transcribing accurately what is said. There is a need for trained transcribers and standardised conventions of transcription and all this labour-intensive work is necessary before the recording takes place. In native-speaker corpora studies, this situation is improving: where the *BNC* has 10% spoken English, the ICE corpora have 60% spoken language.

To make good this lack, The Centre for English Corpus Linguistics at Louvain, in 1995, also began compiling a spoken corpus of informal interviews with learners from a variety of mother tongue backgrounds. This corpus is called the Louvain International Database of Spoken English Interlanguage (*LINDSEI*). The first component of *LINDSEI* contains transcripts of 50 interviews (30 female subjects, 20 male subjects) with French

mother tongue learners of English (c. 100,000 words of learner language) and research has already begun on the phraseology of this type of interlanguage (De Cock et al. 1998). A number of other components are currently being compiled for different mother tongue backgrounds (to date Japanese, Swedish, Spanish, Italian, Bulgarian and Chinese are represented). Alongside these non-native varieties of English, a comparable corpus of interviews with native speakers of English has been compiled, so that interlanguage and native language can be compared and the universal and the L1-specific features of oral interlanguage identified. The cooperative, decentralized nature of *ICLE* and *LINDSEI* means that projects which would be financially unthinkable for one university can be undertaken simultaneously by local teams of academics in different parts of the world, with Louvain acting as the clearing house.

Relative newcomers like myself can learn from investigations of the *ICLE* sub-corpora, as reported in e.g. Granger (ed.) (1998), Granger, Hung and Petch-Tyson (eds.) (2003). The contrastive interlanguage analysis (CIA) approach asks which aspects of learner English can be ascribed to cross-linguistic transfer from the writer's mother tongue and which are shared by all or many *ICLE* sub-corpora.

The investigation of adjective intensification in German advanced learners' writing by Lorenz (1998: 53-66) is an interesting and well-wrought work within the *ICLE* tradition. Comparing the characteristic modifier-adjective associations of German learners and native speakers in corpora of argumentative prose, Lorenz suggests that German learners of English show a tendency to overuse particular modifiers and a propensity to engage in hyperbole. Granger found a highly significant ($p \leq 0.01$) overuse of *very*, which she styles 'the all-round amplifier *par excellence*', among French learners (Granger, 1998b: 151). Section 3.4.8 of this thesis contains a report of an investigation of adjective modification found in *Porticle*, the sub-corpus representing the writing of Portuguese students of EAP.

The words 'under-use' and 'overuse' will often be used in this thesis and it is important to establish from the outset that such assessments of learner English are intended to be purely descriptive: there will be many occasions in the coming pages where it is averred that lexical items or function words appear more or less frequently in the learner corpus compared to a native control corpus. Although this discussion of overuse and underuse of words, phrases, grammatical structures is statistically based and tries to be non-prescriptive, the question arises as to how much freedom NNSs have to transgress the

conventions of NS English. This question, in turn, leads into the debate about the 'ownership' of the English language. The variety of English contained in the *Porticle* learner essays might be classified as English as an international language (EIL) or English as a Lingua Franca (ELF) (Seidlhofer 2001). Hunston and Francis (2000) consider the pedagogical usefulness of a corpus of English as an international language and the specifications according to which it might be constructed. They ask whether native speaker contributions might be included in such a corpus. They predict that, just like NS varieties of English, EIL would exhibit patterns like the ones they discovered in NS varieties of English and recorded in their book, *Pattern Grammar* (2000). This question is discussed in Chapter 5.3.

2.4.2 The *Porticle* research project

I joined the International Corpus of Learner English project in January 2001 by volunteering to compile the Portuguese sub-corpus, which became known as *Porticle*. This was a *quid pro quo* arrangement whereby, in exchange for the contribution of my sub-corpus, I gained access to large amounts of learner and native-speaker data. One attractive feature of this international network of learner corpus compilers is that, although a standardized procedure is pre-established for corpus content, structure and presentation, each compiling team is free, indeed encouraged, to pursue their own research agenda. The *ICLE* community, therefore, provides the CLC investigator with a discourse community with its own conferences, publications, research results and hypotheses to be corroborated or refuted.

Two Portuguese sub-corpora are being compiled: *Porticle* in Portugal and *Bricle* at the Catholic University of Sao Paolo in Brazil. Each national team of compilers can compare their sub-corpus with the others to see which features it shares and where and how it differs from the others. *Porticle* provides the central data of the investigative part of this thesis, or in more traditional terms, the study sample. This collection of argumentative essays written by Portuguese advanced learners of English will appear on a CD-ROM together with at least another 15 mother tongue sub-corpora as the third edition of the ICLE corpus.

2.5 Approaches to extracting prefabs from corpora

In the trawl for prefabs through my corpora, two separate approaches were employed. These two approaches could be typified as, on the one hand, a manual search of small samples from two of the corpora, and on the other hand, an automated search for prefabs followed by the application of various filters to the computer output. The first approach I designated as 'manual' to show that a computer was not used and that the analysis was carried out by a trained human reader.

When human judges of prefabs are used, wide variations have been found in their estimates of the number of prefabs in text. These estimates ranged from seven to 70 in one experiment (Willis 1998 and in personal communication). I carried out an experiment using an argumentative essay written by a Portuguese EAP student. I invited subscribers to the CORPORA distribution list to assess the number of prefabs contained in the 500-word essay. The results obtained from 30 respondents ranged from 10 to 75 prefabs. Confronted with these effects of the subjectivity of human assessors, text retrieval software such as Wordsmith Tools (Scott 1999) offers the corpus linguist a degree of objectivity by showing all the recurrent clusters in a text or corpus. Then, after the manner of Altenberg (1993), the resultant recurrent word clusters can be culled of those sequences which are 'phraseologically uninteresting'. For Altenberg those sequences which are of little phraseological interest are mere repetitions or fragments of larger structures (e.g. *the the, and the, in a, out of the*). He employs an essentially grammatical criterion of 'interestingness'. The resultant set of fixed expressions can be inventoried using the presence and absence of some clearly defined features. The physical appearance of prefabs, e.g. their well-formedness, is largely determined by the definition of prefab used to identify them. Altenberg would probably reject many of what are currently referred to as lexical 'bundles' because they are, in his terms, grammatically incomplete. For example, many of the statistically generated lexical bundles of Biber et al (1999) have a fragmented look, beginning or ending as they do with articles, prepositions, or particles. Some examples of these bundles which Biber et al. found in academic prose are:

the other end of the
the surface of the
the first part of the
in the process of

(Biber et al. 1999:1017)

A closer inspection of these longer bundles shows that they contain collocational frames (see Chapter 1.9). Arguably the nucleus of each of these bundles is the frame with optional adjectives inserted (*the first part of* and *the other end of*) and with an article or preposition added at the beginning or end of the sequence.

My own position on lexical bundles is that they appear hybrid, usually containing elements from several phrases and, therefore, lack semantic unity. Although they recur significantly often in corpora, they are not usually picked out by raters of prefabs because they do not stand out from the surrounding text as integral units and, for this reason, are not particularly memorizable.

In the first manual approach, I worked with a sample of essays from each corpus which was small enough to be assessed for prefabs by a human judge. In this small-scale assessment the presence of each and every prefab was verified and each specimen prefab was then classified as either lexical, grammatical, or pragmatic. It was then possible to calculate the percentage of each sample made up of the different kinds of prefab.

In the second, computerised approach, all the recurrent sequences of words of varying lengths are recorded. Such sequences are called N-grams where N= the number of words in the sequence. Recurrent two-word sequences (bigrams), three-word sequences (trigrams) and four-word sequences (4-grams) are gathered using purpose-built software and specifying a certain *floor* (i.e. minimum threshold to qualify as recurrent). The outcome of the application of this floor is similar to Biber et al.'s (1999) procedure and effectively captures the lexical bundles contained in the corpora. This floor can be varied and the quantities and kinds of prefabs observed. As N increases, the number of N-grams above a certain floor decreases. At the next stage, Hudson's (1998) fixedness criteria (see Section 2.5) were used to separate out those N-grams which are really fixed to some degree or which are, to some degree, conventionalised.

Frequency of occurrence alone does not confer prefab status on a sequence of words. The first and the second word in any text will be a bigram and the second and third word, and so on, producing one bigram less than the total number of words in the text. Thus, if we take any 100,000 words of text, there will be a total of 99,999 bigrams. An N-gram is like a moving window over a text, where N is the number of words in the window.

The kinds of decisions which have to be taken to get from N-grams to prefabs can be illustrated with some examples of bigrams harvested from *Porticle*. The sequence *it is* occurs 423 times and is far and away the most frequent bigram in the sub-corpus. But this frequent recurrence is in no way surprising given that *it* and *is* are in the top 11 most frequent words in the corpus and this sequence forms part of the typical conjugation of the verb *be*. The sequence can prove more interesting, however, when it forms part of an attitudinal stance marker such as the 4-gram 'It is clear that'. The sequence *and to* is third in order of frequency with 75 occurrences and again this is not surprising, given that *and* and *to* are the second and third most frequent words in *Porticle*. This sequence of a conjunction followed by a clause or phrase particle is phraseologically uninteresting in the sense defined by Altenberg (1993). It does not have semantic unity. The bigram *of course* (73 occurrences) combines a very frequent preposition (fourth most frequent word overall) with a less frequent noun *course* (number 115 in order of frequency with 156 occurrences in *Porticle*) to produce a fixed expression: no word could be substituted for *course* without completely changing the meaning; substitutions of the preposition produce similar changes of domain of discourse: *on course, off course, in course*. The phrase *of course* can stand alone and could be used to answer questions or requests affirmatively. *Of course* is employed in academic writing to realize epistemic stance and to remind the reader that something is already known or obvious.

Another similar approach to that outlined in the first part of this section is that adopted by De Cock et al. (1998) in their chapter 'An automated approach to the phrasicon of EFL learners'. Although this was a study of spoken learner and native-speaker English, the methodology can be adopted for the present investigation of written English, with minor changes of emphasis. Most of the caveats and problems of isolating formulaic expressions adduced by Kennedy (1998) are squarely confronted here. De Cock et al. (1998: 67-79) used their own custom-built software to extract N-grams and then applied a filtering procedure.

The first part of this process is essentially formal in nature and consists of

1. the elimination of non-fluencies such as unintentional repetitions and stutterings e.g. *.the the, I I, it was, it was.*

2. the elimination of most combinations of closed-class items that are only phrase or clause fragments; e.g. *in the, and it* and the subjective

elimination of those word combinations which are fragmentary in nature; e.g. *are a lot of, don't know if you.*

3. The next phase of the filtering process is an assessment of whether the remaining candidate formulae have the potential to serve any pragmatic or discourse functions.

4. The final stage examines each occurrence of a potential formula in its co-text.

(De Cock et al. 1998:75)

In this fourth and final stage, candidate formulae which are embedded in other longer formulae were discarded; e.g. *or something* is rejected when it occurs as part of the longer *or something like that.* The candidate *and so on* fails to measure up in the following context: *and so on the twentieth of May there is a festival which is international.* (The four stages and examples are taken from De Cock et al. 1998).

It is to be hoped that, after the application of both the labour-intensive analysis and the corpus-based approach, a reasonable correlation between the results is found. A major difference between the two approaches is that a sequence occurring only once in the corpus would count as a prefab in the manual approach if deemed so by the experimenter but would fail to qualify in the computer-assisted search.

Leibniz (1666) discussed the desirability of a universal language of logic and science which would enable discussants to resolve their differences by translating their propositions into the 'universal character' of symbolic logic and simply calculating or computing a conclusion. Disputants could say 'Calculemus'. Although such a metalanguage of science has yet to emerge, computational linguistics has made various measures of the strength of collocation available to the corpus linguist and in most software packages for concordancing a choice of at least two algorithms is found. Hunston (2002:70-77) gives a careful comparison of two widely used measures of the significance of the collocation of two words: t-score and mutual information. She observes that mutual information or the MI score is more influenced by rare words, but is not particularly dependent on the size of the corpus.

Many of the statistical judgments in this thesis are based on the log likelihood test (sometimes referred to as G^2 or &lambda). Dunning (1993) proposed log likelihood as an alternative to the chi-squared test because it is more reliable with low frequencies and when comparing samples which differ greatly in size. Wordsmith Tools allows the corpus linguist

to choose between the chi-squared test and log likelihood in the calculation of collocational significance or keyness as it is styled in this software package (Scott 1999). For these reasons and also because it is used to compare relative frequencies of occurrence in corpora in the Wmatrix software package, many of the comparisons made in Chapter 3 of this thesis are based on log likelihood. The formula for calculating log-likelihood is given below in Appendix 7.

Another statistical measure which is used in Chapter 3 to process the results of the corpus analysis is the z-test for comparing two independent proportions. The z-test is applied to percentage occurrences of a specific feature in two corpora to assess the likelihood of a genuine difference between the corpora with respect to the feature. The formula used to compute the z -test statistic is given below in Appendix 7. The z-test compares percentages from two samples and allows a decision to be made as to whether there is a difference in the two proportions beyond chance. This is done by setting up the null hypothesis that there is no difference between the two proportions. When the z-test is applied, if the test statistic is greater than the critical value of 1.645, then the null hypothesis can be rejected. It can then be stated that, at the 5% level of significance, there is a significant difference in the two proportions.

Sinclair (2003) points out that the likelihood of two words occurring adjacently in a text cannot simply be calculated by taking the product of the chances of each being in its own position:

> so if two common words each occur on average once in every 500 words of text, the chances of them being next to each other is roughly 1:250,000. For most pairs of words the chances are extremely low. On the other hand, the notion of CO-SELECTION, the simultaneous choice of more than one word at a time, works in the opposite direction, increasing the chances of certain pairs dramatically. As a result we may observe that a two-word phrase, while not as rare as the arithmetical predictions would have it, is still not nearly as common as the words that make it up, and a three-word phrase is even less common
>
> (Sinclair 2003:125)

When statistics are used in this thesis, a certain amount of caution is required. The written corpora used are small samples from much larger populations of writing and so their representativeness needs to be demonstrated. There is an assumption underlying many

statistical tests that words are randomly distributed but obviously an author's choice of one word in a text affects the choice of the next word. Approximately one word in twenty written English words is a repetition of the definite article, *the,* but the probability of *the* co-occurring sequentially with *the* is not twenty times twenty i.e. once in every 400 words. Apart from typos or the name of a rock band the sequence *the the* is highly unlikely to appear in writing. As Kenny (1982:166-167) points out in his handbook on stylometry:

> The methods illustrated in this book take no account of language's serial nature. A sentence length distribution or word frequency list does not preserve the order of words or sentences; it is as if words or sentences were cut out of text and then shuffled together in a bag ... the fact that the statistical methods used in this book ... do not take account of the serial nature of language in no way invalidates [their] use. What it does mean is that a great deal more information about stylistic features remains to be studied even after these techniques have been exploited to the full.

Statistics, then, can reveal tendencies and patterns but the presumption that something has been proven needs to be resisted. At times simple arithmetic can help make sense of a complicated process. A good example of this is where Sinclair (2003:125) draws his reader's attention to an interesting tendency using the phrase *happen to be* as an illustration. The word *happen* occurs 41,484 times in the *Bank of English,* while *happen to* occurs 7,173 times and *happen to be* occurs 1,207 times. 7,173 is 17% of 41,484 and 1,207 is 17% of 7,173. Each extra word which builds the trigram reduces the number of instances by 83%.

2.6 The gathering and processing of the data

For corpus linguists the data usually consists of collections of naturalistically gathered written and/or spoken text. The naturalism referred to here stems from the fact that the producers of the language have no idea that what they are saying or writing will later be subjected to linguistic analysis and have communicative purposes unrelated to the preoccupations of subsequent researchers. For Stubbs (2002), this is part of the definition of corpus data. One area in which it is difficult to achieve this degree of naturalism is in gathering data about learner English. As Scholfield (1995:47) points out, most classroom learners of a language are aware that almost anything they say or write might be recorded electronically or in paper form and be quantified as part of individual, group or course

assessment. This kind of elicited learner data falls short of the ethnographer's ideal of naturalistic data in that the 'donor' is aware that their contribution will be under scrutiny. If students in Portugal are told that a Portuguese sub-corpus of learner English is being compiled and that argumentative essays of 500 words or longer are sought from them, they will intuit that their writing is going to be studied and quantified. They will often be on their best (or what they think is best) linguistic behaviour. The student writers might 'play safe' or adopt error avoidance strategies. The essay sessions were usually set up outside normal class times and the resultant extracurricular nature (for the large majority of *Porticle* writers) of this exercise and the writers' rapid involvement in answering the essay question of their choice meant that fear of error did not figure prominently in the writers' text creation. Figure 2.2 outlines four general approaches to data gathering for quantification in the linguistic sciences.

Figure 2.2 Four main sources of language data (adapted from Scholfield 1995)

1	2	3	4
NATURALISTIC NON-REACTIVE	QUASI-NATURALISTIC REACTIVE	REACTIVE INVOLVING PEOPLE'S OPINIONS	REACTIVE INVOLVING PEOPLE'S MANIPULATION OF VERBAL MATERIAL
observation	*elicitation*	*surveying or questionnaires*	*testing*

Using this idealization of what is really a spectrum, data gathering for corpus compilation can be classified squarely under Type 1 or more rarely under Type 2 of the above table. Native-speaker data can be entirely Type 1, naturalistic and non-reactive. The writers will be writing for their own purposes and there will be no two-way communication between the corpus-builder and the writer. There *may* be a request, on the part of the investigator, for permission to use the text but this will be after the fact, when the text already exists.

According to Granger (2003) the data from students represent a special category of elicitation which is called 'clinical elicitation' by Ellis (1994) as opposed to 'experimental elicitation.' I shall typify this data as Type 2 (see above). To maintain the quasi-naturalistic status for this data, I was careful when presenting *ICLE* and 'explaining' my aims to cooperating students, not to mention set phrases or idioms. This was to avoid alerting the students to my focus on phraseology.

Granger's (2003: 8) latest characterization of the *ICLE* data as 'natural' is to distinguish it from 'experimental' ('data gathered in experimental conditions or in artificial conditions of various kinds' (Sinclair 1996a); but she admits that 'learner data is therefore rarely fully natural' given that some task variables, such as the essay title or the time limit, are often imposed on the writer.

2.7 Conclusion

In this chapter, the epistemological underpinnings of corpus linguistics have been discussed. The relative importance of (1) intuition and (2) attested data, in the investigation of language, have been examined in some detail. Methods of identifying prefabs were examined. The work of applied linguists working in the paradigm of CLC was chronicled and some their insights presented. In the next chapter, Chapter 3, corpus methodology is employed to establish the distinctive features of the prefab use found in a corpus of Portuguese learners.

Chapter 3 Corpus based linguometry

3.0 Introduction

This chapter describes the methodology used in the selection, compilation and analysis of the corpora to be investigated. The research focuses on the phraseology of written English for academic purposes, and in particular on the argumentative essays produced by Portuguese undergraduate students of English. The corpus compiled to represent the academic writing of these students is called *Porticle,* the Portuguese sub-corpus of *ICLE.* The first section of this chapter discusses the argumentative essay as genre and the reasons for choosing it to exemplify academic writing. Section 3.2 contains a description of the students whose essays make up the *Porticle* corpus. Details are given about my relationship to the students and where and when the essays were written.

Section 3.3 describes the three corpora of native speakers' writing, which were adapted or created to provide a basis of comparison with the Portuguese students' writing. These three corpora were designed to represent argumentative writing by native speakers of English manifesting three broad levels of expertise. The methods used to select or compile the three control corpora (*Locness, Bawe and CofE*) are the subject of Section 3.3. The central part of the chapter (3.6-3.11) is taken up with the methods used to interrogate these corpora, in order to establish the amount of prefab use found in each corpus and the functions which these prefabs perform. The differences in quantity of the prefabs and the purposes for which they are used in *Porticle* are compared to those of the three samples of NS writers. The chapter therefore moves from a description of the data collection to an analysis of the data and a presentation of the results of this analysis.

The orientation of this chapter changed as the analysis proceeded. It was originally conceived as a fact-finding one, in which the essays were examined and the discernible surface linguistic features recorded. The intention was to leave deeper analysis until Chapter 4. In the end, this was simply not feasible as it was found necessary to apply certain provisional semantic and pragmatic interpretations to the data as they were gathered. Where possible, however, deeper semantic or pragmatic analysis is left for the following chapters. In the reporting, some attempt is made to show the experimental and

recursive nature of the approach by recounting the thought processes which guided certain decisions. For the same reason, several avenues which were explored and which proved to be *culs de sac* are nevertheless described.

3.1 Methodological issues

The experimental focus of this thesis is a corpus of written essays produced by 215 Portuguese undergraduate students of English and the types and quantity of prefabs in these texts. This corpus is referred to throughout as *Porticle* (or Portuguese sub-corpus of the International Corpus of Learner English). The methodology followed in this investigation has been labelled Contrastive Interlanguage Analysis or CIA (Granger 1998a: 12-14). It is principally empirical: observations and generalizations are made about machine-readable corpora of naturally occurring learner data, which are contrasted with corpora of writing by native or expert writers. To carry out this contrast, there is a need to search for or to develop corpora with texts of a similar length and communicative purpose. As a data-driven investigation, the overall scientific orientation is empirical, involving an iterative cycle of observations, intuitions and hypotheses that lead to further observations, as outlined in Chapter 2. This quantitative analysis is complemented by qualitative analysis to establish whether there is linguistic motivation for the statistically significant differences among the corpora. The first control corpus was *Locness,* 115,642 words written by American undergraduates. Its adaptation is described in Sub-section 3.3.1.

Section 3.6 sets out the results of the manual search of two small samples from *Porticle* and *Locness*. The rest of the chapter is mostly taken up with the results of the computerized search of the four corpora, *Porticle*, *Locness, Bawe,* and *CofE*. The relative merits of these two different approaches to measuring prefabrication is discussed in Chapter 4. At this point it should be observed that certain items which are rare in the corpora are recognizably prefabs. I know this through introspection as a native speaker of English or through consulting collocational or learner dictionaries. Many sequences which I deemed to be prefabs in the two 5000-word samples from *Porticle* and *Locness* occurred only once. Such sequences would not normally be uncovered using text retrieval software operating with a stipulated minimum frequency.

Various automated techniques were used to filter out word sequences which do not comply with the previously specified criteria. For example, the Find and Replace function within Microsoft Word, the word-processing software, proved very useful at this stage. The first filtering of the data was done using syntactic and phraseological criteria. Frequency was chosen as the starting point in the search for prefabs simply because the frequent recurrence of strings of two, three, four or more words suggests that they may be lexicalized sequences which convey conventionalized meanings with some sort of social warrant. A psycholinguistic corollary of this social linkage is that prefabs are recalled rather than composed 'online' by the speaker/writer.

Prefabs are mappings from form to meaning or from meaning to form which, through continued usage, have become accepted as the natural or idiomatic way to say or write something. Undeniably, in most cases, the origin of these collocations can be explained by the syntax governing the words involved, but what is interesting is that they became the preferred way of expressing a meaning against competing word sequences which are also licensed by the syntax. A good example of this 'syntactic packaging,' or colligation, are the collocational frames which abound in language corpora, the phenomena first referred to by Renouf and Sinclair (1991) and discussed in Chapter 2.4. Some such frequent colligations consist of two frequent, usually closed-class words with a paradigmatic choice of open-class nouns occupying the slot between them (e.g. *the* _____ *of*: *the end of, the help of, the purpose of, the husband of*). The instantiations of these frames in each of the four corpora were collected in lists with the slot-fillers in descending order of frequency (Table 3.8).

After carrying out a word-frequency analysis of the four corpora, all the recurrent N-grams (sequences of N words) which occur in each corpus with a frequency greater than a stipulated floor were isolated. Although the Wordsmith Tools suite of programs has a device for detecting recurrent sequences of words from texts, called clusters, I decided to use KFNgram (Fletcher 2004) for this task. One reason for this choice was that KFNgram was custom-built for this purpose and allows various mergings and rearrangements of the bigram, trigram, and N-grams lists with rising value of n. Lists of bigrams, trigrams and Ngrams up to 9 words long are generated both in order of frequency and alphabetical order. KFNgram creates index files and data files and inserts these into the corpus directory. This means that one has either to continually delete these files or else always remember to work

with a copy of the original corpus. The latter option means that one can keep all the N-gram listings for future reference.

The evaluative dimension of the language of academic prose (Hyland 2000) was one of the primary concerns of the study and, in particular, the contribution that prefabs make to the configuration of writer's stance (see Chapter 1.10). If Wray's (2002) contention that self-affirmation is a macro-function underlying prefab use, or is the main function of formulaic sequences in her terminology, then it seems that expressions of stance would be among the most central uses of prefabs. In a similar way, writer self-mention, writer presence in the text, and other facets of interpersonal relations would be key roles of prefabs. In the course of examining the prefabs in the corpora, I was advertent to prefabs being used in this way.

Corpus linguists can test various hypotheses *en passant* or on the fly. Use of computers in corpus linguistics allows initial hypotheses to be tested and then reformulated, thus allowing more real-time flexibility in the analysis carried out. My stated interest in writer's stance appears to run counter to the fairly naïve empiricist programme outlined in this chapter. Nevertheless, many stance prefabs are discernible from surface features and have been well documented in the literature (Aijmer 2002, Biber et al. 1999). While the more directly observable aspects of stance are recorded in this chapter, a more synoptic treatment of stance is found in Chapter 4.2.

The uses of all kinds of prefabs found in the four corpora were examined because I wished to gain an overview of this linguistic phenomenon, preparatory to a series of studies over the coming years. For the purposes of this thesis, a decision was taken to concentrate on those prefabs which contribute to epistemic and attitudinal stance. One aspect of phraseology which attracted my intention was the intensification of adjectives. This combinatoric dimension has been investigated by at least two corpus linguists (Lorenz 1998; Kennedy 2003). Both studies show convincingly that adverb-adjective combinations contribute a great deal to style and the conveyance of writer stance. For this reason and also because it is a fairly circumscribed and therefore manageable aspect of text, I decided to capture occurrences of adjective intensification from each of the four corpora. The data relevant to this study are reported in Section 3.5.6.

In the research for this thesis, a mainly empirical methodology is employed. A corollary of this empiricism is a decision to start from an examination of the surface details

of the word sequences and texts. This permits an assessment of the extent to which such a 'superficial approach' might inform an analysis of the form and function of prefabs.

The shared features of each individual corpus and its frequent and typical prefabs is investigated. Although the Portuguese and the native speaker writers have been referred to in this study as if they formed homogeneous groups, a great deal of variation exists not only across different L1 and L2 groups but also between individuals of the same group and even within the output of any individual. Although corpora are designed to neutralize the effects of such variation, it is wiser to attribute *tendencies* to the writing of certain groups of individuals rather than make categorical statements. When the overuse and underuse of certain terms by *Porticle* writers are mentioned, this should be read as shorthand for 'EAP learners whose mother tongue is European Portuguese tend to use a particular word or phrase more or less frequently than native speakers of English or learners with a different mother tongue.'

3.2 The argumentative essay as data

The argumentative essay is a time-hallowed exercise in rhetoric and dialectic practised by British schoolchildren through the centuries. Writing an argumentative monologue formed part of the Grammar School entrance examination administered in June 2003 to ten-year-old schoolchildren in some parts of England and Northern Ireland (personal communication, Catherine McKenny, a primary school pupil).

Since the genre of academic essay is ubiquitous within higher education, it might be assumed that all NS students are proficient in the writing skills needed to achieve their full potential. There is research, however, that questions the assumption that tertiary level students should automatically be assumed to have acquired the rhetorical skills necessary for effective essay writing. Womack (1993) examines the role which essays play in contemporary education and asks why they have gained their current status. He emphasises the importance of the essay in education as a 'default genre' (Womack 1993). Although there may be other forms of writing required, Womack notes that, as students progress from school to University and through the levels of Higher Education, the essay plays an

increasingly central role as the form of writing required, particularly within the Arts. Womack attempts to dispel the illusion that the essay is an easily acquired genre and shows that it has acquired its dominant position in British academic culture through a shedding of many of the previously practised literary genres, such as colloquies, fables, characters, themes, epistles, orations and declarations.

Womack argues that modern English teaching selected the essay genre as the principal written genre because it was thought to be a more transparent, less complex style, which was suitable for displaying academic expertise and argument. The essay style became increasingly useful for examination assessment purposes during the nineteenth and twentieth centuries when public bodies required a meritocratic recruitment system for the middle classes. The essay was (and still is) generally considered to be a fair means of assessment since it is supposed to be accessible to all as a form of judicious thinking in written form. Womack, however, considers that the essay genre is, in fact, difficult for many students to gain competence in. He describes the essay as a 'universally contradictory sign' and says that it is of such immense value within the British education system because only limited numbers of students are able to use it successfully. It thus creates successes and failures.

As education in many sectors of Britain has moved towards continuous assessment, the dominance of the essay has continued and remains largely unchallenged. Hounsell (1984) notes that, while students may place originality and understanding high in their list of what they *thought* teachers wanted in their essays, these factors were rated as less important by their teachers. Hounsell asserts that teachers want to see students expressing subject knowledge and reasoning which they (the teachers) are already familiar with. In the same study Hounsell also examines how undergraduate history students perceived the essay task and notes three separate approaches (Figure 3.1).

Figure 3.1 Three main student approaches to essays (according to Hounsell 1984)

Argument	Presentation of a coherent point of view backed by evidence (most prized by teachers according to Hounsell)
Viewpoint	A point of view is presented but not backed by data or evidence. According to Hounsell's study, the students who use this style view their learning as complete when they can clearly articulate what they consider to be the central issue(s) of a question.

Arrangement	Students attempt to cover as many points of view on a topic as possible but do not attempt to tie these ideas into a coherent argument because they imagine that they are required to demonstrate the breadth of their learning.

Any attempt to produce a generic structure for the argumentative essay soon brings the realization that such an all-inclusive outline might be extremely difficult to obtain. The essay is a particularly amorphous genre. This elusiveness would please one of its first and greatest exponents, Michel de Montaigne (1575). Few studies of academic writing attempt to define the genre of discursive essay which undergraduates (NS and NNS) need to be proficient in writing. One study which does attempt to produce a model of the generic structure of the essay is Hyland (1991). Hyland emphasizes the need for teachers of EAP to elucidate the structure of the argumentative essay. He recommends that students are familiarized with the different kinds of schema which are used as the organising principles for constructing argumentative texts. Hyland here draws upon Swales' (1981) analysis of the 'moves' within article introductions, discussed in Section 1.13. Hyland analysed NNS student argumentative essays into a three-stage structure which represents, according to him, the organising principle of the genre. He identifies this structure as: Thesis, Argument and Conclusion, and further sub-divides these main stages into smaller segments called 'moves', following Swales' terminology. Hyland further indicates that there may be more sub-genres of argumentative essay.

One outline for an argumentative essay might be: a discussion of the essay question and a defining of terms, a thesis statement, a defence of this thesis, an examination of the alternative viewpoints, a conclusion which decides on the pros and cons and a section containing 'some thoughts to ponder' and perhaps a final resume. Essay writers should adduce proofs, evidence, examples and counterexamples, stories, anecdotes and other opinions to support their position and undermine that of their opponents. But the creation of argumentative prose need not always be so linear. There are other devices available for persuasive writing, such as satire, parable, irony, *reductio ad absurdum*, thought experiments and *argumentum ad hominem*. It can be argued that the genre of essay has fuzzy edges:

> The essay stands apart from both poetry and prose fiction, as well as from other forms of academic writing, in its emphasis upon the actual situation of the writer, and thus upon the personal nature, the 'situatedness' of all writing. Even when an essayist has arrived at a tentative decision and writes to persuade the reader as to the validity of their position, the obligation to persuade reasserts the equality between writer and reader and puts limits on the writer's reliability. These characteristics of fuzziness and riskiness make the genre of argumentative essay a challenging one for writers of all ages.
>
> (Spellmeyer 1989:262)

3.3 The compilation and construction of *Porticle*

The choice of argumentative essays as the constituents of my experimental corpus requires some explanation. The genre is not taught and practised in Portuguese secondary and tertiary education curricula, including in the English language courses. Students are rarely asked their own opinion but are expected to substantiate analysis and commentary by showing familiarity with the received opinions and literature of the field. A minority of the contributors to *Porticle* may have been required to write an argumentative essay for Cambridge University EFL examinations (CPE and CAE), but most were doing so for the first time.

Thus, while the concept of generic structure informs the study of the essays in *Porticle*, it was thought unlikely that a generic structure would be found in the corpus. Nevertheless, there still remained the possibility that a task, through its completion, might generate or necessitate its own generic structure. The decision to use the argumentative essay as the basic ingredient of all the *ICLE* sub-corpora allows comparison of like with like in terms of content and context of use. Although the genre is new to most of the Portuguese writers in *Porticle*, the rhetorical micro-functions of e.g. persuading, defining, illustrating and concluding, which enable the elaboration of argument, are already familiar to them through their English studies in secondary school.

Each essay gathered is a non-technical piece of writing which does not require bibliographical support or extensive research and whose relative brevity takes up few hours of the subjects' time and allows the inclusion of a large range of writers, even in a small corpus. It is based on the assumption that most tertiary level students will have an opinion

on one or more of the topics and will view the invitation to convince the reader of the rectitude of their view as an interesting challenge.

Researchers into other *ICLE* sub-corpora (Kaszubski 1998; Ringbom 1998) observe that learner writers tend to overuse lexis and structures (e.g. core vocabulary and modals) more usually associated with spoken language. Kaszubski (1998) suggests that this is due to the predominantly oral methodology of ELT. A simpler explanation in the case of Portuguese students is that they are unfamiliar with the genre in their first language and in English they have only really practised using persuasive language orally.

From an analytical perspective, an attractive feature of the essay is that each 500 words is an integral text, an accessible discourse. This circumvents the vexed question of whether whole texts or short samples from texts should be used, which has exercised corpus compilers from the *Brown* corpus onwards. For example, *Brown*, *LOB, FLOB* and the *BNC* are all built from samples or excerpts. The *Bank of English*, on the other hand, consists, on principle, of whole texts (e.g. entire novels).

Thus, despite the reservations expressed about the suitability of the essay as a vehicle for testing native and non-native writers, a case can still be made for using the argumentative essay to look at learners' prowess in English academic writing. Firstly, even though the writers who contributed to *Porticle* are in an EFL environment, they are required to produce English essays as assignments and under examination conditions to prove their command of the various sub-disciplines of the English side of their Joint Honours degree. Secondly, most genres of academic writing, with the possible exception of learner diaries, descriptions, taxonomies and reports, contain argumentative steps. Finally, many of the participants in the *Porticle* corpus observed informally, on completion of their essay, that they had enjoyed the challenge to think which the task provided, and that they had discovered what they really thought on the subject they had chosen to write on. This is a good example of 'writing to learn' which so often accompanies 'learning to write' (Murray 1990).

3.3.1 The writers of the *Porticle* essays

Most of the 215 writers whose essays make up *Porticle* were young (19-23) with a greater number of female students (female 77%; male 23%). The majority of them were training to become English or Portuguese teachers in primary, middle or secondary schools, although about a quarter of the subjects were studying English combined with management studies, hoping to take up administrative posts. All of them were Portuguese citizens and had completed at least 7 years' formal study of English. The data was collected throughout 2004 at the universities of Lisbon, Oporto, Evora and Aveiro, the Catholic University of Portugal and at the schools of education in the polytechnics of Viseu, Lamego, Guarda, Castelo Branco, and Leiria. Some essays were contributed by students at ISLA, the Higher Institute of Languages and Administration in Lisbon.

The middle school and early secondary English syllabuses they had studied previous to tertiary education concentrated on grammar and then, in the final four years of secondary school, placed emphasis on learner autonomy and the communicative approach. The syllabus of the last three years of secondary education was also thematic in design, with themes such as environmental awareness, the population explosion, alternative lifestyles, young people and the world of consumerism. However, it should be added that, as in most educational systems, there is a discrepancy between the ideals expressed in the syllabus and the reality of what is taught and learned in the classroom. Although an integrated skills approach is recommended in the syllabus, these themes tend to be realised through texts and text-attack activities such as comprehension and vocabulary work. There is greater emphasis on fluency rather than accuracy, and receptive rather than productive activities, with reading and vocabulary dominating at the expense of speaking and writing. This is certainly the case, for example, in *Zoom* and *Screen 2*, the two coursebooks with highest sales for the 11th year, the penultimate year of High School (Botelho and Silva 2004, Barros et al. 2004). Writing is often done as consolidation of orally learned structures and skills.

Typically, *Porticle* writers were fairly fluent communicators in written English who made few lexical or grammatical errors that impeded their comprehensibility and could produce passable, comprehensible written English texts on a range of subjects in expository, narrative or descriptive modes. They had all passed the national English matriculation examination for university entry. The Portuguese Ministry of Education has

stated that the exit level of English for students finishing secondary school is Common European Framework Level B2 (equivalent to Association of Language Testers of Europe (ALTE) Level 3 and The Cambridge ESOL First Certificate of English). This level of attainment, referred to in the Council of Europe (2001) literature as the Vantage Level, constitutes a good intermediate or pre-advanced level of language proficiency. In the written language of these students, the more basic (and irritating) lexical and syntactic errors tend to decrease in frequency and gravity and this makes phraseological infelicities more noticeable. Pedersen (1995) comments that in EAP written work an accumulation of miscollocations of words can make the reader impatient, or aware of the strangeness of the writer's text. Such oddities can impede comprehension and give the impression that the text lacks quality. Cop (1988:2771) encapsulates this situation well:

> Incorrect collocations have a comical effect on the native language receiver while correct ones are accepted as 'passwords' to native language fluency in a foreign language.

3.3.2 The construction of *Porticle*

As a tertiary-level teacher of English in Portugal, I was well-placed to obtain student essays both from my own students and, through colleagues, from students in other universities. All contributions were from undergraduate students studying English courses in tertiary educational institutions. It was necessary to rely on the good will of lecturers, technicians and students in many different institutions, each with its own distinctive culture. Having obtained fewer than 30,000 words in the first six months, exclusively from northern Portugal, I decided to extend the search for data to most tertiary level institutions throughout Portugal. As with the other Learner Corpora compiled to form part of the International Corpus of learner English (*ICLE*) held at the Catholic University of Louvain, the participants in *Porticle* were all third- and fourth- year English students. Students from ten institutions eventually participated and, as only those students willing to help are represented, the corpus was largely opportunistic. In situations where the lecturer handed the students over to me and withdrew, student participation was less than in situations where the lecturer showed interest. Although the task of writing 500 words in a foreign language within a couple of hours is rather demanding for third- and fourth- year university

students, most participants enjoyed the challenge. One student wrote a sentence and then cut and pasted it repeatedly until he had more than 500 words. This Dadaistic essay showed up very quickly when recurrent clusters of words were sought using Wordsmith Tools, the text retrieval software package. In several institutions, at least one student in a group was unable to achieve 500 words within the 120 minute period. These students asked for more time and some later sent on a completed essay.

After a small presentation about the *ICLE* project and its relevance for future English teachers (which most of the writers in *Porticle* aspire to be), each group of students was given a list of 14 argumentative type questions from which to choose their essay title (see Appendix 2). If participants were currently very interested in another topic, they could write on this instead. Sometimes teachers requested the use of a different title or set of titles more directly related to their curriculum. There were surprisingly few such deviations from the set titles. Students were asked to send their essays as Word attachments to two e-mail accounts. Two addresses were used for security reasons. When there was insufficient time for electronic mail, two copies of the essays were gathered on to floppy disks.

These Word documents together with a learner profile completed by each contributor were the raw data which would make up the *Porticle* corpus. The learner profile was designed to gather important personal and career information about the contributors: their name, age, gender, institution, course, year of study, foreign language learning experience, the languages through which they did their schooling and the language(s) which they and their family mainly spoke at home as they grew up. Significant periods of time spent living, studying or working abroad were also recorded. An important part of each learner profile was the student's signature giving authorization for their work to be used for investigative purposes. A copy of the learner profile is contained in Appendix 2.

The learner essay files contained on the floppy disks collected *in situ* or in attachments received by e-mail from the subjects were individually converted into ASCII or text-only files with carriage return line breaks. The native essays for *Locness* and the essays and assignments contained in *Bawe* were treated in the same way (Section 3.3). The reason for converting all the text to ASCII files is that most text retrieval and tagging software works much better with text data which has little or no formatting. Each student essay was assigned a unique header showing the corpus title, the country of provenance, the contributor's institution, the number of the essay/contributor and the batch to which it

belonged. The first learner essay therefore had <ICLE-PT-ESE-000.1> inscribed at the beginning of the text. This header is entered between angle < > brackets so that it is not ordinarily read by the text retrieval software (word counter, concordancer, wordlist compiler, N-gram finder, type-token ratio calculator and so on). Of course, all of these software tools can be instructed to read the headers, which can then be used to call up specific files from the corpus.

The header was then placed at the beginning of each ASCII file and was also entered on the student writer's questionnaire, as the Text Code, to provide a cross-reference between each essay and the learner and task documentation contained on the corresponding learner profile. It was then necessary to remove all extraneous features such as numbers and essay titles from each ASCII file and then save all the files, using a shortened version of the header contained in the file and learner profile as file-name. The first essay of the corpus was called PTES1001. These file-names, which retain all the essential information from the header and learner profile Text Code, were so designed that when placed in a Windows folder, they would be ordered as a directory which would constitute the *Porticle* corpus.

When the *Porticle* corpus comprised nearly 130,000 running words after eighteen months spent in compilation, I decided that this collection of 215 essays would be the experimental corpus for the investigation. Although the corpus, when it is eventually completed, will contain 200,000 words of essays, each between 500 and 1,000 words in length, with only one essay per student, I confine my research for this thesis to the shorter interim version. The term, *Porticle,* is used at all times in this thesis to refer to the interim corpus of 126,368 words and not to the projected larger corpus of 200,000 words, which will come into existence after completion of the thesis.

In the case of almost all existing NS corpora, the language collected was produced for purposes and in situations totally unconnected with the corpus compilation. As discussed in Chapter 2.7, the learner data in *Porticle* do not meet such a stringent qualification as most of the essays would not have come into existence unless the subjects had been requested to produce them. Within the Portuguese learner corpus there are some exceptions where, for example, the student's writing was graded as part of an end-of-year assessment exercise or the teacher used the essays as part of a writing and research project or teachers who had initially produced student essays when asked to do so, incorporated the essay writing as a formally evaluated course component in the following year(s). In this

minority of cases (70 from 215), the students therefore had a writing purpose other than to become grist to my concordancing mill and so their output could more easily be described as naturalistic data. Nevertheless, at the thematic and informational level, students tend to treat graded work in a certain way, e.g. showing what they know, hiding what they do not. The rest of the data can be characterized as quasi-naturalistic (see Section 2.6), given that the subjects were writing to complete a task and had no clear idea about the research aims. The 'quasi' must remain, however, because the students knew that some kind of analysis was going to be carried out upon their writing and they may therefore have resorted to error avoidance strategies, such as eschewing grammatical constructions which they perceive as difficult or they may have used mainly core vocabulary whose collocational behaviour they felt less unsure of.

3.4 The control and reference corpora

In order to compare and contrast the writing contained in the experimental corpus *Porticle* with other comparable collections of writing, three control corpora of comparable size were adapted (*Locness* and *Bawe*) or created (*CofE*). The three populations of writers represented are (1) American undergraduate students, (2) high-achieving British undergraduates and postgraduates and (3) editorialists writing in British and American broadsheet newspapers. The three control corpora corresponding to these populations are (1) *Locness,* the American component of the Louvain Corpus of Native Essays; (2) *Bawe*, the corpus of British Academic Written English, compiled at Warwick University; and (3) *CofE*, the Corpus of Experts, a collection of US and UK broadsheet editorials, which I compiled from electronic versions of the newspapers. Each of the four corpora used in the study contained approximately 120,000 words.

Two (*Locness* and *Bawe*) were designed using argumentative essays written by contemporaries of *Porticle* writers. The third control corpus, the Corpus of Experts (*CofE*), consisted of editorials and essays published in British and American broadsheets, mostly by professional journalists. The essays in the *Bawe* corpus were chosen from a pilot corpus of British student writing held at Warwick University (Nesi, Sharpling and Ganobcsik-

Williams 2004). All of the writing submitted by students to the *Bawe* corpus had been assessed as good, proficient assignments by subject tutors.

3.4.1 The *Locness* corpus: argumentative essays by American undergraduate writers

The *Locness* corpus was a 115,642 word extract from the 300,000-word Louvain Corpus of Native English Essays, written in the early 1990s. It contained 116 essays written by USA undergraduates, all except six of whom were aged 17-22, of a wide range of ability. The essays represent the writing of native speaker contemporaries of the Portuguese writers in *Porticle*. Upon reading through these essays, I came to appreciate that these American undergraduates should be considered as apprentice writers of English for academic purposes. Although they generally have a higher proficiency in English than the Portuguese writers in *Porticle*, they are still gaining proficiency in academic literacy.

The essays are on different topics, but they are all argumentative rather than narrative, descriptive or expository prose. The essays were untimed and students had access to the reference library of their university. (Any direct quotations in these essays were removed from the corpus and marked thus: <*>.) The *Locness* essays are longer than in *Porticle* (with a mean text length of 1079 vs. 595 words).

At each stage in the corpus investigation of *Porticle*, *Locness* served as the primary control corpus. This decision was taken because the design of *Locness* most closely resembled that of *Porticle,* with the contributors of both being undergraduates of a similar age and educational attainment.

3.4.2 The corpus of British Academic Written English (*Bawe*)

The British Academic Written English corpus, henceforth the *Bawe* corpus (Nesi et al. 2004), consists of writing which has been graded by subject lecturers as being of good quality. It consisted of the work of students in Arts, Social Sciences and Sciences at the University of Warwick. The corpus, begun in 2001, was designed to provide a research database for researchers and writing tutors. The corpus is copyright-cleared and supplemented by appropriate contextual information about the students, their courses of study and the texts they produce. The database is organized in such a way that it is possible

to identify and compare writing performance across a range of parameters, including the gender, year of study, and discipline of the writer.

I was granted access to *Bawe* at an early stage of its development. This pilot version of the corpus contained 501 pieces of written work. The authors whose work was compiled in this early version of *Bawe* were mostly high-achieving undergraduate students and were predominantly native speakers of English. Each of the texts in the corpus contained between 1,000 and 5,000 words. Subjects were paid for their contribution. From this I selected a sub-corpus of 57 texts with 108,912 running words. The shortest pieces of work and writing of a more argumentative nature were chosen, trying, where possible, to use only one piece per student. Clearly, this gives an average of 1893 words per essay, which makes the *Bawe* writings lengthier than the *Locness* essays with their average of 1079 words and the *Porticle* essays averaging 595 words in length. *Bawe* also differs from *Porticle* and *Locness* in that only successful or academically proficient writing was selected.

3.4.3 *CofE* or Corpus of Experts

Granger (1998a:18) suggests that broadsheet newspaper editorials might provide a good control corpus for comparison with a corpus of argumentative essays given the shared persuasive function and similar length. Developing this idea, I compiled The Corpus of Expert English (*CofE*) over the academic year 2002-3, using selections from *The New York Times, The Guardian,* and *The Guardian Weekly* and *The Times Higher Educational Supplement.* This corpus contained 113,101 words. The selections comprised the leaders, editorials, essays and comments written by journalists, politicians and academics, which consisted of between 500 and 1,000 words. Explicitly argumentative pieces were chosen where the text provided sufficient information to enable reader comprehension. Those pieces were chosen which dealt with less localized concerns and which were, as far as could be judged, less ephemeral. A balance was sought between both sides of the Atlantic and the more idiosyncratic text types were filtered out (e.g. pieces written in dialogue or poetic form). The writers of all these short pieces were either professional journalists or academics and therefore could be said to live mainly by their pens or word-processors.

The question of spelling variants is important. In designing *CofE*, it was necessary to decide on using British or American spelling. If the variant spellings are left as they stand this leads to an increase in the number of word forms contained in this corpus. This anomaly in word frequency counts was not considered sufficient reason for tampering with the original texts of either the British or the American writers. The spellings were left as they were. Clearly, the same policy was followed with the essays of the Portuguese writers in *Porticle* in relation to which a similar problem arises. If a student uses the word *literally* this will be counted as one word but there may also occur misspellings such as *literaly* and *littterally* which will be counted as different word types thus inflating the writer's word count. It was decided to let these variant spellings remain: my learner texts were entered into the corpus exactly as they were submitted by the subjects.

3.4.4 The BNC-baby corpus

A further corpus of native speaker English was used on occasion to supplement the three native speaker corpora in this investigation. The *BNC-baby* (Berglund and Wynne 2005) is a four-part corpus constructed by a principled sampling taken from the BNC World Edition. This corpus consists of four million-word samples from the much larger British National Corpus. The samples were chosen to represent academic writing, imaginative writing, newspaper texts, and spontaneous conversation. Texts previously included in the BNC Sampler were excluded from selection. The *BNC-baby* provided a useful yardstick of the usage of a prefab in spoken and written English. In looking up a word or prefab, *the BNC-baby* provides a pie-chart showing how the search word(s) are distributed across the four samples.

3.5 Issues about representativeness and comparability of corpora

There are two key issues in setting up the experimental research at the heart of this thesis. The first requirement is to ensure that *Porticle* is a reasonably representative sample of the argumentative writing of the population of students I wish to study. The second issue is whether the control corpora are comparable with *Porticle* in terms of task undertaken, writer purpose and genre so that the main difference among the corpora is the level of authorial expertise which each corpus represents. My concept of authorial expertise is a fairly intuitive one, based primarily on the NS/NNS divide and secondarily on educational attainment as judged by university tutors in the case of *Bawe* or the professional status and breadth of readership enjoyed by the *CofE* writers. I will argue in Chapter 5.1 that this quasi-experimental design (Cohen et al. 1989:198) provides sufficient equivalence of groups to avoid equivocality of interpretations.

The *Porticle* writers are senior undergraduates at Portuguese polytechnics and universities with widely differing entrance requirements. They range in linguistic ability from near-native speakers of English to intermediate level (IELTS bands 5-8, ALTE 3 and 4). Although they demonstrate a correspondingly wide range of writing ability (from beginner to expert), the majority of the *Porticle* essays show less than full control of the sub-skills of writing and would be assessed by most EAP teachers as exemplifying either an intermediate or post-intermediate level of writing proficiency. I embarked on the research when the number of essays in *Porticle* exceeded 200.

The *Locness* corpus was originally compiled for use as a control for the various *ICLE* sub-corpora. The original research design ensured that the essays were argumentative in nature and not merely discursive or descriptive. Although the essay titles in *Locness* do not coincide exactly with the titles of the *Porticle* essays, there is sufficient correspondence in theme and emphasis to make the two bodies of writing comparable. I compiled my sub-corpus from *Locness* by choosing only from essays written by American students. This was done because the *Bawe* control corpus was compiled to represent successful British undergraduate writers. The essays in *Locness* were longer than those held in *Porticle* (average 1079 words) and were untimed. The American undergraduates were allowed to consult the works in a reference library. Any quotations they made in their essays, however, were excised and shown as <*> in the transcripts. Although only about a third of the *Porticle* essays were untimed, the other writers were given at least two hours to produce

their writing and those who wished took away their essay to finish at their leisure and send on.

The contributors to *Porticle* and *Locness* represent all levels of educational attainment ranging from students destined to barely pass their degrees to students destined for distinction. As a great deal of the comparative corpus analysis focuses on these two groups of students, I feel confident that the greatest difference between them is that of being NNSs and NSs respectively. Although the writers in *Porticle* are the least expert writers in the research, having had only two or three years study of EAP, the *Locness* writers would be closest to them in this respect as they are still learning the conventions of writing academic prose.

The contributors to the *Bawe* corpus had all distinguished themselves as good writers. This was not the case with either the *Porticle* or *Locness* writers. Every effort was made to choose the shortest texts for reasons of comparability with the other corpora in the experimental design: the average length of the *Bawe* texts chosen was 1893 words. Although argumentativeness had not been part of the brief of *Bawe* writers, close reading showed that the writers in *Bawe* had to argue their case. Even in pieces of work of a descriptive nature, the writers had to persuade their reader that the framework or characteristics they had chosen to describe were the right ones. They were university students writing to display their knowledge and understanding of their chosen area and their capacity for conceptual analysis. I contend that there is sufficient overlap between their work and *Porticle* and *Locness* to justify using *Bawe* writers as a control. The three corpora represent EAP in Portugal, the US and in the UK. The two major distinguishing features of *Bawe* are that the contributors are writing in the British variety of EAP and that they are all accomplished writers.

The Corpus of Experts or *CofE* was compiled after a suggestion of Granger (1998) and following work by Neff et al. (2003). Editorials are by their very nature persuasive: they overtly try to influence public opinion and government policy. The chosen editorials fell within the word limits of the *Porticle* essays (500-1,000 words) with few exceptions and the average length of the editorials was 853 words. My selection included British and American English and the writers are professionals: they are judged to be expert writers by their peers and by their readership.

In choosing the control corpora, I hoped to set up a cline of writing expertise in the English language ranging from recently initiated apprentices, the *Porticle* writers, through more experienced learners, the *Locness* contributors, to the successful writers in *Bawe* and the acknowledged experts in *CoE.* The writing tasks undertaken and the texts they produced were sufficiently similar to enable me to attribute the variety of prefab use to the writers' level of authorial skill.

3.6 Small is beautiful: Essaying two samples

As discussed in Chapter 2.1, modern linguistics re-enacts the centuries-old controversy in European thought as to whether sense data or intuitions of the mind should be considered the basis of knowledge and truth. Section 3.5.1 of this chapter, as a prelude to the corpus-based research, describes a 'brain and eye' driven examination of random samples taken from two of the corpora (5% or 5,000 words from each). The outcome of this labour-intensive analysis was subsequently compared with that of the computer-based approach. This comparison provides information about the degree of correspondence between prefabs obtained using a frequency-based semi-automated approach and those deemed to be prefabs on the basis of native speaker intuition.

Random samples (approximately 5%) were taken of the essays in *Porticle* and *Locness*. These two corpora were chosen because they represent the two groups of least expert writers and because *Locness* was originally designed as a control corpus for the *ICLE* sub-corpora.

Using a random number table to select files, the samples were designed so that each contained a closely similar number of words (5143 and 5067 respectively). These samples were then subjected to a close scrutiny to assess the number and kinds of prefabs contained in each. Appendices 3 and 4 contain some of the sample essays from *Porticle* and *Locness* respectively. The purpose of this non-computerized treatment of samples from the two wholly undergraduate corpora was to obtain an independent measure of the prefabs in these corpora which might be used to gauge the accuracy of the computer corpus analysis. This juxtaposition of the brain and the computer is reminiscent of one of the major questions underlying this thesis. Prefabs are copiously produced by native speakers in both spoken

and written language. A major challenge for advanced learners is to memorize and be able to recognise at least some of this wealth of phrases and to be able to weave them into their own language use. A great deal of native-speaker phraseological competence is implicit knowledge. NNSs need to distinguish prefabs from non-idiomatic language in order to learn them. If this identification of prefabs can be made explicit or even automated, NNSs might be helped to learn them more easily (Schmidt and Frota 1986).

It is assumed that native speakers have a large number of these prefabs in their mental lexicon and in this experiment my native speaker intuition was used to recognize prefabs in the text. The sequences that I judged to be prefabs were highlighted so that, later, the amount of prefabrication in the two 5,000-word samples could be calculated. Obviously the outcome of this manual decision procedure revealed prefabs which could not possibly appear in the full-corpus analysis of *Porticle* and *Locness* because they occurred only once or fewer times than the set limit or floor (specified minimum number of occurrences) which was used in the automated search.

This small-scale manual approach should be viewed as complementary to the larger computer-run search. The more exhaustive four-corpora computer analysis is the central part of the investigation. The search of the two 5,000-word samples provides a rough estimate of the number of lexical, grammatical and pragmatic prefabs to be found in the NNS and NS writing. This examination of all the word sequences in a small number of essays also gives a useful overview of the range of prefab types in the larger corpora. The experiment brings out the underlying tension between the two main criteria for prefabhood, that of frequency of occurrence, which the computer handles well, and the degree of conventionalization which the knowing human subject recognizes well. If computer analysis can isolate the majority of prefabs, software might be developed for students to use on corpora to extract prefabs.

For the purposes of this experiment with 5,000-word samples, I adapted the schema employed by Wiktorsson (2003), by simplifying some of her categories, most notably that of grammatical prefabs. In this category, I retained determiners, quantifiers, intensifiers, PRO-forms, and linking prefabs but I eschewed markers of mood, tense and aspect (Figure 3.2). Wiktorsson herself (2003) merely counts these kinds of grammatical prefabs but does not analyse them further, admitting that the learning of such sequences would have more to do with a language learner's acquisition of the grammar of the language than the lexis.

In the count of noun phrase (NP) prefabs and verb phrase (VP) prefabs, I did not distinguish between closed and open prefabs, i.e. where the sequence contained slots filled by open class items marked X in Wiktorsson (2003). I added collocation frames to her lexical prefab category. The adjective and adverb phrase categories were telescoped together because of their small numbers in the final version of the table and are referred to jointly as AP in Table 3.1.

Figure 3.2 Classification of prefabs (adapted from Wiktorsson 2003)

lexical prefabs	**grammatical prefabs**	**pragmatic prefabs**
noun phrase prefabs	determiners	textual prefabs
verb phrase prefabs	quantifiers	epistemic prefabs
adjective phrase prefabs	intensifiers	attitudinal prefabs
adverb phrase prefabs	PRO-forms	
prepositional phrase prefabs	links	
clausal prefabs		
collocational frames e.g. *the* _____ *of*		

Table 3.1 gives the raw frequency of occurrence of each of these kinds of prefabs in the two sample sub-corpora. Although the sub-categories of lexical prefabs are itemized in the table, the figures for the grammatical and pragmatic prefabs are entered in the table without further analysis into sub-categories. Two entries in this table attract immediate attention. There are almost twice as many noun phrases in the sample from *Locness* as compared to the *Porticle* sample. The NNSs deploy far fewer noun phrase prefabs in their writing. On the other hand, they use twice as many pragmatic prefabs compared to their American counterparts.

Table 3.1 Results of non-computerised examination of 5,000 word samples with prefabs arranged according to type

Prefab type	Prefabs found in *Porticle* sample	Prefabs found in *Locness* sample
NP	157	286
VP	133	172
AP	36	37
PP	27	59
Clausal prefabs	16	7
Grammatical prefabs	58	39
Pragmatic prefabs	103	54
Total	530	654

A projection was made of the number of prefabs there might be in the two undergraduate corpora based on the search by hand of the two 5,000-word samples (Table 3.2). As all prefabs contain at least two words, a rough estimate can be made of the percentage of prefabs in the two corpora.

If the projected number of prefabs per 100,000 are divided by 1,000 to convert them into percentages, then at least 21% (10.6 x 2) of *Porticle* and 26% of *Locness* consists of prefabs.

Table 3.2 Projection of the total number of prefabs in the two undergraduate corpora based on the search by hand

Corpus	*Porticle*	*Locness*
Projected total prefabs in corpus	13,393	15,750
Normalized to 100,00 words	10,600	13,080

3.7 Comparing two corpora of apprentice writing

In preparation for the computer analysis of the four assembled corpora, a comparison was made of the two corpora containing writing by the least experienced academic writers. Firstly, a simple comparison was made of the thirty most frequent single words in the *Porticle* corpus and in the *Locness* corpus. The decision to study single words first was an attempt to follow a more bottom-up approach, at least in the early stages of the investigation.

The corpora used in this investigation were relatively small as compared to general reference corpora such as the British National Corpus or Bank of English and, consequently, provided fewer instances of lower frequency words and phrases. With this in mind, I began my study with the most frequent words, that is, the grammatical words, not to investigate these words as such, but rather the phraseologies of which they form a part. In this way, useful overall tendencies could be obtained for each corpus. Word frequency lists from corpora can be compared automatically using text retrieval software and significant differences of frequency can be noted. This comparative analysis is done firstly on single words and then on clusters (sequences) of 2, 3, 4 and more words. The early pages of the word frequency list for most corpora contain mainly function words. When a content or lexical word appears in the first hundred or so most frequent words, it is worth making a concordance of this word to obtain a profile of its typical collocations and colligations. A careful note should be taken of this and other highly frequent lexical words.

The four corpora studied were of varying sizes, that is to say they had differing numbers of running words (or tokens to use the term most popular in computational linguistics). For this reason, the raw figures for the occurrences of tokens and sequences of tokens in each corpus were normalized to express the number of occurrences per 100,000 tokens. This normalization was carried out in order to enable direct comparison among the corpora, using the formula provided by Biber, Conrad and Reppen (1998:263) for normalizing frequencies of occurrences. The decision was taken to normalize to 100,000 tokens instead of the more usual 1,000 or 10,000 tokens because all four corpora contained more than 100,000 tokens. For those comparisons where the size of the corpus affects the results of tests of significance, these normalized statistics have been used. Whenever the

frequency of occurrence was available as a percentage, for instance in the case of highly frequent items, these percentages were directly comparable and normalization was not necessary. In certain comparisons carried out using Wordsmith Tools and the Wmatrix software, the log likelihood measure of significance is obtained. This measure already allows for corpus size and, therefore, obviates the need for carrying out such a process of normalization.

The most striking difference between the two corpora, at first sight, is the apparent underuse by learners in *Porticle* of the definite article in comparison with *Locness* (Table 3.3 line 1). It was not necessary to normalize the results in this table as the occurrences are given both in both raw frequencies and as percentages of the total number of words in the corpus. Occurrences of *the* make up 4.94% of the total words in *Porticle* as compared with 6.28% in the *Locness* corpus. The native speakers were using it 1.38 times more frequently. Johansson's (1985:30) analysis of the word frequencies in the Lancaster Oslo Bergen Corpus (LOB) shows that the definite article occurs with the highest frequency in Category J (learned texts), with a percentage slightly higher than that found in *Locness*. Greater use of the definite article seems to be characteristic of informative written prose. Corroboration of Johansson's finding was obtained from the *BNCBaby* (Berglund and Wynne 2005). Occurrences of *the* make up 6.14% of a corpus of approximately one million words of academic prose, 5.34% of a similar sized corpus of newsprint, 4.53% of a million words of fiction and 2.35% of a million words of transcribed conversation. The descending frequency of occurrence of the definite article from the most to the least informative kind of language in the *BNCBaby* shows the same unmistakable pattern detected by Johansson 20 years earlier.

The usual collocate first to the right when *the* is the node word is typically a noun or nominalization element. Taking this into consideration, the underuse of *the* seems to indicate fewer nouns in the learner corpus. It could be hypothesized that the noun vs. verb balance in the two corpora under comparison is different or that the amount of nominalization is less in the learner corpus. This result was later corroborated using part of speech (POS) tagged versions of the corpora: these results can be found in Table 3.13. An interesting extension to this indirect estimate of noun use emerges when occurrences of *an* in the corpora are compared: the native speakers have a noticeably higher frequency of use

compared to the NNSs. The z-test shows this difference between the NNS and NS use of *an* to be significant at p= .01.

Table 3.3 The thirty most frequently occurring word-forms in *Porticle* and *Locness*

	30 most frequent word-forms in *Porticle*					30 most frequent word-forms in *Locness*		
	Word	**Freq.**	**%**			**Word**	**Freq.**	**%**
1	*THE*	6,244	4.94		1	*THE*	7,564	6.28
2	*TO*	4,499	3.56		2	*TO*	3,811	3.16
3	*AND*	3,854	3.05		3	*OF*	3,678	3.05
4	*OF*	3,378	2.67		4	*AND*	2,899	2.41
5	*THAT*	3,040	2.41		5	*A*	2,658	2.21
6	*A*	2,836	2.24		6	*IN*	2,271	1.89
7	*IN*	2,704	2.14		7	*IS*	2,191	1.82
8	*IS*	2,586	2.05		8	*THAT*	2,186	1.82
9	*WE*	1,791	1.42		9	*FOR*	1,216	1.01
10	*ARE*	1,514	1.20		10	*BE*	1,148	0.95
11	*IT*	1,478	1.17		11	*ARE*	1,111	0.92
12	*I*	1,470	1.16		12	*IT*	1,066	0.89
13	*HAVE*	1,318	1.04		13	*THEY*	1,027	0.85
14	*THEY*	1,241	0.98		14	*THIS*	1,006	0.84
15	*FOR*	1,164	0.92		15	*NOT*	1,001	0.83
16	*BE*	1,144	0.91		16	*AS*	903	0.75
17	*THIS*	1,129	0.89		17	*HAVE*	811	0.67
18	*NOT*	1,008	0.80		18	*WITH*	732	0.61
19	*WITH*	912	0.72		19	*THEIR*	722	0.60
20	*PEOPLE*	858	0.68		20	*ON*	713	0.59
21	*BUT*	845	0.67		21	*OR*	602	0.50
22	*OUR*	770	0.61		22	*BY*	569	0.47
23	*ALL*	728	0.58		23	*PEOPLE*	543	0.45
24	*AS*	716	0.57		24	*WOULD*	506	0.42
25	*CAN*	712	0.56		25	*AN*	486	0.40
26	*OR*	663	0.52		26	*IF*	467	0.39
27	*THEIR*	638	0.50		27	*WAS*	466	0.39
28	*MORE*	600	0.47		28	*ONE*	451	0.37
29	*BECAUSE*	592	0.47		29	*MORE*	446	0.37
30	*WORLD*	553	0.44		30	*FROM*	431	0.36

Hofland and Johansson (1982:22) suggested that the high frequency of *an* found in written informative prose indicated a high proportion of Latinate vocabulary. It might be surmised that my Portuguese writers use fewer Latinate word tokens despite their mother tongue's close filiation with Latin and this again points to a style more used in spoken English than in written academic English. There is a need for caution, however, as learners may misuse 'a' for 'an', thus reducing the number of occurrences of 'an' but not necessarily using fewer words beginning with a vowel. Wildcard concordances on *Porticle*

using searchwords a... a*, a... e*, a... i*, a... o* and a... u* quickly obtains the number of misuses of 'a' in *Porticle.* With words beginning with *a, e, i, o* and *u* there were 16 misuses which does not greatly alter the significance of the difference found between the NNSs and the NSs.

Table 3.4 Search for articles *a, an* and *the* in the four corpora

Corpus	*Porticle*	*Locness*	*Bawe*	*CofE*
occurrences of *an* per 100,000 words	340	486	464	460
occurrences of *a* per 100,000 words	2244	2199	1938	2177
occurrences of *the* per 100,000 words	4941	6281	6844	6745

To continue with a simple visual comparison, another interesting feature of *Porticle* was the appearance of the personal pronoun *we,* with 1,791 occurrences (1.42%), followed closely by the first person singular pronoun *I,* with 1,470 occurrences (1.16%). These two pronouns are noticeably absent from the top 30 most frequent words in *Locness.* Finally, the two lexical words among the thirty most frequent words in *Porticle* were *people* (853 times = 0.68%) and *world* (0.44%). *Locness* writers used *people* somewhat less frequently (0.45%) as the only lexical word in their 30 most frequent words. Ringbom (1998) suggests that *ICLE* contributors of all nationalities tend to overuse vocabulary items of high generality, such as *people* and *things.*

This initial analysis of the target corpora using single words as search words quickly reveals clear lexical differences and paves the way for further analysis of the phraseological patterning. The single-word analysis provides the researcher with lines of enquiry in the search for and discovery of multiword units. A further, more powerful way to compare frequency wordlists is to use the Compare Two Wordlists function on Wordsmith Tools. This approach compares the frequencies of words in each corpus and lists those word-forms which are relatively overused or underused in one of the corpora compared with the frequency of use in the other corpus. Scott (1999) uses the term *keyness* to refer to this relationship between expected frequencies and observed frequencies. In the Wordsmith Tools suite of programs, the Keyword device is based on the log likelihood formula (Dunning 1993). Log likelihood is judged to be a more reliable measure of surprisingness

as it is less influenced by rare events than are other available measures, such as mutual information or Z-score. A log likelihood greater than 6.7 is considered to be statistically significant at 0.05 level of significance (Rayson 2003). Comparing *Porticle* and *Locness* frequency wordlists by this method revealed certain differences between the corpora which then enable further hypotheses to be made. These hypotheses were subsequently tested by examining the texts of the essays. This cyclical approach is an integral part of the methodology followed by many corpus linguists and is clearly chronicled in several books by Sinclair (e.g. Sinclair 1991, 2003).

Table 3.5 Keywords which are more frequent in *Porticle* than in *Locness*

	word	freq.	porticle%	freq.	locness%	keyness
1	*we*	1,791	1.42	273	0.23	1,185.
2	*I*	1,470	1.16	383	0.32	633.5
3	*world*	553	0.44	88	0.07	354.6
4	*our*	770	0.61	206	0.17	321.6
5	*imagination*	202	0.16	2		250.9
6	*dream*	194	0.15	5		220.3
7	*think*	425	0.34	98	0.08	205.4
8	*nowadays*	164	0.13	2		200.9
9	*us*	346	0.27	65	0.05	197.9
10	*don't*	331	0.26	71	0.06	170.4
11	*dreaming*	118	0.09	0		158.0
12	*technology*	166	0.13	11		155.7
13	*my*	428	0.34	129	0.11	155.5
14	*dreams*	115	0.09	1		143.9
15	*things*	314	0.25	77	0.06	143.1
16	*countries*	152	0.12	11		138.8
17	*but*	845	0.67	417	0.35	128.8
18	*degree*	109	0.09	2		128.8
19	*all*	728	0.58	348	0.29	119.9
20	*theoretical*	100	0.08	3		111.1
21	*that*	3,040	2.41	2,186	1.82	104.2
22	*have*	1,318	1.04	811	0.67	99.5
23	*subjects*	73	0.06	0		97.7
24	*and*	3,854	3.05	2,899	2.41	96.0
25	*money*	495	0.39	220	0.18	96.0
26	*opinion*	137	0.11	19	0.02	95.1
27	*practical*	71	0.06	0		95.1
28	*science*	122	0.10	14	0.01	93.3
29	*live*	224	0.18	62	0.05	89.9
30	*course*	156	0.12	30	0.02	87.6
31	*portugal*	65	0.05	0		87.0
32	*kind*	122	0.10	17	0.01	84.5
33	*europe*	68	0.05	1		82.0
34	*degrees*	79	0.06	4		79.4
35	*industrialisation*	59	0.05	0		79.0
36	*prepare*	74	0.06	3		78.0
37	*sometimes*	130	0.10	24	0.02	75.2
38	*can*	712	0.56	401	0.33	74.0
39	*i'm*	98	0.08	12		72.6

40	*european*	72	0.06	4		70.8
41	*like*	340	0.27	147	0.12	69.7
42	*imagine*	70	0.06	4		68.3
43	*future*	137	0.11	31	0.03	67.3
44	*portuguese*	50	0.04	0		66.9
45	*real*	159	0.13	43	0.04	65.4
46	*some*	495	0.39	258	0.21	65.1
47	*say*	230	0.18	84	0.07	63.8
48	*internet*	47	0.04	0		62.9
49	*don*	52	0.04	1		61.1
50	*everybody*	60	0.05	3		60.5

The prevalence of *dream*, *dreaming* and *dreams* and *imagination* among the Keywords in the *Porticle* corpus can be largely attributed to essay title no 14 (see Appendix 2):

> Some people say that in our modern world, dominated by science technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion ?

The overuse of *things* by the *Porticle* contributors, mentioned earlier in this section, is an example of NNS writers' recourse to general delexicalized words when a more precisely lexicalized term would provide the level of formality required. If the first five lines of the concordance made using *things* as a searchword in *Porticle* is examined, it becomes clear that less abstract nouns would give more focus and a greater degree of formality to the texts in which they occur. Suggestions are included in parenthesis at the end of each concordance line.

Concordance of THINGS as searchword in *Porticle* corpus (first five of 314 lines)

1	believe that these are the most important things that we have. I'm going to give (FACULTIES)
2	rights and even self-esteem. But not all things can be solved just with words! I'm (PROBLEMS)
3	about the reality of our country, about the things that happened at the time...So, (EVENTS)
4	Censorship is one of the most terrible things in our society. Nowadays, the (PRACTICES)
5	industrialisation and science are great things. Without them we could not (ADVANCES)

In some of the tables in this chapter, the raw scores have been converted into occurrences per 100,000 words of corpus. As previously mentioned, this conversion is done where it is needed to facilitate easier comparison between corpora of varying sizes. It also enables extrapolation to other corpora beyond the present study. The Portuguese writers in *Porticle* use *I* 1161 times per 100,000, while the American undergraduates in *Locness* use

the first person singular personal pronoun 318 times per 100,000 words. The uses of *our* 609 per 100,000 words vs. 171 per 100,000 words, *us* 273 vs. 54 occurrences per 100,000 words, and *my* 339 vs. 107 occurrences per 100,000 words, follow a similar pattern and show that there is a very different deployment of the pronoun system between the two groups. This overuse of *I* is much greater than in any other language group in *ICLE*. There are 285 occurrences per 100,000 words in *Spicle*, the Spanish sub-corpus of *ICLE*, and 294 occurrences per 100,000 words in *Bricle*, the Brazilian sub-corpus. Thus the number of occurrences of *I* per 100,000 words in *Porticle* is nearly four times as many as that found in *Bricle,* which, to say the least, is a surprising difference between two groups of similar age and educational level with the same first language. A closer analysis of this question will be made in Chapter 4.1.

Opinion is used more than seven times more frequently by the Portuguese learners as compared with the US undergraduates (108 vs. 15 occurrences per 100,000 running words). A quick perusal of a concordance of the Portuguese uses of *opinion* showed that all but a very few occurrences were used in trigrams, 4grams and 5grams in the expressions, *in my opinion*, *in my humble opinion,* and *it is my opinion that*. Again, more will be said on these expressions in the following chapter.

The Portuguese writers greatly underused metalingual terms useful in conducting arguments, e.g. *argument, opponents, evidence, advocates, and proponents* (Table 3.6). The *Locness* writers wielded these terms to good effect in their essays to refer to the different sides in the argument they were conducting. The *Porticle* writers resort to periphrasis, for example *some people say* (7 occurrences) *some people think* (3 occurrences) *some people believe* (2 occurrences) and *some people argue* (1 occurrence).

Table 3.6 Keywords which occur more frequently in *Locness* compared to *Porticle*

	word	freq.	port.lst %	freq.	loc.lst %	keyness
249	*biological*	1	39	0.03		48.0
250	*richard*	0	34	0.03		48.8
251	*players*	0	34	0.03		48.8
252	*support*	35	0.03	115	0.10	48.9
253	*on*	499	0.39	713	0.59	49.3
254	*gun*	1	40	0.03		49.3
255	*claim*	17	0.01	83	0.07	50.7
256	*season*	0	36	0.03		51.7
257	*nuclear*	2	46	0.04		52.1
258	*program*	2	46	0.04		52.1
259	*into*	70	0.06	176	0.15	52.5
260	*water*	8	66	0.05		54.7
261	*flag*	3	52	0.04		55.4
262	*Out*	99	0.08	224	0.19	56.0
263	*athletes*	1	45	0.04		56.3
264	*playoff*	0	40	0.03		57.4
265	*bowl*	0	40	0.03		57.4
266	*aids*	5	60	0.05		57.6
267	*divorce*	1	46	0.04		57.7
268	*genetic*	1	47	0.04		59.1
269	*drug*	27	0.02	112	0.09	60.0
270	*recipients*	0	42	0.03		60.3
271	*drinking*	4	61	0.05		62.9
272	*state*	24	0.02	109	0.09	63.0
273	*advocates*	1	51	0.04		64.7
274	*she*	78	0.06	206	0.17	66.2
275	*age*	19	0.02	102	0.08	66.6
276	*florida*	0	47	0.04		67.5
277	*college*	10	84	0.07		70.3
278	*patients*	0	49	0.04		70.3
279	*whether*	7	76	0.06		70.4
280	*team*	1	57	0.05		73.1
281	*proponents*	0	52	0.04		74.6
282	*these*	203	0.16	400	0.33	75.6
283	*football*	3	70	0.06		79.5
284	*marijuana*	5	78	0.06		80.9
285	*euthanasia*	5	78	0.06		80.9
286	*teams*	0	60	0.05		86.1
287	*her*	59	0.05	198	0.16	86.3
288	*death*	33	0.03	152	0.13	88.9
289	*evidence*	3	77	0.06		89.0
290	*public*	24	0.02	135	0.11	91.0
291	*article*	2	79	0.07		97.3
292	*states*	44	0.03	185	0.15	100.4
293	*opponents*	1	82	0.07		108.2
294	*welfare*	12	132	0.11		123.0
295	*suicide*	9	130	0.11		132.0
296	*sex*	15	0.01	149	0.12	133.7
297	*prayer*	0	132	0.11		189.5
298	*the*	6,244	4.94	7,564	6.28	209.9
299	*argument*	4	194	0.16		244.8

3.8 Using collocational frames for corpus analysis

In the early stages of the investigation described in Section 3.6, two 5,000-word samples from *Porticle* and *Locness* were examined word by word. Although the *Locness* writers appeared to use more prefabs than the NNSs, the *Porticle* writers also used a great deal of them. Both NS and NNS writers used more than 500 prefabs per 5,000 words, which meant that more than 20% of their text was prefabricated, given that each prefab contains at least two words.

The next phase in the research programme described in Section 3.7 was to conduct a computer-driven search of *Porticle* and *Locness* to obtain word frequency lists. The differences in the frequency of certain grammatical words suggested that there might be stylistic differences between the native and non-native undergraduate writers. The two corpora might differ, for example, in relation to the level of formality/informality displayed by their writers. The Keyness or Log Likelihood measure of expected versus observed frequency highlighted this contrast when the *Porticle* and *Locness* frequency lists were compared.

Large differences were found in pronoun use and in the use of general terms such as *things* and *people*. Whenever such differences appeared, further corpus analysis was carried out to compare the *Porticle* corpus and the three NS control corpora. In some situations, I compared *Porticle* with *Locness* and decided not to follow a particular line of enquiry with the other two control corpora. Although the analysis of word frequency lists is a useful method of taking the measure of a corpus or a set of corpora, newer more searching devices have been developed which quickly reveal patterns to the corpus linguist.

Collocational frames, discussed in Chapter 1.8, display the main paradigmatic choices of verbs, nouns and adjectives in the corpus. This provides information about the thematic content of the texts which make up the corpus, aided by the fact that the frames consist of words which occur with very high frequency and are evenly distributed through the text. This thematic content has been referred to as 'aboutness' (Phillips 1984; Scott 1999). The kinds of nouns, verbs and adjectives found in these frames can serve as an indication of what the texts are about. It is important to remember that Phillips (1984) introduced this term to refer to the 'aboutness' of texts (*qua* textbooks) and it is doubtful

whether a corpus can have aboutness in the same way, given its more heterogeneous nature. Using collocational frames produces sufficient examples of less frequent lexical items thus making possible some kind of generalization or prediction about them. In the case of nouns, for example, searching for the frequent collocational frames *a* ____ *of*, *an* ____ *of* and *the* _____ *of*, and sorting alphabetically on the slot filler or varying element in the node, i.e. the word which fills the blank, produces a list of the nouns found in these structures. Individually, few of these nouns occur with a frequency likely to arouse interest but collectively the instances of the frame or pattern merit attention. A search of the four corpora shows that by far the most frequent collocational frames in each corpus are the definite and indefinite article followed by a slot and then by the preposition *of* (Table 3.7). This is not surprising given the very high frequencies in almost all English language corpora of these two grammatical words.

As might be expected from the differences in the occurrence of *the* between the NNS corpus and the NS corpora, *Porticle* writers underuse the collocational frame *the...of* significantly (the z-test for comparing the difference between two proportions gives a significant difference between *Porticle* vs. *Locness* with p= .01).

The collocational frame captures the noun, verb or adjective in its typical environment. Given that the highly frequent function words framing these lexical items constrain the part of speech to be found in the variable slot, collocational frames can be used to access an otherwise unannotated corpus from a grammatical perspective. For example, the frames *a* _____ *of, many* _____ *of* and *the* _____ *of* will normally have nouns in their variable slot, *too* _____ *to* is usually filled by an adjective and *we* _____ *that* almost always has a verb in the variable position. Viewed from the perspective of the individual words in the variable slot, the collocational frame allows colligational information to be obtained.

Table 3.7 Frequency of collocational frames per 100,000 words

Corpus	*Porticle*	*Locness*	*Bawe*	*CofE*
the ... of	758	1042	1394	1128
the ...that	247	242	129	131
a ... of	220	205	274	234
be ... to	93	102	75	78
an ... of	22	40	40	36
we ...that	23	5	2	10
too ... to	6	5	6	13
many ...of	13	2	8	8
for ... of	9	17	6	24
had ...of	1	2	-	-

In the collocational frame *the* _____ *of* the *Locness* writers use 14 types (113 tokens) in plural form, while the *Porticle* writers use only three types (17 tokens) (Tables 3.8 and 3.9). At this exploratory stage, I judged it worthwhile to ascertain whether this 'brute fact' or rough data had any significance. This intuition was based on one of my earliest experiences in corpus linguistics: the discovery that, semantically, the singular and the plural of a noun often behave very differently. For instance, I had noticed that the noun *field* when used in the singular was often used figuratively to mean domain or area of study or interest but, when used in the plural, more often referred to areas of agricultural land. On checking the literature, I became aware that these semantic differences between word-forms of a lexeme had already been commented upon (e.g. Sinclair 1991).

Table 3.8 Collocational frame <the ...of> in *Porticle*

		Freq.				Freq.
1	*the root of*	39		26	*the kind of*	6
2	*the end of*	25		27	*the needs of*	6
3	*the opium of*	20		28	*the number of*	6
4	*the rest of*	19		29	*the purpose of*	6
5	*the idea of*	14		30	*the responsibility of*	6
6	*the beginning of*	13		31	*the sense of*	6
7	*the lack of*	13		32	*the cause of*	5
8	*the majority of*	12		33	*the criticism of*	5
9	*the quality of*	11		34	*the history of*	5
10	*the use of*	11		35	*the language of*	5
11	*the degree of*	10		36	*the middle of*	5
12	*the concept of*	9		37	*the problem of*	5
13	*the love of*	9		38	*the story of*	5
14	*the case of*	8		39	*the area of*	4
15	*the help of*	8		40	*the capacity of*	4
16	*the origin of*	8		41	*the construction of*	4

17	*the power of*	8		42	*the fact of*	4
18	*the consequences of*	7		43	*the invention of*	4
19	*the development of*	7		44	*the learning of*	4
20	*the evolution of*	7		45	*the matter of*	4
21	*the importance of*	7		46	*the nature of*	4
22	*the role of*	7		47	*the people of*	4
23	*the world of*	7		48	*the point of*	4
24	*the declaration of*	6		49	*the process of*	4
25	*the hands of*	6		50	*the question of*	4

Table 3.9 Collocational frame <the ...of>in *Locness*

		Freq				Freq
1	*the use of*	31		26	*the university of*	9
2	*the idea of*	25		27	*the consequences of*	8
3	*the number of*	24		28	*the proponents of*	8
4	*the root of*	20		29	*the types of*	8
5	*the amount of*	18		30	*the winner of*	8
6	*the case of*	18		31	*the love of*	7
7	*the majority of*	17		32	*the people of*	7
8	*the end of*	16		33	*the right of*	7
9	*the effects of*	15		34	*the history of*	6
10	*the question of*	15		35	*the members of*	6
11	*the teaching of*	15		36	*the minds of*	6
12	*the beginning of*	14		37	*the practice of*	6
13	*the rest of*	14		38	*the problems of*	6
14	*the value of*	14		39	*the rate of*	6
15	*the issue of*	13		40	*the rewards of*	6
16	*the lack of*	13		41	*the risk of*	6
17	*the opponents of*	13		42	*the author of*	5
18	*the concept of*	11		43	*the benefits of*	5
19	*the cost of*	11		44	*the center of*	5
20	*the problem of*	11		45	*the heart of*	5
21	*the life of*	10		46	*the lives of*	5
22	*the loss of*	10		47	*the possibility of*	5
23	*the rights of*	10		48	*the presence of*	5
24	*the supporters of*	10		49	*the results of*	5
25	*the age of*	9		50	*the spread of*	5

Versions of both corpora tagged using the TOSCA-ICLE tagger (see Appendix 1) were searched to check whether there was an overall difference in the frequency of use of the plural in the writings of *Locness* and *Porticle*. There appears to be no significant difference in the overall use of the plural between the two groups of writers (Table 3.10).

Table 3.10 Singular and plural use of nouns in each corpus

Corpus	*Porticle*	*Locness*
singular noun tokens	12661	13287
plural noun tokens	6199	6732
Total	18860	20019
% of nouns in plural form	33%	33%

The large amount of overlap between the nouns occurring in the top 50 collocational frames of *Porticle* and *Locness* (shown in Table 3.9) is quite surprising. (The exception, *the opium of,* achieved a high score in the *Porticle* list purely because an essay question had been set for the Portuguese subjects asking whether television had become the new opium of the masses). It seems that argumentative prose requires the recurrent use of a certain set of nouns embedded in frames. This patterning of academic prose opens up a phraseological dimension which could link up with, and build upon, work already done on academic wordlists which usually only concentrate on single words, for example the Academic Word List (AWL) developed by Coxhead (1997).

3.9 Lexical density of the four corpora

One approach to comparing the writing in the four corpora is to measure the lexical density of each corpus to ascertain whether there are lexical differences among them. The concept of lexical density was discussed in Chapter 1.3 and there it was decided that there was no simple correspondence between written and spoken English on the one hand and greater and lesser lexical density on the other. The decisive factor was the degree of interaction possible between the speaker/writer and the listener/reader so that some written texts could have relatively low lexical densities while certain spoken texts could manifest high levels of lexical density. Other things being equal, if there was little or no possibility of interchange then the text needed to be much more explicit and was therefore more lexically dense.

One measure of lexical variation is type-token ratio. The type-token ratio

(henceforth TTR) is the number of word types in each corpus divided by the number of tokens (Table 3.11). One of the limitations of TTR as a measure of the lexical variation of a text is that it is much too sensitive to the size of a corpus. This deficiency of TTR was discussed above in Chapter 1.10. The elaboration of the standardised TTR was an attempt by the designer of Wordsmith Tools (Scott 1999) to counteract this deficiency of TTR. By averaging the TTRs of successive sections of text (choosing from a section size ranging from 10 to 20,000 words), the standardized type/token ratios (STTRs) can be used to compare texts of differing lengths.

Table 3.11 Type-token ratios of the four corpora

CORPUS:	***Porticle***	***Locness***	***Bawe***	***CofE***
tokens	126,368	120,415	108,912	113,101
types	7,676	9,053	10, 634	12,596
type/token ratio (TTR)	6%	7.5%	9.8%	11.1%
standardized type/token ratio (STTR)	57.62	59.02	60.68	64.80

If STTR is used as a measure of lexical variance instead of TTR, the differences across the corpora are less pronounced. The TTR of *CofE* is almost double that of *Porticle,* whereas the STTR of *CofE* is approximately 13% greater than that of *Porticle* (Table 3.11). Although it may be an improvement on TTR as a measure of lexical variance, STTR does not seem to be a suitable measure for the present study. As mentioned in Chapter 1.10, Meunier (1998:32) found that standardized type-token ratio was not a discriminating feature between NS and NNS writers and that lexically rich essays were not necessarily good quality ones.

In view of the dependence of TTR on corpus size and the unsuitability of STTR, I decided to investigate the lexical density of the corpora using the method developed by Ure (1971) in her pioneering work on lexical density. Chapter 1.10 described how Ure used the proportion of lexical to grammatical words in text to measure the lexical density of texts and to situate them on an idealized spoken-written continuum (Ure 1971; Stubbs 1996). Lexical density expresses the proportion of a text or a corpus which is made up of lexical or content words, expressed as a percentage of the total number of words.

The lexical density is calculated by first finding the number of content words in each corpus. The content words are obtained by using a stoplist of the 100 most frequent words in each corpus, almost all of which are grammatical or closed class words. A wordlist of the word types in the corpus, not including these 100 most frequent words, is then made. This gives a word frequency list of the predominantly non-grammatical, i.e. content or lexical words in the corpus. Developing a suggestion from Halliday (1989:64), this count of the content words obtained for each corpus was divided by the total number of tokens in the corpus and the quotient converted to a percentage (Table 3.12).

Table 3.12 Lexical densities of the four corpora (using stoplist of 100 most frequent structural words)

Corpus	***Porticle***	***Locness***	***Bawe***	***CofE***
Lexical density using stoplist	46.51%	51.14 %	54.97%	53.22%

The NS corpora all have a higher lexical density than *Porticle*. They therefore have a significantly greater information load than *Porticle* when the x-test for comparing proportions is applied (see Appendix 7). This finding could be taken as proof of the not too surprising fact that the native speakers have a richer lexis at their command. On the other hand, this finding might point to a fundamentally different approach to meaning by the NS writers. The higher lexical density would be attributable to greater use of less frequent nouns. This interpretation follows Halliday's (1989) suggestion that lexical or content words be divided into high-frequency, relatively delexicalized items and low frequency high information lexical items. Halliday (1989:65) proposed giving the low-frequency lexical items double weighting compared to the weighting given to high-frequency items in a measure of lexical density. There are more noun phrases in the essays in the *Locness*, *Bawe* and *CofE* corpora and such phrases would appear to be core building blocks of written prose. The need for non-native writers to use more noun phrases in their writing will be examined in Chapter 5.

As reported earlier in this section, Meunier (1998) points out that NS writers can display lower lexical density than NNS writers in similar writing tasks. She also states that

lexical density does not necessarily correlate highly with quality of writing in NNS essays (Meunier 1998:32). In a sense, the lexical variation noted in Table 3.12 merely corroborates what seems intuitively to be the case: that while higher lexical density is likely to be related to such stylistic factors as precision and elegance of expression - i.e. overall aesthetic appeal or a certain compelling quality - the relationship is not a necessary one. That is, greater lexical density may coexist with imprecision of expression and with formal inaccuracy (grammatical, semantic, and idiomatic). The number of different words is less important than what is done with the words. In view of these considerations, measures of lexical density as an index of writing proficiency or nativeness will not be pursued further in the present research.

The reservations about TTR and other measures of lexical variance recall the growing realization among certain corpus linguists (Sinclair, 1991; Hunston and Francis 2000) that the distinction between lexis and grammar is hard to maintain. The delexicalization of certain words (*have, get, take*) and the semantic role played by highly frequent ostensibly grammatical words are among the phenomena which led linguists to question the division between lexis and grammar. There is, for example, a modern tendency for EFL dictionaries to become more like grammars and for grammars to become more like dictionaries (Carter 2005).

3.10 Findings from syntactic and semantic tagging

The four corpora were tagged with CLAWS7 and also with the TOSCA-ICLE tagsets. The intention was to investigate whether *Porticle* writers conformed to the pattern already found in the other subcorpora of *ICLE*. Ringbom (1998:43) found that the writers in all of the NNS sub-corpora used more main, modal and auxiliary verbs and fewer nouns, prepositions and adjectives than the NSs in the control corpora. Searching on Part of Speech (POS) tags showed that verbs and adverbs are the only parts of speech which the *Porticle* writers use proportionately more frequently than their native-speaker counterparts in the control corpora (Table 3.13).

Table 3.13 Occurrences of main parts of speech in each of the corpora expressed as a percentage

(based on the TOSCA-ICLE tagged version)

Corpus	*Porticle*	*Locness*	*Bawe*	*CofE*
nouns	33.7%	37.4%	39.9%	38.3%
verbs	32.9%	28.77%	23.6%	23.2%
adverbs	11.69%	9.8%	8.7%	9.1%
adjectives	10.8%	11.4%	12.73%	12.2%
prepositions	13.7%	14.7%	17.5%	17.2%

Phrasal verbs are prefabs par excellence as they have to be memorized as wholes and their meaning cannot usually be deduced from the meaning of their component parts. The semi-automated corpus analyst, Wmatrix, was used to obtain a rapid overview of phrasal verb deployment in the four corpora. This was achieved by searching for the CLAWS7 tag, RP, which stands for prepositional adverb or particle. Only those occurrences of prepositional adverbs or particles collocating with verbs were counted. *Locness* writers noticeably overuse phrasal verbs compared to the other three corpora (Table 3.14).

Table 3.14 Occurrences of adverb or preposition particles in *Porticle* and *Locness*

(frequencies normalized to a basis per 100,000 words of text)

Corpus	preposition or adverb particle
Porticle	97
Locness	200
Bawe	81
CofE	151

The most frequently used phrasal verbs were examined to see if the *Porticle* writers tended to use exemplars which are more usually found in spoken English than in written English as compared with the native-speaker controls. *Porticle* writers use the phrasal verbs

in many more basic senses; e.g. *go back* occurs ten times in *Porticle*. All the uses but the one below are used to refer to a return to a geographical or physical location. This citation from a *Porticle* essay shows *go back* used to organize and signpost within the text:

> Picking up from here I'm going back to our situation of future language teachers.

The writers in *Locness* used verbs followed by *around* 38 times; e.g. *centre around* (4 times) and *focus around* (twice). The two uses of *fall off* in *CofE* show a more figurative use of this phrasal verb:

> In Britain, their arguments just fell off the map.
> Nostradamus is falling off the best seller list.

More exhaustive study of the use of phrasal verbs in the four corpora was decided against, mainly because a great deal of research has already been done on phrasal verbs in linguistics, lexicography and in ELT materials (e.g. Bolinger 1975; Cowie and Mackin 1975; *COBUILD English Grammar* 2002).

In view of the need to specialize, it was decided to focus on adverb+adjective combinations and, more specifically, where the adverbs are functioning as *amplifiers* or *downtoners* to use the terms of Quirk et al. (1972). *The COBUILD English Grammar* (2002) explains this phenomenon to the student of English language employing a simpler nomenclature, namely the terms *modifier* and *submodifier*. These terms are used in what follows to avoid a proliferation of terms (cf. two subcategories of amplifiers: *maximizers* and *boosters* found in Quirk et al. 1985).

As reported in Chapter 2.4.1, Lorenz (1998:53-66) found differences in the frequency and range of combinations of adverbs of degree and adjectives among German writers of different levels of English proficiency and native speakers. Lorenz found that collocational associations between submodifiers and modifiers contribute a great deal to the idiomaticity of written texts. The main function of the submodifier is to assert the centrality or relevance of the quality expressed by the adjective (e.g. *fabulously interesting* (*Bawe*), *critically important* (*CofE*), *quite different* (*Porticle*)); and to add details to its core meaning (*perfectly good* (*Porticle*), *spectacularly bad* (*CofE*), *unusually high* (*Bawe*), *definitely hard* (*Locness*)). Lorenz (1998) shows that German learners of English overuse certain

submodifier-modifier combinations resulting in overstatement and sometimes, hyperbole in their writing (e.g. *very intelligent, absolutely stupid, very delicious, really impossible*).

Table 3.15 Recurrent submodifier-modifier combinations in four corpora

Porticle		*Locness*	
really	*important (7)*	*sexually*	*active (9)*
almost	*impossible (5)*	*sexually*	*transmitted (8)*
extremely	*important (4)*	*terminally*	*ill (7)*
how	*wonderful (4)*	*morally*	*wrong (6)*
completely	*different (3)*	*readily*	*available (4)*
essentially	*due (2)*	*extremely*	*important (3)*
extremely	*serious (2)*	*highly*	*respected (3)*
practically	*impossible (2)*	*virtually*	*impossible (3)*
quite	*different (2)*	*financially*	*stable (2)*
quite	*good (2)*	*how*	*genetic (2)*
really	*serious (2)*	*increasingly*	*weak (2)*
really	*necessary (2)*	*legally*	*wrong (2)*
socially	*acceptable (2)*	*mentally*	*ill (2)*
		morally	*correct (2)*
		nearly	*impossible (2)*
		predominantly	*white (2)*
		quite	*obvious (2)*

Bawe		*CofE*	
analytically	*valid (4)*	*currently*	*illegal (3)*
partially	*automatic (3)*	*politically*	*correct (3)*
privately	*rented (3)*	*critically*	*important (2)*
acoustically	*similar (2)*	*entirely*	*different (2)*
almost	*impossible (2)*	*equally*	*important (2)*
already	*existing (2)*	*especially*	*vulnerable (2)*
already	*known (2)*	*genetically*	*modified (2)*
completely	*different (2)*	*grossly*	*inflated (2)*
entirely	*different (2)*	*how*	*important (2)*
equally	*representative (2)*	*hugely*	*profitable (2)*
fully	*automatic (2)*	*nearly*	*unanimous (2)*
functionally	*distinct (2)*	*politically*	*incorrect (2)*
hugely	*profitable (2)*	*politically*	*motivated (2)*
increasingly	*important (2)*	*politically*	*motivated (2)*
theologically	*impeccable (2)*	*potentially*	*disastrous (2)*
		radically	*different (2)*

In a later study, Biber et al. (1999:545) observed a higher frequency of modifier with submodifier combinations in conversation than in academic prose. On the other hand, they found much more diversity in the modifiers and submodifiers that collocate in academic prose. The POS-tagged versions of the four corpora were used to mechanically extract all the SUBMODIFIER+MODIFIER combinations from the corpora. The results of the machine-search of the tagged corpora were adjusted in order to remove combinations which were due to tagging error. Occurrences of negatives were also filtered out (e.g. *not*,

never). All those submodifier-modifier combinations occurring in the corpora more than once are recorded (Table 3.15).

In *Porticle,* the occurrences of *really important* (7) and *extremely important* (4) are symptomatic of the general overuse of *important* in this learner corpus. *Porticle* writers use *important* much more frequently than the writers in the three control corpora (Table 3.16, Column 1).

Among the most frequent submodifier-modifier collocations found in *Locness* are *sexually active* (9), *sexually transmitted* (8), *terminally ill* (7), and *morally wrong* (6) each of which could be viewed as semi-technical terms in argumentative essays on moral questions. The total number of submodifier-modifier collocations found in each of the four corpora shows that *Porticle* writers use this kind of modification less frequently than the NS writers. The possibility of a positive correlation between writing expertise and the use of submodifier-modifier collocations suggests itself. (The z-test comparing the percentage occurrence in *Porticle* [≈ 0.001%] with *Locness* [≈ 0.002%] gives significance at $p \leq 0.01$).

Table 3.16 Some submodifier-modifier features of the four tagged corpora
(frequencies normalized to a basis per 100,000 words of text)

Column	**1**	**2**	**3**
Feature	**Occurrences of *important* in four corpora**	**Occurrences of *very***	**Frequency of submodifier- modifier collocations**
Porticle	194	275	209
Locness	84	157	304
Bawe	116	116	394
CofE	74	74	359

The contributors to *Porticle* differ in this respect from the German writers of English reported in Lorenz (1998). Whereas the Portuguese writers underuse the submodifier-modifier colligation, the German writers use it more frequently than the

English NS writers. The differences in use might be attributable to the prevalence of the structure in German as compared to English and Portuguese.

If the use of *very* in the four corpora is examined, it becomes clear that the *Porticle* writers use *very* as their all-purpose submodifier, or as Granger (1998b: 51) puts it as 'the all-round amplifier *par excellence'*. Recurrent combinations such as *very important* (37 occurrences), *very difficult* (16), *very good* (11), *very little* (10), *very well* (10), *very theoretical* (9), and *very hard* (9), show the degree to which the *Porticle* writers draw on this versatile submodifier when often, for precision (and elegance), they should be searching their mental lexicon for a nicer submodifier. I shall return to these differences in the use of submodifiers between the NNSs and NSs in Chapter 4 where possible explanations for the differences are sought. Although they use *very* much more frequently than the NS writers (Table 3.16, Column 2) the *Porticle* writers use many fewer pre-modified modifiers (Table 3.16 column 3).

Further information about the way in which premodifiers and modifiers behave can be obtained using Wmatrix (Rayson 2003) to tag the four corpora with POS and USAS semantic tags. As each word is separately tagged both semantically and syntactically, Wmatrix can provide a more sharply focused examination of the behaviour of premodifiers and modifiers. When *Porticle* and the other corpora are searched word by word using Wmatrix and USAS semantic tags the NNS writers are found to use many more maximizers and minimizers (Table 3.17). This corroborates the findings of Milton (1998:191) who writes that NNSs appear to overstate their case by using intensifying and categorical expressions.

Table 3.17 Expressions of degree in the four tagged corpora (using USAS semantic tags)

Corpus	A.13.2 maximizers	A.13.6 diminishers	A.13.7 minimizers
Porticle	*220*	*104*	*71*
Locness	*136*	*146*	*54*
Bawe	*85*	*114*	*36*
CofE	*134*	*103*	*64*

The semantic tags A.13.2, A.13.6, and A.13.7 in Table 3.19 refer to different kinds of adverbials of degree as detected automatically by Wmatrix.

Adverbials of Degree:

Maximizers

e.g : ALL, COMPLETELY, DOWNRIGHT, LARGELY, MAINLY, PREDOMINANTLY, SOUND, WHOLLY, OUTRIGHT ABOVE ALL, BY AND LARGE, FOR THE MOST PART, IN THE MAIN, MOST OF, ALL, ON THE WHOLE

Diminishers

e.g : BUT, FRACTIONALLY, LESS, MILDY, PARTIALLY, PARTLY, SLIGHTLY, SOMEWHAT, UNDER, A BIT OF A, A WEE BIT, A LITTLE BIT, TO SOME EXTENT, TO SOME LIMITED EXTENT, UP TO A POINT

Minimizers

e.g: BARELY, HARDLY, LEAST, LITTLE, SCARCELY

(Adapted from Rayson 2003)

3.11 Contending with N-grams

One form of analysis which can be carried out with relative ease is the measurement of the number of N-grams in each of the four corpora. To begin with, floors were established *a priori.* This setting in advance of minimum frequency levels was merely the first stage of a detailed consideration of the frequency lists of N-grams. After working with the corpora over an extended period, I found the following floors to be the most suitable:

bigrams: minimum of ten occurrences (i.e. raw score of >9)
trigrams: minimum of five occurrences (i.e. raw score > 4)
4-grams: minimum of five occurrences (i.e. raw score>4)
5-grams: minimum of five occurrences (i.e. raw score > 4)
6-grams: minimum of three occurrences (i.e. raw score > 2)
7-grams: minimum of three occurrences (i.e. raw score > 2)
8-grams: minimum of three occurrences (i.e. raw score > 2)

The relative likelihoods of sequences of 2, 3, and 4 words, decreases with each additional word. Meunier et al. (1998:79) observed that Zipf's law, according to which there is an inverse relationship between the length of a word and its frequency, applies equally to word sequences and so the floors were designed to gradually diminish relative to sequence length.

The floor chosen for the first search for bigrams was ten occurrences. *Locness* has fewer bigrams than *Porticle* and the two expert writers' corpora, *Bawe* and *CofE,* have fewer still. Nevertheless, without drastic culling there would be unmanageable quantities of

data. The Portuguese writers tend to greatly overuse certain bigrams and this will become even clearer when the filtered bigrams, i.e. the [2]prefabs, are discussed below in Section 3.13. Upon examining the longer N-grams (Table 3.18), it was found that the NNSs use significantly more types of N-grams and more tokens. An interesting question arises as to whether the apprentice writing contains more repeated sequences because learners know fewer words and, therefore, have to recycle them more often, or whether learners find it easier to write in chunked sequences and, consequently, use fewer word types resulting in lower ratings of lexical density. This question is addressed in Chapter 4.2.

For all values of N, the incidence of N-grams is highest in the NNS corpus and then there is a diminution according to writing expertise, with the professional writers in *CofE* exhibiting the lowest level of recurrent sequences (Table 3.18). When the variation among the N-gram tokens is taken into consideration, an increase is evident this time with the type-token ratio rising from the least to the most expert writers. Again, this holds for all values of N. Although the experts use fewer recurrent sequences of all lengths, they tend to vary more the sequences they use.

Porticle writers use more two-word, three-word, and four-word N-grams than the NS writers (Fig. 3.3). They combine their words in repeated patterns much more than the NSs. There is a clear pattern across the four corpora. The least expert writers contained in *Porticle,* use more N-grams than their NS contemporaries in *Locness* (for all values of N). The *Locness* writers in turn use more N-grams than the successful British writers in *Bawe.* The professional writers in *CofE* use the fewest number of N-grams.

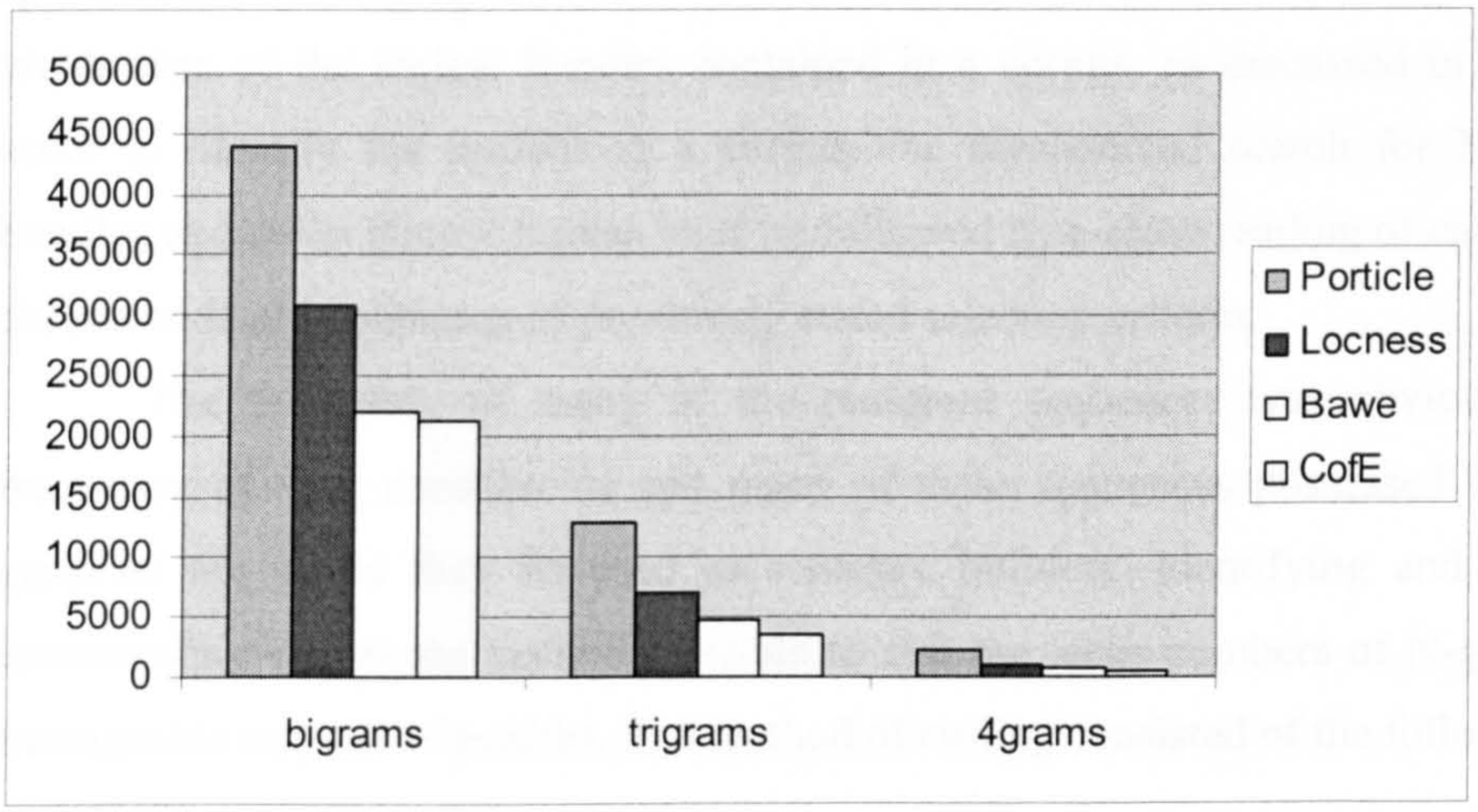
Table 3.18 N-grams and type-token ratios in the four corpora

4-gram T/T ratio	12.07%	12.56%	14.55%	15.44%
5-gram tokens	599	164	178	22
(floor 5)				
5-grams (types)	*Porticle*	*Locness*	*Bawe*	*CofE*
5-gram TTR	12.19%	12.80%	14.04%	18.18%
bigram tokens floor	44,037	30,915	21,928	21,175
6-gram tokens	351	189	57	24
(floor 3)				
bigram types	1528	1218	870	811
bigram T/T ratio	3.47%	3.94%	3.97%	3.83%
trigram tokens	12,267	7,126	4,807	3,603
(floor 5)				
7-gram tokens	170	67	15	6
(floor 3)				
trigrams (types)	1,326	814	610	460
trigram T/T ratio	10.39%	11.43%	12.53%	12.77%
4-gram tokens	2312	1086	304	133
(floor 5)				
8-gram tokens	93	29	3	-
(floor 3)				
4-grams (types)	279	126	117	69

8-gram types	25	8	1	-
T/T ratio	26.88%	27.59%	33.33%	-

A clear correlation between word concatenation and writing expertise is apparent when N-gram occurrences are considered. However, when the N-gram lists are reduced to prefab lists retaining only those sequences which display some degree of conventionalization, this correlation does not hold, as shall be seen in Section 3.13. Except for the longest prefabs with 5 and more words, the NSs record higher frequencies of prefab usage.

Figure 3.3 N-grams contained in the four corpora



The most frequent N-gram in the four corpora of argumentative writing is the sequence *of the.* This provides a very good example of a frequently occurring word combination which is generated by the grammar of the language each time it is used. *Of* is the most frequently used preposition in most corpora of English (Kennedy 1998:139) and is 'sensitive to what precedes more than to what follows' (Sinclair 1991:83). The other constituent word of this highly frequent sequence, the definite article, *the,* is usually the most frequent word in written corpora of English. It is unlikely that NSs or learners of English would commit the sequence *of the* to memory. The *of* might belong to a quantifier (e.g. *a lot of*) or refer back to a noun and *the* might be the determiner for a noun or noun phrase which follows.

The frequency of occurrence of N-grams in a corpus can be swiftly and accurately counted using text retrieval software. There are many thousands of recurrent N-grams in each of the four main corpora used in this study. In the next section approaches to reducing the large number of N-grams to more manageable lists of prefabs are described.

3.12 The procedures used for identifying prefabs

N-gram lists can be automatically extracted from corpora using text retrieval software suites, such as Wordsmith Tools (Scott 1999) and KFNgram (Fletcher 2004). Prior specification of the length and minimum frequency of occurrence (or floor) of the N-grams yields lists of the lexical bundles contained in a corpus, as discussed in Chapter 1.3. In order to identify the prefabs in a corpus, the mechanized search for N-grams, lexical bundles and collocational frames must be followed by a close reading of each sequence and the methodical application of previously stated selection criteria.

The frequency of many of the recurrent sequences was obviously due to the frequency of their constituents and many of these sequences possessed no memorizable qualities nor could they be used as sentence builders. Identifying and removing such uninteresting sequences made it possible to cull the large numbers of N-grams to a more manageable number of prefabs. The method of culling consisted of the following 7 steps:

Criteria used for culling N-grams to obtain prefabs

1. The lists of N-grams were read through and obvious non-idiomatic sequences were eliminated: e.g. most sequences beginning with *and+*, *a +*, *an+*, *he+*, *the+*, and *they+* (with some exceptions such as *a lot* and *the world*).
2. Sequences which are free combinations bound only by morphosyntactic rules (for example, the present simple conjugation of a verb) were only retained if they had some additional pragmatic or textual function. Thus *I think (that)* was retained whereas *they are* (218 occurrences in *Porticle*) was not.

3. Recurrent sequences ending in *a, an,* and *the* were eliminated. On this point I diverge from Biber et al. (1999) rejecting their lexical bundle approach to recurrent sequences (Section 1 3).
4. Substitution tests of constituent words, modelled on those of Hudson (1998:9), were used to ascertain whether there was a degree of fixity or formulaicity in a sequence (see Chapter 1.2).
5. Sequences were classified as prefabs if they appeared to have syntactic unity and to serve a clear grammatical function in their context. Under this criterion, compound verbs, compound nouns, compound adjectives, compound adverbs, compound prepositions, compound connectives and compound quantifiers were counted as prefabs (see Chapter 1.2).
6. If a sequence appears to be an attempt to reproduce a NS prefab or appears to be a calque by NNSs on an L1 prefab, it is counted as a prefab.
7. At each stage, lists of candidate prefabs were made and then each one was checked in context using the Wordsmith Tools concordancer. Following Kjellmer (1994) I was generously inclusive in my selection, so that I could investigate the discourse roles of these candidate prefabs.

All the original lists of N-grams and the results of each successive cull were stored as Excel worksheets. This layered storage was to guarantee that the steps taken in filtering the N-grams could be retraced if necessary. When word sequences were re-examined in their co-text, it was sometimes necessary to re-evaluate the filtering criteria. The first 50 most frequent sequences from each corpus are listed in Table 3.22. Those bigrams which were rejected as prefabs through the application of the seven steps above are marked with a strikethrough. The sequences which qualify as prefabs are not struck through and are marked in blue.

In this process of selecting prefabs from among the N-grams, the corpus linguist has to distinguish patterns in the excerpted sequences and then return to the original linguistic context to see how and why these patterns are used. There is a need, therefore, to shift perspective continually from the lexical level to the discourse level, to examine prefabs as types and then examine each token in its co-text. The investigative focus must range, therefore, from the level of the lexical item and its surrounding phraseology to ever wider

levels of text in which the prefab is employed up to and including the complete essay text from which the word sequence was extracted by the concordancer.

In a preliminary intelligence-gathering approach, in preparation for the analysis of large quantities of N-grams, I used Wordsmith Tools (Scott 1999) to make word frequency lists for each corpus on the basis of clusters of two, three and more words. A comparison could then be made to find clusters which are overused or underused by one group of writers. As was the case with the word level analysis described in Section 3.6, the Compare Two Wordlists function of Wordsmith Tools provides the researcher with a quick way into the data. Wordlists based on two-word clusters (i.e. bigrams) were made for *Porticle* and *Locness* and then compared. Those bigrams which were significantly overused or underused by either group of writers were recorded and the measure of the statistical significance of these events was noted.

The device of comparing two wordlists only highlights the behaviour of those bigrams which differ significantly between the two corpora. Concentrating too much on this comparative data could divert attention from important word sequences common to both, or indeed all four, of the corpora under consideration. Nevertheless, because a comparison between *Porticle* and *Locness* was the main focus, I used this technique extensively in the exploratory stage of the investigation.

Table 3.19 Keyword analysis of bigrams in *Porticle* and *Locness*

	WORD	FREQ	2Porticle.list %	FREQ	2Locness.list %	KEYNESS
1	*WE CAN*	186	0.15	14	0.01	167.7
2	*I THINK*	189	0.15	25	0.02	134.7
3	*IN MY*	161	0.13	18	0.01	124.6
4	*THINK THAT*	188	0.15	39	0.03	99.4
5	*KIND OF*	113	0.09	12		89.5
6	*INDUSTRIALISAT ION*	59	0.05	0		79
7	*WE SHOULD*	74	0.06	3		78
8	*TO DREAM*	55	0.04	0		73.6
9	*REAL WORLD*	56	0.04	4		51.3
10	*I DON'T*	61	0.05	6		49.9
11	*MODERN WORLD*	51	0.04	3		49.4
12	*MASS MEDIA*	33	0.03	0		44.2
13	*EUROPEAN UNION*	32	0.03	0		42.8
14	*WE DON'T*	53	0.04	6		40.8
15	*WE CAN'T*	29	0.02	0		38.8
16	*I CAN*	45	0.04	4		38.3
17	*I BELIEVE*	73	0.06	6	0.01	36.8
18	*MY POINT*	27	0.02	0		36.1

19	*ALL THE*	122	0.1	43	0.04	35.7
20	*WE ALL*	51	0.04	7		35.6
21	*FOR INSTANCE*	50	0.04	7		34.5
22	*WE WANT*	25	0.02	0		33.5
23	*OF COURSE*	73	0.06	18	0.01	33
24	*UNIVERSITY DEGREES*	24	0.02	0		33
25	*ALL MEN*	42	0.03	5		31.5
26	*THE INTERNET*	23	0.02	0		30.8
27	*HAVE MONEY*	23	0.02	0		30.8
28	*THE ONES*	53	0.04	10		30.2
29	*THE SOCIETY*	46	0.04	8		27.8
30	*FOR EXAMPLE*	97	0.08	35	0.03	27.4
31	*THE SAME*	180	0.14	89	0.07	27.2
32	*ALL THIS*	35	0.03	4		26.8
33	*I WOULD*	60	0.05	15	0.01	26.8
34	*THE MASSES*	20	0.02	0		26.8
35	*WE SEE*	39	0.03	6		25.5
36	*OUR WORLD*	33	0.03	4		24.6

With a raw frequency of 189 and a log likelihood (see Chapter 2.5) of 134.7, *I think* is the most salient marker of epistemic stance in *Porticle* (Table 3.19). The sequence *think that* shows 188 occurrences, which suggests the existence of a [3]prefab *I think that.* A concordance showed that this three-word sequence occurred 117 times in *Porticle.* One decision would be to treat all of these occurrences as [2]prefabs and attribute *that* to the complement clause. Such arbitrary decisions may be taken as long as they are recorded. Another way of dealing with this prefab is to view the occurrence of *that* as in some way an affirmation of the speaker/writer's wish to express their attitude and that the non-use of *that* rather than the use of *that* attenuates the force of the assertion. I opted for the first solution and judged *I think* as the core [2]prefab functioning as an epistemic stance marker with *that* as an optional extension when an embedded clause is the continuation. Along with *I think,* there was significant (log likelihood = 36.8) overuse of *I believe* which occurred 73 times and has a similar function to *I think* in the realization of epistemic stance, while at the same time revealing the source of the information and expressing reservations entertained by the writer as to the veracity of the proposition which follows.

The frequency with which the learners use phrases for exemplification is also curious: *for instance* (50 occurrences) and *for example* (97) with the former having a higher log likelihood (henceforth LL) of 34.5. Such apparent under-utilization of exemplifiers by the American undergraduates led me to wonder if they used *such as* more frequently in compensation. I found that they did in fact do so: 82 vs. 64 times; but then, upon totalling

occurrences of *for instance, for example* and *such as* in the two corpora, I found that the NNSs gave examples 167 vs. 103 times per 100,000 words (with raw scores of 211 vs. 124 occurrences). As predicted by the findings of the *ICLE* scholars, Altenberg and Tapper (1998), the adverbial of epistemic stance *of course* showed a high level of overuse (LL=33) by the learners.

Around half of the bigrams listed in Table 3.19 do not qualify for prefab status. Nine of these bigrams are conjugations of the auxiliaries *can, should* and *do*, while two are forms of full verbs *see* and *want*. The collocation *have money* produces a log likelihood of 30.8 as it does not appear at all in *Locness*. (This collocation appears 4 times per million words in the spoken sub-corpus of *BNC-baby* but not at all in the academic sub-corpus).

Table 3.20 Bigrams used more frequently in Locness than in *Porticle*

	WORD	FREQ	2Porticle.list %	FREQ	2Locness.list %	KEYNESS
1	THE FLAG	0		17	0.01	24.4
2	MODEL APPROACH	0		17	0.01	24.4
3	THE AVERAGE	0		17	0.01	24.4
4	NATIONAL CHAMPIONSHIP	0		17	0.01	24.4
5	SURROGATE MOTHERHOOD	0		17	0.01	24.4
6	ADOPTIVE PARENTS	0		20	0.02	28.7
7	THE MILITARY	0		20	0.02	28.7
8	ADOLESCENT SUICIDE	0		20	0.02	28.7
9	CAPITAL PUNISHMENT	5		37	0.03	29.1
10	THIS ARGUMENT	0		23	0.02	33
11	THEIR ARGUMENT	0		23	0.02	33
12	DRINKING AGE	0		23	0.02	33
13	BIOLOGICAL PARENTS	0		24	0.02	34.4
14	CORPORAL PUNISHMENT	0		24	0.02	34.4
15	AFFIRMATIVE ACTION	0		24	0.02	34.4
16	SOUTH CAROLINA	0		24	0.02	34.4
17	STAY HOME	0		25	0.02	35.9
18	IN PUBLIC	5		43	0.04	36.3
19	FLORIDA STATE	0		26	0.02	37.3
20	PLAYOFF SYSTEM	0		26	0.02	37.3
21	WATER POLLUTION	0		27	0.02	38.8
22	GENETIC RESEARCH	0		28	0.02	40.2
23	WELFARE RECIPIENTS	0		28	0.02	40.2
24	COLLEGE FOOTBALL	0		29	0.02	41.6
25	WILD CARD	0		31	0.03	44.5
26	DEATH PENALTY	15		74	0.01	45.6
27	NEW AGE	0		37	0.03	53.1
28	NUCLEAR POWER	0		37	0.03	53.1

29	*THE ARGUMENT*	0		40	0.03	57.4
30	*THE ISSUE*	0		40	0.03	57.4
31	*THE FAMILY*	5		60	0.05	57.6
32	*PUBLIC SCHOOLS*	0		55	0.05	78.9

An interesting feature of the bigram overuse in the *Locness* corpus (Table 3.20) was that almost all of them were compound nouns or nouns modified by adjectives, e.g. *gun control, surrogate motherhood;* or generic use of the definite article before nouns or adjectives *the homeless, supreme court, adoptive parents,* or *the military.* From a total of 25 occurrences of *stay home,* 22 formed part of the trigram *stay home wife* so the bigram, *stay home*, is arguably functioning as a compound adjective. A closer examination of the data in Table 3.19 and 3.20 confirms the suspicion that fewer nominalizations were being used in the *Porticle* corpus. The already noted failure of the NNSs to build noun phrases would appear to be a significant factor contributing to the lexical sparsity and spoken-like quality of their written prose.

The relative overuse by *Locness* writers of explicit referents to the process of argumentation suggests they are more consciously engaged in the rhetoric of argumentation. Their much greater use of *the argument* (LL=57.4) and *the issue* (LL=57.4) displays a grasp of the process of good argument which the writers in the learner corpus still have to achieve. Incidentally, *issue* occurs quite frequently in *Porticle* (37 times in the singular, 24 times in the plural) but almost always with an adjective or demonstrative preceding it. For example, *it's a very complicated issue, a contemporary issue, a very controversial issue* and *this issue* (14 occurrences of this last example) can be found in the concordance of *issue* in the *Porticle* corpus. The NS writers use *issue* preceded by the definite article as a metalinguistic term, to refer anaphorically to the argumentation used by themselves or others.

The overused bigrams in *Locness* appear to be more related to specific semantic content than those in *Porticle* which are often grammatical reducibles or vague general terms (e.g. *the masses, our world, the society, all men*). From the NNSs we see the overuse of auxiliaries, e.g. *I don't, we should, we can* and with high log likelihood values (49.9, 78, and 167.7 respectively). This confirms the findings of Granger and Rayson (1998) who report similar overuse of auxiliaries among learners from various mother tongue backgrounds. These writers hypothesize that this overuse of auxiliaries is related to the speech-like nature of much NNS writing. *Porticle* writers use a high degree of

personalisation in their writing, as noted in their use of the first person pronoun and in phrases like *in my* and *my point,* which usually belong to *in my opinion* and *my point of view* respectively. An alternative explanation would be that the learners have not yet mastered the devices required to write in a formal register. Other auxiliaries which emerge as being significantly more frequent in the learner corpus as compared to *Locness*, the NS apprentice corpus, are *we should, we don't, we can't, I can,* and *I would.*

Table 3.21 Occurrences of modal verbs in the four corpora

(frequencies normalized to a basis per 100,000 words of text)

CORPUS	*Porticle*	*Locness*	*Bawe*	*CofE*
may	79	170	106	149
might	24	40	24	89
can	811	433	387	292
could	193	222	162	133
will	423	405	175	429
would	328	506	273	286
shall	9	4	6	3
should	387	337	143	200
must	160	137	59	105
have to	206	81	28	45

When the frequency of use of the modal verbs is examined, *can, should* and *must* are found to be overused, while *may* and *might* are underused in *Porticle* as compared with the three NS corpora (Table 3.21). These last forms are often used by NSs to express attenuation in the configuration of epistemic stance. In Chapter 4.4, further ramifications of this overuse of *should* and *must* and underuse of *may* and *might* will become apparent. An examination of the modal *must* and the semi-modal *have to* shows a tendency by the Portuguese writers to be more absolute in their affirmations and recommendations (Table 3.21).

3.13 From bigram to ^{2}prefab

^{2}Prefabs are the two-word clusters which were left after the first cull or, put less negatively, they are the first fruits of the prefab harvest. Some of the tokens of the selected prefab types proved chimerical. In the case of each candidate prefab it was necessary to return to the corpus and observe each word sequence in its syntactic, pragmatic and textual context. As a result of this re-contextualization of the candidate prefabs, many of the prefab tokens proved to have a completely non-formulaic role in a larger phrase. For example, when I examined more closely the occurrences of *in that* in the four corpora, I found various examples of free combinations (e.g. **in that** period; **in that** same article). Many others again formed part of trigrams or even longer sequences and, therefore, had to be relegated to the collection of longer prefabs. Most disconcerting of all was the discovery that none of the 31 occurrences of *in that* in the *Porticle* corpus qualified as a prefab (although three of these bigrams formed part of a sentence-initial trigram 'In that sense'.

To understand the journey from N-gram to prefab, or from Figure 3.1 to Table 3.28, the first 50 bigrams in each corpus are examined closely to illustrate the kinds of decisions that are involved in sifting the N-grams in search of interesting prefabs. This sifting process is summarized graphically in Table 3.22 where the ^{2}prefabs are marked in blue and the rejected bigrams are struck through. In the first column showing the most frequent bigrams in *Porticle* in descending order of frequency, four grammatical sequences appear before the first prefab, the semi-modal *have to,* is selected. These four sequences are *it is, we have, they are*, and *we are.* Warren (1999) and Wiktorsson (2003) counted such sequences, which have corresponding contractions, as prefabs. They called these grammatical sequences 'reducibles' and I followed them in this respect in the small-scale experiment described in Section 3.6 above. For the purposes of investigating the four corpora of argumentative writing, I decided to count only those sequences resulting from verb conjugation which carry additional interpersonal or pragmatic meaning. Thus, the semi-modal *have to* attains prefab status within the first 50 bigrams of *Porticle* and *Locness* thanks to its deontic and epistemic meanings. Similarly, *I think,* which occurs in the first 50 *Porticle* bigrams, is given prefab status because the Portuguese writers use this word combination mostly for

expressing epistemic stance. Eleven of the first 50 bigrams in *Porticle* and 14 in *CofE* are sequences ending in *the* and were therefore eliminated in the search for prefabs.

Table 3.22 50 most frequent bigrams from the four corpora in descending order of frequency

Selected prefabs marked in blue
Rejected bigrams marked with ~~strikethrough~~

Porticle	*Locness*	*Bawe*	*CofE*
~~of the~~	~~of the~~	~~of the~~	~~of the~~
~~in the~~	~~in the~~	~~in the~~	~~in the~~
~~it is~~	~~to be~~	~~to the~~	~~to the~~
~~to the~~	~~to the~~	~~and the~~	~~it is~~
~~to be~~	~~it is~~	~~it is~~	~~and the~~
~~in a~~	~~that the~~	~~to be~~	~~on the~~
~~is the~~	~~for the~~	~~that the~~	~~to be~~
~~is a~~	~~on the~~	~~on the~~	~~that the~~
~~that we~~	~~and the~~	~~can be~~	~~for the~~
~~we have~~	~~is a~~	~~for the~~	~~by the~~
~~they are~~	~~is the~~	~~as a~~	~~is a~~
~~we are~~	~~is not~~	~~of a~~	~~of a~~
~~and the~~	~~as a~~	~~as the~~	~~with the~~
have to	~~they are~~	~~is a~~	~~as a~~
the world	~~with the~~	~~with the~~	~~at the~~
~~that the~~	~~this is~~	~~by the~~	~~in a~~
~~think that~~	~~that they~~	~~is the~~	~~is the~~
~~we can~~	~~in a~~	~~in a~~	~~from the~~
i think	~~of a~~	such as	~~there is~~
~~this is~~	~~should be~~	~~from the~~	~~is not~~
~~the same~~	~~by the~~	~~there is~~	~~has been~~
~~on the~~	~~there are~~	~~this is~~	the world
~~there are~~	~~would be~~	~~is not~~	~~will be~~
~~with the~~	~~there is~~	~~at the~~	~~as the~~
~~for the~~	~~can be~~	~~the same~~	~~have been~~
~~in my~~	~~does not~~	~~which is~~	~~the united~~
~~there is~~	~~from the~~	~~the first~~	~~this is~~
~~that is~~	~~at the~~	~~that it~~	united states
~~to do~~	~~do not~~	for example	~~the same~~
~~have a~~	~~is that~~	~~have been~~	~~they are~~
~~that they~~	~~of their~~	~~has been~~	~~is that~~
~~is not~~	~~because of~~	~~it was~~	~~about the~~
~~in our~~	~~the same~~	~~there are~~	~~to a~~
~~should be~~	~~have been~~	~~that is~~	~~more than~~
~~is that~~	~~not be~~	~~in this~~	~~that it~~
~~have the~~	~~that is~~	~~should be~~	~~should be~~
~~say that~~	~~the united~~	~~found that~~	~~it was~~
~~all the~~	united states	~~the most~~	~~the us~~
~~to have~~	such as	~~through the~~	~~would be~~
~~as a~~	have to	~~about the~~	~~there are~~
~~do not~~	~~out of~~	~~does not~~	~~for a~~
~~if we~~	~~has been~~	due to	~~it has~~
~~are not~~	~~of these~~	~~in which~~	~~we are~~
~~can be~~	~~the fact~~	~~to a~~	~~be a~~
~~they have~~	~~if the~~	according to	~~but the~~
~~of a~~	~~are not~~	~~of their~~	~~in which~~
kind of	death penalty	~~between the~~	~~the most~~
~~want to~~	~~to have~~	~~number of~~	~~these who~~
~~be a~~	~~to make~~	~~of his~~	~~can be~~
~~in order~~	~~have a~~	~~the other~~	the british

The bigram, *united states,* as found among the 50 most frequent bigrams in the *Locness* and *CofE* corpus, was originally not considered as a prefab because it was thought to be part of the trigram, *the united states.* On closer inspection of the concordance of the instances of *united states* in the *CofE*, however, there were found to be 12 occurrences where the bigram was functioning as an adjective premodifying a noun out of a total of 87 occurrences of *united states.* When the bigram *the british* was inspected in a concordance from the *CofE*, eight out of the 41 occurrences were being used generically to refer to the British people thus qualifying the expression as a prefab. Incidentally there were only five occurrences of the trigram, *the British people* in the same corpus.

The first 30 candidate two-word prefabs from each of the corpora are arranged in descending order of frequency (Table 3.23). The final prefab in each list, when 30 ^{2}prefabs were yielded, was bigram number 1153, 1128, 899 and 858 from *Porticle*, *Locness*, *Bawe* and *CofE* respectively. Even a cursory look at these figures shows the large number of rejections made from among the bigram candidates and how ruthless the pruning was. Only 30 ^{2}prefabs survive in roughly the first thousand most frequent bigrams in each corpus. In Chapter 4 the contribution some of these ^{2}prefabs make to the configuration of writer's epistemic or attitudinal stance is examined. Prefabs which realize quantification, approximation or precization (see Chapter 1.15), and those which perform metapragmatic or metalinguistic functions are also discussed in Chapter 4.

Table 3.23 Two-word prefabs in the four corpora in descending order of frequency

	Porticle	*Locness*	*Bawe*	*CofE*
(1)	*the world*	*because of*	*a little*	*a few*
(2)	*I think*	*such as*	*a lot*	*a handful*
(3)	*that is*	*have to*	*according to*	*a little*
(4)	*kind of*	*death penalty*	*as cited*	*a lot*
(5)	*in order*	*a chance to*	*appears to*	*above all*
(6)	*for example*	*the family*	*appeared to*	*according to*
(7)	*a lot*	*public schools*	*appear to*	*against terrorism*
(8)	*most of*	*according to*	*as well*	*after all*
(9)	*of course*	*the world*	*at first*	*agree strongly*
(10)	*I believe*	*in public*	*carry out*	*almost certainly*
(11)	*our society*	*the issue*	*carried out*	*and all*
(12)	*such as*	*the argument*	*automatic processing*	*american policy*
(13)	*many people*	*as well*	*at least*	*and more*
(14)	*real world*	*a lot*	*cited in*	*and therefore*
(15)	*no longer*	*nuclear power*	*circadian rhythm*	*and then*

(16)	*in fact*	*capital punishment*	*cultural differences*	*as for*
(17)	*due to*	*for example*	*could be*	*appear to*
(18)	*the fact*	*the media*	*connectionist system*	*apart from*
(19	*modern world*	*high school*	*conditional sentences*	*and yet*
(20)l	*for instance*	*due to*	*communicative competence*	*at least*
(21)	*at least*	*the past*	*cognitive processes*	*at last*
(22)	*I know*	*a part of*	*cognitive development*	*at home*
(23)	*all men*	*wild card*	*due to*	*at all*
(24)	*some people*	*in fact*	*in common*	*as well*
(25)	*instead of*	*a little*	*in all*	*as one*
(26)	*science technology*	*college football*	*in addition*	*as if*
(27)	*even if*	*our society*	*ill health*	*a good society*
(28)l	*a result of*	*look at*	*i wish*	*a free press*
(29)	*mass media*	*water pollution*	*i wanted*	*a few years*
(30)	*a solution to*	*the welfare*	*i thought*	*a debate about*

When those prefabs which configure epistemic stance are examined, much more frequent use of certain prefabs is found in *Porticle* in comparison with the NS corpora. *I believe* as a comment clause is used 73 times to frame the succeeding proposition and to realize epistemic stance in a manner which some composition teachers might consider to be excessively personal. *I guess* had 10 occurrences in *Porticle* and *I mean* 13 occurrences. Many more elaborate uses of *think* to realize writer's stance occurred in *Porticle,* with 46 examples appearing as trigrams with various interpositions between *I* and *think*, and usually an adverb (*truly* (1), *personally* (1), *often* (1), *sometimes* (1), *just* (4), *really* (5), *also* (9),) preceding the node word. There are 20 occurrences of *don't* negativing *think* but there is little doubt that I *don't think* is performing an epistemic stance-taking role in each case. Further evidence of this strong emphasis on expressing epistemic stance can be found elsewhere in *Porticle,* for example in the overuse of the bigrams *we all* and *we know* (18 occurrences each). The question of epistemic and attitudinal stance in the corpora is discussed further in Chapter 4.3.

Of course does not appear among the 30 most frequent prefabs in the control corpora. The prefab expressing certainty occurs 8, 7, and 7 times in *Locness*, *Bawe* and *CofE* respectively. Curiously, the reverse sequence *course of* occurs more frequently in *Bawe* (12 occurrences) and *CofE* (13 occurrences).The overuse of *of course* to mark

epistemic stance by the *Porticle* writers is mirrored in all of the NNS sub-corpora contained in *ICLE* (Granger and Rayson 1998:128). *Of course* occurs only twice in the *BNC-baby,* in the conversation corpus and in the imaginative writing. It is clearly a speech-like adverbial. The prepositional phrase *in fact* is another prefab used frequently to realize epistemic stance. As an illustration of the care which must always be taken with what seem like the most watertight of candidatures, here is a use of *in fact* from the *CofE* corpus which does not express epistemic stance and is not a prefab unless, arguably, it belongs to a much larger six-word prefab:

> And the Chinese did not simply convene the meeting, but participated actively, making the talks multilateral **in fact** as well as form. Moreover, the decision by China to play host to the talks gave it a major stake in a resolution of the issue, and the North Koreans' behaviour…

This constant review of concordance lines for often fairly trite word combinations has brought to my attention some semantic prosodies of some of the prefabs. *In fact* seems to introduce unpalatable truths or surprising or paradoxical facts or admissions. The following excerpts are the first five occurrences of IN FACT numbered in the order of their appearance in *CofE*, the corpus of broadsheet editorials.

> 1 The bad news, however, is that these successes have not won new friends for the United States outside Afghanistan. **In fact,** the effectiveness of the American campaign may have made some parts of the world hate America more than they did before. Critics of

> 5 government has rather notably failed to admit that "some of the most important roots of terrorism in Kashmir" are, **in fact,** of its own making, she said, mentioning the suspension of democratic rights, jailing of intellectual dissidents and the falsifying of election result

> 8 debate about covering up the graffiti was ostensibly a debate about how the rebuilt Reichstag should be designed. **In fact,** it was about something far more significant: inconvenient history. Those who wished to cover the graffiti were espousing

> 11 largely implicit and covert. As some languages are more "international" than others the equality of the 11 languages has **in fact** always been a myth. The existing rights to translation are essential,

because documents emanating from Brussels have the force of law in member states

12 Are markets local or global? Convenience to consumers is also illusory. The time taken to go shopping has **in fact** risen over the last 30 years. Both consumers and the produce clock up more energy-wasteful food miles and the companies take people's uncosted time.

The *Macmillan English Dictionary for Advanced Learners* (2002:495) confirmed my intuitions in its entry for *in fact*:

> in (actual) fact
>
> 1 used for saying what is really true, when this is surprising or different from what people think:
>
> In actual fact, she was quite right.
>
> He was paid money for a job that did not in fact exist.
>
> 2 used when you are adding something to what you have just said, especially something surprising:
>
> I haven't seen him for years. In fact, I can't even remember what he looks like.
>
> She's a friend of mine, a very close friend in fact.

A useful by-product of the investigation of prefabs is more nuanced profiles of their semantic preferences and phraseologies, which in many cases will complement or possibly contradict the information already available in learner and collocation dictionaries.

By creating concordances of this phrase as it occurred in *Porticle* and *Locness*, comparisons could be made of the different utilizations in the two corpora and the extent to which *in fact* realized epistemic stance. As with all adverbials, the position within the sentence is of considerable interest to the student of stance (Biber et al. 1999:770-774). The examples from the *Macmillan Dictionary for Advanced Learners* cited above show sentence-initial, medial and final position. In the two undergraduate corpora, we find a preference for sentence-initial position: in *Porticle*, 29 of the 55 occurrences (53%) of *in fact* are sentence-initial while in *Locness,* 18 of the 31 occurrences (58%) are sentence-initial. In the typical dictionary entry, information about the preferred sentence position of an expression is not directly available, although it may sometimes be inferred from the illustrative examples cited. In contrast, this aspect of the colligability of an expression is immediately discernible in a concordance.

Among the ten most frequently occurring prefabs found in *Porticle,* three prefabs are strongly linked to the configuration of epistemic stance: *I think, I believe*, and *of course.* In comparison, it is necessary to look much further down the frequency list in *Locness* to find a prefab expressing stance (*in fact* with 26 occurrences per 100,000). *Bawe* has *I believe, I found, I think*, and *I thought* as number 47, 48, 49, and 50 in descending order of frequency, with 12 occurrences of each. There are also 21 occurrences each of *appear to*, *appeared to*, *appears to* in *Bawe.* If the three different forms of *appear to* are lemmatised, the total of 63 hits per 100,000 words shows a propensity on the part of the high-achieving British students to hedge their statements. In Chapter 4, the NNSs' inability to hedge is shown to underly a major difference between them and their NS counterparts.

Bawe also contains many prefabs which are exponents of the textual metafunction, such as *as for, cited in, according to, by saying.* It would appear that one reason for the academic success of the *Bawe* writers is their ability to structure and 'signpost' their texts. *CofE* writers use a number of quantifiers which feature in the top four places in the list of prefabs in descending order of frequency: *a few, a handful, a little, a lot.* These last two quantifiers are also the two most frequent prefabs in *Bawe.* The effect of using quantifiers judiciously is to give to the writing the appropriate degree of precision or vagueness.

When the normalized lists of prefabs were examined, several features became apparent. Expressions with *people* showed a very high frequency in the phrasicon of *Porticle*: *many people* registered 46 occurrences and *some people* 32 occurrences per 100,000 words. This almost pronominal use of *people* can be compared to the use of the expression *all men* 33 times per 100,000 words in *Porticle.* The apparent sexism of such frequent use of *men* might surprise the reader given that the majority of *Porticle* writers are female. One writer actually wrote *all men and women.* There does not appear to exist the same pressure within present-day European Portuguese to use gender-neutral nominals. The uses of *homen* (man) and *ele* (he) in Portuguese as generic referents to both sexes do not connote sexism to the extent that their translations do in academic English. Further down the list is the non-sexist, if not quite idiomatic, trigram *all human beings* with a score of 21 per 100,000 words. The accumulation of expressions with *all* reveals pronounced overuse of the universal quantifier.

In *Porticle,* the sequences *many people* and *some people* are often used to refer to an argument or position which the writer intends to refute despite its popularity (Concordance 3.1).

Concordance 3.1

Random selection of concordance lines from *Porticle* using *many people* as searchword.	
1.	or imagine? Have our children time to play and dream? **Many people** consider that our world is too agitated,
2.	selves and to develop their minds. I don't deny that for **many people** tv represents the end of loneliness and
3.	and our lives will revolve around them. For instance, for **many people** watching TV has become a sort of cult,
4.	have an easy life with the help of machines. However, **many people** don't know how to live in better conditions
5.	And things that should be learnt in University. I know **many people** that never attended a University, that don't
6.	bad aspects that will have concequences in our lives. **Many people** like these idea of a modern world, dominate
7.	e is looking for money... just money and more money. **Many people** think that money is the most important thi
8.	mber that they have their childhood to live. There are so **many people** who don't have money and need our hel
9.	speakers of English outnumber native ones. The fact that **so many** people are using English all over the world
10.	programmes that, in fact, affect our subcounscious. **Many people** don't take the time to read a book, a mag

The NS writers tend to use the passive voice or more abstract nominalizations to convey similar ideas. For example, a search on *widely* in the *Bawe* corpus produced 13 concordance lines. Concordance 3.2 contains the first ten lines. The NS use of the passive together with modifier *widely* enables them to express demographic and quantificational estimates similar to those made by the *Porticle* writers but in a more impersonal way, avoiding the overuse of *people.*

Concordance 3.2 Concordance lines from *Bawe* using *widely* as searchword.

Random selection of concordance lines from *Bawe* using *widely* as searchword.	
1	It is probably the most influential and **widely accepted** theory of truth' and has
2	and Britain's attempts to curb this were **widely condemned.** The effects of the
3	sing provision on community health once **widely recognised** by politicians and,
4	of education, schools and learning were **widely resented** and indicated teachers
5	groups become stereotypes when they are **widely shared.** Allport (1954) defined
6	on the basis of the TOEIC test which is **widely taken** in Korea. The discussion
7	resting things in those books which are **widely unknown** and not thought about m
8	so that informal communication is more **widely accepted** than formal methods
9	Behavioristic theories were once **widely accepted** in language teaching
10	The hegemony of the landowners **was widely viewed** by the commoners as self

3.14 Longer prefabs

The main object of the culling process described in the previous sections was to separate recurrent word sequences which result from conventionalization and are memorized by NSs (e.g. *of course* and *such as*) from sequences which recur as a result of the syntactic rules of the English language. The culling of bigrams is a laborious process firstly because of the large quantity of such sequences in each corpus and secondly because of the high proportion of non-prefabs which must be eliminated. In section 3.13 it was found necessary to sift through about 1,000 bigram sequences to obtain about 30 prefabs. When longer sequences are inspected, a smaller overall proportion needs to be eliminated. For example, when the most frequent trigrams in each corpus are inspected using the prefab selection procedures, there are many fewer rejections (Table 3.24) with roughly half of the trigrams being selected as prefabs. 22 trigrams were discounted because they end in *a* or *the*.

Table 3.24 First twenty most frequent trigrams from each corpus with selected prefabs marked in blue

	Porticle	*Locness*	*Bawe*	*CofE*
1	i think that	a good example	in other words	a good society
2	a world of	a great deal	in order to	~~one of the~~
3	a way to	a little more	in most cases	a free press
4	a way that	~~one of the~~	~~one of the~~	a few years
5	all know that	a means of	in line with	~~there is a~~
6	~~one of the~~	~~because of the~~	i believe that	a debate about
7	all human beings	nuclear power	human resource management	~~it is not~~
8	all the time	~~in the united~~	great deal of	a lingua franca
9	all of us	~~should not be~~	goes on to	~~some of the~~
10	~~there is no~~	~~it is not~~	~~due to the~~	~~that it is~~
11	~~the other hand~~	~~in public schools~~	~~part of the~~	~~part of the~~
12	all of them	~~that it is~~	~~that it is~~	~~it is a~~
13	~~it is a~~	~~be able to~~	first of all	~~to be a~~
14	~~most of the~~	~~this is a~~	~~there is a~~	~~most of the~~
15	as I said	capital punishment	~~it is not~~	a group of
16	~~on the other~~	~~that they are~~	~~some of the~~	~~there is no~~
17	~~to have a~~	a member of	~~the nature of~~	a member of
18	~~that we are~~	~~there is no~~	as well as	~~this is a~~
19	~~live in a~~	~~prayer in public~~	as opposed to	a means of
20	as far as	~~there is a~~	as much as	a lot of

When lists are made of the 50 most frequent prefabs in each of the four corpora, regardless of the number of words which constitute them, a number of three-word clusters are found (Table 3.25). It is remarkable that these trigrams occur with sufficient frequency to be counted among the most frequent prefabs, given that they are much less likely to occur, as discussed in Chapter 2.5. The expressions *all the time* found in *Porticle* and *a great deal* found in *Locness* are notable examples.

An examination of the 50 most frequent prefabs in each corpus, shows that the *Porticle* writers are using certain prefabs with a very high frequency to express epistemic stance, e.g. *I think, of course, I believe,* and *I know.* These same writers are also overusing other expressions such as *many people, all men, all know that, all human beings* to state accepted positions or commonly held assumptions. The first of these expressions, *many people,* is a fairly colloquial expression while the others are non-idiomatic translations from the Portuguese. In contrast with this heavy use of epistemic markers in the NNS corpus, there are surprisingly few such expressions among the 50 most frequent prefabs in the control corpora.

Table 3.25 The fifty most frequent prefabs in the four corpora
(in descending order of frequency with frequencies normalized to a basis per 100,000 words of text))

***Porticle* prefabs**		***Locness* prefabs**		***Bawe* prefabs**		***CofE* prefabs**	
the world	156	*because of*	76	*a little*	38	*a few*	52
I think	146	*such as*	68	*a lot*	36	*a handful*	48
that is	122	*have to*	67	*according to*	27	*a little*	45
kind of	89	*death penalty*	61	*as cited*	21	*a lot*	44
in order	87	*a chance to*	56	*appears to*	21	*above all*	29
for example	77	*a good example*	55	*appeared to*	21	*according to*	28
a lot	70	*the family*	50	*appear to*	21	*against terrorism*	27
most of	62	*public schools*	46	*as well*	20	*after all*	27
of course	58	*according to*	43	*at first*	19	*agree strongly*	26
I believe	58	*the world*	40	*carry out*	17	*almost certainly*	24
our society	54	*a lack of*	38	*carried out*	17	*and all*	23
such as	51	*a great deal*	38	*automatic processing*	17	*american policy*	23
many people	46	*a little more*	37	*at least*	17	*and more*	20
real world	44	*a large number*	37	*cited in*	16	*and therefore*	19
no longer	43	*in public*	36	*circadian rhythm*	16	*and then*	19
in fact	43	*a long time*	35	*cultural differences*	15	*as for*	18
due to	43	*the issue*	33	*could be*	15	*appear to*	18
the fact	40	*the argument*	33	*connectionist systems*	15	*apart from*	18
modern world	40	*as well*	33	*conditional sentences*	15	*and yet*	18
for instance	40	*a means of*	32	*communicative*	15	*at least*	17

				competence			
at least	39	*a lot*	32	*cognitive processes*	15	*at last*	17
I know	34	*nuclear power*	31	*cognitive development*	15	*at home*	17
all men	33	*capital punishment*	31	*due to*	14	*at all*	17
some people	32	*a member of*	30	*in common*	12	*as well*	17
instead of	32	*for example*	29	*in all*	12	*as one*	17
science technology	31	*a number of*	29	*in addition*	12	*as if*	17
even if	31	*the media*	28	*ill health*	12	*a good society*	17
a result of	27	*high school*	28	*i wish*	12	*a free press*	17
mass media	26	*due to*	28	*i wanted*	12	*a few years*	17
a solution to	26	*the past*	27	*i thought*	12	*a debate about*	17
this way	25	*a part of*	27	*i think*	12	*aware of*	16
as well	25	*wild card*	26	*i found*	12	*at that*	16
a sort of	25	*in fact*	26	*housing quality*	12	*at present*	16
a little	25	*a playoff system*	26	*health inequalities*	12	*at one*	16
and all	24	*a salary cap*	25	*future research*	12	*a lingua franca*	16
a source of	24	*a result of*	25	*freudian theory*	12	*a group of*	16
a world where	23	*a little*	25	*for instance*	12	*because of*	15
a world of	23	*college football*	24	*for example*	12	*a member of*	15
a way to	23	*a sense of*	24	*new information*	11	*a means of*	15
a way that	23	*a lot of*	24	*neurochemistry of*	11	*a lot of*	15
a way of	23	*welfare recipients*	23	*nervous system*	11	*a long way*	15
so that	21	*our society*	23	*in line*	11	*a long time*	15
at all	21	*look at*	23	*in heaven*	11	*but for*	14
all know that	21	*has to*	23	*in general*	11	*british universities*	14
all human beings	21	*genetic research*	23	*in fact*	11	*british students*	14
a few	21	*water pollution*	22	*in defeat*	11	*big business*	14
all the time	20	*the welfare*	22	*in contrast*	11	*a second resolution*	14
all over	20	*the homeless*	22	*other factors*	10	*a result of*	14
all of us	20	*playoff system*	22	*optic array*	10	*a war against*	13
all of them	20	*on television*	22	*opposed to*	10	*a time when*	13

When *Porticle* [3]prefabs are examined, the most frequent by far is found to be *I think that.* The frequency with which the stance markers, *I believe* and *I think* follow the conjunction *and* (22 per 100,000) in *Porticle* is all the more remarkable when the three control corpora are found to have no occurrences of this sequence. Further down the list are *I really think* and *I know that.* The [3]prefab *all know that* appears to be a calque on the Portuguese expression *todos sabem que* and is thus admitted under criterion 6 of the Criteria used for culling N-grams to obtain prefabs, as discussed in Section 3.12. The prominent position of the four prefabs, I *think that, I really think, but I think,* and *I know that* in the first twenty [3]prefabs extracted from *Porticle* has clear links with the findings of Section 3.13. The writers in *Porticle* overuse certain [3]prefabs to express interpersonal

meanings and epistemic stance compared to their native-speaker counterparts. Only one such expression of interpersonal meaning is found among the twenty most frequent trigrams in the three control corpora, namely *I believe that* in the *Bawe* corpus (Table 3.26).The five trigrams with *all* stand out clearly, occurring as they do in sequence 5-9 in table 3.26.

Table 3.26 First twenty [3]prefabs from the four corpora in descending order of frequency

	Porticle	*Locness*	*Bawe*	*CofE*
1	i think that	a good example	in other words	a good society
2	a world of	a great deal	in order to	a free press
3	a way to	a little more	in most cases	a few years
4	a way that	a large number	in line with	a debate about
5	all know that	a means of	i believe that	a lingua franca
6	all human beings	nuclear power	human resource management	a group of
7	all the time	capital punishment	great deal of	a member of
8	all of us	a member of	goes on to	a means of
9	all of them	a number of	first of all	a lot of
10	as i said	a playoff system	by means of	a long way
11	as far as	a salary cap	be used to	a long time
12	as an example	a result of	at this stage	a result of
13	but i think	a social problem	at this point	a second resolution
14	coal and steel	a way of	at the end	a set of
15	have the chance	a time when	at the beginning	a time when
16	have access to	a symbol of	at a time	a war against
17	get in touch	a strong argument	as well as	and the rest
18	i really think	a way to	as opposed to	business and management
19	i know that	a way that	as much as	as a whole
20	i hope that	this argument	as long as	as a result

The creation of a word frequency list for *all* reveals that *Porticle* writers use *all* at least twice as frequently as the writers in the three control corpora (728 times in *Porticle* vs. 348 times in *Locness*, 190 times in *Bawe* and 327 times in *CofE*). This finding for *Porticle* is confirmed for all of the non-native sub-corpora which form part of *ICLE* (Ringbom 1998: 46). The five trigrams in *Porticle* which begin with *all* show the typical phraseologies that *all* forms part of in the Portuguese learner corpus. It shall be shown in the next chapter that overuse of *all* is very often associated with dogmatism in writing.

When the 20 most frequent 4grams in *Porticle* and *Locness* are inspected a good deal of fragmentariness can be found among the sequences. Compared with the cull of

trigrams, more sequences have to be eliminated. The five selected [4]prefabs are marked in blue in Table 3.27. An examination of the frequency lists of 4grams in *Porticle* and *Locness* shows that the adversative prefab *on the other hand* occurs frequently in both corpora (Table 3.27). The writers in *Porticle*, however, use it at least twice as frequently as the writers in the other three corpora (42 occurrences per 100,000 words in *Porticle* compared to 19, 18, and seven in the three NS corpora). There is a remarkable amount of recycled language found among the most frequent 4grams. For instance, 14 of the 20 most frequent 4grams from *Porticle* contain fragments from the essay titles (Appendix 2).

Table 3.27 The twenty most frequent 4grams extracted from *Porticle* and *Locness* (frequencies normalized to a basis per 100,000 words of text)

	Porticle	Freq.	*Locness*	Freq
1	on the other hand	42	in the united states	42
2	root of all evil	29	prayer in public schools	25
3	i would like to	28	on the other hand	19
4	the root of all	28	the death penalty is	17
5	my point of view	25	lowering the drinking age	15
5	more equal than others	24	the drinking age would	15
6	all men are equal	23	of the death penalty	14
7	are more equal than	23	root of all evil	14
8	at the same time	23	the root of all	14
9	we live in a	23	is one of the	13
10	for dreaming and imagination	21	of prayer in public	13
11	one of the most	21	of the united states	13
12	crime does not pay	20	teaching of new age	13
13	is the root of	20	the fact that the	13
14	all over the world	19	the teaching of new	13
15	place for dreaming and	19	to the fact that	13
16	it is important to	18	as a result of	12
17	money is the root	18	in the case of	12
18	the end of the	18	due to the fact	11
19	for the real world	17	the only way to	11
20	the opium of the	17	a great deal of	

As the size of the N-gram increases, culling becomes more straightforward. Fewer longer sequences need to be eliminated than is the case with bigrams. If a longer sequence recurs with any frequency it is more likely to be a prefab. The larger N-grams yielded relatively few compound adverbs, conjunctions and prepositions and held many more compound or complex noun phrases and complements which revealed more about the content of the essays (e.g. 'the sex and violence on the television', 'the ineffectiveness of the death penalty', both found twice in *Locness*) than about the structure of the

argumentation, the clause relations and the writer's stance. One 5-gram (a five-word sequence) 'what I really hate is' proved to be a configuration of stance which framed and was part of a longer recurrent ten-word prefab:

> **What I really hate is/ when they come in and/** say they're doing too much and they're tired from committee work.
>
> What I really hate is/ when they come in and/ give examples from the sixties.

These latter large prefabrications are extracted from student feedback about university education as reported in *The Times Higher Education Supplement* within the *CofE* corpus.

3.15 Among prefabs

When all of the prefabs have been selected from each corpus and lists of them have been compiled, it is then possible to assess the degree of prefab use for each group of writers. Lists of all the prefabs selected in the four corpora are contained in Appendix 5 and Appendix 6. The *Porticle* writers are found generally to use fewer prefabs than the NS writers except in the case of certain categories. The [5]prefabs present an exception to this overall tendency with the *Porticle* writers using more of them than the writers in any of the other corpora. The *Bawe* corpus contains slightly fewer prefabs but has a greater number of [3]prefabs and [4]prefabs than the *Porticle* corpus.

This underuse of prefabs in the *Porticle* corpus is the overall impression gained from looking at the normalized figures tabulated in Table 3.28. The difference becomes clearer when the numbers of prefabs used in each corpus is translated into the number of words that make up these prefabs. When the percentage of words in each corpus which form part of prefabs is calculated, there is a 1%-2% difference in prefabrication between *Porticle* and the three control corpora (final row, Table 3.28). Given that each of the four corpora contain more than 100,000 words, this percentage difference means that 1,000 to 2000 more words form part of prefabs in each of the three control corpora.

The *Porticle* writers not only use fewer prefab tokens but they also use fewer prefab types (Table 3.28). This results in the NNSs having a significantly lower prefab type-token

ratio compared to the other three corpora. They use fewer prefabs but those that they do use they overuse.

Table 3.28 The total number of prefabs selected from the four corpora
(figures normalized to a basis per 100,000)

	Porticle	**Locness**	**Bawe**	**CofE**
[2]prefabs floor>7	2291	2586	1712	1902
[3]prefabs floor>3	1705	1841	2033	2281
[4] prefabs floor> 1	120	216	330	220
[5] prefabs floor> 1	126	24	51	32
Total no of prefab types	445	562	944	892
Total no of prefab tokens	4129	4667	4124	4447
Prefab type/token ratio	10.49%	12.04%	22.87%	20.11%
% of text made up of prefabs	8.55%	9.68%	10.19%	10.33%

3.16 Conclusion

A projection based on the non-computerised comparative analysis of 5% samples from *Porticle* and *Locness* (reported in Section 3.4.1) predicted that more than 20% of the text of the Portuguese undergraduate argumentative writing would be prefabs. Projections based on this small-scale survey also pointed to the total number of prefabs in *Porticle* being fewer than that contained in the writing of the native speaker controls. In the categorization

of Wiktorsson (2003), adapted for this initial analysis, lexical prefabs refer to the world, textual prefabs to textual relations and pragmatic prefabs are used in the management of interpersonal relationships, especially writer-reader and writer-text. In the comparison of the two samples it was found that the learners in the *Porticle* sample used almost twice as many pragmatic prefabs as the native writers in the *Locness* sample.

The main computer-driven investigation reported in this chapter suggests that, although they recombined words more than the writers in the three NS corpora, the *Porticle* writers used fewer prefabs. They used more N-grams but this resulted in fewer native-like prefabs in their writing. There was marked overuse of certain epistemic stance adverbials such as *I think, I believe, of course,* and *in my opinion.* It was also found that *Porticle* writers more frequently used the collocational frame *it I s____ _____ that,* as a comment clause, usually in sentence-initial position, than the other three NS sub-corpora. As the principal type of pragmatic prefab in the earlier manual analysis was the epistemic stance adverbial, there seems to be a good fit between the small-scale and the larger scale experiments. A fuller analysis of epistemic stance markers is given in Chapter 4.3.

The *Porticle* writers used fewer word types and had a lower lexical density than the NSs (lexical density= 46.51% in *Porticle* vs. 51.14% in *Locness*). An examination of the tagged versions of the four corpora revealed that *Porticle* writers used a smaller proportion of nouns and noun phrases than the NSs. The overuse of pronouns in *Porticle* in comparison with *Locness* was demonstrated using the semantic tagger in Wmatrix. The figures of 14,126 vs. 9661 pronouns was highly significant with a log-likelihood of +1048. The use of such a large number of pronouns by the NNSs suggests that they are using them in contexts where the NSs use NPs to refer to entities, processes, states, and events. As expected in association with fewer nouns, the learners also used fewer prepositions in their texts as compared to the NSs. The overuse of certain auxiliaries as part of the bigrams *we can, I should,* and *I don't* is another feature of the writing in *Porticle* which emerged from this study.

Certain sequences of words are used repeatedly by the Portuguese writers in their argumentative essays in English and not used once in the three native corpora, *Locness*, *Bawe* and *CofE.* Examples of prefabs found only in *Porticle* are: *as I said, as I said before, as far as possible, we can say that, in our days,* and *in what concerns.* The last two sequences in this list, *in our days,* and *in what concerns,* although non-idiomatic, were

classified as prefabs. They are probably calqued on the Portuguese expressions *em nossos dias* and *no que concerne.* Certain other strings, such as *more and more, as far as, in order to, I think that,* and *on the other hand,* although produced by the native speakers, were greatly overused by the non-native learners.

A general impression which I registered while reading the *Porticle* essays is that the style of the language in the learner writing tends to be much more informal than native speaker writing. This sense of the informality of the tenor of the *Porticle* essays is increased by the repeated use of core phrasal verbs, personal comment phrases, and other sequences not often used in academic prose, such as frequent occurrences of *I believe* and *I think* preceded by the conjunction *and*, as noted in the previous section. This question of tenor and style will be addressed in the next chapter where, among other things, the contribution of prefabs to the formality/informality of academic writing style is analyzed. The salient differences in prefab use between *Porticle* and the three control corpora will be examined more closely and, where possible, explained.

In this chapter, clear patterns in prefab use have been found which distinguish the NNS corpus from the NS corpora. The Portuguese undergraduates in *Porticle,* while using more recurrent word sequences (N-grams), used fewer prefabs than the NS writers in the three control corpora. They tended to overuse these prefabs in comparison with NSs and this resulted in a lower type-token ratio of prefabs. The EAP texts which the *Porticle* writers produce are markedly different from the writing in the three NS corpora. The overall effect is that the Portuguese essayists adhere more to the conventions of spoken English than to written English. The spoken-like nature of their writing is apparent in their overuse of pronouns, particularly the first person pronoun, *I,* (Section 3.7) and modal auxiliary verbs (Section 3.12). At the lexical level the choice of core all-purpose verbs like *have* and *get* and the lower lexical density of their writing as compared with the NS corpora (Section 3.9) serves to support this view of *Porticle* writing. At the phraseological level they overused certain submodifier-modifier combinations especially *very important* (Section 3.10). The *Porticle* writers continually deployed a reduced set of epistemic stance markers such as *I think, of course,* and *in my opinion* (Section 3.13) which gives to their writing an overly confessional and intimate tone which has more in common with spoken than with written English.

In the next two chapters the findings from this chapter will be analyzed in order to build up a composite picture of the ways in which the *Porticle* writers' use of prefabs differs from that of native speakers.

Chapter 4 Reading between the lines of NNS and NS writing

> I would just as soon be told that I have used old words. As if the same thoughts did not form a different argument by being differently arranged, just as the same words make different thoughts when arranged differently!
>
> Pascal (1662)

4.0 Introduction

This chapter re-examines the results set out in the previous chapter. The variation found in *Porticle* is interpreted within the context of Second Language Acquisition (SLA) theory, English Language Teaching (ELT) methodology and other areas of applied linguistics which inform the teaching of academic writing. One of the first certainties obtained in this study was that the Portuguese EAP writers in the *Porticle* corpus do actually use prefabs. It soon became clear, however, that they use them differently from their NS counterparts in the control corpora. Earlier CLC researchers, e.g. Kjellmer (1991:124) and Granger (1998b), hypothesized that learners were unlikely to use prefabs to anywhere near the same extent as native-speakers.

> The initial hypothesis was that learners would make less use of prefabs, or conventionalised language, in their writing than native speaker counterparts given that the use of such language is universally presented as typically native like. To use Kjellmer's metaphor (1991:124) I expected learners' building material to be individual bricks rather than prefabricated sections
>
> (Granger 1998b:146)

Some of the differences in the use of prefabs by the non-native writers are statistically significant, and mark genuine differences of compositional style. The next section, 4.1, synthesizes the findings of Chapter 3 and re-examines the main differences between the Portuguese and NS writers of EAP. This further analysis, taking advantage of the latest developments in corpus linguistics software, establishes sharper profiles of the distinctive features of *Porticle* writing and the writing in the control corpora. At certain points in the

investigation, the spoken and academic writing sub-corpora of the *BNC Baby* corpus (Berglund and Wynne 2005) are used to determine which of the four corpora contain more informal prefabs. The corpus linguistics tool, *Wmatrix,* is also deployed to obtain a fuller phraseological characterization of the four corpora.

Section 4.2 examines the role that prefabs play in second language acquisition and the challenges which they present to the language learner. This discussion of the strategies language learners use to build and access their mental lexicon employs ideas from SLA theory and psycholinguistics. Recent approaches to language teaching in the Portuguese educational system, the national syllabus for English in Portuguese secondary-schools and individual learning styles are referred to where relevant.

Research carried out concurrently with this thesis (McKenny 2005b) is presented in the third section of this chapter. This work compared *Porticle* with two of the control corpora using content analysis. This is a distinct methodology, drawn from a separate academic discipline, psychology, which provides an alternative, but complementary, perspective on writer's stance. The configuration of epistemic stance by the *Porticle* writers and the NS writers in two of the control corpora is contrasted using the psychological constructs of 'dogmatism' and 'open-mindedness' (Ertel 1985). A discussion then follows of the implications of this research on dogmatism for the teaching of EAP writing skills.

4.1 Distinctive features of NS and NNS varieties of EAP writing

In an effort to establish what is special about the prefab use in *Porticle*, the texts therein were compared to texts written by groups of NSs who are known, according to corpus design features, to possess varying degrees of expertise in writing. It should not be assumed, when NNS writers deviate from the norm of NS writing, that they are somehow in deficit or need to approximate more to the NS norm. Extremely plausible arguments from contrastive rhetoric (e.g. Grabe and Kaplan 1996:195) would suggest that Portuguese writers bring to the writing of the *Porticle* essays their literacy in Portuguese conventions of discourse and what seems strange to the Anglophone reader is the 'Portugueseness' of their work.

There are some stakeholders in EAP, not least among EAP students themselves, who believe that a better model is provided for learners by the varieties known as English as an international language (EIL) or English as a Lingua Franca (ELF). Promodrou (1997), Flowerdew (2000) and Seidlhofer (2001) have all suggested that such non-native varieties should provide the norm for the student of EAP.

As discussed in the final part of Chapter 3, certain sequences of words are used by the *Porticle* writers in their argumentative essays in English and not used at all in the *Locness* corpus. Other strings, although produced by the NS undergraduates, are overused by the NNS learners. Various theories have been put forward to explain these phenomena. One endearing term used in the literature to characterize certain types of overuse is 'the teddy bear principle' (Hasselgren 1993), which refers to learners coming to depend on what is familiar to them and constantly resorting to words and phrases which they see as tried and tested. They develop a repertoire of such expressions. The term 'fossilization' from SLA characterizes this phenomenon well.

Although the main approach used in this thesis to investigate the phraseology of the essays is 'bottom-up' or empirical there is nonetheless a need to look at these essays as texts and to evaluate their overall effect as discourse. Different characterizations of NNSs' writing considered as discourse can be found in the applied linguistics literature.
Fox (1998) describes over- and underuse in non-native writing:

> When you read the work of even quite advanced students... their language is often stilted, too formal and too high-level; and when it is analysed it is seen that the most common words are used less frequently, and in fewer contexts, than they would be by NSs of English.
>
> (Fox 1998:27)

As a corrective, Fox urges teaching and learning the common words of the language and the most frequent senses of words. However, her comments may relate principally to exceptional students who seek out and learn the most recondite words they can find. Another commentator on the quality of learner writing, Channell (1994:21), claims that advanced learners' English,

> while grammatically... and lexically correct, may sound rather bookish and pedantic to a native speaker. This results in part from an inability to use appropriate vague expressions.

Another criticism often levelled at NNS writing, that it is vague and stereotyped, appears to run counter to the remarks of Fox and Channell. The overuse of *people* was an example, reported in the previous chapter (Section 3.5.2). Words like p*eople* and *thing* often function as semi-structural words in the same way as impersonal uses of pronouns or the passive voice, e.g. 'it is widely believed' and 'you can never trust a stranger'. The use of these vocabulary items of high generality, such as *people* and *thing*, or prefabs like *the fact that* with a very high frequency, contributes to the feeling that NNSs' discourse is vague. The contrary view, that learner writers overstate their case and give too much information, is also found in the literature (e.g. Lorenz 1998).

The comparison of samples from *Porticle* and *Locness* writing reported in Chapter 3.5.1 showed the NNSs using 103 pragmatic prefabs compared to 54 by the NSs. The term 'pragmatic prefab' was taken from Wiktorsson (2003:73). The z-test for comparing the difference between two proportions shows this difference in use of pragmatic prefabs between the NNSs and NSs to be significant at p = .01. Examples of pragmatic prefabs found in *Porticle* are *of course, after all, I mean, in my opinion,* and *I would say.* Flowerdew (2000:151) observes that a corpus of learners' writing compiled in Hong Kong presents underuse of hedging devices leading to writing that is 'too direct'. The findings of the DOTA experiment, as reported below in Section 4 .3, show that the Portuguese contributors to *Porticle*, through inadequate use of hedging devices, create precisely the same effect.

There appear, therefore, to be two contradictory findings. On the one hand, learners' writing is found to be too *informal* in the sense of the term explained in Chapter 1.6. This informality was found in recent investigations of written learner English, e.g. Petch-Tyson, (1998); Wiktorsson, (2003). Other researchers found learners' writing to be bookish and pedantic (e.g. Fox 1998; Channell 1994). One wonders how learner English can be noticeably informal and, at the same time, bookish and pedantic. Is it possible that the same pieces of writing, or at any rate, comparable texts, are being referred to? In the same way that there can be a great deal of variation among the writers in a corpus, as mentioned in the introduction to this chapter, so also can there be a variety of styles instantiated in the same text. An apprentice writer might go from one extreme of (in)formality to the other in the same sentence. The following two sentences taken from two *Porticle* essays illustrate this

mixing of formal and informal styles:

> [It] is definitely true that nowadays the entire world is dominated by science, technology and industrialization but [it] is also true that nobody can live without having a dream, something to fight for. (PTES3006)

> As the great thinker, Karl Marx said, religion is the opium of the people. In my point of view this is also true of TV, but TV is not a Bogey Man. (PTES4002)

Also, in this discussion of learner English, care must be taken to ascertain whether the learners being discussed are all of a comparable level of competence. As is to be expected in a representative corpus, a wide range of writing abilities is contained in *Porticle*. Ringbom (1998:50) observes that the NNSs display a certain prolixity in their writing:

> The limited vocabulary that advanced learners have in comparison with NSs is a main reason for the general impression of learner language as dull, repetitive and unimaginative, with many undeveloped themes. Often they give the impression of verbosity in that many words are, strictly speaking, unnecessary in their contexts

As recorded in Chapter 3, *Porticle* writers, like the other NNSs in *ICLE* studies, have a tendency to use a lower proportion of nouns compared to NSs. It has already been suggested that this different pattern of noun use in *Porticle* fits into a much broader picture, in which *Porticle* writing has many of the characteristics of informal English, while the native-speaker corpora conform more to the norms of formal academic English.

The NNS writers in *Porticle* tend to focus on interpersonal involvement to the detriment of information content. The thematization of many epistemic expressions (e.g. *I think that*) is a good example of this preoccupation with the management of writer-reader relations. The overuse of core nouns (*time, way, society, people, things*), core verbs (*think, get*), auxiliaries (*be, do, have, can*) and most vague quantifiers (*all, some, very*) all contribute to a lack of precision in the writing of all of the *ICLE* sub-corpora, including *Porticle*. As a result, the writing style of the advanced learners may have more characteristics in common with the sub-corpus of conversation and popular fiction contained in the *BNC-baby* corpus than with the sub-corpus of academic prose. A case in point is the use of *think* which occurs 6 times per 10,000 words in the native-speaker corpus, *Locness*, but between 21 and 30 times per 10,000 words in the French, Spanish,

Finnish, Swedish and German corpora and 42 times per 10,000 words in *Porticle*. The currency of certain phrases recurring in *Porticle* was measured in the spoken and academic written English subcorpus of the *BNCBaby* (Berglund and Wynne 2005) to ascertain whether the expressions were more characteristic of spoken British English than of written British English. For completeness, some of the sequences were also checked in the fiction and newsprint sub-corpora of this same corpus.

Most of these phrases do not occur even once in the one-million-word academic sub-corpus of *BNCBaby* (with the exception of *I think*). *A little bit* occurs 13 times per 100,000 words in the spoken sub-corpus (Table 4.1). In corpus linguistics there is always the need for caution when a certain expression under investigation is not attested in a corpus. This can never be taken as positive evidence that the expression in question does not occur in the population of which the corpus is a sample. In other words, absence of evidence is not evidence of absence.

Table 4.1 The occurrence of certain prefabs in *Porticle* and in sub-corpora of *BNCBaby*

	Occurrences per 100,000 in the subcorpora of BNC Baby			
Expression	Academic writing sub-corpus	Spoken sub-corpus	Fiction sub-corpus	Newsprint sub-corpus
I think (189 occurrences in *Porticle*)	7	228	37	13
I think that (117 occurrences in *Porticle*)	-	14	2	-
a little bit (39 occurrences in *Porticle*)	-	13	1	-
a lot of money (28 occurrences)	-	4	2	1
a lot of things (4 occurrences)	-	1	-	-

In her study of Swedish advanced learners of English, Wiktorsson (2003) concludes that her Swedish subjects use an inappropriate style, which results in the appearance of many colloquial or informal prefabs in their writing. The results of the study of Portuguese advanced learners' use of prefabs confirm this same tendency to write in a style more towards the informal end of the formal-informal continuum. Wiktorsson's description of

her Swedish subjects' formal and informal exposure to the English language in school and at home, with a large number of hours hearing British and American English on television, corresponds closely to the experience of the *Porticle* writers in Portugal. Like their Swedish counterparts, most Portuguese learners attain a reasonable command of spoken English thanks partly to their hearing thousands of hours of spoken English in their childhood: as a rule, English language TV programmes are not dubbed into Portuguese. The language learning experience of the parents of the Swedish and Portuguese learners under consideration would have been very different, as they would most likely have had greater exposure to formal written English at school. Portuguese students, in recent decades, have been taught oral English during most of their English learning career and even the written English that they are exposed to tends to be written in an informal style. In EFL textbooks, this informality or 'matiness' is often considered desirable by programme planners and teachers. An input of largely informal language to the learning process may explain why Portuguese students, like students from other language groups, use a smaller proportion of nouns in their writing compared with native speaker writers.

Adopting the formal-informal cline discussed in Chapter 1, a rough categorization of the styles found in the corpora would be that the *Porticle* essayists in general used a more informal writing style than the *Locness* essayists, who in turn tended to use a more informal style than *Bawe* writers. Yet again, there would appear to be a spectrum of writing styles across the three academic corpora and a correlation between writing expertise and formality, with the least expert writers writing the most informally.

One of the findings in Chapter 3 was that the Portuguese EAP writers in the *Porticle* corpus used fewer prefab types and tokens than the NS writers in the other three corpora examined. Nevertheless, the NNS writers did use a substantial quantity of prefabs, tending to overuse certain prefabs to a considerable extent. This is another example of fossilization or the 'teddy bear principle' mentioned above. One explanation for the amount of prefabrication in the learners' writing is that because their writing has a lower lexical density than that of the native speakers, Portuguese writers will inevitably concatenate words more in their creation of EAP texts. Indeed, the question of the direction of causality arises here: do the NNSs draw on fewer word types in their writing because they are using more prefabricated chunks or does their smaller L2 mental lexicon result in their combining fewer lexical items into more frequently recurring prefabs? Other possible explanations are

that learners are less prepared to take risks or have had reduced exposure to a narrower range of possible combinations and are therefore less aware of potential combinations.

These patterns of significant overuse of certain prefabs imply a form of avoidance or under-representation of other possible expressions. In the literature, this kind of language behaviour has been characterised as 'playing safe', opting for a 'safe bet', (Grainger 1998b:148), or depending on 'islands of reliability' or 'fixed anchorage points' (Dechert 1984:227). The metaphor which underlies these expressions, that *language learning is risk taking,* underlies a popular theory of SLA, The Monitor Model (Krashen 1985). Krashen's model might be applied to prefab use in the following way. If learners are intent on monitoring their written output, the flow of prefabs may be impeded as they construct their sentences word by word. An interesting experiment in this connection would be to investigate whether freewriting (i.e. spontaneous stream-of-consciousness writing) contains more prefabs than carefully planned essays with more time for making revisions.

4.2 Cross-linguistic and other influences on the interlanguage in *Porticle*

The term, 'interlanguage', coined by Selinker (1972) has become an indispensable tool for the analysis of learner language. The essays in *Porticle* could be considered as specimens of the written interlanguage of Portuguese third and fourth year university students from 2003. To explain or place in context the overuse and underuse of prefabs in *Porticle* recorded in Chapter 3, two main influences on the learner need to be considered:

1. The influence of the writers' first language, Portuguese, at various levels from that of lexical choice, word formation, the collocability and colligability of words, to the grammatical choices, e.g. the use of plurals, verb tense and verb complementation. These aspects of the learner language can be investigated well through the corpus analytic approach of this thesis.
2. The influence of the Portuguese language on *Porticle* writers at the level of discourse. The textual conventions which prevail in present-day Portuguese society, which are transmitted through the Portuguese educational system, influence Portuguese writers when they write English. The *ICLE* sub-corpora has been investigated

> from a bottom-up perspective, looking at lexis, phraseology and syntax; less research has been done on the influence of L1 discourse conventions on L2 writing.

Second language acquisition can be influenced by the learners' perception of the degree of similarity between the target language and their first language (Kellerman 1983). This learner perception, in turn, affects the amount of cross-linguistic influence (henceforth XLI) from L1 to L2 which takes place at any stage in the acquisition of L2. Other factors which influence a learner's interlanguage include learning style, the teaching emphasis which has brought certain features of L2 to the learner's attention and the informal exposure to L2 which the student has received via travel, living abroad, watching and listening to terrestrial and cable TV, video, the Internet including chat and electronic pen pals and the many forms of modern vocal music in English (rock, rap, soul, hip-hop, dub, etc.). When the term cross-linguistic influence is used, there is no suggestion of a mechanical carryover of L1 structures and lexis to L2 (although this is possible in the case of codeswitching, for example). Learners actively (though not necessarily consciously) choose those aspects of the target language to which they attend. As Larsen-Freeman observed, the second language learner user is 'not merely a black box waiting to be spoonfed NS input' (Larsen-Freeman 1985:443).

Particle writers, for example, use the less formal prefab *go on* in their writing as frequently as *continue* even though this latter verb is lexically congruent with Portuguese *continuar*. Even on this point, caution is required as EFL teachers may well have exhorted these students earlier in their studies to use more phrasal verbs in order to sound more native-like or colloquial.

A simple model of XLI in the development of a learner's interlanguage could be characterised as resulting from competition between two distinct ways of viewing the relationship between L1 and L2:

1. The language learner thinks that L1 and L2 are very similar. When the learner subscribes to this view of the relationship between L1 and L2, XLI tends to occur. Such belief is more likely to be sustained by learners in the early stages of SLA.
2. The language learner thinks that L1 is unique.

Belief in (2) blocks XLI and sometimes results in avoidance. Avoidance is said to occur when specific target language features are under-represented in learner writing in comparison with NS writing. Learners at this stage purposefully avoid any form of XLI. Different Portuguese learners told me in EAP tutorials that they sometimes followed the rule that the greater the difference between Portuguese and the English they were producing, the more accurate the English was likely to be.

A more advanced stage, which many of the *Porticle* writers have attained, is when learners becomes more confident and less reluctant to follow their intuitions about L1-L2 similarities. Three broad stages correspond to these student perceptions about the relationship between L1 and L2 in the learning process.

1. The learner relies heavily on XLI.
2. The learner avoids XLI.
3. The learner realizes that some, but not all, of the similarities between languages are helpful.

They are depicted here as being three distinct phases which occur in succession to one another merely for the sake of clarity. The reality of the learning situation is that learners may be in all three stages at the same time as they struggle to accommodate different aspects of the second language and they may jump from stage one to stage three without giving any indication of going through stage two. This apparently paradoxical behaviour is famously characterised by Kellerman's (1987) U-shaped curve model in which learners use a correct target-language form at one stage, then subsequently replace it with an ungrammatical interlanguage form and finally return to use of the correct target-language form.

Learners may attempt direct translations from their mother tongue and these calques are often opaque to the non-Portuguese speaking reader; e.g. the prefabs *in what concerns X* and *in our day*, found frequently in *Porticle,* are loan translations from the Portuguese. Nonetheless, learners' knowledge of Portuguese usually helps rather than hinders their acquisition of English. The two languages share an extensive array of cognate lexical items. The number of Portuguese-English *faux amis* such as *pretender ≠ pretend*, for example (where Portuguese *pretender* means *to purport to*), is relatively small if compared with the number of Portuguese - English 'real friends' - that is to say, English words that Portuguese

learners have never seen before, but whose meaning they can fairly accurately guess based on their prior knowledge of Portuguese. An example of this would be *interference*, which is a cognate of the Portuguese *interferência*. Other Portuguese prefabs which render their cognate prefabs 'good friends' to the learner are the compound prepositions *em vez de* and *a pesar de* which translate the English compound prepositions *instead of* and *in spite of.*
Skehan (1998) elaborates a robust psycholinguistic theory of how prefabs could be used in a language programme and contribute heuristically to second language acquisition. This model proposes three possible levels of prefab use for adult second language learners, lexicalization, syntacticization and relexicalization (Skehan 1998:90). These stages correspond to stages in children acquiring their first language. This theory would seem to be suitable for use with the mixed ability groups of EAP students.

Another level at which the learner's mother tongue influences their writing is at the level of discourse. It would appear that the discourse conventions of Portuguese are distinct from those of English. Bennett (2006:111), drawing on her experience as a translator of academic Portuguese to English, states that:

> Portugal is one example of culture in which the norms governing the presentation of academic knowledge seem to differ markedly from those employed in the English-speaking world.

According to Bennett (2006:111) the writing in Portuguese journals, and of academic texts produced by scholars and students is largely 'non-analytical, uses language in a non-referential way, and frequently contains an abundance of figurative and ornamental features that would be frowned upon in English texts of the same kind.'

As a translator, Bennett faced a dilemma: whether to translate literally the Portuguese academic texts and run the risk of them not being accepted for publication or to ask the writer's permission to radically alter the structure and style of the text to conform to the expectations and norms of publishers in the English-speaking world. Both options have the same result, according to Bennett, which is 'the silencing of this particular Portuguese way of configuring knowledge' (Bennett 2006:120). Bennett uses a startling term to depict this outcome, *epistemicide*, which she defines as 'the systematic elimination of alternative knowledges that is one of the more sinister symptoms of globalisation'.

Dantas-Whitney and Grabe (1989) found that there was a tendency for Brazilian-Portuguese editorial writing to be more interactionally oriented than the American-English editorial writing with which they compared it. The American-English editorials appeared to be more informationally oriented. The Brazilian-Portuguese writing in this same study was found to be more abstract, formal and elaborated than the American English editorial writing. Contrastive rhetoric has produced many informative and insightful analyses about cross-cultural variation among second language writers (Connor 1999; Grabe and Kaplan 1996). When analyzing interlanguage, and deciding on pedagogic responses including remediation, the writing teacher needs to decide whether a piece of writing is (a) ill-formed, (b) well-formed but unnatural, (c) well-formed and natural or either (a) or (b) and an acceptable manifestation of the writer's cultural identity.

The metaphor of a 'black box' (Larsen-Freeman 1985:443), which recurs in SLA theory, conveys some of the mystery surrounding what goes on inside the language learner's head. The language teacher is sometimes required to intuit, deduce or second-guess what is going on in the mind of the language student. Another approach, sometimes available to the bilingual teacher, is to ask the student to think aloud.

A case in point is the prefab *in what concerns* overused by the *Porticle* writers. This expression is never used by native speakers of English but occurred eight times in *Porticle.* The corresponding Portuguese expression *no que concerne* is relatively rare in Portuguese (644 occurrences in the 180,000,000 word *Cetempublico* corpus of European Portuguese, see Appendix 6). The approximately synonymous expression *no que diz respeito* occurs more than ten times more frequently in *Cetempublico*. The coiners of the calque, *in what concerns* chose, as a mould, a rare and formal expression in Portuguese, which is sentence-initial in more than a third of the occurrences in the *Cetempublico* corpus, possibly associating it with the rather complicated English prefab *as far as …is concerned.*

There exists among teachers a shared and transmitted lore about the most frequent errors committed and pitfalls experienced by students in a country or region on their way to learning English. So well-known have some of these idiosyncratic dialects of English become that they are referred to semi-humorously as *Portuglish, Japlish, Swenglish, Franglais, and Spanglish,* to mention but a few from around the world. Some of these terms, however, e.g. *Singlish* (Singaporean English) need not refer to deficient varieties but to non-standard localized dialects. The publication of papers and books referring to and

analysing these varieties of English attests to the widespread awareness among language teachers of the strength of the influence of the students' mother tongue on their language learning. When such accumulated knowledge is available to the corpus linguist, Tognini-Bellini's (2001:84) methodological imperative to 'respect' and 'accept' the data can be superseded. Hunston (2002:62) refers to the use of 'probes' which are used to seek out specific information believed to be held in a corpus while Carter (2004:137) uses the term 'trigger' for a similar kind of short-cut in searching corpora. Instead of being driven by the data, it was decided, for a change, to be led by the collective teachers' intuitions about the interlanguage variety known as *Portuglish*. Some of the words, phrases and constructions which are considered to be indexical of Portuglish can be used as probes and searched for in the written output of the *Porticle* writers to gauge the amount of cross-linguistic influence, of the simplest kind, there actually is in this writing. Table 4.2 summarises the instances of some of the most well-known examples of Portuglish found in *Porticle*. These expressions could be considered Portuglish prefabs.

As most of my probes consist of verb/noun/adjective + preposition combinations and correspond to relatively superficial characteristics of the text, it is unlikely that their use will uncover all or even most of the XLIs in the Portuguese essays. Much deeper analysis might be needed to reveal XLI at the level of clause, sentence and discourse structure.

Table 4.2 Portuglish expressions found in *Porticle*

Interlanguage expression	LI source	Frequency
I like very much to study English novels	Position of adverbial more flexible in Portuguese	1
the government has afraid	ter medo	3
she has twenty years	ter x anos	2
to stay happy/content	ficar contente	
it depends of your education	depender de	2
they think in money all the time	pensar em	1
this book is of my father	genitive form	
in what concerns...	no que concerne	7

Porticle contained relatively few of these instances of Portuglish. This is to be expected from a group of students who are classifiable as post-intermediate, given the

matriculation requirements for English degrees mentioned in Chapter 3.2. XLI unquestionably plays a significant role in shaping the development of Portuguese EAP students' interlanguage. The collected lore of generations of English teachers (e.g. Fordham, 1995) testifies to this. This can be built upon using corpus linguistic methods. One of the outcomes of my thesis, which I will discuss at greater length in Chapter 5, are publications and workshops to bring to the attention of EAP teachers in Portugal the extent to which 'Portugueseisms' are used by students even at the phraseological level.

Inevitably, the goals and values inherent in the prevailing educational system have an influence on the language student's learning trajectory. There is an intricate web of factors, including teacher training and experience, the pedagogical approach, the syllabus in use, the materials made available to the students, the assessment framework, the classroom ethos, and the value placed on language learning by the student, their family and the wider society which affect the language learning career of those who contributed to *Porticle*.

Milton (1998:190), in a discussion of the teaching of English within the Hong Kong secondary educational system, refers to the way in which the notional-functional approach resulted in certain categories becoming 'overgeneralized, restrictive and dogmatic'. This has resulted in a narrow range of words and phrases being elevated 'to the level of an academic catechism'. Milton is referring, in this instance, to connectives as well as some colourful and complicated (and therefore impressive) expressions. He warns in this piece of the possibly pernicious influences of teaching and language syllabuses. The kinds of overuse (and related underuse) associated with, say, overzealous teaching of discourse markers could be termed 'iatrogenic', extending the medical term (which means *illness induced by the doctor or medicine*) to the classroom. *Porticle* essays are peppered liberally with connectives such as *first of all, on the one hand,* and *however*. These linking devices are often in evidence but somehow they do not always appear to be functioning in the text as they should. This overuse of certain discourse markers might contribute to the sensation of strangeness felt by EAP teachers when appreciating their students' written output and which is tellingly described by Yorio:

> Idiomaticity is a non-phonological "accent", not always attributable to surface language errors, but to a certain undefined quality which many frustrated L2 composition teachers define as "I don't know what's wrong with this, but we just don't say that in English" (Yorio 1989:64).

4.3 Epistemic stance and markers of dogmatism

This section examines writer's stance and ways in which it is realized in argumentative writing. Although the message is important in academic writing, the form in which it is expressed is of equal, or some would say, of more importance. The latter view is illustrated by the famous case of the spoof article published by Alan Sokal (1996), in which bogus content was passed off as an authentic contribution to a peer-reviewed journal. Academic writers have to place themselves and their writing in the context of their disciplinary community. Language learners need to learn how to modulate what they are saying or writing so as to sound neither overly tentative nor too dogmatic or arrogant. As discussed in Chapter 1.10, there are many ways to perform such hedging: lexically, grammatically (e.g. modals) and morphologically (-ish, -y). The terms *precization* and *deprecization* (see above, Chapter 1.10), and their associated verbs *precizate* and *deprecizate* were invented by Naess (1966) to refer to the process of making statements more or less precise. These terms could provide a system for describing how the precision of statements can be varied to a level which is appropriate to the context: this could be by hedging, boosting the statement or making it more vague. J. L. Lemke (1999:1) describes this dimension of language thus:

> We use language to take a stance towards and socially orient ourselves and our texts to others…Whatever we have to say about the world, we can also tell others in the same utterance, to what extent we believe what we say is likely, desirable, important, permissible, surprising, serious or comprehensible.

One of the structures or patterns that Lemke (1999:2) first studied and which she began to see as intimately connected with the evaluative dimension of language was the collocational frame *it is* _____ _____ *that* which usually occurs in sentence-initial position and provides a comment as to the likelihood, desirability, importance, permissibility, surprisingness, seriousness or comprehensibility of the proposition expressed in the main clause. The usual pattern of this [5]prefab is IT +IS + ADV +ADJ+ THAT but there are variations with no adverb or such forms as *it is a shame that*, *it is a pity that*, *it is not fair that* with articles or not in the first open slot and nouns or adjectives in the second open slot. It should be noted that the search for the word pattern *it is* _____ _____ *that* also 'catches' the agentless passive prefab (e.g. *it is well known that*). It is interesting that

Hunston and Francis (2000) working independently, reached very similar conclusions about this sentence-initial structure.

When studying this phenomenon, Oakey (2002) referred to this discourse function as the 'Straw Man' approach in his study of a sub-corpus of economics texts, which he extracted from the BNC. He used this label because the proposition which follows the extraposition is set up by the writer as a superficially plausible point of view which is subsequently undermined or found wanting in the ensuing stretch of text. Often it is swiftly refuted or turned around to suit the writer's purpose. Neff et al. (2003) have observed this phenomenon in a collection of broadsheet editorials.

I searched for examples of this adjective or agentless passive evaluative pattern in the four corpora as indicators of writers' stance. Concordances 4.1, 4.2, 4.3 and 4.4 below shows the results of concordance searches on *Porticle*, *Locness*, *Bawe*, *CofE* using *it is * * that* as the searchword. The concordances were culled to remove lines in which this use of anticipatory *it* did not perform an evaluative function and just happened to be of the same form e.g. clefted sentences such as *it is time that we need.* Interestingly, 17 of the 24 phrases in *Porticle* are to do with the likelihood of the proposition. Five of these are of the form *it is also true that,* with no example of the bare [4] prefab *it is true that.*

The Portuguese writers make use of extraposition or postponement to underline the proposition which follows. This is a framing device similar to the finite peripheral clauses expressing epistemic stance (e.g. *I think,* and *I believe)* which are highly frequent in *Porticle.* Only line 4 of Concordance 4.1 for *Porticle* looks hopeful as a possible representative of the Straw Man device. The refutation is delayed until later in the essay when the writer gives the lie to the widely held belief expressed in the framing of the proposition and suggests that English as a lingua franca may well suffer the same fate as Latin.

Locness has fewer occurrences of this type of comment clause. Only four of the 17 such phrases in *Locness* deal with the likelihood of the proposition of the main clause, while five deal with its desirability. Five of the phrases are agentless passives and nine adverbs are employed.

Concordance 4.1 *it is__ that* expressions culled from 33 found in *Porticle* using IT IS * * THAT as searchword.

1.	issues. Each person has their own personality and **it is now proven that** 60 to 70 percent of a person's
2.	for money and many more are still walking around. **It is incredible** how that little piece of paper or metal
3.	for example those who don't believe in justice because **it is a fact that** she is too slow and sometimes blind.
4.	any ways of thinking, will there ever be a consensus? **It is widely believed that** English is truly the world
5.	are of what was happening around us. In contrast, **it is also true that** crime numbers have not come down
6.	dreamers that destroyed more than have created. **It is the proof that** tecnology allied to sick minds can
7.	who authorise the extradition of teethes and criminals. **It is not fair that** someone does not pay for is crime,
8.	us. Although it is true that it is becoming more efficient, **it is also true that** the justice system does not work
9.	can be so good and, at the same time can be so evil. **It is an issue that** we have to think about. We have to
10.	of distracting themselves and enlarger their horizons. **It is a shame that** many people can only discover new
11.	"some are more equal than others" (Victor Hugo) **it is also true that** all men have the opportunity to dream
12.	and some of them are against crime, however they did it. **It is also true that** all of us have an animal inside us
13.	fact that this issue may have some down sides to it. **It is a pity that** the children must enjoy the good weath
14.	not in touch with reality or is trying to run away from it. **It is also true that** these aspects have positive and
15.	organisation and more wide behaviour. Moreover, **it is more valuable that** one student who has knowledge
16.	country on earth but it wouldn't belong to any nation. **It is well known that** a large number of attempts to
17.	can go beyond our common sense. Nevertheless, **it is also possible that** cultures and identities can
18.	sharing everything and understanding each other. **It is a fact that** humanity is never satisfied, it is constantly
19.	be shocked or troubled by those words or pictures. **It is a fact that** the present TV is not what the viewers
20.	they are not used to put their knowledge in practice. **It is a fact that** the theory is also important but if we
21.	best for both freedom and equality in the long run. **It is a fact, that** discrimination is part of our everyday
22.	students, because there are many ways of using them. **It is a shame** that teachers does not use in their class
23.	in their enterprises someone that they do not trust. **It is also true that** some criminals do not have the will
24.	and other civil rights legislations all around the world, **it is also undeniable that** these laws have not been

A good example of the detective work necessary to track down examples of the Straw Man rhetorical structure is provided by this excerpt from essay number ne000026 from *Locness* where we have to read several paragraphs further into the essay to discover that the statement opening with the extraposed that-clause or agentless passive below (Sentence number 6) is only finally refuted by the writer in Sentence 13 and 14.

6 It is often said that business and ethics don't mix which is why business has its own rules, objectives, ethical standards, and judgments to follow.

7 In business there is more involved than just making a profit and getting ahead; it involves relationships throughout the business environment and there must be a sense of respect and trust involved or else the business will fail.

8 The relationships created in business include the union between the producers and the consumers, the employers and the employees, and the corporation and society.

9 The business system of the United States is a free market system which allows the producers and the consumers control over their end of the system.

10 According to Fortune.

11 This supports my claim because if the producers just worry about their money and not build a relationship with their consumers, the consumers will find someone else to patronize.

12 Without consumer support their business is not making a profit and is forced to close.

13 Business is an economic institution, but like our economy as a whole it has a moral foundation.

14 Lying, fraud, deception, and theft sometimes lead to greater profit and this is morally wrong and unethical according to the book Business Ethics and Common Sense.

Concordance 4.2 *it is __ that* expressions culled from 24 found in search of *Locness* using IT IS * * THAT as searchword.

1	flag flying in memory of our heritage and our ancestors, **it is now believed that** it flies for racial purposes.
2	people of America are not ready for the killing of animals, **it is very possible that** they are not ready for
3	not all of the white population holds racist views, **but it is quite obvious that** it is a very significant problem.
4	the children who have no say in the matter.No, of course **it is not fair that** millions of people in this country do
5	profitable. With a history of tolerance, fame and fortune **it is no wonder that** the statistics rise continuously
6	presented, no clear, cut and dry answer can be found. **It is a shame that** such young children can be the center
7	to insure that products and methods are safe for humans. **It is their belief that** . Animals are needed to teach
8	pain it causes. Social Theory and Practice wrote, **It is also stated that** the state has, Opposition to cap
9	analyze the social problem of criminalization of marijuana **it is clearly shown that** the rewards of legalization
10	use business is a complex fabric of human relationships. **It is often said that** business and ethics don't mix w
11	If Richard is ripped away from his home at this stage **it is highly likely that** he will think it is because of
12	enough and if captured could be raped. They feel that **it is not fair that** men are the only ones. Other women,
13	of the homeless population that they ignore the fact that **it is not right that** this discrepancy exists. What is
14	thousands of gallons of water" (Recycler's 3). Therefore **it is extremely important that** motor oil is disposed
15	the adoption process was never completed. Therefore, **it is only right that** she be returned to her biological
16	petition as serious as men. These claims are very weak. **It is not considered that** maybe the reason why
17	time they reach age sixteen If divorce is widespread, **it is a sign that** we are having a problem with the actual

Bawe has fewer still and four of the prefabs deal with the surprisingness of the main proposition. Five of the phrases are agentless passives and nine adverbs are employed.

Concordance 4.3 *it is__ that* expressions culled from 16 found in search of *Bawe* using IT IS * * THAT as searchword.

1)	ideological principles of freedom and opportunity for all, **it is somewhat inevitable that** the imposition of
2)	However, **it is not always the case** in English as it is in Korean; that is, **it is shown** in the following exam
3)	steadfastness would become his paragon. Certainly **it is not possible that** both men put forward entirely
4)	with which to resolve this conflict remains debatable. **It is generally agreed that** sexual motives are
5)	in speaking as it is indicated. As an illustration, **it is often heard that** ' I wish I have a car' for the second
6)	ware of events taking place. However, through research **it is now known that** sleep is an active state which
7)	are allowed different personalities and outlooks, so **it is not surprising that** some wish to sleep with him,
8)	interrupted sensory pathways in the brain stem. **It is also evident that** the brain stem triggers the muscle
9)	briefly described and discussed above, I suppose **it is quite clear that** in Polish educational institutions,
10)	impaired if more resources are used for the first task. **It is sometimes thought that** the amount of resources
11)	are now proletarians not professionals. Therefore **it is not surprising that** of those who begin to train to

CofE has the fewest number of these constructions and a significant proportion of the stance marker phrases (five) deal with the surprisingness of the main proposition. It should be remembered that this is a corpus of editorials and therefore an important function in this kind of writing is drawing the reader's attention to significant new facts and to connections between the known and the new. No agentless passives are used and this again should not be surprising as editorialists will not usually want to distance themselves so much from their propositions.

The fact that the expert writers use fewer of these adjectival stance markers suggests that a fruitful area of investigation would be to ascertain whether they realize this metapragmatic function through the use of other linguistic resources such as adverbs, adjectives or general nouns.

Concordance 4.4 *it is __ that* expressions culled from 14 found in search of *CofE* using IT IS * * THAT as searchword.

1	The situation is probably the same in the US, and **it is my opinion that** this enabled the hawks to be
2	All have responsibilities they must assume. **It is becoming clear that** the Arab world needs to
3	on pressing challenges such as China and Iraq. But **it is worth remembering that** the United States'
4	fanciful than the nano-meltdown feared by Charles. **It is not surprising that** monarchies are cultural L
5	improved by a desk of Fleet Street sub-editors. **It is also evident that,** in contradistinction to what
6	raised in book-lined homes for many generations, **it is hardly surprising that** some expertise and
7	than almost any company in the world. However, **it is now clear that** Enron used financial engineer
8	small business customers is a case in point. **It is also clear that** the modern capitalist believes in
9	to the ethical dimension in foreign policy, **it is not surprising that** prime minister Blair's speech
10	by the numbers of other pages that link to them, **it is hardly surprising that** blogs are winning over

Applying this collocational frame, *it is __ that*, suggested independently by Lemke (1999) and Hunston and Francis (2000), provides a useful way in to an examination of writer stance in the four corpora. The Corpus of Experts (*CofE*) was so designed in order to provide a representative sample of a genre of English which is self-consciously persuasive. Already it is apparent that *Porticle* writers evaluate their propositions with extrapositioned framing clauses more frequently than the NSs. The technique applied with the collocational frame, *it is __ that* is an example of the use of a probe, mentioned in Section 4.2. In psychological research, this technique has been raised to the level of a globally applied analytic tool.

4.4. Levels of dogmatism and how it relates to writer's stance

I compared the level of dogmatism in three or the corpora, *Porticle*, *Locness*, and *Bawe*. In research comparing the *Porticle* essays with the *Locness* and *Bawe* essays, the level of dogmatism in each corpus was examined (McKenny 2003; 2005a). The instrument used for this investigation was the Dogmatism Text Analysis (DOTA) measure of dogmatism, a content analytic procedure applied to oral and written speech production. DOTA is the creation of the psychologist, Ertel (1985) and his followers (Berth and Romppel 1999). There is potential for cross-pollination between the psychological concepts of dogmatism and open-mindedness and the phraseological study of the expression of evaluation in text and the writer's configuration of stance. DOTA is presented below as an exercise in

contrastive analysis of written texts,which, although it uses different tools, parallels my own linguistic analysis. The focus in both cases is on evaluation and writer's stance.

Corpus linguistics and content analysis are very distinct empirical approaches to the study of language, born within different disciplines and hitherto having had little impact on each other (though see Wilson and Rayson 1993). A brief characterization of these two approaches is that corpus linguistics investigates natural language use in a deliberately atheoretical way and builds theory *a posteriori* or inductively, whereas content analysis does the theoretical or deductive work beforehand in selecting the categories of lexemes and then applying them automatically in the analysis.

Ertel (1985) describes how he sought a linguistic indicator of cognitive closure. He selected categories of lexemes devoid of content, although, of course, not devoid of meaning. Content words were excluded from his coding dictionary. This follows Rokeach's (1960) recommendation that the purely structural features of cognitive or belief systems should be studied rather than their matter. Ertel (1985) isolated six lexeme categories which had common semantic meaning and which seemed to represent something like Rokeach's (1960) 'dogmatism'. An English version of these six categories, which were applied together in the standardised DOTA-tool, is listed in Table 4.1 below together with examples of each.

The English DOTA dictionary, developed by Berth & Romppel (1999), is a translation from the German, supplemented by extensive thesaurus lookups. The German category list contains about 600 terms, while the English dictionary covers 465 words/phrases. The difference in number is due to the fact that around 135 of the items when translated were ambiguous and were interpretable as being either dogmatic or open-minded terms. For example *certain* in *it is certain* is dogmatic but is open-minded in *certain countries are.* An English translation of the complete list of German terms was checked by native speaker colleagues for such ambiguities and the ambiguous expressions held back for further research. I am involved in a project to develop a more representative English DOTA dictionary starting from a representative corpus of English texts and using the categories from a semantic tagger lexicon (both words and phrases) to identify ambiguous terms.

Table 4.3 Six subcategories of the DOTA-dictionary with examples of A-terms and B-terms (adapted from Ertel, 1985:230)

	CATEGORY	A-TERMS (examples)	B-TERMS (examples)
1	FREQUENCY	always forever	not always rarely
2	QUANTITY/AMOUNT	all none every nobody everybody complete	not all several few partly many some
3	DEGREE/MEASURE	absolutely completely totally in every respect extremely principally	very more or less hardly considerably to some degree much
4	CERTAINTY	without question certainly surely in any case naturally	apparently questionable probably doubtful perhaps
5	EXCLUSION	only neither-nor merely nothing-but either-or without exception	not only moreover in addition as well also on the other hand
6	NECESSITY v. POSSIBILITY	must not possible have to cannot impossible	need not can may possible might

Within this framework, closed-minded speakers/writers are predicted to prefer A-words/phrases for example:

frequency (*always, never);*
quantity, amount *(all, each, every);*
degree, measure *(entirely, absolutely);*
certainty *(doubtlessly, necessarily);*
exclusion *(only, either... or*);
necessity vs. possibility (*must, cannot)*

The net result of closed-minded speakers'/writers' tendency to choose more of these A-words/phrases is that their discourse appears to contain exaggerated ideas. Open-minded writers, on the other hand, are expected to prefer B-words/phrases, such as *sometimes, occasionally, not all, some, partly, greatly, possibly, maybe, also, as well, can, need not.* DQ, the *dogmatism quotient,* is based on these predictions. The raw frequencies of A- and B-lexemes obtained from a text unit are related to one another by calculating a quotient: the

frequency of A-lexemes is divided by the frequency of A- plus B-lexemes. Closed-minded writers use a higher proportion of A-words and, therefore, their writing obtains a higher dogmatism quotient than that found in open-minded writers. Closed-minded writers could be said to pre-empt the discussion, while open-minded writers encourage debate with the reader.

Ertel and his followers tested the validity of DOTA by applying it to the speeches and writings of subjects generally believed to be particularly closed or open-minded (Berth and Romppel 1999). The extreme left and right wing of political parties of the Weimar Republic produced higher DQs; Marxian, Hegelian, and Romantic utopian philosophers produced higher DQs than critical philosophers of science (Popper, Albert and Lakatos). It is also significant that individual writers such as Friedrich Hölderlin, Vincent van Gogh and August Strindberg produced higher DQs in their writing when suffering from bouts of psychotic illness (Ertel 1985:234).

In this synoptic list of the main findings in the development of DOTA, dogmatism (greater use of A-terms) and open-mindedness (greater use of B-terms) are attributed to persons rather than to texts. In some applications, the subjects of investigation were political parties or newspapers, but these were viewed as collective identities or personifications. The DOTA content analyst uses texts as data and then, through the quantification of indexical categories contained in the texts, deduces certain cognitive dispositions in the subjects. It is the pivotal importance given to text in both DOTA and corpus linguistics which is the bridge between the two approaches.

For the purposes of this aspect of the research, the following hypothesis was set up:

> The dogmatism quotient of writing is inversely proportional to the level of expertise of the writer, other things being equal.

In other words, the better the writer, the lower the level of dogmatism exhibited in their writing. Obvious counterexamples to this hypothesis are some politicians and religious zealots who, while being overtly dogmatic, are 'good' writers. This concentration on one particular aspect of writing is a simplification or, better, an idealization, which makes it easier to set up a testable hypothesis. If it is established that a writer shows a relatively low degree of dogmatism in his/her writing, then the writing will have expressed writer-reader relations more sensitively. This more finely-tuned modulation of the message suggests that

the writer, having automated most of the mechanics of text creation, has more time to think of the needs of the reader. The experimental data for this work was *Porticle*, and the control corpora were two of the three sub-corpora studied in Chapter 3: (1) *Locness* and (2) The *Bawe* corpus.

Table 4.4 Results of DOTA content analysis for the three corpora (dogmatism quotient {A÷A+B} expressed as a percentage)

CORPUS:	*Porticle*	*Locness*	*Bawe*
(A-terms)	3853	2661	1486
(B-terms)	6093	4991	4761
TOTAL	9946	7652	6247
Average of writers' dogmatism quotients	38.7%	34.9%	28.56%

The results of the DOTA content analysis (Table 4.4) shows the experimental group (*Porticle*) to have a dogmatism indicator of 38.7% vs. 34.9% found in the first control group (*Locness*), which is highly significant (chi square = 392.35 value which results in p=.0001). The results for the high-achieving students in the *Bawe* corpus are even more striking: these students use many fewer dogmatic terms or, more precisely, they underuse them significantly. These DOTA results bear out the original hypothesis that the more expert writers are less dogmatic in their writing. Care should be taken not to press this conclusion too far, however. The apparent dogmatism of the NNSs may be a result of the fact that writers who have not mastered the phraseology and full range of rhetorical devices of the language cannot express the subtleties of argumentation. We accuse them of being dogmatic when perhaps they could not be otherwise, given the current state of their linguistic knowledge. The dogmatism found in *Porticle* may result from a group of factors including confusion about the culturally appropriate register; lack of subtlety of expression; and lack of sufficient exposure to formal written English texts.

For triangulation purposes, a corpus study was made to analyse further the features of *Porticle* revealed by the DOTA test. From word-frequency lists, phrases overused by *Porticle* writers were identified and then carefully examined in concordances to ascertain their function in the text. The browser function of Wordsmith Tools facilitated access to as much co-text as necessary. A closer contextualized reading showed that Portuguese students are not content to aver a proposition but felt the need to put their whole being behind the asseveration. Hunston's (1999:185) concept of the status of a proposition within

a discourse is useful here. She describes how writers give a status to each clause in their text, thus determining the range of responses open to the reader. Clauses are typically averred or attributed.

Porticle contained 157 occurrences of *I* followed by *believe.* Some examples were:
I really believe that this is not a question of theory.
I do believe that the fantasy world is really...
I strongly believe that we must create a better world
Being so I do not truly believe that this solution could be applied.
From my point of view I believe that the years that we, nowadays,
I personally believe that it's important to help others

These examples of 'I-embeddings' (Sanders and Spooren 1997) are comment clauses found to be most frequent in conversation (Biber et al. 1999), which again might suggest that the writing in the *Porticle* corpus tends more towards the informal end of the formal-informal continuum compared to the NS corpora. The Portuguese writers, in their English essays, often give the impression that they are writing in support of the point of view being defended and are putting themselves forward as personal advocates: 'You will believe what I'm telling you because I'm a good person and you, I hope, are a good person.' In a similar way, *in my opinion* occurs 71 times per 100,000 words in *Porticle* compared with only three occurrences in *Locness*. *I'm sure* and *I am sure* both occur nine times in *Porticle* but *I'm sure* occurs three times and *I am sure* twice in *Locness.*

Another striking contrast is the 104 occurrences per 100,000 words of *I think that* used for averral and occasionally for hedging in *Porticle* compared with a mere five occurrences in *Locness*. The simplest (and often the most effective) way of averring something is to state it in a declarative sentence, which usually carries with it the commitment of the writer to its truth or verisimilitude (or acceptability). Aarts and Granger (1998b:137) observed 'striking differences in the way learners and native speakers begin their sentences' through their study of sentence-initial trigrams. They found that the non-native writers tended to begin their sentences with something other than the grammatical subject. The averral systems of the two groups of students (NS and NNS) is worth further study and Hunston's (1999:185) concept of the status of a proposition provides a useful frame of reference for examining how the writers assemble their sentences and the value they give to each in their prose. Petch-Tyson (1998) uses different concepts (writer/reader visibility) but comes to similar conclusions about EFL written discourse: advanced learners

of English tend to be more concerned than native speakers with interpersonal involvement in their writing at the expense of information content.

The findings of the corpus analysis suggest that the variations between the two groups of arguers might be conceptualized on a different level. The Portuguese writers might be thinking more of influencing their readers' feelings and also maintaining a good relationship with them by, for example, writing in an enjoyable way: sometimes they use humour, other times moral indignation or poetic imagery. The *Locness* writers seem to deploy a more impersonal argumentative approach. This may be related to cross-cultural differences in role relationships and peer-solidarity.

The fact that *Locness* writers use *argument* 161 times per 100,000 words compared with only three uses per 100,000 words uses by *Porticle* writers is further evidence for this interpretation. If we lemmatize and include *arguments* and *argue,* the difference in use of these key metadiscursive terms is even more striking: 249 uses by the *Locness* writers vs.15 uses per 100,000 words by the *Porticle* writers.

The structure or pattern, *it...that*, examined in Section 4.3, is a fairly frequent exponent of the evaluative dimension of language. As explained above, it is found more often in sentence-initial position in NNS than in NS writing and comments on the proposition expressed in the main clause. There are variations with no adverb or such forms as it *is a shame that, it is a pity that, it is not fair that* with articles or *not* in the first slot and nouns or adjectives in the second slot. These last three variations are found in *Porticle.*

Ten of the 24 instances of this prefab in *Porticle* are to do with the likelihood of the proposition. Six of these are of the form *it is also true.* This suggests XLI from the Portuguese prefab *também é verdade* and the five occurrences of the five prefabs *it is a fact that* seem also to be loan translations. There is a strong possibility that these are examples of 'padding' beloved of all students who are forced to comply with a minimum word-count for their writing task. Other examples of possible XLI are the four clauses dealing with the desirability of the proposition being framed, namely, *it is a shame that* (x2), *it is a pity that, it is not fair that.* Attributing recurrent expressions to XLI does not assume simple translation. A more layered explanation is that students at an earlier stage of their language-learning career favoured those expressions which had a mother tongue equivalent and so, years later, had readier access to these cognate forms when creating the text which forms part of the *Porticle* corpus.

Table 4.4 reveals that the apprentice writers in *Porticle* overuse or underuse certain expressions compared to the level of usage of the writers in *Locness* and in *Bawe*. This can be interpreted as showing a very clear pattern of distinct levels of dogmatism corresponding to different levels of expertise in EAP writing. A difference is also found between the level of open-mindedness in the two NS corpora. The effect of these patterns is that *Porticle* writers are categorizable, in DOTA terms, as the most dogmatic of the three groups of writers and *Bawe* writers are found to be the least dogmatic.

The next section investigates the functions which the *Porticle* writers are performing through their written language production in order to establish whether they have similar objectives to the NS writers.

4.5 Functions of prefabs in EAP

Chapter 1.0 of this thesis quoted from an aphorism by Pascal (1662) in which he uses an image from tennis to discuss his contribution to theology.

> Let no one say that I have said nothing new; the arrangement of the subject is new. When we play tennis, both players use the same ball, but one of them has a better aim.
>
> (Pascal 1662:247)

On one level, Pascal is claiming that his configuration of the subject matter makes his contribution original. The epigraph to Chapter 4, taken from the same aphorism, confirms this discourse-level interpretation.

> I would just as soon be told that I have used old words. As if the same thoughts did not form a different argument by being differently arranged, just as the same words make different thoughts when arranged differently!
>
> (Pascal 1662:247)

If interpreted at the level of the utterance, however, Pascal's image of tennis ball and delivery could take on a functional interpretation: on the one hand there are words and, on the other, what the language user does with them.

A more recent version of Pascal's suggestions about language functions is found in the theory of Brown and Yule (1983), introduced in Chapter 1.9, where language utterances

are divided into those having a more *transactional* function and those with a more *interactional* function. Brown and Yule (1983:1) emphasize that:

> this division is an analytic convenience. It would be unlikely that, on any occasion, a natural language utterance would be used to fulfil only one function to the total exclusion of the other.

In this chapter, a similarly pragmatic view is taken of the terms chosen to classify prefabs. The 'analytic convenience' is sought which best fits the data obtained from the four corpora. Halliday and Hasan's (1989) functional theory of language, discussed in Chapter 1.9, proved heuristically to be the most tractable system for classifying the prefabs most frequently used by *Porticle* and *Locness* writers. As in all applied linguistic endeavour, the model or theory which best fits the data, or which best explains the practical experience, is to be preferred. After applying each of the systems for arranging the prefabs, discussed in Chapter 1.9, Halliday and Hasan's (1989) model was adopted because it fitted the data most readily and comprehensively. The three metafunctions provided a neat classification of the prefabs found in the four corpora. The model has been described in Chapter 1.9 and is here recapitulated.

The ideational

The prefabs which contribute to the realization of this function are the compound nouns such as *mass media, death penalty*, the phrasal verbs such as *carry on*, and the collocational frames such as *a source of* (all examples taken from *Locness*). Also included under ideational are the compound prepositions (e.g. *in terms of, on behalf of*) and compound connectives (e.g. *with the result that*). An important subset of ideational prefabs are quantifiers such as *a great deal of, many of.*

The interpersonal

The interpersonal function refers to, among other things, the degree of respect and deference shown by the writer to the reader. Another factor in the realization of this function is the level of certainty or tentativeness with which the writer avers each proposition. Thus, Halliday and Hasan's interpersonal function covers expressions of the writer's epistemic and attitudinal stance. Realizations of this function include the choice of pronouns, modal verbs and the use of hedges and boosters. Politeness strategies, face-saving (Brown and Levinson 1987) would also contribute to the interpersonal level.

Another feature of interpersonal meaning is the level of formality used and whether the writer treats the reader as an equal.

The textual

Under this function are included the metalingual references contained in the text, including its signposting, sequencing and references to other writers. A very important dimension, according to Verschueren (2000) is the self-reference or metapragmatic awareness expressed in all texts. NLP scholars have mined from large corpora many thousands of expressions which, in EAP, are used to signal intratextual relationships (e.g. *see above, in the diagram below, in the previous chapter*). Other examples of textual devices are *according to..., as far as... is concerned, with regard to...* and *as for...*

Table 4.5 contains the twenty prefabs occurring most frequently in both *Porticle* and *Locness* divided according to which of the three metafunctions they predominantly serve, within the Halliday and Hasan model discussed above and in Chapter 1.9. If the most frequent prefabs of the two corpora are examined from this functional perspective, it becomes apparent that the NNS writers use prefabs most frequently in their writing to express interpersonal functions whereas the NSs use them mostly to refer to physical, social or political entities in the world (the ideational function) or, to a lesser extent, for creating text (the textual function).

Table 4.5 Twenty most frequent prefabs in *Porticle* and *Locness* arranged according to function. (*Porticle* prefabs in italics; *Locness* prefabs in bold)

IDEATIONAL	INTERPERSONAL	TEXTUAL
a lot	*I think*	*for example*
most of	*I believe*	*such as*
the world	*of course*	*for instance*
real world	*many people*	**such as**
due to	*in my opinion*	**a good example**
because of	*in fact*	**according to**
have to	*modern world*	
death penalty	*I know*	

a chance to	*all men*
the family	*I think that*
public schools	*kind of*
the world	*our society*
a great deal	
a lack of	
a large number	
a little more	
in public	
a long time	
a lot of	
as well	
capital punishment	
nuclear power	

The 20 most frequently occurring prefabs from both *Porticle* (marked in italics) and from *Locness* (marked in bold), as recorded in Chapter 3, Table 3.30 are arranged in descending order of frequency and according to which of the three metafunctions they are mainly serving in the texts they occur in (Table 4.5). Although some of the prefabs could be differently classified, some overall trends in prefab function are clear. The NNSs use prefabs most frequently to manage interpersonal relationships while the NSs use prefabs more frequently to depict the ideational world. There is not a great difference apparent in the number of prefabs the Portuguese and the American undergraduates use to structure their texts. Before this overall divergence in the use of prefabs between the NNS and NS writers could be generalized into a pronouncement that NNSs use prefabs mainly to perform interpersonal functions, further analysis would be needed, taking into account the contribution of the less frequent prefabs,. A possible contributory factor to this divergent deployment of interpersonal prefabs is that the NS is more familiar with the genre of persuasive writing in English and chooses to deal with stance and audience issues less directly. This tendency of NNS writers to be more concerned than native speakers with

interpersonal involvement in their writing at the expense of information content was already registered in Chapter 4.4.

4.6 Conclusion

In this chapter explanations have been sought for the distinctive features of prefab use by the Portuguese EAP writers. A number of factors were shown to contribute to the very special manner in which post-intermediate Portuguese learners design their written texts. These factors include the students' mother tongue; the language learning experience of the apprentice writers; the strategies and hypotheses which inform different stages of their language learning; those aspects of language which their teachers have drawn to their attention; and the cognitive style of the individual learner. The EAP texts which the *Porticle* writers produce as a result of these influences and their own personal style are markedly different from the writing in the three NS corpora.

The result of the DOTA content analysis of *Porticle, Locness* and *Bawe* reported in this chapter showed that there was a highly significant difference in the level of dogmatism found in *Porticle* compared to *Locness* and *Bawe*. The Portuguese writers were much more dogmatic in their writing. When the most frequent prefabs in *Porticle* and *Locness* were examined, it was found that the Portuguese writers used prefabs mainly to express interpersonal meaning while the most frequent prefabs in *Locness* were used to build ideational meaning. The writers in *Porticle* concentrated on writer-reader relations to the detriment of proving their case.

The final chapter of this thesis will make recommendations on how EAP teachers might affirm the communicative value of the Portuguese written interlanguage that has been discussed in this chapter while, at the same time, helping students to write more idiomatically and persuasively.

Chapter 5 Conclusions

5.0 Introduction

It is time to take stock of the position of set expressions in EAP. The previous sentence contains five consecutive prefabs. The first prefab, *it is time to* overlaps with the second, *to take stock of*, illustrating the phenomenon of 'collocation cascade' observed by Gledhill (1995). The reader may have detected that this was an invented example through being aware of the frequency of *it is time to* in journalistic writing (14 times in *BNCBaby* newsprint sub-corpus) and the triteness of *to take stock of* (occurring once in *BNCBaby* in an excerpt from *The Lancet*). Although it is totally prefabricated, the sentence used in this illustrative false start to the chapter would function well in the conclusion of an academic essay. As observed in Chapter 1.12, the imitation of models has long been a valid approach to learning academic writing skills. EAP students, like all writers, need to strike a balance between predictability (prefabness) and creativity (departing from established patterns).

One application of corpora to EAP already suggests itself: learners can use the frequency and range of occurrence of a prefab to decide on its currency and appropriacy for their writing needs. For example, it was found in Chapter 3 that the NSs underuse the epistemic booster *of course* in their writing. While the NNSs use this expression frequently (58 occurrences per 100,000 words in *Particle*), the NS writers seem to be aware that it is no longer used a great deal. The NSs' scant use of *of course* in their writing results from their implicit knowledge of collocational usage and is in line with the usage found in the *BNCBaby* corpus (one occurrence each in the fictional and news sub-corpora). What NNSs lack in terms of implicit knowledge of phraseology due to insufficient exposure to the language could be, to some extent at least, supplemented by information obtainable through corpus analysis.

This chapter discusses applications of the research to the teaching and learning of EAP. Section 5.1 assesses the extent to which the research questions outlined in Chapter 1.0 have been answered. Section 5.2 provides a characterization of the prototypical *Particle* writer based on the research. Section 5.3 examines some of the implications for the design and approach of EAP courses, with particular reference to the teaching situation in Portugal. Section 5.4 discusses pedagogical implications.

5.1 The research questions

As a major part of the research component of this thesis, I compiled the first corpus of Portuguese undergraduate English with the help of a number of Portuguese academics and their students. Each new sub-corpus of essays adds to the number of mother tongues and cultures represented in *ICLE* and opens up new perspectives for investigating the existing sub-corpora and comparing and contrasting EAP writers across a range of linguistic backgrounds.

5.1.1 Quantity of prefab use in the corpora

This sub-section and the following one evaluate the degree to which the research questions have been answered. The first question, set out in the introduction to Chapter 1, underpinned the various pieces of investigation. It is repeated here:

> 1. Do non-native (Portuguese) writers of academic English use a similar quantity of prefabs in their essays in comparison to native speakers?

This question is mainly a quantificational one. A great deal of the research carried out to address that question is contained in Chapter 3. In relation to prefabs consisting of two, three, and four words, the corpus investigation showed that, proportionately, the NS writers used a greater number and a greater range of such prefabs in their writing. On the other hand, the NNSs used a greater number and greater range of five-word and six-word prefabs. Although they used fewer ^{2}prefabs, ^{3}prefabs, and ^{4}prefabs than the NS writers, the differences found between the NNSs and NSs were not great. The NNS writers were using a significant amount of shorter prefabs. 8.5% of their writing was calculated to be prefabs. This was less than the percentage of prefabs in the *Locness* corpus (9.7%) and the percentages of *Bawe* and *CofE* (more than 10%) which consisted of prefabs (Table 3.26). *Porticle* writers' use of prefabs may show that they have attained a fairly good command of English after ten years or more of study. *Porticle* writers compared to *Locness* use a smaller number of prefabs but the ones they use, they tend to use more often. The result is that the *Porticle* corpus has a lower type-token ratio of prefabs than that of the NSs. The difference in the type-token ratio of prefabs is particularly noticeable when *Porticle* is compared with the more accomplished writers in the *Bawe* and *CofE* corpora. This creates an effect of

repetitiveness and limited scope in many of the *Porticle* essays which requires an effort on the part of the reader to discern and follow the line of argument. In answer to Research Question 1: *Porticle* writers overuse a number of prefabs and their writing contains a greater number of longer prefabs (of five words or more). By and large, their use of prefabs is numerically of a similar order of magnitude to that of the NS writers in the three control corpora.

5.1.2 Comparison of the functions of prefab use by NS and NNS writers

The second research question put forward in the introduction to Chapter 1 was concerned with the function of the prefabs used in the corpora.

> 2. Do these non-native writers use prefabs to perform the same functions as native speakers?

In order to address this question, various theories of language function were examined. A version of Halliday and Hasan's (1989) theory of language metafunctions, introduced in Chapter 1.9, was found to work well with the data and had the added advantage of pedagogical utility. To answer the research question, it was first necessary to examine the functions which prefabs perform in the four corpora. The results of the content analysis and the subsequent discussion of corpora findings, recorded in Chapter 4.5, provide the basis for the answer to this research question.

Two corollary questions deriving from this research question are whether NNSs employ different means to perform similar functions or whether they have other linguistic goals. In order to answer these questions, the main functions of prefabs in NNS writing are compared to the NS uses of prefabs and a breakdown of the various functions of the prefabs was made. In the examination of the small samples taken from *Porticle* and *Locness*, reported in Chapter 3.5.1, it was found that the *Porticle* writers used more epistemic prefabs and many fewer lexical prefabs in comparison with the *Locness* writers. This small-scale study would suggest that the NNS writers use prefabs to perform different functions from those of the NSs. The functional analysis of prefabs in *Porticle* and *Locness* reported in Chapter 4.5 pointed to a very distinct deployment of prefabs by the Portuguese EAP writers. Their prefabs serve predominantly interpersonal aims of engaging and convincing the reader through direct and personal intercession by the writer. The *Locness* writers, on

the other hand, use prefabs to refer to the physical or mental world and they produced many compound nouns and noun phrases among their most frequently used prefabs. Two completely different approaches to text creation would appear to be involved here. The NSs concentrate on content and thought (the ideational) while the NNSs concern themselves more with personal relations, stance and feelings (the interpersonal).

A useful feature of the ideational, interpersonal, and textual metafunctions is their suitability as metalingual terms for talking to students about the task of writing argumentative essays. As labels, the three functions are reasonably transparent. With this kind of grid, which clearly delineates the different ways of creating meaning, EAP students can be helped to encode their meanings in texts. These three functions provide a framework for viewing texts from complementary perspectives throughout the writing process. Almost every utterance serves each of the functions simultaneously.

> We cannot pick out one word or one phrase and say this is only experiential meaning, or this is only interpersonal meaning... language is not like that. Every sentence in a text is multifunctional; but not in such a way that you can point to one particular constituent or segment and say this segment has just this function
> (Halliday and Hasan 1989: 23)

The multifunctional nature of most words and phrases notwithstanding, it is often possible to decide the principal function of an utterance in its context. As a term for discussing academic writing, *ideational,* dealing as it does with the subject matter or content of a text, lends itself to advice about getting one's ideas down on paper or on to the computer screen. Although the *interpersonal* and *textual* functions are relevant from the outset, the ideational function is perhaps more basic in generating the raw material or 'stuff' of written text. The *interpersonal* and *textual* functions have more bearing on subsequent drafts and at the editing stage in the production of an academic essay or other piece of persuasive writing. Although utterances making references to other texts are superficially ideational, they have a stronger textual function. This kind of functional approach to the presentation of prefabs makes their utility immediately apparent and aids students in mapping the form on to the function.

5.1.3 Suggestions for further research

A parallel corpus of argumentative essays on the same topics used in *Porticle*, written in Portuguese by the same writers or, at least, by participants with similar educational backgrounds, would make a more direct study of contrastive rhetoric possible. A longitudinal corpus tracing the development of individual learners and groups of learners writing in English throughout the period of their university studies would provide valuable information about Portuguese students' acquisition of EAP writing skills. The corpus work I have conducted so far could be complemented by, and articulated with, other kinds of research including: contrastive rhetorical studies to uncover the distinctive features of discursive and argumentative writing in Portuguese and in English.

Northumbria University, where I now work as a lecturer in EAP, has a large community of international students studying for primary and higher degrees. As the in-sessional EAP tutor for the School of Computer Science, Engineering and Information Science, I have access to large numbers of assignments, reports and dissertations in machine readable form. I intend to build corpora of NS and NNS students' written work with a longitudinal dimension which develops the original idea of the *Bawe* designers of incorporating the grades achieved by each piece of work. Such a corpus could be used to develop materials for writing classes and to provide students with models of successful writing.

5.2 A composite picture of the *Porticle* writer

The use of many different kinds of prefabs in the argumentative writing of third- and fourth-year Portuguese university students majoring in English was examined in Chapter 3. In a range of comparisons with American and British undergraduate writers and with professional journalists, the *Porticle* writers presented a fairly distinctive profile, which suggests that their writing differs from the NSs' writing because of a difference in linguistic competence or writing expertise rather than because of major differences in genre across the corpora. If all the information obtained in this research is collated, a profile of the *Porticle* writer emerges. This 'Identikit' picture depicts the *Porticle* writers as using

significantly more dogmatic expressions in their writing compared to the NS writers, who tend to use more open-minded expressions (Chapter 4.4). Although the *Porticle* writers appear to be much more concerned about building the writer-reader relationship, they fail to show sufficient deference to their readers. Chapter 4.5 reports the results of a functional analysis of the twenty most frequently occurring prefabs in *Porticle* and *Locness*. The Portuguese writers appear to concentrate on the management of interpersonal relations between the writer and reader. This distracts them from the central task of constructing informative sentences written in a fairly impersonal and formal style providing evidence and argumentation in favour of their thesis.

Porticle writers have a tendency to write simply and in a personalized way. They prepose, insert or addend qualificational comments to many sentences, for example the epistemic stance markers *I think, I believe,* and *in my opinion.* When modifying adjectives, they use fewer adverbs types and strikingly overuse *very.* Their prose appears to lack substance due to their underuse of nouns and overuse of pronouns and auxiliary verbs. A more nominal style of writing with a topic-comment information structure needs to be encouraged among Portuguese students of EAP. They must be dissuaded from projecting themselves into their text and from appealing to the better judgement of the reader.

5.3 Implications for EAP course design

If pedagogical is taken in its broadest sense, one implication of my research is that language testing must take cognizance of phraseological competence. One researcher, Wiktorsson (2003), in a longitudinal study of learner writing, found a positive correlation between learners' linguistic competence and the quantity of prefabs in their writing. If the quantity of prefabs in writing proves to be indexical of command of the language, tests measuring the prevalence of prefabs could form part of diagnostic or placement tests. If prefabrication contributes to the idiomaticity and naturalness of language production, as Pawley and Syder (1983) and Wiktorsson (2003) suggest, then prefab use is a new dimension which could usefully be incorporated into a model of communicative competence for language testing.

Several major testing organizations have already begun work to include this aspect of language in their internationally administered language tests (UCLES and TOEFL).

This research contributes to our understanding of phraseology. When parts of it are published or transmitted in seminars, it could help EAP teachers find ways of raising student awareness of phraseology. The seminal work of Howarth (1998) draws attention to EAP students' need to acquire phraseological competence. Although Howarth concentrated mainly on restricted collocations between certain nouns and verbs, I argue that a wider range of prefabs and their communicative functions in text should be brought to the attention of EAP students. Skehan's proposal (1998) that prefabs be used as exemplars in a task-based approach, as discussed in Chapter 4.2, provides a psycholinguistic rationale for incorporating prefabs into the EAP curriculum. Given that EAP has to compete with many disciplines in the Portuguese tertiary curriculum, there is a need to justify any further redistribution of teaching time. The preciousness of time is illustrated by the fact that the student contributors to *Porticle* receive a maximum of 150 contact hours of tuition in English per academic year.

When the functions of prefabs were examined in Chapter 4.5, the intention was not to produce a lexicographical arrangement of prefabs but rather to develop a schema for presenting prefabs to EAP students in a pedagogically useful way. As well as identifying prefabs, ways of making them available to students are needed. L2 students of EAP do not learn prefabs implicitly through exposure as is the case with NSs. Classroom time needs to be devoted to more explicit awareness-raising. Because of the time constraints on university language programmes, explicit learning should normally focus on items with 'high surrender value' (West 1933) and should teach transferable skills which students can use in their independent learning.

The problem of the quality and clarity of the writing in *Porticle* can be addressed on different levels. Howarth (1998) observed that language learners who have mastered the combinatorial rules of English and can produce grammatically well-formed sentences still produce non-native lexical collocations in their academic writing. The problem might be addressed at the curricular level so that a programme of study and teaching materials would focus on the phraseological dimension of written texts. The main insights of this thesis could be transmitted to Portuguese EAP and EFL teachers through in-service teacher

development sessions. While conducting such sessions with experienced Portuguese teachers of English, I discussed the need for raising students' awareness about collocation and phraseology. Several of the teachers acknowledged a certain anxiety about prefabs. These professionals were often unsure about the degree of restrictedness on each of the component words in prefabs. For this reason, they did not always feel competent to teach prefabs to their students. Clearly these teachers would benefit from an accessible set of seminar materials which could possibly be distilled from works like this thesis.

Native speakers of English do not need to study the collocational dimension of language. Their skill in harmoniously combining English words results from their having heard or read a large number of texts from their school days onwards. Learners, on the other hand, do not have recourse to such implicit knowledge and need to build up a learned repertoire of the most frequently used collocations of the target language. This thesis can provide information as to the more frequently used prefabs.

There are several possible approaches to incorporating phraseological considerations in the EAP curriculum. An informal survey of currently available EAP textbooks which I recently conducted showed that few EAP textbooks directly address the issue of prefabs. Several relegated prefabs to a list of links, connectors or sentence adverbials or to an appendix. To date, to the best of my knowledge, no attempt has been made to use prefabs as the vehicle for an EAP course.

The research described in this thesis has shown that Portuguese EAP writers need to consider more carefully the effect their use of prefabs has on their writing style and on their audience. EAP students' attention needs to be drawn to examples of prefabs that they overuse. While this can be done with their own manuscripts or hardcopy texts, it becomes much easier when computers and data projectors are available in the writing class. Human nature is such that no text is more engaging or interesting for students than those texts in the production of which the students themselves participated. This is the basic principle of Seidlhofer's (2003) learning corpus approach to the teaching of writing, which focuses on works in progress by her own students. Sometimes, however, writers have difficulty in reading their own writing with the necessary degree of objectivity. It is this objectivity which a corpus-based approach to text can supply. Students, on seeing overuses (e.g. *in my opinion, I believe, very important*), might become more curious about alternative ways to realize the function served by the prefab. This might, in turn, motivate them to consult a

thesaurus or collocational dictionary for suggestions. Once initiated into concordancing skills, they could examine a corpus, or even use the Internet as a corpus, in order to obtain information about typical collocational choices. Student writers might be encouraged to notice that many of the most frequently overused prefabs (e.g. *of course* and *I think*) partly serve a phatic function which is more appropriate in spoken communication than in written academic text.

NS corpora compiled from texts which exemplify the target genre could be used to point out the moves and phraseological conventions prevalent in the discourse community to which the students aspire. Parallel English and Portuguese corpora could help make students aware of the different discourse conventions in L1 and L2. If the parallel corpora are aligned by sentence, then the students could compare, for example, the various L1 translations of target English prefabs or the different English translations of an L1 prefab.

What I have discovered about the functions of prefabs in creating text has an impact on different levels of the educational system. On the one hand, the analysis of prefabs in student interlanguage, presented in Chapter 3 and Chapter 4, could inform language awareness courses for trainee teachers. In this way, awareness of prefabs might be transmitted by these teachers to their pupils.

The Portuguese EAP students who were the main focus of this thesis needed to develop proficiency in writing formal, impersonal English in order to do well in their degree studies. The informality found in the *Porticle* corpus might be attributable to students' lack of exposure to sufficient quantities of formal written English in their English studies. Misunderstanding of the writing task or insufficient knowledge of the genre are other possible causes of this stylistic mismatch. If the production of formal written English is considered to be a valuable skill for school-leavers, then a longer term solution would be changes to the secondary English syllabus, presenting students with a greater number of formal English texts to read and work with. In the Portuguese context, this would involve, among other things, teachers and teacher educators recommending changes to the secondary and tertiary level English syllabuses.

The educational challenge of empowering students to achieve levels of formality appropriate to a given genre becomes slightly less straightforward when a widely discussed theory from Critical Discourse Analysis is taken into consideration. This theory posits that

the style of public discourse has tended increasingly, in recent years, towards greater informality (Fairclough and Wodak 1997). The discourse of academic English, together with the language used by many public institutions, appears to be changing to a less formal and more personal style. In Chapter 1, the protean nature of academic English was discussed and Elbow (1998) was quoted as being sceptical about its very existence. Steen's (2003) corpus project with his postgraduate students found marked stylistic changes in editorials of *The Times* between 1950 and 2000. This newspaper's editorials had become less formal and more involved and persuasive over the period studied. These findings corroborate Fairclough and Wodak's claim that

> a major change in discursive practices affecting many public institutions in contemporary society is the 'conversationalization' of public discourse.
>
> (Fairclough and Wodak 1997:265)

If this trend is really taking place, the discussion about models for language learning assumes a greater degree of complexity. The assessment of the idiomaticity and persuasiveness of the *Porticle* argumentative essays becomes more relativistic. A cynic might suggest that if the *Porticle* essays are deemed to be too informal, it is merely a matter of waiting some years for the corresponding NS essays and editorials to 'catch up' through continual conversationalization.

5.4 Pedagogical implications

The challenge for the EAP teacher is to find a way to facilitate their students' learning of the productive patterns of prefabs, both fixed and variable. Wiktorsson (2003) requires her prefabs to form a syntactically complete unit. This makes for neater, more intuitively recognizable prefab units. A number of factors need to be taken into consideration: whether prefabs should be arranged in ascending order of length, i.e. according to the number of words they contain; whether they should be arranged in alphabetical order, or according to the semantic field or the wordclass category they belong to; whether grammatical prefabs should be included; and whether they could be arranged according to function. Ideally, the

classification should also make prefab information more accessible to EAP course and syllabus designers.

One notable feature of corpus linguistics is that many of the procedures followed by corpus linguists in their investigation can be transferred with minimal change to the language classroom, resulting in pedagogically valuable activities for experiential learning. The use of text retrieval software to create word frequency lists, lists of multi-word clusters, keyword lists, and concordances from corpora are valuable skills which students can learn and transfer to other areas of their language study. Practice in using these procedures can serve to make language learners more autonomous in their learning. EAP students can be encouraged to play the role of investigators or private detectives in their approach to language learning. Johns (1991), one of the principal advocates of Data Driven Learning (DDL), suggests that this approach empowers the student to become responsible for all aspects of their learning thus 'cutting out the middleman', i.e. the teacher. Language students engage in experiential learning, discover trends, ask their own questions and reach their own conclusions.

Willis (1990) points out in *The Lexical Syllabus*, that language teaching textbooks and pedagogic grammars need to provide help on those language questions which present real difficulties to students, instead of working through all the grammatical topics which are featured in internationally published textbooks. I suggest that Portuguese students of English need help with using a wider variety of prefabs, especially those which are currently underused or avoided in Portuguese learners' English. From what has been learned in this research about the English written by Portuguese undergraduates, a strong case can be made for the preparation of a grammar, bilingual dictionaries and an EAP-style guide prepared exclusively for Portuguese speakers. Although a number of bilingual dictionaries already exist, these 'localised' reference books could carry more guidance on XLI affecting Portuguese writers and on words and phrases which are overused, underused or avoided by Portuguese writers of English. If prefabs are presented in teaching, students are thus provided with units of language which are, by definition, natural sequences of words. As their knowledge and skill with prefabs advance, the students could syntacticize the sequences they have learned and subsequently relexicalize them in the manner described by Skehan (1998).

Various kinds of language corpora can be used to help learners of EAP. Learner corpus material, especially from students with the same mother tongue, can be explored to show students those expressions that Portuguese writers tend to overuse, underuse or avoid. Although objections have sometimes been made against exposing language learners to erroneous English, suggesting that such a teaching approach might instil bad habits in learners, the deployment of a learner corpus is useful to exploit what Tomasello and Herron (1988, 1989) call the garden path technique. In this approach students are induced to make typical interlanguage errors in order raise their awareness about such pitfalls. There are research findings by Tomasello and Herron (1988, 1989) which corroborate the idea that such translation traps are effective ways of addressing negative transfer and overgeneralization errors. They help students notice that there are certain aspects of Portuguese and English which do not have a one-to-one correspondence, and they force students to review their hypotheses as to what is similar and what is different in the two languages. These findings are in accordance with the theory that learning situations that facilitate cognitive comparisons also facilitate learning (Nelson 1987; Tomasello and Herron, 1989).

Corpus analysis of carefully selected learner and native speaker data could provide valuable input to methodology classes for trainee teachers of English, provided that it is used in a structured and sequenced way. Analyses of a learner corpus, such as *Porticle,* or of a corpus of the trainees' own interlanguage, or that of the trainees' own pupils, would also be instructive activities in the language awareness sessions of teacher training courses.

The computerised concordance has been hailed by certain language educators (e.g., Allen 2003) as a powerful new tool which can greatly accelerate language learning. Previous advances in language teaching technology, such as the language laboratory, the video cassette recorder, the computer, CALL programs, and interactive video, were initially received with similarly high expectations. A certain degree of caution is wise in view of the exaggerated hopes placed in the earlier technologies, hopes which could not possibly have been fulfilled. Concordancing can nevertheless contribute to raising student awareness of prefabs. Although words are easily distinguished and recognized in text because they are bounded on either side by white spaces, this is not the case with prefabs. The learner must somehow make a distinction between word sequences which have been generated by the grammatical rules of the language and other sequences which are formulaic or memorized.

Learners may be unaware that they do not know that a given word-string is a prefab. They may have difficulties in descrying a word pattern. A concordance by truncating and realigning text shows very clearly two or more words repeating down the page or screen. Even a casual glance is enough to see that these same words are clustering or 'congealing' to use Cowie's (1998) striking term.

A key-word-in-context (KWIC) concordance of a prefab shows the constituent words of the prefab vertically aligned and in a different colour from the co-text. This increased physical salience can greatly assist learners in recognising and learning the prefabs thus displayed. By sorting the concordance alphabetically on the words immediately to the left or the right of the node word, the student becomes familiar with the principal lexical collocations and the typical colligations of the node word or phrase.

Tognini-Bellini (2001), whose work was discussed in Chapter 2.4, claims that the concordance provides syntagmatic information to its reader when it is read horizontally and paradigmatic information when it is read vertically. Such linguistic knowledge, however, can only be obtained from a concordance when the student has been trained in its use. Language students need to learn this new way of viewing language. The adjustment needed is so great that Sinclair (2003) devoted an entire book to training this skill. The concordance, in order to provide useful knowledge, must have been prepared from a carefully chosen corpus, edited to remove noise, irrelevancies, and duplication and then sorted on the most telling collocates to the left or right of the node word.

Concordancing, having been used to diagnose some of the problems experienced by language learners in Chapters 3 and 4, can rather neatly contribute to a solution. In my own career as an EAP teacher I have used concordancing with students since 1990 and have found that it produces excellent results when well set up and integrated within a balanced programme of activities. The skills learned in carefully conducted classroom sessions can subsequently be used by the student working independently.

Another way in which corpora could have an impact on EAP students' writing is through a technological innovation. A collaborative project between computational linguists and applied linguists could determine the feasibility of an electronic solution to phraseological problems: If students are required to produce idiomatic text and the learning path to the achievement of such writing is arduous (as is well-known) then perhaps a

shortcut could be devised. An add-in electronic phraseology checker could be designed to work within word processing software after the fashion of, for example, the spellchecker or grammar checking devices within Microsoft Word. At each stage of the writing process, this device could warn writers when they produce an unusual sequence of words and might suggest more typical combinations of words, i.e. more frequently occurring ones possibly interconnecting with a thesaurus lookup. They could be shown close matches or be linked to an online collocation dictionary. The sources of typicality could be provided by the World Wide Web used as a 'super-corpus,' or by subject-specific sub-corpora. The challenge in the design would be to avoid prescriptiveness except in a core number of cases say of, say, noun plus preposition or where the collocation is highly restricted.

If the production of idiomatic text could be automated or semi automated, in this way, writers who use this method might, in short- or middle-term, learn to write more idiomatically. By noting the adjustments to their texts suggested by the software program, they could learn the stylistic requirements of their chosen genre. Should such a powerful phraseological *panacea* for the problem of EAP phraseology prove possible, the question as to its desirability arises. How far would EAP students' writing be their own? The already fraught territory of plagiarism would become even more complex.

As discussed in Chapter 3.9, Lorenz (1998:59) analyzed the overuse of adjective intensification (e.g. expressions such as *very important; very interesting)* by advanced German writers of English. These EAP writers showed a tendency towards exaggeration and hyperbole.

> It could, for example, have to do with a certain insecurity among non-native speakers regarding the effectiveness of their own writing. Anxious to make an impression and conscious of the limitations of their linguistic repertoire, they might feel a greater need than native speakers to stress the importance -and the relevance - of what they have to say. This attitude may even partly be induced by writing classes which teach the students to be interesting.

Lorenz's critique of writing teachers who enjoin their students to aim for interestingness recalls Bagnall's (1985:71) criticism of 'the cult of the new' and 'the promotion of self-expression' in the teaching of writing to British schoolchildren in the 1960s (see Chapter 1.12). Perhaps students of EAP writing should be encouraged to be clear and prosaic and to

use the phraseology of those writers who are established in the literature of their discipline. Once they are firmly in control of the mechanics of EAP writing, students can concern themselves with 'being interesting'.

The findings of the present research, if developed further, might suggest a different approach to the teaching of English for academic purposes. It might be pedagogically profitable to pay greater attention to confidence-building through equipping apprentice writers with the lexical and phraseological tools they need to express their meaning and stance comfortably and without undue emphasis. Making novice writers more aware of 'face' (Goffman 1967) and the need to give readers the opportunity to decide for themselves might lead to prose which is more readable and more persuasive. At a more fundamental level, a better understanding of the role prefabs play in language production would enable EAP students to write more natural and idiomatic English.

This thesis is the first corpus-based study of the phraseology of English academic essays written by Portuguese university students. These students were found to overuse a small number of prefabs but generally underused prefabs in their writing in comparison with NS writers. The overused prefabs tended to be more characteristic of spoken English and were often associated with expressions of epistemic stance and writer-reader relations. Content analysis of the essays revealed that the Portuguese writers were much more dogmatic in expressing their views and unable to control features of hedging in their writing. The texts collected in my purpose-built corpus, *Porticle*, resembled stretches of written down speech rather than academic essays. This major varietal difference between the *Porticle* corpus and the control corpora used in this research provides a point of departure for EAP in Portugal. Prefabs need to be explicitly taught as part of the English language curriculum both as building blocks for text but also as vehicles of idiomaticity.

References

Aarts, J. and Granger S. 1998. Tag sequences in learner corpora: a key to interlanguage grammar and discourse. In S. Granger (ed.) *Learner English in Corpora.* London: Longman. 132-142.

Aarts, J., Haan, P. de and Oostdijk, N. (eds.) 1993. *English Language Corpora: Design, Analysis and Exploitation.* Amsterdam: Rodopi.

Aarts, J. and Meijs, W. (eds.) 1984. *Corpus Linguistics*. Amsterdam: Rodopi.

Aijmer, K. 2001. *I think* as a marker of discourse style in argumentative Swedish student writing. In K. Aijmer (ed.) *A Wealth of English: Studies in Honour of Göran Kjellmer.* Göteborg: Acta Universitatis Gothoburgensis. 247-257.

Aijmer, K. 2002. *English Discourse Particles: Evidence from a Corpus*. Amsterdam: Benjamins.

Aijmer, K. and Altenberg, B. (eds.) 1991. *English Corpus Linguistics*. London: Longman.

Aitchison, J. 1987. Reproductive furniture and distinguished professors. In R. Steele and T. Threadgold (eds.) *Language Topics: Essays in Honour Of Michael Halliday*. Amsterdam: John Benjamins. 3-14.

Aitchison, J. 1987. *Words in the Mind*. Oxford: Basil Blackwell.

Allen, R. 2003. *Data-driven learning and vocabulary: investigating the use of concordances with advanced learners of English*. M. Phil. dissertation. Trinity College, Dublin.

Allén, S. 1992. Inaugural speech at the 82nd Nobel Symposium. In J. Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel symposium No. 82, Stockholm, 4-8 August 1991.* Berlin:Mouton de Gruyter. 1-6.

Altenberg, B. 1993. Recurrent verb-complement constructions in the London-Lund Corpus. In J. Aarts, P. de Haan and N. Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation.* Amsterdam: Rodopi. 227-245.

Altenberg, B. and Tapper, M. 1998. The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (ed.) *Learner English on Computer.* London: Longman. 80-93.

Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.) 2003. Corpus Linguistics 2003. *UCREL Technicial Papers Special Issue: Proceedings of the Conference.* Vol 16. Lancaster: University Centre for Computer Corpus Research on Language.

Arnold, I.V. 1973. *The English Word.* Moscow: Vyssaj Skola.

Ascham, R. [1561], 1989. *The Schoolmaster.* L. Ryan, (ed.) London: Associated University Press.

Atkins, S. and Clear, J. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1):1-16.

Austin, J. L. 1962. *How to Do Things with Words*. Oxford: Oxford University Press.

Bagnall, N. 1985. *A Defence of Clichés*. London: Constable.

Baker, M., Francis, G. and Tognini-Bonelli, E. (eds.) 1993. *Text and Technology. In Honour of John Sinclair.* Philadelphia/Amsterdam: John Benjamins.

Bakhtin, M. [1929] 1973. *Problems of Dostoyevsky's Poetics.* Translated by R.W. Rostel. Ann Arbor: Ardis.

Barkema, H. 1993. Idiomaticity in English NPs. Papers from the 13th ICAME conference. Amsterdam: Rodopi. 257-78.

Barros, V., Correia, P. and Pinto, F. 2004. *Screen 2.* Oporto: Porto Editora.

Becker, J. 1975. The phrasal lexicon. In B. Nash and R. Schank (eds.) *Theoretical Issues in Natural Language Processing*. Cambridge, Mass.: Bolt, Beraneck and Newman. 70-73.

Bennett, K. 2006. Critical language study and translation: the case of academic discourse. In J.F. Duarte, A. Assis Rosa and T. Seruya (eds.) *Translation Studies at the Interface of Disciplines*. Amsterdam and Philadelphia: John Benjamins. 111-127.

Benson, M. 1989. The collocational dictionary and the advanced learner. In M. L. Tickoo (ed.) *Learners' Dictionaries: State of the Art.*. Singapore: SEAMEO Regional Language Centre. 84-93.

Benson, M., Benson, E. and Ilson, R. 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations.* Philadelphia, PA: John Benjamins.

Benson, M., Benson, E. and Ilson, R. 1986a. *Lexicographic Description of English.* Amsterdam: John Benjamins.

Berber Sardinha, A. 1996. Writing assessment and corpus linguistics. Paper presented at the Applications of Corpus Linguistics Seminar, Aston University, 19/04/1996.

Berglund, Y. and Wynne, M. 2005. *BNC Baby Corpus.* Oxford: Oxford Text Archive.

Bernstein, B. 1990. *Class, Codes and Control Vol. 4: The structuring of pedagogic discourse.* London: Routledge and Kegan Paul.

Berth, H. and Romppel, M. 1999. Darstellung und Erleben der Wende in Massenmedien. Inhaltsanalytische Untersuchungen am Wendekorpus - zehn Jahre danach/Representation and experience of the "Wende" in mass media. Content analytical inquiries - ten years later. *Medienpsychologie:Zeitschrift fuer Individual und Massenkommunikation*, (1999) Sep; Vol 11(3): 185-199.

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

Biber, D. 1988. *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4):3-57.

Biber, D., Conrad, S. and Cortes, V. 2003. Lexical bundles in speech and writing: an initial taxonomy. In A.Wilson, P. Rayson and T. McEnery (eds.) *Corpus Linguistics by the Lune.* Frankfurt/Main: Peter Lang. 71-92.

Biber, D., Conrad, S. and Cortes, V. 2004. If you look at...: lexical bundles in university teaching and textbooks. *Applied Linguistics* 25: 371-405.

Biber, D., Conrad, S. and Reppen, R. 1994. Corpus-based approaches to issues in applied linguistics. *Applied Linguistics* 15 (2) 169-189.

Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus linguistics: Exploring language structure and use.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

Bizzell, P. 1992. *Academic Discourse and Critical Consciousness.* Pittsburgh, PA: University of Pittsburgh Press.

Blackmore, S. 1999. Meme, myself, I. *New Scientist* 13 March 1999, no 2177. 40-44.

Bloomfield, L. 1935. *Language*. London: Allen and Unwin.

Bloor, T. and Bloor, M. 1991. Cultural expectations and sociopragmatic failure in academic writing. In B. Heaton, P. Howarth and P. Adams (eds.) *Socio-cultural Issues in English for Academic Purposes*. London: Macmillan. 1-13.

Boas, F. 1911. *Handbook of American Indian Languages.* Washington, DC: Smithsonian Institute, Bureau of American Ethnology, Bulletin 40.

Bolinger, D. 1975. *Aspects of language*. Second edition. New York: Harcourt Brace Jovanovich.

Bolinger, D. 1976. Meaning and memory. *Forum Linguisticum* 1: 1–14.

Bolinger, D. 1977. Idioms have relations. *Forum Linguisticum* 2 (2) 157-169.

Botelho, T. and Silva, C. 2004 *Zoom.* Porto: Editora.

Brown, P. and Levinson, S. 1987. *Politeness: Some universals in language use.* Cambridge: Cambridge University Press.

Brown, G. and Yule, G. 1983. *Discourse Analysis.* Cambridge: Cambridge University Press.

Bud, R. and Warner, J. (eds.) 1999. *Instruments of Science: An Historical Encyclopedia.* New York: Garland.

Bühler, K. 1934. *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Jena, Germany: Fischer. Translated by D.F. Goodwin, 1990, as *Theory of Language: The Representational Function of Language (Foundations of Semiotics)* (Philadelphia/Amsterdam: John Benjamins)

Cambridge International Dictionary of English 1995. P. Proctor (ed.) Cambridge: Cambridge University Press.

Carter, R. 2004. *Language and Creativity: The Art of Common Talk*. London: Routledge.

Carter, R. 2005. Word frequency lists and recurrent phrases: new interpretations. Lecture given at Northumbria University. Occasional Durham, Newcastle and Northumbria Universities Joint Seminars. 15/03/2005.

Carter, R. and McCarthy, M.1986. *Vocabulary and Language Teaching.* Harlow: Longman.

Carter, R.A and McCarthy M. J. 2001. Size isn't everything: Spoken English, Corpus and the Classroom. In Research Issues, *TESOL Quarterly* 35 (2) 337-340.

Carter, R. and McCarthy, M.2006. *Cambridge Grammar of English with CD ROM: A Comprehensive Guide.* Cambridge: Cambridge University Press.

Chalker, S. 1990. *English Grammar Word by Word.* Walton-on-Thames: Nelson.

Channell J. 1994. *Vague Language.* Oxford: Oxford University Press.

Chomsky, N. 1957. *Syntactic Structures*. Cambridge, Mass: MIT.

Chomsky, N. 1959. A review of B. F. Skinner's 'Verbal Behavior'. *Language* 35 (1) 26-58.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge: Mass.: MIT.

Chomsky, N. 1966. *Cartesian Linguistics*. Cambridge: Mass.: MIT.

Clear, J. 1993. From Firth principles. Computational tools for the study of collocation. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) 1993. *Text and technology. In honour of John Sinclair.* Philadelphia/Amsterdam: John Benjamins.

COBUILD English Grammar. 2002. London: HarperCollins.

COBUILD English Language Dictionary 1995. J. Sinclair (ed.) London: HarperCollins.

Cohen, L. Mannion, L and Morrison, K. 1989. *Research Methods in Education* (4th edition) London: Routledge.

Cohen, L. J. 1962. *The Diversity of Meaning.* Oxford: Oxford University Press.

Comenius, J.A. 1657. *Magna Didactica.* Translated by M. W. Keatinge. New York: Russell & Russell.

Comenius, J.A. 1658. *Orbis sensualium pictus.* Kila, Montana: Kessinger.

Connor, U. 1999. *Contrastive Rhetoric: Cross-cultural Aspects of Second-language Writing.* Cambridge: Cambridge University Press.

Conrad, S. and Biber, D. 2000. Adverbial marking of stance in speech and writing. In S. Hunston and G. Thompson. *Evaluation in Text.* Oxford: Oxford University Press. 56-73.

Cook, G. 1998. The uses of reality: a reply to Ronald Carter. *English Language Teaching Journal.* 52: 57-63.

Cook, G. and Seidlhofer B. (eds.) 1995. *Principle and Practice in Applied Linguistics: Studies in Honour of H G Widdowson.* Oxford: Oxford University Press.

Cop, M. 1988. The functions of collocation in dictionaries. In Budalex Proceedings: papers from 3rd EURALEX International Congress. 35-56.

Cortes V. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23: 397- 24.

Coulthard, M. (ed.) 1986. *Talking about Text.* Birmingham, UK: English Language Research, Birmingham University.

Council of Europe, 2001. *Common European Framework of Reference for Languages.* Cambridge: Cambridge University Press.

Cowie, A. P. 1988. Stable and creative aspects of vocabulary use. In R. Carter. and M. McCarthy (eds.). *Vocabulary and Language Teaching.* Harlow: Longman. 126-139.

Cowie, A. P. (ed.) 1998. *Phraseology: Theory, Analysis and Applications.* Oxford: Clarendon Press.

Cowie, A.P. and Howarth, P. 1996. Phraseology: a select bibliography. *International Journal of Lexicography* 9(1): 38-51.

Cowie, A. P. and Mackin, R. 1975. *Oxford Dictionary of Current Idiomatic English.* Vol. 1. Oxford: Oxford University Press.

Coxhead, A. 1997. The Academic Wordlist. Available from Paul Nation's website: www.vuw.ac.nz/lals/staff/ paul-nation/vocrefs/vocrefs-15.aspx Last accessed on 25 April 2005.

Cringely, R. X. 1996. *Accidental Empires.* Harmondsworth: Penguin Books.

Crookes, G. 1986. Towards a validated analysis of scientific text structure. *Applied Linguistics* 7: 57-70.

Dantas-Whitney, M. and Grabe, W. 1989. A comparison of Portuguese and English newspaper editorials. Paper presented at the 23rd Annual TESOL Convention, San Antonio, Texas, March 1989.

Dawkins, R. 2000. *The Blind Watchmaker.* [New ed.] London: Penguin.

De Beaugrande, R. 2001. Large corpora, small corpora and the learning of "language". In A. Rosebery, A. Henry and M. Ghadessy (eds.) *Small corpus studies and ELT: theory and practice.* Amsterdam: Benjamins. 3-28.

De Cock, S., Granger, S., Leech, G. and McEnery, T. 1998. An automated approach to the phrasicon of EFL learners. In S. Granger (ed.) *Learner English on Computer.* London: Longman. 67-79.

Dechert, H.W. 1984. Second language production: six hypotheses. In H. W. Dechert, D. Möhle and M Raupach (eds.) *Second Language Productions.* Tübingen: Narr. 211–30.

Dudley-Evans, A. 1986. Genre analysis: an investigation of the introduction and discussion section sections of M.Sc. dissertations. In M. Coulthard (ed.) *Talking about Text.* Birmingham: University of Birmingham (ELR). 128-145.

Dudley-Evans, T. 1989. An outline of the value of genre analysis in LSP work. In C. Lauren and M. Nordman (eds.) *Special Language: From Humans Thinking to Thinking Machines.* Clevedon: Multilingual Matters. 72-79.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19 (1): 61-74.

Elbow, P.1998. *Writing with Power: Techniques for Mastering the Writing Process.* Second edition. New York: Oxford University Press.

Ellis, R. 1994. *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Erman, B. and Warren, B. 2000. The idiom principle and the open choice principle. *Text.* 20(1): 29-62.

Ertel, S. 1985. Content analysis: an alternative approach to open and closed minds. *University of North Carolina High School Journal.* 68: 229-240.

Fairclough, N. and Wodak, R. 1997. Critical discourse analysis. In T. van Dijk (ed.) *Discourse as Social Interaction*. London: Sage. 258-284.

Fernando, C. 1996. *Idioms and Idiomaticity*. Oxford University Press: Oxford.

Feyerband, P. 1987. *Farewell to Reason.* London: Verso.

Fillmore, C. 1978. On the organization of semantic information in the lexicon. In D. Farkas, W. Jacobsen and K. Todys (eds.) *Parasession on the Lexicon: Chicago Linguistic Society.*147-173.

Fillmore C. 1992. "Corpus linguistics" or "Computer-aided armchair linguistics". In J. Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel symposium No. 82, Stockholm, 4-8 August 1991* Berlin: Mouton de Gruyter. 35-60.

Fillmore, C.J., Kay, P. and O'Connor, M.C. 1988. Regularity and idiomaticity in grammatical constructions. *Language* 64(3) 501-538.

Firth, J.R. 1957. *Papers in Linguistics.* London: Oxford University Press.

Fletcher, W. H. 2004. *Phrases in English.* [online]. Available from :http://pie.usna.edu/ [Accessed 7 June 2004].

Flowerdew, J. 1993. An educational, or process, approach to the teaching of professional genres. *English Language Teaching Journal* 47(4): 305-316.

Flowerdew, J. 2000. Discourse community, legitimate peripheral participation, and the nonnative-English-speaking scholar. *TESOL Quarterly* 34: 127-150.

Fordham, M. 1995. *Portuglish.* Lisbon: Plátano Editora.

Fox, G. 1998. Plenary: hocus pocus and graven images. Collocations '98. Talk

given at IATEFL 98, UMIST, Manchester, UK. In P. Grundy (ed.) *International Association of Teachers of English as a Foreign language 1998 Manchester Conference Selections.* Whitstable: IATEFL.

Fox, G. 1998b. Using data in the classroom. In B.Tomlinson (ed.) *Materials Development in Language Teaching.* Cambridge: Cambridge University Press. 25-43.

Garner, M. 2004. *Language: An Ecological Approach.* Oxford: Peter Berg.

Giddens, A. 1991. *Modernity and Self-identity.* Cambridge: Polity.

Gläser, R. 1988. The grading of idiomaticity as a presupposition for a taxonomy of idioms. In W. Hüllen and R. Schulze (eds.) *Understanding the Lexicon.* Max Niemeyer Verlag: Tübingen. 264-279.

Gledhill, C. 1995. Collocational and genre analysis: the phraseology of grammatical items in cancer research abstracts and articles. *Zeitschrift für Anglistik und Amerikanistik.* 43:11-29.

Goffman, E. 1967. *Interaction Ritual: Essays on Face-to-Face Communication.* Garden City, NJ: Erlbaum.

Gold, E. 1967. Language identification in the limit. *Information and Control* 16: 447-474.

Gougenheim, G., Michéa, P., Rivenc, P. and Sauvageot, A. 1956. *L'élaboration du Français Elémentaire.* Paris: Didier.

Grabe, W. and Kaplan, R.B. 1996. *Theory and Practice of Writing: An Applied Linguistic Perspective.* London: Longman.

Granger, S. (ed.) 1998. *Learner English on Computer.* London: Longman.

Granger, S. 1998a. The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (ed.) *Learner English on Computer.* London: Longman. 3-9.

Granger, S. 1998b. Prefabricated patterns in advanced EFL writing: collocations and formulas. In A. P. Cowie (ed.) *Phraseology: Theory, Analysis and Applications.* Oxford: Clarendon Press.

Granger, S. 2003. A bird's eye view of learner corpus research. In S. Granger, J. Hung and S. Petch-Tyson (eds.) 2003. *Computer Learner Corpora, Second Language Acquisition and Foreign Teaching.* Amsterdam: John Benjamins.3-37.

Granger S., Hung, J. and Petch-Tyson, S. (eds.) 2003. *Computer Learner Corpora, Second Language Acquisition and Foreign Teaching.* Amsterdam: John Benjamins.

Granger, S. and Rayson, P. 1998. Automatic profiling of learner texts. In S. Granger (ed.) *Learner English on Computer.* London: Longman.

Halliday, M. A. K. 1978. *Language as Social Semiotic.* London: Edward Arnold.

Halliday, M.A.K. 1989. *Spoken and Written Language.* Victoria: Deakin University.

Halliday, M.A.K. 1994. *An Introduction to Functional Grammar.* Second Edition. London: Edward Arnold.

Halliday, M.A.K. and Hasan R. 1989. *Language, Context and Text: Aspects of Language in a Social-semiotic Perspective.* Second Edition. Oxford: Oxford University Press.

Halliday, M.A.K., McIntosh, A. and Strevens, P. 1964. *The Linguistic Sciences and Language Teaching.* London: Longman.

Harris, R. A. 1993. *The Linguistic Wars.* New York: Oxford University Press.

Harris, Z. 1951. *Methods in Structural Linguistics.* Chicago, Illinois: Chicago University Press.

Hasselgren, A.1993. Lexical teddy bears and advanced learners: a study into the way Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4:237-260.

Hausmann, F. 1979. Un dictionnaire des collocations est-il possible? *Travaux de Linguistique et de Littérature* 17:187-195.

Hindmarsh, R. 1981. *Cambridge English Lexicon.* Cambridge: Cambridge University Press.

Hoey, M. 1983. *On the Surface of Discourse.* London: George Allen and Unwin.

Hoey, M 1997. From concordance to text structure: new uses for computer corpora. In J. Melia and B. Lewandoska (eds.) *Proceedings of PALC 97.* Odz: University Press Odz.

Hoey, M. 1998. Introducing Applied Linguistics: 25 Years on. In *The 31st BAAL Annual Meeting: Language and Literacies* (Plenary paper). The University of Manchester.

Hoey, M. 2000. Persuasive rhetoric in linguistics: a stylistic study of some features of the language of Noam Chomsky. In S. In S. Hunston and G. Thompson (eds.) *Evaluation in Text: Authorial Stance and the Construction of Discourses.* Oxford: Oxford University Press.

Hoffmann, S. and Lehmann, H. M. 2000. Collocational evidence from the British National Corpus. In J. Kirk (ed.). *Corpora Galore: Analysis and Techniques in Describing English.* Amsterdam: Rodopi. 17-32.

Hofland, K. and Johansson, S. 1982. *Word Frequencies in British and American English.* Bergen: The Norwegian Computing Centre for the Humanities.

Hornby, A. S., Gatenby, E. V. and H. Wakefield 1948. *A Learner's Dictionary of Current English.* Oxford: Oxford University Press.

Hounsell, D. 1984. Learning and essay writing. In F. Marton, D. Hounsell and N. Entwhistle (eds.) *The Experience of Learning.* Edinburgh: Scottish Academic Press.

Howarth, P. 1995. *A computer-assisted study of collocations in academic prose, with special reference to grammatical structure and stylistic value.* Unpublished Ph. D. thesis. University of Leeds.

Howarth, P. 1998. Phraseology and second language proficiency. *Applied Linguistics* 19(1): 24-44.

Hudson, J. 1998. *Perspectives on Fixedness: Applied and Theoretical.* Lund: Lund University Press.

Hundt, M. and Mair, C. 1999. "Agile" and "Uptight" Genres: The Corpus-based Approach to Language Change in Progress. *International Journal of Corpus Linguistics* 4 (2) 221-242.

Hunston, S. 2001. Colligation, lexis, pattern, and text. In M.Scott and G. Thompson (eds.) *Patterns of Text.* Amsterdam: John Benjamins. 13-33.

Hunston, S. 2002. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Hunston, S. and Francis, G. 2000. *Pattern Grammar.* Amsterdam: John Benjamins.

Hunston, S. and Thompson, G. (eds.) 2000. *Evaluation in Text: Authorial Stance and the Construction of Discourses.* Oxford: Oxford University Press.

Hyland, K. 1991. A genre description of the argumentative essay. *RELC Journal* 21 (1): 66-79.

Hyland, K. 2000. *Disciplinary Discourses: Social Interactions in Academic Writing.* Harlow: Longman.

Hyland, K. 2002. *Teaching and Researching Writing.* London: Pearson Education

Hymes, D. 1968. The ethnography of speaking. In J.A. Fishman, (ed.) *Readings in the Sociology of Language.* The Hague/Paris: Mouton. 99-138.

Ivanič, R. 1998. *Writing and Identity: The Discoursal Construction of Identity in Academic Writing.* Amsterdam: John Benjamins.

Jackendoff, R. 1997. *The Architecture of the Language Faculty*. Cambridge, MA: The Massachussets Institute of Technology Press.

Johns, T. 1991. From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In T. Johns and P. King (eds.) Classroom Concordancing. *Birmingham University English Language Research Journal* 4: 27-45.

Johansson, S. 1985. Word frequency and text type: some observations based on the LOB corpus of British texts. *Computers and the Humanities* 19: 23-36.

Jones, S. and Sinclair, J. 1974. English lexical collocations. *Cahiers de Lexicologie* 24: 15-61.

Kaplan, R. B. 1966. Cultural thought patterns in intercultural education. *Language Learning* 16: 1-20.

Kaszubski, P. 1998. Enhancing a writing textbook: a national perspective. In S. Granger (ed.) *Learner English on Computer.* London: Longman. 172-185.

Kellerman, E. 1983. Now you see it, now you don't. In S. Gass, and L. Selinker, (eds.) *Language Transfer in Language Learning.* Rowley, MA: Newbury House. 112-134.

Kellerman, E. 1987. *Aspects of transferability in second language acquisition.* Unpublished Ph.D. dissertation. University of Nijmegen.

Kelly, G. and Bazerman, C. 2003. How students argue scientific claims: A rhetorical-semantic analysis. *Applied Linguistics* 24: 28-55.

Kennedy G. 1992. Preferred ways of putting things with implications for language teaching. In J. Svartvik (ed.) *Directions in Corpus Linguistics*: *Proceedings of Nobel Symposium 82 ed.* Berlin: Mouton de Gruyter. 335-373.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics.* London: Longman.

Kennedy, G. 2003. Amplifier collocations in the British National Corpus: implications for English language teaching. *TESOL Quarterly* 37(3): 467-487.

Kenny, A. 1982. *The Computation of Literary Style: An Introduction to Statistics for Students of Literature and Humanities.* Oxford: Pergamon Press.

Kirk, J. M. (ed.) 2000. *Corpora Galore: Analyses and Techniques in Describing English.* Papers from the Nineteenth International Conference on English Language Research on Computerised Corpora (ICAME 1998). Language and Computers: Studies in Practical Linguistics No 30. Amsterdam: Rodopi.

Kirsner, K. 1994. Second language vocabulary learning: the role of implicit processes. In N. Ellis (ed.) *Implicit and Explicit Learning of Languages.* London: Academic. 283-311.

Kjellmer, G. 1982. Some problems relating to the study of collocations in the Brown Corpus. In S. Johansson (ed.) *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities.

Kjellmer, G. 1984. Some thoughts on collocational distinctiveness. In J. Aarts and W. Meijs (eds.) *Corpus Linguistics*. Amsterdam: Rodopi. 163-171.

Kjellmer, G.1987. Aspects of English collocations. In W Meijs (ed.) *Corpus Linguistics and Beyond*. Amsterdam: Rodopi. 133-40.

Kjellmer, G. 1990. Patterns of collocability. In J. W. Aarts and W. Meijs (eds.) *Theory and practice in Corpus Linguistics*. Amsterdam. 163-78.

Kjellmer, G. 1991. A mint of phrases. In K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics*. London: Longman. 111-127.

Kjellmer, G. 1994. *A Dictionary of English Collocations*. 3 volumes. Oxford: Clarendon Press.

Krashen, S. 1985. *The Input Hypothesis: Issues and Implications*, Longman.

Kucera, H. and Francis, W.N. 1967. *Computational Analysis of Present Day English*. Providence, RI: Brown University Press.

Kristeva, J. 1980. *Desire in Language: a Semiotic Approach to Literature and Art*. Translated by T. Gora. New York: Columbia University Press.

Langacker, R.W. 1987. *Foundations of Cognitive Grammar Vol 1: Theoretical Prerequisites*. Stanford, CA : Stanford University Press.

Lakoff, G. 1972. Hedges: a study in meaning criteria and the logic of fuzzy concepts. Papers from the Eighth Regional Meeting, Chicago Linguistics Society, University of Chicago Linguistics Department.

Larsen-Freeman, D. 1985. State of the art on input in second language acquisition. In S.M. Gass, and C.G. Madden, (eds.) *Input in Second Language Acquisition*. Rowley, MA: Newbury. 433-444.

Leech, G. 1991a. The state of the art in corpus linguistics. In K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics: Studies in honour of Jan Svartvik*. London: Longman.

Leech, G. 1991b. Corpora. In K. Malmkjaer (ed.) *The Linguistics Encyclopedia*. London: Routledge. 73-80.

Leech, G.1992. Corpora and theories of linguistic performance. In J. Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel symposium No. 82, Stockholm, 4-8 August 1991.* Berlin: Mouton de Gruyter.

Leech, G. 1998 Preface. In S. Granger (ed.) *Learner English on Computer.* London and New York: Longman. xiv-xx.

Leech, G., Rayson, P. and Wilson, A. *Word Frequencies in Written and Spoken English.* London: Pearson Education.

Leech, G. and Svartvik, J. 1985. *A Communicative Grammar of English.* Cambridge: Cambridge University Press.

Leibniz, G. [1666] 1966. De Arte Combinatoria: On the Art of Combination. In G.H.R. Parkinson (ed.) *Leibniz , Logical Papers.* A selection translated and edited with an introduction.

Lemke, J 1999. Resources for attitudinal meaning: evaluative orientations in text semantics. *Functions of Language* 5(1): 33-56

Lerdahl, F. and Jackendoff, R. 1988. *A Generative Theory of Tonal Music.* London: MIT Press.

Lewis, M. 1993. *The Lexical Approach.* Hove: Language Teaching Publications.

Longman Language Activator.1993. Harlow: Longman.

Longman Essential Activator.1997. Harlow: Longman.

Longman Dictionary of Contemporary English. 1995 (3rd ed.) Harlow: Longman.

Lord, A. 1960. *The Singer of Tales*. Cambridge: Harvard University Press.

Lorenz, G. 1998. Overstatement in advanced learners' writing: stylistic aspects of adjective intensification. In S. Granger (ed.) *Learner English on* Computer. London and New York: Longman. 53-66.

Louw, B. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology. In Honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins. 7-176.

McCarthy, M. J. 1990 *Vocabulary.* Oxford: Oxford University Press.

McCarthy, M. J. and Carter, R. A. 1994. *Language as Discourse: Perspectives for Language Teaching*. London: Longman.

McCarthy, M. J. and Carter, R. A. 2001. Ten criteria for a spoken grammar in E. Hinkel and S. Fotos S. (eds) *New Perspectives on Grammar Teaching in Second Language Classrooms,* Mahwah, NJ: Lawrence Erlbaum Associates, 51-75.

McEnery, T. 1998. Seminar: *Who do the British think they are?* Room 2B13, 2nd Basement, Friday 16 January (1.00-1.55pm) Centre for Computing in the Humanities, Strand Building, King's College London. [online]. Available from: http://www.kcl.ac.uk/humanities/cch/semarch.htm [Accessed 4 October 2003].

McEnery, T. and Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McKenny, J. 2003. Seeing the wood and the trees: reconciling findings from discourse and lexical analysis. Paper at Corpus Linguistics 2003 Conference, University of Lancaster. *University Centre for Computer Corpus Research on Language (UCREL). Technical Papers Volume 16* - Special Edition.

McKenny, J. 2005a. Stance and spin in academic writing. In L Lagerwerf, W. Spooren, and L. Degand. (eds.) *Determination of Information and Tenor in Texts: Proceedings of MAD conference (Multidisciplinary Approaches to Discourse 2003)*. Münster: Nodus Publikationen.115-137.

McKenny, J. 2005b. Content analysis of dogmatism compared with corpus analysis of epistemic stance in student essays. *Information Design Journal + Document Design* 13 (1): 40-49.

Macmillan English Dictionary for Advanced Learners 2002. Oxford: Macmillan.

Malmkjaer, K. 1991. *The Linguistics Encyclopedia*. London: Routledge.

Makkai, A. 1972. *Idiom Structure in English.* The Hague/Paris: Mouton.

Martin, J. R.1984. Language, register and genre. In F. Christie (ed.) *Children Writing: Reader*. Geelong, Australia: Deakin University Press.

Martin, J. R. 1985. Process and text: two aspects of semiosis. In J. Benson and W. Greaves(eds.) *Systemic Perspectives on Discourse. Vol. I: Selected Theoretical Papers from the 9th International Systemic Workshop*. 248-274. Norwood, New Jersey: Ablex.

Meara, P. M. 1980. Vocabulary acquisition: a neglected aspect of language learning. *Language Teaching and Linguistics: Abstracts* 13: 221-46.

Meijs, W. (ed.) 1987. *Corpus Linguistics and Beyond*. Amsterdam: Rodopi.

Mel'čuk, I. 1998. Collocations and Lexical Functions. In A. P. Cowie (ed.) *Phraseology: Theory, Analysis and Applications.* Oxford: Clarendon Press.

Melia, J. M. and B. Lewandoska-Tomaszczyk (eds.) 1998 *Proceedings of PALC 97.* Odz: University Press Odz.

Meunier, F. 1998. Computer tools for the analysis of learner corpora. In S.Granger (ed.) *Learner English in Corpora.* London: Longman. 19-37.

Michael, I. 1987. *The Teaching of English from the Sixteenth Century to 1870.* Cambridge: Cambridge University Press.

Miller, G. 1956. Information and memory. In *Scientific American.* August 1956. San Francisco: W.H. Freeman and Company.

Milton, J. 1996. CALL Design based on analysis of a learner's corpus. Paper delivered at AILA '96, 11th World Congress of Applied Linguistics, Jyvaskyla, Finland 4-9 August 1996.

Milton, J. 1998. Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (ed.) *Learner English on Computer.* London: Longman. 186-198.

Milton, J. and Tong, K. (eds.). 1991. *Text Analysis in Computer Assisted Language Learning.* Hong Kong: Hong Kong University of Science and Technology.

Montaigne, M. de 1580. *The Complete Essays.* Translated and edited with an introduction and notes by M.A. Screech. 2004. London: Penguin.

Moon, R. 1998. *Fixed Expressions and Idioms in English.* Oxford: Clarendon Press.

Moore, T. and Carling, C. 1988. *Limitations of Language* London: Macmillan.

Murray, D. 1990. *Shoptalk: Learning to Write with Writers.* Portsmouth, NH: Heinemann.

Naess, A. 1966 *Communication and Argument: Elements of Applied Semantics.* Oslo: Universitetsforlaget. London: Allen and Unwin Limited.

Nattinger, J. 1980. A lexical phrase grammar for ESL. *TESOL Quarterly* 14: 337-344.

Nattinger, J. 1988. Some current trends in vocabulary teaching. In R. Carter and M. McCarthy (eds.) *Vocabulary and Language Teaching.* Harlow: Longman.

Nattinger, J. and DeCarrico, J. 1992. *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Neff, J., Ballesteros, F., Dafouz, E., Martínez, F. and Rica, J-P. 2003. Formulating writer stance: a contrastive Study of EFL learner corpora. Paper read at Corpus Linguistics 2003 Conference, University of Lancaster. *University Centre for Computer Corpus Research on Language (UCREL). Technical Papers Volume* 16-Special Edition.

Nelson, K.E. 1987. Some observations from the perspective of the rare cognitive comparison theory of language acquisition. In K.E. Nelson and A. van Kleeck (eds.) *Children's Language.* Hillsdale, NJ: Erlbaum (vol. 6, pp 441-445).

Nesi, H., Sharpling, G. and Ganobcsik-Williams, L. 2004. Student papers across the curriculum: designing and developing a corpus of British student writing. *Computers and Composition* 21: 439-450.

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 R. Reppen, S.M. Fitzmaurice and D. Biber (eds.) *Using Corpora to Explore Linguistic Variation.* Amsterdam and Philadelphia: Longman, 111–129.

Ooi, V. 1998. *Computer Corpus Lexicography.* Edinburgh: Edinburgh University Press.

Oxford Advanced Learner's Dictionary of Current English. 1995. Edited by A. S. Hornby. 5th Edition (ed.) J. Crowther. Oxford : Oxford University Press.

Oxford English Dictionary. 1989. 2nd ed / prepared by J. A. Simpson & E.S.C. Weiner. Oxford: Oxford Clarendon Press.

Palmer, H. E. 1917. *The Scientific Study and Teaching of Language.* London: Harrap. Republished by Oxford University Press, 1968, D. Harper (ed.).

Palmer, H. E. 1933. Aids to conversational skill. *Bulletin of the Institute for Research in English Teaching* 90:1-3.

Paltridge, B 1996. Genre, text type and the language learning classroom. *English Language Teaching Journal* 50(3):237-43.

Partington, A. 1998. *Patterns and Meanings*: *Using Corpora for English Language Research and Teaching.* Amsterdam/Philadelphia: John Benjamins.

Partridge, E. 1940 *A Dictionary of Clichés*. London: RKP.

Pascal, B. 1662. *Pensées.* (Translated by A.J. Krailsheimer). Harmondsworth: Penguin.

Pawley, A. and Syder, F. H. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. C. Richards and R. W. Schmidt (eds.) *Language and Communication*. London: Longman. 191-226.

Pedersen, J. 1995. The identification and selection of collocations in technical dictionaries. *Lexicographia* 11: 60-73.

Pennycook, A. 1996. Borrowing others' words: text, ownership, memory, and plagiarism. *TESOL Quarterly* 30:201-230.

Petch-Tyson, S. 1998. Writer/reader visibility in EFL written discourse. In S. Granger (ed.) *Learner English in Corpora.* London: Longman. 107-118.

Peters, A. 1983. *The Units of Language Acquisition*. Cambridge: Cambridge University Press.

Phillips, M. 1984. *Aspects of Text Structure: an Investigation of the Lexical Organization of Text.* Amsterdam, New York, Oxford: North Holland.

Promodrou, L. 1990. English as cultural action. In R. Rossner and R. Bolitho (eds.) *Currents of Change in English Language Teaching.* Oxford: Oxford University Press.

Promodrou, L. 1997. Corpora: the real thing? *English Teaching Professional* 5: 2-6.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1972. *A Grammar of Contemporary English.* Harlow: Longman.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language.* Harlow: Longman.

Raimes, A. 1991. Out of the woods: emerging traditions in the teaching of writing. *TESOL Quarterly* 25: 407-30.

Rayson, P. 2003. *Matrix: a statistical method and software tool for linguistic analysis through corpus comparison.* Unpublished Ph.D. thesis. Lancaster University.

Rayson, P. and Garside, R. 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 (ACL 2000)*. 1-8 October 2000, Hong Kong. 1-6.

Renouf, A. and Sinclair J. 1991. Collocational frameworks in English. In K.Aijmer and B. Altenberg (eds.) 1991. *English corpus linguistics: Studies in honour of Jan Svartvik.* Longman: London.

Ringbom, H. 1998. Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In S. Granger (ed.) *Learner English on Computer.* London: Longman. 191-200.

Rokeach, M. 1960 *The Open and Closed Mind: Investigations Into the Nature of Belief Systems and Personality Systems.* New York: Basic Books.

Rundell, M. and Stock, P. 1992. The corpus revolution. *English Today 30:* 9-14.

Rundell, M. 1998. The corpus of the future and the future of the corpus. Talk at Exeter, special conference on 'New trends in reference science' 29/3/96 [online]. Available from http//www.ruf.rice.edu/~barlow/futcrp.html [Accessed 13 April 2003].

Salager-Meyer, F. 2000. Procrustes' recipe: hedging and positivism. *English for Specific Purposes* 19 (2): 175-187.

Sandberg, C. 1959. Interview published in *The New York Times* [online]. Available from: http://partners.nytimes.com/library/magazine/home/20001217mag-onlanguage.html [Accessed 25 April 2004].

Sanders, J. and Spooren, W. 1997. Perspective, subjectivity, and modality from a cognitive linguistic point of view. In W-A Liebert, G. Redeker. and L.Waugh, (eds.) *Discourse and Perspective in Cognitive Linguistics.* Amsterdam: John Benjamins. 85-114.

Saussure, F. de [1915] 1983. *Course in General Linguistics.* C. Bally and A. Sechehaye (eds.) with the collaboration of A. Reidlinger. Translated and annotated by R. Harris. London: Duckworth.

Schiffrin, D. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.

Schmidt, R., & Frota, S. 1986. Developing basic conversational ability in a second language. A case study of an adult learner of Portuguese. In R. Day (ed.) *Talking to Learn: Conversation in Second Language Acquisition*. Rowley, MA: Newbury House. 237-326.

Schmitt, N. 2000. *Vocabulary in Language Teaching.* Cambridge: Cambridge University Press.

Schmitt, N. (ed.) 2004. *Formulaic Sequences.* Amsterdam: John Benjamins.

Schmitt, N. and Carter, R. 2004. Formulaic sequences in action: An introduction. In N. Schmitt (ed.) *Formulaic Sequences*, Amsterdam: John Benjamins. 1-23.

Schmidt, R., & Frota, S. 1986. Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (ed.) *Talking to Learn: Conversation in Second Language Acquisition.* Rowley, MA: Newbury House. 237-326.

Scholfield, P. 1995. *Quantifying Language.* Clevedon: Multilingual Matters.

Scott, M. 1999. *Wordsmith Tools, Version 3.* Oxford: Oxford University Press.

Scott, M. and Thompson, G. 2001. *Patterns of Text.* Amsterdam: John Benjamins.

Searle, J. 1969. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.

Seidlhofer, B. 2001. Closing a conceptual gap: the case for a description of English as a lingua franca. *International Journal of Applied Linguistics* 11: 133-158.

Seidlhofer, B. 2003. Pedagogy and local learner corpora: working with learner-driven data. In S.Granger, J.Hung and S. Petch-Tyson (eds.) 2003. *Computer Learner Corpora, Second Language Acquisition and Foreign Teaching.* Amsterdam: John Benjamins.

Seidlhofer, B. (ed.) 2003. *Controversies in Applied Linguistics.* Oxford: Oxford University Press.

Selinker, L. 1972. Interlanguage. *International Review of Applied Linguistics* 10: 209-231.

Simon-Vanderbergen, A.-M. 2000. The functions of *I think* in political discourse. *International Journal of Applied Linguistics* 10(1): 41-63.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

International Journal of Applied Linguistics 10(1): 41-63.

Sinclair, J. 1995. Corpus typology- a framework for classification. In G. Melchers and B. Warren (eds.) *Studies in Anglistics.* Stockholm: Almqvist and Wiksell International. 17-33.

Sinclair, J. 1996. Notes from video-conferencing session on lexicology, lexicography and computational linguistics for the Department of English Language and Literature, National University of Singapore, August to October. Manuscript. As reported in V. Ooi, 1998. *Computer Corpus Lexicography.* p.51. Edinburgh: Edinburgh University Press.

Sinclair, J. 1996a *EAGLES. Preliminary recommendations on Corpus Typology* [online]. Available from: http://www.ilc.pi.it/EAGLES96/corpustyp/corpustyp.html

[Accessed 4 November 2004].

Sinclair, J. 2003. *Reading Collocations: An Introduction.* Harlow: Pearson.

Sinclair, J. and Coulthard, M. 1975. *Towards an Analysis of Discourse: The English Used by Teachers and Pupils.* Oxford: Oxford University Press.

Sinclair, J. and Renouf, A. 1988. A lexical syllabus for language learning. In R. Carter and M. McCarthy (eds.) *Vocabulary and Language Teaching.* Oxford: Oxford University Press.

Singleton, D. 1999. *Exploring the Second Language Mental Lexicon.* Cambridge: Cambridge University Press.

Skehan, P. 1998. *A Cognitive Approach to Language Learning.* Oxford: Oxford University Press.

Skelton, J. 1985. The care and maintenance of hedges. *ELT Journal* 42 (1) 37-43.

Skinner, B.F. 1957. *Verbal Behavior.* New York: Appleton-Century-Crofts.

Smadja, F. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1): 143-177.

Sokal, A. 1996. Transgressing the boundaries: toward a transformative hermeneutics of quantum gravity. *Social Text* 46/47: 217-252.

Spellmeyer, K. 1989. A common ground: the essay in the Academy. *College English* 51 (March 1989): 262–76.

Steen, G. 2003. Converationalization in discourse: stylistic changes in editorials of the Times between 1950 and 2000. In L. Lagerwerf, W.Spooren and L.Degand (eds.) *Determination of Information and Tenor in Texts: Proceedings of MAD conference (Multidisciplinary Approaches to Discourse 2003).* 115-124.

Stubbs, M. 1995. Corpus evidence for norms of lexical collocation. In G. Cook and B.Seidlhofer (eds.) *Principles and Practice in Applied Linguistics.* London: Oxford University Press. 245-56.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Stubbs M. 2002. *Words and Phrases: Corpus Studies of Lexical Semantics.* Oxford: Blackwell.

Svartvik, J. 1992. (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel symposium No. 82, Stockholm, 4-8 August 1991* Berlin: Mouton de Gruyter.

Swales, J. 1981. Aspects of article introductions. Aston ESP Research Report No. 1, Language Studies Unit. Birmingham, U.K: University of Aston in Birmingham.

Swales, J. 1984. *Episodes in ESP*. London: Pergamon Press.

Swales, J. 1990. *Genre Analysis*. Cambridge: Cambridge University Press.

Tannen, D. 1982. The myth of orality and literacy. In W. Frawley (ed.) *Linguistics and Literacy.* London: Plenum Press. 37-50.

Tarone, E., Dwyer, E., Gillette, S. and Icke, V. 1981. On the use of the passive in two astrophysics journal papers. *ESP Journal* 1 (2) 1981 (reprinted in Swales 1984).

Thorne, J. 1988. The language of synopsis. In M. Ghadassey (ed.) *Registers of Written English.* London: Pinter Publisher.

Thorndyke, E. L. and Lorge, I. 1938. *A Semantic Count of English Words.* New York: Columbia University Press.

Thurston, J. and Candlin, C.N. 1998. Concordancing and the teaching of the vocabulary of academic English. *English for Academic Purposes* 17 (3) 267-279.

Tickoo, M .L. 1989. (ed.) *Learners' Dictionaries: State of the Art.* Singapore: SEAMEO Regional Language Centre.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work.* Amsterdam: John Benjamins.

Tomasello, M. and Herron, C. 1988. Down the Garden Path: inducing and correcting overgeneralization errors in the foreign language classroom. *Applied Psycholinguistics* 9:237-246.

Tomasello, M. and Herron, C. 1989. Feedback for language transfer errors: the garden path technique. *Studies in Second Language Acquisition* 11:385-395.

Tribble, C. and Jones, G. 1990. *Concordances in the Language Classroom.* London: Longman.

Tribble, C. 1996. *Writing.* Oxford: Oxford University Press.

Tribble, C. 1999. *Genres, keywords, teaching: towards a pedagogic account of the language of project proposals.* Paper read at Teaching and Language Corpora (TALC) Conference 1998. [online].Available from: http://ourworld.compuserve.com/homepagesChristopher_Tribble/Genre.htm#Introduction [Accessed 17 June 2004].

Ure, J. 1971. Lexical density and register differentiation. In J. Perren and J. Trim (eds.) *Applications of Linguistics: Selected Papers of the 2nd International Congress of Applied Linguistics.* Cambridge: Cambridge University Press.

van Ek, J.A. 1975. *The Threshold Level.* Strasbourg: Council of Europe.

van Ek, J.A. and Alexander, L.G. 1980. *The Threshold Level of English.* Oxford: Pergamon Press.

van Ek, J.A., Alexander, L.G., and Fitzpatrick, M.A. 1977. *Waystage English.* Oxford: Pergamon Press.

van Ek, J. A. and Trim, J.L.M. 1996. *Vantage Level.* Strasbourg : Council for Cultural Co-operation.

Van Roey, J. 1990. *French-English Contrastive Lexicology: An Introduction.* Louvain la Neuve: Peeters.

Verschueren, J. 2000. Notes on the role of metapragmatic awareness in language use. *Pragmatics* 10(4) 439-456.

Wales, K. 1999. The British National Spoken Corpus thing and that sort of thing: the interesting thing about 'thing'. In P. Lucko and U. Carls. (eds.) *Form, Function and Variation.* Frankfurt: Peter Lang. 79-88.

Wales, K. 2002. *A Dictionary of Stylistics.* Second edition. London: Longman.

Wallace, M. 1982. *Teaching Vocabulary*. London: Heinemann.

Warren, B. 1999. An alternative view of stored linguistic knowledge and its relevance to text composition. In J.-O. Östman (ed.) *Pragmatic Aspects of Construction Grammar and Frame Semantics.* Amsterdam: John Benjamins.

Weinreich, U. 1969. Problems in the analysis of idioms. In J. Puhvel (ed.) *Substance and Structure of Language.* Berkeley and Los Angeles: University of California Press.

West, M. 1933. *On Learning to Speak a Foreign Language*. London: Longmans, Green.

West, M. 1953. *A General Service List of English Words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longmans, Green.

Widdowson, H. 1979. *Explorations in Applied Linguistics*. Oxford: Oxford University Press.

Widdowson, H. 1984. *Explorations in Applied Linguistics 2*. Oxford: Oxford University Press.

Widdowson, H.G. 1989. Knowledge of language and ability for use. *Applied Linguistics* 10: 128-37.

Wiktorsson, M. 1998. Compositional and non-compositional aspects of written and spoken texts. Paper presented at the Conceptual Structure, Discourse and Language Conference (CSDL-4) in Atlanta.

Wiktorsson, M. 2003. *Learning Idiomaticity: A corpus-based study of idiomatic expressions in learners' written production.* Lund: Department of English, Lund University.

Wilkins, D. 1976. *Notional Syllabuses*. Oxford: Oxford University Press.

Willis, D. 1990. *The Lexical Syllabus: A New Approach to Language Learning.* London: Harper Collins.

Willis, J. 1998. Lexical phrases and prefabricated chunks: a systematic approach to their teaching. In P. Grundy (ed.) *International Association of Teachers of English as a Foreign language 1998 Manchester Conference Selections.* Whitstable: IATEFL.

Willis, D. and Willis, J. 1988. *The COBUILD English Course*. London: Harper Collins.

Wilson, A. and Rayson, P. 1993. The automatic content analysis of spoken discourse: a report on work in progress. In C. Souter and E. Atwell (eds.) *Corpus-based Computational Linguistics.* Amsterdam: Rodopi. 215-226.

Winter, E. O. 1986. Clause relations as information structure: two basic text structures in English. In M. Coulthard (ed.) 1986. *Talking about Text.* Birmingham, UK: English Language Research, Birmingham University. 88-108.

Winter, E. O. 1977. A clause relational approach to English texts: a study of some predictive lexical items in written discourse. *Instructional Science* 6/1: 1-92.

Womack, P. 1993. What are essays for? *English in Education* 27/1.

Wong-Fillmore, L. 1979. Individual differences in second language acquisition. In C. Fillmore, D. Kempler and W. Wang (eds.) *Individual Differences in Language Ability and Language Behavior.* New York: Academic Press.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, A. and Namba, K. 2003. Formulaic language in a Japanese-English bilingual child: a practical approach to data analysis. *Japanese Journal of Multilingualism and Multiculturalism* 9: 24-51.

Yorio, C. A. 1989. Idiomaticity as an indicator of second language proficiency. In K. Hyltenstam and L. K. Obler *Bilingualism across the lifespan*. Cambridge: Cambridge University Press.

Youmans, G. 1991. A new tool for discourse analysis: the Vocabulary-Management Profile. *Language* 67(4): 763-789.

APPENDICES

Appendix 1

Details of corpora and software referred to in the thesis

Bawe	The British Academic Written English corpus is a collaboration between the Universities of Warwick, Reading and Oxford Brookes
Bank of English	This collection of texts aims has the target of 1 billion words of spoken and written English. It is sponsored by HarperCollins Publishers, Glasgow, and conducted by the COBUILD team at the University of Birmingham, UK.
Birmingham Corpus	proto-version of the COBUILD corpus consisting of 20 million words
BNC	British National Corpus: 100,000,0000 word spoken (10%) and written (90%) corpus of British English of the 1990s.
BNCBaby	Four million-word samples from the BNC representing conversation, academic writing, newsprint, and fiction
BRICLE	Brazilian sub-corpus of *ICLE*
BROWN	The Brown Corpus of Standard American English was the first of the modern, computer readable, general corpora, compiled by W.N. Francis and H. Kucera, Brown University, Providence, RI. The corpus consists of one million words of American English texts printed in 1961.
CANCODE	Cambridge and Nottingham Corpus of Discourse in English is a collection of spoken English that has been built up by Cambridge University Press and the University of Nottingham. The recordings are stored in a computerised database which can be searched with specially designed software. CANCODE comprises 5 million words.
CETEMPUBLICO	Corpus de Extractos de Textos Electrónicos MCT/*Público* is a corpus containing some 180 million words in European Portuguese after an agreement between the Portuguese Ministry for Science and Technology and the Portuguese daily newspaper *Público* signed April 2000 Available from: www.linguateca.pt/cetempublico/
CLAWS7	The Constituent Likelihood Automatic Word-tagging System. A grammatical tagger developed at UCREL
COBUILD	Collins Birmingham University Language Database
CofE	Corpus of Experts. 114,000 word corpus compiled from 133 editorials of *The Guardian, New York Times* to be used as control corpus in the research for this thesis
EAGLES	Expert Advisory Groups on Language Engineering Standards
FLOB	The Freiburg-LOB Corpus of British English from 1991. I million words following design of LOB.
FRICLE	French sub-corpus of *ICLE*
HECTOR	18-million word corpus of written English used by Moon (1998) in study of idiom
HKUST	Hong Kong University of Science and Technology learner corpus comprised of the texts of Chinese students of English (mainly Cantonese speakers) at the advanced high school level.

ICAME	International Computer Archive of Modern and Medieval English. ICAME is an international organization of linguists and information scientists working with English machine-readable texts. The aim of the organization is to collect and distribute information on English language material available for computer processing and on linguistic research completed or in progress on the material, to compile an archive of English text corpora in machine-readable form, and to make material available to research institutions. The archive mentioned in the name resides at at University of Bergen, Norway. This acts as a distribution centre for computerized English-language corpora and corpus-related software.
ICE	The International Corpus of English (ICE) began in 1990 with the primary aim of collecting material for comparative studies of English worldwide. Fifteen research teams around the world are preparing electronic corpora of their own national or regional variety of English. Each ICE corpus consists of one million words of spoken and written English produced after 1989. Available from: http://www.ucl.ac.uk/english-usage/ice/
ICLE	International Corpus of Learner English held at the Catholic University of Louvain
KFNgram	A program which generates lists of *n-grams* in text and HTML files develop by Fletcher (2004) [online]. Available from: http://www.kwicfinder.com/KWiCFinder.html
LINDSEI	Louvain International Database of Spoken English Interlanguage
LLC	*Longman Learners Corpus*
London Lund Corpus	The spoken part of the Survey of English Usage Corpus, computerized at the Survey of Spoken English, Lund University under the direction of J. Svartvik. It contains 500,000 words of spoken British English recorded from 1953 to 1987. Available from ICAME in several versions.
LOB	Lancaster Oslo Bergen corpus. One-million-word corpus of written British English from 1967 designed to mirror the *Brown* corpus
Locness	Louvain corpus of native speaker essays: designed as a control corpus for ICLE sub-corpora
MICASE	Michigan Corpus of Academic Spoken English. A collection of transcripts of academic speech events recorded at the University of Michigan. There are currently 152 transcripts (totalling 1,848,364 words) available at this site. [online]. Available from: http://micase.umdl.umich.edu/m/micase/
Porticle	Portuguese sub-corpus of *ICLE*
SPICLE	Spanish sub-corpus of *ICLE*
SWICLE	Swedish sub-corpus of *ICLE*

TOSCA-ICLE	Tools for Syntactic Corpus Analysis designed at University of Nijmegen
UCREL	University Centre for Computer Corpus Research on Language, University of Lancaster
USAS	The UCREL semantic analysis system is a software system for undertaking the automatic semantic analysis of text. Available from: http://www.comp.lancs.ac.uk/ucrel/usas/
Wmatrix	Wmatrix is a software tool for corpus analysis and comparison. It provides a web interface to the USAS and CLAWS corpus annotation tools, and standard corpus linguistic methodologies such as frequency lists and concordances. [online]. Available from: http://www.comp.lancs.ac.uk/ucrel/wmatrix/
Wordsmith Tools	WordSmith Tools is a suite of programs designed by Scott (1999) for text analysis and manipulation. WordList generates word lists from one or more texts by frequency and by alphabet. Concord displays a concordance for any given word or part of word. [online]. Available from: http://www.lexically.net/wordsmith/

Appendix 2

Suggested essay titles and learner profile used in the creation of the *Porticle* corpus

1. Crime does not pay.

2. The prison system is outdated. No civilised society should punish its criminals: it should rehabilitate them.

3. Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.

4. A man/woman's financial reward should be commensurate with their contribution to the society they live in.

5. The role of censorship in Western society.

6. Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.

7. All armies should consist entirely of professional soldiers: there is no value in a system of military service.

8. The Gulf War has shown us that it is still a great thing to fight for one's country.

9. Feminists have done more harm to the cause of women than good.

10. In his novel Animal Farm, George Orwell wrote "All men are equal: but some are more equal than others". How true is this today?

11. In the words of the old song "Money is the root of all evil".

12. Europe.

13. In the 19th century, Victor Hugo said : "How sad it is to think that nature is calling out but humanity refuses to pay heed. "Do you think it is still true nowadays ?

14. Some people say that in our modern world, dominated by science technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion ?

LEARNER PROFILE

Text code : (do not fill in)
Essay :

Title :
Approximate length required : -500 words **0** +500 words **0**
Conditions : timed **0** untimed **0**
Examination : yes **0** no **0**
Reference tools : yes **0** no **0**

What reference tools ?
Bilingual dictionary :
English monolingual dictionary :
Grammar :
Other(s) :

==========

Surname : **First names :**
Age : Male **0** Female **0**

Nationality :
Native language :
Father's mother tongue :
Mother's mother tongue :
Language(s) spoken at home : (if more than one, please give the average % use of each)

Education :
Primary school - medium of instruction :
Secondary school - medium of instruction :
Current studies :
Current year of study :
Institution :
Medium of instruction :
English only **0**
Other language(s) (specify) **0**
Both **0**

==========

Years of English at school :
Years of English at university :

Stay in an English-speaking country :
Where ?
When ? How long ?

Other foreign languages in decreasing order of proficiency :

I hereby give permission for my essay to be used for research purposes.
Date : ..... Signature : .....

Appendix 3

Five of the essays from the 5,000-word random sample from *Porticle* learner corpus discussed in Chapter 3.6

<ICLE-PT-ESE-0006.3>

Is definitely true that nowadays the entire world is dominated by science, technology and industrialisation but is also true that nobody can live without having a dream, something to fight for.

The new technologies and the success of science is vital for the safety of humankind, for the guarantee of new treatments to fight old diseases. Modern world needs the benefices that science, technology and industrialisation give to us.

Another example is the infinite ways of communication. Is quite amazing what we can do with a computer, we can talk with thousand of people at the same time!!!! We can see the world and the world beauties in a simple computer, a usual television,...

All that is true but without a dream, without something to fight for, the man could not have made all this. Imagination gives the world progress, gives hope and faith.

Sometimes, instead of using imagination to produce useful things, the humankind uses that to produce evil things. Weapons, wars, deaths are made of awful dreams, of nightmares, And that is what makes the world stop dreaming. Hope and faith are damage by those people that don't use imagination for built but to destroy.

Although all that, I think that the world needs to built a new world, a place for dreaming and imagination, a place where each of us can find a better day. Dreams are made of hope and desire, ingredients that made good dreams become true.

I have afraid that in a previous future the world becomes more and more selfish. Man don't think about children, don't think about what are they doing with so much advanced technology. They only want to earn to them a nice and well paid present, without imagine how life will be for the next generations. Our kids can not live in a world where bad dreams and desires come true, where persons don't use the heart. They can not live in a world without feelings. Modern world will give our children loneliness.

People use a computer, stay at home instead of going out with their families. Is fun and delicious seeing the real world, see other people, going to the beach, eat an ice-cream.

Making usual things that made life so pleasant. It's difficult to understand how technology, science and industrialisation can be so good and, at the same time can be so evil. It is an issue that we have to think about. We have to wake up and say: my kids have the right to live in a free world, a place where we can dream and imagine a good future, a better day. Hope and faith, that is what the world needs. Hope and faith in the science, that it can bring to us good news.

To finish I need to say that we still live in a good world. We have to be more and more careful, but we can not stop dreaming. I know people that have wonderful dreams, I have good dreams, so I have to make them come true. Let's all together built a place for dreaming and imagination.

<ICLE-PT-ESE-0008.3>

I don't agree with this affirmation. There will always be a place for dreams and dreamers and I think it is absolutely necessary to have imagination nowadays.

If we don't dream we don't live, in this world of wars and selfish people all that is left are dreams. We shall dream, but we also have to be realistic, we live in a cruel world with many injustices. We can't forget that we live in a real world, so it is not good to be always dreaming. There is a time for everything. We live in an industrialised world, full of technology, but imagination it is very important we shall explore it, because it is healthy. All our life we will use our imagination, but it is true that technology gives everything and for instance kids don't need to use their imagination, so teachers like us must help them to use their imagination they have to practise it.

Kids in school should be more obliged to use their imagination writing texts and poems. But they don't want to, they are lazy. About dreams they are very important in our lives, they give us strength to continue, life is very hard. With all the work we have sometimes we don't have the need to continue, so dreams give us some kind of strength, we want to fight for something. It is absolutely necessary to dream.

Technology and industrialisation don't replace dreams, they will always exist. I think that this modern world it is not very good for our children, it is very dangerous. It is a risk to

have kids. This society is full of drugs and alcohol, bad people and hypocrisy. So with all this things don't you think we are obliged to dream and use our imagination?
Imagination is primordial, we have to use it every day to make our lives less boring, interesting and pleasant. The world is being destructed by all this technology and industrialisation. They are polluting rivers and the air, that reduces the quality of life and many lung diseases appear.
There is other thing we can talk about that is the use of technology and industrialisation to make weapons. I can't understand how people who believe in God and have a religion can kill. They say it is in the name of God. How can it be? This is a bad use of technology. Thinking in this things we can stay sad and think that technology is bad, but that's not right people use their imagination badly. They think in killing and not producing anything. So how can we live without dreaming? How can we stay simply without using our imagination? It is impossible!
Science and industrialisation are good things that brought development, richness and a lot of brand new things to the humanity. We can conclude that it is impossible to live without dreaming and it is impossible to live without technology and industrialisation. So both things can exist at the same time, they are both very important to the world.

<ICLE-PT-ESE-0009.3>

One of the most important things to me in the world is being able to dream and imagine things that exist but are hard to come true.
What I love doing in my spear time is reading. Books that talk about wizards, witches, magic and so on. Harry Potter for example was one of the books that I adored reading. I don't think that just because science technology and industrialisation "dominate" the world that there's no space for dreaming and imagination. Every one should have dreams even if they don't come true it doesn't mean that we're losers. Dreams help us live our life to the fullest even if they don't come true.
Most people believe that dreaming and imagining things is a waste of time. People that think that way live their life in such a boring way. For me I think it would be a waste of time living on earth thinking only about things that we can get and have with out any

trouble. There are things that I will never be able to do or have but that doesn't mean that I can't dream or imagine it.

Science technology and industrialisation nowadays is fundamental but to me it's not enough. I need to dream every day. I need to feel that in my dreams I can do what ever I want, dream with who ever I want and have what ever I want. Science can't give this to me. Imagine children not being able to dream or imagine. What would their life be? I can't imagine my childhood without dreams, imagination. I use to love dreaming that I was a famous pop singer and I still do. Even if they didn't come true (and it didn't) it didn't stop me from dreaming.

The film "The Lord of the Rings" or even " Harry Potter" were films that made me imagine what it would be like to be a wizard or even an elf. These films wouldn't even exist if science technology wasn't so advanced but these films were based on dreams and imagination. So these films also exist because the directors used their imagination and dreams to make them. I doubt that these films were made by personal experiences. Elves don't exist, wizard don't exist (I think!). So how could they have invented these characters? Or even the actual story of the films?. There are so many things that people invent and don't exist in real life so: - How can we explain this? I think it's simple. We all have inside of us, hidden deep in side, a world of dreams, dreams that we want to come true but we know most of them don't but we don't quit on life. Life would be so boring if dreams didn't exist but without science technology and industrialisation we wouldn't survive but dreams and the ability to imagine are also important, especially for me. My life would have any meaning without being able to dream.

I love dreaming and I hope that every body feels the same way.

<ICLE-PT-ESE-0010.3>

To begin with, I think that it's not true because our imagination never ends. In spite of our society being considered too much industrialised my opinion is that we can always imagine and invent new things, new solutions. Human imagination is always working and studying about new formulas to help people. For example, thanks to our technology and science people's health is getting better. We take profit from science and technology not only in

medicine but also in many subjects. We can't stop dreaming just because our world is too much industrialised, we always have space and time to dream: The dreams change depending on what we are living presently. Today I have a dream, but tomorrow that dream can be true and after more dreams come and so on... We are always dreaming. I even believe that all these scientific advances that we are living through can make us have more dreams or different dreams. We could have dreamt forty years ago, that science would make it possible to happen; it is no longer a dream, however new dreams appear, as to improve our knowledge, with intention of making our lives better, in areas such as our health, our life expectation, our way of living in this world.

However science can really knock down some dreams...because sometimes people use imagination and science technology not to help people but to destroy them, and consequently, their dreams. Then I cannot think in the same way.

Unfortunately, some countries like U.S.A use their science technology to produce arms to fight aiming always to be a powerful nation. However, as we can see in 11th September all that technology doesn't prevent the taking of thousands of innocent lives. Why do people do that? The first thing I said when this happened was "My God! How could it happen in a country like that, with so much technology! They could avoid that!" After all in few minutes America could be destroyed and they could not do anything to avoid it! After all, sometimes technology does not help. In my opinion, this event, the 11th of September, was a step back in all our dreams. In a time that the dream of a fair and peaceful world was possible it happened, coming to destroy all my beliefs that it's possible to live in a world empty of bad people. But I think that we must keep fighting for one day this almost universal value to come true. After all this heavy attack just gave us more strength to fight for this dream that no science, or society development will erase.

Sometimes when I think of science, industrialisation and technology I get stressed in a moment and I just think of being in contact with nature, listening to the birds. I close my eyes and then I start dreaming of a better world, a better society, not always aiming for power to be " untouchable". No one, anything is untouchable.

To sum up, I think technology is helpful when we use it for a good purpose, and I do not believe that it limits our dream, on the contrary, I believe that science can make us have bigger dreams!!!

<ICLE-PT-ESE- 0012.3>

In my point of view men and women's financial reward is not commensurate with their contribution to the society they work in. Our society is a little bit upside down, because usually people who have worst jobs / occupations , in terms of hardworks, are the poorest ones and receive low payments. People that have what we use to call "clean " and " lazy " jobs are the ones who are better paid. In this order the rich citizens have good jobs and as they are better paid they are always increasing (in terms of social status), and poor citizens that don't have educational qualifications, can only get jobs that are very hard and receive low remuneration.

Most of the times, the hard work that some people have to support all day is not rewarded by the money they receive at the end of the month. Unfortunately, The jobs that are consider the hardest ones are by coincidence (I prefer to think like this!!!) the ones that are badly paid, like people that work in factories doing things that essential to the society (clothes, materials to build houses...) and the ones who clean streets. These jobs and the people who work on them are very helpfully to the society and should be better paid.

With this point of view I'm not excluding or trying to give more importance to some jobs than others, but society should respect and reward everyone by its work and value, according to the contribution their work has to the society. For example scientists that spend lots of money and time trying to find out any theory or invent that wont be useful to our society and at the end of the month they are very well paid.

While I was writing this essay I had an idea that could help people trying to find out jobs or occupations that are socially helpfully, for example people who participate and work in voluntary activities by helping the others who need help like firemen, policemen or helping homeless people shouldn't pay taxes and should be better paid. Maybe with this kind of rewards besides making justice with people that have humanity jobs , people could wake up to the society and notice a little bit to society problems by finding out other kind of jobs.

When we hear people saying : " All men are equal: but some are more equal than others ", they are right, because the world we live in is in a mess in what concerns justice, social status, financial rewards, helping people with difficulties...but what surprises me most is that everyone knows this but no one does nothing to change

it , maybe because people who have enough power to do it are the ones who would lost their positions with the eventual changes.

We are living in a vicious world that will never change for better, only for worst. Martin Luther King once said: " I hope one day everyone can live
in peace and equally in freedom..."

Appendix 4

Three of the essays from the 5,000-word random sample from the *Locness* corpus discussed in Chapter 3.6

<ICLE-US-MRQ-0019.1>

There are many dilemmas that humans must face every day. Some of them easily justified, and some are not. These dilemmas can range from a simple act of choosing which friends side to take in a disagreement, or something more serious, like a priest giving up his reconciliation oath and handing over a confidential confession of a murderer to the police. We make most of these decisions based on our values and beliefs. One's values and beliefs tend to stem from their religious background or the morals and ethics on which they were raised. It is these ethics that contribute to the majority of our decision making tactics. When forming a decision, based on ethics, we must be careful to determine which side is morally correct.

A dilemma in which morals have come up more often than not is the case for animal rights. Should these animals be tested for the protection of human lives, or more importantly, what gives us the right to test these animals? Now, in the Bible, it says that humans were to have <*>. What better of an authority figure than God, could give us this right. Animals were put on earth for humans to do with as they please, as long as we did not take advantage of this gift. I believe animal testing in most cases is trying to save people's lives, one way or the other. The decision must be made that a human's life is more important than an animal's life. Animal testing is a crucial part of our society and must go on, if human's want to continue as the superior species. An example of this, is a three year old, named Timmy Baker has a heart attack from a disease he possibly contracted from eating an undercooked hamburger. He is not responding to resuscitation efforts and his doctor, Dr Michael McManus, decides to give him the epinephrine drug. The dosage recommended by the medical association does not work, but Dr McManus has heard of the research done by Dr Traystman. In research done by Traystman, pigs are electronically subjected to heart attacks, resuscitated, and then given a dose of the drug epinephrine. This drug is used to help revive heart attack victims while simultaneously preventing brain damage from occurring. He also conducts tests determining whether additional doses of the

medication can help a suffering patient. Dr McManus injects additional doses of epinephrine into Timmy following the research done by Dr Traystman. In doing this Dr McManus saves Timmy's life and sends him home six weeks later, a perfectly healthy boy. Animal testing accounts for a substantial part of research today, both in commercial and medical testing. The number of animals involved reflects this point. Seventeen to twenty-two million animals are killed in research labs every year. There are many advantages to being able to experiment drugs on animals before actually administering them to humans. Do these advantages give humans the ethical right to cause pain and suffering on animals for their benefit? Is it ethical to sacrifice the animals so mankind can live in comfort? These questions are asked everyday in the minds of animal lovers, researchers, and animal right activists. They are the basis for the great controversy over animal experimentation. To discuss this issue, the idea of what is ethical and unethical needs to be discussed. For something to be ethical, the person it actively concerns must believe that it is morally correct. On the other hand, when something is unethical, the party activity involved does not agree morally with it. This leads to the point that animal rights activists try focus on.

<*> states Herbert Waddams, protester for animal rights. Animal rights activists believe that it is unethical to inflict pain on animals. They base this argument of experience. Personal human experience takes effect in that humans, as part of nature, try to avoid pain. We center our lives around many principles. One of which is to make our lives as pain less and comfortable as possible. Pain is considered to be evil and in inflicting it on animals, we are said to be evil. Every creature, human or not, has an interest in avoiding suffering. While animals my not be conscious of their interests, activists believe that they should still be given equal consideration. One particularly painful test is the Draize eye Irritancy Test, where chemicals are squirted into a rabbit's eye to see if it causes blindness. How can this be ethical if it is known that there is pain being implemented on the animals? Animal researchers believe that the benefits of animal testing outweigh the harm and pass the research off as ethical. They claim that the amount of information found by testing animals could be discovered by no other method. While Dr Traystman was experimenting with the epinephrine, he was under constant pressure from animal rights activists. Thank God for little Timmy that Traystman continued his experiments, for without those experiments Timmy may have been left with permanent brain damage. The testing may

have killed some pigs, but at least one little boy's life and perhaps many more will have been saved.

There are also examples of where a lack of animal testing comes back to haunt humans. The FDA prohibited the sale of thalidomide, a medicine aides in the healing off birth defects, in the U.S. because of insufficient animal testing. Had they tested this drug on more animals, perhaps many children born with deformations could be living happy, normal lives today.

Animal researchers argue that it is necessary for the testing of animals to progress to insure that products and methods are safe for humans. It is their belief that <*>. Animals are needed to teach medical students techniques in surgery, giving drugs, and taking blood. Dr Traystman argues that practicing on living organisms is essential. <*>. While animal rights activists are attacking the researcher and their centers for causing animals to suffer, the research is actually helping to eliminate human suffering.

There are some experiments on animals that could be eliminated. On animal experiment, labeled as evil and inhumane, included the testing of over one hundred and fifty dogs with electric currents. The dogs were strapped to hammocks and given approximately sixty-four painful shocks lasting five seconds each. While this seemed extremely sadistic, it helped with many human breakthroughs about depression. This depression does not really classify as human pain and discomfort. In this type of testing, the benefit to humans does not substantially outweigh the pain of the animals. This sector of animal testing could be avoided.

The recent outcry against animal testing has resulted in many changes in the living conditions of animals. This turnabout is an effort to make animal research appear to be more ethical to animal rights activists. Dr Raystman's piglets live in ideal living conditions and are more closely monitored than most of the patients in the hospitals. The pigs fill the room with oinks and grunts as though they are living on a farm. <*>, says Dr Traystman. Only these piglets won't ever return to the farm, they are loaded onto crates and sent upstairs to be "sacrificed" in the name of science.

Animal researches also argue that if research was to be banned, where would the line be drawn? Lions, pigs, dogs, and monkeys are all animals, but so are viruses and bacteria and other single cell organisms. These are all beings and have been used in research for years,

but would a ban on animal testing end this testing too? This could lead to enormous setbacks in science.

Animals are being killed in every part of the world today. They are killed for hunting, bullfighting, cock fighting, and other animal sports in the name of entertainment. These methods aren't under the same controversy as testing and testing is for the good of humans. While the Animal Rights coalition is aiming for the total elimination of the use of animals, it is impossible to do at this time. Right now there is simply no alternative to the testing, and research is too important to give up. Allen Bromey, White House assistant to the president for Science and Technology, said that he is worried about attempts to ban animal testing <*>. As said at the beginning of the paper, we must continue the testing in order to insure the safety of humans. After all, who's life is more sacred, the life of a three year old boy, or a pig?

<ICLE-US-MRQ-0021.1>

AIDS is spreading at an enormously fast rate among Americans. It is an especially large concern for the young adults of today. In fact, according to the Centers for Disease Control, <*>. The sheer size of this problem is what prompted Chancellor Joseph Fernandez to develop a plan to combat this horrible epidemic. Under Mr Fernandez's plan, all 120 public high schools in New York City would have staff volunteers hand out condoms along with a booklet explaining their use to any student that wanted them. Although this plan has received much criticism, the distribution of condoms to sexually active teens in high school is a step which needs to be taken in order to combat the spread of the disease called AIDS.

A program like the one suggested for New York City by Chancellor Joseph Fernandez is very new and different compared to anything proposed previously to help combat the AIDS epidemic. Because of this, it has been met with great opposition. One of the main objections to his proposal is that it encourages and condones sexual promiscuity. Many parents believe that condom distributionundercuts schools' unstated purpose of teaching values. Many religious leaders object to the plan on religious grounds to pre-marital sex and believe that the schools should steer away form promiscuous sexual conduct. Dr Irene Imellizzeri, a psychologist, <*>. In response to this Mr Fernandez replied <*>.

A growing body of research and the experience of AIDS educators suggests that condom distribution does in fact decrease sexual activity. These educators have found that condom availability, when combined with AIDS education, can delay and discourage casual sexual activity. Researchers at the John Hopkins Medical School conducted an AIDS education program in Baltimore that included condom distribution in a high school. The results they received consisted of <*>. Other studies in progress are finding similar results. Paul Epstein, the author of the article "Condoms In Schools: The Right Lesson", also found in his interviews with adolescents and college students that the presence of condoms makes them more wary of casual sex, more conscious of the epidemic and more serious about their own risk. The same article also states, <*>. This leads one to believe that if condoms were more readily available, students would be more apt to practice safer sex.

Another criticism of Mr Fernandez's plan is that schools should be teaching students about abstinence rather than freely giving them condoms. The only sure way to avoid infection is abstinence and this is the message that parents, educators and religious leaders want sent to teenagers. These people argue that there is evidence to prove that the goal of abstinence is realistic. In fact, <*>. However, sexual abstinence is not an attractive option for most students. Almost all existing sex and AIDS education classes stress chastity, yet half the nation's high school girls are sexually active; 16% have had four or more partners. According to Time, <*>. This shows that the idea of abstinence is a mere fantasy. School officials argued that students were having sex whether the city liked it or not and that the debate was really over saving lives. The distribution of condoms in schools would help protect these already sexually active teens. For the vast number of teenagers who are going to engage in sex no matter what they are taught in school, abstinence is going to do no good. These students need knowledge and skills to protect themselves and their partners. This means that explicit instruction in what types of sex are most risky and how to reduce risk by using condoms is necessary. Studies have been done to prove that such instruction works. According to an analysis of a 1988 national survey of adolescent male, <*>.

The final argument against condom distribution in schools is concerned with the dependability of condoms. Some oppose distributing condoms because they say it will mislead students into believing there is such a thing as "safe sex". As stated in Time, <*>. To counteract this argument the March 1989 Consumer Report found that only 1 in 65 condoms breaks during vaginal intercourse and only 1 in 105 during anal intercourse.

Epstein writes, <*> This evidence should be reason enough to make condoms readily available to sexually active teens in high schools. If condoms are our only means of protection against this horrible disease, than anyone who engages in sexual activities should have easy access to them. Having condoms readily available in high schools simply means that sexually active teens will better be able to protect themselves and their partners from AIDS.

As of now, there is no evidence available concerning how effective the distribution of condoms in schools is in decreasing the rate in which AIDS is spreading because the plan has only been in effect for such a short time. However, the refutals of arguments presented here should be reason enough to be supportive of the idea. As of now, condoms are the only means to prevent AIDS from spreading among sexually active people. Therefore, by distributing them in our high schools, students will be better able to protect themselves and their partners. Robert Rygor's words are fitting for those people who are against the idea. Mr Rygor, who has AIDS, states, <*>.

<ICLE-US-MRQ-0022.1>

Forty years ago, starting the day off in a public school often meant reciting the pledge of allegiance and a group prayer. Today, the American flag garners little attention and a public prayer contradicts constitutional law. But why does praying in public schools receive such a negative reputation? There is a strong movement in the United States to bring back prayer to the schoolhouse. The argument used to support the use of prayer in public schools, however, relies too heavily on shaky speculation, fallacies, and fanaticism for effectiveness. Those who oppose prayer in public schools focus on fairness and equality to everybody rendering their argument much more convincing and acceptable.

An easy claim to make, and one that frequently appears in arguments for the use of prayer in public school, states that <*>. This may seem like an encouraging thought but it nevertheless displays naïveté and wishful reasoning. A prayer alone cannot subdue major societal problems. Trying to get a troubled youth out of a gang for example, with a prayer in school compares to trying to extinguish a burning house with a thimble full of water. It is just not enough to be effective. A thorough religious awakening may help alleviate significant problems that plague youth but a daily prayer in school hardly qualifies as this.

Also, a side overlooked by supporters of prayer in public schools points out that <*>. This fact strengthens the argument for not allowing prayer in public schools. Supporters often point to the fact that teen pregnancy, violence, and drug abuse among youths is increasing since the ban of school prayer. These facts do not accurately show that prayer lead to moral virtue. Too many uncontrolled variables, like an increase in broken homes and violence on television, may distort results. During the time when prayer existed in public schools moral tarnishes such as segregation, lynchings, and the oppression of women blemished society. It appears as though America has made gains in moral awareness. In short, because no data shows that prayer actually causes a better educational or social environment there is no reason to allow for its inclusion into the public school.

A common strategy used by supporters of the movement to reinstate prayer into the public school classroom attempts to project the ban as a religious straitjacket. The ban, according to Republican state senator Mike Gunn, amounts to <*>. This view holds no validity. By blanketing all students in America as Christians, a large part of the religious population evaporates. Religious groups that do not accept or openly oppose the ideals of Christianity would face an unfair influence by prayer in public schools. A public prayer would in fact constrict the morals of any student who does not believe in God or may have an alternate view of the concept of God. The "Christian commitment" would only nullify legitimate values and beliefs that have equal right for respect.

Opponents of prayer in public schools adopt an attitude that carefully considers the individual beliefs of students. In order to allow all students to retain and establish personal conviction, <*>. A prayer on behalf of an entire student population, in essence, proclaims that one idea of a supreme being is correct or more important than another. Students who hold alternate beliefs would face apostasy in the classroom, sacrificing morals for education. Therefore, the only logical and fair answer would result in not allowing prayer to influence students in any way in the public school system.

In an attempt to get prayer back into the classroom, supporters claim

that praying allows <*>. However, allowing schoolchildren to believe in God does not justify subjecting others to the same belief with a public prayer. A prayer projected to a student population discounts the rights of those who earnestly do not accept the concept of

a supernatural being. Forbidding a prayer in a public school does not discourage any belief system nor does it endorse any particular religious idea. The right of religious freedom that a prayer would supposedly perpetuate in actuality infringes upon it in its most fundamental definition.

The concept of God is not discourage in a public school. Similarly, praying on school grounds is not a crime. The laws of the United States <*>. Students do not have to leave their religion and beliefs at the front door of the schoolhouse. If a student has the desire to pray at any moment during the school day he or she should not encounter any determent. Only when students (or faculty) force any students to join in the prayer does it become a problem. The act of trying to force an unwilling person to digest the religious philosophy of another may lead to an uncomfortable educational setting that would hinder learning and social growth.

Prayer in public schools may continue to gain more popularity in the United States. Popularity, though, does not constitute righteousness. Subjecting students who have non-conventional beliefs to a discriminating prayer causes more harm than benefit. The phrase "public" schools signifies that members of society should have the right to use the facilities without facing a constant threat to their morals or beliefs. Because America's vast religious diversity and the lack of apparent good resulting from discouraging religion diversity, public schools need to remain free of public prayer.

Appendix 5

Prefabs in *Porticle* and *Locness* (normalized to a basis per 100,000)

No.	*Porticle*		*Locness*	
1.	the world	156	because of	76
2.	I think	146	such as	68
3.	that is	122	have to	67
4.	kind of	89	death penalty	61
5.	in order	87	a chance to	56
6.	for example	77	a good example	55
7.	a lot	70	the family	50
8.	most of	62	public schools	46
9.	of course	58	according to	43
10.	I believe	58	the world	40
11.	our society	54	a great deal	38
12.	such as	51	a little more	37
13.	many people	46	a large number	37
14.	real world	44	in public	36
15.	no longer	43	a means of	32
16.	in fact	43	nuclear power	31
17.	due to	43	capital punishment	31
18.	the fact	40	a member of	30
19.	modern world	40	a number of	29
20.	for instance	40	the media	28
21.	at least	39	high school	28
22.	I know	34	the past	27
23.	all men	33	wild card	26
24.	some people	32	in fact	26
25.	instead of	32	a playoff system	26
26.	science technology	31	a salary cap	25
27.	even if	31	a result of	25
28.	a result of	27	a little	25
29.	mass media	26	college football	24
30.	a solution to	26	our society	23
31.	this way	25	look at	23
32.	as well	25	water pollution	22
33.	a sort of	25	the welfare	22
34.	a little	25	the homeless	22
35.	and all	24	playoff system	22
36.	a source of	24	on television	22
37.	a world where	23	Florida State	22
38.	a world of	23	even though	22
39.	a way to	23	a social problem	22
40.	a way that	23	stay home	21
41.	a way of	23	I think	21
42.	so that	21	a way of	21
43.	at all	21	a time when	21
44.	all know that	21	a symbol of	21
45.	all human beings	21	a strong argument	21

46.	a few	21	a great deal of	21
47.	all the time	20	South Carolina	20
48.	all over	20	corporal punishment	20
49.	all of us	20	biological parents	20
50.	all of them	20	affirmative action	20
51.	sort of	19	a way to	20
52.	this essay	18	a way that	20
53.	so much	17	this argument	19
54.	science and technology	17	no one	19
55.	other people	17	no longer	19
56.	human rights	17	had to	19
57.	and i think	17	drinking age	19
58.	and i believe	17	at least	19
59.	too much	16	advocates for censorship	19
60.	this world	16	I feel	18
61.	many things	16	gun control	17
62.	just because	16	an attempt to	17
63.	each other	16	adoptive parents	17
64.	most people	15	adolescent suicide	17
65.	we know	14	supreme court	16
66.	real life	14	an increase in	16
67.	one day	14	an important role	16
68.	the past	13	an example of	16
69.	like this	13	a large number of	16
70.	in mind	13	of course	15
71.	human being	13	appear to be	15
72.	as well as	13	animal rights activists	15
73.	as long as	13	model approach	14
74.	as i said	13	a lot of money	14
75.	as far as	13	this point	13
76.	as an example	13	this issue	13
77.	as a whole	13	premarital sex	13
78.	as a conclusion	13	Notre Dame	13
79.	around the world	13	I know	13
80.	a bit	13	I believe	13
81.	poor people	12	human life	13
82.	modern society	12	battle flag	13
83.	be ready to	12	at all	13
84.	be prepared to	12	as long as	13
85.	at this point	12	as it is	13
86.	at the university	12	as a whole	13
87.	at the end	12	as a symbol	13
88.	at the beginning	12	as a result	13
89.	long time	11	as a deterrent	13
90.	best way	11	welfare system	12
91.	bear in mind	11	think about	12
92.	be seen as	11	strong argument	12
93.	this question	10	opposite sex	12

94.	this issue	10	dealing with	12
95.	little bit	10	black students	12
96.	I mean	10	based on the	12
97.	but i think	10	at this time	12
98.	above all	10	at the time	12
99.	willing to	9	at the end	12
100.	things like	9	at the beginning	12
101.	crime does pay	9	as well as	12
102.	commit a crime	9	as much as	12
103.	coal and steel	9	a way of life	12
104.	after all	9	a stay home wife	12
105.	to continue	8	white students	11
106.	to conclude	8	most people	11
107.	this means	8	in addition	11
108.	my essay	8	human being	11
109.	in general	8	exchange theory	11
110.	I guess	8	due to the fact that	11
111.	have the chance	8	black and white	11
112.	have access to	8	at home	11
113.	get in touch	8	as if	11
114.	first of all	8	as a symbol of	11
115.	find a job	8	as a result of	11
116.	i would say	7	welfare reform	10
117.	i think that	7	the television	10
118.	i strongly believe	7	the system	10
119.	i really think	7	the environment	10
120.	i know that	7	looked at	10
121.	i hope that	7	it is important	10
122.	i guess that	7	gender identity	10
123.	i don't think	7	first amendment	10
124.	i don't know	7	contribution to society	10
125.	i don't believe	7	conservative religious group	10
126.	i don't agree	7	conflict theory	10
127.	i believe that	7	college football season	10
128.	i believe in	7	civil liberties group	10
129.	i am sure	7	church and state	10
130.	i am aware	7	at the same time	10
131.	i also think	7	an issue	10
132.	i agree with	7	young adults	9
133.	i agree that	7	very strong	9
134.	have the opportunity	7	in favor of capital punishment	9
135.	my opinion is	6	distinct gender identities	9
136.	more than ever	6	conflict theory	9
137.	more or less	6	assisted suicide	9
138.	more and more	6	too much	8
139.	men and women	6	sex education	8
140.	it seems that	6	seems to	8

141.	it is true	6	once again	8
142.	it is necessary	6	for that matter	8
143.	it is impossible	6	for many years	8
144.	it is important	6	football players	8
145.	it is good	6	first of all	8
146.	in what concerns	6	crime does not pay	8
147.	in the past	6	brings up	8
148.	in the morning	6	binge drinking	8
149.	in the future	6	animal testing	8
150.	in the end	6	young people	7
151.	in that time	6	West Virginia	7
152.	in terms of	6	training programs	7
153.	in spite of	6	today's society	7
154.	in some cases	6	they think	7
155.	in real life	6	surrogate mothers	7
156.	in other words	6	social problems	7
157.	in order to	6	shield law	7
158.	in order for	6	sexually transmitted	7
159.	in my opinion	6	sexually active	7
160.	in most cases	6	school prayers	7
161.	in front of	6	school integration	7
162.	in a way	6	salary caps	7
163.	if we consider	6	role models	7
164.	the language of	5	rational choice	7
165.	the lack of	5	only to	7
166.	the kind of	5	league championship	7
167.	the industrial revolution	5	leads to	7
168.	the importance of	5	insurance companies	7
169.	the idea that	5	in today's society	7
170.	the idea of	5	in the past	7
171.	the human rights	5	in the morning	7
172.	the human being	5	in the future	7
173.	the history of	5	in the end	7
174.	the help of	5	in the country	7
175.	the hands of	5	in the classroom	7
176.	the feminist movement	5	in terms of	7
177.	the fact that	5	in support of	7
178.	the fact is	5	in some cases	7
179.	the evolution of	5	in recent years	7
180.	the european union	5	in reality	7
181.	the end of	5	in public schools	7
182.	the educational system	5	in our society	7
183.	the development of	5	in order to	7
184.	the degree of	5	in order for	7
185.	the declaration of	5	in jail	7
186.	the death penalty	5	in business	7
187.	the criticism of	5	in addition to	7
188.	the contact with	5	in a way that	7

189.	the consequences of	5	in a way	7
190.	the concept of	5	i think that	7
191.	the chance to	5	i know that	7
192.	the cause of	5	i feel that	7
193.	the case of	5	i believe that	7
194.	the black people	5	his or her	7
195.	the best way	5	hippocratic oath	7
196.	the beginning of	5	he or she	7
197.	the access to	5	have sex	7
198.	the ability to	5	go through	7
199.	that is why	5	general public	7
200.	technology and science	5	football players	7
201.	technology and industrialization	5	financial support	7
202.	technology and industrialisation	5	every day	7
203.	take care of	5	ended up	7
204.	take advantage of	5	divorce rate	7
205.	supposed to be	5	dependent on	7
206.	sooner or later	5	deals with	7
207.	some people say	5	confederate flag	7
208.	so i think	5	college students	7
209.	since the beginning	5	civil liberties	7
210.	quality of life	5	child care	7
211.	portuguese and english	5	Championship Series	7
212.	points of view	5	American people	7
213.	point of view	5	aids patients	7
214.	place for dreaming	5	very well	6
215.	period of time	5	very important	6
216.	on the contrary	5	this theory	6
217.	on one hand	5	this question	6
218.	women and men	4	this paper	6
219.	with this statement	4	this evidence	6
220.	with the statement	4	this claim	6
221.	what i mean	4	terminally ill	6
222.	we see that	4	talk shows	6
223.	we don't know	4	take care	6
224.	way of living	4	stands for	6
225.	way of life	4	spent on	6
226.	to sum up	4	sexual relationships	6
227.	theory and practice	4	set up	6
228.	the whole world	4	residential halls	6
229.	the united states	4	relate to	6
230.	the united kingdom	4	regardless of	6
231.	the truth is	4	professional footbal	6
232.	the sense of	4	organized crime	6
233.	the same thing	4	nuclear power plants	6
234.	the root of	4	no matter what	6

235.	the role of	4	men and women	6
236.	the right to	4	many people feel	6
237.	the rest of	4	male and female	6
238.	the responsibility of	4	looking at	6
239.	the real world	4	living in	6
240.	the real life	4	little or no	6
241.	the quality of	4	listen to	6
242.	the purpose of	4	it is true	6
243.	the problem of	4	it is possible	6
244.	the prison system	4	it is obvious	6
245.	the power to	4	it is hard	6
246.	the power of	4	it is clear that	6
247.	the origin of	4	it is clear	6
248.	the opportunity to	4	in the long run	6
249.	the only thing	4	in the case of	6
250.	the number of	4	in power	6
251.	the needs of	4	in mind	6
252.	the need to	4	in an attempt to	6
253.	the most important	4	I agree	6
254.	the more you	4	human beings	6
255.	the more i	4	honour code	6
256.	the modern world	4	home wife	6
257.	the modern society	4	for years	6
258.	the middle of	4	for instance	6
259.	the mass media	4	family unit	6
260.	the majority of	4	drug testing	6
261.	the love of	4	down to	6
262.	you can see	3	black community	6
263.	when it comes	3	work place	5
264.	whatever they want	3	turned into	5
265.	we can conclude	3	took place	5
266.	ways of thinking	3	they say	5
267.	way of thinking	3	these practices	5
268.	way of seeing	3	these arguments	5
269.	to sum up my position	3	take care of	5
270.	to make money	3	side effects	5
271.	to go on	3	sexually transmitted diseases	5
272.	to fight against	3	sex and violence	5
273.	to each other	3	school and college	5
274.	to commit a crime	3	rules and regulations	5
275.	to be prepared	3	professional football players	5
276.	they don't know	3	over the years	5
277.	the quality of life	3	out of control	5
278.	the most important things	3	on the other hand	5
279.	the love of money	3	most of the time	5
280.	the justice system	3	morally wrong	5
281.	the human nature	3	looks at	5

282.	the french revolution	3	looked upon	5
283.	the entire world	3	it is important to	5
284.	tend to be	3	it is important for	5
285.	teachers and students	3	it is hard to	5
286.	taking care of	3	in turn	5
287.	take care of the children	3	grown up	5
288.	students and teachers	3	get away	5
289.	so i think that	3	gender roles	5
290.	right to vote	3	for life	5
291.	right to live	3	feminist movement	5
292.	right to education	3	end up	5
293.	pay attention to	3	each year	5
294.	our future life	3	commit suicide	5
295.	my personal opinion	3	college campuses	5
296.	my personal experience	3	billion dollars	5
297.	my own experience	3	big business	5
298.	most of the time	3	as for	5
299.	most of the people	3	agree with	5
300.	most of all	3	after all	5
301.	lord of the rings	3	you know	4
302.	large amounts of	3	would seem	4
303.	lack of money	3	world series	4
304.	it must be	3	women athletes	4
305.	it means that	3	welfare program	4
306.	it is a fact that	3	way of life	4
307.	in the classroom	3	wall street journal	4
308.	in that sense	3	violence on television	4
309.	in other countries	3	violence and profanity	4
310.	in my view	3	too many	4
311.	in many ways	3	to have sex	4
312.	in any way	3	this essay	4
313.	i mean that	3	the world series	4
314.	i just think	3	the work place	4
315.	i intend to	3	the work force	4
316.	i do not feel	3	the winner of	4
317.	i can't imagine	3	the wild card	4
318.	i am sure that	3	the white students	4
319.	i am prepared	3	the welfare system	4
320.	i agree with this	3	the welfare program	4
321.	for this reason	3	the value of	4
322.	for many people	3	the use of	4
323.	for all this	3	the university of	4
324.	fight for their rights	3	the united states government	4
325.	declaration of human rights	3	the united states	4
326.	by the way	3	the types of	4
327.	but the fact is that	3	the teaching of	4
328.	but the fact is	3	the surrogate mother	4
329.	but on the other hand	3	the supreme court	4

330.	better living conditions	3	the supporters of	4
331.	as we know	3	the state house	4
332.	as human beings	3	the spread of	4
333.	as a teacher	3	the same as	4
334.	and many other things	3	the rules and regulations	4
335.	and i think that	3	the root of all evil	4
336.	and i believe that	3	the root of	4
337.	and at the same time	3	the risk of	4
338.	an idea of	3	the rights of	4
339.	an excuse for	3	the right to	4
340.	an example of this	3	the right of	4
341.	a waste of	3	the rewards of	4
342.	a very important role	3	the results of	4
343.	a variety of	3	the rest of	4
344.	a type of	3	the rational choice theory	4
345.	a sense of	3	the rate of	4
346.	a part of	3	the question of whether	4
347.	a new life	3	the question of	4
348.	a necessary evil	3	the public schools	4
349.	a near future	3	the public school	4
350.	a lot of things	3	the proponents of	4
351.	a good solution	3	the problems of	4
352.	a foreign language	3	the problem that	4
353.	a bit of	3	the problem of	4
354.	a big problem	3	the presence of	4
355.	what i want to say	2	the practice of	4
356.	what i mean is that	2	the possibility of	4
357.	what i mean is	2	the point that	4
358.	way of seeing the world	2	the people who	4
359.	universal declaration of human rights	2	the people that	4
360.	turn out to be	2	the people of	4
361.	to know what is happening	2	the opposite sex	4
362.	to have a job	2	the opponents to	4
363.	to get in touch with	2	the opponents of	4
364.	to fight for their rights	2	the only way to	4
365.	to fight for one's country	2	the number of	4
366.	the second world war	2	the new york times	4
367.	the rich and the poor	2	the need for	4
368.	the rest of the world	2	the national title	4
369.	the origin of the world	2	the national championship	4
370.	the opium of the people	2	the national champion	4
371.	the non-civilised world	2	the minds of	4
372.	the needs of the market	2	the members of	4
373.	the ministry of education	2	the majority of	4
374.	the lord of the rings	2	the love of	4
375.	the first thing to do	2	the loss of	4
376.	the enlargement to the east	2	the lives of	4

377.	the end of the month	2	the life of	4
378.	the end of the course	2	the league championship	4
379.	the development of science	2	the last day of	4
380.	the declaration of human rights	2	the last day	4
381.	the consequences of their actions	2	the lack of	4
382.	taking care of the children	2	the issue of	4
383.	such a large number of	2	the idea that	4
384.	rest of the world	2	the idea of	4
385.	opium of the people	2	the history of	4
386.	of very little value	2	the hippocratic oath	4
387.	money is a necessary evil	2	the heart of	4
388.	it would be better	2	the general public	4
389.	it seems to me that	2	the first amendment	4
390.	it seems to me	2	the feminist movement	4
391.	it seems logical to say	2	the feminism movement	4
392.	it is well known	2	the family unit	4
393.	it is impossible to	2	the fact that	4
394.	it is good for	2	the fact is	4
395.	it is easier to	2	the exchange theory	4
396.	it is clear that	2	the end of	4
397.	it is also important to	2	the effects of	4
398.	it is also important	2	the drinking age	4
399.	it doesn't mean that	2	the divorce rate	4
400.	in what concerns to	2	the death penalty	4
401.	in such a way that	2	the cotton bowl	4
402.	in such a way	2	the cost of	4
403.	in order for us to	2	the conservative religious group	4
404.	in order for them to	2	the consequences of	4
405.	i wish you good luck	2	the conflict theory	4
406.	i wish to conclude by	2	the confederate flag	4
407.	i think it is important	2	the concept of	4
408.	i like the idea of	2	the civil liberties group	4
409.	i have to say that	2	the center of	4
410.	i have to believe that	2	the case of	4
411.	i have to admit that	2	the case for	4
412.	i don't think that	2	the case against	4
413.	i don't know what	2	the black community	4
414.	i do not think that	2	the biological parents	4
415.	i do not think	2	the best way to	4
416.	i do not believe that	2	the best way	4
417.	i do not believe	2	the benefits of	4
418.	i do not agree	2	the beginning of	4
419.	i do believe that	2	the battle flag	4
420.	i completely agree with this	2	the author of	4
421.	i am not saying that	2	the amount of	4

422.	i also know that	2	the american people	4
423.	i agree with this statement	2	the american league	4
424.	i agree with the statement	2	the american family	4
425.	how wonderful it would be	2	the age of	4
426.	have nothing to do with	2	the advocates for	4
427.	from all over the world	2	the ability to	4
428.	fight for a better world	2	television programs	4
429.	differences between women and men	2	symbolic interactionists	4
430.	by the other hand	2	supporting evidence	4
431.	as a way of	2	support themselves	4
432.	as a source of power	2	suggests that	4
433.	as a source of	2	suggest that	4
434.	as a result of	2	strip joint	4
435.	as a matter of fact	2	sexual activity	4
436.	as a matter of	2	religious group	4
437.	all men are born equal	2	painless death	4
438.	a waste of time	2	one day	4
439.	a source of power	2	motor oil	4
440.	a short period of time	2	most importantly	4
441.	a long time ago	2	moral values	4
442.	a great deal of	2	mercy killing	4
443.	a good example of	2	medical profession	4
444.	a better way of life	2	may seem	4
445.	a better way of	2	main reason	4
446.		4129	main claim	4
447.			Los Angeles	4
448.			legal system	4
449.			in prison	4
450.			in need	4
451.			in conclusion	4
452.			I understand	4
453.			I thought	4
454.			honor codes	4
455.			hard work	4
456.			happen to	4
457.			grow up	4
458.			good idea	4
459.			gender identities	4
460.			foster care	4
461.			fiscal year	4
462.			field advantage	4
463.			few years	4
464.			equal pay	4
465.			effective argument	4
466.			each day	4
467.			drug legalization	4
468.			double standard	4

469.	discriminated against	4
470.	dependent children	4
471.	cotton bowl	4
472.	concerned with	4
473.	comes from	4
474.	come up	4
475.	case studies	4
476.	brought about	4
477.	at times	4
478.	as follows	4
479.	all men	4
480.	wife and mother	3
481.	when it comes to	3
483.	this is why	3
484.	the wall street journal	3
485.	the values and beliefs	3
486.	the value of human life	3
487.	the validity of	3
488.	the unborn child	3
489.	the type of	3
490.	the situation of	3
491.	the shield law	3
492.	the second amendment	3
493.	the role of	3
494.	the return of	3
495.	the rest of their lives	3
496.	the reason that	3
497.	the real world	3
498.	the problem of homelessness	3
499.	the practice of euthanasia	3
500.	the people of america	3
501.	the people in power	3
502.	the misuse of	3
503.	the medical profession	3
504.	the meaning of	3
505.	the level of	3
506.	the legalization of	3
507.	the league championship series	3
508.	the laws of	3
509.	the killing of	3
510.	the importance of	3
511.	the illinois supreme court	3
512.	the illinois supreme	3
513.	the human genome	3
514.	the honor code	3
515.	the belief that	3

516.	strengths and weaknesses	3
517.	sexually active teens	3
518.	sentenced to death	3
519.	people think that	3
520.	people of america	3
521.	on their own	3
522.	on the welfare	3
523.	on the streets	3
524.	marijuana and other drugs	3
525.	it is true that	3
526.	it is obvious that	3
527.	in these instances	3
528.	in other words	3
529.	in my opinion	3
530.	have a child	3
531.	freedom of speech	3
532.	end of the regular season	3
533.	day to day	3
534.	comprehensive sex education	3
535.	both sides of the issue	3
536.	besides the fact that	3
537.	arguments in favor of	3
538.	appears to be	3
539.	an example of this is	3
540.	all aspects of	3
541.	a right to	3
542.	a group of	3
543.	a good idea	3
544.	a form of	3
545.	you don't have to	2
546.	with this in mind	2
547.	to strengthen their argument	2
548.	to have a child	2
549.	the right to vote	2
550.	the right to die	2
551.	the right to choose	2
552.	the pursuit of happiness	2
553.	the love of money	2
554.	the long term effects	2
555.	the fact is that	2
556.	the distribution of condoms	2
557.	the confederate battle flag	2
558.	the civil rights movement	2
559.	survival of the fittest	2
560.	it is up to	2
561.	in the name of	2
562.	in the eyes of	2

563.	a wife and mother	2
564.		4667
565.		
566.		
567.		
568.		
569.		

Appendix 6 Prefabs in *Bawe* and *CofE* normalized to a basis per 100,000

No.	*Bawe*		*CofE*	
1.	a little	38	a few	52
2.	a lot	36	a handful	48
3.	according to	27	a little	45
4.	as cited	21	a lot	44
5.	appears to	21	above all	29
6.	appeared to	21	according to	28
7.	appear to	21	against terrorism	27
8.	as well	20	after all	27
9.	at first	19	agree strongly	26
10.	carry out	17	almost certainly	24
11.	carried out	17	and all	23
12.	automatic processing	17	american policy	23
13.	at least	17	and more	20
14.	cited in	16	and therefore	19
15.	circadian rhythm	16	and then	19
16.	cultural differences	15	as for	18
17.	could be	15	appear to	18
18.	connectionist systems	15	apart from	18
19.	conditional sentences	15	and yet	18
20.	communicative competence	15	at least	17
21.	cognitive processes	15	at last	17
22.	cognitive development	15	at home	17
23.	due to	14	at all	17
24.	in common	12	as well	17
25.	in all	12	as one	17
26.	in addition	12	as if	17
27.	ill health	12	a good society	17
28.	i wish	12	a free press	17
29.	i think	12	a few years	17
30.	i found	12	a debate about	17
31.	housing quality	12	aware of	16
32.	health inequalities	12	at that	16
33.	future research	12	at present	16
34.	freudian theory	12	at one	16
35.	for instance	12	a lingua franca	16
36.	for example	12	a group of	16
37.	new information	11	because of	15
38.	nervous system	11	a member of	15
39.	in line	11	a means of	15
40.	in heaven	11	a lot of	15
41.	in general	11	a long way	15
42.	in fact	11	a long time	15
43.	in defeat	11	but for	14
44.	in contrast	11	british universities	14

45.	other factors	10	british students	14
46.	optic array	10	big business	14
47.	opposed to	10	a second resolution	14
48.	on stage	10	a result of	14
49.	on earth	10	a war against	13
50.	nineteenth century	10	a time when	13
51.	new york	10	a set of	13
52.	new skills	10	consequences of	12
53.	new labour	10	confidence in	12
54.	predictive validity	9	concern about	12
55.	points out	9	committed to	12
56.	pointed out	9	commitment to	12
57.	point to	9	coming home	12
58.	point out	9	comes to	12
59.	physical characteristics	9	combined with	12
60.	to conclude	8	cold war	12
61.	the rest	8	climate change	12
62.	the present	8	civil servants	12
63.	simultaneous tasks	8	civil liberties	12
64.	self-esteem	8	chief executive	12
65.	respond to	8	case for	12
66.	resource management	8	case against	12
67.	repressive masculist	8	carry out	12
68.	rem sleep	8	capable of	12
69.	real-life	8	and the rest	12
70.	raphe nuclei	8	exchange rate	11
71.	public relations	8	even when	11
72.	much of	8	even though	11
73.	less than	8	economic development	11
74.	language school	8	disagree strongly	11
75.	in turn	8	devoted to	11
76.	formal logic	8	development aid	11
77.	far more	8	developing countries	11
78.	comes from	8	democratic appointees	11
79.	circadian rhythms	8	deal with	11
80.	brain stem	8	cultural studies	11
81.	as for	8	at the time	11
82.	storage capacity	7	at the same time	11
83.	role in	7	at that time	11
84.	risk of	7	at a time	11
85.	reticular formation	7	as well as	11
86.	responsible for	7	as a whole	11
87.	response to	7	as a result	11
88.	physical art	7	around the world	11
89.	over time	7	are likely to	11
90.	neoclassic plays	7	gulf war	10

91.	look at	7	greenhouse gas	10
92.	language acquisition	7	global warming	10
93.	interactionist approach	7	foundation hospitals	10
94.	indentured servants	7	foreign policy	10
95.	i believe	7	for years	10
96.	good test	7	for instance	10
97.	french neoclassic	7	expert witnesses	10
98.	fourteenth century	7	expert witness	10
99.	find out	7	business and management	10
100.	english teachers	7	be able to	10
101.	central capacity	7	at the university	10
102.	at once	7	weapons of mass destruction	9
103.	at all	7	in contrast	9
104.	as if	7	in business	9
105.	and all	7	in all	9
106.	working memory	6	i think	9
107.	work on	6	i suspect	9
108.	with practice	6	i know	9
109.	wish clause	6	i hope	9
110.	willing to	6	i believe	9
111.	westward expansion	6	human rights	9
112.	west midlands	6	human beings	9
113.	very important	6	food supply chain	9
114.	verb tense	6	by no means	9
115.	verb forms	6	not even	8
116.	unattended message	6	no one	8
117.	unable to	6	no longer	8
118.	typical of	6	needs to	8
119.	trying to	6	need not	8
120.	try to	6	much more	8
121.	truth tables	6	most people	8
122.	truth conditions	6	mass destruction	8
123.	tried to	6	many years	8
124.	too much	6	lingua franca	8
125.	this means	6	last year	8
126.	this idea	6	last week	8
127.	this essay	6	last month	8
128.	this concept	6	kind of	8
129.	this book	6	involved in	8
130.	this argument	6	instead of	8
131.	this approach	6	in truth	8
132.	third conditional	6	in return	8
133.	think about	6	in public	8
134.	these studies	6	in prison	8
135.	these findings	6	in place	8

136.	these factors	6	in peace	8
137.	the young	6	in particular	8
138.	the world	6	in part	8
139.	the west	6	in front	8
140.	the theatre	6	in fact	8
141.	the classroom	6	in a way	8
142.	the church	6	i believe that	8
143.	that is	6	hard to see	8
144.	teaching context	6	greenhouse gas intensity	8
145.	teacher training	6	freedom of information	8
146.	target language	6	for the moment	8
147.	symbolic interactionists	6	at the university of	8
148.	symbolic interactionist	6	quality assurance	7
149.	such as	6	public opinion	7
150.	strategic competence	6	prime minister	7
151.	storage capacity	6	president bush	7
152.	stage design	6	past decade	7
153.	staff development	6	not enough	7
154.	spoken language	6	no such	7
155.	southern states	6	no more	7
156.	sort of	6	news media	7
157.	sociolinguistic competence	6	new york	7
158.	social welfare	6	need to	7
159.	social psychology	6	muslim world	7
160.	social meaning	6	member states	7
161.	social interaction	6	many years	7
162.	social housing	6	many people	7
163.	social factors	6	make up	7
164.	social context	6	let alone	7
165.	social class	6	last year	7
166.	so that	6	last week	7
167.	smoke alarms	6	last month	7
168.	smoke alarm	6	lack of	7
169.	sleep deprivation	6	in the world	7
170.	short-term memory	6	in the way	7
171.	semantic conception	6	in the past	7
172.	seen as	6	in the middle	7
173.	seems to	6	in the field	7
174.	seemed to	6	in terms of	7
175.	seem to	6	in recent years	7
176.	second language	6	in public services	7
177.	russian theatre	6	in other words	7
178.	roman catholic	6	in order to	7
179.	role of	6	in higher education	7
180.	in other words	6	in front of	7
181.	in order to	6	in favour of	7

182.	in most cases	6	for the first time	7
183.	in line with	6	the rest of the world	6
184.	i believe that	6	The New York Times	6
185.	human resource management	6	stem cells	6
186.	great deal of	6	sort of	6
187.	goes on to	6	so many	6
188.	first of all	6	rather than	6
189.	by means of	6	public services	6
190.	be used to	6	public interest	6
191.	at this stage	6	public debate	6
192.	at this point	6	press freedom	6
193.	at the end	6	poor countries	6
194.	at the beginning	6	policy advice	6
195.	at a time	6	pointed out	6
196.	as well as	6	pay for	6
197.	as opposed to	6	pattern recognition	6
198.	as much as	6	other people	6
199.	as long as	6	ordinary people	6
200.	as it is	6	or more	6
201.	as cited in	6	or less	6
202.	as an illustration	6	opinion polls	6
203.	as a whole	6	one day	6
204.	artist and visionary	6	on the other hand	6
205.	appears to be	6	on the basis of	6
206.	an increase in	6	it may be	6
207.	an example of	6	it is possible	6
208.	an attempt to	6	it can be	6
209.	a way of	6	in a way that	6
210.	a sort of	6	at the heart of	6
211.	a reduction in	6	at a time when	6
212.	a piece of	6	world bank	5
213.	a period of	6	white paper	5
214.	a part of	6	too many	5
215.	a number of	6	this means	5
216.	a means of	6	the way in which	5
217.	a lot of	6	the arab world	5
218.	a lack of	6	the absence of	5
219.	a great deal	6	such as	5
220.	a form of	6	south africa	5
221.	whether or not	5	so much	5
222.	ways of thinking	5	so long	5
223.	this suggests that	5	so far	5
224.	this is because	5	single currency	5
225.	the west midlands fire service	5	set up	5
226.	the west midlands	5	seems to be	5
227.	the way in which	5	seems to	5

228.	the view that	5	seemed to	5
229.	the view of	5	seem to be	5
230.	the value of	5	seem to	5
231.	the validity of	5	see it as	5
232.	the use of	5	security council	5
233.	the united states	5	secretary of state	5
234.	the type of	5	science studies	5
235.	the toeic test	5	science and technology	5
236.	the threat of	5	saudi arabia	5
237.	the target language	5	resulted in	5
238.	the symbolic interactionist approach	5	result in	5
239.	the symbolic interactionist	5	rely on	5
240.	the success of	5	regardless of	5
241.	the style of	5	one of them	5
242.	the southern states	5	on the way	5
243.	the sources of	5	on the contrary	5
244.	the semantic conception of truth	5	on both sides	5
245.	the second language	5	I do not believe	5
246.	the same as	5	years ago	4
247.	the rules of	5	world trade	4
248.	the role of	5	who knows	4
249.	the reticular formation	5	white house	4
250.	the result of	5	western europe	4
251.	the rest of	5	we know	4
252.	the reliability and validity	5	war on terrorism	4
253.	the purpose of	5	war against terrorism	4
254.	the production of	5	war against	4
255.	the process of	5	up to	4
256.	the problem of	5	too much	4
257.	the presence of	5	together with	4
258.	the power of	5	time to	4
259.	the performance of	5	this month	4
260.	the people of	5	this kind of	4
261.	the part of	5	this issue	4
262.	the number of	5	these days	4
263.	the notion of	5	the world of	4
264.	the nineteenth century	5	the world bank	4
265.	the neurochemistry of sleep	5	the work of	4
266.	the neurochemistry of	5	the whole of	4
267.	the nature of	5	the white paper	4
268.	the most significant	5	the white house	4
269.	the most important	5	the west	4
270.	the moscow art theatre	5	the web	4
271.	the meaning of	5	the way to	4

272.	the majority of	5	the war on	4
273.	the loss of	5	the war against	4
274.	the levels of	5	the war	4
275.	the language school	5	the wake of	4
276.	the lack of	5	the view that	4
277.	the key to	5	the use of	4
278.	the introduction of	5	the university of	4
279.	the influence of	5	the united states	4
280.	the importance of	5	the united nations	4
281.	the idea that	5	the united kingdom	4
282.	the idea of	5	the tobacco industry	4
283.	the growth of	5	the threat of	4
284.	the fourteenth century	5	the third reich	4
285.	the foundations of logic	5	the terms of	4
286.	the foundations of	5	the support of	4
287.	the first part	5	the success of	4
288.	the first language	5	the story of	4
289.	the faerie queene	5	the story of	4
290.	the fact that	5	the state of	4
291.	the extent to which	5	the start of	4
292.	the existence of	5	the soviet union	4
293.	the end of	5	the single currency	4
294.	the eighteenth century	5	the security council	4
295.	the effects of	5	the same way	4
296.	the effect of	5	the same thing	4
297.	the distinction between	5	the same as	4
298.	the difficulty of	5	the sake of	4
299.	the difference between	5	the role of	4
300.	the development of	5	the right to	4
301.	the description of	5	the result of	4
302.	the creation of	5	the result is	4
303.	the course of	5	the rest of us	4
304.	the context of	5	the rest of	4
305.	the content of	5	the release of	4
306.	the concept of	5	the question of	4
307.	the components of	5	the question is	4
308.	the classical view	5	the quality of	4
309.	the choice of	5	the purpose of	4
310.	the characteristics of	5	the public interest	4
311.	the category of	5	the process of	4
312.	the case of	5	the problems of	4
313.	the black death	5	the prime minister	4
314.	the beginning of	5	the people who	4
315.	the basis of	5	the people of	4
316.	the amount of	5	the past decade	4
317.	the ability to	5	the only way to	4

318.	that is to say	5	the nuclear issue	4
319.	tarski's semantic conception of truth	5	the need to	4
320.	symbolic interactionist approach	5	the need for	4
321.	some sort of	5	the nature of	4
322.	sleep as an active state	5	the name of	4
323.	seems to be	5	the muslim world	4
324.	seemed to be	5	the most important	4
325.	seem to be	5	the mind of	4
326.	restoration and french neoclassic plays	5	the middle east	4
327.	reliability and validity	5	the majority of	4
328.	reading and writing	5	the lessons of	4
329.	point of view	5	the lack of	4
330.	one can therefore say that	5	the labour party	4
331.	one can therefore say	5	the issue of	4
332.	on the whole	5	the interests of	4
333.	on the other hand	5	the indian government	4
334.	on the basis of	5	the importance of	4
335.	on the basis	5	the idea that	4
336.	it was found that	5	the idea of	4
337.	it seems that	5	the heart of	4
338.	it is necessary	5	the hands of	4
339.	it is impossible	5	the growth of	4
340.	it is important	5	the garda síochána	4
341.	it has been suggested that	5	the future of	4
342.	it has been suggested	5	the form of	4
343.	it can be argued	5	the food supply chain	4
344.	is thought to be	5	the food supply	4
345.	in the main	5	the food industry	4
346.	in the language school	5	the food chain	4
347.	in the development of	5	the first world war	4
348.	in the context of	5	the first world	4
349.	in the case of	5	the first time	4
350.	in terms of	5	the fate of	4
351.	in relation to	5	the fact that	4
352.	in an attempt to	5	the face of	4
353.	for the purpose of	5	the expense of	4
354.	components of communicative competence	5	the existence of	4
355.	at the same time	5	the european union	4
356.	at the end of	5	the european parliament	4
357.	at the beginning of	5	the english language	4
358.	as cited in baddeley	5	the end of	4
359.	a repressive masculist epic	5	the emergence of	4

360.	a piece of art	5	the edinburgh festival	4
361.	with reference to	4	the doha agreement	4
362.	to think about	4	the decision on	4
363.	to look at	4	the consequences of	4
364.	to be able to	4	the cold war	4
365.	this means that	4	the climate of	4
366.	the work of	4	the case of	4
367.	the war of	4	the bush administration	4
368.	the topic of	4	the british people	4
369.	the structure of	4	the british government	4
370.	the stroop effect	4	the behaviour of	4
371.	the strength of	4	the beginning of	4
372.	the state of	4	the basis of	4
373.	the stages of	4	the appointment of	4
374.	the social meaning	4	the ability to	4
375.	the social context	4	that is why	4
376.	the size of	4	remains to be seen	4
377.	the similarity of	4	over the course of	4
378.	the reliability of	4	nuclear arms race	4
379.	the range of	4	nothing to do with	4
380.	the portrayal of	4	in this case	4
381.	the population of	4	in the wake of	4
382.	the people of the west	4	in the name of	4
383.	the opportunity to	4	in the mind	4
384.	the necessity of	4	in the form of	4
385.	the knowledge of	4	in the face of	4
386.	the ideal of	4	in the course of	4
387.	the history of	4	in spite of	4
388.	the functions of	4	in my opinion	4
389.	the function of	4	i don't think	4
390.	the form of	4	i do not believe that	4
391.	the expansion of	4	I do not believe that	4
392.	the example of	4	head of state	4
393.	the diagnosis of	4	for the sake of	4
394.	the development of the self	4	English as a lingua franca	4
395.	the desire to	4	come to terms with	4
396.	the descriptions of	4	both sides of	4
397.	the components of communicative competence	4	being able to	4
398.	the basal forebrain	4	be prepared to	4
399.	the art of	4	at the expense of	4
400.	the appropriateness of	4	as good as	4
401.	the age of	4	as a result of	4
402.	tarski's material adequacy	4	as a lingua franca	4
403.	take into account	4	and so on	4
404.	second language acquisition	4	an attack on	4

405.	roman catholic church	4	all the more	4
406.	people of the west midlands	4	a variety of	4
407.	it is possible	4	a threat to	4
408.	it is necessary to	4	a number of	4
409.	it is impossible to	4	a list of	4
410.	it is important to	4	a language for	4
411.	input and retrieval of information	4	a handful of	4
412.	in this case	4	a couple of	4
413.	in favour of	4	world trade organization	3
414.	in common with	4	world trade center	3
415.	from here to maternity	4	while at the same time	3
416.	foreign language teaching	4	when it comes to	3
417.	bottom-up and top-down	4	we all know	3
418.	believed to be	4	use of english	3
419.	became known as	4	use of drugs	3
420.	be taken into consideration	4	up to now	3
421.	be regarded as	4	top of everest	3
422.	at this time	4	to work with	3
423.	at the time	4	to one side	3
424.	as soon as	4	to let go	3
425.	as an example	4	to find ways	3
426.	as a consequence	4	to engage in	3
427.	as a conclusion	4	to determine whether	3
428.	an important factor	4	to demonstrate how	3
429.	according to piaget	4	to bring about	3
430.	a variety of	4	to be true	3
431.	a time when	4	to be part	3
432.	a state of	4	to be effective	3
433.	a sign of	4	timothy garton ash	3
434.	a sense of	4	the young generation	3
435.	a process of	4	the world trade organization	3
436.	a picture of	4	the whole system	3
437.	a number of factors	4	the vietnam war	3
438.	a combination of	4	the use of english	3
439.	a circadian rhythm	4	the two countries	3
440.	a child's experience	4	the trouble with	3
441.	a chance to	4	the trick is	3
442.	what is more	3	the time to	3
443.	top-down and bottom-up	3	the test of	3
444.	to the extent that	3	the terrorist threat	3
445.	to take part in	3	the spanish basque	3
446.	to some extent	3	the size of	3
447.	to seek employment	3	the simple reason that	3
448.	to a certain extent	3	the simple reason	3

449.	theory of truth	3	the significance of	3
450.	the world of	3	the short term	3
451.	the west indies	3	the shape of	3
452.	the ways of	3	the second world war	3
453.	the usage of	3	the second world	3
454.	the types of	3	the scottish parliament	3
455.	the treaty of	3	the restoration of	3
456.	the town square	3	the reputation of	3
457.	the time of	3	the relevance of	3
458.	the third conditional	3	the release of policy	3
459.	the theory that	3	the relationship between	3
460.	the theory of	3	the quality assurance	3
461.	the test word	3	the prospect of	3
462.	the test takers	3	the product of	3
463.	the t-convention	3	the principle of	3
464.	the survival of	3	the prince of wales	3
465.	the stamp act	3	the prince of	3
466.	the stage-designer	3	the preserve of	3
467.	the stage design	3	the presence of	3
468.	the speed of	3	the practice of	3
469.	the special faculty	3	the outside world	3
470.	the spatial element	3	the other end	3
471.	the spaced practice	3	the opportunity to	3
472.	the social factors	3	the nuclear arms race	3
473.	the second one	3	the nuclear arms	3
474.	the second conditional	3	the northern hemisphere	3
475.	the same way	3	the northern alliance	3
476.	the same age	3	the next time	3
477.	the right hand	3	the needs of	3
478.	the release of	3	the near future	3
479.	the reduction of	3	the most part	3
480.	the recency part	3	the most dangerous	3
481.	the reactions of	3	the long run	3
482.	the proclamation line	3	the list of	3
483.	the problems of	3	the legitimacy of	3
484.	the pineal gland	3	the la times	3
485.	the person who	3	the key to	3
486.	the perception of	3	the journal of	3
487.	the pattern of	3	the iraqi people	3
488.	the past form	3	the introduction of	3
489.	the outbreak of	3	the international community	3
490.	the onset of	3	the ideology of	3
491.	the northern states	3	the house of	3
492.	the nervous system	3	the history of	3
493.	the need for	3	the hereditary peers	3

494.	the mythical arthur	3	the heart of europe	3
495.	the myth of	3	the goal of	3
496.	the most significant thing	3	the general public	3
497.	the most important thing	3	the first to	3
498.	the most basic	3	the fact is that	3
499.	the moral hero	3	the fact is	3
500.	the modern reader	3	the emphasis on	3
501.	the middle ages	3	the diversity of	3
502.	the massed practice	3	the difference between	3
503.	the maiden's presence	3	the developing world	3
504.	the magic fire	3	the decline of	3
505.	the lower classes	3	the decision taken	3
506.	the love of	3	the debate about	3
507.	the local community	3	the culture of	3
508.	the line of	3	the conventions of	3
509.	the level of	3	the context of	3
510.	the layout of	3	the conduct of	3
511.	the last thing	3	the concept of	3
512.	the labouring poor	3	the clinton administration	3
513.	the isolated forebrain	3	the climate of	3
514.	the interaction of	3	the centre of	3
515.	the impact of	3	the case for	3
516.	the immodest damsel	3	the best chance	3
517.	the help of	3	the beijing talks	3
518.	the gaps in	3	the battle for	3
519.	the formation of	3	the bankers vision	3
520.	the first time	3	the argument that	3
521.	the first place	3	the arab-israeli conflict	3
522.	the first half	3	the amount of	3
523.	the fight of	3	the algerian war	3
524.	the female reader	3	the alcohol industry	3
525.	the feelings of	3	the agenda of	3
526.	the extent that	3	the age of	3
527.	the expression of	3	the aftermath of	3
528.	the examination of	3	the afghan people	3
529.	the evaluation of	3	tens of thousands	3
530.	the emergence of	3	tend to be	3
531.	the effectiveness of	3	stop the war	3
532.	the economy of	3	statistical pattern recognition	3
533.	the earthly garden	3	speed of response	3
534.	the dissimilarity between	3	so long as	3
535.	the definition of	3	so far as	3
536.	the current rules	3	side by side	3
537.	the critical period	3	seem to have	3
538.	the consumption of	3	security council	3

			resolutions	
539.	the conclusion that	3	sections of the	3
540.	the combination of	3	second-order implications	3
541.	the character of	3	second world war	3
542.	the central executive	3	says the un	3
543.	the cause of	3	research and development	3
544.	the british government	3	republican and democratic	3
545.	the black report	3	release of policy	3
546.	the bayley scales	3	purposes and values	3
547.	the basis for	3	professor of law	3
548.	the basic conditionals	3	prisoners of conscience	3
549.	the authoritarian personality	3	prince of wales	3
550.	the audio-lingual approach	3	peace and security	3
551.	the aim of	3	over the years	3
552.	the actions of	3	over and over again	3
553.	teaching and learning	3	over and over	3
554.	target language use	3	our use of	3
555.	taking into account	3	on top of	3
556.	take place in	3	on the subject	3
557.	take part in	3	on the other side	3
558.	state pedagogical university	3	on the nuclear	3
559.	so as to	3	on the left	3
560.	sleep-wake cycle	3	on the ground	3
561.	secondary visual cortex	3	on the bench	3
562.	said to be	3	on his own	3
563.	rules of inference	3	on either side	3
564.	right hand side	3	nuclear and missile	3
565.	real-life listening	3	nothing more than	3
566.	plus or minus	3	not to mention	3
567.	piaget and vygotsky	3	not to be	3
568.	perception and evaluation	3	not the first	3
569.	patterns of consumption	3	none of this	3
570.	neuropsychological evidence	3	no such thing	3
571.	nature and nurture	3	new kind of	3
572.	most of all	3	much more likely	3
573.	members of staff	3	most of us	3
574.	member of society	3	more or less	3
575.	long-term memory	3	more and more	3
576.	journal of verbal learning	3	millions of people	3
577.	journal of experimental psychology	3	millions of dollars	3
578.	it seems to	3	many of those	3
579.	it seems reasonable	3	many of them	3
580.	it may be	3	manufacturers and retailers	3
581.	it can be argued that	3	less important than	3
582.	in view of	3	left and right	3

583.	in this work	3	languages in europe	3
584.	in this way	3	language for communication	3
585.	in this respect	3	knowledge of the past	3
586.	in the mind	3	know how to	3
587.	in the memory	3	join the euro	3
588.	in the first place	3	it is vital	3
589.	in return for	3	it is to be hoped that	3
590.	in more specific terms	3	it is time	3
591.	in any way	3	it is possible that	3
592.	impey and underhill	3	it is likely	3
593.	i don't know	3	it is impossible	3
594.	i do not know	3	it is important	3
595.	greenaway and harding	3	it impossible for	3
596.	good or bad	3	it could be	3
597.	for the first time	3	issue of the	3
598.	fixed and absolute	3	israelis and palestinians	3
599.	eysenck and keane	3	is likely to be	3
600.	english language teacher	3	in this regard	3
601.	electrical stimulation of	3	in the united kingdom	3
602.	effects on health	3	in the short term	3
603.	duty over love	3	in the sense	3
604.	duties and responsibilities	3	in the same way that	3
605.	complete simultaneous tasks	3	in the same way	3
606.	communicative listening skills	3	in the possession of	3
607.	communicative language teaching	3	in the number of	3
608.	christian world-picture	3	in the near future	3
609.	cause and effect	3	in the muslim world	3
610.	bottom-up skills	3	in the mind of	3
611.	attitude toward errors	3	in the long run	3
612.	as noted earlier	3	in the history of	3
613.	as far as	3	in the case of	3
614.	as effectively as	3	in the aftermath of	3
615.	as early as	3	in no way	3
616.	as a means	3	in danger of	3
617.	as a hero	3	in an article	3
618.	are able to	3	in a way that is	3
619.	appraisal and training	3	i suspect that	3
620.	and so on	3	i know that	3
621.	and so forth	3	higher education lecturers	3
622.	an understanding of	3	heart of europe	3
623.	an insight into	3	from the start	3
624.	an indication that	3	for too long	3
625.	an important role	3	for those who belive that	3
626.	an important part	3	for the simple reason that	3

627.	an examination of	3	for the simple reason	3
628.	an active process	3	for the most part	3
629.	amounts of information	3	for many years	3
630.	amount of time	3	for a long time	3
631.	amount of resources	3	european central bank	3
632.	a way that	3	enlargement of the european union	3
633.	a time of	3	economic and social development	3
634.	a system of	3	dead or alive	3
635.	a source of	3	conflicts of interest	3
636.	a sleep state	3	code of practice	3
637.	a set of	3	came of age	3
638.	a moral hero	3	at the other end	3
639.	a matter of	3	at the beginning	3
640.	a major role in	3	as part of	3
641.	a major role	3	as far as	3
642.	a long time	3	an expert witness	3
643.	a large number of	3	an end to	3
644.	a large number	3	an effort to	3
645.	a kind of	3	an average of	3
646.	a higher level of	3	an army of	3
647.	a higher level	3	able to see	3
648.	a good test	3	a year ago	3
649.	a fear of	3	a world that	3
650.	a critical period	3	a world of	3
652.	with such devastating effect	2	a single currency	3
653.	we should state that	2	a right to	3
654.	to the extent of	2	a referendum on	3
655.	to take into account	2	a recipe for	3
656.	to seek employment as	2	a range of	3
657.	to provide opportunities for	2	a quarter of	3
658.	to prepare themselves for	2	a polarised world	3
659.	to play a role in	2	a plan to	3
660.	to play a role	2	a philosophical attitude	3
661.	to play a key role	2	a part of	3
662.	to its best advantage	2	a new kind of	3
663.	to improve the situation	2	a man who	3
664.	to hold a conversation	2	a language for communication	3
665.	to get to know	2	a kind of	3
666.	to develop new skills	2	a good job	3
667.	to develop a test	2	a few months	3
668.	to complete the maze	2	a degree of	3
669.	to build a theory	2	a chance to	3
670.	to ascend to heaven	2	a chance of	3

671.	to answer this question	2	a century ago	3
672.	to a great extent	2	would be well advised to	2
673.	this supports the theory that	2	what i really hate is	2
674.	this is not the case	2	what i really hate	2
675.	there is no doubt that	2	were more likely to	2
676.	there is no doubt	2	we need to ensure that	2
677.	there is little evidence	2	turns out to be	2
678.	theories of communicative competence	2	turn out to have	2
679.	theories of cognitive development	2	turn out to be	2
680.	their way of life	2	to the point of	2
681.	the well-being of	2	to the effect that	2
682.	the visuo-spatial sketch	2	to teach the class	2
683.	the use of space	2	to release policy advice	2
684.	the treaty of paris	2	to reach the top	2
685.	the tasks being combined	2	to make use of	2
686.	the synthesis of serotonin	2	to live up to	2
687.	the symptoms of depression	2	to let go of	2
688.	the style of acting	2	to join the euro	2
689.	the students i taught	2	to go to war	2
690.	the structural features of	2	to convince the public	2
691.	the strength of feeling	2	to change the world	2
692.	the special faculty of	2	to bring about change	2
693.	the spaced practice group	2	to be reminded of	2
694.	the social psychology of	2	to be commended for	2
695.	the social and cultural	2	to be able to	2
696.	the sleep-waking cycle	2	to answer that question	2
697.	the sleep-wake cycle	2	through the looking-glass	2
698.	the simon-binet test	2	threat to the planet	2
699.	the sequencing of listening	2	this does not mean	2
700.	the second language classroom	2	the younger-dryas interval	2
701.	the second language acquisition	2	the whole of the web	2
702.	the rules of language	2	the whole concept of	2
703.	the rules of inference	2	the west as a whole	2
704.	the roman catholic church	2	the war on terrorism	2
705.	the role it plays	2	the war in iraq	2
706	the right side of	2	the war against terrorism	2
707.	the right parietal lobe	2	the very moment when	2
708.	the reduction of prejudice	2	the vast majority of	2
709.	the recency part of	2	the use of force	2
710.	the quality of teaching	2	the university of chicago	2
711.	the purpose of sleep	2	the trick is to	2
712.	the professional status of	2	the talks in beijing	2
713.	the processes involved in	2	the stability and growth	2

714.	the process of enclosure	2	the spanish basque country	2
715.	the previous conservative administration	2	the south african experience	2
716.	the potential threat of	2	the shape of things	2
717.	the physical presence of	2	the same as saying that	2
718.	the physical hardships of	2	the same as saying	2
719.	the physical characteristics of	2	the rule of law	2
720.	the past form of	2	the role of universities	2
721.	the other way around	2	the right to freedom	2
722.	the orientation of lines	2	the rich economies of	2
723.	the only cause of	2	the refugee studies programme	2
724.	the notion of arthur	2	the reduction of poverty	2
725.	the next two years	2	the real exchange rate	2
726.	the native english teachers	2	the quality assurance agency	2
727.	the myth of arthur	2	the problem is that	2
728.	the most well known	2	the population at large	2
729.	the modes of writing	2	the political neutrality of	2
730.	the massed practice trials	2	the past two years	2
731.	the major cause of	2	the past few years	2
732.	the main reason for	2	the other end of	2
733.	the line of kings	2	the other concerned parties	2
734.	the left side of	2	the north korean problem	2
735.	the left hand side	2	the moral high ground	2
736.	the last stanzas of	2	the minister for justice	2
737.	the lack of money	2	the majority of whom	2
738.	the initial cognitive representations	2	the last six years	2
739.	the idea of man	2	the interests of people	2
740.	the idea of hierarchy	2	the indonesian armed forces	2
741.	the human way of	2	the hong kong government	2
742.	the human processing system	2	the higher education system	2
743.	the help of others	2	the higher education funding	2
744.	the foundations of semantics	2	the greatest security challenges	2
745.	the explanation for this	2	the great train robbery	2
746.	the enemies of christianity	2	the future shape of	2
747.	the effects of similarity	2	the fact remains that	2
748.	the early stages of	2	the department of health	2
749.	the earlier part of	2	the current generation of	2
750.	the drug parachlorophenylalanine pcpa	2	the culture of violence	2

751.	the definition of intelligence	2	the cry goes up	2
752.	the dangers of fire	2	the conflicts of interest	2
753.	the current rules of	2	the common agricultural policy	2
754.	the concept of constraint	2	the centre of gravity	2
755.	the cocktail party phenomenon	2	the best chance of	2
756.	the christian world-picture	2	the basis of ability	2
757.	the choice of language	2	the arab-muslim world	2
758.	the black report's findings	2	the anti-gay amendment	2
759.	the basic types of	2	the advance of english	2
760.	the authoritarian personality theory	2	somewhere in the world	2
761.	the appraisal and training	2	security of the world	2
762.	the aims and objectives	2	problems caused by alcohol	2
763.	the active nature of	2	part of the solution	2
764.	the activation-synthesis hypothesis	2	ownership of the media	2
765.	the acoustic similarity effect	2	our way of life	2
766.	some of these days	2	on the side of	2
767.	social psychology of prejudice	2	on the one hand	2
768.	social and cultural factors	2	on the issue of	2
769.	sheep and cattle disease	2	on the basis of ability	2
770.	rules of language use	2	nuclear and missile programs	2
771.	right hand side of	2	no such thing as	2
772.	rich and the poor	2	no choice but to	2
773.	responsibility and career advancement	2	new york times	2
774.	research in this area	2	neutrality of civil servants	2
775.	research has shown that	2	membership of the eurozone	2
776.	reliability and validity of	2	members of the garda	2
777.	relevant to their culture	2	means to an end	2
778.	regulation of circadian rhythms	2	long way to go	2
779.	quarterly journal of experimental psychology	2	life of the planet	2
780.	professional and personal growth	2	learning for learning's sake	2
781.	plus or minus two	2	it's worth noting that	2
782.	play an important role in	2	it would be unfair	2
783.	play an important role	2	it was no surprise	2
784.	play an active role	2	it remains to be seen	2
785.	play a major role in	2	it might have been	2
786.	play a major role	3	it may not be	2
787.	performance of the tasks	2	it matters a lot	2

788.	perform as effectively as	2	it is worth remembering	2
789.	parts of the brain	2	it is vital that	2
790.	object of courtly love	2	it is not surprising that	2
791.	new directions in research	2	it is not surprising	2
792.	nature of the land	2	it is likely that	2
793.	nature of physical art	2	it is impossible not	2
794.	moments in the play	2	it is important that	2
795.	modes of writing course	2	it is hardly surprising that	2
796.	models of abnormal behaviour	2	it is hardly surprising	2
797.	model of selective attention	2	it is hard to see	2
798.	metaphor of the pearl	2	it is a debate	2
799.	medulla and spinal chord	2	is a case in point	2
800.	material influences on health	2	influx of commercial funds	2
801.	long- term memory	2	in the world of	2
802.	located in the brain	2	in the words of	2
803.	it seems unlikely that	2	in the way of	2
804.	it seems to be	2	in the southern hemisphere	2
805.	it seems reasonable to conclude	2	in the right direction	2
806.	it must have seemed	2	in the past decade	2
807.	it is very easy to	2	in the opposite direction	2
808.	it is very easy	2	in the northern hemisphere	2
809.	it is thought to be	2	in the middle of	2
810.	it is thought that	2	in the hands of	2
811.	it is quite clear	2	in the first world	2
812.	it is often the case	2	in the first place	2
813.	it is not surprising that	2	in the european union	2
814.	it is not surprising	2	in the context of	2
815.	it is not possible	2	in the campaign against	2
816.	it is likely that	2	in the business of	2
817.	it is important that	2	in the arab world	2
818.	it is hard to	2	in the age of	2
819.	it is evident that	2	in such a world	2
820.	it is easier to	2	in front of the telly	2
821.	it is desirable to	2	in an open letter to	2
822.	it is asserted in	2	in a world that	2
823.	it has been shown that	2	i don't think so	2
824.	it has been found that	2	i do believe that	2
825.	it has also been suggested	2	have it both ways	2
826.	investigation of the market	2	has nothing to do with	2
827.	introduction by donald roy	2	genuine freedom of information	2
828.	internal to the organism	2	freedom of the press	2
829.	interact with each other	2	for thousands of years	2
830.	intact secondary visual cortex	2	end of the beginning	2
831.	individuals with mental	2	elimination of nuclear	2

	disorders		weapons	
832.	increase in the population	2	declaration of human rights	2
833.	increase in demand for	2	change for the better	2
834.	in their existing roles	2	but the fact remains that	2
835.	in the target language	2	but the fact remains	2
836.	in the style of	2	but it is worth remembering	2
837.	in the southern states	2	but at the same time	2
838.	in the second language	2	bureau for lesser used languages	2
839.	in the same way	2	both sides of the atlantic	2
840.	in the same form	2	borne out by events	2
841.	in the same article	2	being a member of	2
842.	in the reticular formation	2	be well advised to	2
843.	in the real world	2	be the first to	2
844.	in the production of	2	be in full agreement	2
845.	in the power of	2	back on the agenda	2
846.	in the pattern of	2	at the white house	2
847.	in the opening stanzas	2	at the time of	2
848.	in the nineteenth century	2	at the cutting edge	2
849.	in the importance of	2	at the core of	2
850.	in the history of	2	at the beginning of	2
851.	in the hands of	2	as the product of	2
852.	in the growth of	2	as soon as possible	2
853.	in the form of	2	as he put it	2
854.	in the following part	2	as a means of	2
855.	in the first section	2	as a last resort	2
856.	in the first part	2	arguments in favour of	2
857.	in the early stages	2	anything to do with	2
858.	in the correct order	2	answer to all problems	2
859.	in the area of	2	another common feature of	2
860.	in such a way that	2	and in doing so	2
861.	in such a way	2	an open letter to	2
862.	in order to answer this	2	an integral part of	2
863.	in its own right	2	an equitable language policy	2
864.	in eysenck and keane	2	an edited version of	2
865.	in english language teaching	2	an all-republican panel	2
866.	in a sort of	2	american and british troops	2
867.	in a number of ways	2	all over the world	2
868.	in a number of	2	aid and know-how	2
869.	in a different way	2	academic freedom and autonomy	2
870.	important for cognitive development	2	about the nature of	2

871.	idea of staff development	2	about a quarter of	2
872.	i would say that	2	abortion and capital punishment	2
873.	human way of perception	2	a world in which	2
874.	his material adequacy condition	2	a wide range of	2
875.	from time to time	2	a two-state solution	2
876.	french and indian war	2	a threat to life	2
877.	foundations of formal logic	2	a steady fall in	2
878.	for a long time	2	a small group of	2
879.	english language teaching context	2	a sad day for	2
880.	english language teacher education	2	a pre-emptive strike	2
881.	end of the war	2	a piece of history	2
882.	due to the fact that	2	a more realistic view	2
883.	diagnosis of mental disorders	2	a long way to go	2
884.	development of the personality	2	a long way to	2
885.	come to the rescue	2	a long way from	2
886.	christian point of view	2	a leading role in	2
887.	british journal of psychology	2	a hundred years ago	2
888.	both of the books	2	a home-grown network	2
889.	binet and wechsler tests	2	a group of institutions	2
890.	belarusian socio-cultural context	2	a first-rate intelligence	2
891.	belarusian language and literature	2	a few years ago	2
892.	awareness of the process	2	a couple of years	2
893.	at the top of	2	a child at birth	2
894.	at the heart of	2	a case in which	2
895.	as the presence of	2	a case in point	2
896.	as the moral hero	2		4447
897.	as the ideal knight	2		
898.	as the idea of	2		
899.	as the basis for	2		
900.	as effectively as possible	2		
901.	as cited in eysenck	2		
902.	as an example of	2		
903.	as a way of	2		
904.	as a source of	2		
905.	as a process of	2		
906.	as a means of	2		
907.	as a form of	2		
908.	art of the theatre	2		
909.	appropriate to the context	2		
910.	an intrinsic part of	2		

911.	an increased number of	2		
912.	an important role in	2		
913.	an important factor in	2		
914.	an evolving set of	2		
915.	an essential part of	2		
916.	an active state of	2		
917.	america and the west	2		
918.	address terms in korean	2		
919.	acting on the environment	2		
920.	acoustic coding in stm	2		
921.	acceptable definition of truth	2		
922.	ability to acquire language	2		
923.	a year or two ago	2		
924.	a year or two	2		
925.	a substantial amount of	2		
926.	a strong loyalist contingent	2		
927.	a staff development policy	2		
928.	a sociology of childbirth	2		
929.	a social psychological phenomenon	2		
930.	a significant effect on	2		
931.	a set of rules	2		
932.	a redrawing of boundaries	2		
933.	a normal rectangular room	2		
934.	a new home for	2		
935.	a lot of time	2		
936.	a large criticism of	2		
937.	a far more complex character	2		
938.	a direction of treatment	2		
939.	a course of treatment	2		
940.	a certain amount of	2		
941.	a cause of development	2		
942.		4124		
943.				
944.				
945.				
946.				
947.				
948.				
949.				
950.				

Appendix 7 Formulae used in the research

1. Formula for Log Likelihood

Log likelihood is calculated by constructing a contingency table as follows:

Note that the value 'c' corresponds to the number of words in corpus one, and 'd' corresponds to the number of words in corpus two (N values). The values 'a' and 'b' are called the observed values (O), whereas we need to calculate the expected values (E) according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

In our case N1 = c, and N2 = d. So, for this word, E1 = c*(a+b) / (c+d) and E2 = d*(a+b) / (c+d). The calculation for the expected values takes account of the size of the two corpora, so we do not need to normalize the figures before applying the formula. We can then calculate the log-likelihood value according to this formula:

$$-2\ln \lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

This equates to calculating log-likelihood G2 as follows: G2 = 2*((a*ln (a/E1)) + (b*ln (b/E2)))

The higher the G2 value, the more significant is the difference between two frequency scores. For these tables, a G2 of 3.8 or higher is significant at the level of $p < 0.05$ and a G2 of 6.6 or higher is significant at $p < 0.01$

Adapted from the log-likelihood wizard page at the University Centre for Computer Corpus Research on Language (UCREL): Available from: http://ucrel.lancs.ac.uk/llwizard.html Accessed 13 April 2005

2. Formula for Z-test

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\overline{pq}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$\overline{p} = \frac{X_1 + X_2}{n_1 + n_2} \qquad \hat{p}_1 = X_1 / n_1$$

$$\overline{q} = 1 - \overline{p} \qquad \hat{p}_2 = X_2 / n_2$$

Formula used to carry out the *z* test for comparing two proportions (the critical value is 1.645 at p= 0.05).