INTERLANGUAGE LEXICOLOGY OF ARAB STUDENTS OF ENGLISH A COMPUTER LEARNER CORPUS-BASED APPROACH

by

MOUSA ABDELGHANI AL-BTOOSH

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2004

To my parents, wife and daughter with love

ACKNOWLEDGEMENTS

This dissertation owes its existence to many people whose expert guidance, great encouragement and strong support have made it possible. First of all, I would like to express my deep gratitude to my supervising professor, Laurel S. Stvan, for her valuable insights, helpful suggestions and infinite patience during the course of the writing of this work. Her sense of involvement has made it easier to achieve the goals of the research.

I would like also to extend grateful thanks to the members of my committee, Donald A. Burquest, David J. Silva, Jerold A. Edmondson, and Peter E. Unseth for their helpful and constructive comments on the draft of this dissertation. The dissertation is much better for their suggestions and advice.

The chair and faculty members of the Department of Linguistics and TESOL at the University of Texas at Arlington deserve warm thanks for their tremendous and determined efforts to make our academic dreams come true. I would like to take the opportunity to thank you all for your effective teaching, sheer inspiration, endless patience and fair and favorable treatment.

A word of appreciation should go to my professors and colleagues at Mutah, Yarmouk, Al al Bayt and Al-Hussein Bin Talal universities, Jordan. Special thanks go to Professor Adel E. Twessi, President of Al-Hussein Bin Talal University and Professor Muhammad R. Zughoul, Yarmouk University, whose intellectual efforts have had a profound impact on my academic life.

A word of gratitude must also go to Al-Hussein Bin Talal University, Jordan, for granting me a generous three-year scholarship to pursue my Ph.D. in the true academic world of the USA.

I owe deep gratitude to all my colleagues at UTA. In particular, I would like to thank Mark Miller and Mahmoud Smadi for the wonderful times we spent together.

I am forever indebted to all people who helped me in the course of data collection. Special thanks go to Prof. Mahmoud Kanakri (Mutah University), Dr. Hashem Al-Taweel, Dr. Abbas Dikaan, Dr. Mousa Al-Kurdi, Mr. Mahdi Elwi (Al-Hussein Bin Talal Universty), Dr. Ahmad Oleimaat (Al al-Bayt University) Dr. Bassam Btoush (Zarqa National University) and Mr. Kifah Omari (Hashemite University). Thanks also to Dawn Reice, Michael Daily and Jashua David for their native-speaker help.

I would like to extend my deepest gratitude to the people of the International Office at UTA. In particular, I would like to thank Dr. Joanna McClellan, Director of International Student and Scholar Services, for her effective cooperation.

I am thankful to my parents, brothers, sisters and my father-in-law and mother-in-law for their consistent encouragement and substantial help. I close with heartfelt thanks to Asia, my wife, and Sireen, my daughter, and I apologize to you both for being always busy with my research and for spending much time away from home.

August 2, 2004

ABSTRACT

INTERLANGUAGE LEXICOLOGY OF ARAB STUDENTS OF ENGLISH
A COMPUTER LEARNER CORPUS-BASED
APPROACH

Publication No. _____

Mousa Abdelghani Al-Btoosh, Ph.D.
The University of Texas at Arlington, 2004

Supervising Professor: Laurel S. Stvan

Since the very early emergence of machine-readable corpora into the linguistics scene in the 1960s, the direction of a considerable body of linguistic research began to shift from syntax and phonology, the, by then, focus of linguistic research, to a number of domains that remained mostly neglected under the umbrella of traditional approaches. Fortunately, lexicology, the target of this research, was a major beneficiary of that dramatic shift.

By employing a computer learner corpus-based approach, this study addresses multidimensional lexical aspects of a machine-readable corpus of the writing of Arab students of English as a foreign language. Lexical investigation of this corpus, which was solely compiled to serve the objectives of this study, required the existence of a similar sized authentic corpus, which was, in turn, methodically selected from *Louvain Corpus of Native English Essays* (LOCNESS). Via the computerized contrastive and analytical methods employed here, this dissertation aims at exploring: (1) learners' lexical complexity and

richness, (2) how far the learner corpus is deviant from the reference corpus in terms of the features and percentages of the top most 200 frequent tokens and hapax legomena, (3) how far the learner corpus is influenced by learners' L1, (4) the most salient lexical and stereotyped features of the learner corpus, (5) learners' lexical and collocational errors and (6) whether learners' collocational knowledge is on a par with their lexical knowledge.

Findings show that: (1) the learner corpus is much less complex in terms of lexical diversity and density than the reference corpus, (2) learners' top 200 tokens are markedly characterized by vague lexica, excessive overuse of the most frequently used words and L1 transfer, (3) rhetorically speaking, learners' writing is much closer to their L1 than to L2, (4) no source of lexical errors is more confusing for learners than near-synonyms, (5) a significant degree of diversity in terms of the incorrect use of collocations is obviously ascribed to the method of investigation, (6) a considerable body of collocational errors occurs as a result of the learners' limited word stock rather than from their ignorance of the collocability between the target lexical items, and (7) learners' free writing collocations are well-governed by their L1 collocations and thus, the degree of success in the use of the target collocations depends heavily upon the degree of similarity between the two languages (positive transfer).

TABLE OF CONTENTS

Appendix

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The rapid and progressive advancement of the artificial intelligence revolution during the last six decades has led to the introduction of a number of interdisciplinary fields in several realms of knowledge including linguistics. A quick look at such newly-established fields shows that they have a common feature, namely, they share the use of software programs as tools to examine the theories of their subjects. For this very reason, all these fields start with the word computational, e.g., computational physics, computational chemistry, computational linguistics. For Hausser (1999:13), computational linguistics is "a highly interdisciplinary field which comprises large sections of traditional and theoretical linguistics, lexicology, psychology of language, analytical philosophy and logic, text processing, the interaction with databases, as well as the processing of spoken and written language."

Research on the applicability of the ever-growing number of artificial intelligence software products has continued and succeeded in expanding to nearly all domains of linguistics. Consequently, computational linguistics has evolved into a number of subfields that reflect the different themes and methods of linguistics. Among the most important and widely studied topics that have grown out of the ongoing attempts to use computers in describing and analyzing language is corpus linguistics (CL, henceforth). Etymologically speaking, the word *corpus* (pl. *corpora*) is a Latin word meaning *body.* In a recent comprehensive account of the term, Hladka (2000:3) defines a corpus as a vast electronically processed, uniformly structured and continually added to collection of language

texts (written and oral) containing a variety of information the corpus might provide. The word *electronically* is used here to distinguish the pre-electronic corpora (e.g., *Survey of English Usage*) from the new machine-readable ones (e.g., the *British National Corpus*). Prior to the machine-readable age, corpora were used as reference books, and thus they were of more limited value. Oostdijk (1991:4) throws some light on the advantages of machine-readable corpora:

> Unlike earlier corpora, the corpora that are currently used are computer readable and lend themselves to automatic analysis. As a result, larger quantities of data can be processed at a greater speed, while consistency in the analysis is warranted through the use of a formalized description contained in the grammar.

Tribble and Jones (1990) argue that the central idea of CL, providing contextual evidence, is as old as linguistics itself. As they claim, this idea reaches back to the Middle Ages, when a number of scholars tried to make lists of all the words in particular texts together with their contexts- -what is today called concordancing. However, the history of the specific term CL, in its current sense, is relatively new, dating back to the beginning of the 1960s. The first attempt at computerized compiling of corpora was carried out by Nelson Francis and Kenry Kucera producing the well-known *Brown Corpus* in 1964. Since then, much research has been done in several languages all over the world (e.g., *Corpus of Spoken Bulgarian, Contemporary Portuguese Corpus*, and *Hypermedia Corpus of Japanese Conversation*).

Despite the significant criticism (by Chomsky 1962 and Abercrombie 1963, among others), of the first generations of corpora, electronic texts or corpora have attracted much attention and formed the nuclei for a large body of contemporary linguistic research in a remarkably short period of time. Besides the flexibility of data that can be easily reproduced and processed for different purposes, corpora allow for economy of time and effort.

Owing to the various functions that general corpora serve in linguistic research--e.g., providing linguistics evidence (phonological, morphological, lexical, syntactic), and use in producing dictionaries- -there have recently been numerous attempts to move from general corpora to more specific ones. As a result, it is quite common nowadays to have what is called *corpora for specific purposes.* For translation purposes, for instance, free-translation, parallel, comparative and bilingual corpora are much more useful than monolingual ones. In creating such corpora, we find that a language may need dozens of corpora or even more to satisfy the different application domains such as law, commerce, discourse analysis, rhetoric and second language acquisition (SLA, henceforth).

Though the above-mentioned ends were achievable by classical approaches, it is perhaps the corpus-based approach that can provide the most verifiable representative data about different aspects of language. Close inspection of the corpus-based studies conducted so far shows that the lexicography as well as lexicology, which remained almost neglected in the traditional approaches, are the major beneficiaries of the advent and development of corpora. As immediate results of the introduction of corpora, word frequency, word in context (concordancing) and collocations- -the likelihood co-occurrence between words- -have been recently targeted for intensive research worldwide.

CL relies chiefly on the notion of practical evidence, which is also the backbone of much of the SLA research. Consequently, SLA scholars have found corpora, specifically what are known as computer learner corpora, to be particularly useful to objectively investigate learners' interlanguage, a term coined by Selinker (1972) to refer to a separate linguistic system based on the observable output that results from a learner's attempted production of a target language norm. Furthermore, such corpora have made it possible to compare and contrast the interlanguage of language learners with similar authentic (native) corpora; and they have enabled researchers to examine the various stages of development in language learning and how the goals of the learners have progressed.

Such uses, therefore, explain the growing interest in and attempts to compile learner corpora in several languages worldwide, as evidenced by the *International Corpus of Learner English* and the *Corpus of English by Japanese Learners.*

Previous attempts to compile an English learner corpus of Arab students' writings are next to nonexistent. So, this study aims to fill in the gap by developing a representative machine-readable corpus of the written interlanguage of Arab students of English as a foreign language. The primary goal behind compiling this corpus is to investigate the learners' interlanguage lexicology (e.g., lexical complexity, text-profiling and lexical and collocational errors) represented in the corpus. Achieving this aim requires comparing and contrasting this corpus with a similar-sized authentic (native) English corpus. As can be clearly seen from the literature ahead, three reasons influenced the preference of lexicology as a target subject in this study. First, the paucity of literature devoted to this subfield vis-a-vis syntax, phonology, discourse analysis, etc. Secondly, the recent universal shift of interest from syntax and phonology to lexicology. Thirdly, the recent developments in software that have made studying a large body of data quite feasible.

At this point, it should be made clear that the term *lexicology*, which was originally a concern with the study of word stocks, must not be confused with the term *lexicography*, which is concerned with the compiling and producing of dictionaries. The terminological distinction between lexicology and lexicography has not been well-established in the literature yet (Klein 2001). According to Jackson (1988, cited in Jackson and Amvela 2000), lexicographical compilation might be considered to be derived from lexicological theory and thus, lexicography can be regarded as applied lexicology.

Being the first machine-readable corpus combined from the interlanguage of Arab students of English, this corpus is expected to be highly influential for future research on second language research, linguistic theory, natural language processing, lexicology, morphology, syntax, semantics, discourse analysis, speech and language learning, teaching

and testing. Furthermore, this corpus is expected to be an initial encouraging step towards compiling further corpora on different aspects of language teaching.

1.2 Objectives

Creating a machine-readable corpus is by no means an end in itself. Rather, it is simply a means of achieving the objectives behind its compilation and annotation. Thus, the type and size of the corpus, as discussed below, are governed by the research objectives. As for this study, six main objectives have been set for consideration.

- To explore learners' lexical complexity and richness vis-a-vis that of the native speakers (NSs, hereafter). Achieving this aim requires comparing and contrasting this corpus with a similar-sized authentic corpus.
- To identify the axial features characterizing the learner corpus(e.g., word categories, overproduced items, underproduced items). Special attention is paid to the features and percentages of the top 200 frequent tokens and to the hapax legomena, words used only one time in the corpus.
- To provide a detailed list of all lexical errors presented in the learner as well as the lexical translation corpora.
- To provide taxonomies for both lexical and collocational errors presented in learners' writing and translation.
- To examine the number as well as the percentage of each of the learners' lexical and collocational errors categories.
- To check if lexical knowledge is on a par with collocational knowledge.

As far as this study is concerned, learners' lexical and collocational errors are thoroughly investigated to identify all kinds of deviation, regardless of the source. Achieving such objectives will make it possible for learners, teachers and researchers to get accu-

rate and reliable information about the degree of deviation between subjects' output and native speakers' norms. Also, it will provide them with the areas of strengths and weaknesses and thus, enables syllabus designers to make needed corrections.

1.3 Research Questions

Despite the tremendous need for investigating several aspects of interlanguage lexicology of Arab students of English, it is often recommended that researchers not scatter their attention and lose focus, no matter how accessible their aims are. So, in order to avoid divergence or dispersing, this study has been limited to exploring and attempting to answer the below-mentioned seven questions. The first three questions deal with lexical complexity and text profiling while the rest are concerned with lexical and collocational errors.

1. To what extent does the learner corpus deviate from the reference corpus in terms of lexical complexity?
2. To what extent does the learner corpus deviate from the reference corpus in terms of the features and percentages of the top 200 frequent tokens and of the hapax legomena? And how can learners' lexical stereotypes be captured through word frequency?
3. What are the most salient and stereotyped features of the learner corpus? And how far is the learner corpus influenced by the learners' L1?
4. What are the most problematic words that Arab students of English encountered by the corpus?
5. What are the categories of the learners' lexical errors? And what is the contribution of each category to the total number of errors?

6. What are the categories of the learners' collocational errors? And what is the contribution of each category to the total number of errors?
7. Is learners' collocational knowledge on a par with their lexical knowledge?

1.4 Significance of the Study

The significance of this study stems from being one of the first attempts to compile and electronically analyze a representative computerized corpus of the written interlanguage of Arab students of English and also to establish the objectives set for it. Due to the scarcity of research done on this topic, this study is expected to contribute positively by providing an objective and reliable electronic text that can be a nucleus for various relevant corpora expected to be compiled in the future. In view of this, the benefit of such a corpus is potentially vast. Linguists, no matter what their interests, can approach it from different angles and thus its significance has the potential of not being restricted to a particular domain or phase.

Additionally, the findings of this study are expected to be of multifaceted significance. First, the study delineates learners' lexical complexity and richness in comparison with the reference (authentic) corpus. Secondly, it enhances teacher as well as student awareness of the importance of the proper use of lexical items and collocations in the mastering of the target language. Thirdly, it provides curriculum designers with areas of weaknesses in student writing and thus, enables them to make the revisions. Fourthly, it uncovers the differences between the subjects' output and the English norm. Fifthly, it provides students as well as researchers with contextual evidence of common interlanguage lexical and collocational errors. Sixthly, it may inspire other researchers to develop further corpora for different purposes.

1.5 Overview of Chapters

This dissertation has been divided into six chapters. The first chapter introduces the reader to the topic and several other key issues pertinent to the research objectives, research questions, significance of the study and definition of terms. The second chapter provides an overview of the relevant literature, which is, in turn, divided into four sections. The first section (2.1) *Introduction* briefly introduces the reader to the contents and objectives of the chapter. The second section (2.2) *Perspectives on Language Learning and Lexicology* reports on the status of lexicology within the framework of a number of theories that dominated the linguistic scene during the last six decades. The next section (2.3) *Lexicology* comments on the recent and contemporary recognition and development of lexicology as a vital component of language mastery. Section four (2.4), *Corpus Linguistics* focuses on the genesis, development and use of corpus linguistics as a new linguistic method. Furthermore, this section provides a description of corpora creation and analysis, along with what has become known as *interlanguage learner corpora.*

The details of the present study are established in the next chapters. In addition to the background information about the subjects, the setting of the study, data filtering, sociolinguistic variables and native speakers' judgments, Chapter Three describes the procedures employed in gathering and analyzing the data. Answers to the first three research questions concerning learners' lexical complexity and text-profiling are set forth in Chapter Four. Chapter Five provides a comprehensive and detailed analysis of the outcomes of the second half of the study concerning lexical and collocational errors. Finally, Chapter Six presents conclusions, implications and recommendations for further research.

1.6 Definition of Terms

- *Hapax legomena*: words that occur only once in a given corpus.
- Concordancer: a kind of search engine designed to present an index to the words in a text.
- Lexical complexity: a cover term for both lexical density and lexical diversity.
- Learner fluency: the learner's ability to keep pen to paper(measured by the number of words) without breaks in thought and cohesion.
- Lexical density: a lexical measure calculated according to the following formula: $\frac{\textit{the total number of content words X 100}}{\textit{the total number of all tokens in the given corpus}}$
- Lexical diversity: a measure of the spread or richness of the vocabulary in a text calculated according to the following formula: $\frac{\textit{the number of types (different words) X 100}}{\textit{the number of all tokens (instances of each word)}}$
- Part-of-speech (POS) tagging: The process of assigning lexical categories (that is, part-of-speech tags) to words in linguistic data.
- Span: this is the measurement, in words, of the co-text of the node. A span of -5, +5 means that five words on either side of the node word will be taken to be its relevant verbal environment.
- Text file: this is the simplest form of file on which words are stored. There is no formatting. A text file can be read by any computer regardless of operating system. In the Windows environment, the name given to any text file must end in '.txt'.
- Types and Tokens: the 'tokens' of a corpus refers to the simple word count, the number of running words in the corpus. The number of 'types' in a corpus refers to the number of different words in the corpus. These are the words that appear in a word index.
- Tag set: in computational linguistics, a set of possible tags for a given annotation task. For example, a part-of-speech tag set is a list of lexical syntactic categories which may be associated with lexical items.

CHAPTER 2

REVIEW OF RELATED LITERATURE

2.1 Introduction

Various subfields of linguistics immediately come to mind when one writes on research dealing with the use of corpus linguistics to examine learners' lexicology. For this reason, a deliberate attempt has been made to narrow the scope of the related literature by selecting and reporting only and synoptically on the areas most relevant to the topic being investigated, viz. language learning, lexicology and corpus linguistics.

2.2 Perspectives on Language Learning and Lexicology

2.2.1 Introduction

Over the past five decades, the SLA domain, as the literature shows, has been the target of active ongoing research worldwide. Close inspection of the research conducted on this field shows that various divergent arguments, hypotheses and theories have been proposed to account for the process of SLA. Such divergence reflects the different schools of thought that have attempted to facilitate and provide an explanation for language learning. However, not all aspects of SLA have been treated equally in terms of research and investigation. Lexicology, until fairly recently, for instance, has been largely neglected in most of the approaches that dominated the SLA scene during the last five decades. In what follows, an attempt is made to shed some light on how language learning and lexicology, in particular, were conceptualized by these schools and then, the recent recognition of the importance of lexicology in contemporary research.

2.2.2 From the Behaviorists' Perspective

Despite the lack of a precise date for its beginning, evidence in the literature indicates that the initial influential revolutionary seeds of SLA research originated in the behaviorists' attempts to describe second language learning. While there are certainly other possible starting points, a realistic history of this field goes back to the publication of Fries' *Teaching and Learning English as a Foreign Language* in 1945 and, then later, Lado's *Linguistics Across Cultures* in 1957. Although both authors are professed behaviorists by approach, the tenet of their works is blended in content. They mix certain aspects of behaviorist psychologists, who see language acquisition as a product of habit formation, and structuralist linguists who emphasize the detailed description of the two languages involved in the study (the mother tongue and the target language). The result of this blending was the emergence of the highly regarded Contrastive Analysis Hypothesis (CAH, hereafter) and the subsequent extensive contrastive analysis research. However, before attempting to engage in a discussion of Contrastive Analysis (CA henceforth), one should mention that language, from the behaviorists' perspective, is a part of human behavior and language learning is no more than a process of habit formation built through imitation and reinforcement. What happens in SLA, they claim, is that habits of L1 interfere in the learning of L2 habits (Rodriguez 2000). Such beliefs were the very cradle into which CAH was born.

CAH, which largely dominated the scene of SLA research for slightly more than two decades, claims that the principle barrier to SLA is the interference of the mother tongue or language transfer, the automatic, uncontrolled and subconscious use of the previously-learned behaviors in new situations. Lado (1957:2) states that similarities between native and target languages lead to ease in learning and differences lead to difficulty:

> We assume that the student who comes in contact with a foreign language will find some features of it quite easy and others extremely difficult. Those elements that are similar to his native language will be simple for him, and those elements that are different will be difficult.

Such an assumption led to controversy over learners' errors. As a result, proponents of CAH, in its heyday, were classified into two different groups, *purists* and *rationalists.* Consequently, this led to two simultaneous versions of the same hypothesis: (i) the strong version advocated by purists and (ii) the weak version advocated by rationalists. In the preface to *Linguistics Across Cultures: Applied Linguistics for Language Teachers,* Lado (1957) summarizes the principle ideas of the strong version: "we can predict and describe the patterns that will cause difficulty in learning, and those that will not cause difficulty, by comparing systematically the language and culture to be learned with the native language and culture of the student" (p.vii). On the basis of the strong version, a structural analysis of any two linguistic systems will enable a linguist to predict the kinds of difficulties a learner would encounter. The weaker version, which seems more realistic and practicable, claims that some errors are traceable to the influence of the mother tongue, and that CA is only valid to explain errors rather than predict them. In so doing, the weak version "begins with what learners do and then attempts to account for those errors on the basis of NL-TL differences" (Gass and Selinker 2001:73). Thus, it is rather obvious that CA, within the weak version framework, works together with error analysis.

Syntax and phonology, within the CAH framework, were the most popular in terms of attention and research. Lexica and collocations, on the other hand, were largely ignored. Fries (1945), whose ideas deeply influenced CA's researchers, argues that language learning does not mean learning vocabulary but rather mastering the sound system and syntactic structures of the target language. Lado (1957) links the difficulty in learning a new vocabulary item to the extent to which that item resembles or differs from the

learner's L1. Ramsey (1981) has attributed the lack of research on lexicon to the prevailing teaching method of that time, namely, the audiolingual method, which considers phonology and syntax as primary and lexicon as secondary: "teachers and syllabus makers still follow the precepts of the audiolingual approach in which vocabulary is relegated to a secondary status in comparison to phonology and grammar" (p.15). Since mastering considerable vocabulary is necessary to obtain proficiency in a target language, behaviorists assert that bilingual word lists are the most efficient technique to master a second language (Weinreich 1953). However, recent research pertinent to second language vocabulary has verified that decontextualized bilingual word lists are inadequate for long term mastery (Groot 2000:61).

The behaviorists' domination of SLA research, however, did not go unchallenged. Various empirical studies pointed to CA's failure to account for the existence of non-interference errors in language learners (Brooks 1960; Corder 1967; Olsson 1974, among others). Such studies also stressed that the percentage of language transfer is much less than what CA had claimed before. These findings, together with the new positive attitude towards learner's errors, hastened the emergence of the Error Analysis movement and Interlanguage Theory, both of whose findings, as illustrated below, would refute most of the findings of the earlier hypothesis.

2.2.3 From the Mentalists' Perspective

The emergence of Chomsky's article *A Review of B. F. Skinner's Verbal Behavior* onto the linguistics scene in 1959 shook behavioristic ideas to the roots and subjected them to increasing suspicion and criticism. Concepts such as *stimulus-response*, *habit-formation* and *reinforcement*, which were the heart of the behaviorists' tenets, were supplanted by Chomsky's *stimulus-free* proposition. Building on children's ability to produce sentences that have never been spoken before and to understand sentences that

they have never heard before, Chomsky concluded that the behaviorists' claims about language acquisition are logically and practically groundless. To account for the gap between the input and output in children's performance, Chomsky (1975) proposed the idea of Innate Knowledge. He defines innate knowledge as "the system of principles, conditions and rules that are elements or properties of all human languages not merely by accident but by necessity" (p.29). The principles, conditions and rules that comprise innate knowledge are often referred to as Universal Grammar (UG, henceforth). While principles apply to all human languages, variations among languages are accounted for in terms of parameters. More importantly, since principles are innate, children are presumed to learn only the parameters.

Though it was originally concerned with first language acquisition, Chomskyan linguistics has been extended to areas of SLA. A great number of SLA researchers found in Chomsky's revolutionary tenet a convincing tool to resolve part of the SLA riddle by claiming the full or partial accessibility of UG to L2 learners (White: 2000). However, opponents of this view argue that second language learners' knowledge of UG is mediated through L1. These divergent opinions evolved into two hypotheses divided sharply over the nature of the internal linguistic knowledge with which learners begin the SLA process (Gass and Selinker 2001:174). Access to UG and transfer are two variables in these hypotheses. First, the *Access To UG Hypothesis* claims that the innate language facility is operative in SLA and constrains the grammar of second language learner. Intensive research has been done, until fairly recently, to examine the accessibility of UG in adult L2 acquisition. Findings as summarized by White (2000) show five divergent arguments, which are still targets for intensive research worldwide:

(i) Full transfer/ partial (or no) access

(ii) No transfer/full access

(iii) Full transfer/full access

(iv) Partial transfer/full access

(v) Partial transfer/partial access

Proponents of the other view, the Fundamental Difference Hypothesis, claim "the learner constructs a pseudo-UG, based on what is known of the native language. It is in this sense that the NL mediates the knowledge of UG for second language learners" (Gass and Selinker 2001:176). They argue that a child's first language and adult SLA are totally different. The differences between first and second language acquisition, according to this hypothesis, are attributed to the four aspects of difference: (i) age, (ii) necessity, (iii) attitude and (iv) the existence of the previous knowledge.

The mentalists' priority of explanatory adequacy over the descriptive adequacy (Meyer 2002:2-3) explains the priority of syntax and phonology at the expense of other branches (e.g., lexica and collocations) in the literature of this approach. Furthermore, it should be borne in mind that even the attention paid to lexicology within the mentalist approach is attributed to the lexicon's vital role in determining the distribution of syntactic categories and subcategorization frames. Much of the contemporary research within the mentalist approach shows that the lexicon, which is not innate, is studied for the sake of syntax (Ouhalla 1999; Burquest 1999, among others). Haegeman (1999:36) states that "Words belong to different syntactic categories, such as nouns, verbs, etc., and the syntactic category to which the word belongs determines its distribution, that is in what contexts it can occur." This view also justifies the small amount of research done on the lexicon when compared with the extensive research carried out on syntax and phonology.

Thus, for lexica and collocations to be adequately investigated, a language should be approached from a new perspective that emphasizes language use rather than language

structure. In this sense, a corpus-based approach, which emphasizes language use, is perhaps the most effective method to be employed for this purpose, as will be illustrated below.

2.2.4 From the Autonomous Discipline Perspective

An overwhelming consensus among second language scholars indicates that SLA as an autonomous discipline began with the influential ideas and pioneer works of Corder and Selinker in the late 1960's and the beginning 1970's (Sharwood-Smith 1994; Ellis 1994; Gass and Selinker 2001, to name just a few). While both figures have associated themselves with what is known in the literature as Interlanguage Theory, Corder's research on error analysis makes him also the leader of the Error Analysis movement, which was the primary source of the Interlanguage Theory.

2.2.4.1 Error Analysis

In no previous publication on SLA are the learners' errors more positively highlighted and approached than in Corder's (1967) influential article, *Significance of Learner's Errors*, which is widely recognized as the cornerstone in a new phase that overturned the, by then, prevailing hypotheses and arguments of SLA research. Four significant findings of this article have been often used to refute the behaviorists' claims: (i) errors are not random, (ii) input, stretch of the target language available to the learner, should not be equated with intake, the portion of input that actually enters the cognitive process of the learner, (iii) mother tongue is not the only barrier to SLA and (iv) second language learners pass through certain stages of acquisition and thus, many errors are attributed to levels of development rather than negative transfer. Over and above such findings, the negative attitude towards errors which were prevalent during the heyday of CAH were supplanted by a new positive attitude. According to Error Analysis (EA, hereafter),

learners' errors are considered of great significance to the teacher, learner and researcher (Corder 1967):

1. Errors provide the teacher with evidence if s/he undertakes systematic analysis, and show how far towards the goal the learner has progressed and, consequently, what remains for her/him to learn.
2. Errors provide the researcher with evidence of how language is learned or acquired, and what strategies and procedures the learner uses in her/his discovery of the language.
3. Errors tell the learner about her/his weaknesses, and they provide him with an accurate way to test her/his hypotheses about the nature of the language s/he is learning.

Another crucial issue that Corder brings to light is the distinction between *errors* and *mistakes*. Systematic deviation made by learners who can't correct themselves because they have not yet acquired the rules pertinent to such structures are called errors and these, according to him, are worthy of investigation and explanation. Learners' errors, he argues, reflect lack of competence and cannot be self-corrected. Unsystematic performance slips, on the other hand, are caused by excitement, lack of attention or fatigue. These slips have nothing to do with competence; they are called mistakes and can be self-corrected. The concern of the EA researchers (Corder (1967, 1971), Richards 1974 and Jain 1974, to name just a few) with lexicology as a major target for investigation did not go far beyond what we saw in the previous approaches.

While the predecessors of learner corpora can be traced back to EA era, there are several distinctive features that make learner corpora compiled during this period different from the current generation of computer-based corpora (Granger 1998:5). First,

EA learner corpora were very small (sometimes no more than 2,000 words) in comparison with the current generation of corpora (e.g., the *British National Corpus* consists of 100,0000,000 tokens). Secondly, EA researchers paid minimum attention to a variety of factors that might influence learners' output (Ellis 1994, cited in Granger 1998:5). Thirdly, EA learner corpora were neither compiled nor used in the same manner as computer-learner corpora. Rather, they were frequently discarded once the errors were extracted.

2.2.4.2 Interlanguage Theory

Empirical research on learner errors has shown that the output of a language learner is almost always characterized by a considerable body of deviant forms that can be attributed neither to L1 nor to L2. Such a conclusion led Selinker (1972) to postulate the existence of transitional system called *interlanguage.* As defined in chapter 1, interlanguage is a separate system based on observable output that results from a learner's attempted production of a target language norm. This system, according to Selinker, is the output of five cognitive processes:

1. Language transfer- -the automatic, uncontrolled and subconscious use of the previously-learned behaviors in new situations. In this case, the learner uses her/his L1 as a resource.
2. Transfer of training- -fossilizable items, rules and subsystems that occur as a result of identifiable items in training procedures.
3. Strategies of language learning- -fossilizable items and rules that occur as a result of an identifiable approach by the learner to the material to be learned.
4. Strategies of communication- -deviant items that result from the learner's strategy to communicate with native speakers of the target language.

5. Overgeneralization- -errors that result from overextension or overgeneralization of rules and semantic features of the target language.

In brief, this system is basically attributable to developmental learning stages and fossilization, the cessation of learning. In addition to the aforementioned five cognitive processes underlying interlanguage knowledge, this theory has a number of other features (Yang 1999, 323-36):

(i) Interlanguage is independent- -the term *independent* is used here to indicates "the separateness of a second language learner's system, that has a structurally intermediate status between the native and target languagees" (Selinker 1972:16).

(ii) Interlanguage is dynamic- -L2 learners pass through stages of development and, thus, their in-between system is continually changing.

(iii) Interlanguage is permeable- -learners' interlanguage rules and features are open to amendments; they are not stable or fixed.

(iv) Interlanguage is systematic- -learners' interlanguage is not random. Rather, it is based on existing systematic rules and features.

(v) Interlanguage is a process reflecting learning psychology- -this indicates that learners' systems or varieties involve assimilation, accommodation and creative-construction processes that echo language learning.

Historically, the evolution of Interlanguage Theory coincided with the new revolutionary attitudes towards the lexicon, which emphasized the importance of the lexicon in language teaching (Wilkins 1972, Lord 1974, Richards 1976, Judd 1978, among others). However, interlanguage research was not influenced by such attitudes, rather its concerns were merely a juxtaposition of the previous theories. Interlanguage literature was pri-

marily devoted to syntax and phonology and secondarily to discourse and pragmatics. The great portion of the limited interlanguage research conducted on lexicon is devoted to the acquisition order of morphemes (Dulay and Brute 1974, Ellis and Roberts 1987, among others).

Perhaps no single area has been as unstudied, within the Interlanguage Theory framework, as interlanguage lexicology. The situation is weaker when one speaks of interlanguage lexicology of Arab students of English as a foreign language. The semi-dearth of attention devoted to lexical errors did not preclude the recent research conducted on collocations, however. In their attempt to examine the collocational knowledge of junior and senior English majors at Yarmouk University and language teachers at the Higher College for the Certification of Teachers, Farghal and Obiedat (1995:315) concluded that both students and senior English majors "are seriously deficient in collocations." Shakir and Shdeifat (1996) shed light on the ability of the learners to translate collocations as an indicator of development of foreign language competence. In his doctoral dissertation, Al-Zahrani (1998) examines the knowledge of English collocations among four groups of Saudi EFL students representing four academic levels. Zughoul and Hussein (2003) study the collocational strategies used by Arab students of English when they use English collocations.

In view of what we have seen in the preceding sections, second language lexical acquisition has been of peripheral concern in almost all of the schools that dominated linguistics and language teaching up to the end of the twentieth century. A remedy for this gap was not totally inaccessible, however. Numerous serious initiatives to bring lexicology onto the scene were intermittently seen in the literature as illustrated below.

2.3 Lexicology

2.3.1 Recognition and Development

Having briefly examined language learning and lexicology within the framework of a number of traditional approaches, this dissertation will now proceed to examine the roots of the neglect of lexicology in modern linguistic research in general and specifically the genesis of its renaissance in contemporary research.

Clear-cut evidence concerning the reasons behind the absence of lexicology in modern linguistic research as an independent domain investigated for its own sake comes from a number of leading figures such as Bloomfield (1933), Fries (1945) and Chomsky (1965). According to Koenig (1999), both Bloomfield (1933) and Chomsky (1965) assume that a lexicon consists of a theoretically uninteresting repository of idiosyncrasies. Such a proposition, which prevailed for several decades, was considered the defining reason behind the priority of syntax and phonology. Whereas syntax and phonology, within the Chomskyan framework, are governed by a number of universal principles and parameters, the lexicon goes ungoverned. It is worth reiterating that Fries (1945) states that language learning does not mean learning vocabulary but rather mastering the sound system and syntactic structures of the target language. Such arguments proposed by influential and leading figures have led linguists and SLA scholars to sacrifice lexicology on the altar of syntax and phonology.

Recent studies in SLA have shown that no linguistic impropriety is more likely to lead to misunderstanding than errors in lexical choice. This explains the increasing trends in SLA that have called for the preference of lexicology over syntax and phonology. Such calls are largely based on the high percentage of lexical errors observed in language learners vis-a-vis phonological and syntactic errors. Politzer (1978:257) states that errors of vocabulary are the most serious errors for the language learner and they outnumber any

other type of error. As a sign of full recognition of the importance of lexicon, Gass and Selinker (2001) allotted a separate chapter entitled *The Lexicon* in the most recent edition of their book *Second Language Acquisition: An Introduction.* In this chapter, the authors cite different arguments concerning the vital role of lexicon in SLA. They also propose that although the lexicon has received the least attention in interlanguage literature in comparison to other parts of language, the picture is quickly changing. Furthermore, they argue that the recent research on SLA has shown that the most neglected part, the lexicon, "may be the most important language component for learners" (p. 372).

Perhaps the importance of lexicology in contemporary research is no more clearly stated in the literature than in Laufer (1997:147):

> Vocabulary is no longer a victim of discrimination in second language learning research, nor in language teaching. After decades of neglect, lexicon is now recognized as central to language acquisition process, native or non-native.

Though its concerns are different from the concerns of pure lexicology and the aims of this study, the current concerns of Chomskyan linguistics with lexicon could open the door to further serious research on this domain. Theoretically, language acquisition, from the Minimalist Program perspective, should be totally concerned with lexicology. Chomsky (1991, cited in Cook 1996:87) argues that "there is only one human language apart from the lexicon, and language acquisition is in essence a matter of determining lexical idiosyncrasies." This quotation indicates that language acquisition is, in its core sense, the learning of vocabulary. The Lexical Parameterization Hypothesis states that "the values of a parameter are associated not with particular grammars, but with particular lexical items" (Manzini & Wexler 1987). Such improvement in the status of the lexicon in theoretical and applied linguistics led Groot (2000:61) to state that viewing vocabulary as a set of irregularities is a näive view and long outdated.

In her attempt to examine the attitudes of English-speaking professors towards university ESL students, Wright (2000) examined several variables including the interactivity between professors' judgements and learners' fluency in lexicon (writing). Her findings show that professors form a relatively more positive judgement of learners who write longer and larger sentences. This, of course, reveals that learners' proficiency in lexicon and syntax are crucial factors in writing, which are, in turn, crucial factors in the professors' assessments.

Furthermore, applied research on lexicology has also emphasized the importance of lexical knowledge, (knowledge of individual words or relations between words) in mastering different aspects of the target language. Zhang (1993) argues that proficiency in second language writing is directly connected to the degree of lexical mastery. The greater the word stock a learner has the better. Saville-Troike (1984, cited in Willis, 1998) states that vocabulary is the most important aspect of L2 knowledge for academic achievement. For Zughoul (1991), the lack of the right lexicon may lead to misunderstanding between interlocutors. From a more general standpoint, errors of lexicology result from either an inappropriate use of a lexical item or from the ignorance of the collocabilty among the lexical items in question.

2.3.1.1 Lexical Choice

According to Edmonds (1999:2), *lexical choice* refers to "the process of determining which word in a given language most precisely expresses a meaning specified in some formalism other than the given language itself." As he argues, the goal of lexical choice is to "verbalize the exact denotation and connotation desired, and nothing else" (p.2). In this sense, a lexical choice error means that an item is used inappropriately in a particular context due to an error or misuse in its semantics, connotation, register, vagueness, generality, specificity, etc. In his attempt to propose a new model for lexical choice

architecture, Reiter (1990:23) states that "the lexical choice process should be regarded as a constraint satisfaction problem: the generation system must choose a lexical unit that is accurate (truthful), valid (conveys the necessary information), and preferred (maximal under a preference function)."

Various studies devoted to lexicology and communicative competence have explicitly indicated that lexical choice errors often lead to misunderstanding either locally or globally. Recently, however, some scholars have asserted that ungrammatical utterances with accurate vocabulary are much more understandable for native speakers than those utterances with grammatical but inaccurate vocabulary (Widdowson 1978, cited in Lafford et al. 2000). Lexical errors, according to Gass and Selinker (2001), are numerous and disruptive and both native and non-native speakers of a language recognize the importance of getting the appropriate word. Lexical choice errors in both spoken and written discourses, as the literature shows, make up a considerable percentage of the grand total of all kinds of errors (Petrarca 2002:64). In a relevant empirical study that gives full credit to such argument, Politzer (1978:257) states that statistically native speakers of German judge lexical errors by English speakers to be the worst type of errors, as shown in Table (2.1).

Table 2.1. Native speakers' judgement of errors type

Type	Number	% of NSs' Judgment
Vocabulary	2234	77
Verb Morphology	1600	55
Word Order	1562	54
Gender Confusion	1502	51
Phonology	1045	36
Case Ending	821	28

Carter (1987:65) states that lexical choice errors in the early stages of learning, in particular, are attributed to several sources including interlingual and intralingual ones. He writes that:

> errors may result from a mismatch in morphophonemic correspondence (the fit between sound and written form), from inserting the word in the wrong grammatical slot or from failing to locate grammatical dependencies, from inaccurate first language transfer (often leading to specific semantic errors), and from intralingual confusion, that is, as a result of failing to distinguish appropriately between and among lexical items in the target language.

Unlike syntactic or phonological errors, lexical errors and learners' level are reversely interactive. Martin (1984) argues that "as the fluency of advanced language learners increase, so too does the number of vocabulary errors generated, both in speaking and writing." The majority of learners' lexical errors, she argues, "reflects confusion between and among lexical items in the target language itself." For her, there are four types of dissonance between a lexical item and its appropriate use: (i) stylistic, (ii) syntactic, (iii) collocational and (iv) semantic.

The increasing awareness of the centrality of lexicology in SLA research is revealed in the discovery that learners' lexical richness and errors are determinant factors in second language proficiency in general and in evaluating their writing in particular (Linnarud 1986, Engber 1995, to name just a few). Based on learners' judgments of the difficulties they encounter in the course of their second language acquisition, Meara (1982:100) argues that lexicon, which suffered from long-term absence of research in second language learning literature, is the most problematic area for learners:

> vocabulary acquisition is part of the psychology of second-language learning that has received short shrift from applied linguistics, and has been very largely neglected by recent developments in research. This neglect is all the more striking in that learners themselves readily admit that they experience considerable difficulty with vocabulary, and once they have got over the initial stages of acquiring their second language, most learners identify the acquisition of vocabulary as their greatest single source of problems.

Regardless of the lack of a universal taxonomy for lexical errors, empirical research on lexicology worldwide has revealed several common sources of lexical errors, not least of which are the influence of L1, near-synonymity, paraphrasing, idiomaticity and avoidance, ideas to which I will return in Chapter Five.

2.3.1.2 Collocations

The observation of repeatedness in the occurrence of words or what is called *collocations* is by all means not new; indeed the phenomenon of collocation reaches back to 300 BC (Robins 1967). However, the term *collocation* itself is new, dating back to the fifth decade of the last century. Firth (1957:196), the coiner of the term, states that the meaning of collocation has nothing to do with the conceptual approach to the meaning of words:

> The statement of meaning by collocation and various collocabilities does not involve the definition of word-meaning by means of further sentences in shifted terms. Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*.

It is well-known that words of a language do not combine randomly. Rather, their occurrence is sometimes predictable or even governed. In the introduction to the *Oxford Collocations Dictionary*, Deuter et al. (2002:vii) distinguishes between three types of words combinations:

> Combinations of words in a language can be ranged on a cline from the totally free-*see a man/car/book* - to the totally fixed and idiomatic - *not see the woods for the trees.* This idiom is not only fixed in form, it also has nothing whatever to do with wood or trees. Between these two extremes, there is a whole range of nouns that take the verb see in a way that is neither totally predictable nor totally opaque as to meaning. These run from the fairly 'weak' collocation *see a film* (which elementary students learn as a 'chunk' without pausing to reflect that this is not quite the literal meaning of *see*) through the 'medium strength' *see a doctor* to the 'stronger' collocations of *see danger/reason/the point.* All these combinations, apart from those at the very extremes of the cline, can be called collocations. And it is combinations such as these - particularly in the 'medium-strength' area-that are vital to communicative competence in English.

For Choueka (1988), collocations or predictable patterned combination of words share some semantic and syntactic features peculiar to them.

> A sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived from the meaning of connotation of its components.

Broadly speaking, collocations are of two types: *lexical* and *grammatical.* Lexical collocations, according to Sinclair (1991), refer to the occurrence of two words or more within a short space of each other in a text. This kind of collocation is divided into two sub-categories: paradigmatic collocations and syntagmatic collocations. The distinction between such sub-categories is stated clearly in Doerr (1994: 8-9):

> Paradigmatic collocations are defined as *lexical items* in which one lexical item (the collocate) can occur instead of another lexical item (the basis) in a particular collocation in a particular language. In this type of collocation, both the basis and the collocate must be of the same taxonomic class, such as synonyms, antonyms or hyponyms, and identify semantically related words. In "ladies and gentlemen", both are antonyms."Fire and flame" is an example of a paradigmatic collocation containing hyponyms. Paradigmatic collocations are frequently found in syntactic structures such as the conjoined phrases "left and right" and "boys and girls" in which the basis and the collocate are separated by a closed-class items.
> Syntagmatic collocations are defined as lexical items in which one lexical item (the collocate) can occur next to another lexical item (the basis) in a particular natural language. In this type of collocation, both the basis and the collocate together express a collocational phenomenon, that is, they co-occur frequently in a given text. For example, the collocates "flips" and "tosses" can be adjacent to the basis "a coin" ("flips a coin", "tosses a coin") but "rotates" (*"rotates a coin") cannot in the context of meaning to-flip-a-coin.

Grammatical collocations, on the other hand, as explained in Hsu (2002:36), consist of a content word (verb, noun, or adjective) and another grammatical word (preposition, or certain structural patterns, i.e. a "that"-clause, *to* + infinitive, or gerund). Though it is not easy to distinguish between fixed word combinations and other combinations of words that are open to creative formulation, there is a set of features that make collocations or fixed combinations of words differ from other unfixed ones. Setting up criteria for classifying group of words that function as units (collocations) has been the

target of intensive research during the last few decades (e.g., Kjellmer 1984). Nation (2001:328-332) presents a set of scales, which indicate what is primarily involved in learning collocations:

1. Frequency of co-occurrence
2. Adjacency
3. Grammatically connected
4. Grammatically structured
5. Grammatical uniqueness
6. Grammatical fossilization
7. Lexical fossilization
8. Collocational specialization
9. Semantic opaqueness
10. Uniqueness of meaning

In their influential model of cohesion, Halliday and Hasan (1976) assert that lexical cohesion is of primary importance in producing a coherent text. For them, "lexical cohesion embraces two distinct though related aspects which we referred to as reiteration and collocation" (p. 318). Whereas reiteration refers to the repetition of a lexical item or the occurrence of a synonym, collocation, on the other hand, refers to "a word that is in some way associated with another word in the preceding text, because it is a direct repetition of it, or in some sense synonymous with it, or tends to occur in the same lexical environment, coheres with a word and so contributes to the texture." (p. 319).

In their attempts to examine collocations, researchers have employed three approaches-- the lexical composition approach, the semantic approach, and the structural approach (Gitsaki 1999). *Lexis*, from the lexical composition approach perspective, is "an autonomous entity, choosing its own collocates which in turn can be enumerated and classi-

fied in lexical sets" (p. 26). While the semantic approach, in order to predict their collocates, focuses on the meaning of the semantic features, the structural approach provides patterns of collocations that join together grammatical and lexical words. Collocations in this study are approached from the lexical composition perspective.

Keller (1991) argues that the human mental lexicon is made up of both single words and larger fixed and more variable phraseological units. Thus, knowledge of collocations is deeply rooted in our minds, and its role in second language learning should not be disregarded. Brown (1974), Channell (1981), among others, emphasize the importance of collocations in language learning, and they argue for intensive teaching of collocations for learners. The importance of collocations stems from the roles they play in the language skills: listening, communication, reading and writing. Some other scholars (e.g., Aghbar 1990, cited in Al-Zahrani 1998) assert that the knowledge of formulaic language is essential for language fluency of both native and non-native speakers. Drawing on the findings of their research concerning the German advanced EFL students' productive knowledge of English collocations, Bahns and Eldaw (1993:101) conclude that "EFL teaching should be concentrated on those collocations which cannot be paraphrased." Among the significant points raised in this article is the distinction between vocabulary knowledge and collocational knowledge. This procedure, to which I will return below, shows that the learners' low productive knowledge of collocations is sometimes attributed to the lack of lexicon rather than the inability to use the multi-word lexemes.

Since the introduction of corpora into the linguistic scene, collocations have attracted increasing interest, especially by lexicographers, whose enthusiastic attempts have evolved recently into producing a number of dictionaries and corpora that are devoted either completely or partially to collocations. Such works include *Oxford Collocations Dictionary for Students of English*, *The Kenkyusha Dictionary of English Collocations*, *A*

Dictionary of English Collocations, *Student's Dictionary of Collocations*, *Collins Cobuild Learner's Dictionary*, etc.

2.3.2 Lexical Competence

It is still an open question as to what competence really means. A short review of the relevant literature indicates that Chomsky views competence as knowledge while it is knowledge and ability for Hymes (1972). As far as lexicon is concerned, competence is directly connected to knowledge and appropriateness. Meara (1996, cited in Lafford et al. 2000) proposes that lexical competence is measured by both the size of a learner's store of lexical items as well as the organization of such items. As to size, it is commonly believed that the learner's reading and writing abilities depend solely upon the learner's lexical repertoire (the number of lexical items that a learner has, at least, some knowledge of). Organization, on the other hand, refers to all types of knowledge that result from the knowledge of a word. Nation (1994:121-122) states that lexicon knowledge entails several other relevant components and skills . As can be readily seen from the criteria listed below, these skills can be reclassified into three broader categories of knowledge: (i) knowledge of form, (ii) knowledge of meaning and (iii) knowledge of use.

1. Being able to recognize the spoken form of the word.
2. Being able to pronounce the word.
3. Being able to spell the word.
4. Being able to write the word.
5. Knowing the underlying meaning of the word.
6. Knowing the range of meanings of the word.
7. Knowing the grammatical patterns the word fits into.
8. Knowing the affixes the word stem can take.

9. Knowing the words that fit into the same lexical sets.
10. Knowing the typical associations of the word.
11. Knowing the range of collocations of the word.
12. Knowing whether the use of the word is limited by considerations of politeness, gender, age, country, formality, and so on.
13. Knowing whether the word is commonly used or not.
14. Being able to use the word receptively and productively.

Similarly, Pawley and Syder (1983) argue that native-like command of the target language requires both native-like selection and native-like fluency. Native-like selection refers to "the ability of the native speaker to convey his meaning by an expression that is not only grammatical but also native-like" (p.191). Native-like fluency, on the other hand, refers to "the ability to produce fluent stretches of spontaneous, connected discourse" (p.191).

Like any other aspect of language, vocabulary acquisition and competence are affected by several intralexical factors. Laufer (1991:305) presents several phonological, morphological and semantic facilitating and difficulty-inducing intralexical factors that affect vocabulary learning shown in Table (2.2):

Table 2.2. Intralexical factors which affect vocabulary learning

Adapted from Laufer (1990: 305)

Facilitating factors	**Difficulty inducing factors Intralexical**	**Non-effective factors**
unproblematic pronunciation	: difficult pronunciation :(presence of foreign sounds)	: :
inflexional regularity	:inflexional complexity : :	: : :word length
derivational regularity	:derivational complexity :	: :
morphological transparancy	:deceptive morphological :transparency :	: : :part of speech
generality	:specificity :idiomaticity	: :concreteness/ :abstractness
nonidiomaticity	:	
one form representing one meaning	: one form representing : several meanings : (polysemy/homonomy)	: : :
register neutrality	:register restrictions	:

To sum up, the preceding sections have substantiated the contention that interlanguage lexicology until fairly recently, has been mostly neglected. This fact, together with the vital importance of lexicology in SLA acquisition, makes it obvious that this largely neglected topic should garner further research and be made a priority in language learning. However, with the emergence of the corpus-based approach into the scene, it has become feasible to give lexicology its due. For Biber et al. (1998), the weaknesses

of traditional approaches turn out to be the strengths of corpus-based approaches. Some of these strengths are attributed to its ability to examine several domains that remained unaccounted for under the previous approaches.

2.4 Corpus Linguistics

2.4.1 Attitude and Use

A survey of the corpora developed worldwide so far shows a wide gap among languages in the concern with corpora, and with CL in general. While some languages, e.g., English, have been of increasing interest in CL, others, such as Arabic, have seen confined interest in this respect. This explains the rapid growth of English corpora compared with Arabic corpora. The following samples of corpora provide a finely-focused picture of the concern of English with corpus linguistics and corpora during the past five decades (source: *Gateway to Corpus linguistics on the Internet*):

1. Brown University Corpus

Org: Brown University, Rhode Island,U.S.

Time: 1960s

Size: ca. 1 million words

Contents: American written English; 500 text samples of approximately 2,000 words distributed over 15 text categories

Access: available on the ICAME CD-ROM

2. LLC London-Lund Corpus of Spoken English

Org: Time: 1960s-mid-1970s

Size: 500,000 words

Contents: spoken British English

Access:

Notes: The LLC is the result of two projects: SEU (1959) at University College London and SSE at Lund University in 1975

3. FROWN - Freiburg BROWN Corpus of American English

Org: University of Freiburg, Germany

Time: 1991-92

Size: ca. 1 million. words

Contents: "The ultimate aim was to compile parallel one-million-word corpora of the early 1990s that matched the original LOB and Brown corpora as closely as possible"

Access: available on the ICAME CD-ROM;

Notes: SGML Markup; FROWN was created as a parallel corpus to the BROWN corpus but with data from the 1990s.

4. BNC - British National Corpus

Org: Led by an industrial/academic consortium lead by Oxford University Press

Time: completed in 1994; first release in 1995; second release in 2001

Size: over 100 million words (4,125 texts)

Contents: multigeneric; 90 percent written and 10 percent spoken materials

Access: Licensed; Guest account available by using the SARA Client at the BNC Online Service or conduct a simple search at the BNC.

Notes: SGML Markup according to the TEI guidelines; POS tagging carried out with CLAWS

A cursory look at the above corpora, together with other regional, general and specific corpora developed during the past five decades reveals three crucial aspects. First, the concern with corpora has been constantly increasing since the creation of the *Brown Corpus* in 1964. Secondly, corpora have substantially benefited from the continuous progress in artificial intelligence. This benefit is evident in the ever growing software products used today in corpus analysis as well as the huge gap in storage capacity between the first generation of corpora (e.g., *Brown Corpus*, 1,000,000 tokens), and the current generation (e.g., *British National Corpus*, 100,000,000 tokens). Thirdly, the existence of regional corpora (e.g., *British National corpus*, *The Australian Corpus of English*), authentic (native) corpora and learner corpora (*LOCNESS*), *The International Corpus of Learner English*), spoken corpora (*Corpus of Spoken Professional American English*) and written corpora (e.g., *Longman Written American Corpus*) bears witness of the divergent functions of corpora in language and linguistic research.

It should be made clear that CL is still oscillating between the ideas of empiricists and those of rationalists. Chomsky, the founder of the modern rationalistic school of linguistics, argues that a linguist should rely on the reality of competence rather than on performance. For this reason, rationalists feel that the nonoccurrence of X and Z items in a corpus does not prove the nonexistence of such items in the internalized system of the speaker or writer; in short, a linguist should describe grammar rather than enumerate sentences (McEnery & Wilson 1996).

Empiricists, on the other hand, argue that CL is a fertile field and is the best method developed thus far to reflect competence and to provide researchers with large bodies of naturally occurring data. Some linguists, on the other hand, have attempted to bridge the gap between theoretical and descriptive linguistics by emphasizing their complementary roles in linguistic research. Leech (1992:27) states that both types are

mutually contributory:

> Both types of linguistics are valid in their own terms, and should be regarded as mutually contributory. Descriptive linguistics can be just as answerable as the "theoretical linguistics" of language universals. In fact, descriptive linguistics is more amenable to theory construction and testing in accordance with the tenets of scientific method, because the nature of its data (i.e. utterances in a particular language) is less abstract and more directly observable.

In fundamental agreement with Leech's view about the status of CL in the theoretical investigation of language, Halliday (1992:41) states that the evidence that CL can provide has important implications for several areas of theoretical inquiry:

> Corpus studies have a central place in theoretical investigations of language. There are many ways in which a corpus can be exploited, of which the one considered here - by no means the only one–is that of providing evidence of relative frequencies in the grammar, from which can be established the probability profiles of grammatical systems. These in turn have implications for at least five areas of theoretical inquiry: developmental, diatypic, systemic, historical and metatheoretic.

Taking the empirical view of language one step further, one may conclude that the heart of empirical linguistics lies in the notion of evidence. It should be born in mind that evidence within a CL framework is based on experience and observance rather than prediction or guessing. Kennedy (1998:7-8) states that CL is not a theory in competition with other linguistic theories but rather a source of evidence that comprises the core of any linguistic study.

> Linguists have always needed sources of evidence for theories about the nature, elements, structure and functions of language, and as a basis for stating what is possible in a language. At various times, such evidence has come from intuition or introspection, from experimentation or elicitation, and from descriptions based on observations of occurrence in spoken or written texts. In the case of corpus-based research, the evidence is derived directly from texts. In this sense corpus linguistics differs from approaches to language, which depend on introspection for evidence.

Importantly, corpus-based studies have shown extraordinary capabilities of uncovering certain linguistic aspects (particularly those related to language use and colloca-

tions) that have remained unattainable by traditional approaches. For example, due to the scarcity of corpora for Modern Standard Arabic, one can hardly provide reliable answers to questions related to word order patterns, dialectical differences, collocations or percentages of loan words.

Passing to matters more closely related to internalized linguistics, Chafe (1992: 79-95) argues that corpora "are an absolutely crucial part of the linguistic enterprise" and he adds that a corpus linguist is one who aims to "understand language and behind language the mind by carefully observing extensive natural samples of it and then, with insight and imagination, constructing plausible understandings that encompass and explain those observations."

From an empirical perspective, the naturally occurring data that a corpus provides us with are believed to be superior to any hypothetical and non-natural (inauthentic) data. As Aarts (1992) points out, CL can be efficiently used to produce *observation-based* instead of *intuition-based* grammar. At this stage, CL can claim to be a better or, at least, an equally useful tool in linguistic analysis, be it syntactic or semantic, than the intuition of the native speaker can provide. For Aijmer and Altenberg (1992:2), corpora have become "excellent resources for a wide range of tasks." This, they claim, is due to two main reasons:

1. Language corpora have provided a more realistic foundation for the study of language than earlier types of material, a fact which has given new impetus to descriptive studies of English lexis, syntax, discourse and prosody.
2. Language corpora have become a particularly fruitful basis for comparing different varieties of English, and for exploring the quantitative and probabilistic aspects of the language.

Biber et al. (1998: 233) argue that a corpus-based approach takes advantage of several things that contribute positively to making it more powerful and applicable to the study of individual linguistic features:

> This approach takes advantage of: computers' capacity for fast, accurate, and complex analyses; the extensive information about language use found in large collocations of natural texts from multiple registers; and the rich descriptions that result from integrating quantitative findings and functional interpretations. For these reasons, the corpus-based approach has made it possible to conduct new kinds of investigations into language use and to expand the scope of earlier investigations.

Some of the continuing success of corpus-based approaches is attributed to a concordancer's ability to process a large body of information that would require thousands of tedious hours by other approaches in a short period of time. For example, it has become possible to identify the discourse markers or the distribution of tenses in a hundred million-word corpus in minutes. Such a work may take months or even years to complete by traditional approaches.

Recent empirical research conducted on corpora, including learner corpora (Kennedy 1990; Tognini-Bonelli 2001; Hunston and Francis 2000, to name just a few) has pointed out that a well-compiled and annotated corpus can provide researchers and learners with comprehensive knowledge of lexical features. First, it shows the different contextual meanings associated with a particular word. Secondly, since words do not occur or group together in a text randomly, a corpus provides a description of the commonly found words that co-occur with a particular word (collocations). Thirdly, the frequency of a word can be shown relative to other related words. This, of course, provides teaching material designers with sufficient background about the main and frequently used vocabulary in the language. Fourthly, the non-linguistic association patterns that a particular word has to a register or dialect can be easily found. Fifthly, the use and the distribution of seemingly synonymous words can be detailed.

A corpus is also extremely useful in investigating the mismatch between the rules of prescriptive grammarians and the linguistic facts in language teaching. For example, Kennedy (1991, cited in Tognini-Bonelli 2001) points out that it is not always easy to draw a distinction between words depending upon the grammatical terms: "various meanings of the words sometimes overlap regardless of whether they function as prepositions or adverbs." Thus, he argues that the basic grammatical distinction between prepositional and adverbial uses of *between* and *through* lies in the word class they each most frequently associate with: nouns before *between* and verbs before *through.* This indicates the importance of grammatical collocations to distinguish between the two words. Another explicit example of the mismatch between what is believed and taught and what it is real and practiced is the traditional equation between *if not* and *unless* (Berry 1994; cited in Tognini-Bonelli 2001:17).

Corpora have also played a significant role in meaning disambiguation. According to Tognini-Bonelli (2001:25-33), corpora help learners "identify and distinguish between particular meanings which may be neither reported in reference dictionaries nor explained with reference to grammatical structures." The author provides evidence from the positive answer he made to a question raised by one of his English class student "whether *all but* is the same as *except*" (p. 25). Though both dictionaries and reference grammars failed to provide the accurate distinction between them, the corpus did succeed in doing so.

The deep concern with lexicon within this approach has led Francis and Sinclair to argue vehemently against the traditional separation between lexis and grammar. As they argue, lexis and grammar should be treated as one category. Francis (1995, cited in Hunston and Francis 2000:30) explicitly express this complementary relationship:

> Particular syntactic structures tend to co-occur with particular lexicon items, and -the other of the coin - lexicon items seem to occur in a particular range of structures. In short, syntax and lexis are co-selected, and we cannot look at either of them in isolation.

Other immediate results of the introduction of corpora in linguistic research are clearly seen in historical linguistics as well as sociolinguistics. By employing corpora in comparative studies, it is now feasible to examine various issues related to vocabulary loss, borrowing and semantic change. The same method, in sociolinguistics, on the other hand, has provided reliable results concerning regional and class variation, jargon and register.

The scope of CL research can be expected to continue expanding to cover most of the linguistics disciplines. For Biber et al.(1998), corpus-based methods can be used to study a wide variety of topics including individual words, grammatical features, men's and women's language, children's acquisition of language, author style, register patterns and distribution of features across dialects and time periods. They add that a corpus-based approach "can be applied to empirical investigations in almost any area of linguistics"(p.11).

2.4.2 Applications of Corpus Linguistics in SLA Research: Learner Corpora

A result of the widespread use of computer services worldwide is a growing interest in corpus-based approaches in SLA research. Since it is open to objective verification of results, a corpus-based approach, according to Leech (1992), is a powerful methodology. Another feature that makes corpus study more powerful and plausible than many other approaches is its availability to the public and thus, its ability to be investigated objectively from different angles and for different purposes.

Emphasizing the importance of authentic texts in teaching EFL, de Beaugrande (2001) claims that "learners of EFL, and some non-native teachers of EFL too, suffer not from exposure to non-standard English, but partly from exposure to non-authentic English and partly from lack of exposure to authentic standard English." This argument reinforces the need for CL and corpora in second language learning and teaching. Thus,

learners' exposure to standard, but not authentic materials is not enough to enable them to master the target language. Learners must be exposed to authentic texts to acquire collocations and other grammatical, semantic, discursive and pragmatic features.

However, the divergent themes of linguistics, along with the incapability of general corpora to meet all of linguistics' subfields' demands have pushed the idea of specialized corpora to the fore. This, therefore, explains the existence of what are called *learner corpora*, a collection of texts or essays produced by learners of a language. Engwall (1994) and Hunston (2002), among others, attribute the divergent types of corpora to the divergent objectives and purposes that lie behind creating them. However, producing such corpora has enabled all those interested in the SLA domain to obtain specific and comprehensive information about language learning that has remained unaccounted for in previous literature. Such information includes all kinds of collocations, syntactic structures, word frequency, contextual overgeneralization, word category, etc.

Furthermore, learner corpora have enabled researchers to compare and contrast native and non-native speaker performance- -what is now known in the literature as Contrastive Interlanguage Analysis (CIA, hereafter). Unlike CA (which is based on a comparison between the source language and the target language), CIA, according to Granger (1998:12), involves two major types of comparison:

1. Native language vs. interlanguage, i.e. comparison of native language and interlanguage;
2. Interlanguage vs. interlanguage, i.e. comparison of different interlanguages.

Such studies have provided teachers and researchers with all kinds of learners' errors and areas of weaknesses and also they enabled them to investigate the differences between native and non-native performance. Again, they enabled researchers to examine vari-

ous aspects of learners' developmental stages that were not or hardly accessible via the previous methods. Writing development, for instance, until fairly recently, was primarily measured in terms of the syntactic errors, but now is examined in terms of lexical density, diversity, sophistication, word frequency, word category, etc.

Four obvious indicators concerning the importance of corpora in studying and teaching lexicology have recently arisen in contemporary research. First, it is now possible to see the gradual development of first and second language learners by comparing different corpora that represent different stages of growth or education. Secondly, by providing consistent indications of the high percentage of learners' lexical errors, corpora have contributed to changing the researchers' concern from the extensively studied topics (syntax and phonology) to the least studied ones (lexicology). Meara (1984), cited in Gass and Selinker (2001:372), states that "lexical errors outnumbered grammatical errors by a three to one ratio in one corpus." Yet, based on the preceding sections, it would be possible to state that lexica and collocations in the pre-corpora era were for the most part neglected. Thirdly, unlike the isolated bilingual word lists, corpora provide learners with the context of usage and consequently with syntactic, semantic register and collocational features of a particular word. Fourthly, due to their over-representing of concrete words to the determent of abstract and social terms, traditional intuition-based materials fail to prepare students for a variety of tasks including reading newspapers and report-writing (Ljun 1991, cited in Granger 1998:7). This denotes the preference of text materials based on authentic native English corpora to those traditional intuition-based materials.

Biber et al. (1998:197) argue that the use of learner corpora in SLA research is quite useful in investigating "the frequency and persistence of errors in groups of second language students. Such studies increase our understanding of second language acquisition, provide data for other perspectives on errors (e.g., as interlanguage and non-standard target forms), and provide evidence for instructional decisions".

Hunston (2002: 212) states that using learner corpora in contrastive interlanguage studies has two main advantages:

> Firstly, it makes the basis of the assessment entirely explicit: learner language is compared with, and if necessary measured against, a standard that is clearly identified by the corpus chosen. If that standard is considered to be inappropriate (if, for example, the appropriate target for Norwegian schoolchildren is considered to be expert Norwegian speakers of English rather than British speakers of English), then the relevant corpus can be replaced. Secondly, the basis of assessment is realistic, in that what the learners do is compared with native/expert speakers actually do rather than what reference books say they do. Many of the parameters of difference noted, such as vocabulary range, or word-class preference, do not appear in most grammar books.

Biber (2001) argues that empirical analyses of representative corpora provide reliable information that is often surprising even to TESL professionals. For example, corpora have proved that the use of simple aspect verbs in conversation is more than 20 times as common as the use of progressive verbs. Such a finding, he claims, is surprising to TESOL professionals who, until fairly recently, kept emphasizing the use of progressive verbs in conversation textbooks for a long period of time.

Before going any further, it is important to bear in mind that there are, at least, four reasons that show how CL differs from the traditional approaches:

- its dependence on representative naturally occurring data
- its objective analysis and results
- its dependence on qualitative and quantitative analysis
- its dependence on the artificial intelligence products

2.4.3 Corpus Compiling

A well-compiled and annotated corpus is presumed to provide its users with much more reliable information about the target language than a blind or raw corpus. In as much as corpora depend on evidence or observation rather than intuition, there is

concern with the notion of quantification (representativeness and statistics), which, as will be shown, constitutes the core of corpus-based studies.

2.4.3.1 Representativeness

Recent proposals and results within the corpus framework have revealed that special attention should be paid by corpus linguists to the notion of representativeness, the types of texts comprising the database for a corpus. It is, therefore, necessary to have a corpus that is not restricted to one register or domain. More precisely, the selected texts should come from different fields of knowledge. McEnery & Wilson (1996:22) state that a corpus should respect all aspects of the quality notion:

> In building a corpus of a language variety, we are interested in a sample which is maximally representative of the variety under examination, that is, which provides us with as accurate a picture as possible of the tendencies of that variety, including their proportions. We would not, for example, want to use only the novels of Charles Dickens or Charlotte Bronte as a basis for analyzing the written English language of the mid-nineteenth century. We would not even want to base our sample purely on texts selected from the genre of the novel. What we would be looking for are samples of a broad range of different authors and genres which, when taken together, may be considered to 'average out' and provide a reasonably accurate picture of the entire language population in which we are interested.

The representativeness criterion is not always constant for all corpora. Learner corpora and corpora for specific purposes, for instance, are almost always much more restricted in size as well as type of texts providing their database. For this corpus, the representativeness criterion is reflected in the number and themes of texts providing the database of this study. It should be borne in mind that the principal idea behind representativeness lies in the notion of evidence, and since this corpus is concerned with interlanguage lexicology of Arab Students of English, it is expected to provide evidence relevant to this particular issue and not to the language as a whole. However, if the idea behind compiling this corpus were to produce a dictionary, then the current size and type of texts would be definitely insufficient.

2.4.4 Corpus Annotation

Over the past few decades, there has been ongoing research and progress in corpus annotation, the automatic or manual assignment of tags covering particular information or features of the sampled language. Such tags, as a matter of fact, play a central role in retrieving the data in question. Traditionally, most of the work on annotation has been devoted to the categorization of linguistic information rather than identifying information related to the source, author, genre, register etc. McEnery & Wilson (1996:36-57) distinguishes between eight types of linguistic annotation.

2.4.4.1 Part of speech annotation

Part-of-speech (POS, hereafter) annotation, which aims at attaching to each lexical unit or token in the corpus a code indicating its part of speech, is the most essential foundation for corpus analysis. During the POS enriching phase, a corpus passes through two subsequent stages, viz. tokenization and annotation. During the tokenization stage, a tokenizer breaks the text into tokens and then categorizes each token. Lexical units are then labelled or named (as a result of the POS tagging) according to their contextually-defined word classes.

As far as this study is concerned, the C7 tagset developed by Lancaster University is used. Further information about this tagset is illustrated in chapter 3. The tags themselves are listed in Appendix I.

2.4.4.2 Lemmatization

Despite the different tags assigned to 'sleep', 'slept', 'sleeps' and 'sleeping' at the morphosyntactic level, they are assigned the same tag at the lemmatization level. As a result of this, all variant forms of a related lexical unit are treated as occurrences of

the same unit. Unfortunately, most concordance programs developed so far treat words according to their inflections rather than their lemmas, which can pose limitations on paradigmatically-oriented analyses.

2.4.4.3 Syntactic annotation or parsing

This type of annotation comprises both syntactic recognition and syntactic analysis, assigning constituent structure analysis to the sentence. According to Kennedy (1998:231), parsing involves both annotation and linguistic analysis simultaneously:

> Parsing is a more demanding task involving not only annotation but also linguistic analysis, according to some particular grammatical theory, to identify and label the function of each word or group of words in a phrase or sentence. A word tagged as a noun can function as the subject, object or complement of a verb, for example. A parsed corpus is necessary if we wish to retrieve, say, relative clauses identified by labelled bracketing of the syntactic function of these clauses in texts. Corpora which have been analyzed in this way are often called **treebanks** because they are collections of labelled constituent structures or phrase markers.

In other words, the parsing phase involves the procedure of combining morphosyntactic categories into high-level syntactic relationships with one another (McEnery & Wilson (1996:42). In addition to the syntactic labels (subject, NP, VP), words or tokens (during this phase) get their semantic role annotations (e.g., agent, goal, beneficiary).

2.4.4.4 Semantic and pragmatic tagging

Besides the POS and grammatical annotations, a corpus could also undergo an interpretive analysis to make connections between linguistic reality and extra-linguistic reality. For Leech (1987:12), this level of annotation aims to provide both natural or literary meaning (semantics) and non-natural meaning (pragmatics):

> ...concerned with the assignment of an interpretation or meaning to a text or a part of a text. The distinction between semantics (dealing with uncontextualized meaning) and pragmatics (dealing with contextualized meaning) is not universally accepted in linguistics, but it is a useful division for the purposes of computer text comprehension. Semantic analysis is the assignment of a meaning to a text (-sentence) independently of the local knowledge-resources to which the computer system has access. Pragmatic analysis is the integration of the meaning (as determined by semantic analysis) into those knowledge - resources, including the identification of references, and the modification of beliefs.

2.4.4.5 Discoursal and text linguistic annotations

To keep abreast of all types of linguistic analysis, annotation is not restricted to word or sentence level. Rather, it might involve the entire corpus or text in question. During the discoursal and text linguistic tagging phase, a corpus is enriched with two main kinds of annotations, viz. (i) Anaphoric annotations: the marking of pronoun reference and (ii) Discourse tags: the functions of elements in the discourse: 'good evening': greetings, 'please': politeness, etc.

2.4.4.6 Phonetic Transcription

This type of annotation is peculiar to spoken corpora, and it is usually carried out by persons skilled in the perception and transcription of speech sounds. This means that it cannot be done automatically as is the case for most other kinds of annotations.

2.4.4.7 Prosodic annotations

Like phonetic annotation, prosodic annotation, which is concerned with the sound system above the segmental level, is relevant only for spoken corpora. The London-Lund Corpus (LLC) was the first corpus to have prosodic annotation.

2.4.4.8 Problem-oriented tagging

Unlike all the previous types of annotations, problem-oriented tagging depends solely upon the research's goals and, thus, it is subject to variation from one study to another. The idea behind this type of tagging, which can be applied to a tagged or even raw corpus, is to retrieve the data in question easily using a specific type of codes. Also, this type is restricted to the items in question and not to the entire corpus. As far as this study is concerned, problem-oriented tagging is used for retrieving and establishing frequency count of the lexical and collocational errors found in the corpus.

Despite the availability of several tagging software programs which have been developed over the past few decades, only POS and problem-oriented annotations are employed in this study. The idea behind employing POS tagging stems from the need to provide reliable quantitative and qualitative information concerning the learners' lexical complexity, word-category and text-profiling lexicology (lexical vs. grammatical errors). Furthermore, such tagging makes it possible to compare and contrast this corpus with reference/authentic English corpora.

In sum, the aforementioned sections have outlined different aspects relevant to the status of the lexicon over the past few decades as well as the advent and development of corpus linguistics and corpora, which are considered the best methods ever employed to serve the ambitions of lexicology and lexicography. Overall, a close look at the first two sections (2.2) and (2.3) shows the dramatic shift that has taken place recently in the worldwide concern with lexicology, which, as a result, has become the central issue of language learning. Section (2.4), on the other hand, clarifies the crucial role of CL and machine-readable corpora in lexicon research.

CHAPTER 3

METHODOLOGY

3.1 Introduction

Along with the data analysis procedures, this chapter reports on the corpus compilation method, corpus size, subjects and setting, sociolinguistic variables, native speakers' judgments, data filtering procedures, platform, tools and quantitative analysis measures used in this study. It should come as no surprise from the preceding sections that the preference here for the corpus-based approach over other traditional approaches is due to the objectives of this study, which can be best approached and achieved by emphasizing observation and real-life language rather than intuition and hypothetical data.

3.2 Corpus Method

Attaining the first aforementioned aim, compiling the essay writing corpus of Arab students of English as a foreign language, entails that this study follows the choice method rather than the chance method. Proponents of the choice method, according to Engwall (1994:49), adhere "to a careful definition of the specific subpopulation of natural language that is object of their study." This means that the relevant data is selected methodically to represent the target topic of the study. Chance method, on the other hand, relies on the compiler's own criteria while reading books, newspapers, listening to conversations, etc. In other words, the data of this kind of corpora are collected haphazardly and therefore they lack the representativeness criteria. Engwall cites four criteria for creating a choice corpus: (i) *category* (literary works, scholarly works, newspapers or conversation), (ii) *genre* (imaginative prose, drama, scientific texts or dialogues), (iii) *time* (diachronic or

synchronic) and (iv) *whole texts* or *sample excerpts* (if a whole text is not selected, then sampled selections must be justified and not be haphazardly selected).

On the basis of such criteria, this corpus is academic in terms of category, expository prose in terms of genre, synchronic in terms of time, and sample in terms of text. The subsequent sections will also report on three crucial requirements of corpus design (Biber 1993 and Haichour 1999, among others):

- the language and sublanguage (topics) of the corpus
- the corpus size or constituent texts
- the range of the constituent texts genres

3.3 Corpus Size and Representativeness

Despite the lack of any universal principles pertinent to corpus size and representativeness, corpus linguists tend to agree that a corpus ought to be as vast and representative as possible. Obviously, there have been ongoing changes in the notion of size since the days of the first-generation of corpora. The steady developments in information technology unquestionably have made what was tedious and time-consuming to compile three decades ago quite accessible and easy today. As a result, the first-generation corpora of one million words have become quite small in comparison with the current generation such as the *British National Corpus*, which consists of a hundred million words.

The preceding sections have substantiated that the most obvious starting point in linguistic enquiry within the realm of corpus work is dealing with delineating the types, numbers and percentages of texts providing the database of a corpus. Biber et al. (1998:246) state, "a corpus is not simply a collection of texts. Rather a corpus seeks to represent a language or some part of a language." Oftentimes representativeness is linked to diversity, that is, to provide a balanced corpus that is not restricted to one type or field. In other words, the data should come from different domains or fields of

knowledge (Biber 1993). A related crucial issue in corpus composition is the weighting among different corpus constituents.

Corpus size and representativeness, as the literature shows, are subject to two criteria: (i) corpus objective(s) and (ii) availability of resources. Learner corpora and corpora for specific purposes, for instance, are almost always smaller in size than general corpora simply because they intend to serve and represent specific aspects of a given language and not the language as a whole as the general corpora do. Lexicographic corpora, whose aim is to provide a dictionary of a language, are expected to be much larger than any other corpora developed for specific purposes.

3.4 Subjects of the Study

In the course of data collocation, 450 Jordanian undergraduate students enrolled in their second to fourth year of English Language and Literature at five Jordanian universities volunteered to participate in this study. The native tongue of all the subjects, who, at the time, were considered to be at the intermediate to advanced level of English proficiency, was Arabic. Each of the subjects participated in at least one of the tasks set for this study, namely, the essay writing task (from which learner corpus was compiled), the translation task, or the collocational task.

The subjects were told before their voluntary participation that no information jeopardizing their anonymity directly or indirectly would be included in the corpus. Additionally, they were told that their participation and answers would not affect their grades and would never be used for course assessments.

Learner and task variables used in this corpus (as shown in Figure 3.1) are quite similar to those used in *The International Corpus of Learner English* (Granger 2003).

As is apparent from Figure (3.1), the learners who contributed to the current learner corpus have three characteristics in common and they differ in three other respects, as

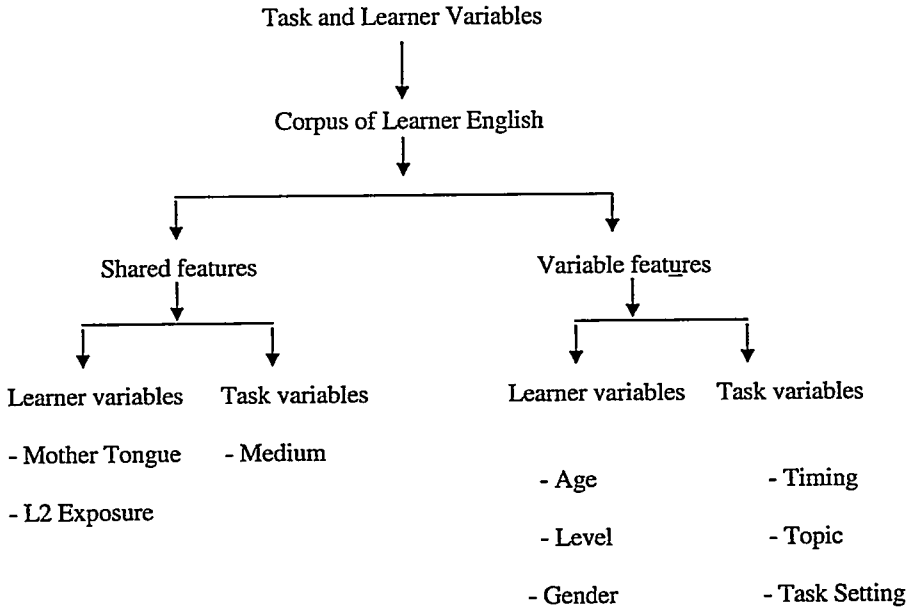


Figure 3.1. Learner and task variables in the learner corpus.

well. All learners are native speakers of Arabic and have never lived in an English-speaking country prior to their participation in this corpus collection. In addition, they are all majoring in English Language and Literature.

As far as the variations are concerned, learners differ slightly in age. However, this divergence is minimal since more than 98% of them are between 19-23 years old (See Appendix H). Likewise, they differ in their level; the word *level* is used here to refer to the learner's year in the major (1st, 2nd, 3rd or 4th). Also, learners differ by gender (males and females).

Though the entire corpus consists of only written essays, there are, however, some divergences in timing (timed vs. untimed essays), conditions (exam conditions vs. home-work assignments) and topics (wide variety of topics are used). Further details on these variables are given below.

Generally speaking, the motivation behind majoring in English for Jordanian students is chiefly instrumental- -to get a good job. Students start learning English at the age of 12 (fifth grade) at public schools with five English classes conducted weekly. However, a couple of years ago, a decision was made to start teaching English at the age of six (first grade). This means that, prior to the university or college phase, students would have studied English for at least eight years.

English is one of the core courses throughout school study, and it makes up 30% of the total grade of the General Secondary Education Certificate (GSEC, hereafter). No student can enter a university or community college program unless s/he has passed the GSEC. This criterion insures that mastering a considerable level of English is a prerequisite to a university or even a community college education. At the university or community college level, students are also required to pass at least one course in general English. Students majoring in English Language and Literature are expected to complete approximately 100 credit hours in linguistics, translation and English language and literature.

3.5 Setting of the Study

The data of the entire essay-writing corpus, along with the translation and collocations corpora, were conducted in the departments of English Language and Literature at five public and private Jordanian universities countrywide. These included Al-Hussein Bin Talal University, Hashemite University, Mutah University, Al al-Bayt University and Zarqa National University.

Al-Hussein Bin Talal University is a newly-established public university located on Ma'an city in the southern part of Jordan. Hashemite University is a public university located in the outskirts of the city of Zarqa (about 15 miles to the north-east of the capital Amman). Mutah University is the third largest Jordanian university. It is located in the

Karak governorate in the middle of Jordan. Al al-Bayt University is a public university located on the outskirts of the city of Mafraq to the northeast of the capital Amman. Finally, Zarqa National University is a private university located on the desert highway (about 10 miles to the northeast of the capital Amman).

3.6 Data Gathering Procedures

The corpus providing the database of this study consists of three components comprising about 130,000 tokens. Given the range of goals for the research, these components or corpora were collected by various techniques to serve the overall aims of the research.

3.6.1 Learner Corpus

The goal of this component, which provides 70,307 tokens (54.08% of the entire data), was to collect representative samples of the subjects' writing and get reliable statistical information concerning the learners' lexical richness, lexical and collocational errors, words frequency and the other most salient features of the learner corpus (e.g., word category, overproduction, underproduction). This part is made up of 269 timed and 160 untimed essays, tests and homework assignments; the untimed essays were written as homework assignments without a strict time limit. Timed essays, on the other hand, were written under exam conditions.

Prior to writing the untimed essays, the subjects were asked to write, at minimum, a 500-word essay on any of the given topics or on any other topic they might come up with. Having known that their participation would neither affect their grades nor be used for course assessments, the subjects were discouraged from worrying about errors or mistakes and thus, discouraged from accessing reference tools.such as grammar books or dictionaries.

The essays and tests comprising the database of this part are of different themes (e.g., everyday activities, opinions, literature). To uphold the principle of representativeness as much as possible, the subjects of the study were given a guide list of topics to help them choose the topics of their writing. The subjects were not restricted to these topics, however. It should be mentioned here that some additional topics also came from the instructors of writing and literature courses (See Appendix F for a complete list of the topics used in the essay writing corpus). Diversity, in terms of writing framework types, was given due attention, too. As we look at the list of the topics of the corpus, we see samples of narrative, argumentive, literary, procedural and explanatory texts.

3.6.2 Lexical Translation Corpus

Reliable quantitative testing of lexical choice errors required the creation of another corpus where 300 lexical items that represent different fields of knowledge were deliberately selected. The purpose of this part was to get specific information about the subjects' knowledge of common lexemes in the L2. For purposes of ease and simplification, this test was divided into fourteen forms (with 15 to 24 target lexical items in each form). Each form was translated by fifteen students who, by then, were in their second to fourth years in the English major. The entire test was administrated in class under exam conditions. To avoid ambiguity, vagueness and imprecision, each of the target lexical items was contextually given in full and meaningful sentences (See appendix B).

To avoid testing the same subject twice, the test was administrated in several classes that would meet simultaneously. However, since there were sometimes an insufficient number of classes that would be meeting at the same time, another procedure was employed whereby several classes were selected at different times when one of the target courses is a prerequisite to the other. So, to the maximum extent, no student attended both classes together. In this component and the subsequent one, learners' errors were

first tagged with different labels (carrying the initials of errors categories) and then a quantitative analysis was performed.

3.6.3 Collocational Corpus

This component consists of two parts which aim at measuring the subjects' knowledge of and ability to use English lexical collocations. To this end, guided and unguided tasks were developed and employed in this study. The word *guided* is used here to distinguish the carefully designated tests, whereby the subjects have limited options, from the free writing used in the learner corpus. Variation in terms of the selected nodes, subjects and techniques was given due attention over the entire study. Except for the free writing collocations, each collocational item in the guided task was attempted by fifteen students.

3.6.3.1 Guided Collocational Task

Three main procedures, namely, translation, multiple choice, and semi-cloze and cloze tasks were employed to achieve the goals of this section. In order to avoid any post-test effects, special attention was paid to the sequential organization of these tasks, which share the same collocational items. To this end, two groups of students participated in carrying out this section. The first group was first asked to translate into English the given sentences and then to do the multiple choice task. The members of the second group, on the other hand, were only asked to do the cloze and semi-cloze tasks. These tasks were strictly conducted in the following sequence.

(i) Translation Task

Again, this section consists of two parts. The first part, on the one hand, consists of ten different widely used Arabic collocations, where the noun is the node and the adjective is its collocate. While the ten different lexical nodes collocate with ten different adjectives in Arabic, they, in order to convey the equivalent lexical meaning, collocate with one adjective in English. In other words, the ten different Arabic adjectives should be translated into one word in English- -the adjective *heavy.* Seven of these collocations were previously checked against authentic Arabic and English dictionaries (Heliel 1989). The other three were also checked here against the *Oxford Collocations Dictionary for Students of English*, along with *Al-Mawrd Dictionary* (a bilingual English-Arabic and Arabic-English Dictionary) (See Appendix C (part A) for a complete list of these collocations). The second part, on the other hand, consists of ten sentences with ten different collocate-node pairs in the source and the target languages (See appendix C (part B)).

(ii) Multiple-Choice Task

This section consists of the same sentences given in the lexical translation task. After the researcher had translated them into English, the subjects were asked to determine the missing collocate by choosing one of the four alternatives provided after each sentence (See appendix D parts A & B). A double check of the translation of both Arabic and English was made by two native speakers of both languages.

(iii) Cloze and Semi-Cloze Task

This two-part task consists of 10 sentences (the same sentences used in part (B) of Appendix C). As seen in the second part of Appendix F, the target collocations in this section are of different nodes. In the first five sentences, the subjects were asked to identify the missing collocate, knowledge of which, however, requires a minimum exposure to the

target language on the part of the examinees. The semi-cloze test, on the other hand, consists of 5 sentences. Each of these sentences has a missing collocate (see Appendix F part B). To narrow down the possible answers, the first letter of the missing collocate was given. So, in addition to the context, the examinees were provided a helping clue to identifying the intended collocate.

3.6.3.2 Unguided Collocational Task

This section examines the learners' use of collocations in their free writing. In so doing, 100 node-collocate pairs were investigated in the learner corpus. Table (2.1) presents the part of speech combination and their frequencies.

Table 3.1. Word categories of the nodes and collocates investigated in the learner corpus

Node	Collocate	Frequency
adjective	noun	35
noun	adjective	35
noun	verb	15
adverb	verb	15

The extracted collocates were tabulated into two columns. While the first one includes the correct collocational uses, the second one includes the incorrect ones.

3.7 Lexical Knowledge vs. Collocational Knowledge

Learners' collocational knowledge ought not be separated from their lexical knowledge. Support for this assertion comes from numerous instances of collocational blending, where a learner, due to the lack of the appropriate lexical item rather than the ignorance of the collocability between the items in question, successfully uses either the node or the collocate but not the two together. In order to map out the percentage of colloca-

tional errors relative to the lexical errors, a statistical comparison between the results of the lexical translation component and the guided collocational task was conducted. The total number of expected lexical and collocational items is illustrated in Table (3.1). It is important to mention that lexical errors refer to content word errors rather than function word errors.

Table 3.2. Percentage of collocations relative to lexical items

No.	Type	Total number of items
1.	Lexical words	4500
2.	Collocations	750
3.	Percentage of collocations relative to lexical errors	16.67%

3.8 Data-Filtering and Constraints

All writing samples and tests employed in this study have met the requirements of data selection. So, to the maximum extent, no part of this corpus has violated any of the following constraints:

1. Only students majoring in English Language and Literature at Jordanian universities whose native tongue is Arabic and who have never lived in an English-speaking country were eligible to participate in this study.
2. Writing samples that were illegible or irrelevant to the assigned task were immediately eliminated. Additionally, untimed essays (those assigned as homework) that sounded inauthentic (not written by the learner) were excluded, too.

Prior to running the learner corpus on *Wordlist* software, all spelling errors were corrected. The reason behind doing this was to avoid obtaining incorrect results in terms of lexical diversity (type-token ratio) if such errors were left uncorrected. Again, this kind of errors, if left uncorrected, would also affect lexical density.

3.9 Sociolinguistic Variables

The subjects' gender and level of education (See the demographic questionnaire (Appendix g)) were the only two sociolinguistic variables considered during the course of compiling the three corpora. Though gender plays no relevant role in this particular study, such a variable might be important for any future studies utilizing this corpus.

3.10 Native Speakers' Judgement

Though corpus linguistics relies heavily on the notion of evidence rather than native speaker's intuitions, this corpus takes advantage of both; the native speakers' job in this study was exclusively devoted to examining the lexical and collocational errors after being sorted out and categorized by the researcher. It should be obvious here that there are often differences between the native speaker's intuition and the corpus' usage. For example, whereas the native intuition prefers *if I were* instead of *If I was*, several engines (e.g., Google) show that *if I was* sometimes outnumbers *if I were* in several domains.

3.11 Quantitative Analysis

It is hopefully apparent from the previous discussion that the ultimate goal behind the compilation and annotation of corpora is to provide accurate and reliable descriptions of how languages are structured and used. In so doing, one can conclude that corpus-based analysis depends primarily on the quantitative techniques of analysis or what is

often referred to as probabilistic analytical methods. In spite of the sharp criticism that such methods have encountered (e.g., Chomsky 1962), corpus-based statistical methods remain the most convenient ways to describe and understand language structure and use (Hladka 2000:3).

The most widespread corpus-based methods are the statistical (or probabilistic) methods. The statistical methods offer a good theoretical background, an automatic estimation of probabilities from data and a direct way to disambiguate the particular information. It is also worth adding that the growing interest in quantitative studies goes beyond the identification of the most frequent or rarest entities to provide researchers with reliable information (e.g., on the interactivity between lexemes and genres) and to entreat that bad or unscientific guessing never sets foot in analysis. For Feynman et al. (1963:6-1) the growing tendency of using statistics is mainly employed to avoid guessing and to provide justification for claims:

> By chance, we mean something like a guess. Why do we make guesses? We make guesses when we wish to make a judgment but have incomplete information or uncertain knowledge. We want to make a guess as to what things are, or what things are likely to happen. Often we wish to make a guess because we have to make a decision. For example: Shall I take my raincoat with me tomorrow? For what earth movement should I design a new building? Shall I build myself a fallout shelter? Shall I change my stand in international negotiations? Shall I go to class today?
> Sometimes we make guesses because we wish, with our limited knowledge, to say as much as we can about some situation. Really, any generalization is in the nature of a guess. Any physical theory is a kind of guess work. There are good guesses and there are bad guesses. The theory of probability is a system for making better guesses. The language of probability allows us to speak quantitatively about some situation which may be highly variable, but which does not have some consistent average behavior.

In this study, statistics plays a central role in all kinds of lexical analysis (lexical diversity, lexical density, lexical errors, etc.). The findings of this study are compared and contrasted with reference corpora to provide crucial information pertinent to word frequency, overuse of words, richness and poverty of lexicon, etc. The t-Test and the

automatic statistical analysis carried out by *WordSmith* were employed in analyzing this corpus.

3.12 Data Processing and Analysis Procedures

The past four decades have witnessed giant strides in the development of tools used in compiling, retrieving and parsing corpora. One of the strengths of modern corpora is the quantity of being machine-readable, which makes corpora more accessible to all users. Doubtless, the long days that one might spend in compiling and computerizing a corpus are relatively minor in comparison to the tedious analytical procedures that followed. Of critical importance at this stage is to bear in mind that data analysis procedures in corpus linguistics do not usually start as soon as corpus compiling and computerization is done. Oftentimes, there is a transitional enriching phase, during which the raw corpus is tagged and/or parsed. What determines this intermediate phase is solely the research objectives. Fortunately, this a phase, which was the most exhaustive phase several years ago, has become the easiest one due to the recent development in artificial intelligence products.

Data analysis in this study was divided into two phases. The first phase precedes annotation while the other one comes after annotation.

1. Pre-Tagging Phase

The inability of raw corpora to provide some additional information that tagged corpora can provide should not call into question their validity; raw corpora still provide learners and researchers with insights that would otherwise be impossible or at least difficult to obtain. Information pertinent to word frequency, word diversity, lexical fluency and lexical and collocational errors, which require no additional tags, are better provided

by raw corpora.

(i) Word Frequency

It has been long noted that the principal format used historically in displaying linguistic elements in a corpus is by means of listing and counting (Kennedy 1998:244). Software technology makes it possible to display corpus contents in three different forms, namely, alphabetical order, frequency order or appearance order. For convenience, all the data of this corpus were displayed in frequency order. However, for partial comparative goals, alphabetical order was also employed. For the purpose of this study, *Wordlist*, one of *WordSmith's* tools, and the *Frequency Indexer*, developed by Catherine Ball at Georgetown University, were used.

(ii) Lexical Diversity

The availability of software programs concerned with quantitative analysis, as noted earlier, has explicitly affected the direction of much new linguistic research. Fortunately, lexicology has been a major beneficiary in this regard. This explains the frequent use of a variety of lexical measures (e.g., lexical diversity, lexical density, lexical sophistication) in much of the recent research conducted on lexicology worldwide (e.g.,Granger 1998).

As far as this study is concerned, lexical complexity, an umbrella term for both lexical diversity and lexical density, was used as a quantitative measure of learners' lexical richness in comparison with the NSs. Lexical diversity, a measure of the spread or richness of the vocabulary in a text, requires no annotations and thus is carried out prior to POS tagging. This measurement is calculated according to the following formula:

$$\frac{\textit{the number of types (different words)} \ X \ 100}{\textit{the number of all tokens (instances of each word)}}$$

(iii) Lexical Fluency/Proficiency

Composition length has been used relatively recently in numerous studies as a reliable measure of learners' proficiency or fluency (Larsen-Freeman and Strom, 1977, Larsen-Freeman 1978, Linunard 1986, Reid 1990, Engber 1992, Wright 2000, among others). Following Engber (1992), the length of each essay was measured in orthographic words, a string of alphanumeric characters bounded by spaces.

Though most of the subjects of this study have minimal differences in terms of their exposure to L2, diversity, in terms of their fluency in the L2 writing, is expected for different reasons, not the least of which are the learner's aptitude and motivation towards the L2. Some researchers have pointed out that some learners face a serious lexical problem in retrieving the lexicon they already know.

(iv) Lexical and Collocational Errors

Extraction of lexical and collocational errors via the *WordSmith'* concordancer preceded the tagging process. However, in numerous cases it is better to access a tagged corpus for extracting collocations, especially the grammatical ones.

2. Post-Tagging Phase

The information obtained from tagged corpora depends on the type of tags that a corpus has already received during the enriching phase. It is hopefully apparent from Chapter Two that there are various kinds of tags that we can supply a corpus with during the enriching phase (e.g., POS tags, semantic tags, phonetic tags). As far as this corpus is concerned, only POS and problem-oriented annotations have been used.

(i) Lexical Density

Unlike the proficiency measure, lexical density seems to be much more consistent and well-established in the literature (particularly in measuring the differences between spoken and written discourses). Lexical density is calculated according to the following formula: $\frac{\textit{the total number of content words X 100}}{\textit{the total number of all tokens in the given corpus}}$

(ii) Word category

A great deal of recent research on corpus linguistics has centered on characterizing texts according to word categories. Thus, it has become possible to investigate various aspects of language (grammatical, discoursal, lexical, etc.). It is crucial to know that many of the aspects concerned with word categories remained unaccounted for, at least in large corpora, in all of the methods that dominated the linguistics scene during the last century. In addition to all the major word categories, this study devotes special attention to coordinating conjunctions, subordinating conjunctions, pronouns and articles.

To sum up, the analytical procedures of this study were carried out in the following sequence:

1. Compiling and computerizing the essay-writing corpus, lexical translation corpus, and collocational corpus according to the aforementioned criteria.
2. Establishing an automatic frequency count of the reference as well as the learner corpora.
3. Extracting and tabulating lexical form errors.
4. Correcting all lexical form errors.
5. Establishing (another) automatic frequency count of the reference as well as the learner corpora.
6. Comparing and contrasting the frequency count findings in the learner corpus with those of the reference corpus.

7. Examining, via *WordSmith* tools, lexical diversity in the learner as well as the reference corpora at both corpus level and individual level.
8. Examining the lexical size in the learner and reference corpora. In so doing, it was possible to examine the mean values as well as the standard deviation in both corpora.
9. Extracting the lexical and collocational errors.
10. Obtaining the native speakers' judgments concerning lexical and collocational errors.
11. Providing taxonomies for lexical and collocational errors.
12. Providing part of speech annotation for the essay-writing corpus as well as the reference corpus.
13. Examining lexical density in learner and reference corpora.
14. Providing qualitative as well as quantitative analyses for lexical and collocational errors presented in both lexical translation and collocational corpora.

3.13 Data Computerization

3.13.1 Data Entry

Once the data are selected and collected, the second phase, the conversion of the non-electronic data into an electronic form, begins. Typically conversion is accomplished by one of three methods (Sinclair 1991:14):

(i) Adaptation of material already in electronic forms;

(ii) Conversion by optical scanning (machine reading);

(iii) Conversion by keyboarding.

Since all the texts of this study are in handwritten form and due to the large body of errors expected to occur as a result of scanning, keyboarding was the only method used for inputting the texts of the corpus.

3.13.2 Platform and Tools

Since *WordSmith*, an integrated suite of software programs, does not run on the *Apple Macintosh*, *Windows* was used as the platform for this study. Inasmuchas the tools of the *WordSmith* software perform varied functions (e.g., concordancing, wordlisting, splitting, text converting, controlling) no additional software programs were needed to accomplish the purposes of this study. However, for comparison purposes, other software programs (e.g., Georgetown University *Frequency Indexer* and *Simple Concordance Program*) were used.

A concordance, according to Sinclair (1991:32-35) "is a collection of occurrences of a word-form, each in its textual environment." In a previous work, Sinclair (1986) states that the use of concordancing programs helps to provide "explanations that fit the evidence, rather than adjusting the evidence to fit a preset explanation" (p. 202). Although it is closely connected with computer-based studies, the actual use of concordancing in linguistic research dates back to the 13th century (Tribble and Jones 1990:7). However, the use of concordancing in its current sense is relatively new. The heavy reliance on concodancing in corpus-based studies perhaps makes it the most important of all the software tools used in the corpus analysis. One of the most well-known formats for concordancing in the literature is what has been termed the KWIC (Key Word in Context) in which the key word appears at the center of the page with a designated number of characters to the right. *WordSmith's* concordancer makes a concordance using DOS, Text only, ASCII or ANSI text files. This concordancer has the ability to:

- make concordances of a search-word

- find collocates of the search-word
- display a map plotting where the search-word occurs in each text file
- identify common phrases (clusters) in the concordance e.g., "give it up"
- show the most frequent words to left and right of the search-word

Wordlist, which is one of the three main tools in the *WordSmith* software package, generates word lists on one or more ASCII or ANSI text files. This tool has the ability to:

- generate word lists based on one or more text files.
- generate individual word lists or batches of them to save time.
- display word lists in alphabetical and frequency order.
- carry out lexical comparison of two texts.
- provide output for use by KeyWords.

As for the POS tagging, this study has utilized the current standard C7 Tagset (in CLAWS). C7 tagset consists of 137 tags (See Appendix I for a complete list of the part of speech tags used in this Tagset). As we end this chapter, it is best to mention that some variations are usually noticed by running the same data on the same software if such data are saved as different file formats (Word document, Text.file, etc.) This is the reason for saving the entire data as plain text files.

To sum up, this chapter has delineated the methodological procedures employed in the study. As can be seen, a deliberate attempt has been made to minimize fragmentation and maximize clarity and focus. As such, a strict adherence to the data gathering and analysis procedures stated in this chapter is expected to achieve the goals behind the research in an insightful manner.

CHAPTER 4

LEXICAL COMPLEXITY AND TEXT-PROFILING RESULTS AND DISCUSSION

4.1 Introduction

This chapter is designed to present and explain in a step-by-step way the outcomes of the first three research questions concerning learners' lexical complexity and text-profiling. For the sake of organization, the chapter is made up of three sections, which appear in exactly the same order as the research questions posited earlier. The results of each question are addressed with reference to the findings of previous literature.

4.2 Results Related to Research Question (1)

Research Question (1): To what extent does the learner corpus deviate from the reference corpus in terms of lexical complexity?

Due to the extremely frequent occurrence of the lexical errors in learners' inter-language and the apparent negative effect they have on communication between senders (speakers or writers) and receivers (listers or readers) (cf. chapter 2), lexical errors are considered the worst among all types of learners' errors. Thus, the availability and accessibility of a considerable body of lexicon, as the literature shows, is vital for learners' productivity and communication. All measures of fluency, accuracy and complexity, according to Wolfe-Quintero et al. (1998:106), relate primarily to the lexicon. This indicates the centrality of the lexicon and of lexical complexity, in particular, in language learning and proficiency.

Following Li (2000), *lexical complexity* is used in this study as an umbrella term for both *lexical diversity* and *lexical density*. For this reason, the results of this part are presented in two subsections 4.2.1 and 4.2.2. While *lexical sophistication*, the ratio of sophisticated word types to the total number of word types, is often included under the umbrella of lexical complexity, this study, for convenience, is limited to exploring the first two measures and will not consider lexical sophistication.

4.2.1 Lexical Diversity

A critical factor adversely affecting lexical diversity is corpus size/length. So, in order to avoid its converse role when the analysis is carried out on individual essays, which vary in their length, this measure was carried out on a full corpus basis (equal basis). Figures (4.1) and (4.2) present the findings of lexical diversity in both the learner and reference corpus respectively.

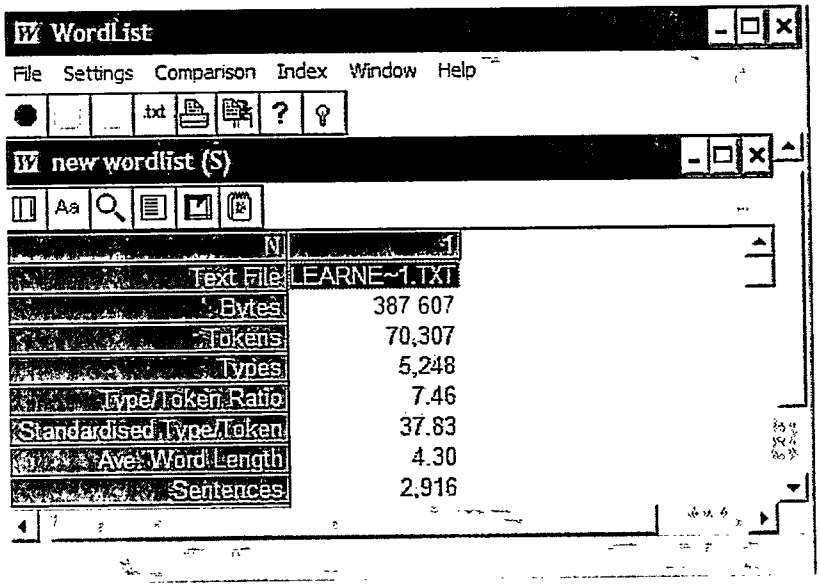


Figure 4.1. Type-token ratio in the learner corpus.

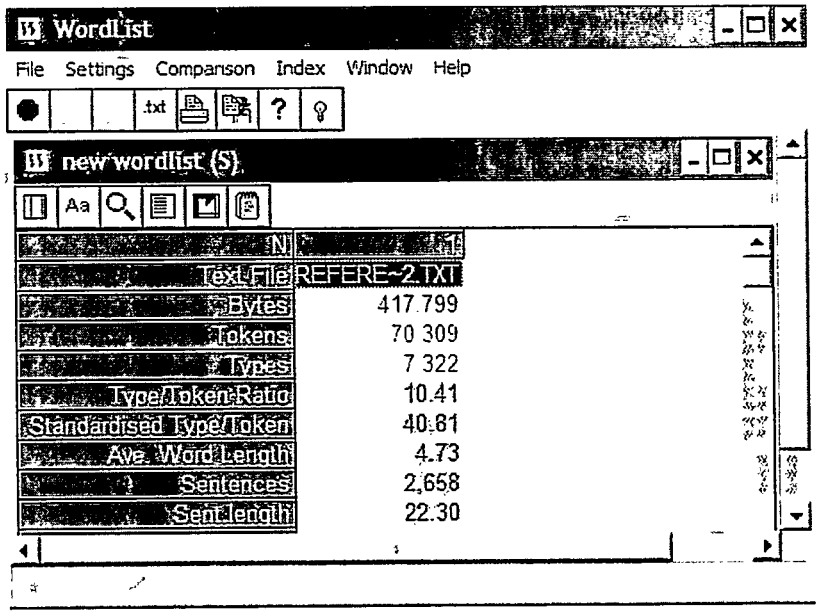


Figure 4.2. Type-token ratio in the reference corpus.

As shown in Figures (4.1) and (4.2), this study found, in regard to lexical diversity in the learner corpus in comparison to the reference corpus (7.46 vs. 10.41), that differences are highly suggestive. The substantial disparity in the number of the types (unique words) shown above properly indicates that the lexical diversity in the reference corpus exceeded considerably the learner counterpart (7,322 vs. 5,248). While it was not unexpected for the type-token ratio in the reference corpus to outnumber the learner counterpart, the marked percentage of diversity (28.3%), which favored the reference corpus, goes far beyond expectations. By taking into account the five times higher diversity of the learner corpus over the reference corpus in regard to the number of the subjects (See Appendix F), and three times higher in regard to topics and themes, the findings become potentially striking. However, a look at all of these results, together with the findings of previous research, reveals the learners' limited word stock and their excessive reliance on repetitive lexemes and patterns to convey messages in the target language.

Are we to ascribe the disparity in type-token ratio only to the learners' developmental stages for such a wide disparity in type-token ratio? Such a conclusion would be unwarranted if we proceed and consider three crucial aspects: (1) the oversimplification of L2 input, (2) the high tolerance of learners' lexical errors (particularly those resulting from near-synonyms, word match, and incorrect collocations), and (3) the rhetoric of learners' L1, where repetition or the *beating around the bush strategy*, see below, is a prevalent rhetorical device. While the first two points deal with the prevailing L2 teaching strategies, the third one, on the other hand, reflects the unconscious transfer of the linguistic and cultural rhetorical patterns and strategies to L2.

Oversimplification of L2 lexicon, which is harshly criticized by Fox (1979:68), often leads learners to rely on a limited number of lexical items that could be incorrectly used to stand for other presumably difficult lexical items of which they share a common meaning. However, oversimplification is not limited to lexicon. Rather, it is extended to other aspects of L2 (e.g., syntax). Building on previous research (e.g., Blau 1982:525, Yano et al. 1994), Oh (2001) argues that input simplification has several problems. First, the use of limited vocabulary and short and simple sentences is likely to result in "choppy, unnatural" discourse that may deviate significantly from authentic texts. Secondly, the elimination of unfamiliar linguistic items shields learners from exposure to items that they should learn. Consequently, learners who experience an oversimplified lexicon would graduate with a minimal word stock and an apparent inability to deal with authentic texts or discourses. The high tolerance of L2 lexical errors is of no less destructive effect than oversimplification. Given these aspects in addition to the universal learners' repetitive dominant tendency, the gap between the two corpora in terms of type-token ratio may be more understandable.

Although the sizable gap between the two corpora in terms of lexical diversity is evident, in favor of the reference corpus, this ratio might be misleading if such analysis

were carried out on different sized corpora. To be more precise, lexical diversity is strongly influenced by size; the more tokens a corpus has, the less diversity it possesses. For this reason, no analysis would be more accurate than one carried out on the equality basis. By displaying type-token ratio carried out on individual essays, which vary in their length, Table (4.1) provides powerful evidence for the sensitivity of lexical diversity to size. The means of type-token ratio shown below have been computed with the *Wordlist.*

Table 4.1. Means values of lexical diversity (carried out on individual bases) in learner and reference corpora

	L.C.	R.C.	Difererence[a].
Mean	56.43	42.70	13.73**
SD	7.0098	7.886	

[a]**The difference between the two corpora is significant at the $\alpha = 0.05$ *level* ($t = 14.3907, p <$ .0001) using two-sided parametric t-test assuming equal variance

However paradoxical it might appear at first glance, the learners' lexical diversity ratio is significantly outnumbers that of the native speakers. Yet, this would be an ungrounded conclusion and should be immediately questioned since it ignores the size of the essays, which plays a significant role here. Knowing that the mean of learners' fluency/essay length (illustrated below) is 163.89 tokens and the mean of the native speakers' fluency is 889.99 tokens, then, it becomes obvious that lexical deversity and size almost always stand in an inverse relation and this is why we see the reverse ratios shown in Table (4.1). (As a precaution against overgeneralizing, it should be noted that evidence from the literature indicates that there are some cases where lexical size and lexical diversity correlate. This occurs when much of the text in question includes a list of items, such as names of persons, machines, items on restaurant menus, etc.).

But in order to determine if this argument carries any weight, it needs to be examined more closely. In order to do so, one essay from the reference corpus was run on the *Wordlist* a second time. The first time the entire essay (1,362 tokens) was run, while only 500 tokens (36.68% of the entire size of the essay) were run in the second time. The results are shown in Figures (4.3) and (4.4).

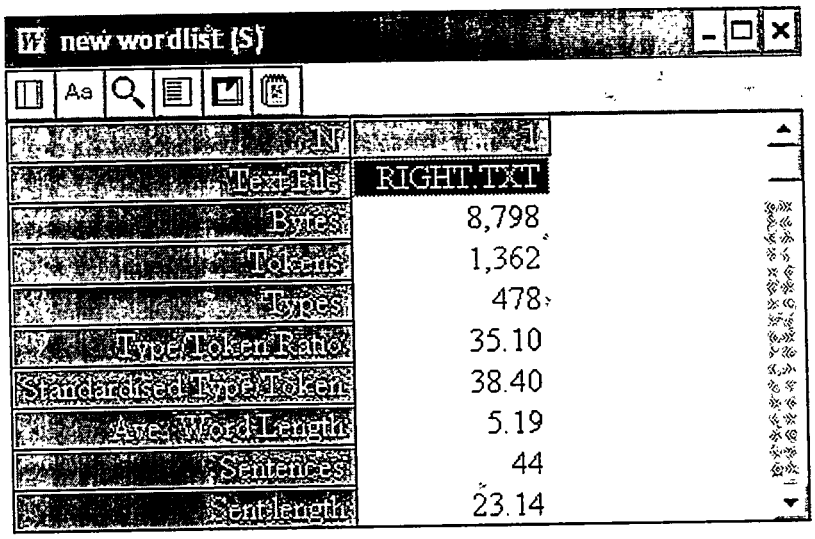


Figure 4.3. Type-token ratio in a 1,362-token essay.

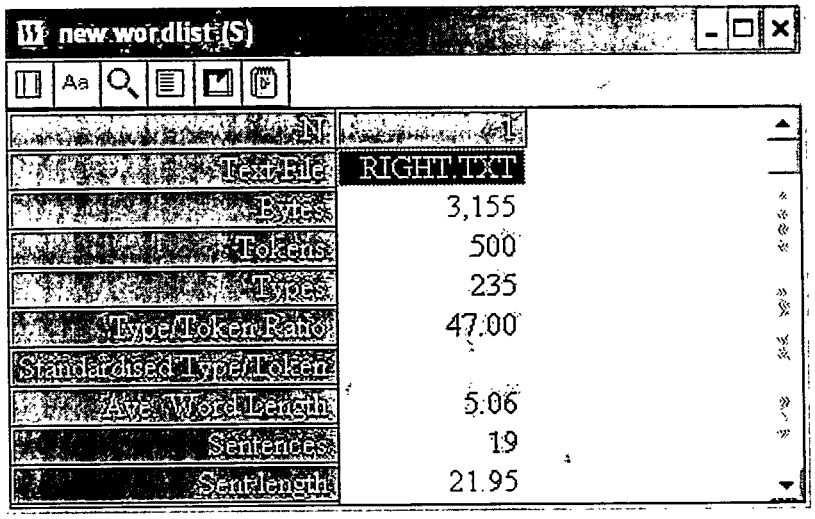


Figure 4.4. Type-token ratio in a 500 token essay.

Figures (4.3) and (4.4) furnish unambiguous support for the initial conclusion concerning the role of size on type-token ratio. The ratio of type-token in the 500 token essay is 47% while it decreases markedly to 35.10% in the 1,362 token essay. Such findings cast doubt on the conclusion of some previous studies (e.g., Miller: 1981), which argued for the independence of type-token ratio and sample size.

A question that arises here is whether the findings of this study agree with the findings of the previous research. The answer to this question would, to some extent, be affirmative. Menunier (1998:32) argues that "A 1,000-word corpus may have a type/token ratio of about 40 percent, whereas a 100,000-word corpus of essays may have a type/token ratio of about 10 percent". Empirical support for these findings, at least those related to the reference corpus, comes from Barnbrook (1996). After running the entire text of the novel *Frankenstein*, which is made up of 75,214 tokens, on a frequency indexer, Barnbrook found that the type-token ratio in the text is 10.8 (compared with 10.41 in the current reference corpus) and the unique words are 6,986 (compared with 7,322 in the current reference corpus). By comparing the findings of the *Frankenstein* corpus

analysis with the findings of the current reference corpus, which is almost 5,000 tokens smaller, it is clear that the reference corpus outscores *Frankenstein* corpus in terms of lexical diversity. However, this percentage of divergence, which lies within the expected range, is attributed to the repetitive use of names and patterns in the literary texts, in general, and also to the diversity in themes or topics in the reference corpus.

In the context of the interactivity between size and type-token ratio, it is relevant to mention that there are 669,505 different word forms in the *British National Corpus*, (100,000,000 word corpus). This indicates that the type-token ratio in this huge corpus is 0.67% (less than 1%) (Hoffmann and Lehmann 1998:17).

Research on learners' lexical diversity, which is still in its infancy, shows no significant relationship between learners' level and word variation (Cumming and Mellow 1996). What makes most of the findings of previous studies rather difficult to compare with the findings of this one is a difference in size. It is appropriate, at this juncture, to question whether this measure, lexical diversity, has any value. According to Wolfe-Quintero et al. (1998:106), there are two problems with this measure.

1. It does not discriminate between a writer who uses a few types in a short composition and a writer who uses more types in a longer text.
2. It does not respond appropriately to length of the sample; the scores gets lower as a text gets longer since the types repeat more often.

The conclusion drawn from the above discussion merely confirms that size plays a negative role in the diversity literature and thus, the preference for carrying out this measure on the entire corpus rather than on individual essays was justified and necessary to avoid the aforementioned pitfalls. Also, the mentioned above results show a large gap between NNSs and NSs in terms of lexical diversity.

4.2.2 Lexical Density

Results pertaining to lexical density, which is calculated by dividing the total number of content words X 100 by the total number of all tokens in the given corpus, in the learner as well as the reference corpora are reported in Table (4.2). Unlike lexical diversity, which is extremely sensitive to the size or length notion, lexical density is completely independent of size (McCarthy 1990). This entails that an individual-by-individual analysis was needed to get reliable results.

Table 4.2. Mean of lexical density and standard deviation in learner and reference corpora

	L.C.	R.C.	Differerence [a]
Mean	44.75	47.51	2.76**
SD	4.85	5.00	

[a]**The difference between the two corpora is significant at the $\alpha = 0.05$ *level* ($t = -4.6311, p < .0001$) using two-sided parametric t-test assuming equal variance.

Owing to its insignificance as a discriminating measurement between the interlanguage of the NNSs and the language of the NSs, and also between different stages of learners' development in much of the previous literature, the debate over the reliability of lexical density has not yet been settled. However, this measurement has been typically and successfully used as a discriminating factor between spoken and written texts.

Lexical density percentage, according to Ure (1971), generally tends to be over 40% in written texts and less than 40% in the spoken ones. By contrasting written and spoken versions of one and the same text, Eggins (1994:61) furnished reliable support for Ure's argument. Lexical density, according to the findings of her study, was 9% higher in favor of the written text (33% vs. 42%). Knowing that the percentages of lexical density

in the learner corpus and the reference corpus were 44.34% and 46.93% respectively, it becomes manifest that both corpora support Ure's findings in this respect.

In his article *A Window on Lexical Density*, Beber-Sardinha (1996) raises several interesting and valuable points concerning lexical density in speech and writing including the influence of nominalization and redundancy. By examining lexical density in intervals (not the whole text), Beber-Sardinha found that dialogues "had very high portions, contrary to what the ratios for the whole text would suggest" (p. 1).

When it comes to comparing and contrasting the reference corpus and the learner corpus in regard to the literature, the reliability of this measure becomes weaker simply because of the nearly identical results found in literature. Yet, this is not to deny the existence and validity of such a measure in lexical studies. As far as the findings of this study are concerned, learners have a lower percentage of lexical density than native speakers, as illustrated in Table (4.2). The percentage of diversity between means (though statistically insignificant) is not unprecedented in literature. Linnarud (1986) found that native language speakers had higher lexical density (44%) than second language learners (42%). In most of other studies (e.g., Hyltensstam 1988), the percentage of difference was almost insignificant.

The question that one might ask is whether the lexical density percentage in the reference corpus consistently outnumbers the lexical density percentage in the learner corpus in four major word classes. To this end, annotated versions of both corpora were run on *WordSmith's* concordancer.

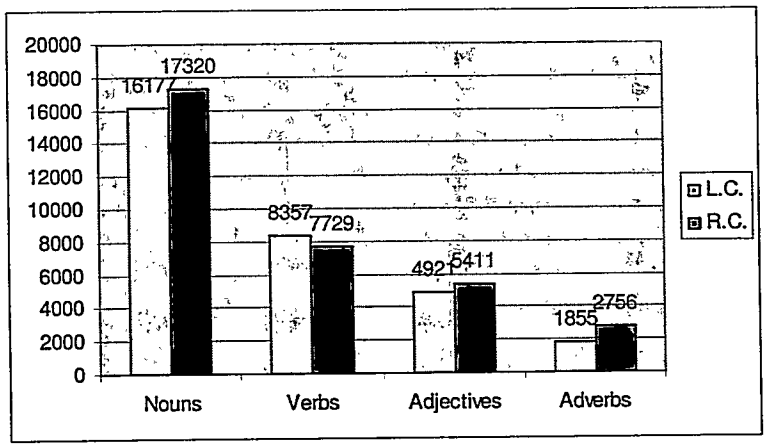


Figure 4.5. Overall frequency of content words in learner and reference corpora.

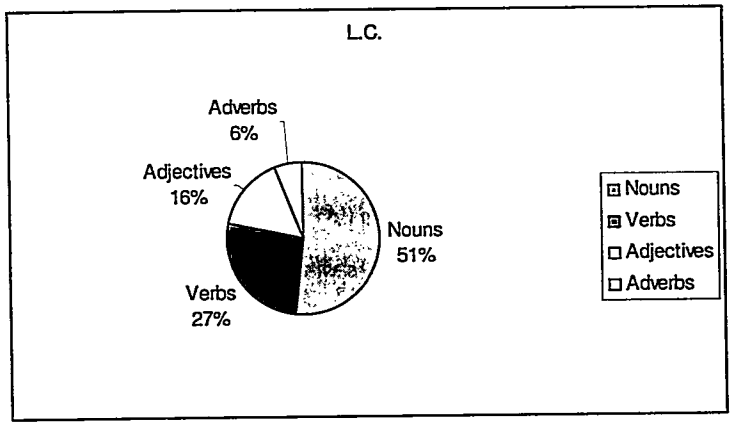


Figure 4.6. Percentage of content words in the learner corpus.

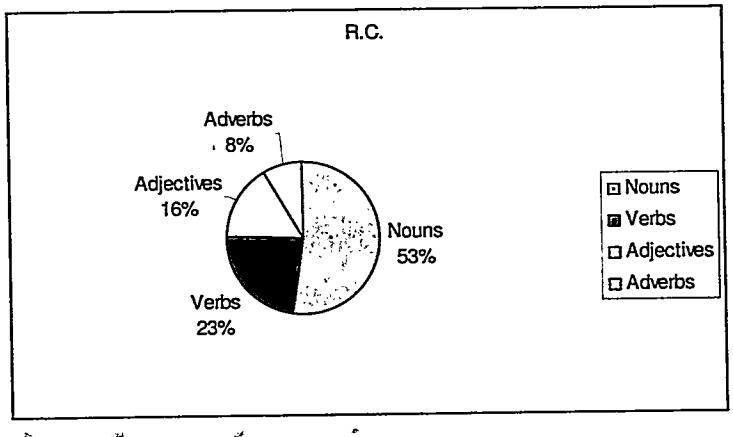


Figure 4.7. Percentage of content words in the reference corpus.

Figures (4.5, 4.6 and 4.7) reveal that lexical density in the reference corpus outpaces its learner counterpart in the number of nouns, adjectives and adverbs while it is less in the number of verbs. Do such percentages seem reasonable? It is obvious that the high percentage of nouns is quite normal for three reasons. First, a high percentage of nouns vis-a-vis other parts of speech has been attested in the literature (Biber 1998, Biber et al. 1999, Connor 1990, Halliday 1989, Grant and Ginther 2002, among others). Despite the wide gap between the number of content words, Biber et al. (1999) found that in overall frequency nouns are the most frequent category among all the word classes though nouns are the least frequent in conversation (Guo 2003:1). Secondly, by examining excerpts from Bertrand Russell's wrtings to check the use of nominalization in modern English, Halliday (1989) concludes that modern English is really "highly nominalised" and that "lexical meaning is largely carried out in the nouns"(p.72). Thirdly, in the context of academic writing, it is relevant to mention that the more proficient writers use more nominalizations than do the less proficient writers (Grant and Ginther 2002:135). Thus, the learners' underuse of nouns in comparison with the NSs might be attributed to their low level of proficiency in L2. It is relevant to mention that the percentage of nouns (in the total number of all word categories) in the reference and learner corpora (24.53%) and (23.01%) respectively supports most of the previous research findings. For example, the percentages of noun categories in *Brown* and *LOB* corpora (1,000,000 tokens in each) are 26.80% and 25.2% respectively.

The learners' overuse of content verbs in comparison with the NSs is also attested in the previous literature. In a comparison between a sampled LOB corpus (S-LOB) and the corpus of the Chinese EFL learners' written production (ILC), Dafu (1994) found that "native speakers use more nouns, adjectives, wh-determiners, articles and prepositions while the Chinese EFL learners prefer verbs, adverbs, pronouns, general determiners and conjunctions..."

Recent research on learners' use of word classes has also attested learners' underuse of nouns and overuse of verbs. In a contrastive article, *Between Verbs And Nouns And Between the Base Form and the Other Forms of Verbs- -A Contrastive Study into COLEC and LOCNESS*, Guo (2003), examines the use of 25 verbs and their noun equivalents in COLEC (a corpus of learner English mainly composed of Chinese university students' essays in national exams) and LOCNESS (native) corpora. Findings show that learners mainly use verbs whereas native speakers prefer nouns. By examining the frequency of the same 25 verbs and their noun equivalents used in Guo (2003), this study, as shown in Table (4.3), furnishes clear-cut support for learners' preference of using verbs where NSs use nouns.

Table 4.3. The frequency of 25 verbs and their equivalent nouns in learner and reference corpora

Word	(verbs) L.C.	(nouns) L.C.	(verbs) R.C.	(nouns) R.C.
accept	16	0	26	7
apply	1	0	7	1
argue	0	0	25	77
assume	0	0	4	0
believe	33	5	54	15
choose	37	13	7	23
commit	3	1	10	3
communicate	10	29	0	1
compare	0	0	5	4
complete	22	0	16	4
create	5	1	9	19
enter	7	0	10	0
examine	4	2	4	0
express	17	4	4	0
include	9	0	8	0
indicate	1	0	2	0
introduce	0	3	2	6
involve	0	0	4	4
manage	0	0	1	2
occur	7	1	12	0
produce	10	3	7	1
realise	0	0	5	7
realize	6	1	12	0
refuse	3	0	3	5
survive	1	0	11	8
Total	192	63	248	187
Percentage	75.29	24.71	57.01	42.99

The ratio of nouns to verbs in the reference corpus (42.99% vs. 57.01%) seems to be normal, and it is much closer than the ratio between nouns and verbs in the learner corpus (24.71% vs. 75.29%). This clearly indicates that the percentage of difference between the two corpora in terms of the use of nouns (18.28%), which favored the reference corpus, indicates that learners prefer to use verbs where NSs' use nouns.

While there was an overall similar use of content and modal verbs between French learners and American native speakers, Ringbom (1998:43-44) found that the most frequent main verbs surface more frequently in the learner corpora than in the NSs corpus as shown in Table (4.4).

Table 4.4. High-frequency main verbs forms-occurrences per 10, 000 words

Adapted from Ringbom (1998:43)

Word	NS	FRE	SPA	FIN	FINSW	SWE	DUTCH	GER
think	6	21	21	22	30	30	16	22
get	6	7	18	18	16	16	14	19
make	14	12	16	15	17	17	12	10
become	8	14	7	13	9	9	13	5
want	6	11	11	9	14	12	11	14
take	9	10	6	9	11	12	8	11
find	5	9	7	7	11	11	6	10
know	4	7	9	9	11	11	9	10
use	13	4	13	9	9	11	6	6
go	5	8	7	8	10	8	12	12
live	3	11	12	6	8	11	6	10
Total	79	114	127	125	146	148	113	129

From these figures, it becomes apparent that the high-frequency verbs are almost always higher for learners than for NSs. Though there is significant diversity in terms of verbs frequency, the average occurrence of each verb in the learner corpora is still higher than that of the NSs'. The wide gap between the NSs and the NNSs in terms of the uses of the content words suggests that NSs' language and NNSs' interlanguage are fairly heterogenous. It is worth considering whether the findings of Ringbom (1998) concerning the learners' overuse of the main verbs, in particular, are applicable in the case of our corpora.

Table 4.5. High-frequency main verbs forms-occurrences per 70,307 words in learner and reference corpora

Word	L.C.	R.C.
think	138	13
get	71	48
make	186	77
become	168	69
want	166	41
take	90	43
find	87	30
know	143	25
use	42	57
go	133	29
live	82	31

It is interesting to note that Tables (4.4) and (4.5) demonstrated that the overuse of the above-mentioned main verbs is a general tendency in learners' writing samples, no matter what their L1 is. Also, we should note that the diversity in terms of the number of occurrences of the verbs shown in Tables (4.4) and (4.5) is attributed to the size of the examined corpora (10,000 vs. 70,307). Finally, what could explain the overuse of these verbs in almost all learner corpora is the learners' limited word stock and their belief in the complete synonymity between these verbs and other close ones (e.g., *think* vs. *believe, become* vs. *turn*).

The percentage of adjectives (to other content word classes) in the learner and reference corpora is identical (16%). However, the proportions of adjectives (to all other word classes) in both learner and reference corpora are (7.7% and 7%,) respectively. These portions appear to be normal when compared with other corpora. For example, the percentage of adjectives in the 1,000,000 token corpora of *Brown* and *LOB* were (7.07%) and (7.3%) respectively.

A striking diversity between the two corpora is clearly seen in the number of adverbs (1,691 vs. 2,786), which favored the reference corpus. By comparing compositions written by Swedish learners of English and NSs' writing, Linnarud (2,638) attested that the largest differences between the groups lie in the adjectives and adverbs. While there is surprisingly little research on this particular aspect, it is possible to attribute the divergence in the number of the adverbs between the two corpora to the following causes.

- Learners' use of adverbs is somewhat different from that of the NSs; for learners, the use of adverbs is largely restricted to intensification and (quasi-nominal adverbs of) time. However, for NSs adverbs are multifunctional (e.g., adjuncts, conjuncts, cohesive and referential devices, hedges, evidentials, amplifiers) (Hinkel 2002:121-22). This means that NSs use more adverbs than NNSs.
- The overemphasis of textbooks, together with teachers, on lexical items that express or describe actions (verbs) is another primary reason behind the huge disparity between the two corpora in terms of the use of adverbs.
- L1 influence, where adverbs are used less commonly than in English (Smith:1987).

Overall, the results so far show that the reference corpus is much more complex in terms of lexical diversity than the learner corpus.

4.3 Results Related to Research Question (2)

Research Question (2): To what extent does the learner corpus deviate from the reference corpus in terms of the features and percentages of the top 200 frequent tokens and hapax legomena? And how can learners' lexical stereotypes be captured through word frequency?

There is a strong consensus among corpus linguists on the importance of word frequency lists in corpus analysis (McEnery & Wilson 1996, Kennedy 1998, among others). Drawing on its multifunctional uses, creating a word frequency list is a fruitful and productive technique, in the sense that it might be used for various purposes ranging from designing syllabuses to text analysis. This technique has also shown great reliability in revealing the nature of the subject matter of a text or corpus and several other lexical aspects such as active or inactive vocabulary, the differences between spoken and written discourses, hapax legomena (words used one time in the corpus) and the influence of L1. Moreover, frequency lists provide unique insights into the repetitive mechanism and other rhetorical aspects including the overuse or underuse of lexemes in learner corpora compared to the authentic (native) ones.

Beyond the previous uses, recent research on SLA has shown the centrality of frequency lists in measuring learners' vocabulary. Lexical Frequency Profile (LFP), proposed in Laufer and Nation (1995), is now considered the most reliable and powerful measure of learners' vocabulary proficiency or knowledge. Likewise, frequency lists help determine the number of vocabulary items learner needs to become proficient or fluent in L2. Laufer and Nation (1999) argue that 79.9% of written English uses only the top 2000 most frequent words in the language. This indicates that mastering such words guarantees a good command of the target language.

Apart from its normal use in examining catches, frequency lists have also been used in this study as a preliminary tool to select and then examine lexical and collocational errors via concordancing. Figures (4.8) and (4.9) present the top 100 frequent tokens (in a version of the list arranged in descending frequency order) in the learner and reference corpora respectively.

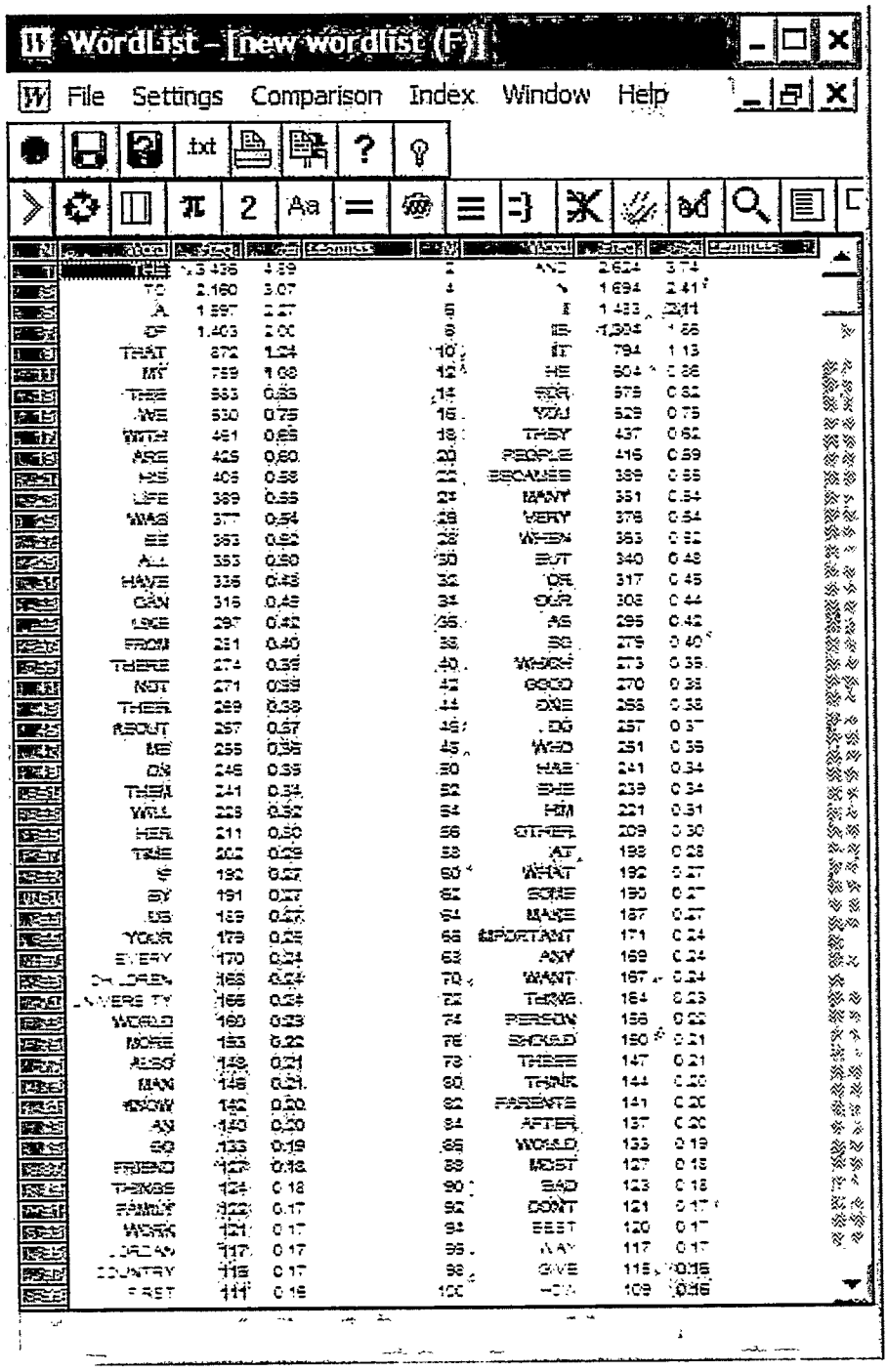


Figure 4.8. Top 100 frequent words in the learner corpus.

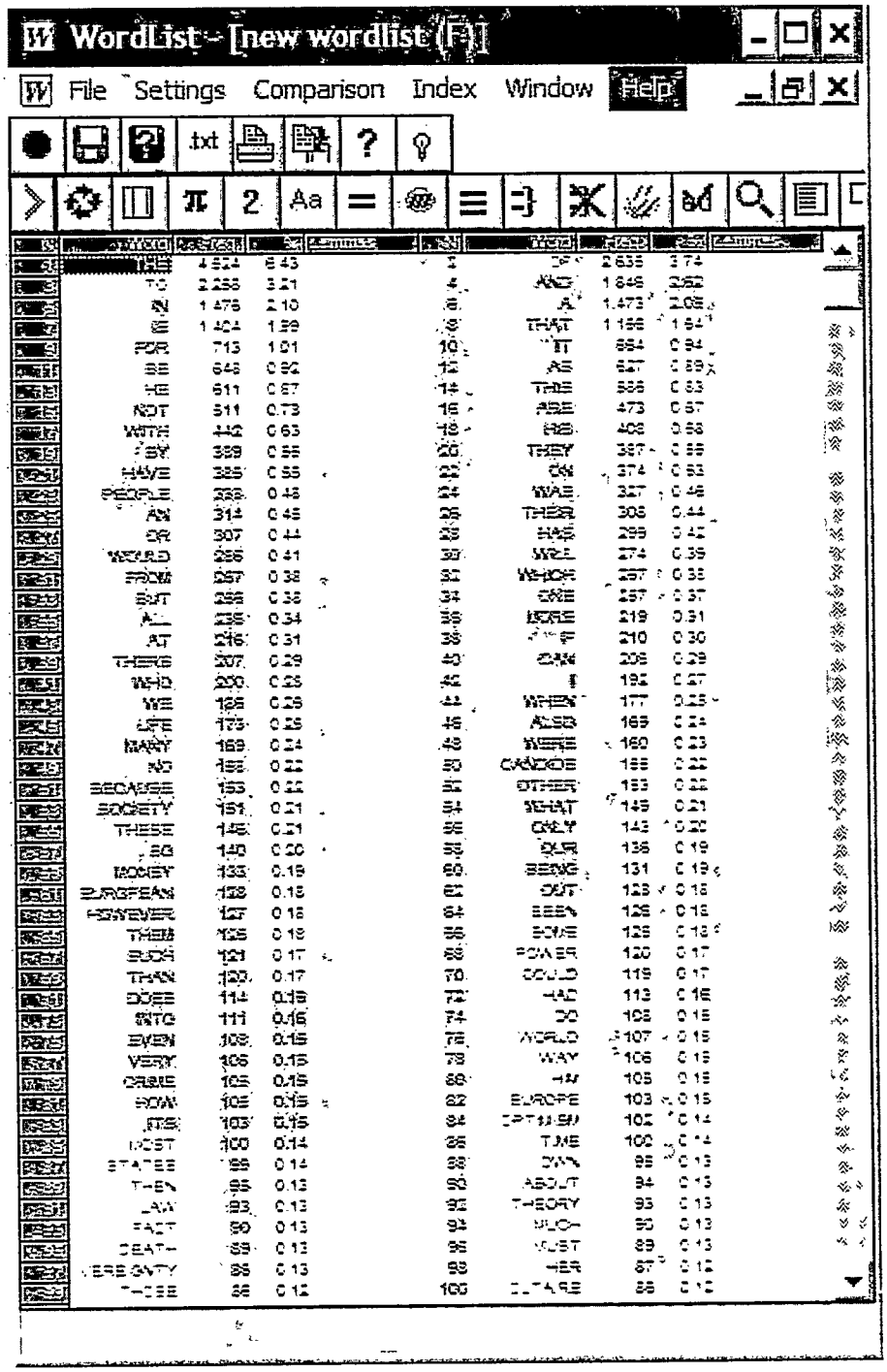


Figure 4.9. Top 100 frequent words in the reference corpus.

Along with the tokens, seven notable points immediately emerge from Figures (4.8) and (4.9). First, function words occupy the top positions in terms of frequency in both corpora. Out of the 200 tokens used in the above extracts, only 51 tokens were content words. Secondly, due to the excessive use of some vague nouns and generic adjectives (e.g., *people, way, life, important, great, old*), which are attributable to lexical developmental stages and the influence of the L1, the learner corpus leads the reference corpus by 7% in terms of the content words in the top 100 frequent tokens. Thirdly, learners' L1 rhetorical devices (e.g., overstatement, writer visibility *I*) have a noticeable effect on word frequency. Fourthly, the two corpora share approximately two-thirds of the used tokens. Fifthly, as one scrolls down, frequency, together with the percentage of tokens, consistently, but sharply, declines in both extracts (from 4,525 to 85). Sixthly, though it is much higher in the learner corpus than in the reference corpus, the top 100 tokens in both corpora take up more than 50% of the total number of the tokens in the entire corpora.

Two questions immediately come to mind while looking at the extracts shown in Figures (4.8) and (4.9): what is the importance of word frequency lists in this study? Which factors are likely to be responsible for the differences in frequency between the two corpora (learner corpus and reference corpus)?

The central role of a frequency count has recently become an established tenet in much of the linguistic research. Doubtlessly, its advantages are large and varied. As for this study, in particular, a frequency count provides us with fruitful information that otherwise would be difficult to reveal. First, by displaying the contents of a corpus in an isolated word list, the frequency lists provide us with the lexical repertoire of the subjects and what remedies they might need to in order overcome their lexical difficulties or gaps. This, in turn, enables us to put forward generalizations concerning the subjects' lexical richness or impoverishment. Such lists also give syllabus designers a fine-grained

picture of the missing or inactive (less frequently used) vocabulary that the learners might urgently need. Secondly, using the word lists, it was possible to select the items to be run on the concordancer to investigate lexical and collocational errors. Thirdly, the word lists provide us with crucial information concerning the percentage of hapax legomena, rhetorical and stereotyped features of learners' writing.

While (65%) of the lexical items in the top 100 frequent tokens are shared between the two corpora, it appears to be unsound to rely on this ratio as an indicator of similarity or difference between them. There are, at least, two reasons that may justify this statement. First, the high percentage of the shared types between the two corpora is misleading since more than 70% of these tokens or types are grammatical words, which always occupy the top positions in any corpus, whether native or learner. This is what led Halliday (1989:65) to categorize lexical items into three categories rather than two: (i) grammatical words, (ii) high frequency lexical items and (iii) low frequency lexical items. By so doing, Halliday (1989) assumed that grammatical words are always high in terms of frequency. Secondly, in most cases, the top unshared frequent types reflect the divergent themes of the texts providing the database of the corpora.

A close look at the percentage of the number of content words to the grammatical words in the top 100 frequent tokens in the learner and reference corpora shows some variation in the proportion of each corpus in the total number of content words as shown in Figure (4.10).

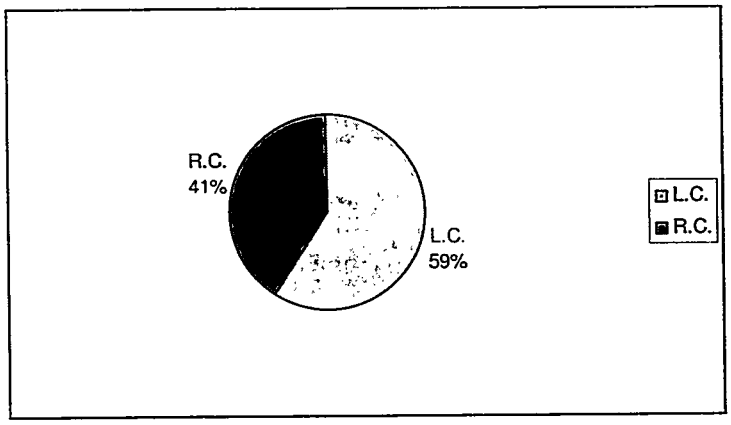


Figure 4.10. Proportion of the learner and reference corpora in the total number of the content words in the top 100 frequent tokens.

It should be made clear that this percentage depends on the size of the corpus in question and the type of texts comprising its database. In his article, *Vocabulary Frequency in Advanced Learners English: A Cross-Linguistic Approach*, Ringbom (1998) compared the top 100 frequent words in seven learner corpora, whose participants belong to seven different language groups. The findings show that learners' use of the 100 most frequent words was almost 4 to 5 percent higher than native speakers. A close look at the Figure (4.11) shows that the percentage of the top 100 tokens to the total number of tokens in both corpora was 5.3% higher in the learner corpus. Thus, this percentage goes in the same direction as in previous research (e.g., Ringbom 1998).

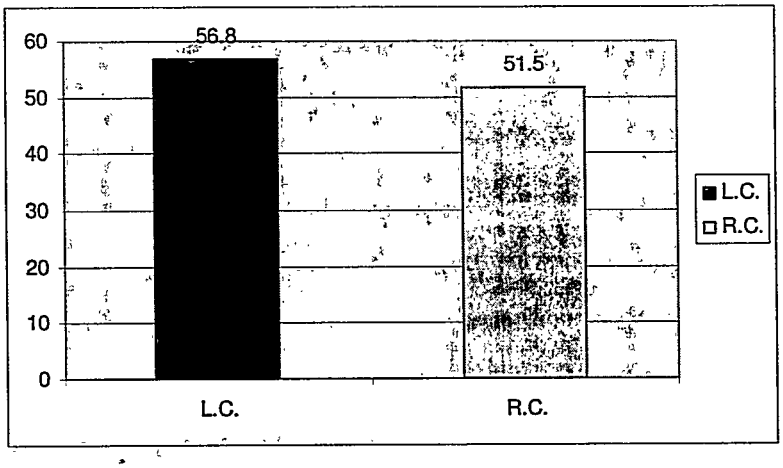


Figure 4.11. Percentage of the top 100 frequent tokens in the learner and reference corpora.

The percentage of the top 100 frequent tokens (shown in Figure 4.11), which accounts for more than 50% of the total number of all tokens in each corpus should come as no surprise here. In research on the approximate percentage of different word types at different word frequency in texts, Kennedy (1998) states that "between 50 and 100 English words typically account for half of the total word tokens in any text" (p. 97).

By comparing the number of the content words with the total number of the tokens in the top 100 frequent tokens, it becomes apparent that the tokens of the learner corpus outnumber the reference corpus by 3,748 tokens. As illustrated in Figures (4.13, 4.13 and 414), the ratio of the content words frequency to that of the grammatical words is 7% in the reference corpus while the equivalent ratio in the learner corpus is 14%.

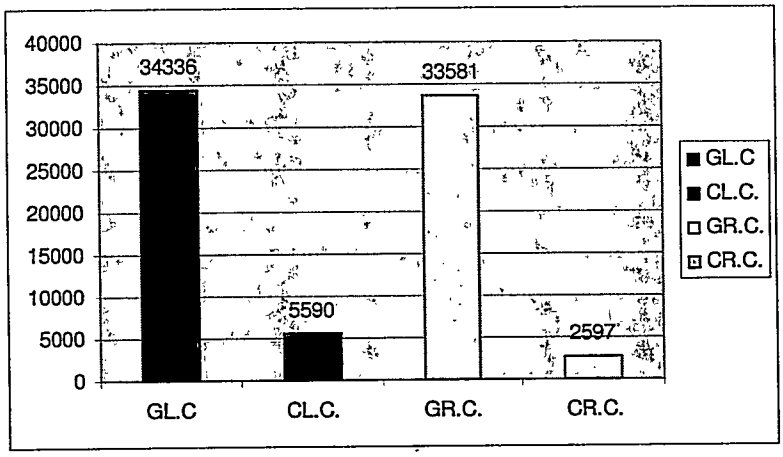


Figure 4.12. Frequencies of the content and grammatical words in the top 100 frequent tokens in learner and reference corpora.

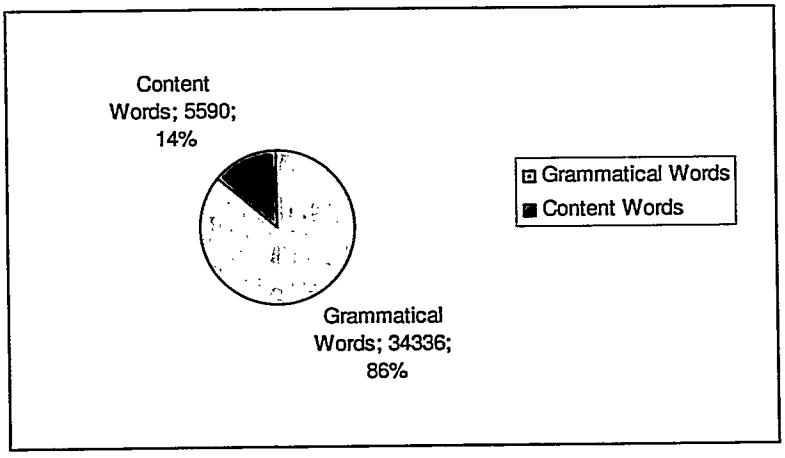


Figure 4.13. Ratio of the content words frequency to that of the grammatical words in the top 100 frequent tokens in the learner corpus.

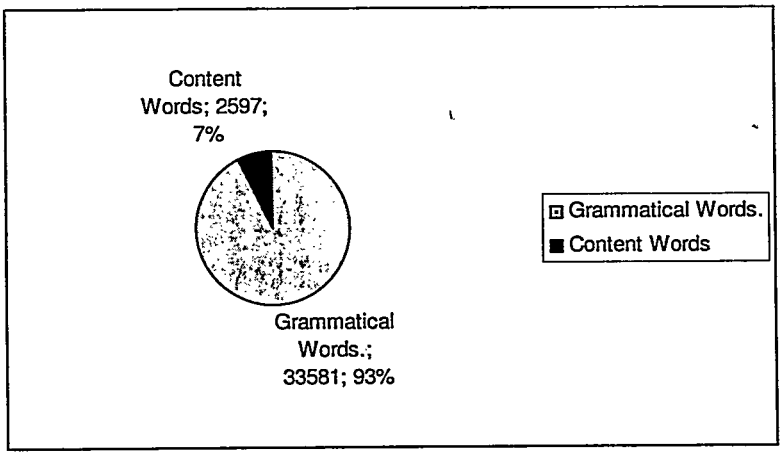


Figure 4.14. Percentage of the frequency of the content words to that of the grammatical words in the top 100 frequent tokens in the reference corpus.

Drawing on the learners' heavy use of some common tokens, Ringbom (1998) argues that advanced learner language is vague and stereotyped. To garner satisfactory empirical support for this argument, he provides numerous examples of learners' overuse of the less common grammatical words (e.g., *which, into, because*), along with some vague content words (e.g., *way, people, thing(s)*). The first person pronoun *I* and the verb *think*, for instance, were overused by learners between three to five times (in comparison with the NSs' use of these items). More often than not, the use of vague lexica is attributed to the lack of target vocabulary in the learner's lexical repertoire.

It is striking to find that the percentage of the top 10 frequent tokens in learner and native corpora appears to be similar regardless of their size. A close look at Figure (4.15) makes it clear that the present learner and reference corpora, the *Quebec Learner Corpus* (QLC), and the *Brown Corpus* are alike in in terms of the percentage of the top 10 frequent tokens (relevant to the total number of all tokens in the corpus), though BC (1,000,000 words) is almost seven times as big as that of the present learner and the reference corpora combined.

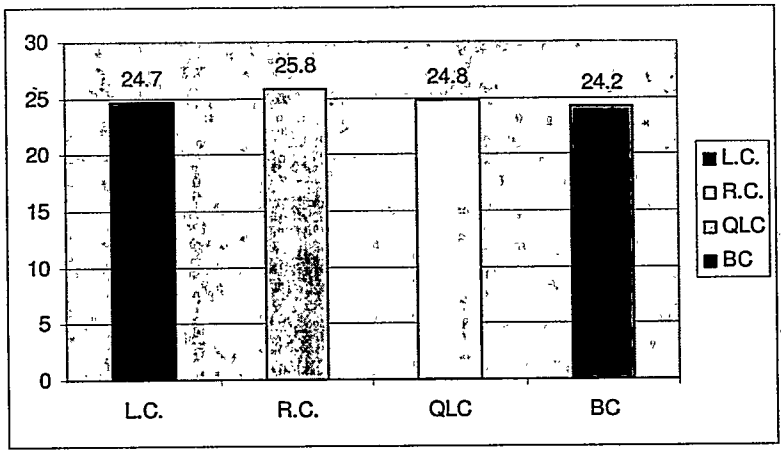


Figure 4.15. Percentage of the top 10 frequent tokens learner and reference corpora.

Ringbom (1998:42) furnishes additional support for the percentage of the top 10 frequent words, which seems to be universal; the percentage of the top 10 frequent words in the seven corpora, according to his study, is almost 25% of the total number of tokens in each corpus.

Additionally, frequency lists have provided a reliable tool to examine the textual features (linguistic and rhetorical) of both corpora. More concretely, the use of concordancing depends on the types (different words) and frequency percentages displayed by the frequency indexer. Among the textual features examined in the coming sections are parts of speech, coordination, hedges and emphatics.

The previous analysis might immediately raise issues of similarities and consistency, that is, whether the behavior of the second 100 top frequent tokens is similar to the first top 100 frequent ones. A look at Figures (4.16)and (4.17) suggests tremendous diversity between the first top 100 frequent tokens and the second 100 frequent tokens in both corpora.

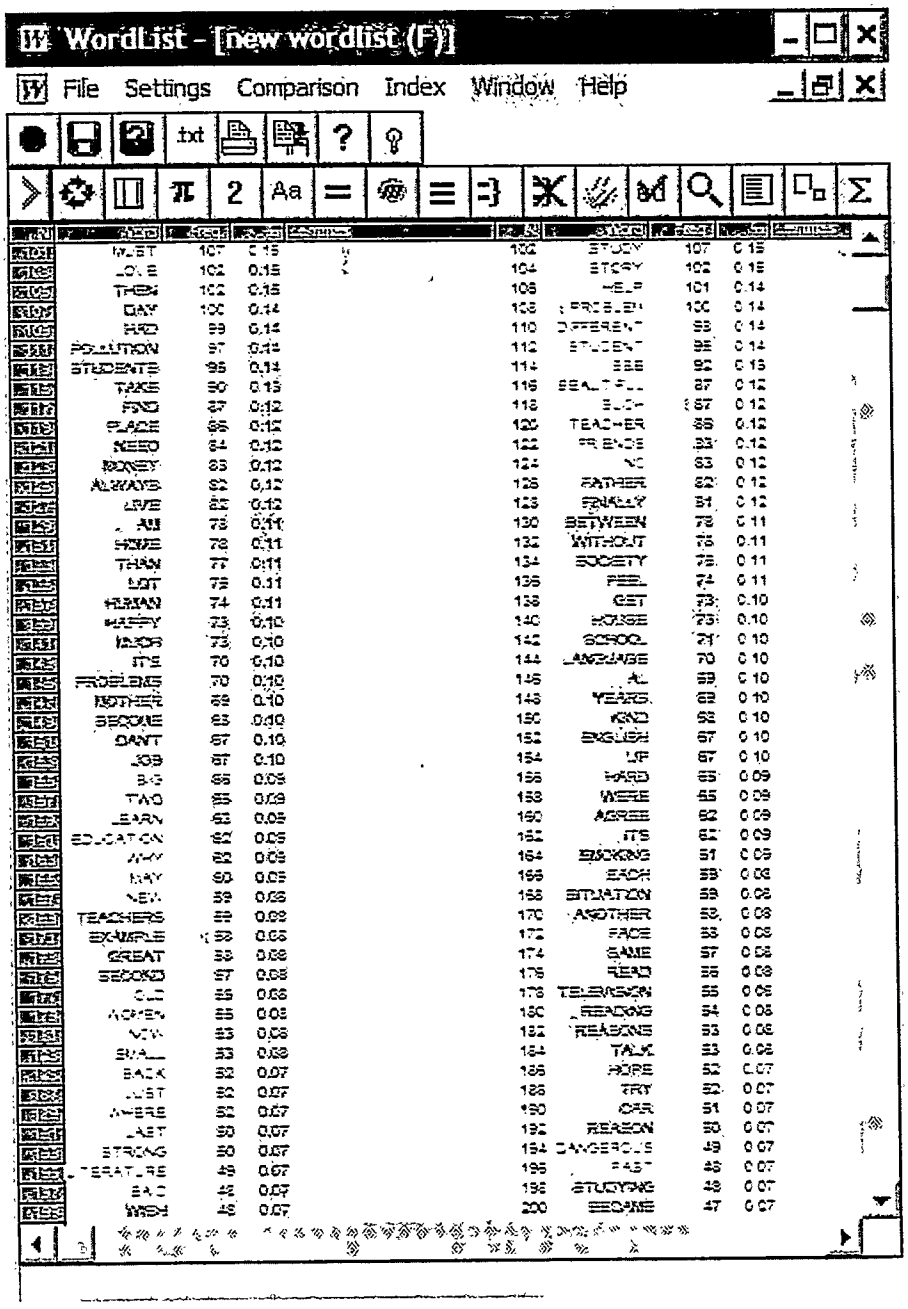


Figure 4.16. The second 100 frequent words in the learner corpus.

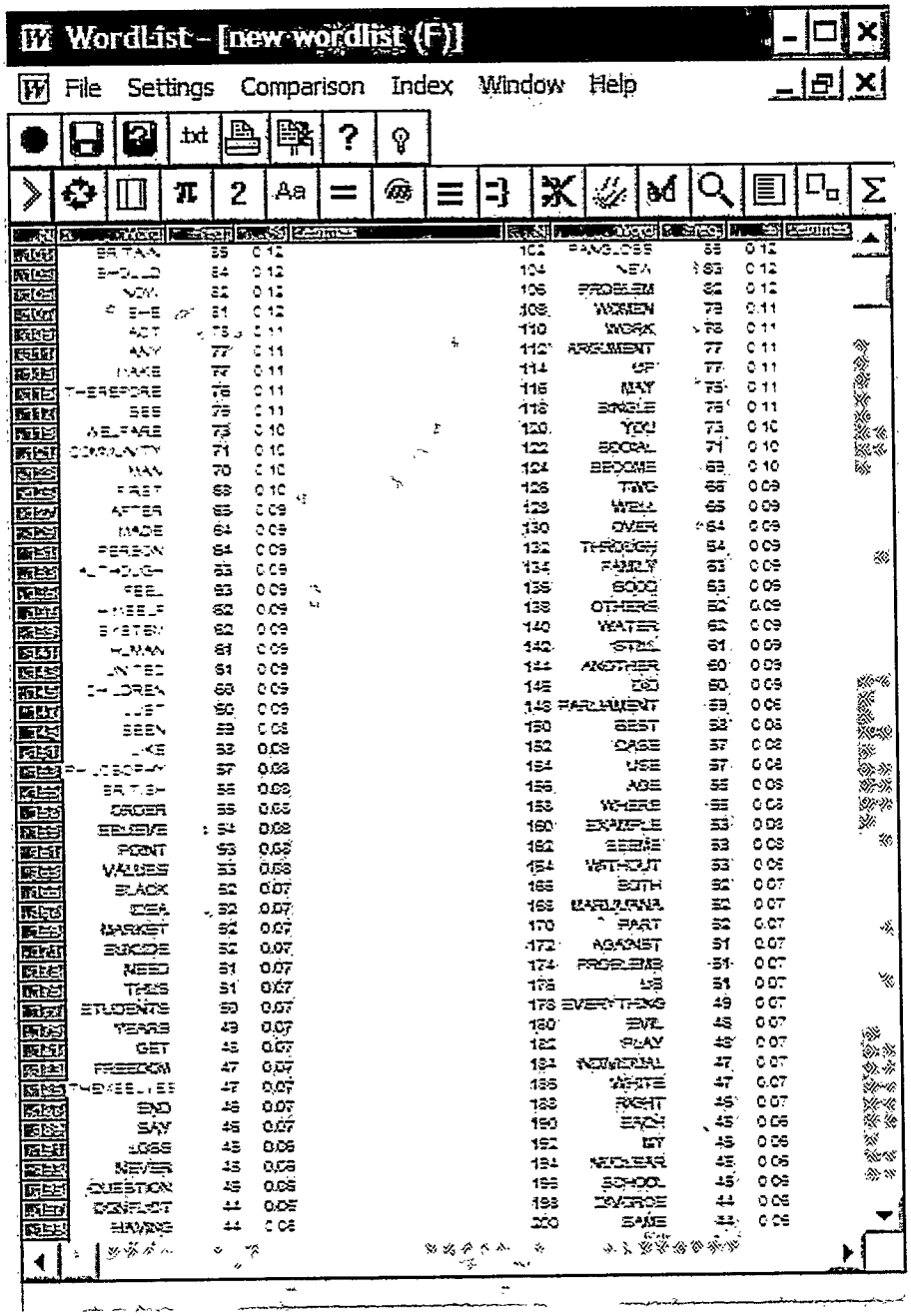


Figure 4.17. The second 100 frequent words in the reference corpus.

A careful examination of the second 100 tokens in each corpus shows three crucial points:

1. A marked increase in the number of content words in the second 100 frequent tokens:

   Unlike the first top 100 frequent tokens, where more than (70%) of the tokens in both corpora are grammatical words, the proportion of the content words in the total number of tokens in the second 100 frequent tokens in the learner and the reference corpora are (81%) and (74%), respectively, as shown in Figures (4.18) and (4.19).

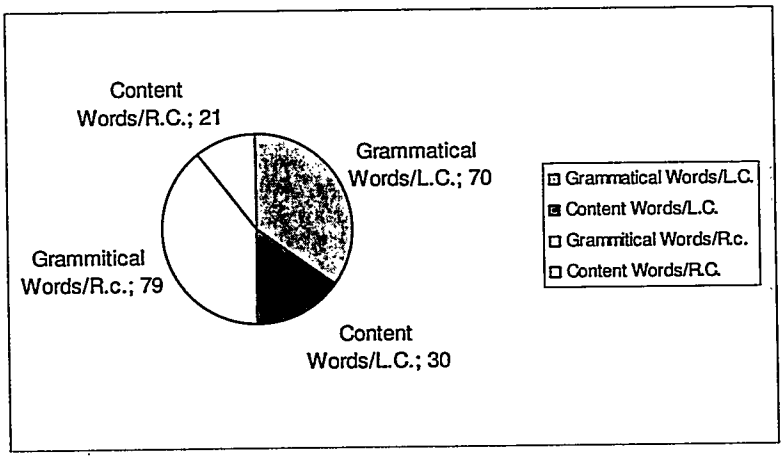


Figure 4.18. Number of content and grammatical words in the top 100 frequent token in the learner and reference corpora.

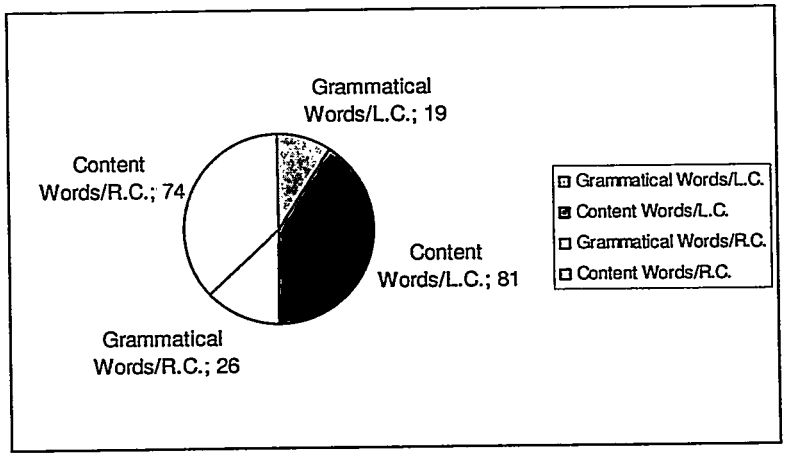


Figure 4.19. Number of content and grammatical words in the second 100 frequent tokens in the learner and reference corpora.

2. A marked decrease in the contribution of the second 100 frequent tokens to the total number of corpus tokens:

The sharp decline in the percentage of the grammatical words provides powerful evidence for the continuous decrease in the number of grammatical words as we scroll down. While the percentage of the first top 100 frequent words claims over (50%) of all the tokens in both corpora, the percentage of the second top 100 frequent tokens in the learner and reference corpora constitues only (9.5%) and (8.5%) respectively. However, the high percentage of the second 100 frequent tokens in the learner corpus compared with the reference corpus supports Goodfellow's at al. (2002) argument concerning learners' high lexical frequency at the early stages:

> we could expect vocabulary knowledge at an early stage of development to consist mainly of high frequency words and at a later stage to have a higher proportion of low frequency words.

3. The learners' marked use of generic adjectives (e.g., *strong, old, dangerous, great*).

Another disparity between the learner corpus and the reference corpus lies in the percentage of the hapax legomena. Plag (2000) argues that hapax legomena and high frequency are conversely proportional. This indicates that the more hapax legomena a corpus has, the more productive (varied) it will be. Table (4.6) shows that the reference corpus contains many more hapax legomena than the learner corpus.

Table 4.6. Percentage of hapax legomena in learner and reference corpora

corpus	# of hapax legomena	# of types	% of hapax legomena
R.C.	3,417	7,322	46.7%
L.C.	2,361	5,248	45%

Drawing on such findings, we discover that the number of active vocabulary items (lexemes used more than once) in the reference corpus is 3,905 while it is 2,887 in the learner corpus. From the evidence of these lines, the new measure of productivity (hapax legomena) has unquestionably raised the NSs' lexical productivity percentage. Diversity in terms of the number of hapax legomena between the NSs and the learners corpora conveys the idea that learners tend to rely on the more repetitive lexemes than on unique words.

Like lexical diversity, hapax legomena are sensitive to corpus size; the larger and more representative a corpus is, the fewer hapax legomena it contains. This explains the high percentage of hapax legomena in the learner and reference corpora vis-a-vis the percentage of hapax legomena (39.5%) in the *American Heritage Intermediate Corpus*, which consists of 5.09 million words. Yet, it is clear that the percentage of hapax legomena is not always consistent in all corpora. For, example a text that lists items (names of

persons, places, machines, etc.) is expected to have more hapax legomena that a literary text that has repetitive lexemes or patterns.

In view of these findings, it is crucial to note that the percentage of hapax legomena has widened the gap between learners and NSs concerning lexical richness. A more likely explanation for the diversity resulting from the percentage of hapax legomena is the interaction between the developmental stage and word frequency.

As we leave our discussion of this question, it is important to note two relevant points. First, the task of the frequency count is not yet complete. Rather, it is, as shown below, an important tool to investigate other features, particularly those related to the underused and overused lexemes. Secondly, the power of word frequency lists, which decontextualizes tokens, is rather limited since the lists provide no access to the usage context. For this reason, lexical and collocational error analysis, which is at the heart of the objectives of this study, is not established until the context is available. Thus, to access the environment of lexicon, another tool, namely, the concordance software, is required.

4.4 Results Related to Research Question (3)

Research Question (3): What are the most salient and stereotyped features of the learner corpus? And how far is the learner corpus influenced by the learners' L1?

Research on CL has recently witnessed the extension of Crystal's (1991) notion of *profiling*, which was originally concerned with stylistics, to the interlanguage domain (Granger 1998:119). *Text-profiling* was used in this study to refer to the identification of the most salient lexical and stereotyped features of the learner corpus; identification of such features requires continuous use of the reference corpus for comparative and contrastive purposes. Despite the various lexical and stereotyped features that might be included under this title, this section is limited to exploring four main areas: (i) word

categories, (ii) overproduced lexical items, (iii) underproduced lexical items and (iii) non-lexical measures (learners' proficiency in L2, paragraphing and word and sentence length).

4.4.1 Word Categories

Research on CL has been deeply influenced by the constant productivity of artificial intelligence, which has, so far, evolved into numerous tools that have shown outstanding capabilities in processing huge corpora. Tagged corpora, as mentioned earlier, have some capabilities that raw corpora do not. Via the codes/tags used in the corpus tagging, for instance, it is possible to investigate various features of the corpus in question, regardless of its size, in a remarkably short period of time. Among the features whose investigation was tedious in the near past is the proportion of word categories. Investigation of such categories, as shown in Figure (4.20), exemplifies further advantages of the tagged corpora.

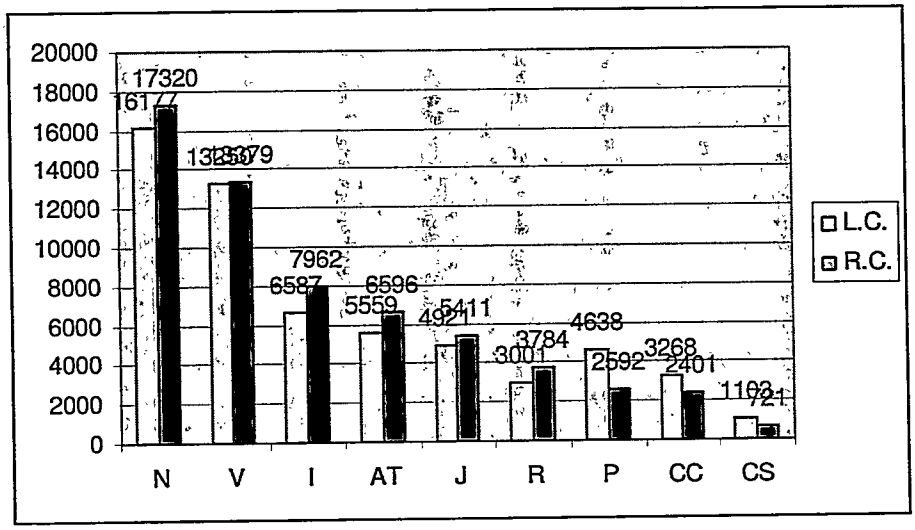


Figure 4.20. Word category in learner and reference corpora.

Table 4.7. Reduced word category tag list

N	Nouns
V	Verb
I	Prepositions
AT	Articles
J	Adjectives
R	Adverbs
P	Pronouns
CC	Coordinations (adversative) coordinating conjunctions
CS	Subordinate conjunction

Variation in word category between authentic corpora and learner corpora, as the literature shows (e.g., Granger 1998), is likely to occur more often than not in a systematic way. As it is shown in Figure (4.20), word categories in the learner corpus (relative to the reference corpus) can be classified into three groups: (i) underuse, (ii) overuse and (iii) similar use.

Table 4.8. Learners' use of lexical categories in comparison with the NSs

1.	Underuse	nouns, prepositions, articles, and adverbs
2.	Overuse	pronouns, coordinating conjunctions and subordination conjunctions
3.	Similar use	verbs and adjectives

(i) Underused categories

(a) Nouns

Drawing on the aforementioned discussion, learners' underuse of nouns is anticipated in all learner corpora regardless of the learners' native tongue. The divergence in word categories between the learner and reference corpora, in particular, is attributed to

several factors such as: (i) the learners' low proficiency in the L2; proficient writers use more nominalizations in their writing (Grant and Ginther 2002), (ii) a general tendency, where NNSs prefer to use verbs in places where NSs choose nouns (Guo: 2003), (iii) the NSs' excessive use of nominalization in contemporary English (Haliday 1989). Learners' underuse of nouns vis-a-vis NSs has been attested in the previous literature (e.g., Granger and Rayson 1998, Guo 2003, Grant and Ginther 2002)

(b) Prepositions

Prepositions present another area of divergence between the learner corpus and the reference corpus. Explanation for the learners' underuse of prepositions, which has been also attested in previous research, might involve one or both of the following factors.

(1) interlingual factors

The influence of L1 is clearly seen when Arabic uses a zero preposition in a context where English requires the use of a preposition as exemplified in the following phrasal verbs:

1. Learner's sentence: I am waiting him (English norm: *waiting for* him)
2. Learner's sentence: We always listen our parents' advice. (English norm: *listen to* ...)

(2) a general tendency

Research on learners' use of prepositions shows that learners' underuse of prepositions is a general tendency. In his article, *Where have the prepositions gone? A study of English prepositional verbs and input enhancement in instructed SLA*, Kao (2001) found that "the null-preposition construction does occur in SLA." Granger and Rayson (1998) present further evidence of the French learners' omission of prepositions. Omission of English prepositions by Arab students of English, in particular, was also attested in Scott and Tucker (1974). In the context of our discussion of the learners' underuse of prepositions,

it is relevant to mention that learners face serious problems with preposition resulting from the negative transfer of the fixed prepositions (with adjectives and verbs) in their L1 as shown in the following examples:

- **prefer on* instead of *prefer to*
- **addicted on* instead of *addicted to*
- **proud in* instead of *proud of*
- **afraid from* instead of *afraid of*

(c) Articles

The divergence between the learner and reference corpora in terms of the use of the articles is basically attributed to the L2 richness in this category. Whereas Arabic uses either the definite or zero article, English uses four articles (*a, an, the* and *zero* article). This explains the learners' use of a zero article instead of an indefinite one when the noun in question is indefinite in their L1. Such a case was attested in Scott and Tucker (1974:86):

> Arabic marks nouns as definite or indefinite by the presence or absence of the article. Errors of omission of the indefinite article in English are attributable to MT interference. About 30 percent of the errors made with articles were omissions of the indefinite article. This occurred with equal frequency in all four samples and were the most frequent type of error within the article system.

Again, the omission of the definite article where the following noun is not definite in Arabic is very noticeable throughout the learner corpus. Another obvious and previously attested problem, in the learners' use of the definite article *the*, stems from the Arabic genitive construction as shown in the following example:

- **Jordan team* instead of *the Jordan team/the team of Jordan*

(d) Adverbs

The marked divergence in the number of adverbs, which favored the reference corpus, can be attributed to two main factors, viz. (i) L1 influence and (ii) teaching strategies and priorities.

Empirical research has shown that adverbs in Arabic are used less commonly than in English (Smith 1987: 152). This denotes that a considerable portion of the learners' underuse of adverbs is likely to be attributed to the influence of their native tongue. Additionally, the high concern of text materials and instructors with tokens expressing actions explains their overuse of verbs and underuse of adverbs. Such divergence in the number of adverbs between NSs and NNSs was also attested in literature; it is worth reiterating that this result is consistent with Linnarud (1986), who found that the largest differences between Swedish learners of English and the NSs lie in the adjectives and adverbs.

(ii) Overused categories

(a) Pronouns

The excessive overuse of pronouns in the learner corpus is primarily attributed to learners' preference for visibility in the text. Support for this argument comes from the excessive use of the first person pronoun *I* in the learner corpus (1,433 times) compared to only (184 times) in the reference corpus. The huge gap between the two corpora in the use of the first person pronoun *I* reflects the extension of the subjectivity of the Arabic discourse to the target language, where objectivity rather than subjectivity is almost always the optimal candidate.

In coming to understand the discourse subjectivity and learners' preference for visibility in the text, an attempt is made here to compare the use of the conjugations of the first person pronouns in the learner and reference corpora.

Table 4.9. Analysis of features of writer visibility in the learner and reference corpora

Feature	L.C.	R.C.
First person singular pronouns (I, I'x, me, my, mine)	2,441	250
First person plural pronouns (we, we'x, us, our, ours)	1,023	370
Total first person pronouns	3,464	620

From this brief comparison, it becomes manifest that learners' subjectivity vastly outweighs that of the NSs. It might be argued that the overuse of the first person pronouns is a general tendency rather than a language specific feature. While this is unquestionably true, the stigmatized use of such pronouns in the learner corpus compared with other learner corpora makes these pronouns attributable to the L1 rhetoric, too. Support for this conclusion comes from Petch-Tyson (1998). In an analysis of the features of writer/reader visibility, Petch-Tyson (1998:112) found that Dutch, Finnish, French and Swedish learners of English markedly overused more first and second person pronouns in comparison to NSs as shown in Table (4.10).

Table 4.10. Analysis of features of writer/reader visibility

Adapted from Petch-Tyson (1998:112)

Feature	Dutch (55,314)	Finnish (56,910)	French (58,068)	Swedish (50, 872)	US (53,990)
First person singular pronouns (I, I'x, me, my, mine)	391	599	364	448	167
First person plural pronouns (we, we'x, us, our, ours)	484	763	775	1,358	242
Second person pronouns ((you, you'x, your,yours)	447	381	257	227	76
Total first/second person pronouns	1,322	1,743	1,396	2,033	485
Total first/second person pronouns per 50,000 words	1,195	1,531	1,202	1,998	449

By taking the number of token in each corpus into consideration, none of the learner corpora shown in Figure (4.10) above outnumbers the present learner corpus in terms of the frequency of the first person pronouns. This appears to indicate that ascribing the overuse of the first person pronouns solely to the general tendency or developmental stages is ungrounded.

(b) Coordinating conjunctions

While the evidence provided here concerning the learners' overuse of coordinating conjunctions supports the previous research (e.g., Kharma 1985, Kaplan 1966), it is important to mention that such a conclusion is sometimes misleading. Support for this argument comes from numerous examples of *and*, where it is used as a sentence opener rather than as a coordinating conjunction as shown in Figure (4.21). Further analysis of the use of and as a sentence opener is illustrated in the coming sections.

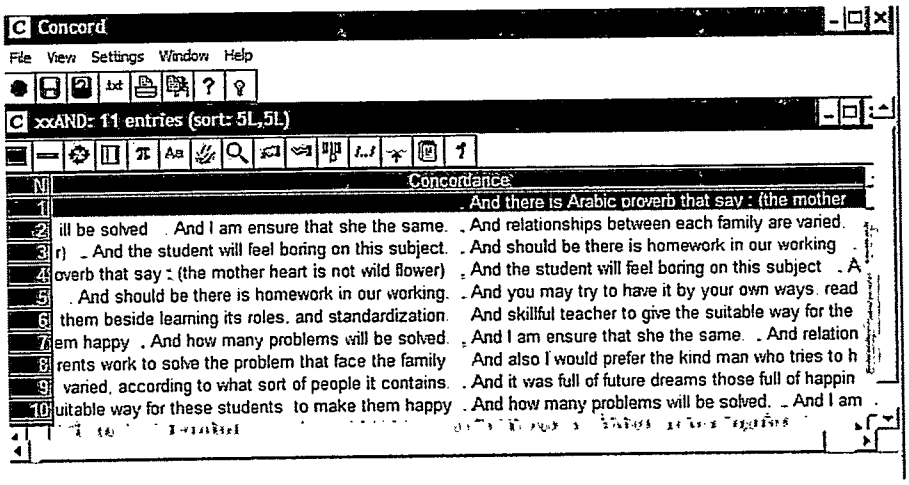


Figure 4.21. Examples of the use of *and* sentence initially.

(c) Subordinating conjunctions

In much of the previous research (e.g., Kaplan 1966), it was argued that Arab students of English, due to the influence of the L1, overuse coordination and underuse subordination. While this seems to be partially true for coordinating conjunctions (*and* and *but*), it is still questionable for the subordinating conjunctions, particularly, because such studies were based on limited samples of texts. As Figure (4.20) shows, subordination use, contrary to previous claims, was found to be higher in the learner corpus.

4.4.2 Overproduction and Verbosity

The advent of modern software programs, as mentioned earlier, has made it possible to examine, compare and contrast the number of occurrences of lexical items between corpora no matter how large they are. A subsequent advantage of this development is the ability to examine the use, misuse, underuse or even overuse of lexical items in learners' speech or writing compared with a corpus of a similar-sized native corpus. Before going

any further, it is worthwhile to reiterate that the term *overproduction* is used in this study to refer to lexical and grammatical items that are used excessively by learners across the corpus (on a full corpus basis). *Verbosity*, which is sometimes used to refer to a high style of lexicon or pretentious words (e.g., Zughoul 1991), is used here to refer to the words unnecessary in a given context (Ringbom 1998:50).

By running the *Wordlist* tool for text comparison on the two corpora, it was possible to see numerous instances of divergence in the marked overuse of lexical items. While there are numerous instances of overused items that might be classified under the general tendencies of learners that are confirmed in previous research (such as vague expressions e.g., *people, thing(s)*), there are also various instances attributed to the learners' L1 rhetoric. For the sake of clarification, Figure (4.22) presents some of the divergence between the two corpora in this aspect.

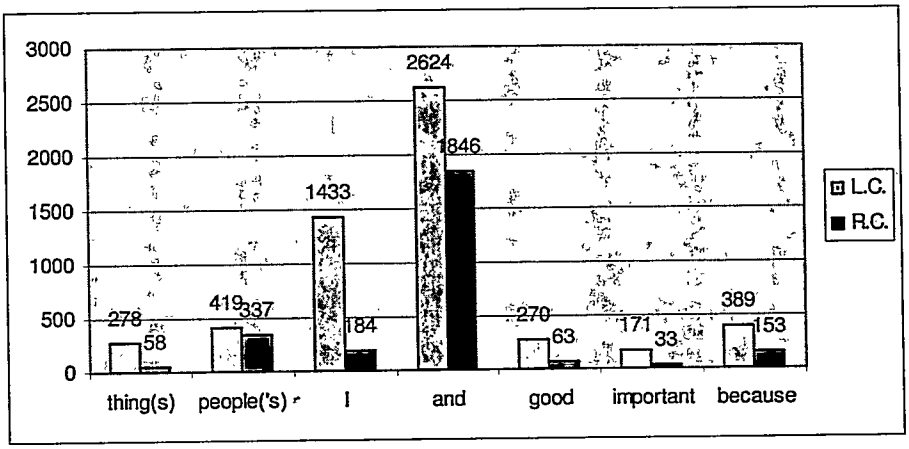


Figure 4.22. Samples of overproduction.

The above brief comparisons provide further evidence that learners' interlanguage and NSs' writing are heterogenous. There are two possible reasons for such heterogenous results. First, in the situation where there is neither a daily contact with the NSs of the

target language, nor much exposure to authentic texts, learners' interlanguage tends to rely heavily on their L1 rhetoric. Thus, these overproduced items reflect the rhetoric of their L1. Secondly, some of the vague overproduced lexemes tend to be general tendencies. This explains the overuse of words (e.g., *things, people, way, world*), which are also found in the output of other English learners (Halliday 1989, Hinkel 2002, to name just a few).

Support for the first argument comes from lexemes such as *and, I, must* and *good*, which prevail in the learners' L1. *And*, for example, has multiple functions in Arabic, not least of which are for coordination and as a sentence opener. Pertinent to this is the Arabic preference for parallel structures and coordination over subordination. Likewise, it is justified to tie the overuse of the emphatics and intensifiers, as illustrated below, to the learners' L1 rhetoric, where emphasis and overstatement are preferred to hedging and understatement.

Does the literature support or counter the findings of the present study? Based on the findings of seven learner corpora examined by Ringbom (1998:45-49), it appears that learners overuse all these lexemes, no matter what their L1 background. Thus, the findings of the present study agree with the previous research. However, it is necessary to mention that going in the same direction does not imply getting the same result. As far as the coordinating conjunction *and* and the first person pronoun *I* are concerned, we see that the use of these items by Arab students of English greatly exceeds the use of the same items in the reference corpus or even all other learner corpora. Clearly, this suggests that Arab learners are heavily influenced by the rhetoric of their L1. It should be acknowledged that the use of *and* sentence initially is possible, though not as common in English. By examining its use sentence initially in the two corpora, the results indicated the L1 transfer as shown in Table (4.11). It is obvious that the use of *and* as a sentence opener is more than six times more frequent in the learner corpus than in the reference corpus.

Table 4.11. Use of *and* as a sentence opener

item	L.C.	R.C.
and	66	10

In the context of comparison between NSs and NNSs, there seems to be no escaping the frequency of overstatement in the NNSs corpora. While it is feasible to attribute adjectival intensifiers and emphatics used in the learner corpus to a universal learner tendency, it is more appropriate (due to stigmatized use) to blame the L1, where such items are prevalent.

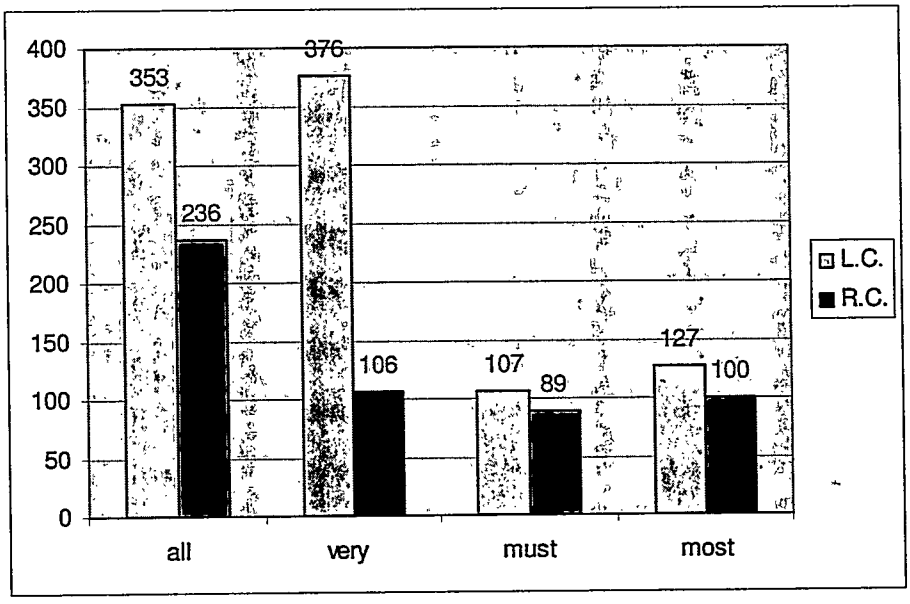


Figure 4.23. Intensifiers and emphatics in learner and reference corpora.

The primary reason that learners use emphatics and intensifiers is to strengthen the force of their propositions, a feature highly favored in Arabic discourse. Again, these findings are well-grounded in the previous literature. By examining the non-native/native differences in the actual inventories of adjective intensification in two native corpora and

two non-native ones (around 100,000 words each), Lorenz (1998:54) found that the most prominent difference between German learners of English and NSs' usage lies in the overall intensifier counts. That is, learners use more intensifiers than the NSs.

As we leave this subcategory, it should be noted that only inappropriate and stigmatized lexical items were counted as errors. This indicates that appropriate uses of lexical items, even if excessively overused, cannot be counted as errors.

4.4.3 Underproduction

One key result that might be also cited here to shed light on the differences between the learner and reference corpora is the learners' underuse of some lexical items compared with the NSs. Since divergence in terms of frequency is expected among homogeneous (between two groups of NSs) or heterogeneous groups (between NSs and NNSs), it is important to keep in mind that the examples cited in (4.3.2) and (4.3.3) of this section represent only those items markedly divergent in the two corpora. In order to exemplify some aspects of the underused lexical items in a corpus characterized by the excessive overuse of emphatics and intensifiers, it is reasonable to resort to hedges, as a polar opposite. Figure (4.24) presents some of the underproduced items between the two corpora.

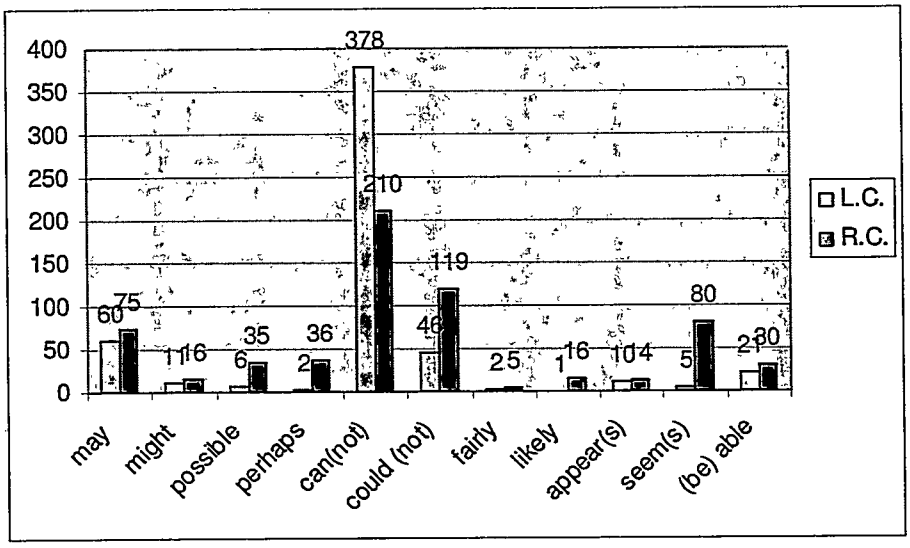


Figure 4.24. Hedges in learner and reference corpora.

The well attested data given as examples of overproduction or underproduction reveal that learners' lexicology lies between two extremes (overuse or underuse). While Figure (4.23) presents the markedly overproduced lexemes in the interlanguage corpus, Figure (4.24), on the other hand, shows some markedly less frequent lexemes. Again, the explanation of the underused lexemes shown above might feasibly be understood with reference to the learners' L1 rhetoric. The criteria used in sorting out and counting the errors in the previous subcategory were applied to this subcategory, as well.

4.4.4 Non-Lexical Measures

(1) Essay length

Although this measure is charged with having the inability to ensure writing quality, (Reid 1990) argues that "in several studies with native and nonnative speaker writers, length of essay has correlated highly with quality writing" (p. 195). Essay length has attracted much attention as a quantitative measure to ensure proficiency or fluency of both learners and native speakers. As argued earlier, this measure is originally and primarily

concerned with speech and "is used as a synonymy of overall oral proficiency" (Chambers 1997:535). By extending it to writing, this measure is meant to include smoothness and continuity in writing:

> it demonstrates an ease of writing, a "scribal fluency" of keeping pen to paper with the obvious "halting" (Galvan 1986) that can characterize breaks in thought and coherence on the part of the writer; often, then, fluency is demonstrated by overall length of essay.(Reid 1990:195).

Relevant literature (e.g., Enber 1995, Larsen-Freeman and Strom, 1977, Larsen-Freeman 1978, Linunard 1986, Reid 1990) has shown powerful evidence of the interactivity between essay length and writing quality. Whether the findings presented in Table (4.12) are ascribed to the learners' limited word stocks, retrieval inaccessibility or carelessness, it is obvious that learners' fluency in L2 lies far below satisfactory levels. This conclusion comes from the findings of the untimed essays, where only one student (out of 160) met the minimum task requirement concerning the number of tokens (a 500-word essay). It should be mentioned that none of the learners in timed or untimed essays has matched the average of the NSs (889.99 tokens). Since this measure directly addresses the subject's fluency/proficiency, it should be carried out on individual bases rather than on a corpus basis. Table (4.13) presents the subjects' fluency/proficiency means in the two corpora.

Table 4.12. Mean of lexical proficiency in learner and reference corpora

	L.C.	R.C.	Difererence[a]
Mean	163.89	889.99	726.1**
SD	75.38	694.42	

[a]** Significantly different from zero (a = 0.05) using two-sided parametric t-test assuming equal variance.

As Table (4.13) reveals, there can be no doubt that learners' fluency/proficiency in L2 writing was moderate or even weak compared to the native speakers of the language. Whereas the mean of lexical items in the native corpus is 889.99, the mean of the lexical item in the learner counterpart is 163.89. This sharp drop in the means indicates that the lexical fluency of learners is only (18.4%) compared to the native speakers. It should be mentioned that the mean of the untimed essays (homework assignments), in particular, is 171.23. Having known that the subjects were asked to write a 500-token essay on the topic they prefer, the mean (171.23) bears witness of learners' inability to keep pen to paper.

It would be more appropriate to link the essay length average in both corpora with the learners' lexical knowledge. In other words, learners' lexical fluency, which in its simplest sense means keeping the pen to the paper (Engber 1992), is obviously far below the norm. It is possible, at this juncture, to argue that the learners' fluency percentages is likely to be far less than the percentage shown above. Two salient features of learners' writing might support this argument. First, the frequent repetitive tendency found in learners' writing in general and in the writings of Arab students of English, in particular. Second, learners prefer to overuse of lexical bundles, recurrent expressions, regardless of their idiomaticity, and regardless of their structural status (Biber et al. 1999:990).

(2) Sentence and Word Length

Another commonly stereotyped feature of the writing of Arab students' of English deals with sentence length. Educators usually complain about the marked length of learners' sentences compared to the NSs' norm. Oftentimes, the blame is placed over the coordinating conjunction and parallelism. However, by running the learner and the reference corpora on the *Wordlist*, it turned out to be that NSs' sentences are longer than

those of the learners. Figures (5.18) and (5.19) present the findings of sentence length in learner and reference corpora respectively.

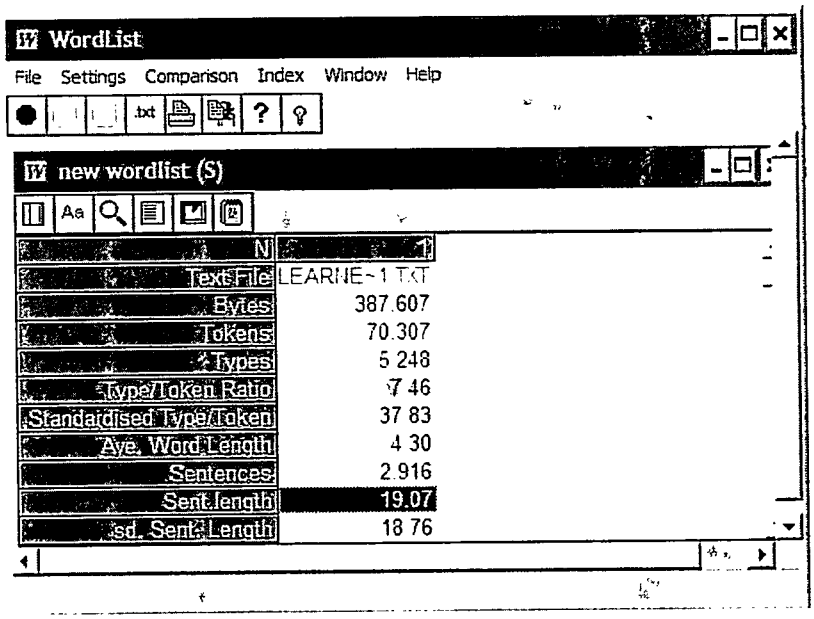


Figure 4.25. Sentence length in the learner corpus.

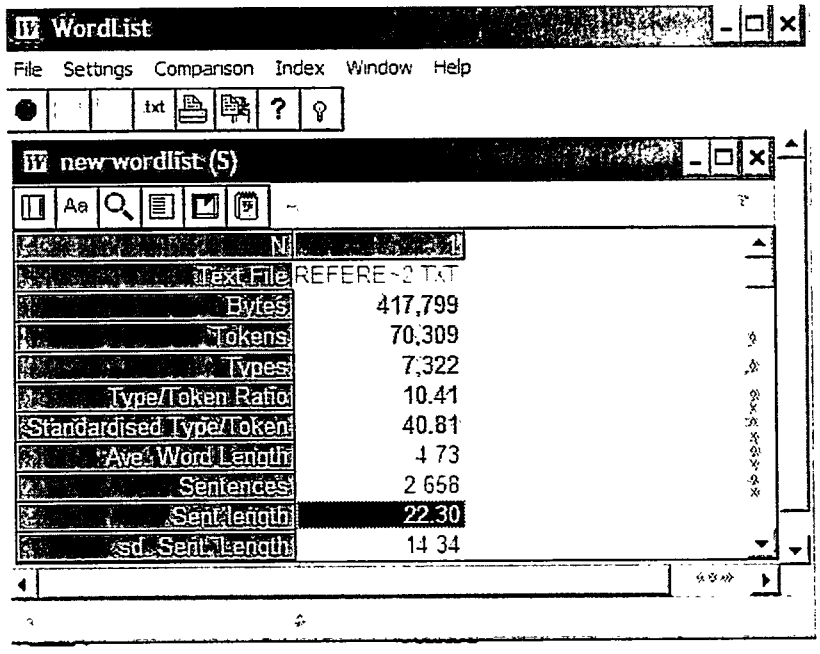


Figure 4.26. Sentence length in the reference corpus.

From a rapid scan of the figures, it becomes apparent that sentence length in the reference corpus (22.30) is longer than that of the learner corpus (19.07). Consequently, this subject calls the long-held erroneous impression among language educators about learners' (particularly Arab students of English) sentence length into question. Furthermore, as far as word length is concerned, it is obvious from the figures above that the average word length in the learner corpus (4.30) is shorter than that of the reference corpus (4.73). These figures resonate with the findings of previous literature (e.g Dafu: 1994).

(3) Paragraphing

Without going deeply into other technical aspects of their writings, learners' serious violation of the English paragraphing rules makes the learner corpus paragraphing closer to Arabic than to English. However, this, as illustrated below, shouldn't deny the

involvement of other factors such as an insufficient exposure to the L2 and teaching and learning strategies, where L2 proficiency is often sacrificed in favor of language simplification. Such a conclusion is based on the learners' violation of the following paragraphing rules:

1. Oneness of aim or unity: A paragraph in the learner corpus does not have a single thought as would be the case in the target language. Rather, it is widely noted that a paragraph may have several thoughts together, and it is also possible to have two or more paragraphs share the same thought.
2. Proportion: while the English paragraph rules require that "enough to be said to exhibit fully the purpose and ideas of the paragraph" (Scott and Denney 1909:18), it seems to be quite customary in the learner corpus to have a one-sentence paragraph, where no minimum evidence of elaboration is shown.

The attribution of the learners' writing inability in L2 (e.g., short essays, less complex sentences, less elaboration) to developmental stages and to insufficient exposure to the L2 (Kamel 1989, cited in Kubota 1998) might provide convincing explanation for these aspects. Yet, this explanation might not be taken for granted when it comes to other rhetorical features such as repetition (lexical and content), textual organization, parallelism, absence of paragraphing, under-paragraphing and overamplification, which are better linked to the influence of linguistic, cultural and rhetorical patterns of the L1 than to the developmental stages. The use of L1 rhetoric in the L2 immediately reminds us that learners are not aware of the linguistic and cultural differences in writing (Kaplan 1966, Buckingham 1979, among others).

3. Absence and under paragraphing: Though there is abundant evidence to the contrary, the absence of paragraphing in essays that have different topics, which often results in disorganization and the incoherence of ideas, does not seems to be stigmatized in the learner corpus. The figures presented in Table (4.13) present a brief paragraphing comparison between reference and learner corpora.

Table 4.13. Paragraphing in learner and reference corpora

Corpus	# of essays	# of paragraphs after eliminating the titles and numbers of essays	# of words/ paragraph	Average of paragraphs/essay
L.C.	429	1,275	52.72	2.97
R.C.	79	710	97.14	8.99

Compared with the number of essays in the learner corpus (429), the number of paragraphs found (1,275) is obviously very low. Figure (4.27) presents a random sample of a learner's essay that exemplifies a lack of paragraphing, incoherence and undeveloped ideas:

In-depth analysis of the one-paragraph essay cited above provides clear-cut evidence of several features, not least of which are the three points mentioned above (absence of paragraphing, disorganization of paragraphs and incoherence of ideas). How is one to explain these three points? Does the learner start writing without thinking about what s/he is going to mention in the next sentence? Is it an inevitable result of the minimal exposure to authentic texts in L2? Or can it be ascribed to L1 rhetoric? Even if we assume that all these factors are likely to be responsible in part, the frequent occurrence of such fragmentation throughout the corpus regardless of the academic level of the learner strongly suggests it is due to the L1, particularly when such features are less

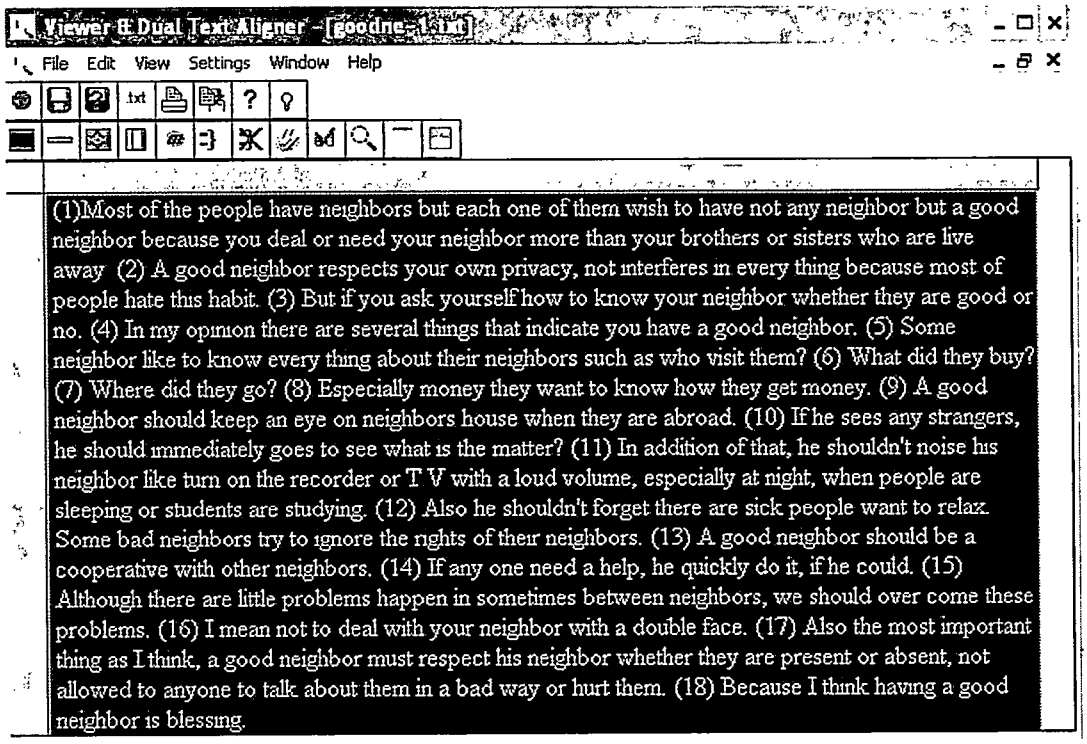


Figure 4.27. A sample of learners' writing.

valued in Arabic texts than in English. While lack of paragraphing is apparent (e.g., Figure (4.27) consists of one paragraph), other features need to be closely examined.

It is evident that since the learner is enumerating the traits of a good neighbor, an idea that should be preceded by an introductory paragraph, then sentence (2) should come immediately after (4). This means that sentence (5) is haphazardly placed. Sentences (9-14) should be placed after sentence (2). Sentence (17) should come after (14). Sentences and questions (5-8), together with (15) should be placed in a separate paragraph. Yet, it is possible to have them in the same paragraph if the comparison and contrast technique were meant to be employed.

As far as the idea of coherence is concerned, it is obvious that the learner is not using a strategy of linear development for the theme, which is the norm of the L2. The incoherence of ideas results in a scattering of focus and a loss of concentration. Support for these arguments stems from the sudden shift from enumerating the traits of a good neighbor (sentence 4) to talk about the traits of a bad neighbor (sentence 5) in a rather unorganized way.

Overall, this chapter has explored numerous aspects of lexical complexity and text-profiling in the learner corpus in comparisons with the reference corpus. Comparative corpus evidence has shown numerous areas of convergence and numerous areas of divergence between the two corpora. By addressing the results with reference to the previous literature, it was possible to delineate the features attributed to the influence of the L1 and the features that are likely to be classified under general tendencies.

CHAPTER 5

LEXICAL AND COLLOCATIONAL ERRORS RESULTS AND DISCUSSION

5.1 Introduction

Lexical analysis is a daring enterprise no matter what language is involved. Among other things, there are two notable reasons for such a statement. First, lexical errors, which tend to consistently outnumber any other type of errors in most of the recent studies conducted on SLA worldwide, are context-dependent. This signifies that much time and effort are required to go back and forth to the context of usage to identify such errors and then to categorize and quantify them. Second, insufficient research on lexicology, in general, and on interlanguage lexicology of Arab students of English, in particular, requires that each of the lexical errors, whether local or global, needs to be examined with extreme caution.

For the sake of organization, this chapter is made up of four sections, which appear in exactly the same order as the research questions (4-7) pertinent to learners' lexical and collocational errors. Again, the results of each research question are addressed with reference to the findings of the previous literature.

5.2 Results Related to Research Question (4)

Research Question (4): What are the most problematic words that Arab students of English encountered in the corpus?

Despite the centrality of the lexicon in language learning, previous attempts to provide lists of the problematic words that learners are likely to encounter in the course

of their mastering of the target language are almost nonexistent. As discussed previously, this is attributed to the gross negligence of the importance of lexicon, in general, and the, by then, prevailing beliefs of some influential figures (e.g., Fries 1945, Hockett 1958, Chomsky 1965), who believe that language learning means learning syntax and phonology but not learning vocabulary. In addition to the negative role of such beliefs on lexicology and its status in language learning, the prevailing methods of the past were unable to access a large body of representative naturally-occurring data. However, the advent of learner corpora, which are still in their infancy, is expected to open new pathways for the study of learners' lexicology and the investigation of the most problematic words that students are likely to encounter at different phases of language learning.

In view of these remarks, together with the descriptive explanatory objectives of this study, providing a lexical error list of the most problematic words in the learner as well as lexical translation corpora has been given priority here. Tracking the theme and lexical diversity in the learner corpus, together with the lexical translation corpus, has clearly made this study much more informative and appropriate for this end. Consequently, this anticipates the presence of lexical choice erors that belong to different domains of knowledge. Another positive aspect of the research is the huge number of students who participated in this study. This, of course, provides representative samples of lexical errors frequently committed by Arab students of English, though a relatively high percentage of these errors are likely to occur in learners' production no matter what their native tongue is.

There are two common ways to examine if a word is used appropriately in a given context: (i) by reading, word by word, the entire corpus/text and (ii) via concordancing. Oftentimes, the second method requires running the data on a frequency indexer to obtain a list of the words of the corpus and then examining the contextual use of the items in question. What primarily distinguishes concordancers from frequency indexers

is the environment, which enables users to easily go back to the context of usage and thus to examine the correct or incorrect usage of the item in question. Via this tool, it was possible here to examine the context thoroughly and then to extract the lexical and collocational errors presented in the learner corpus and the lexical translation corpus, as well.

Table (5.1) presents 100 randomly selected word types from the 761 most problematic lexical items found in both the learners as well as the lexical translation corpora. (See Appendix A for a complete list of learner' lexical errors.) This list, which is probably the longest list done on the interlanguage lexicology of Arab students of English, provides an insight into the compensation strategies employed by the learners in their attempt to evoke the needed lexeme.

Table 5.1. Samples of learners' lexical errors

No.	Target lexicon	Learner's lexicon
1	abbreviations	shorts, contracts, cuts, reductions
2	absorb	suck, drink, swallow, take
3	accredited	independent, trusted, reliable adopted, authorized, commissioned
4	acquired	earned, gained, obtained
5	adjourn	raise, delay, finish, postpone, move, close, defer, lift, put up
6	advanced	high, old
7	application	request, order, demand
8	appreciate	estimate
9	assassinate	abdicate, murder, kill
10	balance	arrange, coordinate, stabilize
11	bald	without hair on his head, bold
12	beggars	not rich, poor people who keep asking others for assistance
13	beneficiary	advantager, useful person, user, benefiter
14	betray	break oath, lie, perjury, cheat
15	big	great, old, large
16	bills	counts, invoices, vouchers, fawateer, fees
17	board	council, group, members
18	calm (sea)	quite, smooth, not noisy, relax
19	capital	head money, beginning money
20	challenge	resist
21	chamber	room
22	chivalric	heroic, great, brave, horsical , knight
23	climate	weather, atmosphere
24	coma	shock, comma, unconscious, not awake, absence of mind
25	compensate	pay back, repay

Table (5.1) Continued

No.	Target lexicon	Learner's lexicon
26	cruel	tough, aggressive, hard, rough, rock heart, merciless, hard-hearted, rigid, without emotions
27	debate	negotiate, discus, talk, argue, dialogue
28	decree	will, wish, wanting, order, intend, permission
29	deliberate	slow, unhurried, careful, quite, leisure, rational, late
30	discount	cheap, reduction, cut of prices, low down, sale, decrease
31	discriminate	separate, distinguish
32	donate	give, offer, grant, gift
33	drop	increase, decrease, low, reduce, fall, got down,
34	duty	homework, job
35	editor	liberator, author, director
36	environment	nature
37	escalate	rise, increase, aggravate, rise, elevate, make high
38	exclusively	only, private, limited, especially, on the face to monopolization
39	exempt	exceptional, free, pardoned
40	exercise	practice, sport, play sport
41	expire	end, finish
42	faithful	fixed
43	fatal	leads to death, killed, deadly, lethal
44	fetus	child, baby
45	fiscal	money, financial
46	float	spread over water, over flow, swim
47	foot	leg
48	forgiveness	excuse, amnesty, tolerance, pardon, mercy, excuse
49	gap	space, hole, distance, dash
50	grow	increase

Table (5.1) Continued

No.	Target lexicon	Learner's lexicon
51	harmful	so bad to health, unhealthy, dangerous for health
52	heat	warm
53	hospitality	generosity, welcoming visitors, receiving guests, hostility
54	illegal	unpermitted, against the law
55	immature	children put in bottles, not complete, pre-time, incomplete, minor, children who are in glass house
56	immunity	power, protection, security
57	infected	effected
58	intermittent	from to time, non-continuous
59	invent	find, discover, create
60	job	task, work, assignment, function, career, work
61	listen	Hear
62	lose face	loose the water of his face, mis his shame,
63	mammals	animals whose children depend on milk, creatures that are born by eggs
64	might	possible
65	missed	lost
66	needy	people who need money, poor
67	opponent	opposition, competitor, enemy, antagonist, rival, against
68	overcome	finish, pass, exceed, get over, cross, (over) step
69	pantry	store
70	patent	invention innocence, invention purification
71	peak (times)	climax, top, summit, afternoon, first, hard, great, difficult
72	plain	bitter, unsweetened, black, without sugar, dark coffee, sick
73	poll	questionnaire, opinion search, gather opinions
74	polluted	not clean, dirty, spoiled, unclean
75	prescribe	describe, give, write formula

Table (5.1) Continued

No.	Target lexicon	Learner's lexicon
76	prey	Victim, weak creatures that are easily eaten by other animals
77	priest	Church man
78	promotion	preferment, become upper in his position, raise, lift
79	quit	leave
80	rate	average, scale
81	regain	return, give back, come back
82	retail	separate, single, partial selling, individual, small, alone
83	return	repair
84	scattered (showers)	few, small amount, little, different, separate
85	self-determination	fate deciding, final destination, end decision
86	separated	divorced
87	session	cycle, circulation, meeting, round, period
88	sever (v)	cut (of), stop
89	sick leave	ill(ness) vacation, illness permission, sick holiday, illness rest, ill absence
90	slavery	godless, worshipping, idol, lack of freedom
91	smuggling	transportation, passing illegally, escaping, running. trading
92	solar	sun
93	spread	Separate
94	stimulate	develop, encourage, courage, helps, motivate, trigger, give, support, enhance, activate
95	superficial	minor, easy, external, surface, shallow, flat, ceiling
96	tires	wheels
97	trustee	honestee, safer
98	unsurpassed	unequal, unrival, unbelievable, unique, unseen, incomparable
99	warranty	guarantee, wheels
100	younger	small

A quick examination of these errors reveals the dominance of errors attributed to near-synonymy, paraphrasing, word-match, and literal translation in comparison to other error sources. Perhaps the excessive number of errors attributed to these subcategories over others is due to the confusability of near-synonymous words and the learners' attempt to fill in a lexical gap while, at the same time, remaining closer to the intended meaning.

Though providing a blind word list has some advantages for both researchers and learners, it is better for both of them to gain access to the environment or context of usage, where they get a comprehensible idea about the error in question. To this end, Figure (5.1) displays an edited concordancer of some of the learners' errors found in both learner and lexical translation corpora. A close look at the correct lexica given between parentheses shows that these errors represent mots types of errors represented in the corpus.

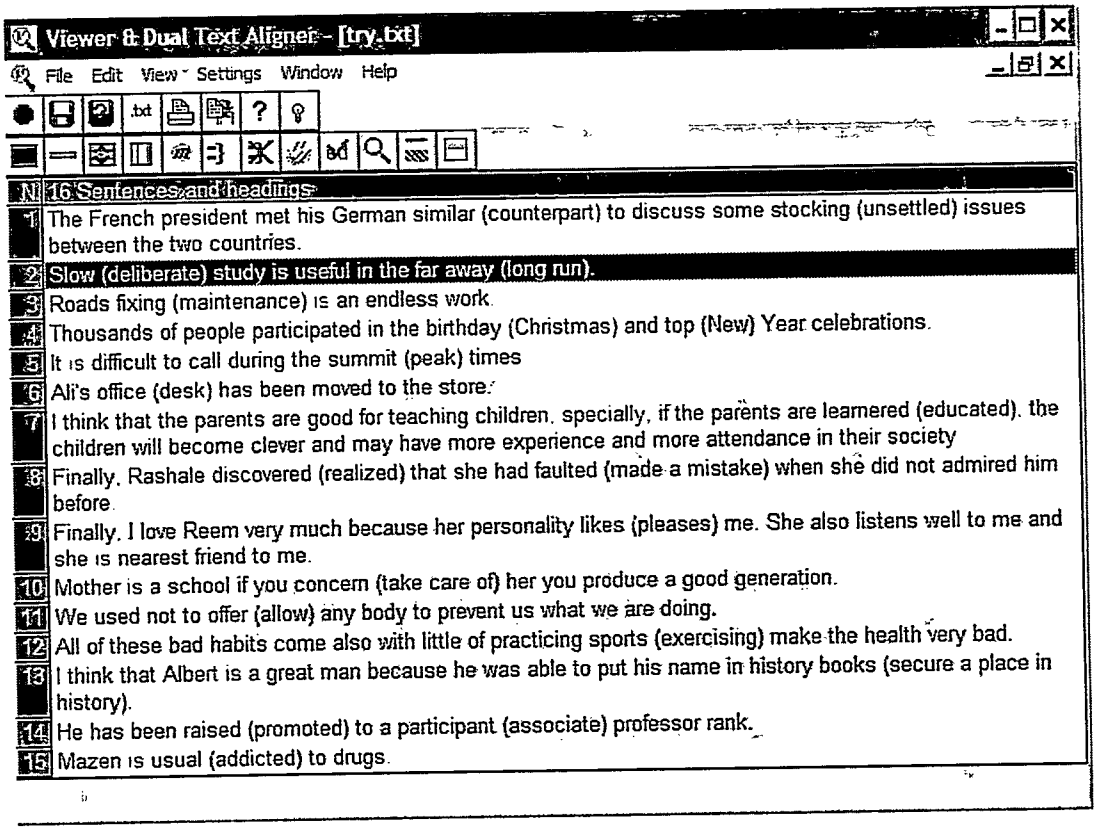


Figure 5.1. Samples of learners' errors.

A check back with the corpus demonstrates that what might appear to be easy to infer or understand in the list above is more often than not misleading, either locally or globally. This results shows the large gap between the learner's interlingual and the native speaker's norm. As a general precaution against broad generalities, we avoid such guessing in favor of using the statistics presented in section 5.3 below.

5.3 Results Related to Research Question (5)

Research Question (5): What are the categories of learners' lexical errors? And what is the contribution of each category to the total number of errors?

Indeed, it would be inconvenient to start discussing learners' lexical errors individually as presented in Table (5.1) or Appendix A. Rather, a more powerful tool for understanding and gaining control over the investigation of such divergent errors lies in providing a taxonomy for the types of errors and compensation strategies employed by the learners in their attempt to come up with the intended lexeme. A point to be made clear is that the reason behind not allotting a separate category for collocational errors in Table (5.2) is that such errors do not share one source or strategy. Rather they are, as shown in section (5.3), attributed to a number of lexical sources and strategies (e.g., near-synonymity, word-match, synforms, creativity, paraphrasing). For this reason and in order to avoid redundancy, it was more appropriate to classify them directly under the sources or strategies they share.

Once the types of errors are classified, the next phase, dealing with frequency counts begins. Categorizing and quantifying lexical errors as shown in Table (5.2) helps researchers as well as readers identify:

- the errors that occur most frequently in learners' performance
- the errors that are likely to impair comprehension significantly

A comparison of the numbers presented in Table (5.2) demonstrates that learners use a varied set of compensation strategies to bridge lexical gaps. As intimated earlier, the percentage of near-synonym errors is markedly higher than any other type in the taxonomy. Another important distinction to note is that intralexical errors significantly outnumber the interlexical ones. This provides further evidence against the erroneous assumption of the strong version of the CA hypothesis, which claims that the mother tongue is the principal barrier to SLA. Although word-match and literal translation occupy second place in terms of the number of errors, they are still far less frequent than the number of those attributable to near-synonymy. Avoidance and overproduction and verbosity characterize advanced levels of learners, too.

Table 5.2. Taxonomy of Lexical Errors

Category				Freq.	% of Total
intra-lexical	lexical meaning (sense relations)	near-synonymy		1,786	28.1
		high lexicon		85	1.3
		hypernym, hyponym converse and metonymy		123	1.9
	lexical form	synforms and homophones		267	4.2
		close forms		71	1.1
	creativity			217	3.4
inter-lexical	negative transfer	word-match and literal translation		992	15.6
		simple word transfer		6	0.1
		rhetoric	repetition	436	6.86
			overproduction and verbosity	411	6.5
			underproduction	64	1.01
		overdifferentiation		3	0.05
		intention match		289	4.5
paraphrasing	circumlocution and approximation			813	12.8
idioms and idiomaticity				174	2.7
avoidance				621	9.8
Total				6358	100

The foregoing discussion suggests that understanding learners' errors requires four consecutive steps: *identification* or *extraction*, *categorization*, *quantification* and *explanation*. Having already commented on the first two steps, it is now time to address the third step. Much of the remainder of this section is taken up with examining qualitatively the data presented in Table (5.2). As seen from the long list of lexical errors, along with Table (5.1), these errors are divided into five major categories.

- Intralexical errors
- Interlexical errors
- Paraphrasing

- Idioms and idiomaticity
- Avoidance

5.3.1 Intralexical Errors

This section, which consists of three major subcategories, *lexical meaning, lexical form* and *creativity*, addresses intralexical errors, deviant items that occur as a result of the influence of one or more of what Laufer (1990a) calls *intralexical factors* (a set of intrinsic features related to the form and/or meaning of a given word) or the overextension of learners' previous knowledge to new rules where it is inapplicable. Like intralingual errors, the occurrence of intralexical errors is not limited to the NNSs. Rather, they are likely to be found in the performance of NSs in some early stages of development or as a result of fatigue or rashness. The importance of the lexical meaning or sense relations subcategory stems from its being the highest in terms of the percentage of errors, not only within the intralexical category, but also within all other categories in the taxonomy.

5.3.1.1 Lexical Meaning/Sense Relations

Evidence from neurolinguistics suggests that human beings store words in their mental lexicon in terms of sense relations (James 1998: 151). Consequently, it is feasible to approach and categorize a considerable number of lexical errors in terms of such relations. In order to make this idea a bit more concrete, it will be useful to analyze a sample of learners' errors attributable to five sense relations, namely, near-synonymy, high lexicon, hyponym-hyperonym, metonymy and converses. Altogether, as Table (5.2) demonstrates, these errors claim (31.36%) of the total number of errors.

(1) Near-synonymy

It seems that imperfection is irradicable in all aspects of life, including language. This failing explains the recent attempt to seek out idealism in semantics, which might be captured by the "one form-one meaning" slogan, which was discussed by Geeraerts (1997), who recommends the "isomorphic principle" for fulfilling this purpose. However, the widespread occurrence of polysemy and synonymy and other sense relations make this seem illusory.

Synonymy in its strict sense- -two words that can (in a given context) express the same meaning including all meaning variants for two polysemous lexemes and all meaning parts, i.e. descriptive, social and expressive meaning- -is difficult to find in any variety worldwide. Yet, partial synonymy, two words that have one meaning variant in common, is quite possible (Lobner 2002:46). Crystal (1996:164) states that "there may be no lexemes which have exactly the same meaning." Rather, there is usually some nuance that separates them, or there is a context where one of the lexemes can appear while the other cannot. *Autumn* and *fall*, for instance, are synonymous, but the former is British English and the latter is American English. Also, while *salt* and *sodium chloride* are synonymous, "the former is everyday and the latter is technical" (p. 164).

Jackson and Amvele (2000:93) furnish further support for the aforementioned arguments by rejecting the possibility of having complete synonymity between any two existing lexemes in a living variety:

> Strict synonymy is uneconomical; it creates unnecessary redundancy in a language. To have a completely free choice between two words for a particular context is a luxury that we can well do without. Indeed, it would appear that where, historically, two words have been in danger of becoming strict synonyms, one of them has either changed its meaning in some way or fallen out of use.

As the findings indicate, no factor, whether intra- or interlexical, has had more negative influence on learners in terms of incorrect lexical choices than near-synonymity.

The frequent occurrence of near-synonymous errors is mainly, but not exclusively, attributable to the learners' insufficient exposure to native speakers or authentic texts, where they could acquire sufficient knowledge of the contextual and collocational use of related lexical items. The occurrence of near-synonymous errors in the output of all learners (regardless of their native tongue) makes this sense relation a universal challenge for learners. What makes the problem of synonymy seemingly insuperable is its connection with learners' level. That is, it increases in the performance of advanced learners (Martin 1984). The following figures present various instances of near-synonymous errors, which are tagged with NS.

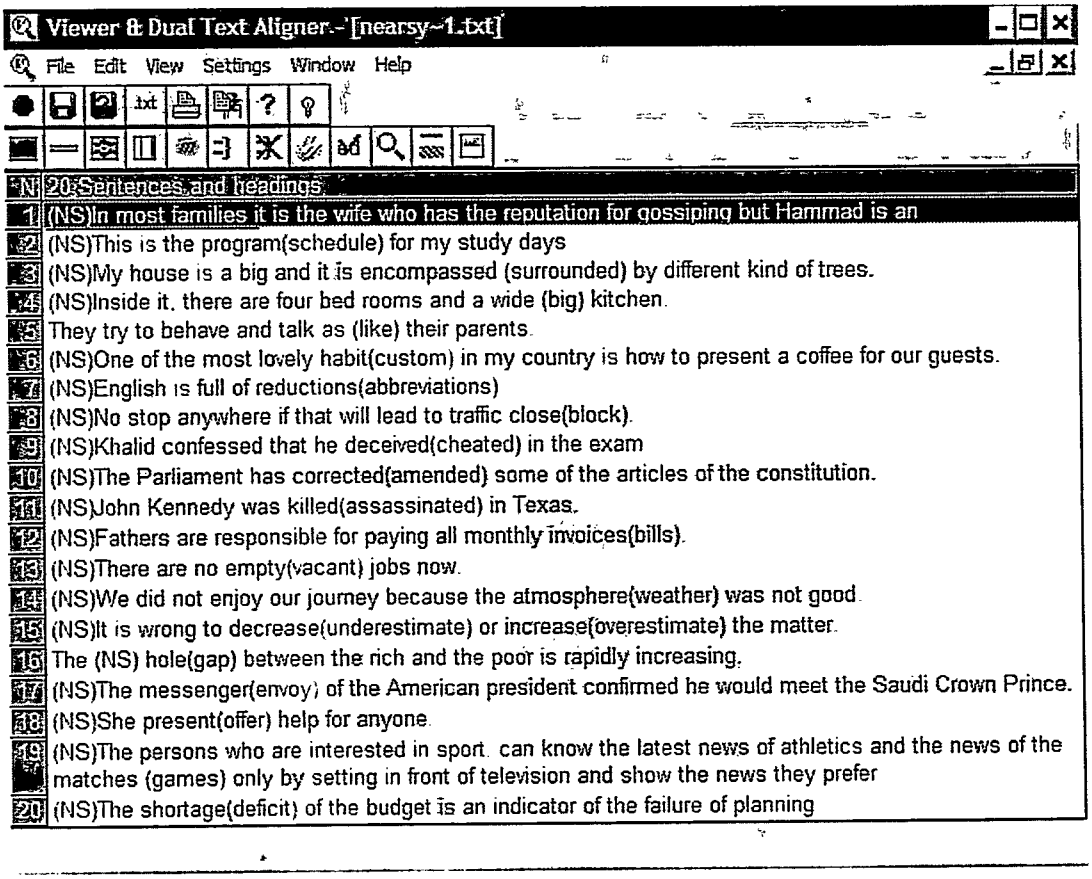


Figure 5.2. Errors attributed to near-synonymity.

It is relatively easy to see that each of the sentences in Figure (5.2) contains at least one lexical error attributed to the near synonymity between synonymous lexemes in the target language. Sentence (1), for instance, shows that the learners has incorrectly used *excluding* to stand for *exception* based on the assumed synonymity between the two lexemes. It is probable that the learner has successfully identified that the missing target word should be a noun and was close enough to the target meaning, but s/he failed to distinguish between the two lexemes contextually. Yet, it would be incorrect to assume that the meaning of near-synonymous errors is always easy to guess and can be immediately obvious from the context without affecting the intended message as we saw in (1). Rather, in many situations, such errors lead to a partial or global misunderstanding as is easily seen in (10). Whereas *correct* and *amend*, the correct target, are generally synonymous, the receiver of (10) seems to favor two different interpretations rather than one if s/he is given the same sentence with these two different lexemes. Due to the inherent differences between the two lexemes, the average native speaker knows that the synonymity between *amend* and *correct* is partial. Again, semantically speaking, sentence (15) does not run smoothly. The substitution of *underestimate* and *overestimate* for *decrease* and *increase*, respectively, might create a momentarily global misunderstanding. What drives the learner to use *decrease* and *increase* instead of *overestimate* and *underestimate* is the complete synonymous relationship between them, at least, from the his/her perspective.

The high percentage of errors attributed to near-synonymity (28.1) is on par with results in the literature. Zughoul (1991) found that this category claims the highest percentage of all lexical errors. Also, this category has the highest percentage of collocational errors in Farghal and Obeidat (1995). The higher percentage of its occurrence in learners' performance, in general, justifies its being the focus of numerous studies during the last two decades (e.g., Martin 1984).

While it is justifiable to attribute near-synonymous errors to the insufficient exposure of learners to authentic texts, one should consider other possible factors, such as *transfer of training*, whereby the lack of differentiation between near-synonyms is related to teachers (Selinker 1972). Evidence in the literature also shows that bilingual dictionaries, which often lack contextualized examples, participate negatively in the spread of the phenomenon. While dictionaries are indispensable tools for learners they are, according to Kjellmer (2003), misleading, especially when demonstrating how words are used idiomatically by native speakers. Kjellmer argues that dictionaries define synonymous words in terms of each other, while, in fact, they are not totally synonymous. There are three aspects to consider when it comes to the distinction between near-synonymous words: (i) their frequency, (ii) their style and text type preference and (iii) their collocability. A check back with the near-synonymous pairs discussed above provides consistent evidence that none of them are completely synonymous or interchangeable in all usage contexts.

A final observation on synonyms might be made about the harmfully oversimplified vocabulary lists that learners use throughout their study phases. More often than not, tolerance of the learners' lexical errors results in paying less attention to semantic as well as contextual differences among the related words. The problem of oversimplification has other harmful sides effects, as well. Fox (1979:68) argues that:

> It is my observation that many English as a Second Language Programs are harming their students by dealing for too long only with simplified structures and simplified vocabulary. The result is that when the students leaves our programs, they are actually far from being able to read unsimplified English which they are expected to read. The gap between the academic English they are now expected to understand and the simplified English they have been taught is too great.

(2) High lexicon

In her influential article, *Advanced Vocabulary Teaching: The Problem of Synonyms*, Martin (1984:131) argues for the connection between culture and high style of writing:

> students from cultures that require "high" style of writing produce a prose resembling this excerpt from a composite of student efforts to describe their ride to school: we board a bus, not waiting for the tardy ones, who rush lest they should miss it, vantage seats are sought and occupied, and much advice is tendered as over half a century of souls swarm into the vehicle.

Indeed, the diglossic situation in Arabic (the side by side existence of a prestigious or formal variety and a common or colloquial variety of a language) feeds the interactivity between learners' linguistic and cultural backgrounds and their attempt to use a high lexical style in the L2. Consequently, learners, as illustrated below, prefer to use high lexical items instead of the common ones, assuming that such words will have the same effect on the L2 audience that Arabic pretentious classical big sounding words have on their own L1 audience.

The use of high lexicon in the interlanguage literature of Arab students of English is confirmed in Zughoul (1991), who uses the term *verbosity* to refer to this subcategory. In their attempt to come up with lexical items that make their writing seem more impressive and literary, learners, he argues, tend to incorporate long and artistic words. Crucial support for to this argument might come from learners' L1 where the use of pretentious words is a key factor in the evaluation of a piece of writing. Examples of a high lexicon, which comprises 1.34% of the total errors, are shown in Figure (5.3). The extracting of lexical items classified under this subcategory would be tedious without the use of both the frequency indexer, which enabled us to identify all such pretentious words, and the concordancer, which enabled us to examine the contextual use of such items.

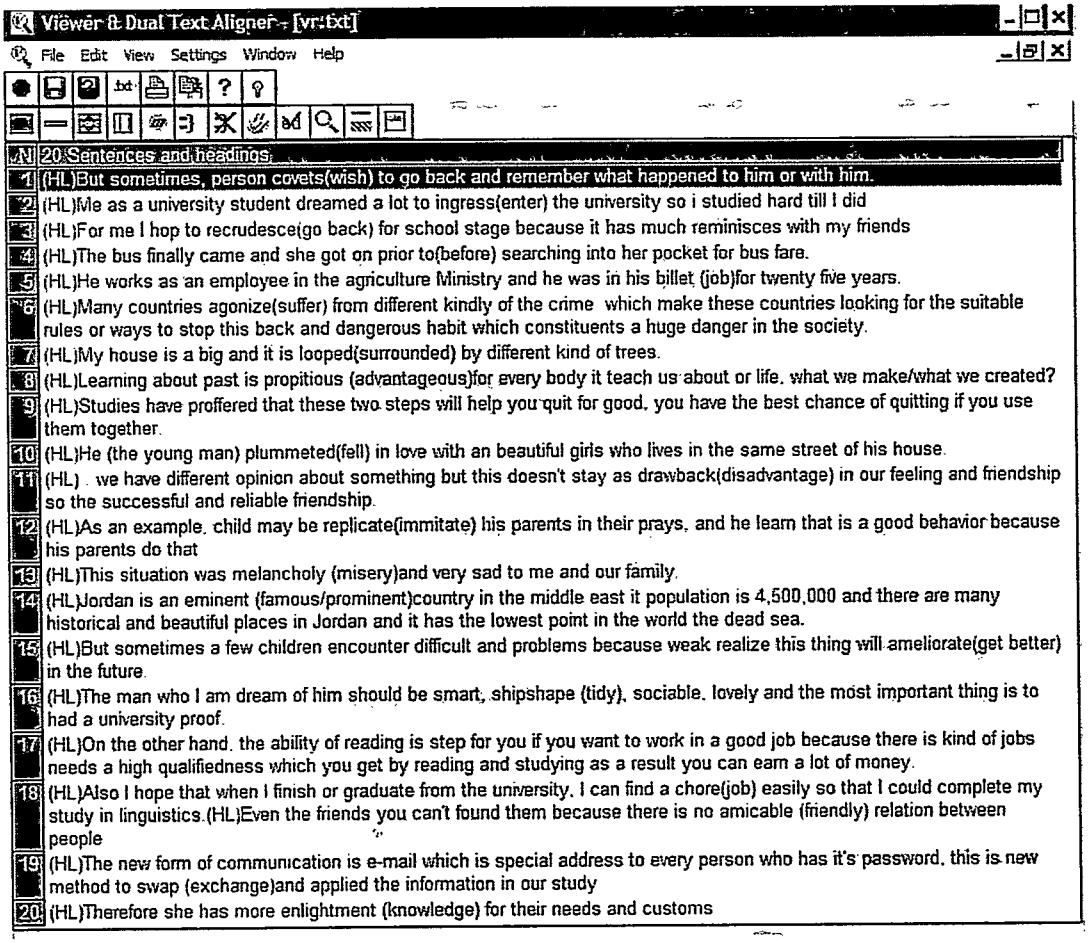


Figure 5.3. Example of errors attributed to high lexicon.

From the data given in Figure (5.3), it seems incontrovertible that each lexical item next to the lexical item in parentheses is an example of high lexicon. Due to the similarity between this category and the preceding one, then, it is useful to question whether it is possible to classify high lexicon errors under the near-synonymy category. The answer to this question is affirmative, but it is not preferred. To justify the classification of these errors under a separate subcategory, a questionnaire with five multiple-choice questions was used (See Appendix J). In each question, thirty randomly selected students were asked to choose lexical item that they prefer to use in formal communication, assuming

that all three items are completely synonymous. Two of the given items (e.g., *strength* and *force*) are well-known to them, while the other (e.g., *potency*) is somewhat unfamiliar. 104 out of 150 obtained answers favored the items classified under *high lexicon* category. Such findings (which explains the use of *covets* instead of *wishes* in sentence (1), *melancholy* instead of *misery* in (13), etc.) influenced the preference for classifying these errors under a separate category.

In summary, learners' use of high lexicon might stem from their L1, where employing classical and impressive words is widely preferred in writing as well as formal speech. Yet, this reason is not likely to exclude another possible reason, whereby using a high lexicon aims at demonstrating the learner's advanced level in the target language.

(2) Hypernym-hyponym, metonymy and converse relations

Learners' writing is characterized by the frequent use of generic lexical items (e.g., *people*) to stand for more specific ones (e.g., *men*) and vice versa. Such senses are often referred to in literature as *hypernym* and *hyponym*. For Lobner (2002:85), "an expression A is a hyponym of an expression B iff the meaning of B is part of the meaning of A and A is a subordinate of B." Crystal (1996) illustrates this relation with the following tree diagram.

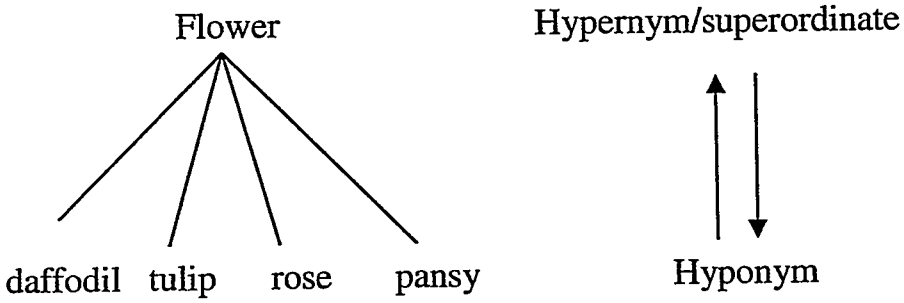


Figure 5.4. Hypernym-hyponym relation.

Drawing on the tree diagram, *daffodil, tulip, rose* and *pansy* are hyponyms and the top lexeme *flower* is the hypernym. Thus, the meaning of the more specific items (*daffodil, tulip, rose* and *pansy*) includes that of the top/superordinate lexeme *flower*.

Though they do not match the hyponym-hyperonym sense relation in terms of the number of occurrences, the occurrence of converses, a sense relation which refers to the same relationship from opposite viewpoints, indicates that they are, to some extent, problematic for learners. Figure (5.5) presents some of the examples of hyponym-hypernym, metonymy and converse relations found in the learner corpus.

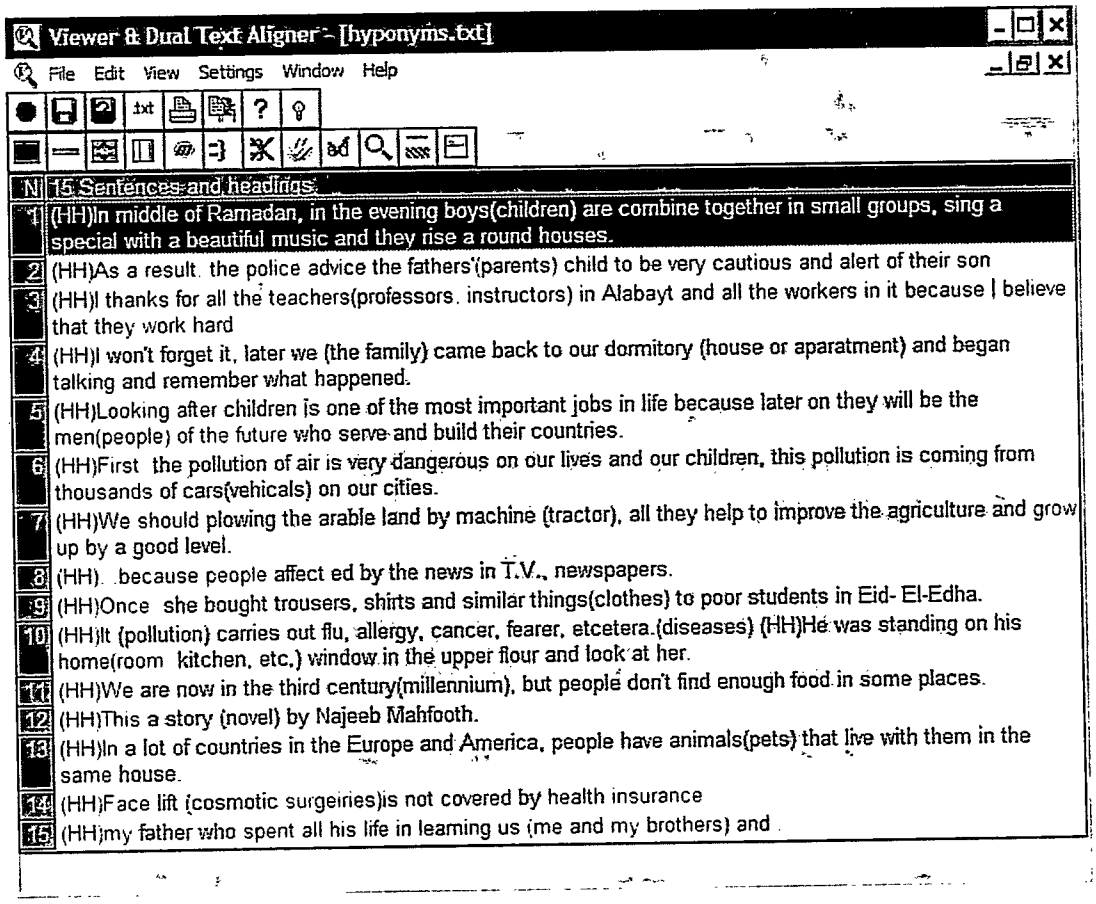


Figure 5.5. Errors attributed to hyponym-hypernym, metonymy and converse relations.

A careful investigation of the given data shows that all types of errors exemplified in Figure (5.5) can be accounted for in terms of the aforementioned sense relations. In sentence (1), for example, the learner has used the included *boys* to stand for the superordinate lexeme, *children*. While the classification of this category under intralexical factors is consistent with the literature, much evidence indicates the possibility of classification of such errors under interlexical factors. A clue to how this kind of error might be classified under interlexical factors rather than intralexical factors comes from sentences (1) and (5). The use of the lexemes *boys* instead of *children* and *men* instead of *people* might be attributed to the deeply rooted masculinity of the learner's L1, where masculinity is almost always the default case in all language domains. For instance, all Arabic verb roots have an inherent masculine gender. In the context of language masculinity, it is important to mention that some Arabic words change their meanings when feminized. For example, the meaning of the masculine form *muSib* 'he is correct' becomes 'she is a diaster' when feminized to *muSibah*. Also, the meaning of the *naaib* 'he is a representative' (masculine) becomes 'she is a diaster' when feminized to *naaibah*. Furthermore, the use of the masculine forms to stand for feminine ones is possible (e.g., *'abawein* 'fathers' is used to stand for 'parents' while it is impossible to to use *mothers* to stand for the 'parents').

Unlike sentences (1) and (5), sentence (7) presents a counter example, where the hypernym *machine* is used to stand for the hyponym *tractor*. A possible explanation for the use of *machine* instead of *tractor* might be that the learner was just literally translating from his spoken variety of Jordanian Arabic, where the word *makineh* (machine) is often used to stand for the word *tractor*.

Sentence (11) exemplifies another sense relation called *metonymy* (a part-whole relationship), whereby the token *century* is used to stand for *millennium*. Another example of this relation is clearly seen in (14), where *face-lift* (hyponym) is used to stand

for *cosmetic surgery* (hypernym). The last sentence (15) exemplifies a converse relation, where *learn* is erroneously used to stand for *teach.*

5.3.1.2 Lexical forms

Errors attributed to incorrect selection of forms are apparent and tend to appear regularly in the learner corpus. Such errors are classified here into three categories, namely, 'synforms' (similar lexical forms), 'homophones' (similar lexical phones) and 'close forms'. Laufer (1992, cited in James 1998:145) identifies six features in which pairs of synforms (e.g., *floor* vs. *flour*) can be similar: they can (i) have the same number of syllables, (ii) have the same stress pattern, (iii) be of the same word class, (iv) have the same initial part, (v) have some phonemes in common, and (vi) have phonemes with shared features. Homophones, on the other hand, refer to a falsified use of a lexical item that is phonetically similar to another one in the target language (e.g., *buy* vs. *by*). However, in most cases neither synforms nor homophones lead to any kind of global misunderstanding. All other instances of incorrect selection of forms (e.g., *fluence* vs. *influence*) are classified under close forms. It should be made clear that the classification of *suite* and *suit*, and *buy* and *by*, which are homophones, under two different categories (synforms and homophones) is ascribed to their different word classes; whereas *suite* and *suit* belong to the same word class (nouns), *buy* and *by*, on the other hand, belong to different word classes (verb vs. preposition).

In all the examples shown in Figure (5.6), there is an obvious confusion over the lexical form of the given words. From a semantics perspective, misunderstanding resulting from these confusing words is often resolved from the verbal or nonverbal context. The substitution of *flour* for *floor* in sentence (2), and *suite* for *suit* in (6) are ascribed to similarity among the forms of these lexemes (synforms). In numerous cases, however, errors of lexical forms are ascribed to the learners' inability to distinguish between close

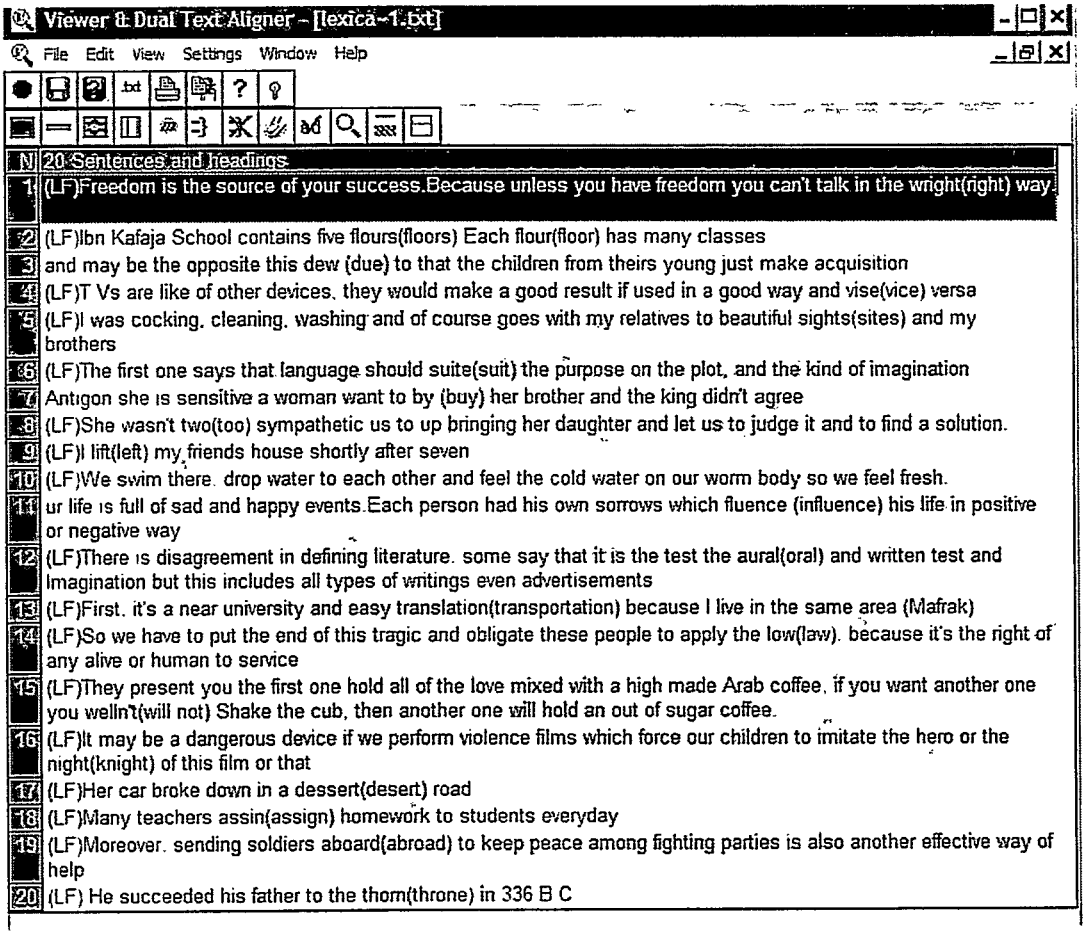


Figure 5.6. Errors attributed to lexical forms.

vowels in the target language as shown in the word *lift* in (9). Homophones are clearly shown in (3) *dew* vs. *due* and (7) *by* vs. *buy*. Close forms, on the other hand, are shown in (11) *fluence* vs. *influence* and (13) *translation* vs. *transportation*.

In the context of lexical errors, although beyond the scope of this study, it is relevant to mention the numerous distortion errors that occur due to the application of one of four potential actions described by Dulay et al. (1982). These include errors attributed to *omission* (e.g., *comftable*), *overinclusion* (e.g., *dinning*), *misformation* (e.g., *delitous*) and *misordering* (e.g., *littel*). James (1998), who is not satisfied with the label of *Surface*

Structure Taxonomy suggested by Dulay et al. (1982), adds another type called *blends*, which is a "typical of a situation where there is not just one well-defined target, but two" (p. 111). But can all errors that result from the application of these types be classified under the *lexical forms* of the given taxonomy in Table (5.2)? Obviously, the answer is negative, because that taxonomy includes only the misselected lexemes resulting from synforms, homophones or close forms.

5.3.1.3 Creativity

Creativity in language learning, which seems to be an axiomatic aspect of learners' intelligence, often leads learners to create new nonexistent lexemes in the L2. Connor (1996) relates "creative construction" in language learning to several sources such as "learners' limited knowledge of L2, knowledge of L1, and knowledge of communication, the world, and other human communicators" (p.12). Creation of nonexistent lexemes is not restricted to learners, but rather it is also found in NSs' performance in early developmental stages (Politzer:1973:48). Errors of creativity are sufficient to justify Selinker's coined term, *interlanguage*, which is based on an in-between system that belongs neither to the L1, nor to the L2.

Figure (5.7) brings into focus the fact that learners' errors are not limited to the confusing existing synonyms, synforms or close forms. Rather, in their attempt to fill in a lexical gap, learners sometimes create, by false analogy, new nonexisting words in the target language. At the center of these strategies are false derivation, overgeneralization and word coinage, which together claim 3.41% of the total percentage of errors. The extension of a previously learned rule or strategy to new situations where it is inappropriate is the common feature among these three strategies. The term *creativity* is used here to capture the essence of these innovative processes. The examples shown in Figure (5.7) (tagged with CR) are instances of learners' creativity.

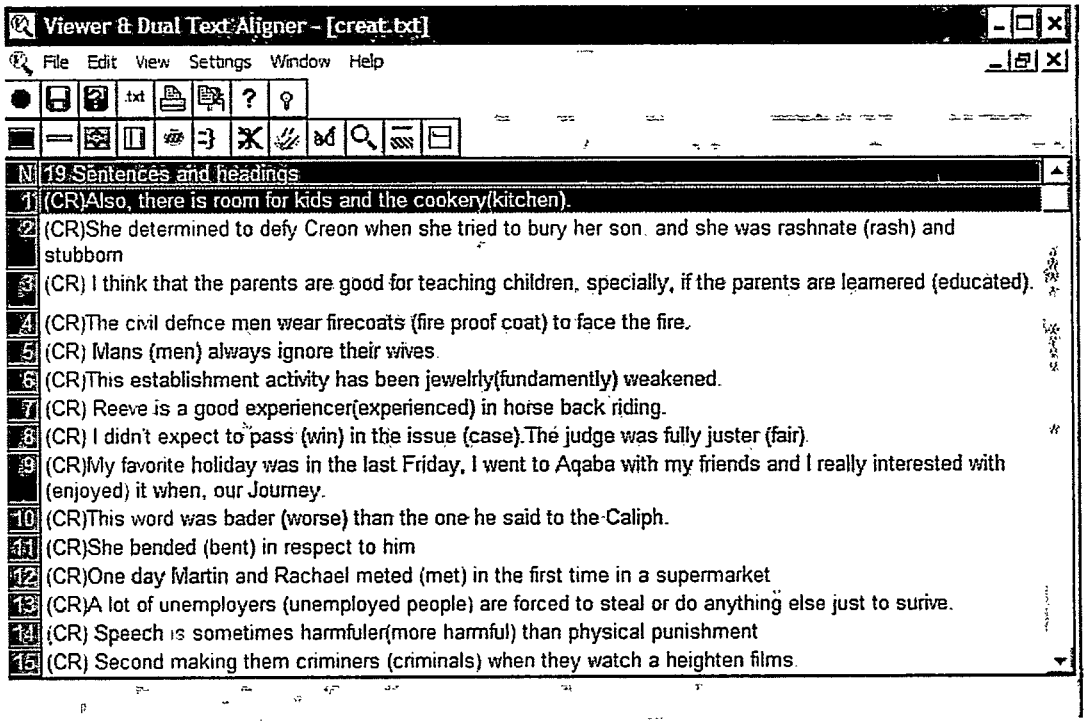


Figure 5.7. Errors attributed to creativity.

As example (1) demonstrates, the learner, by analogy, has derived the nonexistent noun *cookery* from the verb *cook*. The learner may have erroneously applied his or her previous knowledge of derivation of similar nouns from verbs such as *bake* (e.g., *bake-bakery*) to the verb *cook*. It should be mentioned that, though it is not used in American English, the noun *cookery* is used in British English to refer to the act or art of cooking, but not a place where *cooking* occurs.

Likewise, the extension of a previously learned rule concerning the addition of the *-er* suffix to derive nouns from verbs, may explain the incorrect output in (8). However, ascribing errors of creativity entirely to the influence of the rules of the target language on each other and fully freeing the L1 from the blame might be questioned. The abundance of errors attributed to morphological derivation in the learners' language, in particular,

demonstrate that learners are, to some extent, affected by their L1, whose productivity heavily depends on derivation. In view of these remarks, it is rather easy for most Arab students of English to overcome lexical gaps in the target language by deriving many unknown lexemes once one member of a word family is available for them.

In (4), the learner resorts to 'word-coinage' strategy to compensate for his/her lack of the lexeme in question. The creation of *firecoat* to stand for *fireproof coat* has resulted via a false analogy from a similar previously learned lexeme *raincoat.* Apparently learners produced numerous ill-formed lexemes by simply applying a similar previously learned rule or strategy in a new situation where it is inapplicable.

However, in sentence (6), the situation turns out not to be as clear as it might be since the derivation process was preceded by an intermediate stage. The learner apparently derived *jewelrly* from *jewelry* by a false analogy attributed to the native derivational process, whereby the learner derives the adverb *jawhary* 'fundamentally' and the noun *jawaaher* 'jewelry' and the adjective *jawhar* 'essential/fundamental' from the same root *jwhr*, and then by another false analogy s/he extended the previous knowledge of adverbial derivation in the target language (adding the *-ly* suffix) to the derived form. The result was a nonexistent lexeme *jewelrly.* In (14), it is clear that the learner has erroneously extended his previous knowledge of creating comparative adjectives with a one-syllable base to the two-syllable adjective.

Errors of creativity have received much attention in the literature under different titles such as overgeneralization, word-coinage, derivativeness and creativity (Richards, 1974, Zughoul, 1991, Yang and Xu, 2001, among others)

5.3.2 Interlexical Errors

To put it at its most basic, interlexical factors refer to the direct or indirect influence of previous linguistic knowledge, whether resulting from the mother tongue or any other previously acquired or learned language, on the lexemes of the language being learned. In numerous previous studies conducted on SLA, language transfer has been attested to correlate to the learner's level of proficiency. Several studies conducted on SLA (e.g., Pouslisse and Bongaerts 1994) found that the more transfer involved, the lower the proficiency in the L2.

Though it reaches back to the 1940s (e.g Fries 1945), the concern with interlexical errors, in its current sense, emerged in the literature that followed the pioneering article of Selinker, *Interlanguage*, in 1972. It is important to emphasize that although all the following subheads address the issue of transfer from different angles, their direct involvement in learners' deviant writings varies widely. This set of interlexical factors has witnessed, and is still witnessing, numerous changes due to ongoing research on contrastive rhetoric as well as contrastive learners' corpora. Throughout this section, we will explore various instances of interlexical errors.

5.3.2.1 Word-Match and Literal Translation

Among the six interlexical factors shown in Table (5.2), this subcategory, statistically speaking, is the highest in terms of the percentage errors 15.60%. As the title reveals, this subcategory consists of two strategies, namely, word-match and literal translation. *Word-match* is used here to refer to two words with distinctly separate meanings and uses in the target language that are expressed in the source language by one single word (Lombard 1997:61). As will be illustrated ahead, errors of this subcategory oftentimes result in lexical items that are neither the target words nor synonymous of the target

words. To expand on this a little, this subcategory may reflect the richness of one language compared with another one. As an exceptional borrowing language, English is clearly very rich in lexicon vis-a-vis any other language worldwide. When it comes to Arab students of English (whose native tongue is not as rich in lexicon as is English), this subcategory claims a huge number of errors, which are attributable to the lexical richness of the target language. Sometimes, however, it is apparent that the blame should be placed on the learners or their spoken variety, especially when the standard variety of their L1 makes a distinction between the two lexemes in question.

Literal translation, on the other hand, has received more focus in the literature than any other subcategory listed under interlexical factors. The learner simply replaces a word in the source language with another equivalent one in the target language by a direct translation. Oftentimes, this strategy results in obscuring the intended meaning as shown in Figure (5.8), where each error is tagged with (LT).

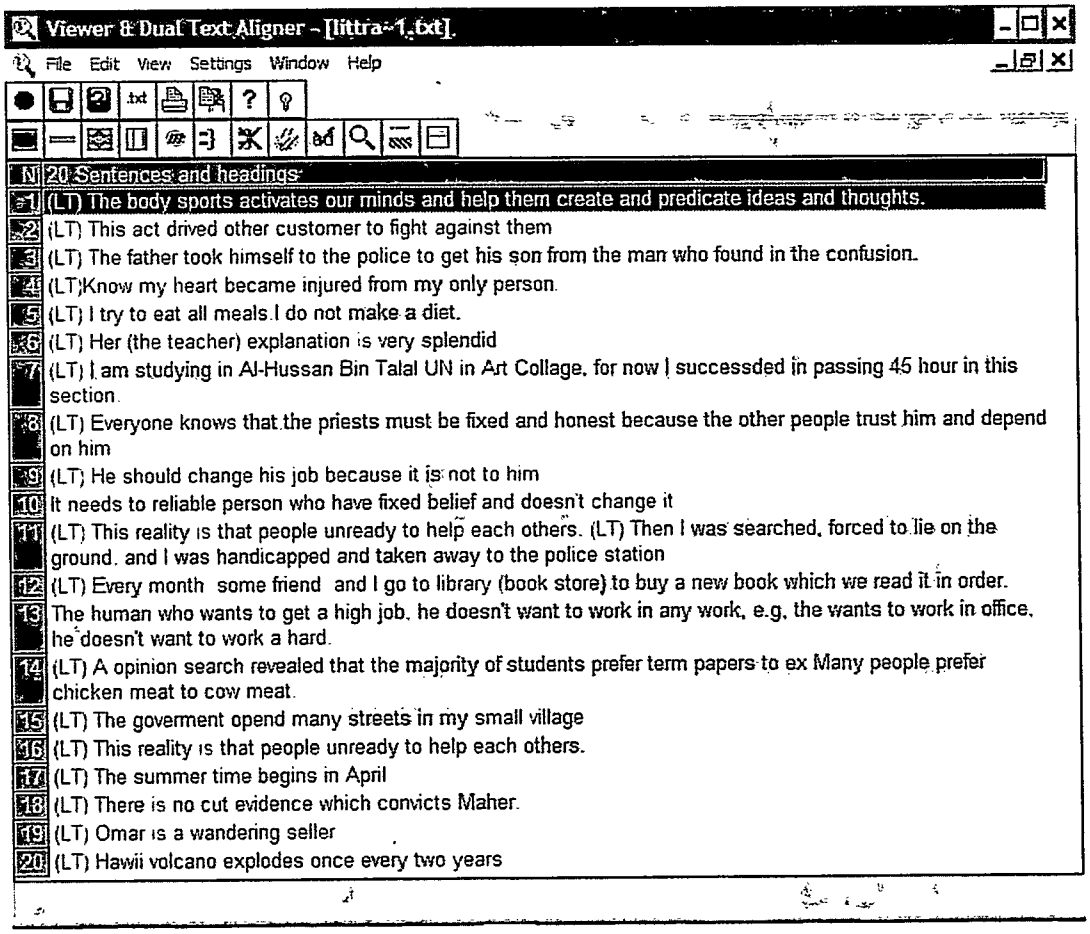


Figure 5.8. Errors attributed to word-match and literal translation.

As can be readily observed from the examples cited in Figure (5.8), each sentence has at least one lexical error attributable to either word-match or literal translation. Semantically speaking, many of these examples sound awkward to the native speaker, or any receiver whose native tongue is not Arabic. Accounting for the differences between the learners' sentence and the English norm requires a comprehensive understanding of the context of usage as well as a good command of the two languages.

(1) Learner's sentence: The *body* sport *activates* our minds and help them...

(1) Intended: *Physical* sports *stimulate* our minds and help us...

(2) Learner's sentence: The father *took* himself to the police to get his son from the man who found in the *confusion.*

(2) Intended: The father *went* to the police department in order to get his son from the man who took him during the *fight.*

Lexical choice errors in both examples are attributed to either literal translation or word-match. In sentence (1), for instance, the learner's use of *body* and *activate* instead of *physical* and *stimulate* (respectively) is attributed to word-match. In the second sentence, however, the use of *took* instead of *went* is attributed to literal translation. It is obvious that the learner was literally translating from his spoken variety into the target language.

wa 'akhath Halu lashurTah

and took himself (the father) to the Police.

'He (the father) went to the police department.'

A close look at the first eleven examples in Figure (5.8) indicates the learner was just translating from his native spoken variety without paying any attention to the notion of contextual appropriateness. Sentence (12), on the other hand, exemplifies a frequently committed error attributable to word-match. In Arabic 'library', 'bookshop' and 'book-store' are referred to as *maktabah* (library). So, the learner makes this error due to the lack of differentiation between the two target lexemes in his native tongue. Word-match is once again responsible for the production of the lexical error in (20). In Arabic, the verb *yathoor* collocates with 'volcano' to mean 'erupt', with 'soldiers'/'fighters'/'people'

to mean 'rebel', and with 'explosive' to mean 'bomb'.

5.3.2.2 Simple Word Transfer

As the title shows, this category, which claims only 0.1% of the errors, represents another compensation strategy, whereby the learner directly resorts to his/her native tongue to replace the missing (equivalent) word or expression in the target language in order to maintain communication. The use of this strategy, which is more common in speech than in writing, reflects not only a gap in the lexicon, but also a failure in employing any of the compensation strategies mentioned in Table (5.2).

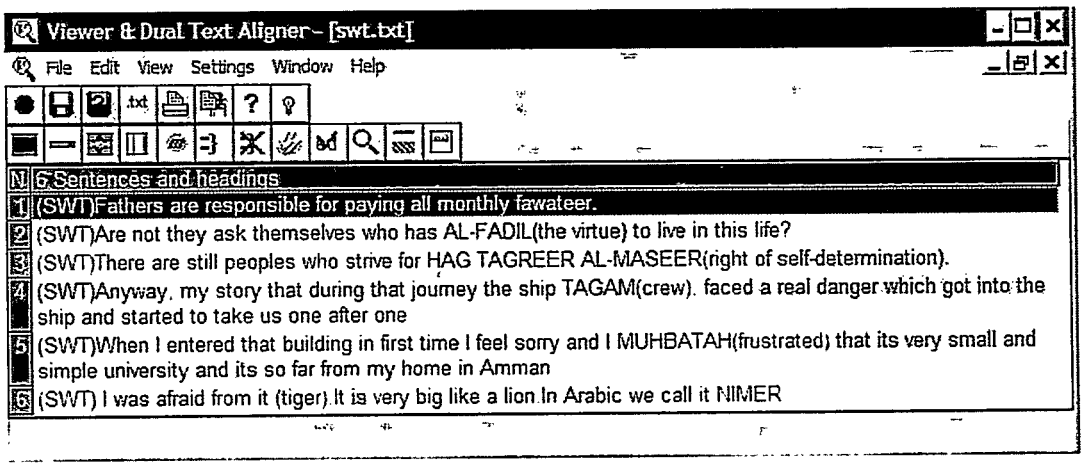


Figure 5.9. Errors attributed to simple word transfer.

In sentence (1), the learner uses the word *fawateer* to stand for the missing target word 'bills'. *Haq tagreer al-maseer* in the third example stands for 'the right of self-determination', which appears to be difficult for learners whose exposure to political texts is minimal.

Whatever the reason behind the variety switching in the cited examples (lack of lexicon, difficulty in retrieving the lexicon, etc.), it is apparent from the very low number of such errors compared to other categories that this is the learners' least favored strategy. Support for this argument is drawn from the learners' apparent preference for avoidance over resorting to their L1 as illustrated below.

It seems that the learner, who might have found himself unable to paraphrase the target lexeme or expression, considered two options: either to avoid it or to resort to his/her L1. It may be possible that learners, when they resort to their mother tongue, know that their use of L1 lexemes would be understandable simply because their receivers share with them the same linguistic background. It follows that, if the learners had known that their texts are addressed to non-native speakers of Arabic, they would have not resorted to such a strategy. Instances of simple word transfer are attested in several studies conducted using the learners' spoken discourses (e.g., Al-Khaniji 1996). It is relevant to note in this context that word transfer or code switching in the written and spoken discourses is viewed from different angles. While it is a means of solidarity, facilitation in the spoken discourse (e.g., Rolin-Ianziti 2002), it is a sign of lexical gap in written discourse.

5.3.2.3 Rhetoric

While superficially appearing to be English, close inspection of the learners' interlanguage shows that learners, to some extent, use English orthographically, but still think and organize their thoughts in their L1. It would be rare to find a piece of writing in the corpus that avoids this phenomenon. Strictly speaking, employing L1 rhetoric in the production of the L2 has shown serious negative consequences. Kaplan (1966) indicates that:

> Foreign students who have mastered syntactic structures have still demonstrated inability to compose adequate themes, term papers, theses, and dissertations. Instructors have

> written, on foreign-student papers, such comments as:"The material is all here, but it seems somehow out of focus," or "lacks cohesion." And these comments are essentially accurate. The foreign-student paper is out of focus because the foreign student is employing a rhetoric and a sequence of thought which violate the expectations of the native reader.

Such findings provide clear-cut evidence that L1 influence goes beyond simple devices to involve other larger aspects such as rhetoric. Support for this argument comes from the documented use of various features characterizing learners' L1 rhetoric (repetition, parallelism, subjectivity, excessive use of emphatics, meager use of hedges, lack of paragraphing and vague expressions, etc.) throughout the learner corpus. Having already discussed most of these features in chapter 4, much of this section will be devoted to commenting on both repetition and parallelism. However, for statistical purposes, instances of overproduction and verbosity and underproduction discussed in chapter 4 are included under this category 'rhetoric'.

(1) Repetition

The overuse of repetition at various levels (e.g., morphological, phrasal, syntactic and semantic), as shown below, furnishes powerful evidence of the continuous presence of L1 rhetoric during different phases of L2 acquisition. It should be noted that the heavy reliance of learners on repetition is ascribed to their L1, where repetition has multiple rhetorical functions. Abdullah (2001) argues that repetition serves several functions in Arabic such as emphasis, exaggeration, or the creation of parallel structures. In his article, *The Discourse of Arabic Advertising: Preliminary Investigations*, Gully (1996-1997: 22) states that the main effect of parallelism "would seem to be a reinforcement of the qualities of a product in almost mnemonic fashion through repetition of linguistic patterns." Repetition in the learner corpus surfaces primarily in three forms, namely, lexical couplets, simple repetition and content repetition. For consistency, this study distinguishes between overproduced lexical items on the full corpus basis, resulting from the

word list comparison between the two corpora and the simple repetition on an individual (essay) level based on the contextual use. For instance, the excessive use of the first person pronoun *I* (1,433 times) in the learner corpus is an example of overproduction whereas the repeated use of the word *future* in one essay, as illustrated below, is an example of simple repetition.

(i) Lexical Couplets

An additional frequently overlooked feature that characterizes learners' writing is the lexical couplet. *Lexical couplet*, in its simplest sense, refers to two near-synonymous words or expressions connected by either the coordinating conjunction *and* or the disjunctive *but*. Although lexical couplets are not as productive a structure in contemporary English writing as they once were (Johnstone 1991:37), then, their frequent use in the learner corpus is attributed to the influence of L1 rhetoric, where this phenomenon is widespread. Johnstone (1991:37) defines a lexical couplet as a structure of the form A/B which meets the all following three criteria:

1. X is a coordinating conjunction, usually additive (and; Arabic *wa*) but occasionally disjunctive (or; *'aw*).
2. A and B are synonyms, if they are single words, paraphrases if they are phrases...
3. The structure A×B has a single referent; it is used to refer to a single object, action, or state, rather than two temporally or logically discrete objects, actions or states.

Figure (5.10) presents three examples of lexical couplets in the learner corpus.

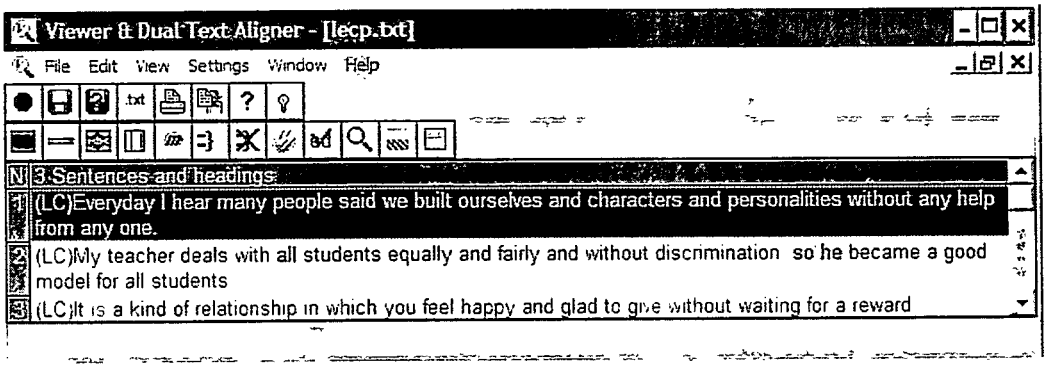


Figure 5.10. Examples of lexical couplets in the learner corpus.

A cursory reading of the three examples set out in Figure (5.10) shows that each sentence contains a pair of near-synonymous words conjoined by the coordinating conjunction *and*. One sees in these instances the possibility of omitting either of the near-synonymous pairs with no or a minimal effect on the overall meaning of the sentence. In sentence (2), for example, it is possible to omit *equally* or *fairly*, or even both of them, with no violation of the intended message since the meaning is preserved by paraphrasing *without discrimination*. Sentence (3) also shows another example of lexical couplets (*happy and glad*), where, semantically speaking, the deletion of either one will have no impact on the overall message.

(ii) Simple Repetition

Simple repetition (repeating the same lexical item) is another feature characterizing the learner corpus. In addition to its function as a cohesive device, simple repetition is used chiefly for persuasion. A close look at lexical reiteration in the learner corpus shows that this type is used more frequently in emotional and serious contexts to persuade the receiver regarding the sender's message. In the learners' writing, there are numerous

occasions where simple repetition serves several functions simultaneously, as illustrated in Figure (5.11).

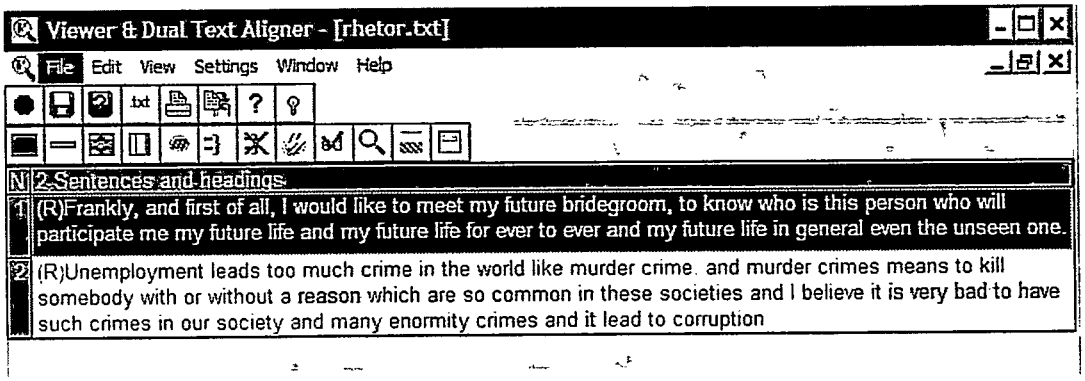


Figure 5.11. Examples of simple repetition in learner corpus.

Sentence (1) might be seen as explicitly a functional repetition. The learner, who is discussing a very important aspect of her future life, is trying to convince the receiver (via repetition) that a *husband*, for her, means *future* and thus, her selection of an ideal husband is a key to success in her life. In this sense, the deliberate repetition of the word *future* is aimed at emphasizing the relationship between the two variables (*husband* and *future*).

A similar analysis applies to sentence (2), though the excessive repetition of the word *crime* is better considered emotionally rather than literally. Taken together, these examples (1) and (2) offer empirical evidence that emphasis is often imparted, at least from the learners' perspectives, via repetition.

(iii) Content Repetition

Content repetition, which has been also attested in previous literature as a rhetorical device in Arabic (Johnstone 1991), is too obvious to be missed in most essays providing

the database for the learner corpus. In numerous cases, the entire message of a paragraph or an essay is nothing more than the repetition of a message conveyed by the topic sentence as shown in Figure (5.12).

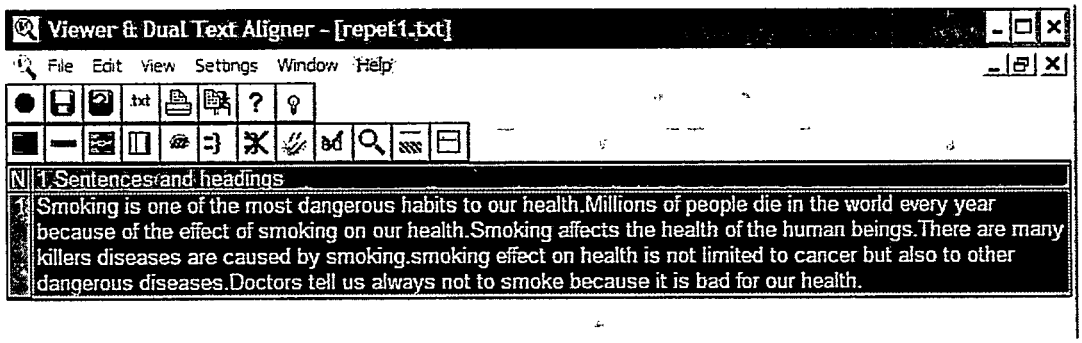


Figure 5.12. An example of content repetition carried out by lexical repetition.

It is obvious that there is a single message in this extract, viz. *smoking is dangerous for one's health.* Thus, it seems quite likely that the learner has kept massaging and repeating the topic sentence instead of developing or supporting it. Content repetition, as exemplified in Figure (5.12), is largely carried out by lexical repetition (smoke+people+health+affect+death/die). However, in many situations, content repetition is largely accomplished by variation of words and expressions.

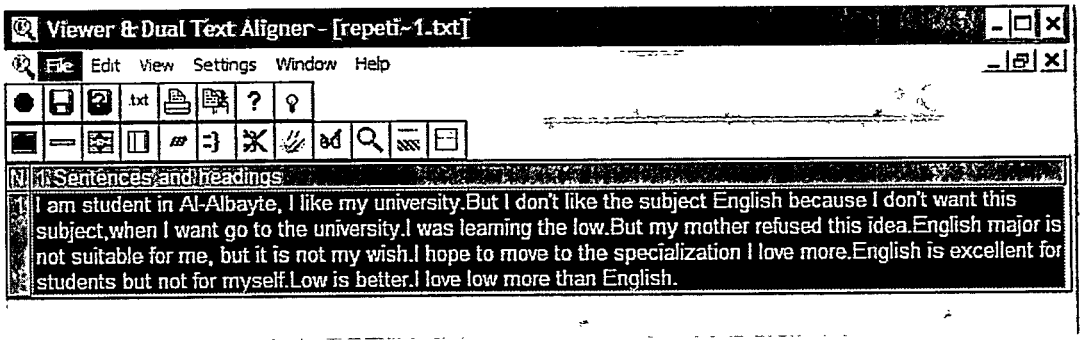


Figure 5.13. An example of content repetition carried out by word variation.

As can be readily seen in the above figure, there is one message repeated by word variation: *I don't like my major (English), but my mother wants it.* So, except for the message expressed by the topic sentence, there is no new information, just a repetition of what has been already stated.

(2) Parallelism

Another feature characterizing the learner corpus, which is also attributable to L1 influence, is parallelism, a similarity in the syntactic structure of a set of words in successive phrases, clauses or sentences. This feature, which is created by the repetitive patterns, has been noted in previous literature about Arabic (Kaplan 1966, 1972, Al-Jubouri (1984), among others). Kaplan (1966, 1972) argues that parallelism in Arabic writing was a result of the influence of classical Arabic and the language of the Holy Koran. This, according to him, explains the preference of coordination over subordination in Arabic. In this sense, the excessive use of parallelism and coordination is considerably attributable to L1 influence.

The widespread use of parallelism in the learners' L1 could explain the presence of the strings of grammatically parallel syntactic structures in the learner corpus. Empirical evidence presented in the corpus supports Henry's (1993) argument concerning the general tendency of using parallelism in sentence predicates (Henry 1993). Figure (5.14) exhibits four instances of syntactic parallelism. Three of these involve the same pattern (*avoid + gerund + and + gerund*) conjoined by the coordinating conjunction *and.*

avoid+smoking+and+eating

avoid+eating+and+consuming

avoid+drinking+noun+and+noun

avoid+sleeping+and+working+adverb

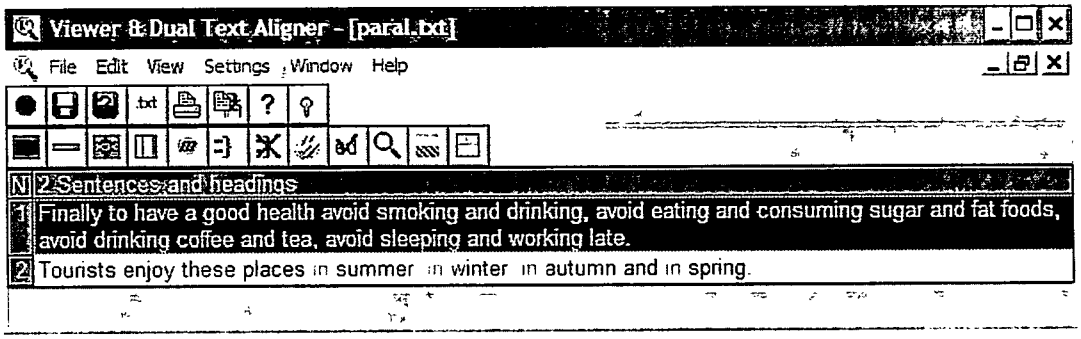


Figure 5.14. Examples of parallelism in the learner corpus.

Notice that simple repetition (*avoid* and *and*) and parallelism go together and bolster each other. In the course of quantification, only stigmatized repeated lexemes whose use has no function in the L2 have been tagged as unfavorable.

5.3.2.4 Overdifferentiation

It is axiomatic that a learner's L1 keeps revealing itself at various stages of L2 development. While this influence is frequent in the first stages, through time, it decreases to a minimal level. It is important to note that no part of L1 reveals itself in the L2 more clearly than the lexicon. In the context of corpus analysis, there are three examples, at least, where some lexical items and expressions are used erroneously not as a result of a lexical gap, but rather as a result of overdifferentiation. The term *overdifferentiation* is used here to refer to the extension of lexical distinction of the L1 into the L2, in which no such distinction is used. Though it comprises only 0.05% of the total percentage of errors, the occurrence of this category at the lexical level may indicate its occurrence at other levels, too. This type of error occurs where the L1 is lexically richer than the L2 in the field where the errors occur. This explains the presence of such errors in kinship

terms, where Arabic is unquestionably richer than English. Figure (5.15) presents two errors attributed to overdifferentiation.

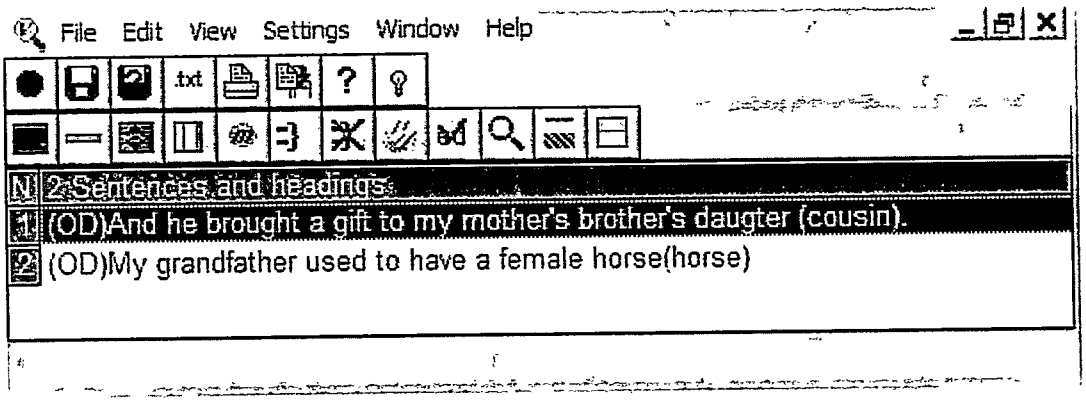


Figure 5.15. Errors attributed to overdifferentiation.

Unless the context is taken into consideration, errors of this category as shown in sentence (1) occasionally appear to be a type of paraphrasing. It is crucial to go through the entire text in question to examine whether such errors are used here due to the ignorance of the target lexicon or to overdifferentiation. In sentence (1), for example, the learner's use of *my mother's brother's daughter* is not likely to be a type of paraphrasing when we find out that the same learner used *cousin* (to refer to his uncle's son) and *uncle* in the previous paragraph. What explains the use of this lengthy expression is the learner's attempt to extend the kinship term distinction in his/her L1 to the target language where no such distinction is made. Again, the second sentence exemplifies another error of overdifferentiation. By using the word *horse*, the learner no doubt has full knowledge of the target lexeme, but s/he resorts to what seems to be paraphrasing here just to convey a lexical distinction used in his/her L1.

5.3.2.5 Paraphrasing

Maintaining their intention via the words and expressions they use is a difficult and consistent challenge for learners. For this and other relevant reasons, we might prematurely assume that the excessive reliance of learners on paraphrasing in this striking manner is attributable to the learners' attempts to express their ideas and thoughts in the target language with a minimal violation or obliteration of the intended message. As the figures of Table (5.2) show, this subcategory comprises 12.79% of the total number of learners' lexical errors. More often than not, circumlocution results in empolying several lexical items to stand for one word or expression. Approximation, on the other hand, results in a vague or imprecise word or expression whose meaning depends on the context of usage.

As the data in Figure (5.16) show, the primary feature of paraphrasing and circumlocution is wordiness. Though counter examples are not unlikely, all instances of paraphrasing shown result in using more words than necessary.

The question that arises is whether it is possible to rely on the context to define the intended meaning of the paraphrased lexeme. Empirical evidence given in Figure (5.16) reveals that context is not always likely to be helpful. In sentence (1), it is obvious that the intended meaning (abortion) cannot be discerned at sentence level. By paraphrasing *abortion*, the target item, into *losing a young baby*, it is obvious that the receiver is likely to understand a different message from the one intended by the learner. Generalizing from sentence (1) and applying the generalization to the subsequent examples can lead us astray. A look at (5) restore our confidence that paraphrasing does not necessarily result in destroying the intended message either totally or partially. In this sentence, we see that the learner has paraphrased the word *abroad* into *outside the country.* So, the meaning is preserved.

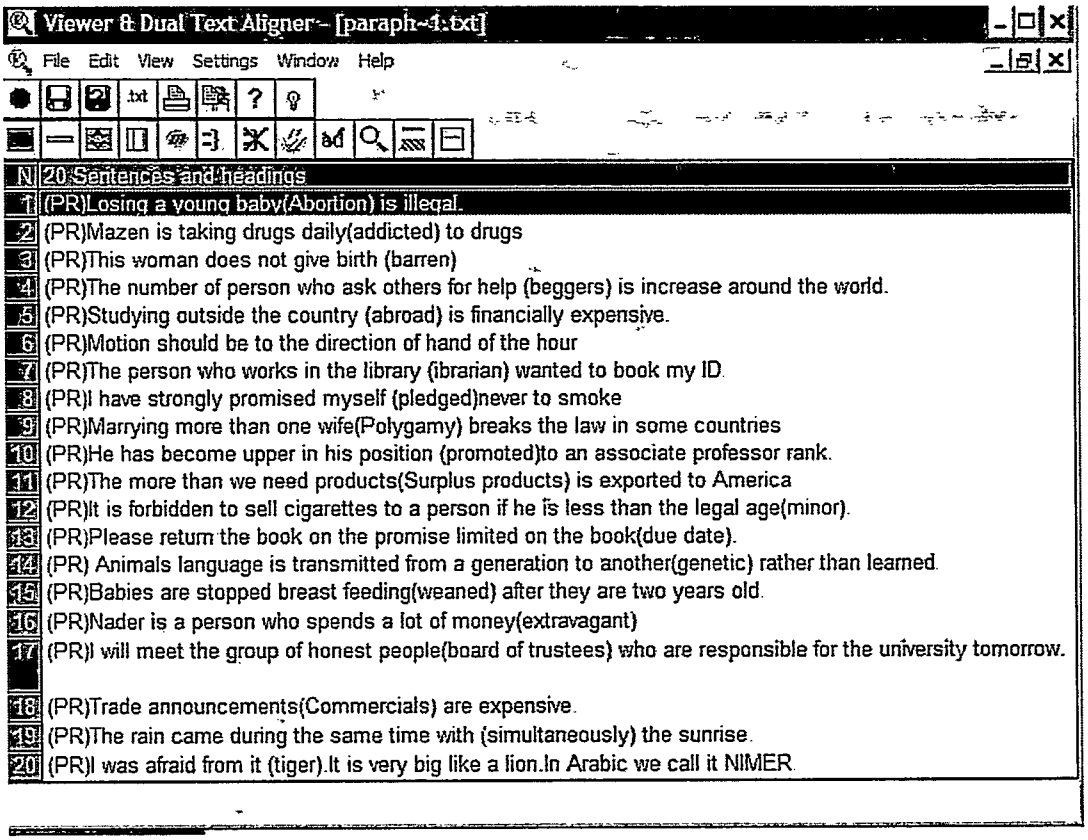


Figure 5.16. Examples of circumlocution and approximation in the learner corpus.

It is immediately obvious from the context of sentence (20) that the learner intends a *predatory animal* to be one that belongs to the cat family. However, in order to avoid ambiguity generated by approximation (like lion), the learner resorted to language switch (*nimir* 'tiger') to make the meaning clear-cut.

5.3.2.6 Avoidance

In his article, *An Error in Error Analysis*, Schacter (1974:210) argues that some difficulties that learners are likely to encounter in the L2 may not surface in the number of errors they make, but rather in the number of times they avoid using the problematic

structures in the target language. This indicates that avoiding the use problematic structures or vocabulary in the L2 is a popular strategy employed by learners whenever they feel that their use of the target item is likely to lead to an error.

Avoidance often surfaces in three varieties- -lexical avoidance, message avoidance and topic avoidance. There is plenty of evidence to indicate that lexical avoidance, which constitutes 9.77% of the total lexical errors in the corpus, is, to a great extent, associated with translation rather than with free writing. Support for this argument lies in the learners' frequent employment of topic and message avoidance in speech and writing while this option is unavailable for them when it comes to translation. Lexical avoidance, as the data reveal, could be classified as a lack of or difficulty in retrieving the lexical item in question. Albeit not as frequent as lexical avoidance, message avoidance manifests itself whenever the learner leaves the intended message unspoken or uncompleted due to lexical difficulties (more than one lexeme). Topic avoidance, on the other hand, surfaces in two ways: (i) avoiding topics that pose lexical difficulties and (ii) writing on a topic that is clearly different from the one assigned to them.

Should we blame time constraints for the high percentage of avoidance in the translation task? Such a claim may be ungrounded if we point out that less than 1% of the subjects asked for extra time to get the job done. This makes it clear that lack of knowledge or difficulty in retrieving the target lexicon is primarily responsible.

5.3.2.7 Intention Match

Meaning, as the literature shows, has been a controversial issue that has attracted the attention of a great number of linguists and text analysts during the past few decades. Various views, however, have been proposed to account for this notion. The first major view, on the one hand, believes in what Grice (1957) calls *natural meaning*, the literal meaning of words. Jordan (1992, cited in Btoosh 1999:4) argues, "...we cannot possibly

know what was in the writer's mind...we must instead analyze what she [any writer] did write." What this view suggests is the semantic interpretation of texts. The second view, on the other hand, stresses the importance of non-natural meaning, a meaning based on the sender's intention. For Hofmann (1997:273), there are usually gaps between the literal meaning of words and the sender's intention: "Whenever language is used, there is a speaker and his intent, and more often than not, the ultimate intent is hidden behind the literal meaning (i.e. "between the lines") of what is said." Meaning, according to this view, is context-dependent.

In view of these remarks, issues related to the mismatch between what a sender writes/says and what the receiver understands are of eminent importance in language learning. Research on SLA shows that the mismatch between what the learner writes/says and the receiver understands is rather more complicated than among the NSs of the language (Politzer 1978, Gass and Selinker 2001, etc.). Since analysts are supposed to be neutral in order to avoid debate over intentionality and the possibility of retrace the sender's intention, our analysis is based solely on the context of usage. Giving the paramount significance to the text rather than to the user's intention strictly complies with principles of literary criticism (Taylor 1986). In other words, errors classified under match intention category are those clearly recognized from the context.

Oftentimes, errors of intention match involve the whole phrase or sentence and thus, their classification under global errors is justified. As shown in Figure (5.17), the message of the given sentence is totally different from the intended message which is understood from the context. Since it is impossible to explain what is going on in the mind of the learner when s/he produces a given piece of language, for this particular end, we rely heavily on the context to obtain findings.

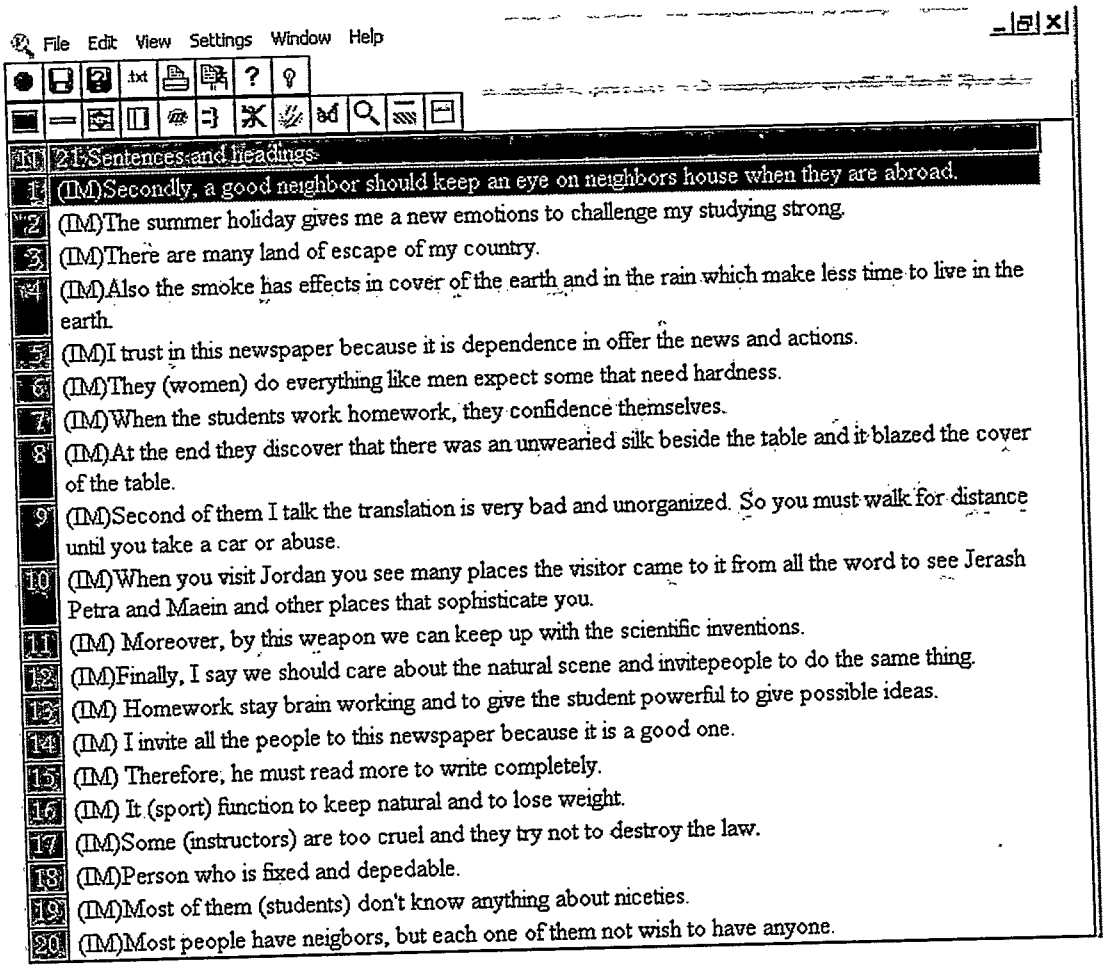


Figure 5.17. Examples of errors attributed to intention match.

This category, which comprises 4.54% of learners' errors, allows us to look closely at another kind of error, where the learner's intention is entirely or at least partially different from what s/he is writing in the target language. The substantial difference between the sender's intention and the receiver's comprehension of the above examples requires going back and forth throughout the given text to understand the gap between the two. In several cases, however, the message goes uncomprehended because the context is not clear enough to convey the learners' intention. In order to assess the semantic gap

between what learners' sentences mean and what they really mean (as understood from the global context), three examples are given.

(1) Learner's sentence: They can do everything like men expect some that need hardness.
(1) Intended: They can do everything that men do except some difficult jobs that requires great physical strength.

(2) Learner's sentence: Therefore, he must read to write completely.
(2) Intended: Therefore, he must read to write well.

(3) Learner's sentence: Most people have neighbors, but each one of them not love to have anyone.
(3) Intended: Most people have neighbors, but everybody wishes to have a good neighbor.

A close look at the learners' sentences, together with the given intended messages, shows that there is an obvious mismatch between what the learner intends and what the receiver gets. This shows that errors of this category require more effort on the part of receivers to comprehend the message correctly.

5.3.2.8 Idioms and Idiomaticity

One area in which there is an overwhelming consensus among language educators as to its importance and its difficulty in the language learning environment is the category of idioms (strictly governed expressions) and idiomaticity (acceptable usage). Research on this subfield shows that idioms and idiomaticity are not synonyms. Fernando (1996) argues that both idioms and idiomaticity share a predictable co-occurrence with specific

words, but such co-occurrence is somewhat stronger in idioms than in idiomaticity. As argued earlier, the relationship between lexemes might be predictable, as in the case of collocations, governed, as in the case of idioms, or free, neither collocations nor idioms. Errors of idioms often emerge in learners' performance due to their ignorance of a strict or governed relationship among the elements of a given idiom. In most cases, errors of idioms and idiomaticity result in a fundamental misunderstanding. Figuring out learners' problems with these subheads is rather difficult for two reasons. First, a deep knowledge of the idioms in the source language and in the target language is required. Secondly, the word-for-word or literal translation of idioms is a bit more difficult, especially when it comes to idioms whose meaning cannot be retained if their components are separated from each other.

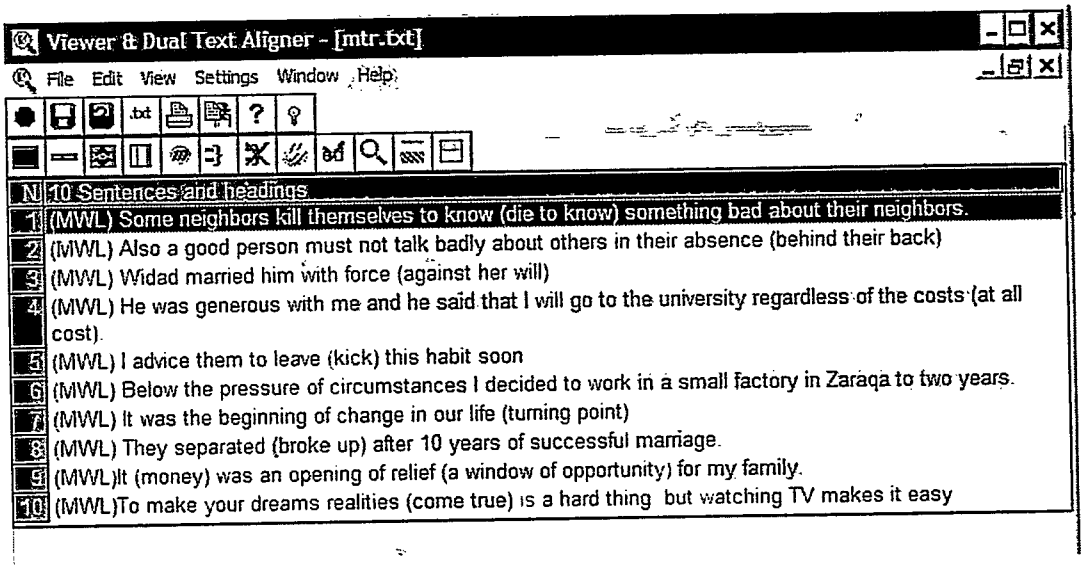


Figure 5.18. Examples of learners' use of idioms and idiomatic expressions.

In coming to understand the learners' deviant expressions, we need the English norm, as shown in the following examples.

(1) Learner's sentence: Some neighbors *kill themselves to know* something bad about their neighbors.

(1) English norm: Some neighbors *are dying to know* something bad about their neighbors.

(2) Learner's sentence: Also a good person should not talk badly about others *in their absence.*

(2) English norm: Also a good person should should not talk about other people *behind their back.*

(3) Learner's sentence: It (money) was *an opening of relief* for the family.

(3) English norm: It opened *a window of opportunity* for the family.

(4) Learner's sentence: *below the pressure of circumstances* I decided to work in a small factory in Zarqa to two...

(4) English norm: *under the circumstances,* I decided to work in a small factory in Zarqa to two...

(5) Learner's sentence: I advise them *to leave this habit.*

(5) English norm: I advise them *to kick the habit.*

For the typical native speakers of Arabic, these sentences sound good since they literally state what is used in their native tongue. Although errors attributed to this category happen infrequently, their remedies require tremendous effort on the part of the learners, instructors and syllabus designers. In stark contrast to synonymy, lexemes of this cate-

gory, particularly idioms and lexical phrases, are not freely combinable. Consequently, any change in lexemes results in altering the semantic component of the given sentence.

5.4 Results Related to Research Question (6)

Research Question (6): What are the categories of the learners' collocational errors? And what is the contribution of each category to the total number of errors?

There have been several lines of research exploring learners' collocations during the past decades. The aims of such studies have centered on measuring learners' knowledge and ability to use the target language collocations, which is considered an integral part of proficiency in the target language. However, the findings of such studies are neither consistent in terms of the learners' strategies nor in the percentage of collocational errors in general (Farghal and Obeidat 1995, Zughoul and Abdel-Fattah 2003, Al-Zahrani 1999). Such marked divergence raises the question of whether the methodological procedures employed in such studies have any impact on the performance of learners. To this end, this study has simultaneously employed several techniques to examine learners' collocational knowledge as shown in Table (5.3).

Table 5.3. Percentage of collocational errors per each method

Task	No. of target collocations	No. of correct collocations	% of the correct collocations
Translation	300	83	27.67
Multiple Choice	300	115	38.33
Semi Cloze	75	16	21.33
Cloze	75	13	17.33
Free writing	100	42	42
Total	850	269	31.65

By taking into consideration that the same test is used in the translation, multiple choice, cloze and semi-cloze tasks and that they were conducted under the same conditions, it becomes apparent that the methodological procedures and measurements of the learners' performance are strongly interactive. The highest percentage of collocational errors, as shown in Table (5.3), was found in the cloze task and the least in the free writing task. A possible explanation for the better results in the multiple-choice task than in the translation, cloze and semi-cloze tasks may involve the notion of accessibility. That is, the learners' difficulty in retrieving the target lexeme in the translation exercise is solved by having the lexeme presented in the multiple choice test. It should be mentioned that the interactivity between accessibility or retrieval problems and learners' performance have been attested in a considerable body of recent research (e.g., Lennon 1996). It is interesting, though it might be misleading, that the highest percentage of correct usage of collocations was scored in the learners' free writing. It is crucial to note that the relatively good performance of the learners' usage of collocations in free writing is attributed to either positive transfer, where the source and the target languages share the same collocates or to the learners' excessive use of general or vague adjectives (e.g., important, strong) that closely collocate with a wide variety of nouns as illustrated below. For this reason, the results of the free writing collocations cannot be generalized to the rest of the corpus. Extensive research on a wide variety of node-collocate pairs is needed to get a clear idea about this aspect.

Contrary to expectations, students' performance in translation was better than their performance in the cloze and semi-cloze tasks. Perhaps this is attributed to the learners' comprehension or lack of comprehension of the target sentence. After getting an idea about the interactivity between learners' results and the methods of investigation, it is now the time to return to the research question (6) concerning the categories of collocational errors.

It is important to be realistic in our expectations of the similarities and differences between the compensation strategies employed in section (5.2) and the strategies employed here. In striving for simplicity, an attempt is made here to use the same taxonomy used for lexical errors, except for the categories not represented in the collocational corpus. Such similarities sometimes go beyond the types of categories to include the percentage of the contribution of such categories to the total number of errors.

Table 5.4. Taxonomy of Collocational Errors

Category			Frequency	Percentage
intra-lexical	lexical meaning	near-synonymy	182	31.3
		high lexicon	4	0.7
	lexical form	synforms & homophones	11	1.9
		close forms	7	1.2
	creativity		15	2.6
interlexical	negative transfer	word-match and transfer	103	17.7
paraphrasing	circum. and approx.		141	24.3
avoidance			118	20.3
Total			581	100

Table (5.4) presents descriptive statistics for the frequencies and percentages of the sources of collocational errors. It should be made clear that Table (5.4) shows the compensation strategies of learners used in this corpus and does not imply the impossibility of employing other strategies if larger representative data were investigated. In other words, a larger corpus might employ an even wider range of possible strategies. Again, the four most often used types that have had negative weight in lexical errors are still active here, namely, near-synonym, word-match and literal translation, paraphrasing and avoidance.

Due to the similarity between lexical and collocational taxonomies shown earlier, and in order to avoid redundancy, the discussion here is primarily centered on the ex-

amples rather than on the categories of errors. Various examples cited in the following sections come from the learner corpus. In the following subsections, collocations, as argued earlier, are examined from the lexical composition approach perspective.

5.4.1 Intralexical Errors

The influence of the lexemes of the target language on each other also manifest itself in collocations. A brief look at Table (5.4) shows that this category is still the highest in terms of the total percentage errors.

5.4.1.1 Near-Synonymy

As we have by now come to expect, near-synonymous errors, as Table (5.4) demonstrates, claim the highest percentage 31.3% of errors. However, since the meaning of collocations is not a straightforward composition of the meanings of its components, collocational near-synonyms errors are expected to be more problematic than what was seen in free combinations. Inkpen and Hirst (2002) argue that "in order to convey the desired nuance of meaning and to avoid unwanted implications, knowledge about the differences among near-synonyms is necessary." What makes this subcategory a bit clearer in terms of the total percentage of collocational errors is the possibility of its occurrence in either the node or the collocate or even both of them simultaneously. In numerous cases (e.g., sentence 10 below), near-synonymy threatens the intended meaning. Consequently, partial or total destruction of the meaning in some of the examples shown in Figure (5.19) should come as no surprise.

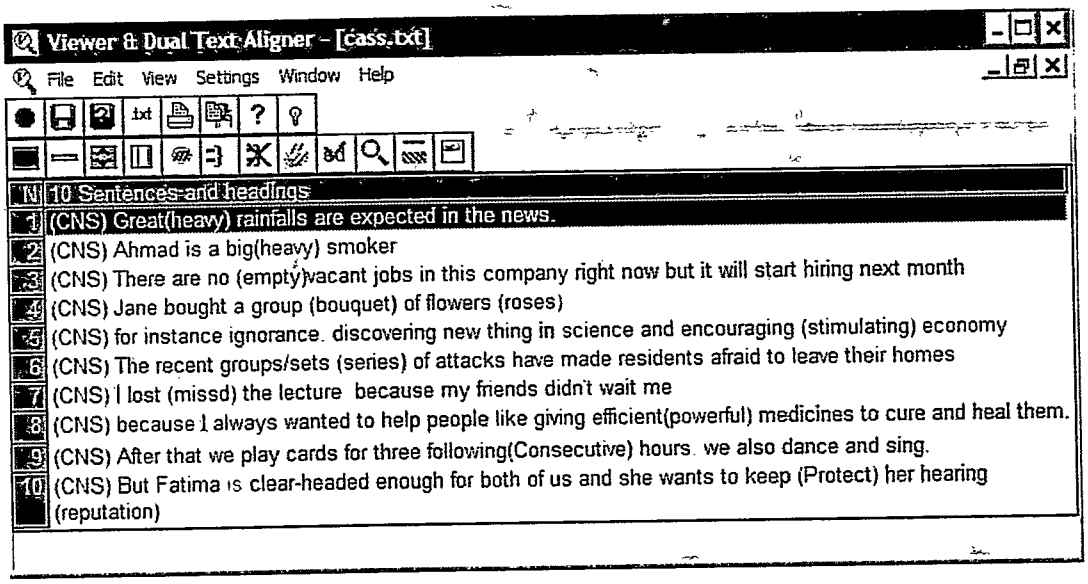


Figure 5.19. Collocational errors attributed to near-synonymity.

What explains the use of *great* instead of *heavy* in sentence (1), *empty* instead of *vacant* in (3) and *lost* instead of *missed* in (7) is the learner's belief that the each pair of these lexemes is fully synonymous.

Examples (4) and (10), in particular, show that near-synonymy might simultaneously occur in both the basis and the collocate. While the meaning of (4) is still preserved, the meaning of sentence (10), as noted above, is different from the originally intended meaning.

5.4.1.2 Lexical Form

As long as there are writing exercises for students, errors of lexical forms are likely to occur. Although fatigue, rashness and slips of the pen are possible reasons, the systematic occurrence of these kinds of errors offers independence to this subcategory and proves that it is not incidental. Rather, learners systematically commit lexical form errors whenever

they deal with confusing/unfamiliar lexemes in the target language. Figure (5.20) sheds further light on these errors, which are tagged as CLF.

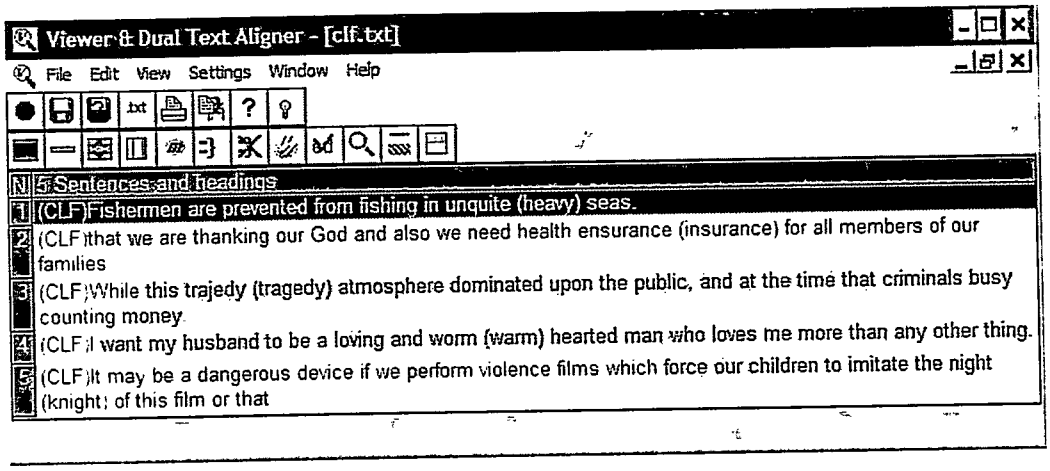


Figure 5.20. Collocational errors attributed to lexical form.

Due to the orthographic similarities between *quiet* and *quite*, the learner, in sentence (1), has used the second form to stand for the first though neither of them is the target collocate. Sentence (2) exemplifies another synform error (*ensurance* vs. *insurance*). It is worth reiterating that the lack of the full correspondence between the spoken and written forms in the target language causes a considerable body of lexical form errors (e.g., *worm* vs. *warm*). Related to this problem is the distinction among vowel phonemes; the orthographic, together with the close articulation of *i* and *e*, is also problematic for learners. Lexical form errors, as intimated previously, occur in the early developmental stages of NSs, too.

5.4.1.3 Creativity

Since the learners' intelligence is more likely to emerge at any stage of language development, the occurrence of errors of creativity is likely to continue until very advanced

levels of language proficiency. Once again, what stimulates the reliance of learners on this subcategory is their previous knowledge of their L1 and their knowledge of some similar rules in the target language. Consequently, learners overgeneralize these rules to new situations where their applicability in a new context is inappropriate as shown below.

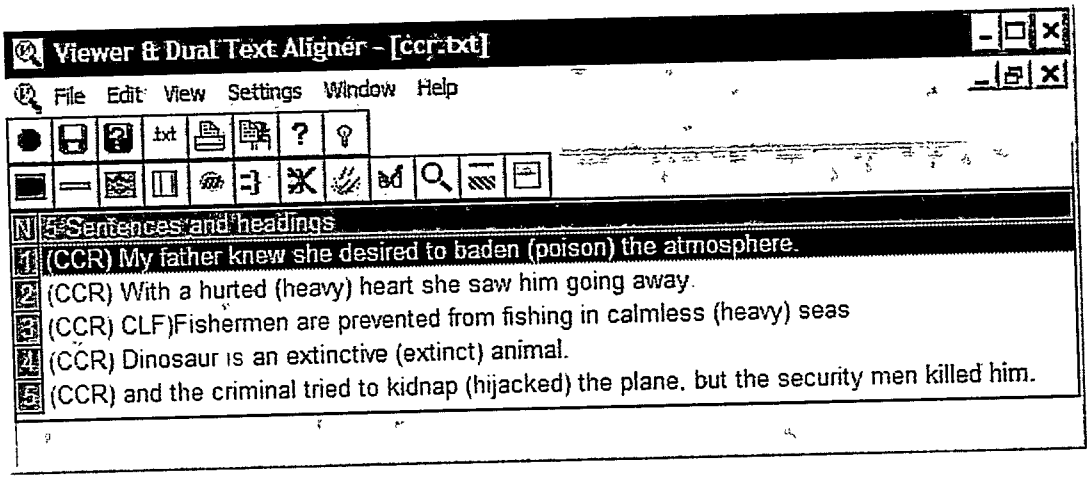


Figure 5.21. Collocational errors attributed to creativity.

The evidence provided in Figure (5.21) suggests that the learners' interlanguage exhibits various examples of creativity errors. Two notable creative errors are presented in sentence (1). First, the derivation of a nonexistent verb (*badden*) (by analogy *sad-sadden*) and the overextension of the derived form to a new situation, where *poison* is the target collocate. In sentence (5), the learner has also erroneously extended the use of *kidnapped* instead of *hijacked* to collocate with *plane.*

5.4.2 Interlexical Errors

There is no doubt that errors attributed to the L1 are hard to avoid even in fairly advanced stages of development. As shown in Table (5.4), collocational interlexical errors surface in one subcategory, viz. word-match and literal translation.

5.4.2.1 Word-Match and Literal Translation

Among the widely recurrent errors found in the learners' corpus are those attributed to word-match and literal translation. Not surprisingly, this subcategory is dominant in learners' performance whose exposure to native speakers, along with the authentic texts, is insufficient.

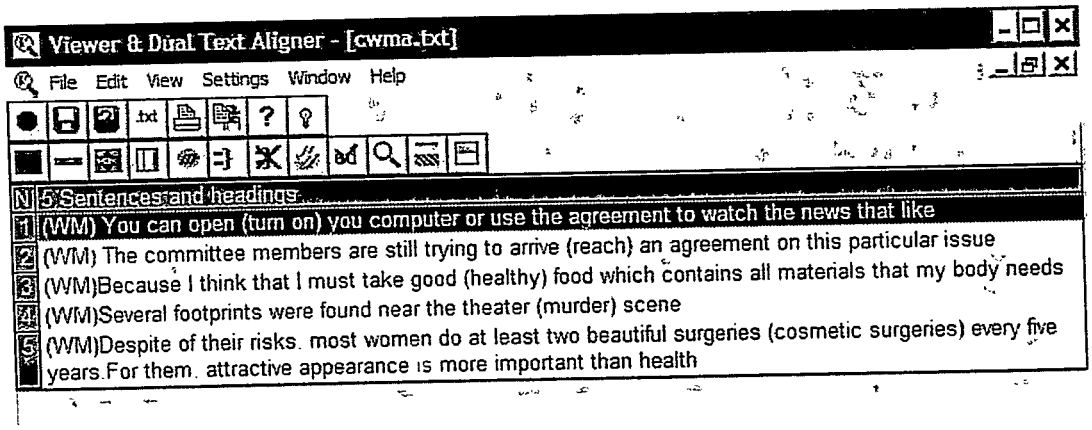


Figure 5.22. Collocational errors attributed to word-match and literal translation .

The first sentence exemplifies a common collocational error in spoken as well as written discourses of Arab students of English. Clearly, the learner is directly translating from his/her L1, where the verb *fataha*, together with all its conjugations is used indiscriminately in all the contexts, where English uses *open*, *inaugurate*, *turn on*, *begin*, etc. In sentence (3), the learner seems to be translating from his/her spoken variety; in spoken Jordanian Arabic, we often say *kweiyis* to stand for *good* and other related words such as *acceptable*, *healthy*, etc. It is thus apparent that the learner is overextending the adjective *good* to a situation where both standard Arabic and English use *healthy*.

The last example (5), presents evidence of literal translating. The equivalent expression of *cosmetic surgeries* in Arabic is *9amaliyyaat tajmiliyeh*, where the noun

9amaliyyaat means *surgeries* and the adjective *tajmiliyeh* is derived from *jamal* 'beauty'. By resorting to literal translation, the learner erroneously produced the deviant collocation *beautiful surgeries.*

5.4.2.2 Paraphrasing

Whatever compensatory techniques they might resort to, learners still find in paraphrasing the safest resort in dealing with lexica and collocations. This explains their excessive reliance on this subcategory, in particular.

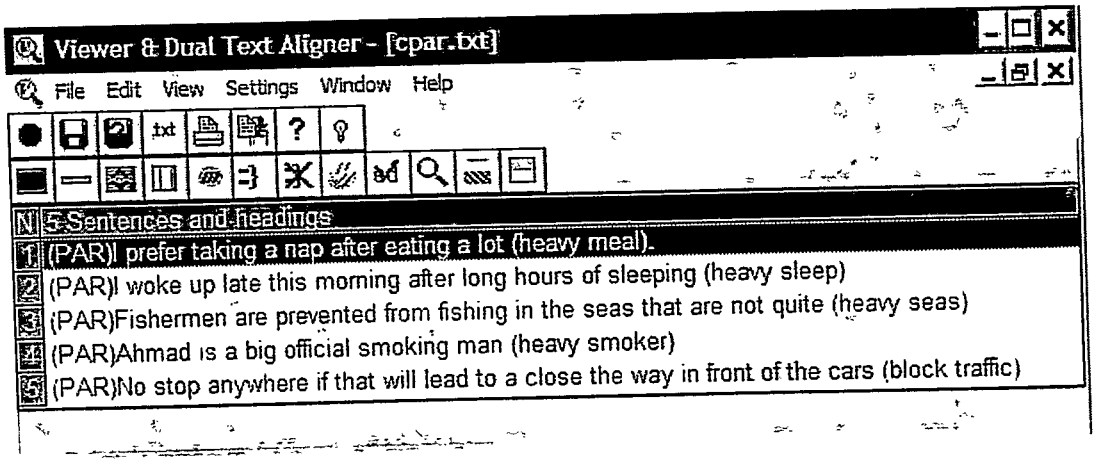


Figure 5.23. Collocational errors attributed to paraphrasing.

The shared feature of the examples cited in Figure (5.23) is the learners' heavy reliance on paraphrasing. More often than not, paraphrasing results in a lengthy answer. In sentence (1), *eating a lot of food* is clearly a paraphrase of *heavy meal.* Paraphrasing exemplified by sentence (5), *close the way in front of the cars* instead of *traffic jams* provides more evidence of lengthy alternates. Sometimes paraphrasing comes at the expense of the meaning of the sentence as shown in sentence (5).

5.4.2.3 Avoidance

As it is most general, avoidance, if no other reason intervenes (e.g., time constraint), is a direct confession of the learner's failure to come up with the target item. Due to the lack of options, as was argued previously, avoidance is almost always associated with the translation task more than any other technique (e.g., free writing task). An overview of the findings listed in Table (5.6) shows that this subcategory comprises 20.31% of the of errors.

5.4.2.4 Free Writing Collocations

In the context of our discussion of learners' collocations and in conformity with the aims posited earlier, this section examines the use of 100 node-collocate pairs in the learner corpus. In so doing, a number of adjective, noun and verb collocates have been thoroughly investigated. Empirical findings presented below show that the use of collocations in learners' free writing is characterized by four features: (i) the heavy reliance on L1 collocations, (ii) the excessive use of general and vague collocates (e.g., *good, strong*), (iii) the use of unique or creative collocations (collocations neither found in the L1 nor in the L2, e.g., *clean education*) and (iv) the substitution of a collocate for another one due to the near-synonymity between the two lexical items.

(1) Noun collocates

The collocational profile of the nouns that collocate with the adjective *strong* as shown in the concordance below (Figure 5.24) provides a detailed picture of L1 influence on the target language.

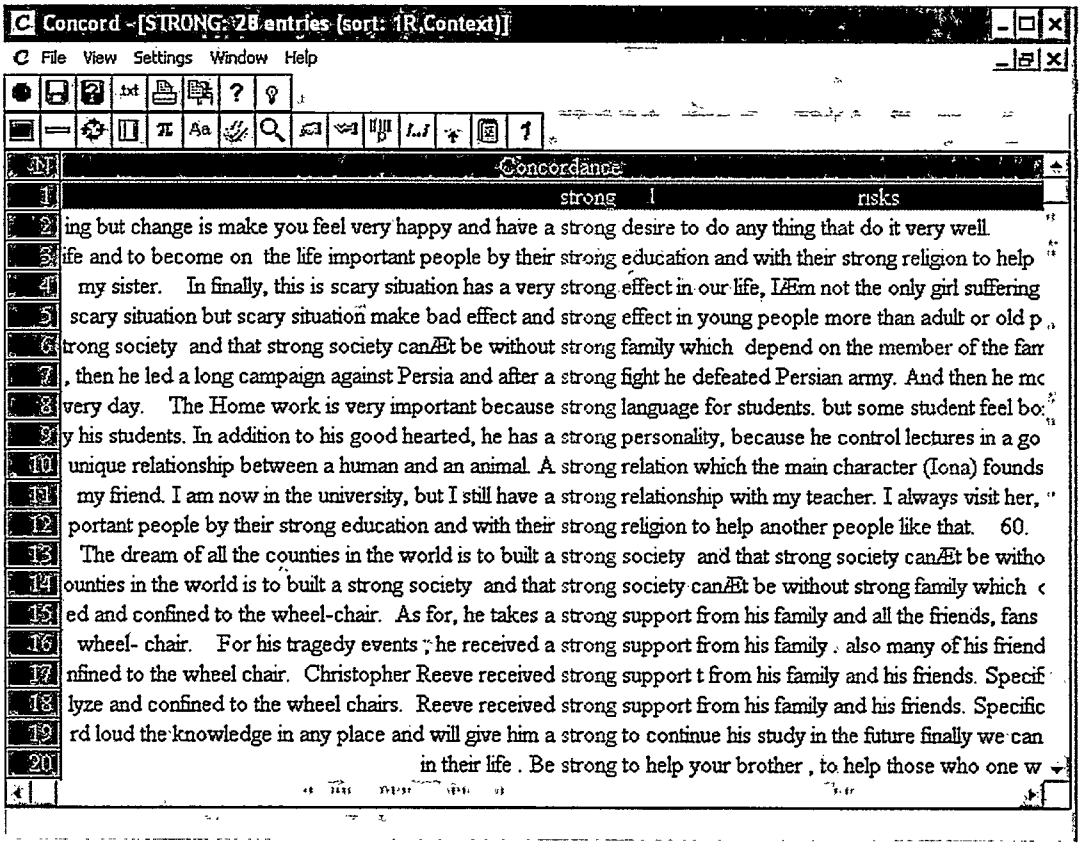


Figure 5.24. Nouns that collocate with the adjective *strong* in the learner corpus.

A careful investigation of the collocates of *strong* in Figure (5.24) reveals numerous instances of *positive transfer*, transfer resulting in correct performance due to similarity between the L1 and the L2. The Arabic adjective *qawi* 'strong' co-occurs with most of the given nouns in the above concordancer. In this sense, it is feasible to ascribe the correct uses of the collocates of *strong* to *positive transfer* while ascribing the incorrect ones to *negative transfer*, transfer resulting in errors due to differences between the L1 and the L2. For a better understanding of the cognitive process of transfer, Tables (5.5) and (5.6) reclassify the noun collocates into two categories reflecting the two types of transfer.

Table 5.5. Collocates of *strong* attributed to negative transfer

Arabic Collocates	English Collocates	Noun
strong	excellent/good	education
	full/complete	support
	close	relation
	big	fight
	Significant/substantial	change

Table 5.6. Collocates of *strong* attributed to positive transfer

Arabic Collocates	English Collocates	Noun
strong	strong	desire
	strong	effect
	strong	personality

Before we proceed to another concordance, it should be noted that the ability to categorize all nouns that collocate with the adjective *strong* into two categories reflecting the language transfer supports the previously cited argument that learners use English orthographically but still think and organize their thoughts in their L1.

(2) Verb collocates

The concordance below (5.25) shows immediately that the learners' use of verbs that collocate with the noun *university* are quite similar to the nouns that collocate with the adjective *strong* in terms of the L1 influence.

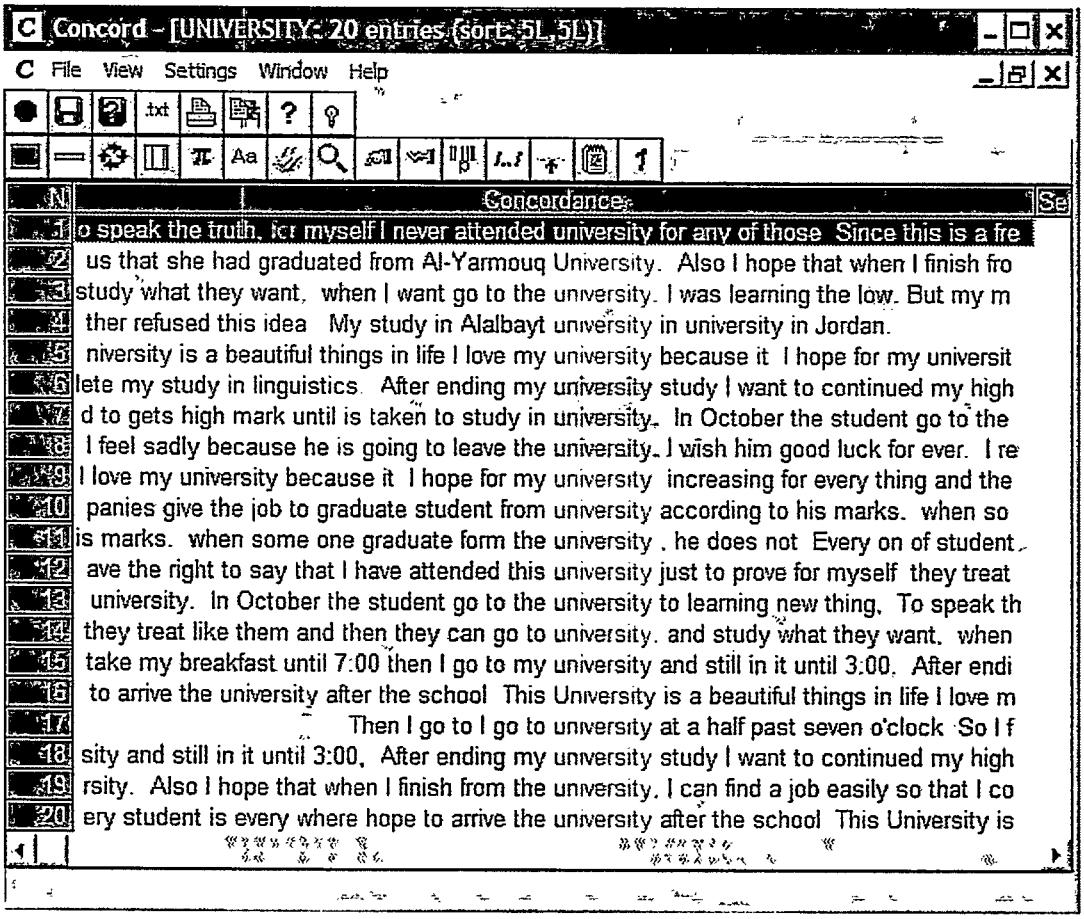


Figure 5.25. Verbs that Collocate with the noun *university* in the learner corpus.

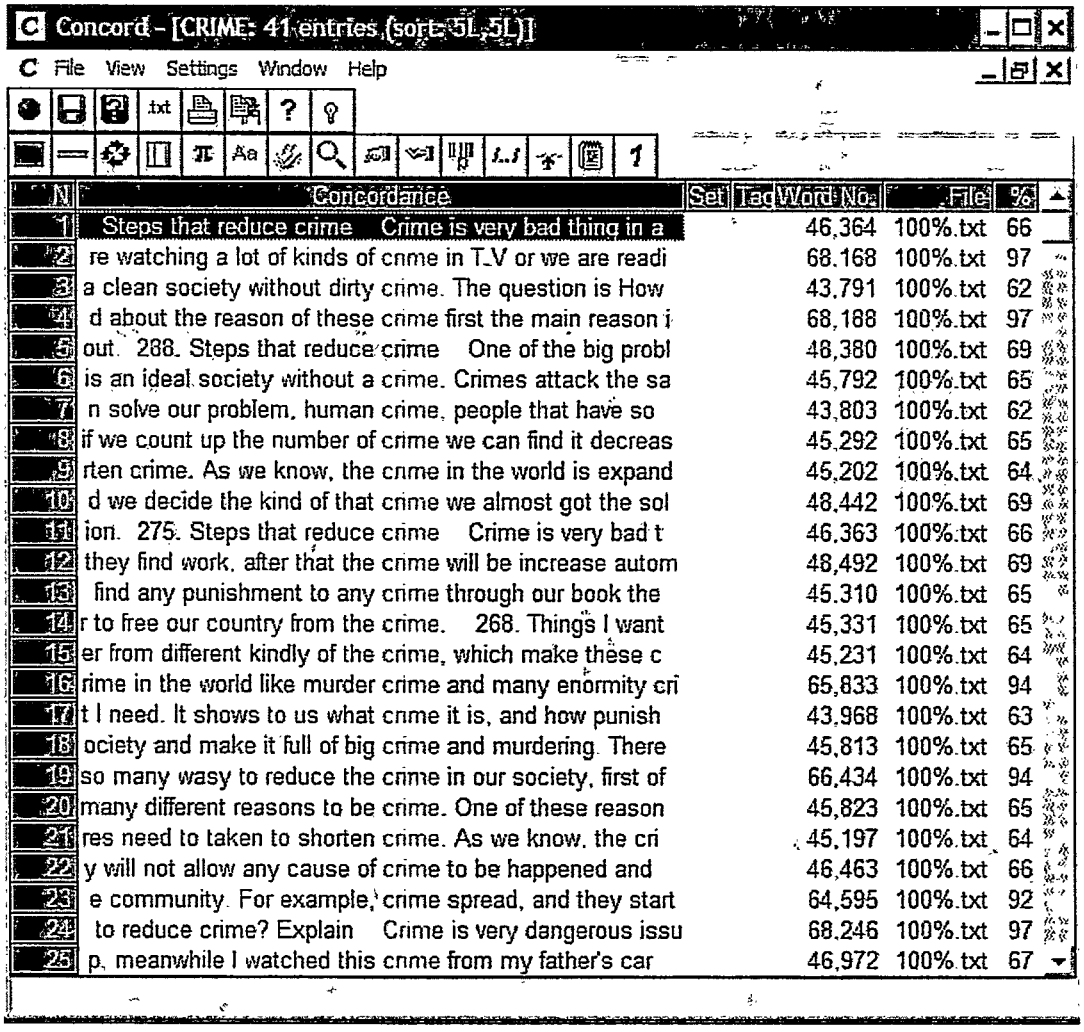
Table 5.7. Verbs that collocate with *university* as a result of negative transfer

Arabic Collocates	English Collocates	Noun
end	graduate/finish	university
leave	graduate/finish	university
reach	enter	university

Table 5.8. Verb that collocate with *university* as a result of positive transfer

Arabic Collocates	English Collocates	Noun
finish	finish	university
graduate	graduate	university
go to	go to	university
attend	attend	university

Having shown samples of adjective-noun and verb-noun collocates, it is now time to proceed to view a sample of noun-adjective collocates. To this end, the following concordance (Figure 5.26) shows some of the adjectives that collocate with the noun *crime.*



N	Concordance	Set	Tag	Word No.	File	%
1	Steps that reduce crime Crime is very bad thing in a			46,364	100%.txt	66
2	re watching a lot of kinds of crime in T.V or we are readi			68,168	100%.txt	97
3	a clean society without dirty crime. The question is How			43,791	100%.txt	62
4	d about the reason of these crime first the main reason i			68,188	100%.txt	97
5	out. 288. Steps that reduce crime One of the big probl			48,380	100%.txt	69
6	is an ideal society without a crime. Crimes attack the sa			45,792	100%.txt	65
7	n solve our problem, human crime, people that have so			43,803	100%.txt	62
8	if we count up the number of crime we can find it decreas			45,292	100%.txt	65
9	rten crime. As we know, the crime in the world is expand			45,202	100%.txt	64
10	d we decide the kind of that crime we almost got the sol			48,442	100%.txt	69
11	ion. 275. Steps that reduce crime Crime is very bad t			46,363	100%.txt	66
12	they find work, after that the crime will be increase autom			48,492	100%.txt	69
13	find any punishment to any crime through our book the			45,310	100%.txt	65
14	r to free our country from the crime. 268. Things I want			45,331	100%.txt	65
15	er from different kindly of the crime, which make these c			45,231	100%.txt	64
16	rime in the world like murder crime and many enormity cri			65,833	100%.txt	94
17	t I need. It shows to us what crime it is, and how punish			43,968	100%.txt	63
18	ociety and make it full of big crime and murdering. There			45,813	100%.txt	65
19	so many wasy to reduce the crime in our society, first of			66,434	100%.txt	94
20	many different reasons to be crime. One of these reason			45,823	100%.txt	65
21	res need to taken to shorten crime. As we know, the cri			45,197	100%.txt	64
22	y will not allow any cause of crime to be happened and			46,463	100%.txt	66
23	e community. For example, crime spread, and they start			64,595	100%.txt	92
24	to reduce crime? Explain Crime is very dangerous issu			68,246	100%.txt	97
25	p, meanwhile I watched this crime from my father's car			46,972	100%.txt	67

Figure 5.26. Adjectives that collocate with the noun (*crime*)in the learner corpus.

A close look at the concordance lines shows once again the strong influence of the L1 collocations on the L2. In line (3), for instance, the learner uses the Arabic collocation *jarima gaDirah* 'dirty crime' to stand for the target 'awful crime'. Sentence (19) also presents another example of language transfer. The learner uses his/her native collocate *kabirah* 'big' with the noun *jarima* 'crime' to stand for the English collocate 'great crime'. Another sample of negative transfer that is not as straightforward as the previous ones

is shown in line (7). The learner uses *human crime* 'jarima insaneeyah' to stand for *crime against humanity.* Sentence (21) presents evidence of the intralexical errors, where *shorten* is used instead of *reduce* due to the near-synonymity between them.

In sum, we have seen that the correct or incorrect uses of target collocations is, to a certain extent, governed by the similarities or differences between the two languages. This explains the high percentage of correct collocations in the first concordance (the adjective *strong*) and the terrible results in the third concordance (the noun *crime*).

5.5 Results Related to Research Question (7)

Research Question (7): Is learners' collocational knowledge on a par with their lexical knowledge?

In much of the previous research conducted on interlanguage collocations, there was a strong presumption that the learners' failure in the use of collocations is solely attributable to the learners' ignorance of the collacability between the target base and the collocate in question. While this argument is not totally imprecise, it is even more appropriate to tie the collocational errors to the lack of the target lexicon. In coming to understand and examine this relation, the following statistical comparison between lexical and collocational items was conducted based on the results of the lexical translation corpus and guided collocational corpus.

Table 5.9. Percentage of collocational errors relative to lexical errors

No.	Type	Expected	Correct	Incorrect	% of incorrect
1.	Lexical words	4,500	2,024	2,476	55.02
2.	collocations	750	227	523	69.73
3.	Perc. of collocations	16.67%	11.22%	21.12%	

A comparison of the figures reveals three crucial insights. First, the percentage of learners' lexical errors 55.02% is worse than what is generally expected, and it indicates that learners face serious problems not only with collocations but also with the lexicon. Secondly, learners' lexical knowledge is likely to support the assumption that learners' collocational errors are, to some extent, attributable to the learners' lack of either the collocate or both of the collocate and the base rather than to their ignorance of the collocability between them. Thirdly, while the overall ratio of the target collocations relative to the lexical items is 16.67%, the ratio of the incorrect collocations to the incorrect lexical items is 21.12%. This means that learners face a much wider gap in collocations than in lexicon as illustrated in Figure (5.27).

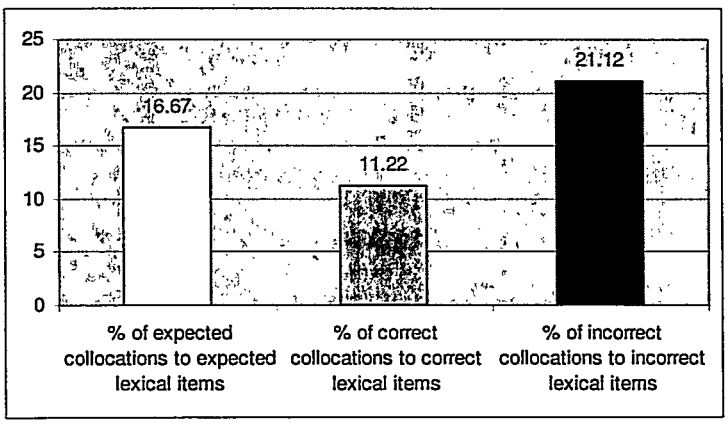


Figure 5.27. Percentage of learners' collocational errors to lexical errors.

A look at the figures demonstrates that learners' knowledge and use of collocations is worse than that of the lexical items. Yet, learners' poor performance in collocations relative to lexical items should not hide learners' problem in lexicon, too. These findings support a suggestion already made by Laufer (1996) concerning the necessity of having special courses devoted to vocabulary teaching.

In brief, lexical knowledge is a consistent challenge for learners and thus, it is not surprising to get such a low level of learners' lexical and collocational proficiency in L2. What might be directly inferred from such figures is that lexical knowledge, which is a neglected variable in the language curriculum of most English departments at Jordanian universities, is not in a much better situation than collocational knowledge. In view of these remarks, it is fair to conclude that much urgent attention should be devoted to this vital aspect of language learning.

CHAPTER 6

CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS

6.1 Introduction

This concluding chapter consists of three sections. Section (6.2) summarizes by reviewing the research questions and findings of the study. Section (6.3) presents the limitations of the study. Section (6.4) provides recommendations for future research.

6.2 Summary

Using empirical methods to examine lexical complexity, text-profiling, lexical and collocational errors in the writing and translation of Arab students of English, this study has addressed multiple research questions: (1) To what extent does the learner corpus deviate from the reference corpus in terms of lexical complexity? (2) To what extent does the learner corpus deviate from the reference corpus in terms of the features and percentages of the top 200 frequent tokens and hapax legomena? And how can learners' lexical stereotypes be captured through word frequency? (3) What are the most salient and stereotyped features of the learner corpus? And how far is the learner corpus influenced by learners' L1? (4) What are the most problematic words that Arab students of English encountered by the corpus? (5) What are the sources of learners' lexical errors? And what is the contribution of each category to the total number of errors? (6) What are the sources of learners' collocational errors? And what is the contribution of each category to the total number of errors? (7) Does learners' collocational knowledge go on a par with their lexical knowledge?

The use of the corpus-based approach to answer the above-mentioned research questions required the availability of three component parts: (i) a machine-readable representative corpus of the written interlanguage of Arab Students of English, (ii) a similar-sized authentic machine-readable reference corpus, and (iii) a number of software programs (e.g., *Concordancer*, *Wordlist*).

The following multiple findings, which come in the same sequence as the aforementioned research questions, reveal that some of the research results resonate with the previous literature while others show counter results. Yet, it should be mentioned that some counter results presented here are ascribed to the differences in methodology, data or the influence of the learners' cultural, linguistic and rhetorical background.

Findings of Research Question (1): The reference corpus is much more complex in terms of lexical diversity and density than the learner corpus. The divergence in lexical diversity between the two corpora reflects the learners' limited word stock. Since deficiency in lexicon results in an overall deficiency in language learning, such findings convey an urgent need for a serious revision of the curriculum.

Findings of Research Question (2): Learners rely more heavily on grammatical words than NSs do. Also, the learner corpus is characterized by excessive frequency of the top 200 frequent tokens and the use of vague and general expressions. As for the hapax legomena, learners use a lower percentage of unique tokens than the NSs.

Findings of Research Question (3): the differences between NSs' and NNSs' use of word categories are attributable to either the learning developmental stages or the influence of learners' L1. Also, the findings show that the learner corpus is characterized by excessive overproduction of some lexica (coordinating conjunction *and*, first person

pronoun *I*, etc.) and excessive underproduction of other lexica (may, perhaps, etc.). The divergence between the learner and reference corpora in terms of the overused and underused lexica, as the figures show, results from the profound influence of the linguistic and rhetorical features of learners' L1. Again, although it is unlikely for learners to match NSs' proficiency level, learners' proficiency in L2 writing is far beyond satisfaction.

Findings of Research Question (4): After eliminating the redundant and unique errors, it was possible to sort out 761 lexical errors, which, in turn, fall into sixteen subcategories. It is obvious from these subcategories that learners use divergent strategies to compensate for their lexical gaps in L2.

Findings of Research Question (5): Errors of near-synonymy, paraphrasing and word match, and literal translation come first in terms of frequency.

Findings of Research Question (6): The marked divergence in the percentage of collocational errors in learners' performance indicates a strong link between the percentage of errors and the methods being employed. As far as the collocational errors categories are concerned, near-synonyms, paraphrasing and avoidance are the most commonly occurring types.

Findings of Research Question (7): Learners' limited word stock, together with their ignorance of the collocability between items, are the main reasons behind the high percentage of collocational errors vis-a-vis lexical errors.

6.3 Limitations of the Study

Despite the accessibility of approaching a wide range of topics (e.g., discourse markers, cohesion), this study has been strictly limited to investigating a few lexical aspects of the writing of Arab students of English as a foreign language. This means that no other aspects (e.g., pragmatics, discourse markers, syntax) has been targeted in this study. Furthermore, this study has been devoted solely to the learners' written interlanguage. So, no attempts have been made to get the spoken discourse involved in any part of this study.

Subjects' residency is another limitation to the study; no writing samples or tests have been employed in this corpus if the participant ever lived in an English-speaking country. By testing volunteer participants in classes that would meet simultaneously or when one course is a prerequisite to another, no subject, to a maximum extent, could sit twice for the same test.

In compliance with the Office of Research protocol (03.309) concerning the use of human subjects in academic research, no part of the data used in this study was collected before or after the agreed upon dates.

6.4 Future Research

Since this is, as far as can be discovered, the first study of its kind conducted on the interlanguage lexicology of Arab students of English as a foreign language via a corpus-based approach, then, it is reasonably expected that the research on this field is still immature and there are still vast areas that have not been yet taken into consideration. Additionally, the data in this study, which are not representative in terms of genres, points to a strong need for building further corpora. Furthermore, the findings of this study, which is strictly limited in its scope, are not predicting absolutes for other corpora

that might incorporate new compiling criteria. In other words, much research is needed to uncover different scopes of learners' lexicology.

In view of the previous remarks, further research is definitely needed to: (i) investigate the interactivity between learners' lexicology and the level of education, sex or specialization, (ii) examine learners' lexical complexity in the spoken discourse, (iii) investigate lexical and grammatical collocations in learners' free writing, (iv) create a dictionary of the problematic words that Arab students of English are likely to encounter at different phases of their second language mastery, (v) build a syllabus that meets learners' lexical need, and (vi) examine the interactivity between input modification and proficiency in L2.

As for curriculum and syllabus designing, it is sufficiently evident from the preceding chapters that learners have serious problems in literacy and this, in turn, calls on curriculum and syllabus designers to review their objectives to keep up with the recent developments in the theories of learning and teaching. However, the term *literacy* is not used here in the same traditional sense, the ability to read and write. Rather, it means the amount, type and scope of activities that academic institutions provide learners with. Cooper (online) argues that schools need to broaden their concept of theme and the materials that constitute themes:

> Typically, themes of study have focused on literature in the traditional sense, including narrative and expository texts, with a heavy emphasis on stories. However, a "real world" literacy perspective calls for themes that are much broader in scope and content (Walmsley & Walp 1990). These themes need to be built around a combination of high-quality literature in the traditional sense and high-quality "real world" resources, including such things as posters, letters, magazines, maps, brochures, charts, journals, computer resources, and so forth. In essence, broadening our concept of literacy leads us to broaden our concept of literature to include all possible things that individuals might need to learn to read and respond to in life.

The question that might come to mind now is why we should blame the first component of literacy (reading) while examining the second component (writing). Krashen

(1993:72-72), who believes in the vast and divergent advantages of reading (e.g., improving vocabulary, spelling, and grammar) provides an answer for this question:

> The research reviewed earlier strongly implies that we learn to write by reading. To be more precise, we acquire writing style, the special language of writing, by reading. We have already seen plenty of evidence that this is so: In Chapter 1 we saw that children who participate in free reading programs write better (e.g., Elley and Mangubhai 1983; McNeil in Fader 1976), and those who report they read more write better (e.g., Kimberling et al. 1988 as reported in Krashen 1978, 1984; Applebee 1978; Alexander 1986; Salyer 1987; Janopoulos 1986; Kaplan and Palhinda 1981; Applebee et al. 1990.

While the use of literacy in L1 involves numerous activities other than reading books and writing papers (e.g., solving problems – they read signs or advertisements; for social activities – writing letters, bumper stickers, posters; for gaining news and information – reading newspapers and magazines; for remembering things – messages to self and others; and so forth.)(Brice Heath 1983, cited in Cooper), the use of literacy in L2 is largely restricted to reading books and writing papers. This, of course, leaves learners with a minimum opportunity to use literacy L2 in comparison with L1. Again, the oversimplification of L2 input and the selection of non-authentic materials, make the situation worse than ever expected. Beyond these unpleasant facts, a considerable body of learners who have access to translated materials (particularly plays, novels, novellas, etc.) prefer to read the assigned texts in their L1.

In the light of these statements, it is highly recommended that academic institutions: (i) maximize the number of activities that encourage learners to develop literacy in L2, (ii) minimize oversimplification of L2 input, (iii) select authentic text materials and (iv) discourage learners from resorting or referring to translated text materials (by assigning new text materials that have not been translated into learners' L1).

APPENDIX A

LEARNERS' LEXICAL ERRORS

No.	Target lexical item	Learners' used items
1.	abandon	drop, quit, leave
2	abbreviations	shorts, contracts, cuts, reductions
3.	abductor	thief
4	abortion	failure, drop, killing babies, miscarriage, losing a young baby
5	abound	increase, grow
6	abroad	outside the country, outdoor, in a foreign country, out country, overseas
7	absorb	suck, drink, swallow, take
8	abstain	not, refuse, prevents, reject, restrain, not agree
9	account	calculation
10	accredited	independent, trusted, reliable, adopted, authorized, commissioned
11	achieve	reach
12	acquired	earned, gained, obtained
13	actions	doings
14	activity	movement
15	addicted	chronic, habitual, drugger, accustomed to drugs, a person who takes drugs daily,
16	address	title, setting, residence, place of dwelling
17	adhered to	keep, committed to, very close, strict, conservative
18	adjourn	raise, delay, finish, postpone, move, close, defer, lift, put up
19	adjust (their beliefs)	fix
20	admission	agree(ment), accept(ance), reception, entrance, approval, agree (of enlisting), enlisting, indoor,
21	adopt	adapt
22	advanced	high, old
23	advise	advice
24	affect	effect
25	affect	power, influence
26	against	opposite
27	allow	offer
28	alter	altar
29	altogether	all together
30	amend	reform, change, correct, fix, equal, modify, adjust, alternate, improve, repair
31	among	between
32	amount	number
33	ancient	archaic, old, big
34	angle	corner

35	annoy	noise,
36	appendix	additional part, attached section
37	appetizer	help to eat
38	application	request, order, demand
39	appoint	name, employ, assign, put eye
40	appreciate	estimate
41	appreciation	taste, enjoyment
42	argument	conflict
43	arid	dry, without plants,
44	around the clock	all the time, all around the day, around the hour, all the time,
45	arouse	arose
46	arrest	jail, capture, internee, prison
47	assassinate	abdicate, murder, kill,
48	associate	participant, sharing, subscriber
49	astronavigation	space (shiping), sailing,
50	asylum	political refuge
51	attend	come, present, join
52	attract	draw, pull, take
53	average	rate
54	bachelor	without a wife, unmarried, not married, alone
55	bad	dirty
56	bakery	oven
57	balance	arrange, coordinate, stabilize
58	bald	without hair on his head, bold
59	banned	forbidden, prohibited, not allowed, obscenity, prevented
60	barren	does not give birth, can't have children, sterility, childless
61	bathe	wash
62	beef	cow meat, cow flesh
63	beer	bear
64	before	prior to
65	beggars	poor, not rich, poor people who keep asking others for assistance
66	beginning	origin, source, start
67	behaviors	doings, manners, conducts
68	behind	back, after
69	believe	think
70	beneficiary	advantager, useful person, user, benefiter
71	besiege	surround
72	betray	break oath, lie, perjury, cheat

73	better	bitter
74	between	among
75	beverages	drinks, liquids
76	bicycle	air bike
77	big	great, old, large
78	big (heart)	warm (heart)
79	bills	pills, counts, invoices, vouchers, fawateers, fees
80	blank verse	rhythmical, well-built poetry, balanced poetry
81	blanks	empty places, spaces, dashes,
82	block	obstruct, close, impede, stop
83	board	council, group, members
84	body/dead body	coffin
85	book shop	library
86	border	edge
87	bribery	commission, giving a person some money that he shouldn't take to serve him, illegal money
88	bridal (money)	marriage (money)
89	bring	take
90	broke	out of money, penniless, bankrupt, insolvent, does have no money, poor,
91	brother-in-law	my wife's brother,
92	brutal	hot, burning, warm, harsh
93	budget	balance, money plan, scale
94	burn	spend
95	busy	working, occupied, engaged, have no free time,
96	buy	purchase
97	cabinet	council of ministers, head ministry,
98	caliber	mental ability, cleverness, intellectuality
99	calm (person)	can always hide his anger
100	calm (see)	quite, smooth, not noisy, relax,
101	cancel	omit, delete, erase, clean
102	capacity	ability
103	capital	head money, beginning money
104	carry out	achieve, work, do,
105	catering	serving foods and drinks
106	cattle	gaggle, flock, troops, group, herds, kettle,
107	cease	end, pause
108	cell phone	small carry telephone/mobile
109	censor	watch, observe
110	century	millennium
111	challenge	resist
112	chamber	room

113	characteristics	adjectives, qualities, features, properties, traits,
114	charity	money offered to help the poor, offer money to the poor, alms, donate, zakat, giving, free money, present, help
115	cheat	to do something without permission or knowledge, deceive, betray
116	chicken	chicken meat, fowl
117	childhood	child's period, period from 1-6 years
118	children	childs
119	chivalric	heroic, great, brave, horsical , knight
120	Christmas	birthday
121	class	college/university
122	classic motif	historic type
123	climate	weather, atmosphere
124	clinical death	coma, bed death, sleep death, die in bed, mind death
125	clip	cut
126	clock	watch, o'clock
127	clock wise	in the direction of the hand of the clock, towards the hour pointer, from left to right, in the same direction of the clock
128	close	near
129	closely tied	fixed tightly
130	coach	trainer, exerciser, team leader, teacher
131	coincide	come also with, during /at the same time,while, accompany
132	cold	chilly, icy, freezing
133	coldhearted	hard heart
134	collapse	fall
135	colleague	friend
136	collect	calculate, count
137	coma	shock, comma, unconscious, not awake, absence of mind
138	comfortable	easy
139	commercials	advertisements, trade announcements
140	committee	group
141	compassion	sympathy
142	compensate	pay back, repay
143	complete	finish, do, meet, end, continue
144	compliments	complements
145	conceited	pride, deceived, proud, jealous
146	conceive	become pregnant
147	conclusive (evidence)	cut, perfect, clear, explicit, strong

148	confer	discuss, talk, negotiate
149	conference	session, meeting, summit
150	confidence	courage, trust, belief, certainty,
151	confines	admit, say the fact, give knowledge of, speak
152	consist	contain, comprise, involve
153	constitution	law, regulation
154	consult	negotiate, discuss, take the other's opinion, talk, argue
155	continue	complete
156	continue	keep
157	continuous	rapid, unlimited, running
158	contribute	participate
159	conventions	traditions, values, customs, imitations, habits
160	conversely	on the other way
161	convinced	admit,
162	coordinate	arrange, harmonize, design, decorate
163	core	important, main
164	corporal	physical, body
165	corpse	dead man
166	correct	see
167	correspondent	reporter, transfer, messenger, sender
168	cosmetic surgery	beautiful process, face lift, cosmetic operation, reface, beauty surgery
169	count	consider
170	counterpart	similar, the same, mate, friend, opposite, the other side, reflect, equal, contrary, head
171	create	exist
172	credibility	truth(fullness), honesty, trustfulness
173	credit	commissioner, adopted, recognized, dependable,
174	criminals	guilties, thieves,
175	critical	dangerous
176	criticism	punishement
177	crook	bad, rude, tricky, playful, deceiver, scandalous, cheater, shrewd
178	cruel	tough, aggressive, hard, rough, rock heart, merciless, hard-hearted, rigid, without emotions
179	cultivable	suitable for plants,
180	cure	treat
181	current events	happening things
182	curriculum vita (cv)	autobiography, self bibliography, self story
183	cut	slice, chip
184	dangerous	enemy

185	dark	bold
186	daylight saving time	summer time, summer clock
187	dead line	last date, final appointment, end time, maximum date, final period
188	debate	negotiate, discus, talk, argue, dialogue
189	deceive	cheat
190	decrease	shortening, declining,
191	decree	will, wish, wanting, order, intend, permission
192	deep	profound
193	deficit	shortage, incapacitate, weakness, lack, failure, lack of ability
194	degree	study
195	degree	grade
196	delete	cancel, erase, omit, clean
197	deliberate	slow, unhurried, careful, quite, leisure, rational, late
198	deliver	say, declare
199	dentists	doctors of teeth
200	deodorant	body spray, smell remover
201	deploy	spread
202	desires	customs
203	desk	office, table
204	desk	disk
205	devastating	hard, strong
206	develop	change
207	developed (countries)	preceded
208	developing (countries)	growing
209	dictation	imitation, memorization, spoon-feed, oral taking, drill teaching
210	didactic	educational, taught
211	died (in a road accident)	went
212	differentiate	divorce
213	dinner	evening meal
214	disappoint	let me down
215	disaster	destroy, catastrophe, trouble, damage
216	discount	cheap, reduction, cut of prices, low down, sale, decrease
217	discriminate	separate, distinguish
218	discuss	talk
219	disguise	mask
220	dismiss	dispel, quit, separate, release?
221	dispersed	disappeared

222	diverse	different
223	do	finish, complete, achieve, work, solve
224	do (homework	take
225	donate	give, offer, grant, gift
226	dormitory	university home, university city, university housing, university residence, university hostel, place where students live,
227	down payment	deposit, first money, initial payment, forward payment, pay in advance
228	draft	project
229	drop	increase, decrease, low, reduce, fall, got down,
230	due date	right time, certain date, fixed date, exact time, particular time, the same time, limited promise, specific date, accurate date, definite time, on time
231	dust	rust
232	duty	homework
233	earn	Take, collect
234	eat	take
235	editor	liberator, author, director
236	educate	make
237	egalitarian	balanced, equal
238	employees	workers,
239	empower	give power
240	encounter	meet, face, confront
241	encourage	drive, invite, advice
242	encourage person	encourage human
243	endeavor	struggle
244	enjoyable	beautiful
245	enroll	register, write
246	ensanguined	red handed
247	entire	all, whole, hole, complete, total,
248	environment	nature
249	envoy	sender, messenger
250	erupt	rebel, explore, revolt, raging, pump,
251	escalate	rising, increase, aggravate, rise, elevate, make high
252	especially	specially
253	evacuation	be left, movement, transfer, be got, be disserted, be emptied, be immigrated
254	evidence	proof, proof that can't be changed, sign, approve, indication

255	exaggeration	over, addition
256	exalted	high, raised, tall, noble, majestic
257	exclusively	only, private, limited, especially, on the face to monopolization
258	executed	killed, murdered, slaughtered
259	exempt	exceptional, free, pardoned
260	exercise	practice, sport, play sport
261	exile	get out of country, forced to leave his country, negative(d), neglected, banished, expelled, punished, get rid of
262	expand	get larger and larger
263	expire	end, finish
264	explore	discover, find out
265	express their message	pass their message
266	extinct	dead, not found, dead from a long time, no longer exist, unique, non-existent, finished, rodent
267	extraordinary	irregular, up normal, unusual, exceptional, excluding, alternation
268	extravagant	spends a lot of money, exaggerator, luxurious, lavish, prodigal, wasteful, not mean
269	fad	style, dresses popular nowadays
270	failure	non-success, unsuccessful, defeated, fall, fail,
271	fairness	justice, legal
272	faithful	fixed
273	farther	further
274	fatal	leads to death, killed, deadly, lethal
275	favorite	best, preferred
276	feed	eaten
277	fetus	child, baby
278	filed	study
279	find	discover
280	fine	money paid as a punishment
281	finished (one year)	aroused
282	fiscal	money, financial
283	fitting (room)	measure
284	fix	mend, correct, repair
285	fix (eggs)	prepare/cook (eggs)
286	flat	part
287	float	spread over water, over flow, swim
288	floor	flour
289	foam	cream, paste

290	follow/on (a diet)	make
300	foot	leg
301	force	obligate
302	forceful (speech)	tough, strong, loud, hard, sharp, oratory
303	forecast	weather broadcast, weather report,
304	forge	fake, unreal, counterfeit
305	forgiveness	excuse, amnesty, tolerance, pardon, mercy, excuse
306	foundation	construction
307	free	not busy, empty, unoccupied, not working, have space time/leisure time/nothing/no work, without money
308	fresh	new, newly cut, modern, recent
309	friends	company
310	frozen	iced, frosted, snowed, frozen
311	fruits	crops,
312	full	rich
313	full of	crowded
314	fun	enjoyment
315	fundamentally	jewelrly, essentially, substantially, principally
316	games	matches
317	gap	space, hole, distance, dash
318	get dressed	wear my clothes
319	get surprised	suddened
320	give (all the money)	put
321	go	arrive
322	go through	cut, cross, walk, pass, jump,
323	good	pleasant
324	grade	level
325	gradually	step by step
326	graduate	high level study, degree after bachelor degree
327	gray	as if it is covered snow
328	great	large, big, huge, beg, old, grand
329	grow	increase
330	hall	class, room
331	halt	stop, put an end to, block, interrupt, cease
332	hard	strong
333	hardness	be strong
334	harmful	so bad to health, unhealthy, dangerous for health
335	harmony	agreement
336	have?	do, make
337	healthy	good

338	heartless	hard, tough, cruel, rock heart, aggressive, acute,
339	heat	warm
340	heavy drunker	chronic, drunken man, a man who drinks a lot of wine,
341	held	caught
342	help	give, make
343	high speed	very fast, big speed, in a hurry
344	highway	motorway
345	hold	contract, set, cast, knot, tie
346	hold (id)	take, book, keep, restrain, reserve, put off, arrest
347	hold (tongue)	keep silent, knot, catch, complicate, tie, un speak
348	honestly	carefully
349	honorary	proud
350	horn	pipes, car sound, noise sound, clarinet
361	hospitability	generosity, welcoming visitors, receiving guests, hostility
362	hot	warm
363	hot-tempered	warm warmed
364	humans	men
365	icebergs	iced/icy/frozen/frosty mountains
366	ideal	perfect, real, good
367	identify	catch
368	ignore	neglect, without concern, not care, never mind, forget, carless
369	illegal	unpermitted, against the law
370	illiteracy	uneducation, inability to read and write, ignorance, lack of education, motherhood
371	imitate	adopt
372	immature	children put in bottles, not complete, pre-time, incomplete, minor, children who are in glass house
373	immersing	emeressing
374	immortal	has no end, will never die
375	immunity	power, protection, security
376	impatience	rashness
377	improve	increase
378	in favor for	toward, for, for good, for the benefit of
379	in other meaning	in other words
380	in spite of	although
381	inauguration	opening
382	increase	addition, growth, become large
383	indispensable	can't live without it, you need it,

384	infected	effected
385	influential	effective
386	infrastructure	under building, basement, underground building, lower rank, facilities
387	ingredients	contents
388	inhale	breathe, suck air, take air in
389	inherited	in born, natural
390	initial	first, elementary, beginning, primitive
391	installments	payments, stages, payment on several times, parts, stations, credit, means
392	instructor	constructor
393	insurance	security, care, sake
394	insure	ensure
395	integrate	involve
396	intensive	heavy, hard
397	interest	benefit, advantage, price, useful price
398	interior	inside
399	intermediate	middle
400	intermittent	from to time, non-continuous
401	interrupt	cut, stop, pause, make me stop, cross on my talk, let me continue, cut my words,
402	invent	find, discover, create
403	inventor	devise
405	inventor	devisor
406	irrational/impatient	hot warmed
407	irritation	pain
408	issue	problem
409	job	task, work, assignment, function, career, work
410	justify	explain
411	keep	stay
412	kind (hearted)	white, love all people, clean, good man, blank heart, pure, simple-heart
413	kitchen	room for cooking, cookery
414	know/learn	grasp
415	knowledge	weapon
416	lack	lose, miss, does not have
417	land	nature
418	landlocked	no opening on the sea
419	large	a lot of, spread, enormity
420	last	take
421	launch	lunch
422	laundry	washer/washing machine, sink, dry-clean

423.	league	university
424	leak (water)	escape, move, loose, pump, infiltrate
425	learn	adopt, deal
426	learners	educated people
427	letter	message
428	librarian	the library man, responsible of library man, library keeper, library honest, library security, secretary of the library, the person who works in the library, safe library
429	license	permission
430	like	love, prefer, seem, as, desire
431	lioness	wife of lion
432	listen	hear
433	live	stay
434	live up to	promote
435	lock (the door)	close
436	long	tall, big
437	long run	remote time/future, far future, in the far, on long time, long away, far place, highway, far extent, for a long turn, far goal
438	long time ago/many years ago	old years, past
439	loose face	loose the water of his face, loose his shame,
440	loud	huge
441	low	under
442	lucidity	clarity, flexibility, simplicity, easiness, not undifficulty, smoothness, clearance, absolve
443	luxury	excellent, enoyment
444	made a mistake	faulted
445	maintenance	fixing, repair, refreshment, keeping new, exchange, conservation, preservation, rebuilding, make up,
446	major	study
447	make	Fall, do
448	make up (test)?	retest, return the test, make another test, give the exam again, repeat the exams
449	mammals	animals whose children depend on milk, creatures that are born by eggs
450	many persons	many people
451	marital	martial
452	masses (of workers)	herds
453	math	calculator
454	mature	adult, riped, adult, strong, responsible,

455	memoirs	memories, diaries, note book, notices, auto bibliography
456	method	way
457	might	possible
458	migrate	leave
459	minors	a person under/low/less than regular age, teenager, children, underage teenagers, less than, teens
460	missed	lost
461	mistake	error, wrong, false
462	model	ideal, form
463	moisturize	to put liquid or cream on the dry skin
464	monument	moment
465	mood	way
466	multilingual	speaks different languages
467	multiply	hit
468	narrow-minded	fixed belief
469	near	under, beside
470	neat	tidy
471	needy	people who need money, poor
472	negotiations	arguments, conversations,
473	neutral	isolated
474	new	recent, modern
475	new (year)	head of the year, beginning of the year, year top
476	nice	darling, comfortable
477	no house **lacks** a t.v. set	**empty**
478	notary public	justice/faire writer, clerk
479	nourishing	feeding
480	nowadays	in these days
481	obstacles	problems, difficulties
482	obstacles	difficulties, barriers, troubles
483	offer	present
484	oil	petrol
485	old	big, large
486	operation	process
487	opponent	opposition, competitor, enemy, antagonist, rival, against
488	options	favor
489	oral	aural
490	ordinary	normal, regular, usual
491	organs	parts, elements
492	orphan	does not have parents,

493	out of order	broken, turned off, broken, does not work, off,
494	outstanding	imaginary
495	overcome	finish, pass, exceed, get over, cross, (over) step
496	overlook, ignore	forget
497	overtime	overwork, extra work, additional work
498	overwhelming	almost all of people, whole, majority, most numbers, total, more
499	pale	white
500	pantry	store
501	park	bark
502	participants	contributors, members
503	parts	departments
504	paste	cream
505	patent	invention innocence, invention purification
506	pay	spend
507	peak (times)	rush, climax, top, summit, afternoon, first, hard, great, difficult
508	peddler	hawker, walked seller, move seller, wandering seller, sales man, travel salesman
509	pedestrian (bridge)	walker
510	peers	friends, colleagues, companies, partners
511	penultimate	pre-last, before the last, before the end,
512	people	men, creatures
513	personality	trait
514	persuade	coax
515	physical	body
516	physical (education)	athletics, sport, playing
517	pick up (fruit)	elicit, harvest, collect, pluck, cut, elect, amaranth
518	pile up, accumulate	getting over each other
519	pillars	elements, principles, basics, components, roots, structures
520	pitiful	merciful, sympathy, kindness, sorrowful, exciting
521	plain	bitter, unsweetened, black, without sugar, dark coffee, sick
522	plan	draw
523	please (v)	like
524	pleasure	enjoyment
525	pleasure	joke
526	pledge	cut a promise on my self, make a vow, declare, take a promise, make a promise, promise
527	plow	dig
528	political man	politician

529	poll	questionnaire, opinion search, gather opinions
530	polluted	not clean, dirty, spoiled, unclean
531	polycot	speaks several languages, multi-tongues
532	polygamy	several/many/multi wives, marrying more than one woman, multiple marriages, variety of wives
533	portion	part, department
534	portions	places
535	poultry	chicken
536	power	energy, source
537	powerful	strong, brave
538	predict	analyze
539	prefer	like, prefer
540	prescribe	describe, give, write, formula
541	presence	attendance, existence
542	president	major, chief, boss, manager
543	prey	victim, weak creatures that are easily eaten by other animals
544	priceless	expensive
545	priest	church man
546	principles	rudiment, doctrines, notions
547	produce	perform
548	professional	high level
549	professors/teachers	doctors
550	profit	prophet
551	prominent	the most important one, famous one, distinguished, outstanding, noticeable, brilliant, well-known, prime, popular,
552	promotion	preferment, become upper in his position, raise, lift
553	proponents	supporters, defendants, agreed persons, encouragers
554	protect	keep, safe, peace, save
555	provide	supply
556	public	masses, general
557	punch	bundle, group, bouquet,
558	punishment	cruelty
559	quick-witted	intelligent, clever, apprehensive, impulse
560	quietly	deeply
561	quit	leave
562	quote	mention, citation, martyrdom, give example,
563	racial	ratial
564	raise	Build up, bring up

565	rate	average, scale
566	reach	arrive at, extend
567	realistic	actual, objective
568	realize	discover
569	rear	grow
570	reasons	considerations
571	receive	study
572	recently	modernly, lately, newly, up-to-date,
573	recite	read (with the same rhyme), sing, tell
574	reflecting	mirroring
575	refuse	reject, didn't accept,
576	regain	return, give back, come back
577	registrar	writer, recorder
578	reinforce	encourage, motivate, enhance, support, emphasis, raise, intensify, increase, make, push, strengthen, support
579	reject	refuse, disagree
580	remain to this day	stay to this day
581	remember	review
582	remind (me)	remember
583	repair	correct, mend
584	require	need
585	residence	living, stay, sitting,
586	resignation	stop working, retirement, give up working, request to end his work, permission to leave the work
587	rest	nap
588	retail	separate, single, partial selling, individual, small, alone
589	return	repair,
590	rhymed (poetry)	balanced, measured, scaled
591	ride	drive, lead
592	rinse	wash, clean,
593	risk	danger
594	robbed	stolen, plundered
595	robbers	stealers
596	rotate	change
597	rug	carpet
598	rule	control
599	rural	ruler
600	safety	protection
601	scattered (showers)?	few, small amount, little, different, separate

602	scenery	nature
603	scent	smell, odor, musk, odor
604	schedule	program
605	secretary	a person who keeps secrets, secrete saver, safer, faithful secretary
606	secular	universal
607	secure	gain, insure, maintain, save, keep, get, guarantee
608	see	sea
609	self-determination	fate deciding, final destination, end decision
610	self-sufficient	no need for help, self-enough
611	semi-final results	non-final conclusion
612	seminar	session,
613	senior (citizens)	old (people)
614	separated	divorced
615	session	cycle, circulation, meeting, round, period
616	sever (adj)	strong, high, terrible, keep, sharp, chronic
617	sever (v)	cut (of), stop
618	several	a number of
619	share	arrow
620	share	participate
621	show	teach
622	show (us something)	see
623	showers	rain, drops
624	shrewd	intelligence, quick understanding, very clever,
625	sick leave	ill(ness) vacation, illness permission, sick holiday, illness rest, ill absence
626	sides	faces
627	sit	set
628	site	sight
629	sitting	setting
630	skinny	slim, thin,
631	slavery	godless, worshipping, idol, , lack of freedom
632	smoothness	easy, without facing probles
633	smuggling	transportation, passing illegally, escaping, running. trading
634	snowball	things getting over each other
635	soap	soup
636	sociable	mixed
637	solar	sun
638	solar (year)	sunny
639	sometimes/occasionally	in special cases
640	spacious	wide, deep, big, large, broad, expansive

641	special	especial
642	specialist	expert
643	spirit	moral
644	split	separate, divide, divorce
645	spread	separate
646	stay	retain
647	step mother	father's wife, mother-in-law,
648	stick	wood
649	stimulate	develop, encourage, courage, helps, motivate, trigger, give, support, enhance, activate
650	storage	deposit, lodging,
651	story (building)	floor
652	strength	power
653	strict	cruel
654	strong	beautiful, hardness
655	structure	control
657	stubborn	hard, strong head, boneheaded, arrogant, hard-headed, difficult
658	study	make, work
659	studying	education
660	subcommittee	branch committee, secondary/subgroup
661	submit	give, deliver
662	suit	suite
663	summary	summery
664	summit	peak, climax
665	sunrise	rising sun, sun rise,
666	superficial	minor, easy, external, surface, shallow, flat, ceiling
667	supply	give
668	supreme (court)	high, top, head, super, highest
669	surplus	extra, enough, addition, not necessary
670	surrendered	gave up, decline, receive, be arrested
671	surrounded	encompassed
672	suspend	cancel, stop, hang, defer,
673	swelling	lump, enlargement, increasing, expanding, growing, tumor, disease, growing, bulge, enlargement, expanding,
674	symptoms	features, reasons, feature, indications
675	taboo	forbidden, prevented, prohibited, unacceptable, banned, barred, unlawful
676	take	spend
677	take care of	concern

678	tall	long
679	tasty	delicious
680	teach	learn, feed
681	teaching	explanation
682	team	club, group of players
683	tear	break, cut into pieces
684	tell	show
685	teller	box keeper, box protect, safer, box reserve, box honest, saving box, cashier
686	temperature	degree temperature, heat
687	tempt	attract, seduce
688	tender	giving, offer, project
689	tie	close, stop, hold
690	tight/close group	squeezed group
691	times	situation
692	tires	wheels
693	topic	subject
694	tow	got my for the truck
695	traditions	customs, habits, usual doings
696	traffic	vehicles, cars
697	traffic lights	light signals, flares, lights traffic, traffic lamps
698	transcript	mark lists/tables/sheets, degree/average lists
699	transmit	transport, translate
700	transparent	clear, visible
701	transportation	transaction
702	travel	abroad, go, visit
703	traveling	vacation
704	treasurer	keeper of the box
705	treaty	agreement, lease, contract, document, guarantee, conference
706	trim	cut, shave, clip
707	tropical	flat, not mountainous
708	trustee	honestee, safer
709	tuition	fees, taxes, money
710	turn	become
711	turn on	start
712	turn up	raise, rise, increase, high, speak up, loud, upper
713	unanimously	by all members, agree all people, totally, wholly
714	underestimate	decrease, reduce, lessen, belittle, decline, simplify, reduce, minimize
715	unpleasant	bad, not good
716	unquestionable love	unconditional love

717	unsettled	hanging
718	unstable	shaking
719	unsurpassed	unequal, unrival, unbelievable, unique, unseen incomparable
720	until recently	until now
721	unwilling	unready
722	unwired	unwearied
723	vacant	empty, lack, unoccupied, unused
724	vacation	holiday
725	vacation/weekend	holiday
726	values	advantages
727	vanish	disappear,
728	varied (books)	wide (books)
729	vast	great
730	vehemently	strongly
731	veil	cover of the face, hijab
732	vets	doctors of animals,
733	viceroy	crown prince, vice king,
734	view	look
735	vocational (school)	training, practical
736	vomit	puke, return, throw, disgorge, sick
737	vote	show agreement, elect, voice, give your voice, select, elect, sound, show,
738	wages	salaries, rents, fares, prices
739	wait	still
740	want	covet, like
741	warranty	guarantee
742	water proof	keep from the water, water preventive, water resistance, against water, blocking, antiwater
743	weaken	reduce, decrease
744	wean	stop sucking milk from their mothers, not to be allowed to suck milk, ablactate
745	weather	whether
746	wedding (party)	marriage (party)
747	well	easily, completely
748	whenever	in all case
749	whether	weather
750	while	distance
751	white	clear
752	whole	hall
753	wholesale	complete sale, group/huge/balk sell
754	will	well

755	wish	dream
756	without	empty
757	wonderful	very splendid
758	write	compose
759	wrong	incorrect
760	young	children, beginning of age
761	younger	small

APPENDIX B

LEXICAL TRANSLATION TEST

Translate the following bold words and expressions into English.

(a)

1. إلتقى الرئيس الفرنسي **بنظيرة** الألماني لبحث بعض المسائل **العالقة** بين البلدين.

1. The French president met his German **'counterpart'** to discuss some **'unsettled'** issues between the two countries.

2. لقد **قطعتُ عهداً** على نفسي أن لا أدخن أبداً.

2. I have **'pledged'** never to smoke.

3. من فضلك **لا تُقاطعني** وأنا أتحدث.

3. Please **'don't interrupt me'** while I'm talking.

4. لا أحد يعلم **عنوان** إقامة أخي في البرازيل.

4. Nobody knows my brother's **'address'** in Brazil.

5. **تورم** (إنتفاخ) العيون واحدة من **أعراض** هذا المرض.

5. The **'swelling'** of eyes is one of the **'symptoms'** of this disease.

6. **قطع** أحمد **الإشارة الضوئية** وهي حمراء.

6. Ahmad **'went through/ran'** the red (traffic) light.

7. لقد **صوت** المجلس **لصالح** القرار الجديد.

7. The Councilman **'voted in favor of'** the new resolution.

8. **الدراسة المتأنية** مفيدة على **المدى البعيد**.

8. **'Deliberate'** study is useful in the **'long run'**.

9. أراد **أمين المكتبة أن يحجز** هويتي.

9. The **'librarian'** wanted to **'hold'** my ID.

10. **رفضت** وداد الدواء الذي **وصفه** لها الطبيب.

10. Widad **'refused'** to take the medicine the doctor **'prescribed'** for her.

11. انا الان **مشغول**. لوسمحت اتصل معي لما أكون **فاضي**.

11. I am '**busy**' now. Please call me when I am '**free**'.

12. لقد **أستعاد ثقتي** به.

12. He '**restored my trust**' in him.

13. **صيانة** الطرق عمل لانهاية له.

13. Road '**maintenance**' is (an) endless work.

14. هو **ناضج**.

14. He is '**mature**'.

15. ان الاوان **لقطف** ثمار الأشجار.

15. It is time to '**pick/ gather**' fruit from trees.

(b)

16. هذة سيارة **قوية وواسعة**.

16. This is a '**powerful**' and '**roomy/spacious**' car.

17. لقد **تخطوا عقبات** كثيرة للوصول إلى هذا الهدف.

17. They have '**overcome**' many '**obstacles**' in order to reach this goal.

18. أظهر **إستطلاع للرأي** إن غالبية الطلاب **يفضلون** الورقة البحثية على الإمتحانات.

18. A '**poll**' revealed that the majority of students '**prefer**' term papers to exams.

19. اللغة الإنجليزية مليئة **بالإختصارات**.

19. English is full of '**abbreviations**'.

20. الدراسة في **الخارج** مكلفة.

20. Studying '**abroad**' is expensive.

21. السارس مرض **مُميت**.

21. SARS is a '**fatal**' disease.

22. من المتوقع أن **يعقد** مجلس النواب **دورة إ ستثنائية** لمناقشة بعض المواضيع التي لم تُناقش في **الدورة العادية**.

22. It is expected that the Parliament will '**hold**' a '**special/irregular session**' to discuss some issues that were not discussed during the '**regular session**'.

23. الاسفنج أفضل مادة **لإمتصاص السوائل**.

23. The sponge is the best '**liquid-absorbing**' material.

24. تهريب **المخدرات** مشكلية دولية وحلها يتطلب جهدا دوليا كبيرا.

24. Drug '**smuggling**' is an international problem and solving it requires a great international effort.

25. كثير من الناس يُفضلون **لحم الدجاج** على **لحم البقر**.

25. Many people prefer '**chicken**' to '**beef**'.

26. الرجل الذي يقف **خلف** السيارة أخي.

26. The man who is standing '**behind**' the car is my brother.

27. يتزايد عدد **المتسولين** (الشحادين) في العالم.

27. The number of '**beggars**' is increasing around the world.

28. إشتريت **باقة** (ضمة) من الورود.

28. I bought a '**bunch**' of flowers.

29. يُمنع تناول **المشروبات الكحولية** في الاماكن العامة.

29. Drinking alcoholic '**beverages**' is forbidden in public places.

30. لايُسمح ببيع الدخان لاي **شخص أقل من السن القانونية**.

30. It is not allowed to sell tobacco to '**minors**'.

(c)

31. أقص أضافري كل اسبوع.

31. I '**clip**' my **nails** every week.

32. علي **يقود** دراجته الهواية بسرعة كبيرة.

32. Ali is '**riding**' his bicycle very fast.

33. تعاني كثير من دول العالم الثالث من مشكلة **الامية**.

33. Many of the Third World countries suffer from the problem of '**illiteracy**'.

34. يعتبرحزب المحافظين **الخصم** الرئيسي لحزب العمال الحاكم في بريطانيا.

34. The Conservative Party is considered the main '**opponent/rival**' of the ruling Labor Party in Britain.

35. أنا بشتري الخبز من **المخبز** الي بجنب المدرسة.

35. I buy bread at the '**bakery**' near the school.

36. هذا المحل دائما بعمل **خصم** على أسعار الملابس.

36. This store always offers '**discounts**' on clothing prices.

37. أنا تماما ضد **العقاب البدني** في المدارس.

37. I am absolutely against '**corporal**' punishment in schools.

38. يُمنع الوقوف لتجنب **إعاقة حركة السير**.

38. No stopping to avoid **traffic jams**.

39. **الفجوة** بين الاغنياء والفقراء في **إزدياد مستمر**.

39. The '**gap**' between the rich and the poor is **continuously widening**.

40. حصل أحمد على منصب رئيس النادي **بالإجماع**.

40. Ahmad was elected president of the club **'unanimously'**.

41. **الصدقة** تساعد الفقراء.

41. **'Charity'** helps the poor.

42. تحدثت مع **أخو زوجتي** هذا الصباح.

42. I talked to **'my brother-in-law'** this morning.

43. تم **حديثا** إجراء **تعديلات** على نظام القبول و التسجيل.

43. Several **'amendments/changes'** in the Admission and Registration System have been made 'recently'.

44. المريض الان في حالة **غيبوبة**.

44. The patient is now in a **'coma'**.

45. لقد **عامت** الشوارع بالمياة نتيجة **إنسداد** في الأنابيب.

45. The streets have **'overflowed'** with water because of **'blockage'** in drainpipes.

(d)

46. ناقش **مجلس الوزراء** الخطة السنوية **للبينة التحتية**.

46. The **'cabinet'** discussed the annual **'infrastructure'** plan.

47. البحر **هادى** والجو مناسب للرحلة.

47. The sea is **'calm'** and the weather is suitable for picnic.

48. **الملاحة الفلكية** معقدة.

48. **'Aerospace'** is complex.

49. هذا شخص **مغرور**.

49. He is an **'egotistical/arrogant'** person.

50. الماء **متلج** (مجمد) ولايمكن شربة الان.

50. The water is **'frozen'** and cannot be drunk.

51. هذا **عمل فروسي**.

51. This is a **'chivalrous'** deed.

52. شارك الالاف في إحتفالات **عيد الميلاد ورأس السنة**.

52. Thousands of people participated in the **'Christmas'** and **'New Year's'** celebrations.

53. مازن **مُدمن** على **المخدرات**.

53. Mazen is **'addicted'** to drugs.

54. **إستشهد** الطالب بعبارات المؤلف.

54. The student **'quoted'** the author's words.

55. وافق **المدرب** على إضافة ثلاثة طلبة جدد إلى الفريق.

55. The **'coach'** agreed to add three new students to the team.

56. ربما **يتزامن** عيد ميلادي **مع** بداية شهر رمضان هدا العام.

56. This year the beginning of Ramadan will probably '**coincide with**' my birthday.

57. إن **التقليل أو المبالغة** في هذا الأمر خطأ كبير.

57. It is a big mistake to '**underestimate**' or '**overestimate**' the matter.

58. خالد الان **مُعتقل** في السجن.

58. Khalid is now '**detained**' in prison.

59. ذهب عمر **لإصلاح** سيارتيه.

59. Omar has gone to get his car '**repaired**'.

60. غداً إمتحان **التذوق الادبي**.

60. The Literary '**Appreciation**' test is tomorrow.

(e)

61. أنا **مُفلس** (لا أملك نقود).

61. I am '**broke**'.

62. التدخين **ممنوع** في **الاماكن العامة**.

62. Smoking is '**forbidden**' in '**public**' places.

63. هذة المرأة **لاتُنجب**.

63. This woman is '**infertile/barren**'.

64. يجب لا أن لا نسمح لإحد بأن **يُوقف** عملية السلام.

64. We should not allow anybody to '**halt**' the peace process.

65. تم نقل **مكتب** علي إلى **المستودع**.

65. Ali's '**desk**' has been moved to the store.

66. يُمنع إستخدام **الزامور** بالقرب من المستشفيات.

66. '**Horns**' shouldn't be honked near hospitals.

67. هذة الافكار **تُصعد** أعمال العنف.

67. These ideas '**escalate**' the violence acts.

68. هذا اللقاء خاص بقناة الجزيرة **على وجة الحصر**.

68. This interview is '**exclusive**' for Al Jazeera Satellite Channel.

69. نبيل **يتيم**.

69. Nabeel is an '**orphan**'.

70. لا أحبذ **العمل الاضافي**.

70. I do not like to work '**overtime**'.

71. **معجون الأسنان** أغلى من **معجون الحلاقة**.

71. '**Toothpaste**' is more expensive than '**shaving cream**'.

72. **العجز** في الموازنة مؤشر على **فشل** في التخطيط.

72. The '**budget deficit**' is an indicator of the failure of planning.

73. أحضر لإمتحان **التربية الرياضة**.

73. I am preparing for '**Physical**' Education test.

74. هذة أمور **سطحية**.

74. These are '**superficial**' issues.

75. هذة الجامعة غير **معتمدة** لدى وزارة التعليم العالي.

75. This university is not '**accredited**' by the Ministry of Higher Education.

(f)

76. **مؤيدوا** المشروع في **تناقص**.

76. The '**proponents**' of the project are decreasing.

77. خضار هذا السوبرماكت **طازجة**.

77. The vegetables in this supermarket are '**fresh**'.

78. تمت زيادة **أجور** الموظفين.

78. The employees '**wages**' have been raised.

79. سوف **يستغرق** العمل في هذا **المشروع** شهراً كاملا.

79. Work on this '**project**' will '**take**' a whole month.

80. . **يثور** بركان هاواي كل عامين.

80. The Hawaii volcano '**erupts**' once every two years.

81. مياة هذا النهر **ملوثة** بمجلفات الحيوانات.

81. The water of this river is '**polluted**' by animal dung.

82. أسعار الكتب في **المكتبات** التجارية مُرتفعة.

82. Books prices are high in '**bookstores**'.

83. **المغسلة** (غرفة غسيل الملابس) في هذه المنطقة السكنية معطلة.

83. The '**laundry**' in this '**housing area**' is out of order.

84. ليس من السهل تحقيق **الإكتفاء الذاتي**.

84. It is not easy to achieve '**self sufficiency**'.

85. تم **فصل** المانيا إلى دولتين عقب الحرب العالمية الثانية.

85. Germany was '**divided**' into two countries after World War II.

86. نجاحك في الإمتحان **يضمن** قبولك في الجامعة..

86. Passing the exam will **'guarantees'** your admission to the University.

87. **النتائج الأولية** للا نتخابات تختلف عن **النتائج شبة النهائية**.

87. **'The elections initial results'** differ from **'the semi-final results'**.

88. من الصعب الإتصال في **أوقات الذروة**.

88. It is difficult to call during **'the peak hours'**.

89. سعد **مُعفى** من الرسوم.

89. Saa`d is **'exempt'** from fees.

90. تم **إعفاء** صالح من خدماتة.

90. Saleh has been **'dismissed'** from his job.

(g)

91. رائحة هذة الورود سئية.

91. The scent of these flowers is rather '**unpleasant**'.

92. عين أحمد أمينا للصندوق.

92. Ahmad was appointed to a '**teller**'.

93. تم ختم كشوف العلامات.

93. The '**transcripts**' have been '**stamped**'.

94. سوف التقي أعضاء مجلس الأمناء غداً.

94. I will meet with the '**board of trustees**' tomorrow.

95. التبرع بالأعضاء أو الدم صدقة جارية.

95. '**Blood and organ donation**' is an everlasting charity.

96. سيتم تخصيص مبلغ لهذا الامر في السنة المالية القادمة.

96. A sum of money will be '**allotted**' for this matter during the next fiscal year.

97. هو نحيل الجسم (غير سمين).

97. He is '**skinny**'.

98. يمتلك الوزراء حصانة دبلوماسية.

98. Ministers have diplomatic '**immunity**'.

99. سعر الفائدة في هذا لبنك مُرتفعة.

99. This bank's '**interest rate**' is high.

100. تُعاني كوثر من صداع حاد.

100. Kawther is suffering from a '**sever**' headache.

101. هذا احد **أبرز** شخصيات هذا القرن.

101. He is one of the most '**prominent**' figures of the century.

102. هذة الجهاز يعمل على **الطاقة الشمسية**.

102. This device is solar-'**powered**'.

103. **الرشح** (تسريب المياة) و **التبخر** أخطر التحديات للزراعة.

103. Water '**leakage**' and '**evaporation**' are the most serious challenges for agriculture.

104. مازالت **العبودية** موجودة في عدد من دول أفريقياء.

104. '**slavery**' still exists in some African countries.

105. تكثر **الإجازات المرضية** في أوقات الإمتحانات.

105. '**Sick-leaves**' abound during test times.

(h)

106. يجب أن تكون الحركة **بأتجاه عقارب الساعة**.

106. Motion should be in a '**clockwise**' direction.

107. تأسست **جامعة** الدول العربية في أربعينيات **القرن** المُنصرم.

10 7. The Arab '**League**' was founded in the forties of the last century.

108. يبدأ **التوقيت الصيفي** في شهر نيسان.

108. '**Daylight saving time**' begins in April.

109. يصدر **الفائض** من الإنتاج إلى أميركا.

109. The '**surplus**' production is exported to America.

110. **التأمين الصحي ضروري**.

110. '**Medical insurance**' is necessary..

111. قررت **اللجنة الفرعية** إلغاء العطاء.

111. The '**subcommittee**' decided to cancel the bid.

112. تعاني الحوامل من **التقيؤ** المتكرر أثناء الحمل.

112. Expectant mothers (pregnant women) suffer from frequent '**vomiting**' during '**pregnancy**'.

113. يحصل **الأطباء البيطريين** على مرتبات تفوق تلك التي يحصل عليها **أطباء الاسنان**.

113. '**Vets**' get higher salaries than '**dentists**'.

114. تتم **فطامة** الأطفال بعد العام الثاني.

114. Babies are '**weaned**' after they are two years old.

115. تمتد **كفالة إطارات السيارات** إلى أكثر من سنتين.

115. Car tires '**warranty**' extends for more than two years.

116. أسعار الجملة أفضل من أسعار المفرق (التجزئة).

116. **'Wholesale prices'** are better than **'retail prices'**.

117. هذا الجاكيت واقي من الماء.

117. This jacket is **'waterproof'.**

118. هذة الأرض صالحة للزراعة.

118. This land is **'arable.**

119. صوم رمضان أحد أركان الإسلام.

119. **Fasting** during Ramadan is one of the **'pillars'** of Islam.

120. الضرب من العمليات الحسابية.

120. **'Multiplication'** is a mathematical operation.

(i)

121. أتشاور مع أصدقائي حول هذا الموضوع يومياً.

121. I **'confer'** with my friends every day on this issue.

122. ألقى الرئيس الفرنسي **خطاباً شديد اللهجة** حول **الهجرة غير الشرعية**.

122. The French president delivered a **'forceful'** speech about **'illegal immigration'**.

123. يمكن شراء سيارة بدون **دفعة أولى**.

123. You can buy a car without a **'down payment'**.

124. لايوجد **دليل قطعي** لادانة ماهر.

124. There is no **'conclusive evidence'** to convict Maher.

125. وبعد ذلك **رفع** الرئيس الجلسة إلى يوم الأحد القادم.

125. After that the President **'adjourned'** the session.

126. تحدثت مع **زوجة أبي** الأسبوع الماضي.

126. I talked to my **'stepmother'** last week.

127. يهتم المزارعون **بالنشرة الجوية** كثيراً.

127. Farmers have great interest in the **'weather forecast'**.

128. **سلّم** عمر نفسة إلى الشرطة.

128. Omar has **'surrendered'** to the police.

129. يجب إرفاق السيرة الذاتية عند تقديم **طلب وظيفة**.

129. The CV should be submitted along with the **'application'**.

130. **الأشارات الضوئية** مُعطلة اليوم.

130. The **'traffic lights'** are out of order today.

131. هذة الاحوال **مثيرة للشفقة**.

131. These conditions are **'pitiful'.**

132. الديناصور حيوان **مُنقرض**.

132. The dinosaur is an '**extinct**' animal.

133. أنا سهل **الإغراء**.

133. I am easily' **tempted**'

134. يتحمل الأباء مسؤلية دفع جميع **الفواتير** الشهرية.

134. Fathers are responsible for paying all monthly '**bills**'.

135. **إستنشاق** الهواء الملوث **مضر لصحة** المرء.

135. **Breathing** polluted air is '**harmful**' to one's health.

(j)

136. تم إنتخاب عادل رئيسا **فخريا** للنادي.

136. Adel has been elected as an **'honorary** 'president of the club.

137. **إمتنع** النائب عن التصويت.

137. The representative **'abstained'** from voting.

138. ما زالت هنالك شعوب تطالب بحق **تقرير المصير**.

138. There are still peoples who strive for **'self-determination'**.

139. **الجبال الجليدية** تهدد حركة الملاحة.

139.**' Icebergs'** threaten navigation.

140. **التلقين** أسلوب تعليمي خاطىء.

140**. 'Spoon feeding** is a bad teaching method.

141. يُعرف العرب بأهتمامهم في **الضيافة**.

141. The Arabs are famous for their **'hospitality'**.

142. ماهر **مدمن** على المخدرات.

142. Maher is **'addicted'** to drugs.

143. تصدر القوانين **بأرادة ملكية سامية**.

143. Laws are issued by an **'(exalted) Royal decree'**.

144. **نُفي** نابليون أكثر من مرة.

144. Napoleon was' **exiled'** more than once.

145. تم **إجلاء** سكان المناطق المنكوبة.

145. The inhabitants of the disaster areas were **'evacuated'**.

146. **الإجهاض** عمل غير مشروع.

146. '**Abortion**' is an illegal act.

147. **أغتيل** جون كندي في ولاية تكساس.

147. John Kennedy was **'assassinated'** in Texas.

148. من الايام التي **لاتُنسى** في تاريخ الجامعة **يوم الإفتتاح**.

148. An unforgettable day in the history of the university is its **'inauguration'** day.

149. **الغالبية المُطلقة** من أعضاء مجلس العمداء **تؤيد** القرار.

149. The **'vast majority'** of the Dean's Council is in favor of the decision.

150. تمت **ترقيتية** إلى رتبة أستاذ **مشارك**.

150. He has been **'promoted'** to the rank of **'associate'** professor.

(k)

151. **تجاهُل** هذا الأمر أدى إلى **كارثة** كبيرة.

151. Ignoring this issue led to a big **'disaster'**.

152. **العمليات التجميلية** ليست **مغطاة** في التأمين الصحي.

152. **'Cosmetic surgery'** is not covered by health insurance.

153. عمر **بائع متجول**.

153. Omar is a **'peddler'**.

154. ألغي **التجنيد الإجباري** في كثير من الدول.

154. **'Compulsory military service'** has been **'abolished'** in many countries.

155. أصبحت **امينا للسر** و خالد أميناً للمكتبة.

155. I have become a **'secretary'** and Khalid has become a librarian.

156. سوف يفقد عميد الكلية **مصداقيته** اذا لم يُنفذ ما وعد به.

156. The Dean will lose his **'credibility'** if he does not keep his promise.

157. حصل أسامة علي **براءة إختراع**.

157. Usama has taken out a **'patent'**.

158. **الحنث** (الكذب) في اليمين أصبح شائعا هذة الايام.

158. **'Perjury'** is common nowadays.

159. من المُخجل أن يفقد المرء **ماء وجهه**.

159. It is shameful to **'lose face'**.

160. **عُلقت** المحاضرات هذا اليوم.

160. Classes have been **'suspended'** today.

161. تركيا دولة **علمانية**.

161. Turkey is a **'secular'** country.

162. غنياً عن القول زوحتي **غيورة**.

162. It goes without saying that my wife is '**jealous**'.

163. **عدل** البرلمان بعض مواد **الدستور**.

163. The Parliament has '**amended**' some of the articles of the '**constitution**'.

164. عزمت **عديلي (زوج اخت زوجتي)** على العشاء.

164. I invited my '**wife's brother-in-law**' to dinner.

165. لو سمحت **إرفع** صوت التلفزيون.

165. Could you please '**turn up**' the volume of the T V.'?

(l)

166. **تعدد الزوجات** أمر **غير قانوني** في بعض البلدان.

166. **'Polygamy'** is illegal in some countries.

167. مازالت عواطف **ملتزمة بمبادئها**.

167. A`awatif sill **'adheres' to** her principles.

168. فؤاد **متعدد الألسن** (يتكلم عدة لغات).

168. Fuad is **'multilingual'**.

169. يُعرض الجنود حياتهم للمخاطر من أجل **سلامة** أوطانهم.

169. Soldiers expose their lives to' **risks'** for the protection of their country.

170. قرر مجلس الإدارة **فصل** عدد من الموظفين.

170. The administrative board decided to **'dismiss'** some of the employees.

171. هذا شخص **محتال**.

171. This man is a **'crook'**.

172. **إنخفضت** درجة الحرارة.

172. The temperature suddenly **'dropped'**.

173. لغة الحيونات **موروثة** وليست **مكتسبة**.

173. Animals language is **'inherited'** rather than **'acquired.'**

174. يوم الأحد هو **اخر موعد لتقديم** الورقة البحثية.

174. Sunday is the **'deadline'** for **'submitting'** the term papers.

175. **التمّيز** العرقي عادة سئية يُعاني منها الكثير من الناس.

175. Racial **'discrimination'** is a bad practice from which many people suffer.

176. تؤكد **مسودة القرار** الجديد **عدالة** الرئيس.

176. The new **'draft of the decision'** shows the president's **'fairness'**.

177. زخات **متفرقة** من المطر متوقعة اليوم.

177. **'Scattered showers** (of rain) are expected today.

178. يجب ان **لاتقلل** من خطورة الأمر.

178. You should not' **underestimate'** the danger of the issue.

179. **طرحت** دورة **مكثفة** لتعليم اللغة العربية في الجامعة.

179. An '**intensive'** Arabic course was '**offered'** at the University.

180. نادر شاب **مسرف** (مبذر).

180. Nader is '**spendthrift'**.

(m)

181. نُشرت **حديثاً مذكرات** هيلاري كلنتون.

181. Hilary Clinton's **'memoirs'** have been **'recently'** published.

182. يحاول الكثير من الباحثين **إكتشاف** أسرار الحضارة الفرعونية.

182. Many scholars try to **'discover'** the secrets of the **'civilization'** of Pharaohs.

183. لايوجد **وظائف شاغرة** الان.

183. There are no job **'vacancies'** now.

184. أتُهم طارق **بتزوير** عدة **وثائق**.

184. Tariq was accused of **'forging'** many documents.

185. ياسر هو **المستفيد** الوحيد من الميراث.

185. Yaser is the sole **'beneficiary'** of the **'inheritance'**.

186. **ترفيع** مراد كان **على حساب** أسامة.

186. The promotion of Murad was **'at the expense of'** Usama.

187. **نشر الجنود** في هذا المكان كان خطأ كبيراً.

187. The **'deployment'** of troops to this area was a big mistake.

188. وسوف يؤدي هذا الامر إلى **قطع** العلاقات البلوماسية بين البلدين.

188. This will lead to the **'severing'** of diplomatic relations

189. **سدد** دينه **على أقساط**.

189. He paid his debt in **'installments'**.

190. كان هذا الموضوع مدار بحث **لدراسات مكثفة**.

190. This issue has been the main focus of several **'in-depth'** studies.

191. هذا الدواء يُساعد على **فتح الشهية**.

191. This medicine helps to '**whet the appetite**'.

192. رفض **الكاتب العدل** توقيع الوثيقة.

192. The '**notary public**' refused to sign the document .

193. القرار المالي الجديد **يُحفز** الإقتصاد الوطني.

193. The new financial decision '**stimulates**' the national economy.

194. لقد **واجهنا** العدو ببسالة **منقطعة النظير**.

194. We faced the enemy in an '**extreme/unsurpassed valor**'.

195. **بإختصار** تعتبر **الرشوة** أسواء الظواهر الإجتماعية.

195. In short, '**bribery**' is regarded the worst social behavior.

(n)

196. طلبت رنا حق **اللجوء** الساسي في بريطانيا.

196. Rana applied for political '**asylum**' in Britain.

197. رأيت **اللبوة** (أنثى الاسد) تلاحق قطيع من الإبقار.

197. I saw the '**lioness**' chasing a herd of cows.

198. أكد **مراسل** محطة CNN أن **مؤتمر القمة** سوف يعقد في موعدة.

198. The CNN '**reporter**' said the '**Summit Conference**' will be held at its appointed time.

199. من هو **محرر** هذا الكتاب؟

199. Who is the '**editor**' of this book?

200. غادر **المشاركون** في الندوة.

200. The '**participants**' in the symposium (have) left.

201. أنا لدي الرغبة ولكنني **افتقد** القدرة على القيام بذلك.

201. I would like to but I '**lack**' the ability to do so.

202. **إعترف** خالد بأنة **غش** في الإمتحان.

202. Khalid '**confessed**' that he '**cheated**' on the exam.

203. أعلن **المسجل العام** قبول عدد من الطلبة الأجانب هذا العام.

203. The '**Registrar General**' announced the admission of a number of foreign students this year.

204. الإعلانات التجارية مُكلفة.

204. **'Commercials'** are expensive.

APPENDIX C

COLLOCATIONAL TRANSLATION TEST

Part (A)

* ترجم ما يلي الى الانجليزية.

*** Translate the following sentences into English**

1. أمطار غزيرة متوقعة اليوم.

1. **Heavy rainfalls** are expected today.

---

2. إغلق المطار بسبب الضباب الكثيف.

2. The airport was closed because of the **heavy fog.**

---

3. إستيقضت متأخرا هدا الصباح من نوم عميق.

3. I woke up late this morning from a **heavy sleep.**

---

4. أحمد مدخن مفرط.

4. Ahmad is a **heavy smoker.**

---

5. أفضل النوم قليلا بعد كل وجبة دسمة.

5. I prefer taking a nap after every **heavy meal.**

---

6. يمنع الصيادون من صيد السمك في البحار الهائجة.

6. Fishermen are prevented from fishing in **heavy seas.**

---

7. صناعة السيارات واحدة من أهم الصناعات الثقيلة في الولايات المتخدة الامريكية.

7. Automobile industry is one of the most important **heavy industries** in America.

---

8. بقلب مثقل (بالاسى) شاهدتة يغادر.

8. With a **heavy heart**, she watched him go.

---

9. عانى الجيش من **خسائر فادحة** الاسبوع الماضي.

9. The army suffered **heavy casualties** last week.

---

10. محاصيل وافرة من الفواكة متوقعة هذا العامز.

10. **Heavy crops** of fruits are expected this year.

Part (B)

1. لقد جعلت سلسة **الهجمات** الاخيرة السكان يخافون مغادرة بيوتهم.

1. The latest **series of attacks** have made residents afraid to leave their homes.

2. تاكد بأن النوافد **مغلقة بإحكام**.

2. Make sure that the windows are **firmly closed**.

3. أثر أقدام مختلفة وجدت قرب مسرح **جريمة القتل**.

3. Several footprints were found near the **murder scene**.

4. تأجل **الافتتاح الرسمي** للجامعة.

4. The **official inauguration** of the university has been postponed.

5. لا يوجد **وظائف شاغرة** في هده الشركة الان ولكنهم سيبدأون التوظيف الشهر المقبل.

5. There are no **vacant jobs** in this company right now but it will start hiring next month.

6. مازال أعضاء اللجنة يحاولون **الوصول الى اتفاق** على هدة القضية.

6. The committee members are still trying to **reach** an **agreement** on this particular issue.

7. اصطدمت ثلاثة سيارات في **حادث مميت** على الطريق السريع (الصحراوي) اليوم.

7. Three cars were involved in a **fatal accident** on the desert highway today.

8. سوف تقضي خالدة **اجازاتها المرضية** في عجلون. هي تعبانة جدا هدة الايام.

8. Jane will spend her **sick leave** in Ajloun. She is very tired these days.

9. **صوت** الاعضاء **بالإجماع** لصالح القرار الجديد.

9. The members **unanimously voted** in favor of the new resolution.

10. المرسيدس **سيارة قوية**.

10. Mercedes is a **powerful car**.

APPENDIX D

MULTIPLE CHOICE TEST

Part (A)

Circle the letter of the best answer.

1. _____ **rainfalls** are expected in the forecast.

a. strong b. abundant c. huge d. heavy

2. The airport was closed because of the _____ **fog.**

a. ample b. heavy c. huge d. large

3. I woke up late this morning from a _____ **sleep.**

a. strong b. long c. huge d. heavy

4. Ahmad is a _____ **smoker.**

a. excessive b. waster c. huge d. heavy

5. I prefer taking a nap after every _____ **meal.**

a. large b. big c. heavy d. fatty

6. Fishermen are prevented from fishing in _____ **seas.**

a. waving b. heavy c. high d. wild

7. Automobile industry is one of the most important _____ **industries** in the USA.

a. heavy b. huge c. weighty d. big

8. With _____ **heart,** she watched him go.

a. big b. huge c. hot d. heavy

9. The army suffered _____ **casualties** last week.

a. big b. large c. heavy d. abundant

10. _____ **crops** of fruits are expected this year.

a. heavy b. numerous c. large d. big

Part (B)

1. The latest _____ of **attacks** have made residents afraid to leave their homes.

a. scraps b. sets c. series d. sacks

2. Make sure that the windows are _____ **closed.**

a. closely b. totally c. firmly d. severely

3. Several footprints were found near the **murder** _____.

a. theater b. arena c. setting d. scene

4. The _____ **inauguration** of the university has been postponed.

a. official b. formal c. stamping d. careful

5. There are no _____ **jobs** in this company right now but it will start hiring next month.

a. empty b. blank c. vacant d. busy

6. The committee members are still trying to _____ an **agreement** on this particular issue.

a. arrive b. do c. reach d. perform

7. Three cars were involved in a _____ **accident** on the highway today.

a. killing b. big c. strong d. fatal

8. Jane will spend her _____ **leave** in Ajloun. She is very tired these days.

a. ill b. patient c. sick d. medical

9. The members _____ **voted** in favor of the new resolution.

a. wholly b. fully c. unanimously d. completely

10. Mercedes is a _____ **car.**

a. powerful b. strong c. hard d. forceful

APPENDIX E

CLOZE AND SEMI CLOZE TEST

Cloze

1. The committee members are still trying to _____ an **agreement** on this particular issue.
2. Jane will spend her _____ **leave** in Ajloun. She is very tired these days.
3. Several footprints were found near the **murder** _____.
4. The _____ **inauguration** of the university has been postponed.
5. There are no _____ **jobs** in this company right now but it will start hiring next month.

Semi Cloze

11. The latest s_____ of **attacks** have made residents afraid to leave their homes.
12. Make sure that the windows are f_____ **closed.**
13. Three cars were involved in a f_____ **accident** on the highway today.
14. The members u_____ **voted** in favor of the new resolution.
15. Mercedes is a p_____ **car.**

APPENDIX F

TOPICS OF THE LEARNER CORPUS

1. Parents are the best teachers. Do you agree or disagree?
2. Television has destroyed communication among friends and families. Do you agree or disagree?
3. Independence is the symbol of dignity
4. Nothing is more important than freedom
5. Which place would you most like to visit—USA, Africa, China, UK, Alaska? Why?
6. Describe the best teacher you ever had.
7. I wish I had a million... Then I would...
8. What do you like to do in your free time?
9. Smoking
10. Drinking alcohol…drugs can harm one's health
11. What is your favorite book? And why?
12. What is a good neighbor?
13. The Late King of Jordan
14. The new 21st century has begun. What changes do you think that this new century will bring?
15. People do many different things to stay healthy. What do you do for your health?
16. Watching television is bad for children. Do you agree or disagree?
17. Alone on a desert island
18. It is better to be a member of a group than to be a leader of a group. Do you agree or disagree? Why?
19. My Favorite foods

20 My favorite game

21 Unemployment

22 Some items (such as clothes or furniture) can be made by hand or machine. Which do you prefer-items made by hand or those made by machine? Why?

23 There is nothing that young people can teach older people. Do you agree or disagree?

24 A special Birthday

25 In some societies women and men have almost the same social roles and duties. However, in other societies the idea is completely different. Describe the situation in your country.

26 Discrimination against others (on the basis of religion, race, geography, etc.) is always fatal.

27 Neighbors are the people who live near you. What are the qualities of a good neighbor?

28 Only people who earn a lot of money are successful. Do you agree or disagree?

29 The importance of education

30 Agriculture

31 Mother

32 You have the opportunity to visit a country for two weeks. Which country would you like to visit? Why?

33 It is sometimes said that borrowing money from a friend harms or damages the friendship. Do you agree or disagree? Why?

34 Ten people I would like to meet.

35 Many people visit museums when they travel to new places. Why do think that people visit museums?

36 It is better for children to grow up in the countryside than in a big city.

37 Things that make you cry.

38 Things I want to accomplish by time I am 40 years old.

39 Christopher Reeve

40 Describe a scary situation you passed through during your life.

41 My School

42 Pollution is a dangerous enemy for us. Why?

43 People attend universities for different reasons (e.g. getting knowledge, getting experience, getting a job, etc.). What about yourself?

44 A book you have recently read

45 Describe your daily schedule (form waking up till sleeping).

46 People do many different things to stay healthy. What do you do for your health?

47 A baby sees with his ears

48 What is your favorite holiday or vacation? What makes it special?

49 My father

50 People are never satisfied with what they have; they always want something more and different.

51 Some people prefer spending their time alone. Others like to be with friends most of the time. To which category you belong?

52 If you could go back to some time and place in the past, when and where you would like to go? Why?

53 Describe a custom from your country that you would like people from other countries to adopt.

54 Telephones and emails have made communication between people much easier.

57 Many teachers assign home works to students every day. Do you think that homework is necessary for students?

58 One should never judge a person by external appearances. Do you agree or disagree? Why?

59 What are the important qualities of a good son or daughter? Why?

60 Some people prefer to travel with a companion. Others prefer to travel alone. Which do you prefer?

61 Some people believe that university students should attend classes. Others believe that going to classes should be optional. Which point view do you agree with? Why?

62 Learning about our past has no value. Do you agree or disagree?

63 Which of the following transportation vehicles has changed the lives of people? Bicycles, buses or airplanes.

64 It has been announced that a new movie theater is going to be built in your neighborhood. Do you support or oppose this plan? Why?

65 When people succeed, it is because of hard work. Luck has nothing to do with success. Do you agree or disagree?

66. Grades (marks) encourage students to learn. Do you agree or disagree? Why?

67 Some people enjoy change and they look forward to new experience. Others like their lives to stay the same, and they don't change their usual habits. Which do you

prefer? Why?

68 An extraordinary creature

69 Helping poor people

70 Good planning leads to success.

71 Describe a vacation you enjoyed.

72 Tourism in Jordan

73 What do you think are the major causes of divorce in our society? Explain.

74 What steps need to be taken in order to reduce crime? Explain

75 If you could have a conversation with a famous person (living or dead), whom would you choose? Discuss.

76 Do you think that sports help develop good character? Discuss.

77 My daily schedule.

78 What being a friend means to you.

79. A sad event in my life

80. The woman I would like to marry

81. The man I would like to marry

82. A date with the death

83. My sole wish

84. My country

85. An important invention

86. My university

87. Petra

88. A goal to achieve

89. Friendship
90. Reading
91. Peace
92. Albert Einstein
93. The Queen Mother
94. Jordan first
95. Working at home
96. Describe a trip you enjoyed
97. The newspaper or magazine I like best
98. A scary picnic
99. Which h story do hold to be your favorite? Give reasons for your answer.
100. Taj Mahal
101. Two American Presidents
102. My room
103. A goal to achieve
104. Charles Dickens
105. My sister
106. Describe a person you know
107. Criminals know no mercy
108. Olympics games
109. University study
110. The Original Oak
111. Alexander the Great

112. Selfishness

113. Unforgettable moments

114. Black Gold

115. What do you like to do in your free time?

116. My brother

117. Consider a time when you bought something because an advertisement convinced you to buy it. Were you disappointed with the product? Or were you happy with it?

118. If you were offered a job that requires telecommunicating, would you accept it? Why or why not?

119. My purpose when I attended university

120. An interesting experience

121. Tests have no value without the reader

122. Lancaster

123. What's happiness?

124. Some difficulties at my university

125. Ramadan

126. Double Mood

127. Sport in Jordan

128. Difficulties in learning English

129. Cars: advantages and disadvantages

130. Tourism

131. Freedom

132. What a miserable life!

133. I' m different

134. Transportation

135. Long life

136. Types of library

137. Life

138. Problems we face

139. My family

140. Learning

141. The most important thing in life

142. Is anger useful?

143. The importance of education

144. My favorite holiday

145. Agriculture

146. My favorite story

147. Al-Arab Al-Yawm news paper

148. The Rose Red City

149. A letter

150. University life

151. Building a university in the desert

152. Horror world

153. Habits and traditions

154. Organs of speech

155. Shopping!!!

156. Old people

157. Alps

158. Getting a job

159. A trip to the Dead Sea

160. Quality of university foods

161. What are the important qualities of a good son or daughter? Why?

162. Ants

163. Difficulties of Registration

164. A wish that you have every body in this life

165. Poverty

166. Gambling

167. Real Friends

169. A collection of final exams in literature

APPENDIX G

DEMOGRAPHIC QUESTIONNAIRE

1. **Academic Level:** 1st Year 2nd Year 3rd Year 4rth Year

2. **Sex:** Female 2. Male

3. **Native Tongue:** Arabic Other_____

4. **Have you lived in an English speaking country (i.e. USA, UK, Australia, etc)?**

YES NO

If yes, from grade_____ till grade_____.

5. **Have you studied in a private school?** YES NO

6. **Age:** under 18 18 and over

APPENDIX H

LEARNER AND REFERENCE CORPORA

A. REFERENCE CORPUS: ***LOCNESS CORPUS OF NATIVE ENGLISH ESSAY WRITING***

Total number of essays selected for the purposes of this study is 79.

Total number of tokens is 70, 309

1. American Universities 1. R. 1.1 Marquette University (codes: ICLE-US-MRQ)			
Selected tokens	**Time**	**Type**	**pantsPartci**
10,125	March 1995	Untimed Argumentative essays	NSs of English Age: from 18 to 21 (+ 1 of 30, 1 of 31 and 1 of 40)

2. R. 1.2 Indiana University at Indianapolis (codes: ICLE-US-IND)			
sSelected token	**Time**	**Type**	**Partcipants**
13,629	March 1995	Timed Argumentative essays	NSs of English Age: from 22 to 48

3. R. 1.3 Presbyterian College, South Carolina (codes: ICLE-US-PRB)			
Selected tokens	**Time**	**Type**	**Partcipants**
12,447	April 1995	Untimed Argumentative essays	NSs of English Age: from 2o to 22

R. 2. BRITISH ESSAYS: University students			
1. R. 2. 1. brsur.cor			
Selected tokens	**Time**	**Type**	**Partcipants**
11, 570	March 1991	Exams, literary, historical and expositry essays	NSs of English

2. R. 2. 1. brsur.cor			
Selected tokens	**Time**	**ypeT**	**Partcipants**
11, 405	----------	Exams, literary,	NSs of English

2. R. 2. 1. brsur.cor			
Selected tokens	**Time**	**Type**	**Partcipants**
11, 133	----------	argumentive	NSs of English

B. LEARNERS ARGUMENTATIVE, NARRATIVE, PROCEDURES, AND LITERARY ESSAYS AND EXAMS

Total Number of essays: 429

Total number of tokens: 70,307

1. Al-Hussein Bit Talal University			
Selected tokens	**Time**	**Type**	**Partcipants**
17, 012	August-December 2003	Timed and untimed Argumentive, procedurers, literary, expository essays and exams	NSs of Arabic majoring in English language and literature Age: 18-25

2. Mutah University			
Selected tokens	**Time**	**Type**	**Partcipants**
9, 347	August-December 2003	Timed and untimed Argumentive, procedurers, literary, expository essays	NSs of Arabic majoring in English language and literature Age: 18-25

Hashemite University			
Selected tokens	**Time**	**Type**	**Partcipants**
21, 216	August-December 2003	Timed and untimed Argumentive, procedurers, literary, expository essays and exams	NSs of Arabic majoring in English language and literature Age: 18-25 (+3 students over 25)

Al al-Bayt University			
Selected tokens	**Time**	**Type**	**Partcipants**
18, 516	August-December 2003	Timed and untimed Argumentive, procedurers, literary, expository essays	NSs of Arabic majoring in English language and literature Age: 18-25

Zarqa National University			
Selected tokens	**Time**	**Type**	**Partcipants**
4, 216	August-December 2003	Timed and untimed Argumentive, procedurers, literary, expository essays	NSs of Arabic majoring in English language and literature Age: 18-25

APPENDIX I

UCREL CLAWS7 TAGSET

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
BCL	before-clause marker (e.g. in order (that),in order (to))
CC	coordinating conjunction (e.g. and, or)
CCB	adversative coordinating conjunction (but)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	as (as conjunction)
CSN	than (as conjunction)
CST	that (as conjunction)
CSW	whether (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)
DA1	singular after-determiner (e.g. little, much)
DA2	plural after-determiner (e.g. few, several, many)
DAR	comparative after-determiner (e.g. more, less, fewer)
DAT	superlative after-determiner (e.g. most, least, fewest)
DB	before determiner or pre-determiner capable of pronominal function (all, half)
DB2	plural before-determiner (both)
DD	determiner (capable of pronominal function) (e.g any, some)
DD1	singular determiner (e.g. this, that, another)
DD2	plural determiner (these,those)
DDQ	wh-determiner (which, what)
DDQGE	wh-determiner, genitive (whose)
DDQV	wh-ever determiner, (whichever, whatever)
EX	existential there
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (' or's)
IF	for (as preposition)
II	general preposition
IO	of (as preposition)
IW	with, without (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. older, better, stronger)

JJT	general superlative adjective (e.g. oldest, best, strongest)
JK	catenative adjective (able in be able to, willing in be willing to)
MC	cardinal number,neutral for number (two, three..)
MC1	singular cardinal number (one)
MC2	plural cardinal number (e.g. sixes, sevens)
MCGE	genitive cardinal number, neutral for number (two's, 100's)
MCMC	hyphenated number (40-50, 1770-1827)
MD	ordinal number (e.g. first, second, next, last)
MF	fraction,neutral for number (e.g. quarters, two-thirds)
ND1	singular noun of direction (e.g. north, southeast)
NN	common noun, neutral for number (e.g. sheep, cod, headquarters)
NN1	singular common noun (e.g. book, girl)
NN2	plural common noun (e.g. books, girls)
NNA	following noun of title (e.g. M.A.)
NNB	preceding noun of title (e.g. Mr., Prof.)
NNL1	singular locative noun (e.g. Island, Street)
NNL2	plural locative noun (e.g. Islands, Streets)
NNO	numeral noun, neutral for number (e.g. dozen, hundred)
NNO2	numeral noun, plural (e.g. hundreds, thousands)
NNT1	temporal noun, singular (e.g. day, week, year)
NNT2	temporal noun, plural (e.g. days, weeks, years)
NNU	unit of measurement, neutral for number (e.g. in, cc)
NNU1	singular unit of measurement (e.g. inch, centimetre)
NNU2	plural unit of measurement (e.g. ins., feet)
NP	proper noun, neutral for number (e.g. IBM, Andes)
NP1	singular proper noun (e.g. London, Jane, Frederick)
NP2	plural proper noun (e.g. Browns, Reagans, Koreas)
NPD1	singular weekday noun (e.g. Sunday)
NPD2	plural weekday noun (e.g. Sundays)
NPM1	singular month noun (e.g. October)
NPM2	plural month noun (e.g. Octobers)
PN	indefinite pronoun, neutral for number (none)
PN1	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)
PNQO	objective wh-pronoun (whom)
PNQS	subjective wh-pronoun (who)
PNQV	wh-ever pronoun (whoever)

PNX1	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	3rd person sing. neuter personal pronoun (it)
PPHO1	3rd person sing. objective personal pronoun (him, her)
PPHO2	3rd person plural objective personal pronoun (them)
PPHS1	3rd person sing. subjective personal pronoun (he, she)
PPHS2	3rd person plural subjective personal pronoun (they)
PPIO1	1st person sing. objective personal pronoun (me)
PPIO2	1st person plural objective personal pronoun (us)
PPIS1	1st person sing. subjective personal pronoun (I)
PPIS2	1st person plural subjective personal pronoun (we)
PPX1	singular reflexive personal pronoun (e.g. yourself, itself)
PPX2	plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY	2nd person personal pronoun (you)
RA	adverb, after nominal head (e.g. else, galore)
REX	adverb introducing appositional constructions (namely, e.g.)
RG	degree adverb (very, so, too)
RGQ	wh- degree adverb (how)
RGQV	wh-ever degree adverb (however)
RGR	comparative degree adverb (more, less)
RGT	superlative degree adverb (most, least)
RL	locative adverb (e.g. alongside, forward)
RP	prep. adverb, particle (e.g about, in)
RPK	prep. adv., catenative (about in be about to)
RR	general adverb
RRQ	wh- general adverb (where, when, why, how)
RRQV	wh-ever general adverb (wherever, whenever)
RRR	comparative general adverb (e.g. better, longer)
RRT	superlative general adverb (e.g. best, longest)
RT	quasi-nominal adverb of time (e.g. now, tomorrow)
TO	infinitive marker (to)
UH	interjection (e.g. oh, yes, um)
VB0	be, base form (finite i.e. imperative, subjunctive)
VBDR	were
VBDZ	was
VBG	being

VBI	be, infinitive (To be or not... It will be ..)
VBM	am
VBN	been
VBR	are
VBZ	is
VD0	do, base form (finite)
VDD	did
VDG	doing
VDI	do, infinitive (I may do... To do...)
VDN	done
VDZ	does
VH0	have, base form (finite)
VHD	had (past tense)
VHG	having
VHI	have, infinitive
VHN	had (past participle)
VHZ	has
VM	modal auxiliary (can, will, would, etc.)
VMK	modal catenative (ought, used)
VV0	base form of lexical verb (e.g. give, work)
VVD	past tense of lexical verb (e.g. gave, worked)
VVG	-ing participle of lexical verb (e.g. giving, working)
VVGK	-ing participle catenative (going in be going to)
VVI	infinitive (e.g. to give... It will work...)
VVN	past participle of lexical verb (e.g. given, worked)
VVNK	past participle catenative (e.g. bound in be bound to)
VVZ	-s form of lexical verb (e.g. gives, works)
XX	not, n't
ZZ1	singular letter of the alphabet (e.g. A,b)
ZZ2	plural letter of the alphabet (e.g. A's, b's)

APPENDIX J

HIGH LEXICON

1.	a. depend on	b. confide in	c. count on
2.	a. fortunate	b. lucky	c. prosperous
3.	a. flawless	b. complete	c. perfect
4.	a. simple	b. easy	c. facile
5.	a. strength	b. force	c. potency

REFERENCES

Aarts, Jan. 1991. Intuition Based and Observation Based Grammars. English Corpus Linguistics, ed. by Karin Aijmer and Bengt. Altenberg, 44-62. London: Longman.

Abercrombie David. 1963. Studies in Phonetics and Linguistics. London: Oxford University Press.

Aijmer, Karin and Bengt Altenberg. (eds.) 1991. English Corpus Linguistics. London: Longman.

Al-Zahrani, Mohammad Said. 1998. Knowledge of Lexical Collocations Among Male Saudi College Students Majoring in English at Saudi University. Unpublished Ph.D. Dissertation, Indiana University of Pennsylvania.

Bahns, Jens & Moira Eldaw. 1993. Should we teach EFL students collocations? System 21. 101-114.

Barnbrook, Geoff. 1996. Language and Computers: A Practical Introduction to the Computer Analysis of Language. Edinburgh: Edinburgh University Press.

Beaugrande, Robert de. 2001. 'If I were you...': Language Standards and Corpus Data in EFL. Revista Brasileira de Lingüística Aplicada 1. 117-154.

Berber, Sardinha Tony. 1996. A window on lexical density in speech. (Unpublished?) Paper presented at the 8th Euro-International Systemic Functional Workshop, Nottingham Trent University.

Biber, Douglas. 1993. Representativeness in Corpus Design Literary and Linguistic Computing 8. 243-257.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge; New York: Cambridge University Press.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, & Finegan, Edward. 1999. Longman grammar of spoken and written English. London: Longman.

Biber, Douglas. 2001. Using Corpus-Based Methods to Investigate Grammar and Use: Some Case Studies on the Use of Verbs in English. Corpus Linguistics in North America: Selections from the 1999 Symposium, eds. by Simpson and Swales, 101-115. Ann Arbor: The University of Michigan Press.

Bloomfield, Leonard. 1933. Language. New York: Henry Holt and Company.

Brooks, Nelson. 1960. Language and Language Learning. New York: Harcourt Brace & World.

Brown, Dorothy. 1974. Advanced Vocabulary Teaching: The Problem of Collocation. Regional Language Center Journal (RELC) 5. 1-11.

Btoosh, Mousa. 1999. Hedging in Journalistic Arabic Political Discourse during the Third Gulf War. Unpublished M.A. dissertation, Yarmouk University, Jordan.

Burdine, Stephanie. 2001. The Lexical Phrase as Pedagogical Tool: Teaching Disagreement Strategies in ESL. Corpus Linguistics in North America: Selections from the 1999 Symposium, ed. by Simpson and Swales, 195-210. Ann Arbor: The University of Michigan Press.

Burquest, Donald. 1999. A Field Guide for Principles and Parameters Theory. Dallas, SIL International.

Butler, Christopher. 1985. Statistics in linguistics. Oxford: Oxford University Press.

Carter, Ronald. 1987. Vocabulary: applied linguistic perspectives. London; Boston: Allen & Unwin.

Carter, Ronald and Michael McCarthy. 1988. Vocabulary and Teaching. New York: Longman.

Chafe, Wallace. 1992. The Importance of Corpus Linguistics to Understanding the Nature of Language. Directions in Corpus Linguistics Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991, ed. by Svartivik, Jan, 79-97. Berlin: Mouton de Gruyter.

Chambers, Francine. 1997. What Do We Mean by Fluency? System 25. 535-544.

Channell, J. 1981. Applying Semantic Theory to Vocabulary Teaching. English Language Teaching Journal 35. 115-122.

Chomsky, Noam. 1962. Paper given at the University of Texas, 1958. Third Texas Conference on Problems of Linguistic Analysis in English. Austin, Texas.

Chomsky, Noam. 1965. Aspects of the theory of syntax. M.I.T. Press.

Chomsky, Noam. 1968. Language and Mind. Harcourt Brace Jovanovich, Inc.

Chomsky, Noam. 1975. Reflections on Language. New York: Pantheon

Choueka, Y. 1988. Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases, Proceedings of the RIAO

Conference on User-oriented Context Based Text and Image Handling. 609-623, Cambridge, MA.

Cobb, Thomas. 2003. Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. Canadian Modern Language Review 59. 393-423.

Connor, Ulla. 1996. Contrastive Rhetoric: Cross-cultural aspects of second-language acquisition. Cambridge: Cambridge University Press.

Cook, Vivian. 1996. Minimalism, Vocabulary and L2 Learning. Paper given at AILA, Jyvaskyla.

Corder, Stephen Pit. 1967. The Significance of Learner's Errors. International Review of Applied Linguistics 5.161-9.

Corder, Stephen Pit. 1971. Idiosyncratic errors and Error Analysis, International Review of Applied Linguistics 9. 147-159.

Crystal, David. 1996. The Cambridge Encylopedia of Language. Cambridge: Cambridge University Press

Cumming, A. and Mellow, D. 1996. An Investigation into the Validity of Written Indicators of Second Language Proficiency. Validitation in Language Testing, eds. By Cumming, A. and Berwick, R. 72-93. Clevedon, Avon: Multilingual Matters.

Dafu, Yang 1994. Interlanguage Errors and Cross-linguistic Influence: A Corpus-based Approach to the Chinese EFL Learners' Written Production. Unpublished Ph.D. Dissertation. http://www.clal.org.cn/baseinfo/PHD/yangdafu_eng.htm

David, Cooper. (online). Literacy, Literature,and Learning for Life. http://www.eduplace.com/rdg/res/literacy.html

Deuter, Greenan, Nobel and Philips. 2002. Oxford Collocations Dictionary. Oxford: Oxford University Press.

Doerr, Rita McCardell. 1994. A Lexical-Semantical Approach to Lexical Collocation Extraction for Natural Language Generation. Unpublished Ph. D. dissertation, University of Maryland.

Dulay, Heidi and Mariana Brute. 1974. Natural sequences in child second language acquisition. Language Learning 24. 37-53.

Dulay, Heidi, Mariana Burt, and Stephen Krashen. 1982. Language Two. New York; Oxford: Oxford University Press.

Dulay, Heidi C. and Mariana K. Burt. 1973. Should we teach children syntax? Language Learning 23, p235-252.

Edmonds, Philip Glenny. 1999. Semantic Representations of Near-synonyms for Automatic Lexical Choice. Unpublished Ph. D. dissertation, University of Toronto.

Eggins, Suzanne. 1994. An Introduction to Systemic Functional Linguistics. London: Pinter Publishers.

Ellis, Rod. 1994. The Study of Second Language Acquisition: Oxford; New York: Oxford University Press.

Ellis, Rod. and Roberts, Celia. 1987. Two approaches for investigating second language acquisition in context. Second language acquisition in context, ed. by Rod Ellis, 3-29. Englewood Cliffs, NJ: Prentice Hall.

Engber, Cheryl A. 1992. A Study of Lexis and the Relationship to Quality in Written Texts of Second Language Learners of English (ESL). Unpublished Ph.D. Dissertation, Indiana University.

Engber, Cheryl A. 1992. The Relationship of Lexical Proficiency to the Quality of ESL Compositions. Journal of Second Language Acquisition 4. 139-155.

Engwall, Gunnel. 1994. Not Chance But Choice: Criteria in Corpus Creation. Computational Approaches to the Lexicon, eds. by Atkins, Sue B. T. and Antonio Zampolli, 49-82. Oxford: Oxford University Press.

Farghal, Mohammad and Hussein Obiedat. 1995. Collocations: A Neglected Variable in EFL. International Review of Applied Linguistics 33.315-331.

Fernando, Chitra. 1996. Idioms and Idiomaticity . Oxford: Oxford University Press.

Feyman, R. P., R. B. Leighton and M. Sands. 1963. The Feyman Lectures on Physics 1. Addison-Wesley.

Firth, John R. 1957. Papers in Linguistics: 1934-1951. London:Oxford University Press.

Fox, Len 1979. On Acquiring an Adequate Second Language. Journal of Writing 2. 68-79.

Fries, Charles. 1945. Teaching and learning English as a foreign language. Ann Arbor: University of Michigan Press.

Garside, Roger, Geoffrey Leech and Anthony McEnery (eds.) 1997. Corpus Annotation: Linguistic Information from Computer Text Corpus. London; New York: Longman.

Garside, Roger, Geoffrey Leech and Geoffrey. Sampson. 1987. The Computational Analysis of English: a corpus-based approach. London: Longman.

Gass, Suzan. M. and Larry Selinker. 2001. Second Language Acquisition: An Introductory Course. (2nd Ed.). Mahwah; New Jersey: Lawrence Erlbaum.

Gateway to Corpus linguistics on the Internet (website): http://www.corpus-linguistics.de/corpora/corp_nav_open.html

Geeraerts, Dirk. 1997. Diachronic prototype semantics. Oxford: Oxford University Press.

Gitsaki, Christina. 1999. Second Language Lexical Acquisition: A Study of the Development of Collocational Knowledge. Bethesda: International Scholars Publications.

Grice, Herbert Paul. 1957. Meaning. The Philosophical Review 66. 377-388.

Granger, Sylviane. 2003. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition. TESOL Quarterly 37. 538–546.

Granger, Sylvian and Paul Rayson. 1998. Automatic profiling of learner texts. Learner English on Computer, ed. by Sylviane Granger, 119-131. London; New York: Longman.

Grant, Leslie and April Ginther. 2000. Using Computer-Tagged Linguistic Features to Describe L2 Writing Differences. Journal of Second Language Acquisition 9.123-145.

Groot, Peter J. M. 2000. Computer Assisted Second Language Vocabulary Acquisition. Language Learning and Technology. 4.1: 60-81.

Gully, Adrian. 1996-1997. The Discourse of Arabic Advertising: Preliminary Investigations. Journal of Arabic and Islamic Studies. 1.1-49.

Guo, Xiaotian. 2003. Between Verbs And Nouns And Between The Base Form and the Other Forms of Verbs-A Contrastive Study into COLEC and LOCNESS. English Language Postgraduate Seminars Autumn Term, University of Birmingham.

Haegeman, Liliane. 1999. Introduction to Government and Binding Theory. Oxford: Blackwell Publishers.

Haichour, EL Houcine. 1999. A Corpus Linguistic Analysis of English and Arabic Parallel Business Discourse Domains. Unpublished MA. Thesis, Georgetown University.

Haliday, M. A. K. 1992. Corpus Studies and Probabilistic Grammar. In Aijmer, Karin and B. Altenberg (eds.) 1992. English Corpus Linguistics. London: Longman.

Haliday, M. A. K. 1989. Spoken and written language. Oxford University Press.

Halliday, M. A. K & Ruqaiya Hasan. 1976. Cohesion in English. London: Longman.

Hausser, Roland. 1999. Foundations of Computational Linguistics, Human-Computer Communication in Natural Language, 2nd Edition. Berlin, New York: Springer.

Heliel, Mohamed Hilmi. 1989. Collocations and Translation. Proceedings of the FIT Round Table Professional Arabic Translation and New Technologies. Tanger.

Hinkel, Eli. 2002. Second Language Writers' Text: Linguistic and Rhetorical Features. London: Lawrence Erlbaum Associates, Publishers.

Hladka, Barbora. 2000. Czech Language Tagging. Unpublished Ph.D. Thesis, Charles University, Prague.

Hockett, Charles. 1958. A course in Linguistics. New York: The Macmillan Company.

Hoffman, T. R. 1993. Realms of Meaning: An Introduction to Semantics. London: Longman.

Hsu, Jeng-yih (Tim). 2002. Development of collocational proficiency in a workshop on English for General Business Purposes for Taiwanese college students (China). Unpublished Ph.D. dissertation, Indiana University of Pennsylvania.

Hunston, Susan and Gill Francis. 2000. Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English (Studies in Corpus Linguistics). Amsterdam/Philadelphia: John Benjamins.

Hunston, Suasan. 2002. Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

Hyltenstam, K. 1988. Lexical characteristics of near-native second-language learners of Swedish. Journal of Multilingual and Multicultural Development 9. 67-84.

Hymes, D. 1972. On communicative Competence. Sociolinguistics, ed. by J. B. Pride and J. Holmes. Harmondsworth; England: Penguin Books.

Inkpen, Diana Zaiu and Graeme Hirst. 2002. Acquiring Collocations for Lexical Choice between Near-synonyms. RANLP-2003, Bulgaria.

Jackson, Howard and Etienne Ze Amvela. 2000. Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology. London: CASSELL.

Jain, M. P. 1974. Error Analysis: Source, Cause and Significance. Error Analysis: Perspectives on Second Language Acquisition, ed. by Jack C. Richards, 189-215. London: Longman.

James, Carl. 1998. Errors in Language Learning and Use: Exploring Error Analysis. Longman: London and New York.

Johnstone, Barbara. 1991. Repetition in Arabic Discourse Paradigms, syntagms and the ecology of language. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Judd, Elliot. L. 1978. Vocabulary teaching and TESOL: a need for re-evaluation of existing assumptions. TESOL Quarterly 12.71-6.

Kao, Rong-Rong. 2001. Where have the prepositions gone?: A study of English prepositional verbs and input enhancement in instructed SLA 39.195-215

Kaplan, Robert. B. 1966. Cultural thought patterns in inter-cultural education. Language Learning 16.1-20.

Kennedy, G. 1990. Collocations: Where grammar and vocabulary teaching meet. Language teaching methodology for the nineties, ed. by S. Anivan, 215-229. Singapore: RELC Anthology Series 24. [LOB].

Kennedy, Graeme. 1998. An Introduction to Corpus Linguistics. London; New York: Longman.

Kharma, Nayef. 1985. Advanced Composition in EFL. Abhath Al-Yarmouk 3. 7-22.

Kjellmer, Göran. (2003). Synonymy and corpus work: On almost and nearly. *ICAME Journal* 27. 19-28. http://helmer.aksis.uib.no/icame/ij27/kjellmer.pdf

Klein, Wolfgang: Lexicology and Lexicography. 2001. The International Encyclopedia of the Social & Behavioral Sciences eds. by Smelser, N. J. & P. B. Baltes, 8764-8768. Amsterdam; Paris; New York: Elsevier.

Koenig, Jean-Pierre 1999. Lexical relations. California: CSLI Publications.

Kreuz, Roger. J., & Roberts, Richard. M. 1993. When collaboration fails: Consequences of pragmatic errors in conversation. Journal of Pragmatics 19. 239-252.

Kroll, Barbara. 2003. Exploring the Dynamics of Second Language Writing. Cambridge: Cambridge University Press.

Kubota, Ryuko. 1998. An Investigation of L1 L2 Transfer in Writing among Japanese University Students: Implications for Contrastive Rhetoric 7. 69-100.

Lado, Robert. 1957. Linguistics Across Cultures: Applied Linguistics for Language Teachers. Ann Arbor: University of Michigan Press.

Lafford, Barbara A., Collentine, Joseph. G., Karp, Adam. 2000. The Acquisition of Lexical Meaning By Second Language Learners: An Analysis of General Research Trends with Evidence from Spanish. http://jan.ucc.nau.edu/~jgc/research/vocabstate/

Lager, Torbjörn. 1995. A Logical Approach to Computational Corpus Linguistics. Ph.D. Dissertation, Goteborg University, Sweden.

Larsen-Freeman, D. & V. Strom 1977. The construction of a second language acquisition index of development. Language Learning 27. 123-134.

Larsen-Freeman, D. 1978. An ESL index of development. TESOL Quarterly 12. 439-448.

Laufer, B. (1988), 'The concept of 'synforms' (similar lexical forms) in vocabulary acquisition', Language and Education, 2, 2, 113-132

Laufer, Batia. 1990a. Why some words are more difficult than others? Some intralexical factors that affect the learning of words. International Review of Applied Linguistics 28. 293-308.

Laufer, Batia. 1990b. Sequence and order in the development of L2 lexis: some evidence from lexical confusions. Applied Linguistics. 11. 281-296.

Laufer, Batia. 1994. The lexical profile of second language writing: does it change over time? Regional English Language Center Journal 25.21–33.

Laufer, Batia and Paul Nation. 1995. Vocabulary size and use: lexical richness in L2 written production. Applied Linguistics 16. 307-322.

Laufer, Batia. 1997. What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words. In Vocabulary: Description, Acquisition and Pedagogy. N. Schmitt and M. McCarthy. Cambridge: Cambridge University Press.

Laufer, Batia and Paul Nation. 1999. A vocabulary size test of controlled productive ability. Language Testing 16. 33-51

Lawler, John and Helen Aristar Dry (eds.). 1998. Using Computers in Linguistics. London; New York: Routledge.

Leech, Geoffrey. 1987. General Introduction. The Computational Analysis of English: A Corpus Based Approach, ed. by Garside, Roger, Geoffrey Leech and Geoffrey. Sampson, 1-15. London: Longman.

Leech, Geoffrey. 1992. Corpora and Theories of Linguistic performance. Trends in Linguistics: Direction in Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991, ed. by Jan Svartvik, 125-148. Berlin; New York: Mouton de Gruyter.

Li, Yili. 2000. Linguistic characteristics of ESL writing in task-based e-mail activities. System 28. 229-245.

Liaurui, Yang. 1999. Interlanguage Development: Language Analysis of Some ESL Learners' Language. Unpublished M.A. Thesis. Gonzaga University.

Linnarud, Moira. 1986. Lexis in composition. Lund Studies in English. Malmo, Sweden: Liber Forlag.

Lobner, S. 2002. Understanding Semantics. Arnold; New York: Oxford University Press.

Lombard, Robin Janine. 1997. Non-Native Speakers Collocations: A corpus-Driven Characterization from the Writing of Native Speakers of Mandarin. Unpublished Ph.D. dissertation. The University of Texas at Arlington.

Lord, R. 1974. Learning Vocabulary. International Review of Applied Linguistics 12.239-47.

Lorenz, Gunter 1998. Overstatement in advanced learners' writing: stylistic aspects of adjectives intensification. Learner English on Computer, ed. by Sylviane Granger, 53-66. London; New York: Longman.

Martin, Marilyn. 1984. Advanced Vocabulary Teaching: The Problem of Synonyms. The Modern Language Journal 68. 130-137.

McEnery, Tony & Andrew Wilson. 1996. Corpus Linguistics. Edinburgh: Edinburgh University Press.

McCarthy, Michael 1990. Vocabulary. Oxford: Oxford University Press.

Meara, P. 1983. Vocabulary in a second language. London: Center for Information in Language and Research.

Meunier, Fanny 1998. Computer tools for the analysis of learner corpora. Granger, Sylviane (ed.) 1998. Learner English on Computer. London; New York: Longman.

Meyer, Charles F. 2002. English Corpus Linguistics. Cambridge: Cambridge University Press.

Miller, Jon. 1981. Assessing language production in children. Baltimore: University Park Press.

Nation, Paul. 2001. Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.

Nation, Paul (1994) Morphology and language learning. The Encyclopedia of Language and Linguistics, ed. by Asher, R. E. and Simpson, J. M. Y., 2582-2585. Oxford; Pergamon Press.

O'Grady, William, Michael Dobrovolsky and Francis Katamba (eds.) 1997. Contemporary Linguistics: An Introduction. London: Longman.

Oh, Sun-Young. 2001. Two Types of Input Modification and EFL Reading Comprehension: Simplification and Versus Elaboration. TESOL Quarterly 1. 69-96.

Olsson, M. 1974. A Study of Errors, Frequencies, Origin and Effects. Goteborg, Sweden: Pedagogiska Institution.

Oostdijk, Nelleke. 1991. Corpus Linguistics and the Automatic Analysis of English. Amsterdam-Atlanta: Rodopi.

Ouhalla, Jamal. 1999. Introducing Transformational Grammar: from Principles and Parameters to Minimalism. London: Arnold; New York: Oxford University Press.

Pawley, A. & F. Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. Language and communication, ed. Richards, Jack. C. & Richard. W. Schmidt, London: Longman

Petch-Tyson, Stephanie. 1998. Writer/reader visibility in EFL written discourse. Learner English on Computer, ed. by Sylviane Granger, 119-131. London; New York: Longman.

Petrarca, Mark Paul. 2002. Machine translation: A tool for understanding linguistic challenges facing the second language student. Unpublished Ph.D. dissertation, Indiana University of Pennsylvania.

Politzer, Robert L. 1978. Errors of English Speakers of German As Received and Evaluated by German Natives. The Modern Language Journal 62. 253-261.

Poulisse, Nanda and Theo Bongaerts. 1994. First Language Use in Second Language Production. Applied Linguistics 15. 36-57.

Ramsey, Robert 1981. A Technique for Interlingual Lexico-Semantic Comparison: The Lexigram. TESOL Quarterly 15. 16-25.

Reid, Joy. 1990. Responding to different topics types: A quantitative analysis from a contrastive rhetoric perspective. Second language writing: Research insights for the classroom, ed. by Barbara Kroll, 191-210. Cambridge: Cambridge University Press.

Reiter, E. 1990. A New Model of Lexical Choice for Open-Class Words. In Proc of the Fifth International Workshop on Natural Language Generation (INLGW-1990), pages 23-30. Philadelphia: Dawson.

Richards, Jack C. 1974. Word List: Problems and Prospects. Regional Language Center Journal (RELC) 5. 69-74.

Richards, Jack C. 1976. The Role of Vocabulary Teaching. TESOL 10. 77-89.

Richards, Jack C. 2003. Second Language Writing. Cambridge: Cambridge University Press.

Ringbom, Hakan. 1998. Vocabulary frequency in advanced learner English: a cross-linguistic approach. Learner English on Computer, ed. by Sylviane Granger, 41-52. London; New York: Longman.

Roberts, I. 1993. Verbs and diachronic syntax, Studies in natural language and linguistic theory, Kluwer, Dordrecht.

Robins, Robert H . 1967. A Short History of Linguistics, London, Longman

Rodriguez, Sara. 2000. Universal Grammar and the Acquisition of Clitic Conditions in Spanish as a Second Language. Unpublished Ph.D. Dissertation, State University of New York at Buffalo

Rolin-Ianziti, Jeanne. 2002. Teacher Use of Learners' Native Language in the Foreign Language Classroom. Canadian Modern Language Review 58. 1-22.

Scott, Fred Newton and Joseph Villiers Denney. 1909. Paragraph-Writing: A Rhetoric for Colleges. Boston; New York; Chicago: Allyn and Bacon.

Scott, Margaret Sue. & G. Richard Tucker. 1974. Error analysis and English language strategies of Arab students. Language Learning 24. 69-97.

Selinker, Larry. 1972. Interlanguage. International Review of Applied Linguistics 10. 209-31.

Shakir, Abdullah , and Omar Shdeifat. 1996. The Translation of Collocations as an Indicator of Development of FL Competence. Al. Manarah 1. 9-27.

Sinclair, John. 1991. Corpus Concordance Collocation. Oxford University Press.

Sinclair, John. 1986. Basic computer processing of long texts. Computers in English Language Teaching and Research, eds. by Leech Geoffrey. & Candlin Christopher, , Harlow, Essex: Longman

Smith, Bernard 1987. Learner English: A teacher's guide to interference and other problems. Cambridge: Cambridge University Press.

Sharwood-Smith, Michael. 1994. Second Language Learning. London: Longman.

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. Functions of Language 2.23-55.

Stubbs, Michael. 1996. Text and Corpus Analysis. Blackwell Publishers: London.

Tribble, Chris. and Chris. Jones. 1990. Concordances in the Classroom. London: Longman

Tognini-Bonelli, Elena. 2001. Corpus Linguistics at Work. Amsterdam; Philadelphia: John Benjamins.

Taylor, Gordon. 1986. Errors and Explanations. Applied Linguistics 7. 144-166.

Ure, J. 1971. Lexical density and register differentiation. Applications of linguistics: Selected papers of the Second International Congress of Applied Linguistics, eds. by G. E. Perren & J. L. M. Trim, 443-452. Cambridge: Cambridge University Press.

Weinreich, Uriel. 1953. Languages in Contact; Findings and Problems. New York: Linguistic Circle of New York.

White, Lydia. 2000. Second Language Acquisition: From Initial to Final State. Second Language Acquisition and Linguistic Theory, ed. by John. Archibald, 130-155. Oxford: Blackwell.

Wilkins, David Arthur. 1972. Linguistics in Language Teaching. London : Arnold.

Willis, Martin. 1998. Development of Second Language Lexical Organization: A Semantic Space Approach. Unpublished Ph.D. dissertation. Temple University.

Wolfe-Quintero, Kate, Inagaki, Shunji, Kim, Hae-Young. 1998. Second Language Development in Writing: measures of fluency, accuracy & complexity. Honolulu, Hawai'i: University of Hawai'i Press.

Wright, Shirley. 2000. Attitudes of Native English-Speaking Professors Toward University ESL Students. Unpublished Ph.D. Dissertation. The University of Texas at Arlington.

Yang, Xio-ming and Huaxin Xu. 2001. Errors of Creativity: An Analysis of Lexical Errors Committed by Chinese ESL Students. University Press of America.

Zughoul, Muhammad R. 1991. Lexical Choice: Towards writing problematic word lists. International Review of Applied Linguistics 29. 45-60.

Zughoul, Muhammad R. (forthcoming). Developing Computer Based Corpora of Arabic: A Preliminary Proposal.

Zughoul, Muhammed R. and Abdul-Fattah, Hussein. 2003. Translational collocational strategies of Arab learners of English: a study in lexical semantics. Babel (Amsterdam, The Netherlands) 491. 59-81.

Worldwide Web

Tagset

http://www.comp.lancs.ac.uk/ucrel/claws/trial.html

Concordancing Glossary

http://www.nsknet.or.jp/~peterr-s/concordancing/glossary.html

Topics

http://kazuo.fc2web.com/English/TOEFL-essay.htm

BIOGRAPHICAL STATEMENT

Mousa A. Btoosh was born in Karak, Jordan, in 1972. He received his B.A. degree in English Language and Literature from Mutah University, Jordan in 1994; his M.A. degree in linguistics from Yarmouk University, Jordan in 1999. In the fall of 2001, he was admitted to the Ph.D. program of linguistics at The University of Texas at Arlington, USA.

He worked as a full-time lecturer at Al-Balqa' Applied University, Jordan from 1999 till 2000. In the spring of 2000, he started teaching as a full-time lecturer in the English Dept. at Al-Hussein Bin Talal University, Jordan. In the fall of 2003, he was employed by Brookhaven College as an adjunct faculty. While attending the University of Texas at Arlington and in recognition of his outstanding merit and accomplishment, he was elected to *Who's Who among Students at American Universities.* Upon his return to Jordan, he will be appointed as an assistant professor in the Dept. of English Language and Literature at Al-Hussein Bin Talal University, Jordan.