

**ÇUKUROVA UNIVERSITY  
THE INSTITUTE OF SOCIAL SCIENCES  
ENGLISH LANGUAGE TEACHING**

**A LEARNER CORPUS BASED STUDY ON SECOND LANGUAGE  
LEXICOLOGY OF TURKISH STUDENTS OF ENGLISH**

**Fahrettin ŞANAL**

**DOCTOR OF PHILOSOPHY**

**ADANA / 2007**

**ÇUKUROVA UNIVERSITY  
THE INSTITUTE OF SOCIAL SCIENCES  
ENGLISH LANGUAGE TEACHING**

**A LEARNER CORPUS BASED STUDY ON SECOND LANGUAGE  
LEXICOLOGY OF TURKISH STUDENTS OF ENGLISH**

**Fahrettin ŞANAL**

**Supervisor: Asst. Prof. Dr. Cem CAN**

**DOCTOR OF PHILOSOPHY**

**ADANA / 2007**

To Directorate of the Institute of Social of cukurova Universty ,  
We cartify that this dissertation is statisfactory fort he award of degree of Doctor of  
Philosofhy in the subject of English Language Teaching

Chairperson : Asst. Prof. Dr. Cem CAN  
Supervisor

(Member of Examining Committee) : Prof. Dr. A. Necmi YAŞAR

(Member of Examining Committee) : Assoc. Prof. Dr. Hatice SOFU

(Member of Examining Committee) : Asst. Prof. Dr. Abdurrahman KİLİMCİ

(Member of Examining Committee) :  
Asst. Prof. Dr. Şaziye YAMAN

I certify that this dissertation conforms to the formal standards of the Institute of Social  
Sciences

Prof. Dr. Nihat KÜÇÜKSAVAŞ  
Director of Institute

PS: The uncited usage of the reports, charts, figures, and photographs in this dissertation, whether original of quoted from  
other sources, is subject to the Law of Works of Art and thought NO: 5846.

NOT: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge, şekil ve fotoğrafların kaynak  
gösterilmeden kullanımı, 5846 sayılı Fikir ve Sanat Kanunu'ndaki hükümlere tabidir.

## ÖZET

### İNGİLİZCE ÖĞRENEN TÜRK ÖĞRENCİLERİNİN İKİNCİ DİL SÖZCÜK BİLGİSİ ÜZERİNE ARADİL BUTUNCE TABANLI BİR ÇALIŞMA

Fahrettin ŞANAL

Doktora Tezi, İngiliz Dili Eğitimi Anabilim Dalı

Danışman: Yrd. Doç. Dr. Cem CAN

Ağustos 2007, 94 sayfa

1960larda, dil bilim alanında bilgisayarda işlenebilir bütüncenin ortaya çıkması, dilbilim araştırmalarının yönünü büyük ölçüde sözdizimi ve sesbilim araştırmalarından çoğunlukla geleneksel yaklaşımlar kapsamı altında ihmal edilen bir çok alanlara çevirmiştir. Ve şu anki araştırmanın hedefi olan sözlük bilimi bu değişimden asıl faydalanan olmuştur.

Bilgisayarlı aradil bütünce tabanlı yaklaşımını kullanan bu çalışma, İngilizceyi yabancı dil olarak öğrenen Türk öğrencilerinin yazılı örneklerinden elde edilen bütüncenin (TICLE) çok yönlü bir şekilde sözlüksel açıdan bilgisayarda analizi üzerine kuruludur. Bu bütüncenin sözlüksel açıdan incelenmesi Louvain Corpus of Native English Essays (LOCNESS) veritabanından derlenen aynı büyüklükte bir bütüncenin hazırlanmasını gerektirmiştir. Çalışma, bilgisayarda karşılaştırmalı ve analitik yöntemleri kullanarak şunları hedeflemiştir: (1) aradil kullanıcısının sözcük bilgisinin zorluk derecesi ve zenginliği. (2) aradil bütüncesi ilk 200 en sık kullanılan sözcük açısından yüzde ve özellik bakımından ne dereceye kadar referans bütünceden farklılık göstermektedir? (3) aradil bütüncesinin en dikkatçekici ve basmakalıp özellikleri nedir? Aradil bütüncesi aradil kullanıcılarının anadillerinden ne dereceye kadar etkilenmektedir. Bulgular göstermektedir ki: (1) aradil bütüncesi sözcük çeşitliliği ve yoğunluğu açısından referans bütüncesinden çok daha az karmaşık bir yapıya sahiptir. (2-3) aradil kullanıcılarının ilk 200 sözcüğü belirgin bir şekilde, belirsizlik ifade eden sözcükler, anadil kullanıcısına göre bazı sözcüklerin az kullanımı, veya çok kullanımı açısından belirgin bir şekilde farklılık göstermektedir. Bu farklılık aradil kullanıcısının birinci dilinin dilbilgisel ve anlatım özelliklerinden kaynaklandığı tespit edilmiştir.

**Anahtar sözcük:** Bütüncedilbilim, aradil, sözcüksel çözümleme, aradil bütüncesi, azkullanım/çok kullanım

**ABSTRACT****A LEARNER CORPUS BASED STUDY ON SECOND LANGUAGE LEXICOLOGY  
OF TURKISH STUDENTS OF ENGLISH****Fahrettin ŞANAL****PhD Dissertation, English Language Teaching Department****Supervisor: Asst. Prof. Dr. Cem CAN****August 2007, 94 pages**

The emergence of machine-readable corpora in the linguistics field in the 1960s shifted the direction of a considerable body of linguistic research from syntax and phonology, which was by then the focus of linguistic research to a number of domains that were mostly neglected under the umbrella of traditional approaches. And lexicology, which is the target of this research, was a major beneficiary of that shift.

By utilising a computer learner corpus-based approach, this study addresses multidimensional lexical aspects of a machine-readable corpus of the writing of Turkish students of English as a foreign language (TICLE). Lexical investigation of this corpus, required a similar sized authentic corpus, which was compiled from Louvain Corpus of Native English Essays (LOCNESS). Employing the computerized contrastive and analytical methods, 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? (3) what are the most salient and stereotype features of the learner corpus? And how far the learner corpus influenced by the learners' L1? Findings show that: (1) the learner corpus is much less complex in terms of lexical diversity and density than the reference corpus. (2-3) Learners' top 200 tokens are markedly characterized by vague lexica, underuse and overuse of some lexica, resulting from the influence of the linguistic and rhetorical features of learners' L1.

**Key words:** Corpus linguistics, interlanguage, lexical analysis, learner corpus, underuse/overuse

## ACKNOWLEDGEMENTS

Working on this dissertation was not an easy task. Nevertheless, the process was a great pleasure thanks to intellectually stimulating discussions as well as continuous support from a considerable number of people. This work would not have been completed without their help.

First of all, I would like to express my sincere gratitude to my advisor and dissertation committee chair, Asst. Prof. Dr. Cem Can. He patiently spent a considerable amount of time discussing the study with me throughout the process, and those discussions provided me with new insights. He always raised very interesting and significant theoretical questions about second language acquisition and corpus linguistics, and the exploration of those questions deepened my knowledge. He lent me invaluable support not only as an academic scholar who is knowledgeable about the field but also as a friendly and approachable professor. I learned so many things from him, including how a scholar should be and how to explain content to students.

I am also grateful to Asst.Prof.Dr. Abdurrahman Kilimci, who gave me feedback on the statistical analyses of the dissertation as well as on the research about the nature of corpus linguistics. Consultations with him on statistical analyses helped me understand what I should be careful with when interpreting the results. Whenever I talked to him about my studies, including my dissertation, he asked me a question: "So what?" This is sometimes a tough question, but it is a significantly important question to consider because it relates to the meaningfulness of the research. I am sure that this question will always be on my mind when I conduct research in future.

I would also like to thank Assoc.Prof.Dr. Hatice Sofu. She spent a few hours whenever I had an appointment with her, commenting on details of my study. She always pointed out various issues that I had been concerned with on my research, followed by helpful solutions to the problems. She paraphrased her explanations when I could not capture the main idea, and that way she ensured that I understood the concepts. Her insightful comments helped me investigate the topic from various angles.

Lastly, I would like to express my deepest gratitude to my wife Hacer, my sons Mesut and Uğur, my daughter Faika who continuously supported me in every single aspect of my graduate study at Çukurova University, Adana. Without their encouragement, I would not have reached this far. I am grateful to them from the bottom of my heart for loving me and supporting me throughout my life.

## TABLE OF CONTENTS

TÜRKÇE ÖZET.....	i
ABSTRACT.....	ii
ACKNOWLEDGEMENT.....	iii
TABLE OF CONTENTS.....	iv
LIST OF TABLES.....	vii
LIST OF FIGURES.....	viii

### CHAPTER I

INTRODUCTION.....	1
1.1. Introduction.....	1
1.2. Objectives.....	3
1.3. Research Questions.....	4
1.4. Significance of the Study.....	4
1.5. Definition of Terms.....	5

### CHAPTER II

REVIEW OF RELATED LITERATURE.....	6
2.1. Introduction.....	6
2.2. Perspectives on Language Learning and Lexicology.....	6
2.2.1. Introduction.....	6
2.2.2. From the Behaviorists' Perspective.....	6
2.2.3. From the Mentalists' Perspective.....	8
2.2.4. From the Autonomous Discipline Perspective.....	10
2.2.4.1. Error Analysis.....	11
2.2.4.2. Interlanguage Theory.....	12
2.3 Lexicology.....	14
2.3.1. Recognition and Development.....	13
2.3.1.1 Lexical Choice.....	16
2.3.2. Lexical Competence.....	18
2.4. Corpus Linguistics.....	19
2.4.1. Attitude and Use.....	19
2.4.2. Applications of Corpus Linguistics in SLA Research:	
Learner Corpora.....	24

2.4.3. Corpus Compiling.....	27
2.4.3.1. Representativeness.....	27
2.4.4. Corpus Annotation.....	28
2.4.4.1. Part of speech annotation.....	28
2.4.4.2. Lemmatization.....	29
2.4.4.3. Syntactic annotation or parsing.....	29
2.4.4.4. Semantic and pragmatic tagging.....	29
2.4.4.5. Discoursal and text linguistic annotations.....	30
2.4.4.6. Phonetic Transcription.....	30
2.4.4.7. Prosodic annotations.....	30
2.4.4.8. Problem-oriented tagging.....	30

### CHAPTER III

METHODOLOGY.....	32
3.1. Introduction.....	32
3.2. Data Gathering Procedures.....	32
3.2.1. Learner Corpus.....	32
3.2.2. Referance Corpus.....	32
3.3. Quantitative Analysis.....	32
3.4. Data Processing and Analysis Procedures.....	33
3.5. Software.....	36

### CHAPTER IV

#### LEXICAL COMPLEXITY AND TEXT-PROFILING

RESULTS AND DISCUSSION.....	38
4.1. Introduction.....	38
4.2. Results Related to Research Question (1) .....	38
4.2.1. Lexical Diversity .....	38
4.2.2. Lexical Density.....	40
4.3. Results Related to Research Question (2) .....	45
4.4. Results Related to Research Question (3).....	66
4.4.1. Word Categories.....	66
4.4.2. Overproduction and Verbosity.....	71
4.4.3. Underproduction.....	73



## CHAPTER V

CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS.....	77
5.1. Introduction.....	77
5.2. Summary.....	77
5.3. Limitations of the Study.....	78
5.4. Future Research.....	79
REFERENCES.....	81
LIST OF APPENDIX.....	90
CURRICULUM VITAE .....	95

## LIST OF TABLES

Table 2.1. Native speakers' judgement of errors type.....	17
Table 4.1. Mean of lexical density and standard deviation in learner and reference corpora.....	41
Table 4.2. Reduced word category tag list .....	67
Table 4.3. Learners' use of lexical categories in comparison with the NSs .....	68
Table 4.4. Analysis of features of writer/reader visibility Adapted from Petch-Tyson (1998:112) .....	70

## LIST OF FIGURES

Figure 4.1 Type-token ratio in the learner corpus.....	39
Figure 4.2 Type-token ratio in the reference corpus.....	39
Figure 4.3 Overall frequency of content words in learner and reference corpora.....	42
Figure 4.4 Percentage of content words in the learner corpus.....	43
Figure 4.5 Percentage of content words in the reference corpus.....	43
Figure 4.6 Top 100 frequent words in the learner corpus.....	47
Figure 4.7 Top 100 frequent words in the reference corpus.....	50
Figure 4.8 Proportion of the learner and reference corpora in the total number of the content words in the top 100 frequent tokens.....	54
Figure 4.9 Percentage of the top 100 frequent tokens in the learner and reference corpora.....	55
Figure 4.10 Frequencies of the content and grammatical words in the top 100 frequent tokens in learner and reference corpora.....	56
Figure 4.11 Ratio of the content words frequency to that of the grammatical words in the top 100 frequent tokens in the learner corpus.....	56
Figure 4.12 Percentage of the frequency of the content words to that of the grammatical words in the top 100 frequent tokens in the reference corpus.....	57
Figure 4.13 Percentage of the top 10 frequent tokens learner and reference corpora....	58
Figure 4.14 The second 100 frequent words in the learner corpus.....	59
Figure 4.15 The second 100 frequent words in the reference corpus.....	62
Figure 4.16 Number of content and grammatical words in the top 100 frequent token in the learner and reference corpora.....	65
Figure 4.17 Word category in learner and reference corpora.....	67
Figure 4.18 Examples of the use of and sentence initially.....	71
Figure 4.19 Samples of overproduction.....	72
Figure 4.20 Hedges in learner and reference corpora.....	74
Figure 4.21 Sentence length in the learner corpus. ....	75
Figure 4.22 Sentence length in the reference corpus. ....	76

## 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*).

Owing to the various functions that general corpora serve in linguistic research - e.g., providing linguistic 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, it is clear 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 (ICLE)* and its subcorpus including the *Turkish International Corpus of Learner English (TICLE)* (Grange et al., 2002).

Previous attempts to compile an English learner corpus of Turkish students' writings prior to TICLE are next to nonexistent. The primary goal behind compiling this corpus is to investigate the learners' interlanguage represented in this written corpus.

Being the first machine-readable corpus combined from the interlanguage of Turkish 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, two main objectives have been set for consideration.

- To compare and contrast learners' lexical complexity and richness with 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 lexical 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.

Achieving such objectives will make it possible for learners, teachers and researchers to get accurate 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 Turkish 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 three 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 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?

### **1.4. Significance of the Study**

The significance of this study stems from being one of the first attempts to electronically analyze a representative computerized corpus of the written interlanguage of Turkish students of English. First, the study delineates learners' lexical complexity and richness in comparison with the reference (authentic) corpus. Second, it may provides curriculum designers with areas of weaknesses in student writing and thus, enables them to make the revisions. Third, it uncovers the differences between the

subjects' output and the English norm. Finally, it may inspire other researchers to conduct studies on other aspects of interlanguage.

### 1.5. Definition of Terms

- 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:

$$\frac{\text{the total number of content words} \times 100}{\text{the total number of all tokens in the given corpus}}$$

Lexical diversity: a measure of the spread or richness of the vocabulary in a text calculated according to the following formula:

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

- Part-of-speech (POS) tagging: The process of assigning lexical categories (that is, part-of-speech tags) to words in linguistic data.
- 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 noninterference 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*, *habitformation* 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 transferI partial (or no) access
- (ii) No transferI 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 CAR 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 fa-

tigue. 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,

#### **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 indicate "the separateness of a second language learner's system, that has a structurally intermediate status between the native and target languages" (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 primarily 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 Bruton 1974, Ellis and Roberts 1987, among others).

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 naive 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 collocability 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.

### **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).

To sum up, the preceding sections have substantiated the contention that inter-language 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 Turkish, have seen confined interest in this respect. This explains the rapid growth of English corpora compared with Turkish 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 collocations) that have remained unattainable by traditional approaches. For example, due to the scarcity of corpora for Modern Standard Turkish, one can hardly provide reliable answers to questions related to word order patterns, dialectal differences, collocations or percentages of loan words.

Passing to matters more closely related to internalized linguistics, Chafe (1992: 7995) 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 (1991:2), 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.

40 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 various 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 detriment 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 nonstandard 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 Turkish 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, 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. Data Gathering Procedures

The database of this study consists of two corpora; Learner Corpus (LC) TICLE and the reference corpus (RC) LOCNESS.

##### 3.2.1. Learner Corpus

The learner corpus comes from the International Corpus of Learner English (ICLE). The Turkish component of the ICLE corpus (TICLE) contains essays written by Turkish university students of English. All the essays are expository and argumentative in character and the selected sample for the present comparative study total 176,171 words.

##### 3.2.2. Reference Corpus

The control corpus of similar writing is taken from the Louvain Corpus of Native English Essays (LOCNESS) database. This native speaker corpus consists of argumentative essays written by American University students and contains 175,612 words.

### 3.3. Quantitative Analysis

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, corpus). The findings of this study are compared and contrasted with reference corpus 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.

### 3.4. 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 is 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, 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 was 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:

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

## **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:

$$\frac{\text{the total number of content words} \times 100}{\text{the total number of all tokens in the given corpus}}$$

### **(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. Establishing an automatic frequency count of the reference as well as the learner corpora.
2. Comparing and contrasting the frequency count findings in the learner corpus with those of the reference corpus.

3. Examining, via WordSmith tools, lexical diversity in the learner as well as the reference corpora at both corpus level and individual level.
4. 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.
5. Providing part of speech annotation for the essay-writing corpus as well as the reference corpus.
6. Examining lexical density in learner and reference corpora.

### 3.5. Software

WordSmith, an integrated suite of software programs, was utilised for the lexical analysis in this study. Inasmuch as 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.

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 concordancing 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).

To sum up, this chapter has delineated the methodological procedures employed in the study.



## 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 textprofiling. 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.

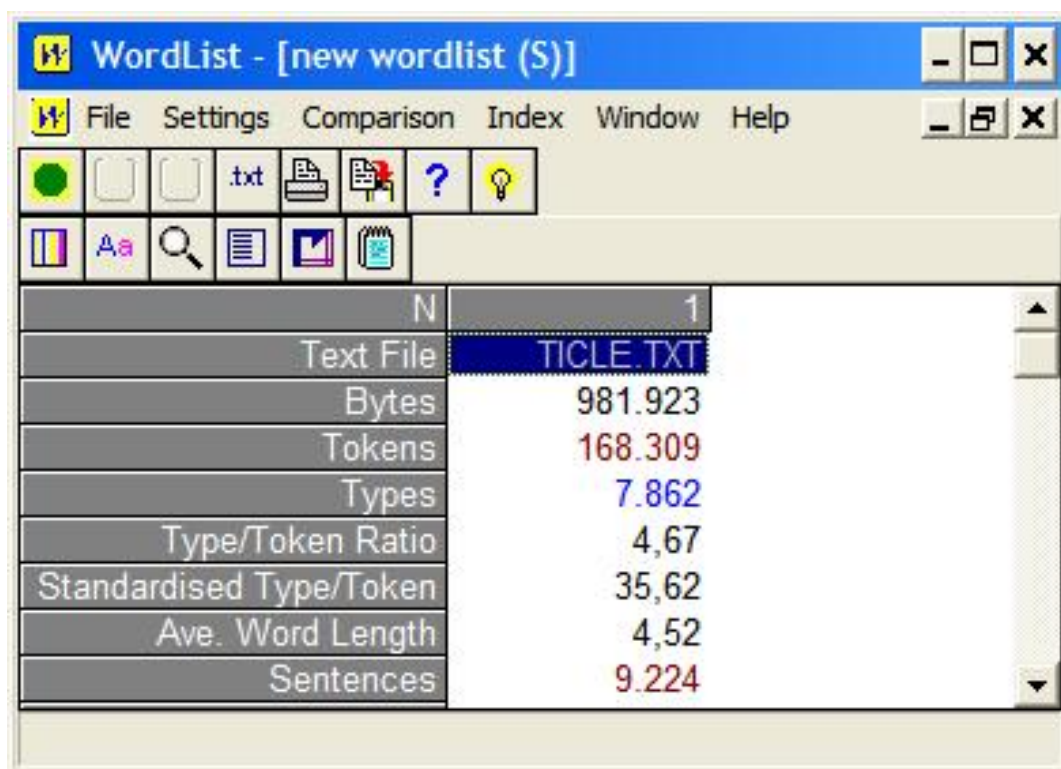
#### 4.2. Results Related to Research Question (1)

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

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

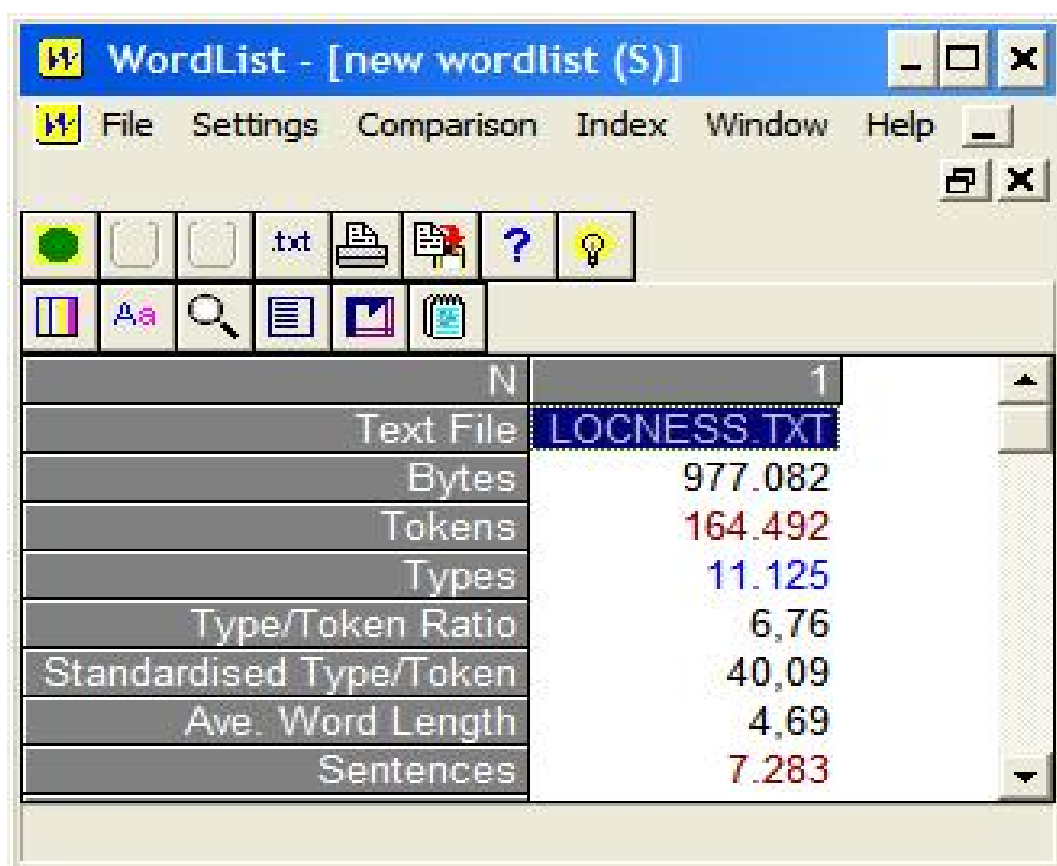
A critical factor adversely affecting lexical diversity is corpus size/length. So, in order to avoid its converse role when the analysis is carried out on individual essays, which vary in their length, this measure was carried out on a full corpus basis (equal basis). Figures (4.1) and (4.2) present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



The screenshot shows the WordList application window titled "WordList - [new wordlist (S)]". The menu bar includes File, Settings, Comparison, Index, Window, and Help. Below the menu is a toolbar with icons for file operations and analysis. The main display area shows a table of statistics for the file "TICLE.TXT".

	N
Text File	TICLE.TXT
Bytes	981.923
Tokens	168.309
Types	7.862
Type/Token Ratio	4,67
Standardised Type/Token	35,62
Ave. Word Length	4,52
Sentences	9.224

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



The screenshot shows the WordList application window titled "WordList - [new wordlist (S)]". The menu bar includes File, Settings, Comparison, Index, Window, and Help. Below the menu is a toolbar with icons for file operations and analysis. The main display area shows a table of statistics for the file "LOCNESS.TXT".

	N
Text File	LOCNESS.TXT
Bytes	977.082
Tokens	164.492
Types	11.125
Type/Token Ratio	6,76
Standardised Type/Token	40,09
Ave. Word Length	4,69
Sentences	7.283

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 (4.67 vs. 6.76), 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 (11,125 vs. 7,862). While it was not unexpected for the type-token ratio in the reference corpus to outnumber the learner counterpart, the marked diversity, which favored the reference corpus, goes far beyond expectations. 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.

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 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 (4.1). 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 4.1. Mean of lexical density and standard deviation  
in learner and reference corpora

	L.C.	R.C.	Difference <sup>''</sup>
Mean	49.20	52.10	2.74**
SD	5.30	5.45	

<sup>''\*\*</sup>The difference between the two corpora is significant at the  $Q = 0.05$  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%).

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). Nouns Verbs Adjectives Adverbs

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 (4.1). 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.

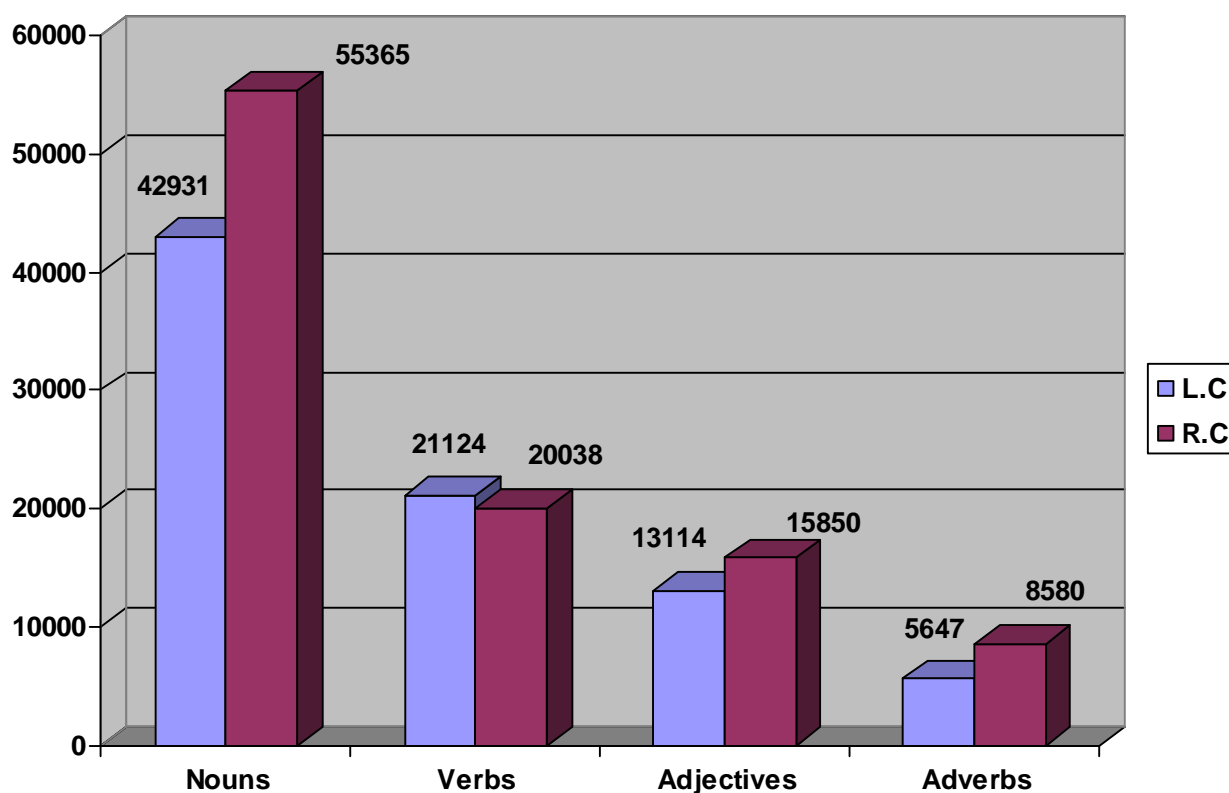


Figure 4.3. Overall frequency of content words in learner and reference corpora.

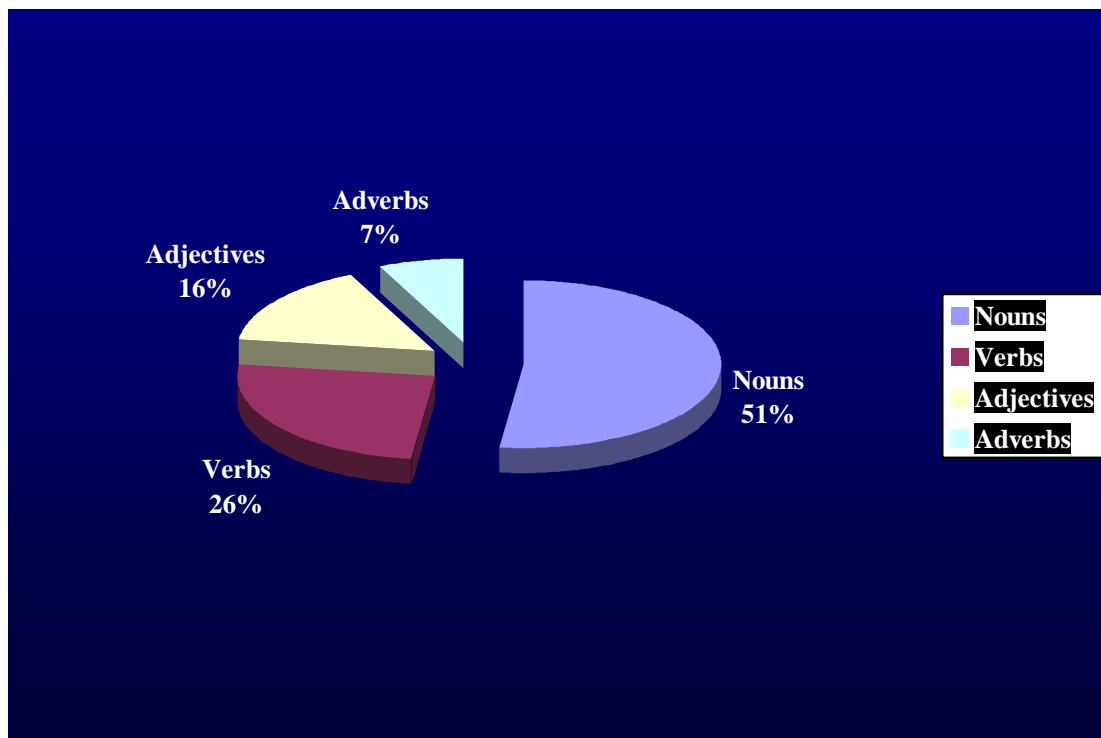


Figure 4.4. Percentage of content words in the learner corpus.

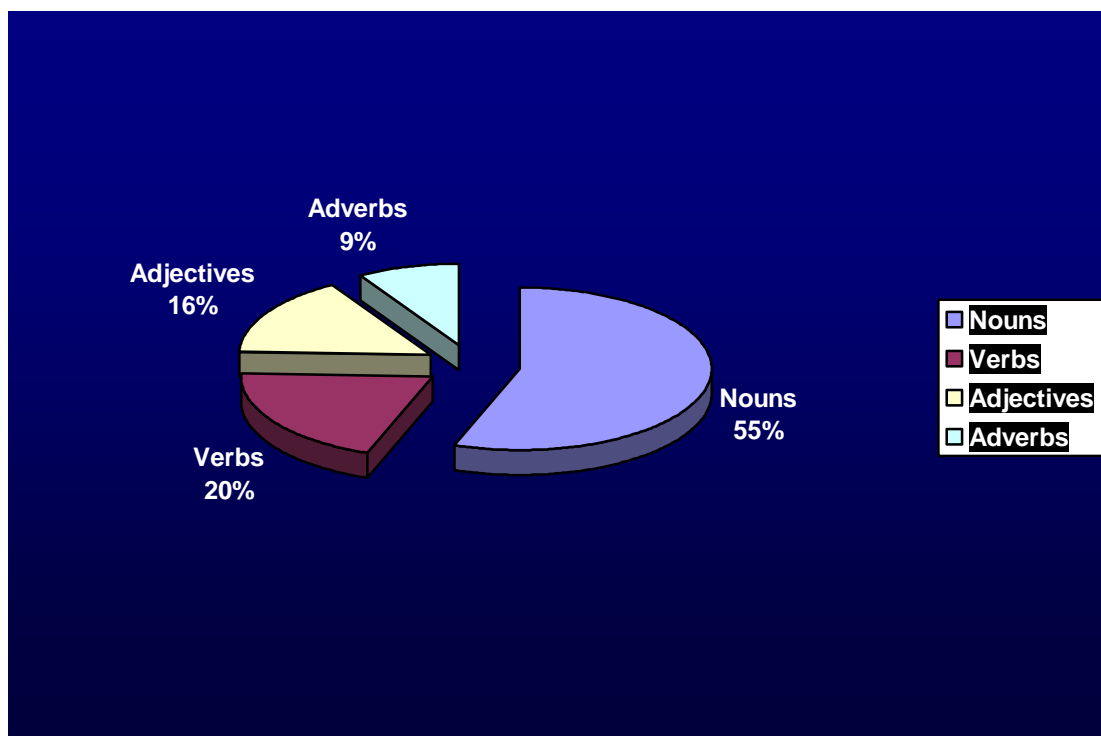


Figure 4.5. Percentage of content words in the reference corpus.

Figures (4.3, 4.4 and 4.5) 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 writings 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 (51%) and (55%) 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.

A striking diversity between the two corpora is clearly seen in the number of adverbs (5647 vs. 8580), 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 2002:12122). 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 how can learners' lexical stereotypes be captured through word frequency?

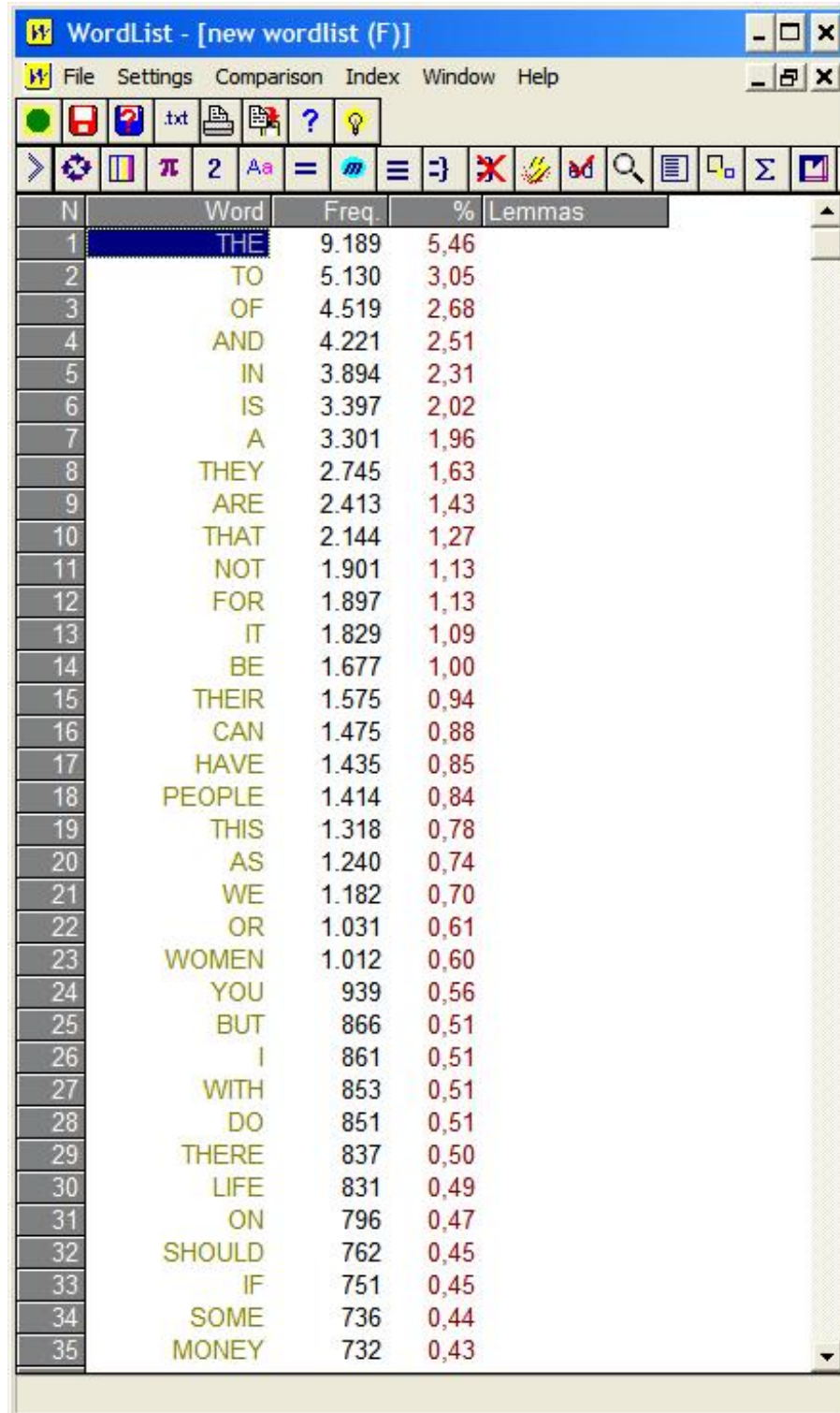
There is a strong consensus among corpus linguists on the importance of word frequency lists in corpus analysis (McEnery & Wilson 1996, Kennedy 1998, among others). Drawing on its multifunctional uses, creating a word frequency list is a fruitful and productive technique, in the sense that it might be 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, 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 (4.6) and (4.7) present the top 100 frequent tokens (in a version of the list arranged in descending frequency order) in the learner and reference corpora respectively.

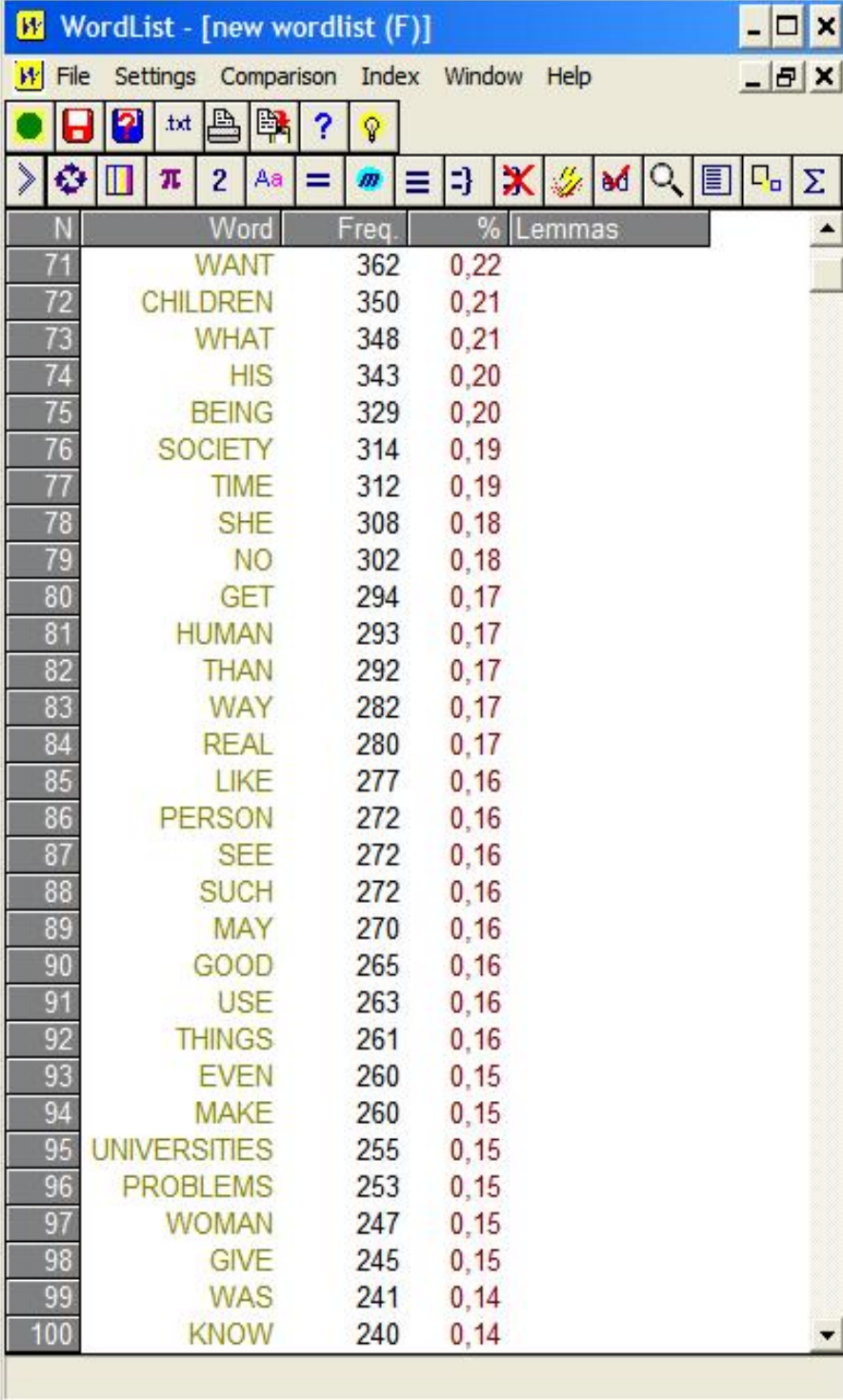


N	Word	Freq.	%	Lemmas
1	THE	9.189	5,46	
2	TO	5.130	3,05	
3	OF	4.519	2,68	
4	AND	4.221	2,51	
5	IN	3.894	2,31	
6	IS	3.397	2,02	
7	A	3.301	1,96	
8	THEY	2.745	1,63	
9	ARE	2.413	1,43	
10	THAT	2.144	1,27	
11	NOT	1.901	1,13	
12	FOR	1.897	1,13	
13	IT	1.829	1,09	
14	BE	1.677	1,00	
15	THEIR	1.575	0,94	
16	CAN	1.475	0,88	
17	HAVE	1.435	0,85	
18	PEOPLE	1.414	0,84	
19	THIS	1.318	0,78	
20	AS	1.240	0,74	
21	WE	1.182	0,70	
22	OR	1.031	0,61	
23	WOMEN	1.012	0,60	
24	YOU	939	0,56	
25	BUT	866	0,51	
26	I	861	0,51	
27	WITH	853	0,51	
28	DO	851	0,51	
29	THERE	837	0,50	
30	LIFE	831	0,49	
31	ON	796	0,47	
32	SHOULD	762	0,45	
33	IF	751	0,45	
34	SOME	736	0,44	
35	MONEY	732	0,43	

Figure 4.6. Top 100 frequent words in the learner corpus.

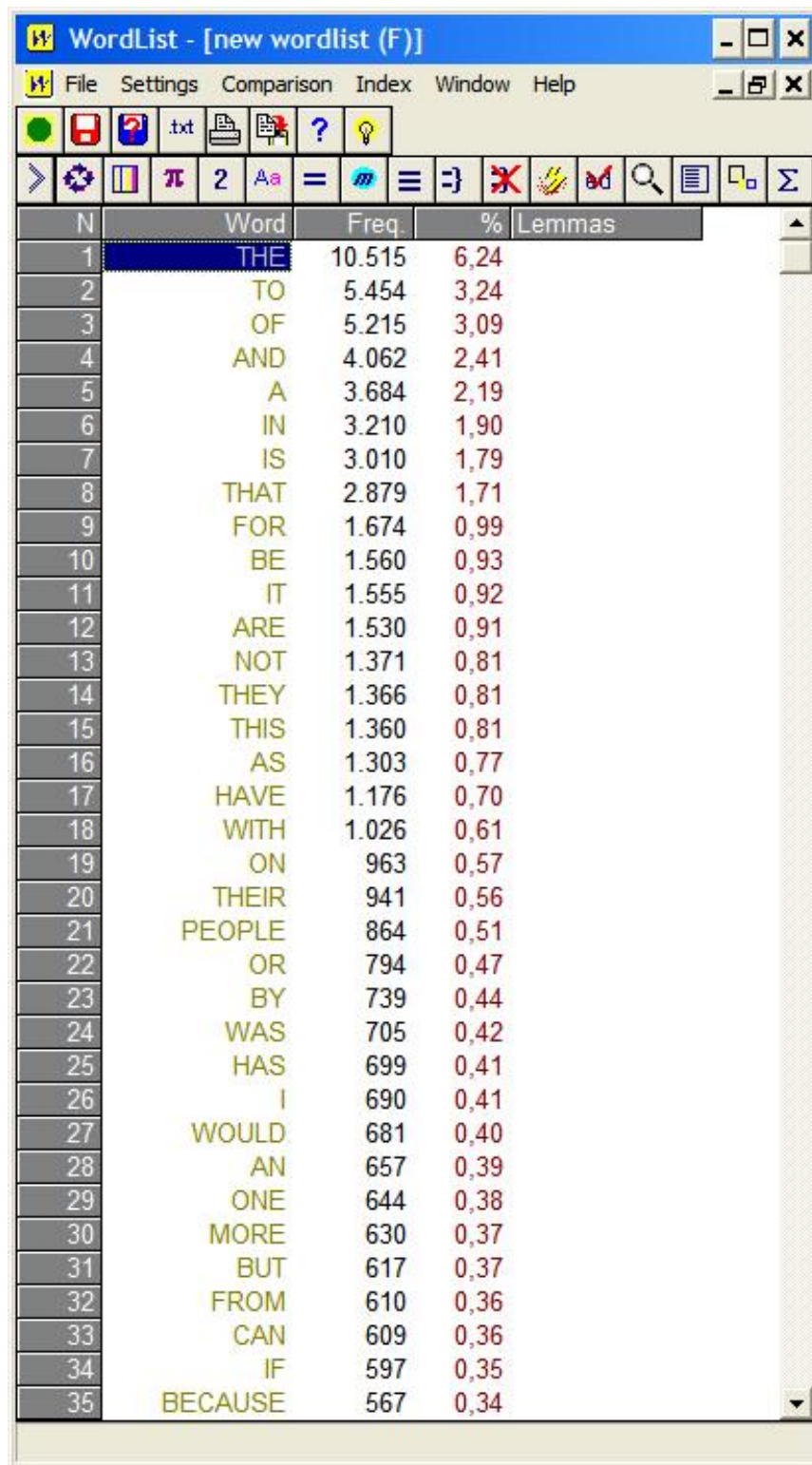
N	Word	Freq.	%	Lemmas
36	STUDENTS	725	0,43	
37	ALL	702	0,42	
38	THESE	699	0,42	
39	SO	685	0,41	
40	BY	682	0,41	
41	MEN	675	0,40	
42	BECAUSE	652	0,39	
43	THEM	650	0,39	
44	WILL	650	0,39	
45	FROM	640	0,38	
46	OTHER	617	0,37	
47	S	598	0,36	
48	WHO	592	0,35	
49	T	587	0,35	
50	MORE	575	0,34	
51	AT	553	0,33	
52	WHEN	522	0,31	
53	ALSO	521	0,31	
54	OUR	516	0,31	
55	MANY	514	0,31	
56	AN	505	0,30	
57	THINK	503	0,30	
58	ONE	485	0,29	
59	WORLD	478	0,28	
60	HAS	477	0,28	
61	ABOUT	457	0,27	
62	MOST	455	0,27	
63	WHICH	451	0,27	
64	HE	439	0,26	
65	HER	397	0,24	
66	EDUCATION	391	0,23	
67	UNIVERSITY	382	0,23	
68	ONLY	368	0,22	
69	VERY	368	0,22	
70	IMPORTANT	364	0,22	

Continuid Figure 4.6



N	Word	Freq.	%	Lemmas
71	WANT	362	0,22	
72	CHILDREN	350	0,21	
73	WHAT	348	0,21	
74	HIS	343	0,20	
75	BEING	329	0,20	
76	SOCIETY	314	0,19	
77	TIME	312	0,19	
78	SHE	308	0,18	
79	NO	302	0,18	
80	GET	294	0,17	
81	HUMAN	293	0,17	
82	THAN	292	0,17	
83	WAY	282	0,17	
84	REAL	280	0,17	
85	LIKE	277	0,16	
86	PERSON	272	0,16	
87	SEE	272	0,16	
88	SUCH	272	0,16	
89	MAY	270	0,16	
90	GOOD	265	0,16	
91	USE	263	0,16	
92	THINGS	261	0,16	
93	EVEN	260	0,15	
94	MAKE	260	0,15	
95	UNIVERSITIES	255	0,15	
96	PROBLEMS	253	0,15	
97	WOMAN	247	0,15	
98	GIVE	245	0,15	
99	WAS	241	0,14	
100	KNOW	240	0,14	

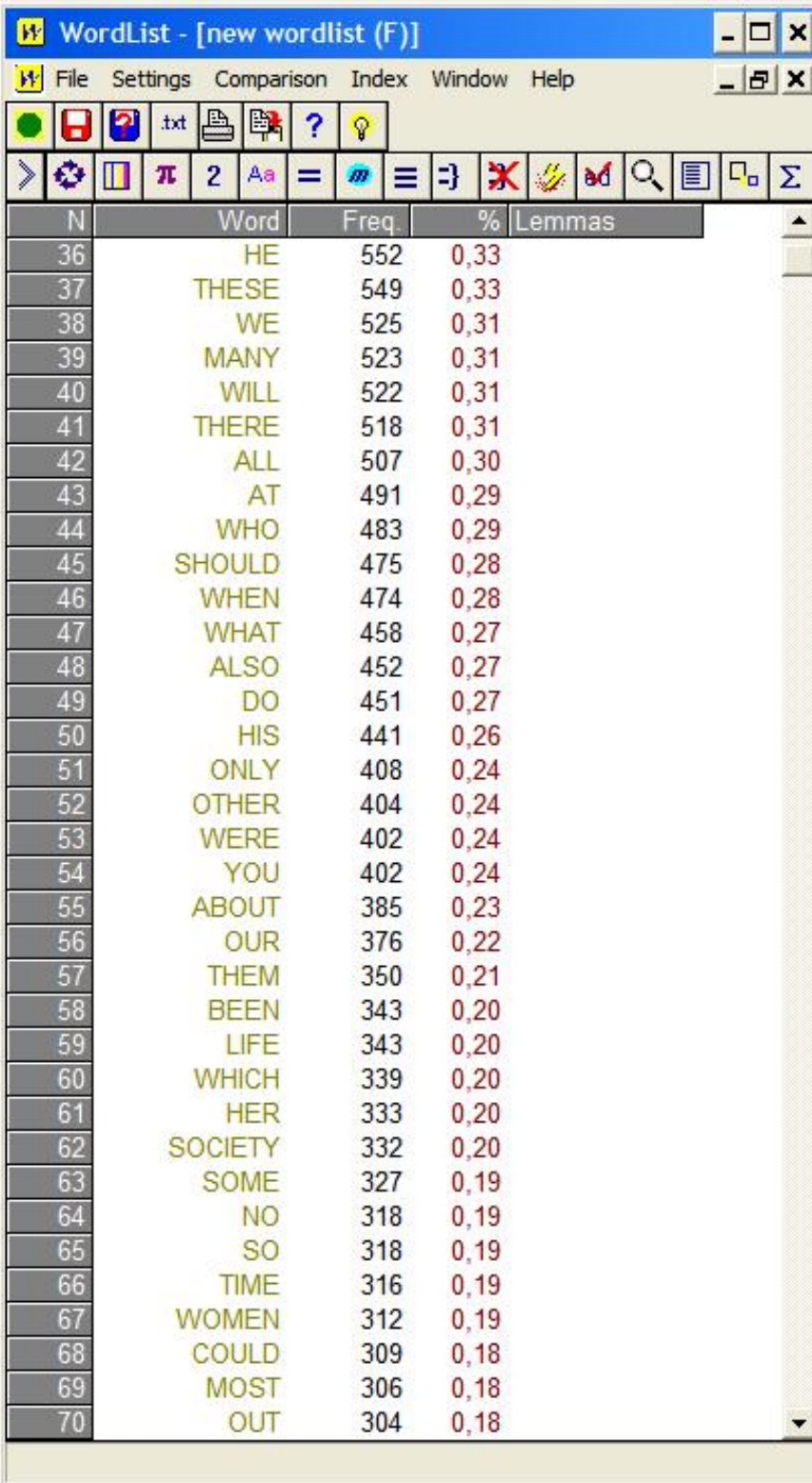
Continuid Figure 4.6



N	Word	Freq.	%	Lemmas
1	THE	10.515	6,24	
2	TO	5.454	3,24	
3	OF	5.215	3,09	
4	AND	4.062	2,41	
5	A	3.684	2,19	
6	IN	3.210	1,90	
7	IS	3.010	1,79	
8	THAT	2.879	1,71	
9	FOR	1.674	0,99	
10	BE	1.560	0,93	
11	IT	1.555	0,92	
12	ARE	1.530	0,91	
13	NOT	1.371	0,81	
14	THEY	1.366	0,81	
15	THIS	1.360	0,81	
16	AS	1.303	0,77	
17	HAVE	1.176	0,70	
18	WITH	1.026	0,61	
19	ON	963	0,57	
20	THEIR	941	0,56	
21	PEOPLE	864	0,51	
22	OR	794	0,47	
23	BY	739	0,44	
24	WAS	705	0,42	
25	HAS	699	0,41	
26	I	690	0,41	
27	WOULD	681	0,40	
28	AN	657	0,39	
29	ONE	644	0,38	
30	MORE	630	0,37	
31	BUT	617	0,37	
32	FROM	610	0,36	
33	CAN	609	0,36	
34	IF	597	0,35	
35	BECAUSE	567	0,34	

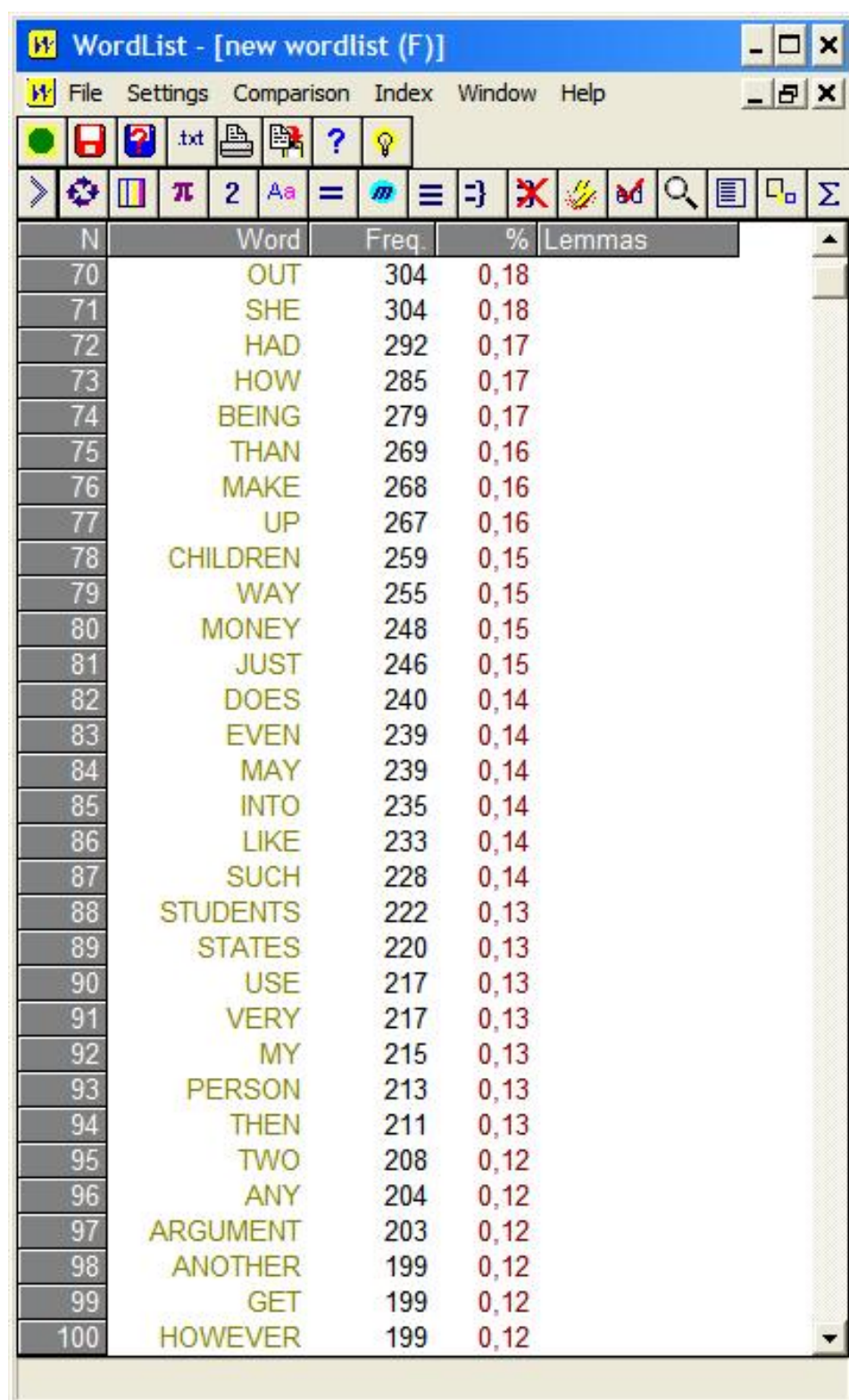
Figure 4.7. Top 100 frequent words in the reference corpus.





N	Word	Freq.	%	Lemmas
36	HE	552	0,33	
37	THESE	549	0,33	
38	WE	525	0,31	
39	MANY	523	0,31	
40	WILL	522	0,31	
41	THERE	518	0,31	
42	ALL	507	0,30	
43	AT	491	0,29	
44	WHO	483	0,29	
45	SHOULD	475	0,28	
46	WHEN	474	0,28	
47	WHAT	458	0,27	
48	ALSO	452	0,27	
49	DO	451	0,27	
50	HIS	441	0,26	
51	ONLY	408	0,24	
52	OTHER	404	0,24	
53	WERE	402	0,24	
54	YOU	402	0,24	
55	ABOUT	385	0,23	
56	OUR	376	0,22	
57	THEM	350	0,21	
58	BEEN	343	0,20	
59	LIFE	343	0,20	
60	WHICH	339	0,20	
61	HER	333	0,20	
62	SOCIETY	332	0,20	
63	SOME	327	0,19	
64	NO	318	0,19	
65	SO	318	0,19	
66	TIME	316	0,19	
67	WOMEN	312	0,19	
68	COULD	309	0,18	
69	MOST	306	0,18	
70	OUT	304	0,18	

Continuid Figure 4.7



N	Word	Freq	%	Lemmas
70	OUT	304	0,18	
71	SHE	304	0,18	
72	HAD	292	0,17	
73	HOW	285	0,17	
74	BEING	279	0,17	
75	THAN	269	0,16	
76	MAKE	268	0,16	
77	UP	267	0,16	
78	CHILDREN	259	0,15	
79	WAY	255	0,15	
80	MONEY	248	0,15	
81	JUST	246	0,15	
82	DOES	240	0,14	
83	EVEN	239	0,14	
84	MAY	239	0,14	
85	INTO	235	0,14	
86	LIKE	233	0,14	
87	SUCH	228	0,14	
88	STUDENTS	222	0,13	
89	STATES	220	0,13	
90	USE	217	0,13	
91	VERY	217	0,13	
92	MY	215	0,13	
93	PERSON	213	0,13	
94	THEN	211	0,13	
95	TWO	208	0,12	
96	ANY	204	0,12	
97	ARGUMENT	203	0,12	
98	ANOTHER	199	0,12	
99	GET	199	0,12	
100	HOWEVER	199	0,12	

Continuid Figure 4.7

Along with the tokens, three notable points immediately emerge from Figures (4.6) and (4.7). 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 34 tokens were content words. Secondly, due to the excessive use of some vague nouns and generic adjectives and topic related words (e.g., *people, person things, life, important, women, money*), which are attributable to lexical developmental stages, the learner corpus leads the reference corpus by 24% in terms of the content words in the top 100 frequent tokens. Thirdly, 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 (4.6) and (4.7): 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 to 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 most 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 75% 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 (4.8).

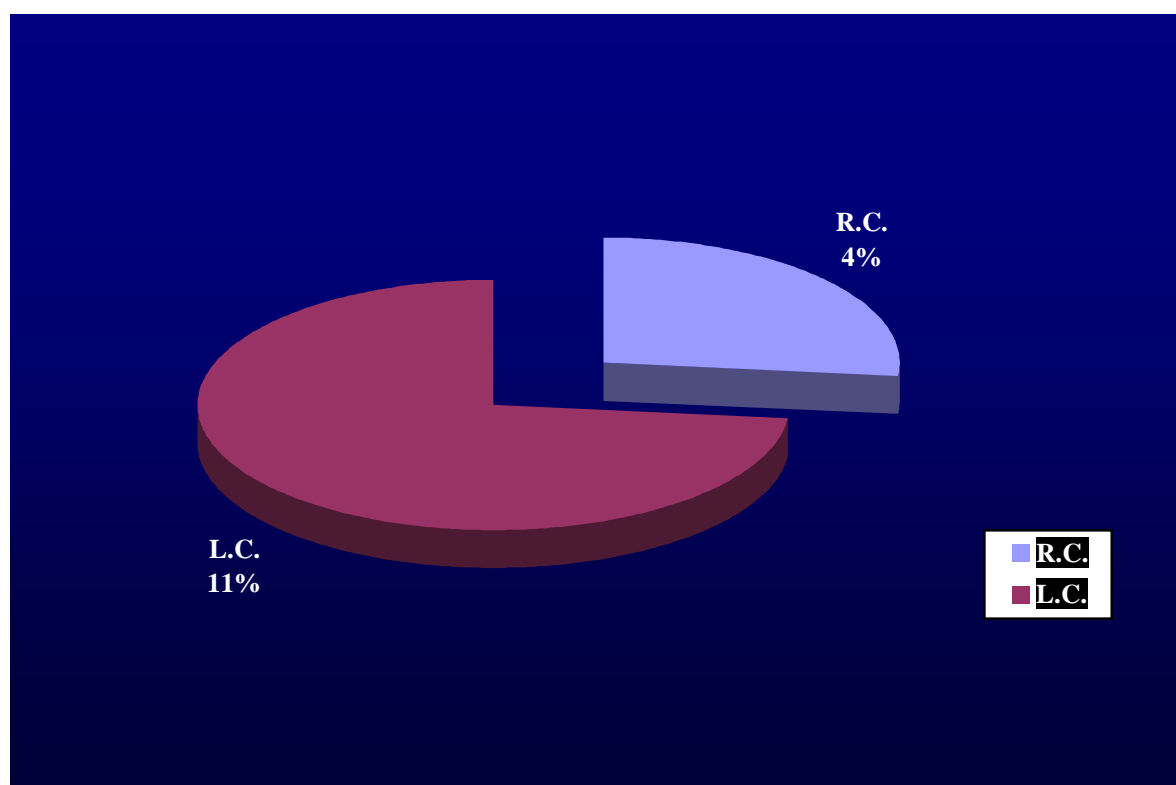


Figure 4.8. Proportion of the learner and reference corpora in the total number of the content words in the top 100 frequent tokens.

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 (4.9) 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).

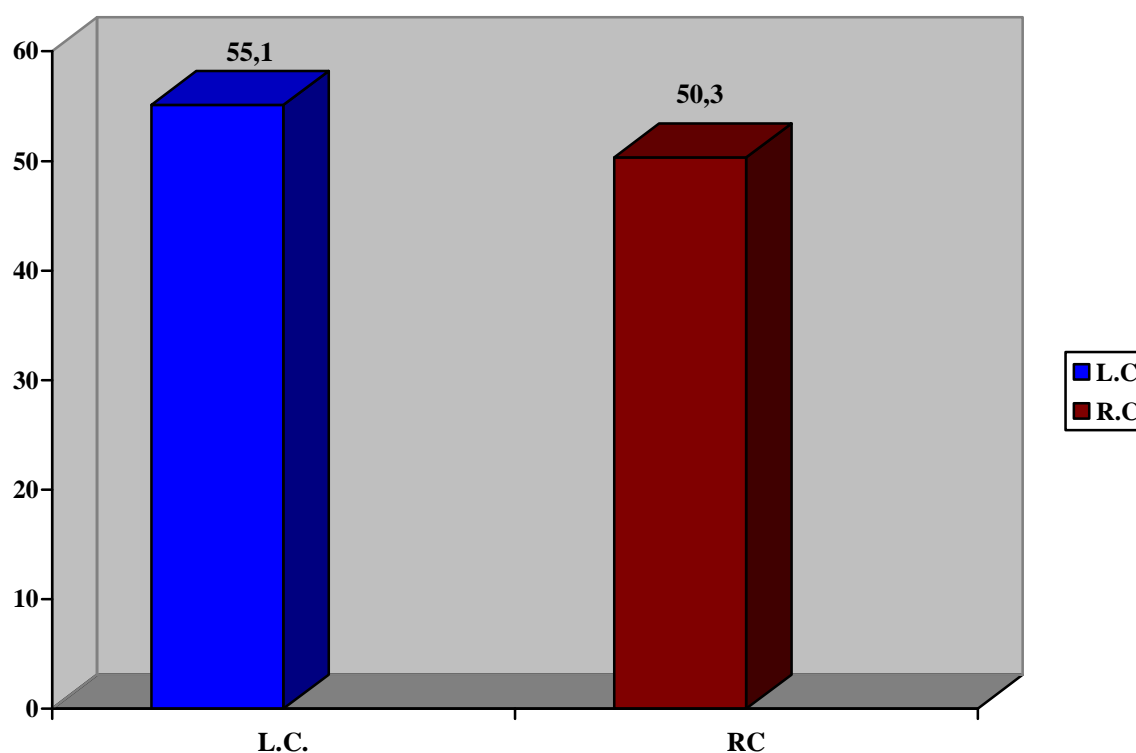


Figure 4.9. Percentage of the top 100 frequent tokens in the learner and reference corpora.

The percentage of the top 100 frequent tokens (shown in Figure 4.9), 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 11, 439 tokens. As illustrated in Figures (4.10, 4.11 and 4.12), 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%.

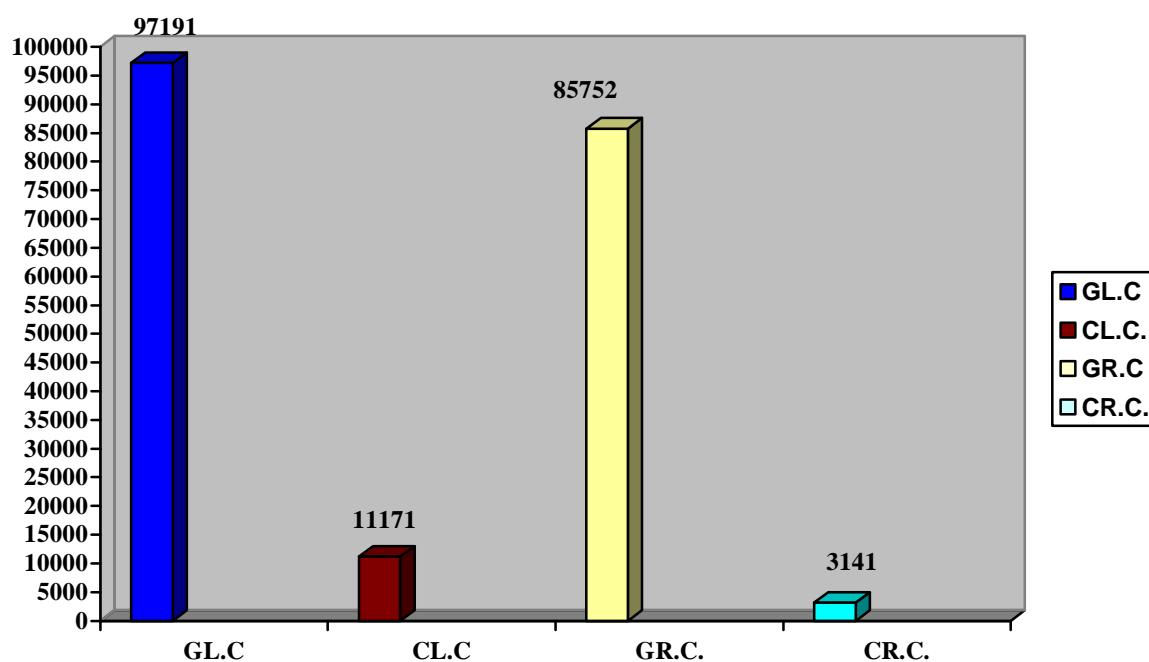


Figure 4.10. Frequencies of the content and grammatical words in the top 100 frequent tokens in learner and reference corpora.

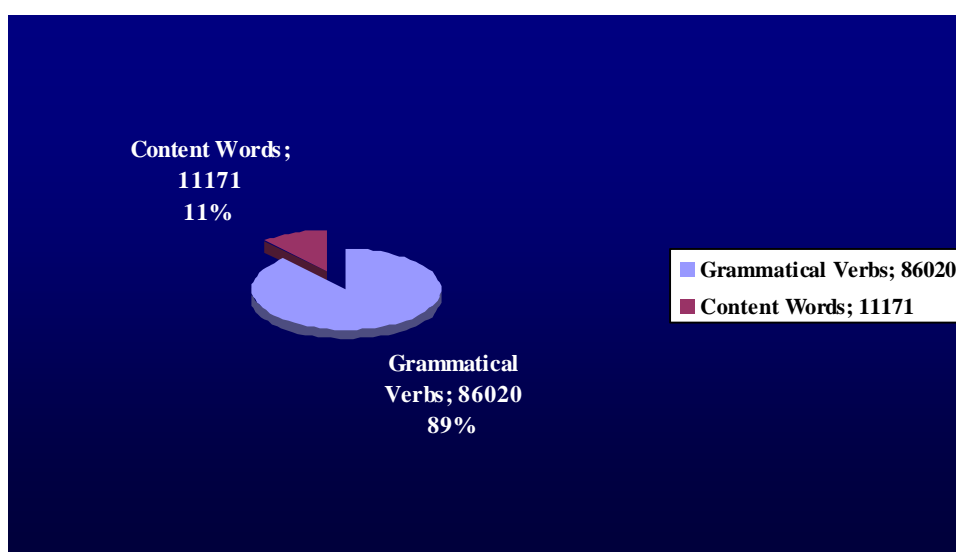


Figure 4.11. Ratio of the content words frequency to that of the grammatical words in the top 100 frequent tokens in the learner corpus.

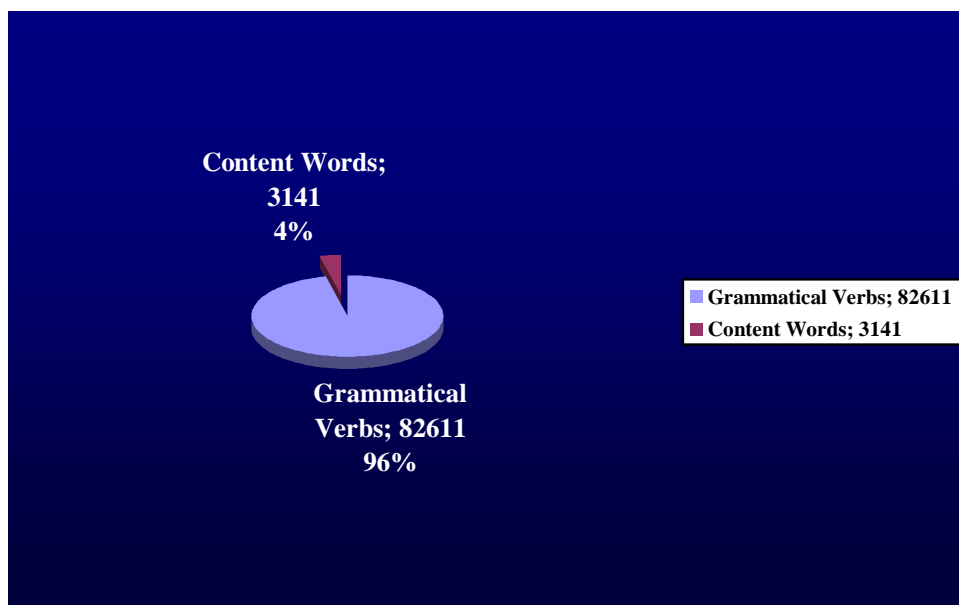


Figure 4.12. Percentage of the frequency of the content words to that of the grammatical words in the top 100 frequent tokens in the reference corpus.

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 NBs' 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 (4.13) makes it clear that the present learner and reference corpora, the *Quebec Learner Corpus* (QLC), and the *Brown Corpus* are alike in terms of the percentage of the top 10 frequent tokens (relevant to the total number of all tokens in the corpus), though Be (1,000,000 words) is almost seven times as big as that of the present learner and the reference corpora combined.

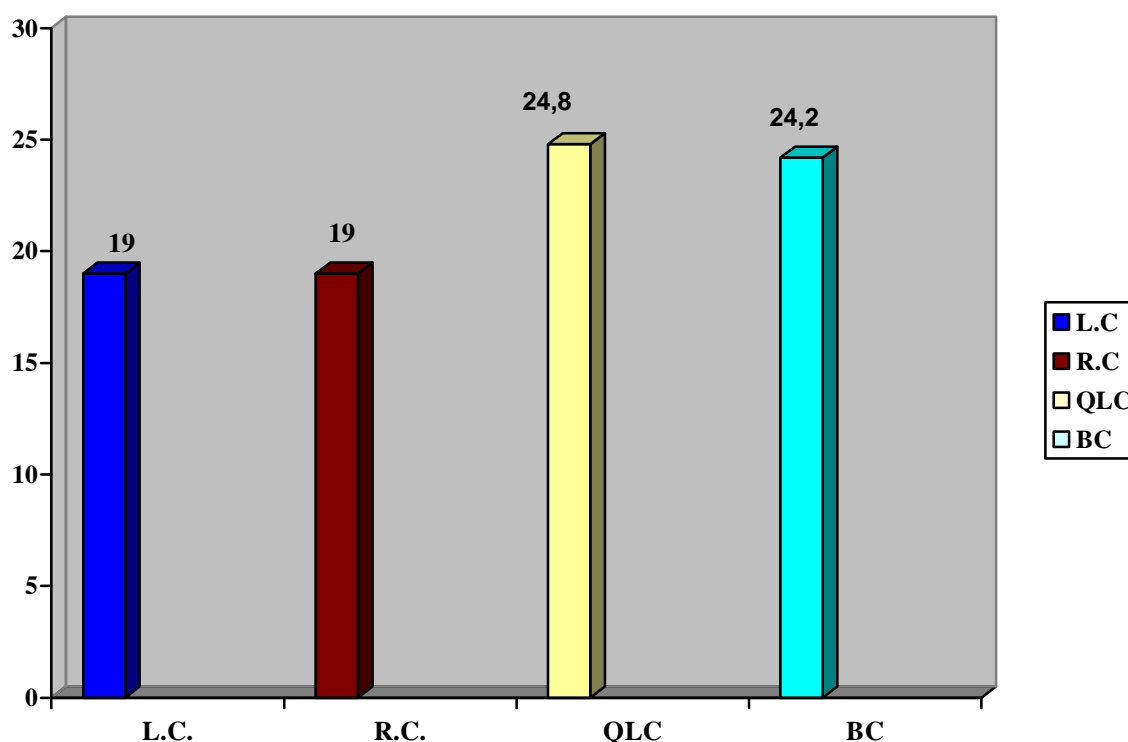



Figure 4.13. Percentage of the top 10 frequent tokens learner and reference corpora.

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 (4.14) and (4.15) suggests tremendous diversity between the first top 100 frequent tokens and the second 100 frequent tokens in both corpora.



N	Word	Freq.	%	Lemma
100	KNOW	240	0,14	
101	EXAMPLE	239	0,14	
102	ANY	235	0,14	
103	FAMILY	234	0,14	
104	MAN	230	0,14	
105	DIFFERENT	227	0,13	
106	YOUR	227	0,13	
107	DON	224	0,13	
108	MUCH	224	0,13	
109	RIGHT	222	0,13	
110	AFTER	220	0,13	
111	WORK	220	0,13	
112	RIGHTS	218	0,13	
113	HOW	217	0,13	
114	LIVE	217	0,13	
115	MUST	216	0,13	
116	FIRST	213	0,13	
117	BOTH	211	0,13	
118	ABORTION	210	0,12	
119	GO	208	0,12	
120	KNOWLEDGE	206	0,12	
121	SAME	206	0,12	
122	MY	202	0,12	
123	WERE	201	0,12	
124	EVERY	200	0,12	
125	JUST	199	0,12	
126	JOB	198	0,12	
127	COUNTRIES	197	0,12	
128	EACH	197	0,12	
129	THEN	196	0,12	
130	COUNTRY	193	0,11	
131	PROBLEM	193	0,11	
132	REASON	192	0,11	
133	SEX	187	0,11	

Figure 4.14. The second 100 frequent words in the learner corpus.



WordList - [new wordlist (F)]

File Settings Comparison Index Window Help

WordList - [new wordlist (F)]

N	Word	Freq.	%	Lemma
134	BEEN	186	0,11	
135	SAY	186	0,11	
136	UP	186	0,11	
137	NEED	182	0,11	
138	WITHOUT	181	0,11	
139	US	180	0,11	
140	USED	178	0,11	
141	ALWAYS	175	0,10	
142	ANIMALS	175	0,10	
143	EQUALITY	175	0,10	
144	HAD	175	0,10	
145	LEARN	174	0,10	
146	POWER	174	0,10	
147	EQUAL	173	0,10	
148	CHEATING	172	0,10	
149	WHILE	171	0,10	
150	ANOTHER	170	0,10	
151	BETWEEN	170	0,10	
152	SCHOOL	170	0,10	
153	TEACHERS	169	0,10	
154	THING	169	0,10	
155	TAKE	167	0,10	
156	FIND	162	0,10	
157	ENOUGH	159	0,09	
158	DAY	157	0,09	
159	HOWEVER	157	0,09	
160	WOULD	157	0,09	
161	WELL	154	0,09	
162	BAD	153	0,09	
163	LIVES	153	0,09	
164	OWN	151	0,09	
165	NEW	148	0,09	
166	FACT	146	0,09	
167	SOCIAL	146	0,09	

Continued Figure 4.14.





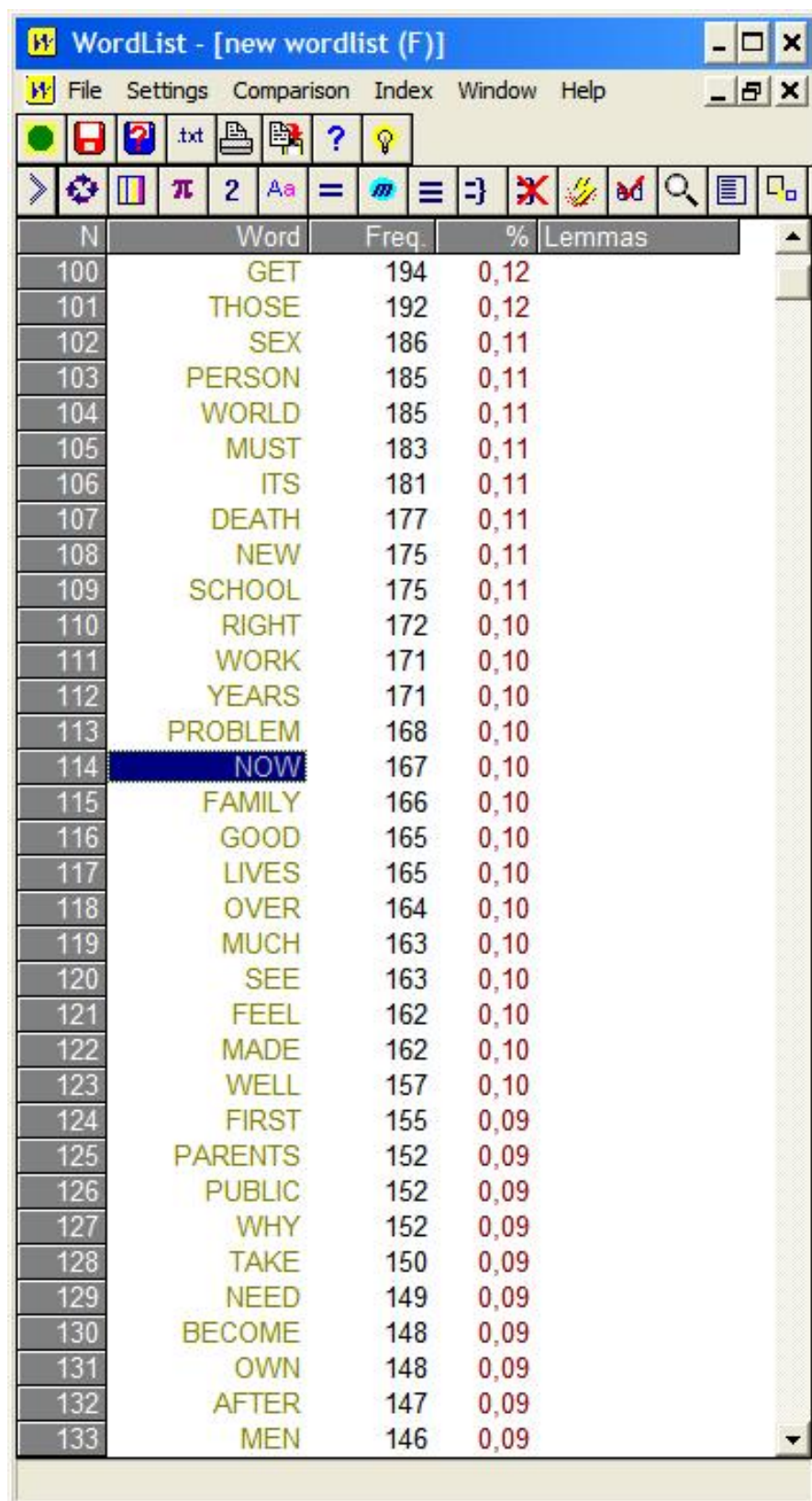
WordList - [new wordlist (F)]

File Settings Comparison Index Window Help

N	Word	Freq	%	Lemma
168	EUTHANASIA	145	0,09	
169	DOES	141	0,08	
170	BECOME	140	0,08	
171	SITUATION	139	0,08	
172	TEACHER	137	0,08	
173	DEATH	135	0,08	
174	LIVING	135	0,08	
175	NOW	135	0,08	
176	WHY	135	0,08	
177	STUDENT	134	0,08	
178	COURSE	133	0,08	
179	EVERYTHING	133	0,08	
180	SOMETHING	133	0,08	
181	CHILD	132	0,08	
182	THEMSELVES	132	0,08	
183	ESPECIALLY	131	0,08	
184	TOO	131	0,08	
185	ALTHOUGH	130	0,08	
186	ORDER	130	0,08	
187	REASONS	129	0,08	
188	SINCE	128	0,08	
189	TRY	128	0,08	
190	HELP	127	0,08	
191	GIVEN	126	0,07	
192	INFORMATION	126	0,07	
193	HAVING	125	0,07	
194	ITS	124	0,07	
195	OUT	124	0,07	
196	ETC	123	0,07	
197	TWO	123	0,07	
198	YEARS	123	0,07	
199	PARENTS	122	0,07	
200	MEANS	121	0,07	

Continued Figure 4.14.





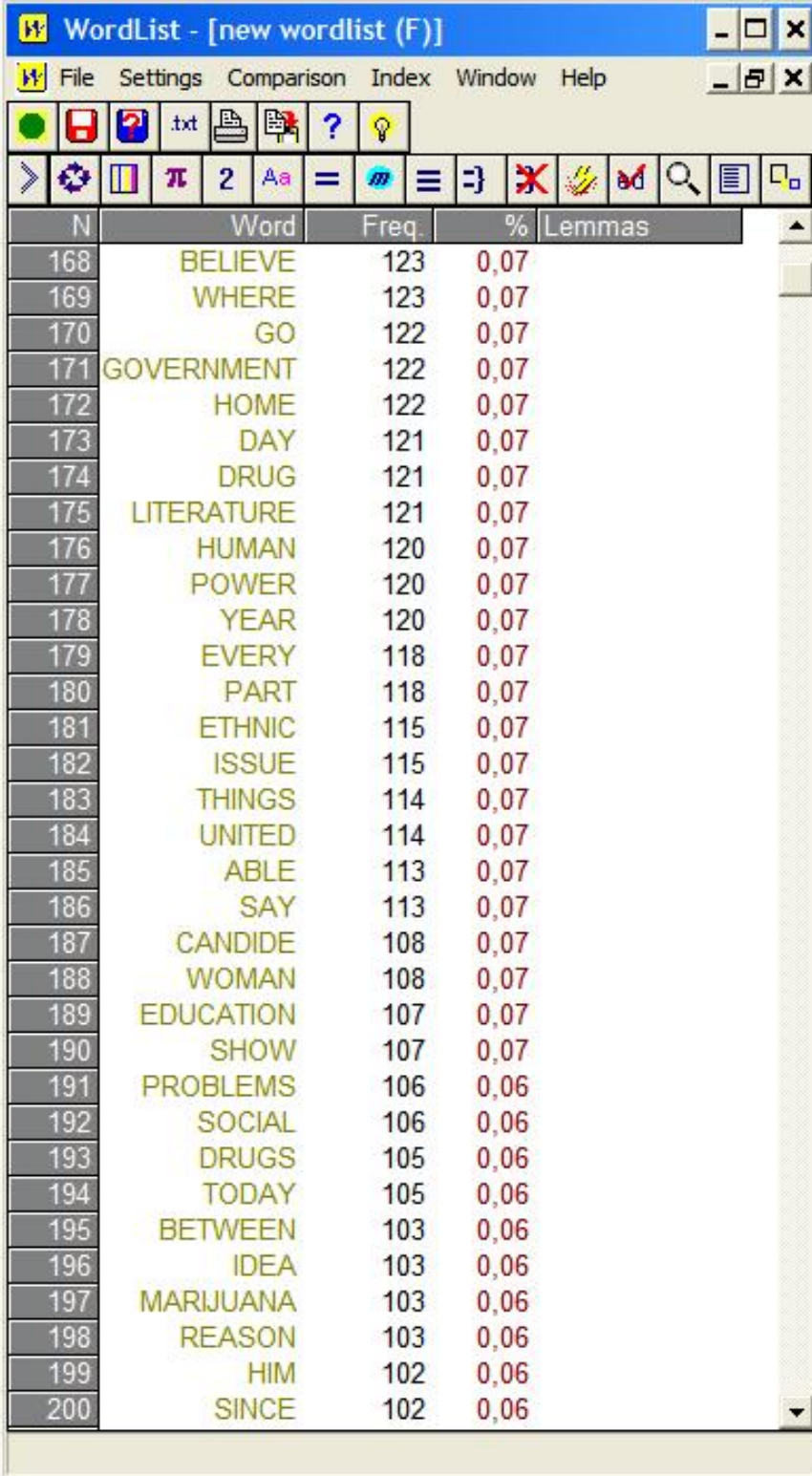
WordList - [new wordlist (F)]

File Settings Comparison Index Window Help

N	Word	Freq.	%	Lemmas
100	GET	194	0,12	
101	THOSE	192	0,12	
102	SEX	186	0,11	
103	PERSON	185	0,11	
104	WORLD	185	0,11	
105	MUST	183	0,11	
106	ITS	181	0,11	
107	DEATH	177	0,11	
108	NEW	175	0,11	
109	SCHOOL	175	0,11	
110	RIGHT	172	0,10	
111	WORK	171	0,10	
112	YEARS	171	0,10	
113	PROBLEM	168	0,10	
114	NOW	167	0,10	
115	FAMILY	166	0,10	
116	GOOD	165	0,10	
117	LIVES	165	0,10	
118	OVER	164	0,10	
119	MUCH	163	0,10	
120	SEE	163	0,10	
121	FEEL	162	0,10	
122	MADE	162	0,10	
123	WELL	157	0,10	
124	FIRST	155	0,09	
125	PARENTS	152	0,09	
126	PUBLIC	152	0,09	
127	WHY	152	0,09	
128	TAKE	150	0,09	
129	NEED	149	0,09	
130	BECOME	148	0,09	
131	OWN	148	0,09	
132	AFTER	147	0,09	
133	MEN	146	0,09	

Figure 4.15. The second 100 frequent words in the reference corpus.





N	Word	Freq.	%	Lemmas
168	BELIEVE	123	0,07	
169	WHERE	123	0,07	
170	GO	122	0,07	
171	GOVERNMENT	122	0,07	
172	HOME	122	0,07	
173	DAY	121	0,07	
174	DRUG	121	0,07	
175	LITERATURE	121	0,07	
176	HUMAN	120	0,07	
177	POWER	120	0,07	
178	YEAR	120	0,07	
179	EVERY	118	0,07	
180	PART	118	0,07	
181	ETHNIC	115	0,07	
182	ISSUE	115	0,07	
183	THINGS	114	0,07	
184	UNITED	114	0,07	
185	ABLE	113	0,07	
186	SAY	113	0,07	
187	CANDIDE	108	0,07	
188	WOMAN	108	0,07	
189	EDUCATION	107	0,07	
190	SHOW	107	0,07	
191	PROBLEMS	106	0,06	
192	SOCIAL	106	0,06	
193	DRUGS	105	0,06	
194	TODAY	105	0,06	
195	BETWEEN	103	0,06	
196	IDEA	103	0,06	
197	MARIJUANA	103	0,06	
198	REASON	103	0,06	
199	HIM	102	0,06	
200	SINCE	102	0,06	

Continued Figure 4.15.

A careful examination of the second 100 tokens in each corpus shows three crucial results.

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 (75%) 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 (51%) and (70%), respectively, as shown in Figure (4.16).

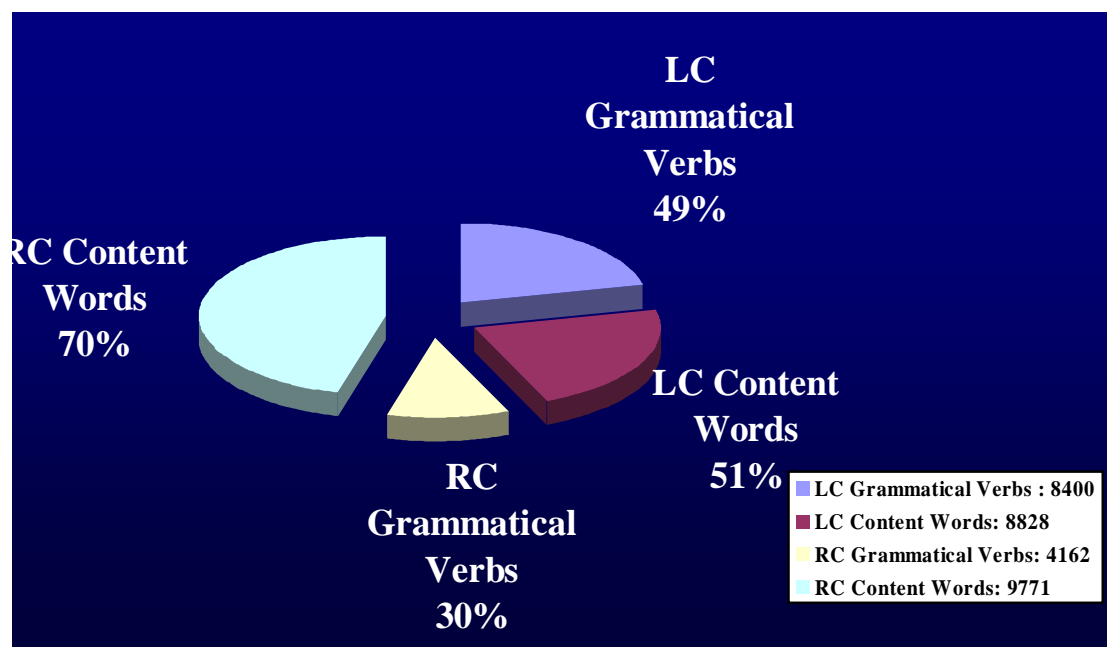


Figure 4.16. Number of content and grammatical words in the top 100 frequent token in the learner and reference corpora.

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 constitutes only (9%)

and (7%) respectively. However, the high percentage of the second 100 frequent tokens in the learner corpus compared with the reference corpus supports Goodfellow's et 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.

#### **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 (4.17), exemplifies further advantages of the tagged corpora.

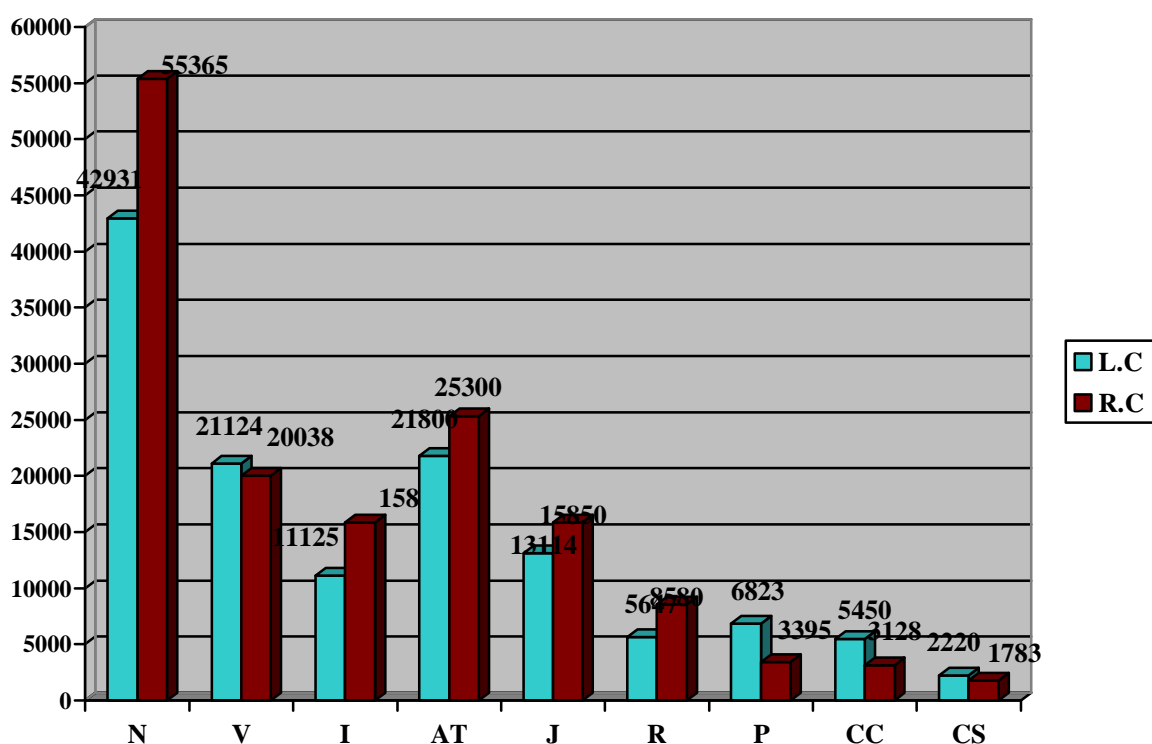


Figure 4.17. Word category in learner and reference corpora.

Table 4.2. 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 (4.17), 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 4.3. 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 Turkish uses case markers 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.

#### (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 Turkish uses only case markers, 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.

#### (d) Adverbs

The divergence in the number of adverbs, which favored the reference corpus, is conspicuous.

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 (861times) compared to (675 times) in the reference corpus.

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 (4.4).

Table 4.4. 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

The frequency of the first person pronouns above 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 (4.18). Further analysis of the use of and as a sentence opener is illustrated in the coming sections.

Concord

File View Settings Window Help

xxAND: 11 entries (sort: 5L,5L)

N	Concordance	Set	Tag	d No.	File	%
1	ofessors and teachers who are qualified . And the worst thing as the universitie			2.377	1\ticle.txt	91
2	could not be anything but an "Almighty Lord". And the Almighty ( Zeus or Al-l			78.397	1\ticle.txt	39
3	es even they know the answer very well . And most of this kind students regard			6.995	1\ticle.txt	73
4	n system of Turkey correct these ideas . And another thing that make universi			2.195	1\ticle.txt	91
5	ves to suffer pain until to eventual death . And there is no rule such as 'Euthana			52.062	1\ticle.txt	26
6	to such a question, I can't say "Yes" . And I believe many relatives think an			3.172	1\ticle.txt	56
7	life , recreating from night until morning . And the education system of Turkey			2.185	1\ticle.txt	91
8	e society from those dangerous people . And we are living in a country where t			68.896	1\ticle.txt	34
9	ty lead the students to be more passive . And also, they don't have any proble			84.835	1\ticle.txt	42
10	ple believe the saying "blood for blood" . And some say that the people are ge			68.775	1\ticle.txt	34

Figure 4.18. Examples of the use of and sentence initially.

#### (c) Subordinating conjunctions

Figure (4.17) shows, subordination use was found to be higher in the learner corpus (Word category CS; 2,220, 1,783 for Learner and reference corpus, respectively).

#### 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., Zughouli 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 (4.19) presents some of the divergence between the two corpora in this aspect.

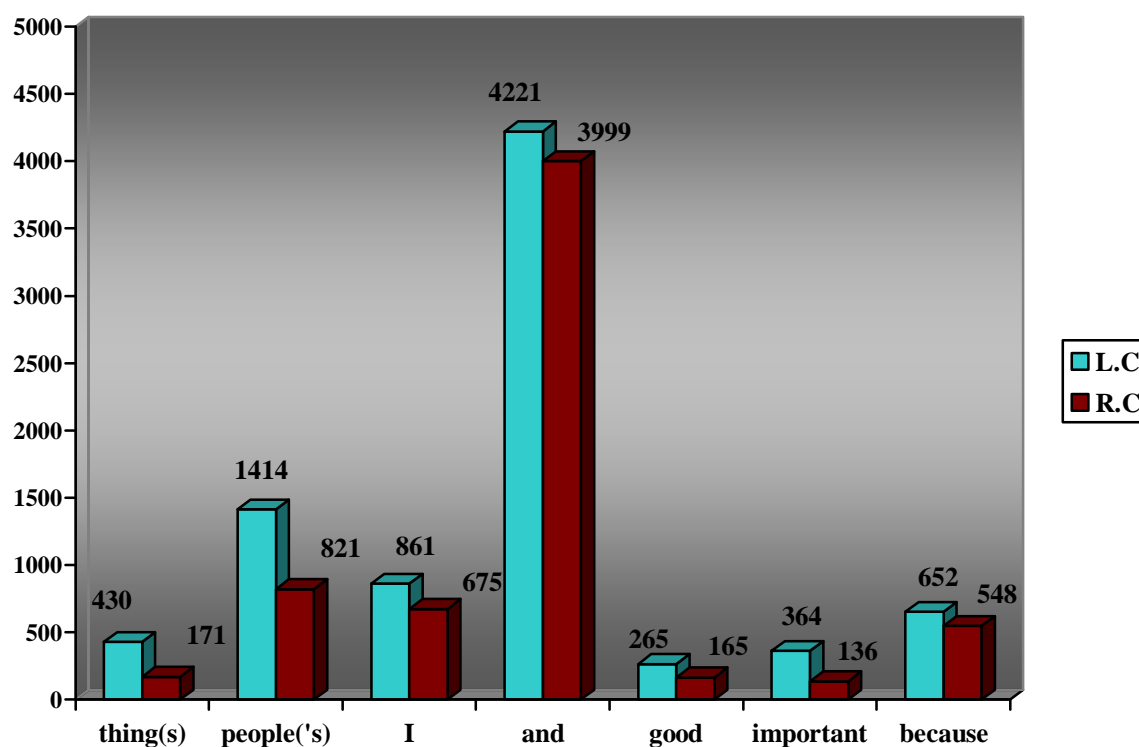


Figure 4.19. Samples of overproduction.

The above brief comparisons provide further evidence that learners' interlanguage and NSs' writing are heterogeneous. There are two possible reasons for such heterogeneous 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).

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 Turkish students of English greatly exceeds the use of the same items in the reference corpus or even all other learner corpora.

#### **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 (4.20) presents some of the underproduced items between the two corpora.

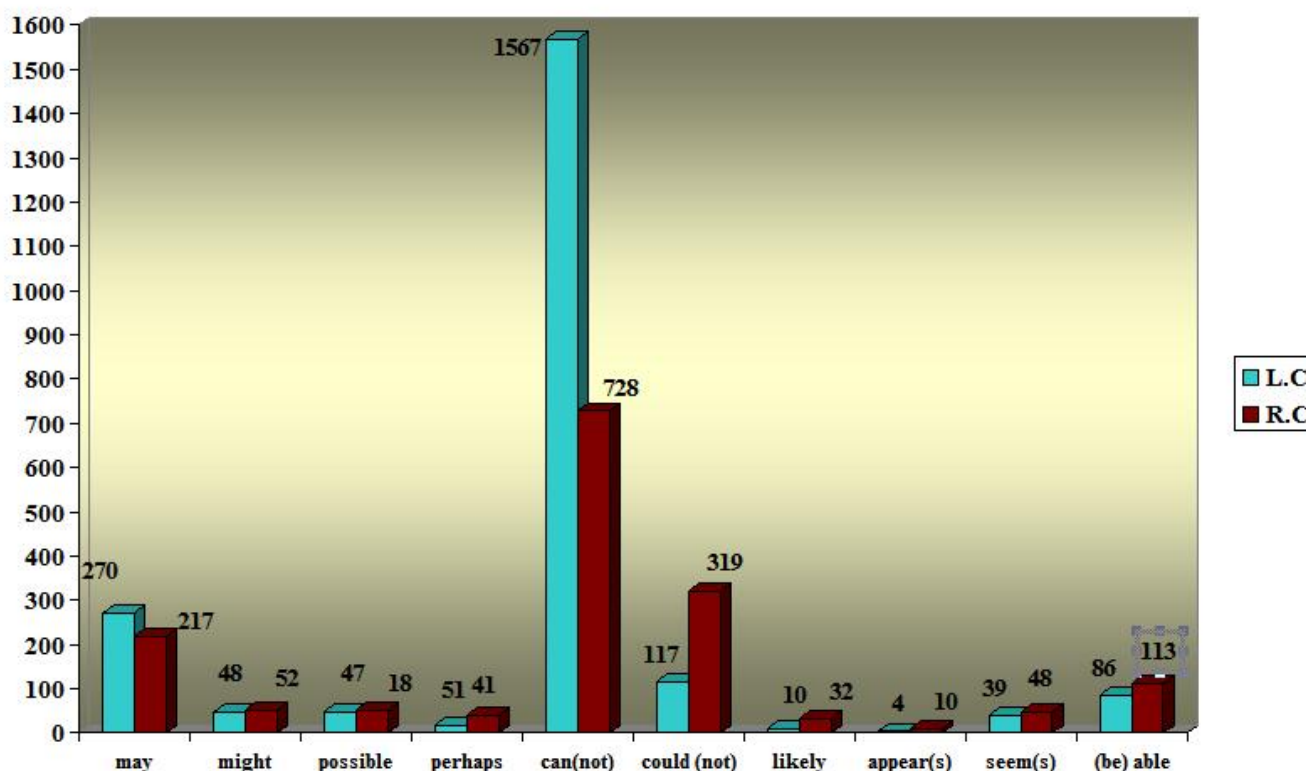


Figure 4.20. Hedges in learner and reference corpora.

The well attested data given as examples of overproduction or underproduction reveal that learners' lexicology lies between two extremes (overuse or underuse). As is seen Figure (4.20) shows some markedly few 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.

## (2) Sentence and Word 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 Word list, it turned out to be that NSs' sentences are longer than those of the learners. Figures (4.21) and (4.22) present the findings of sentence length in learner and reference corpora respectively.

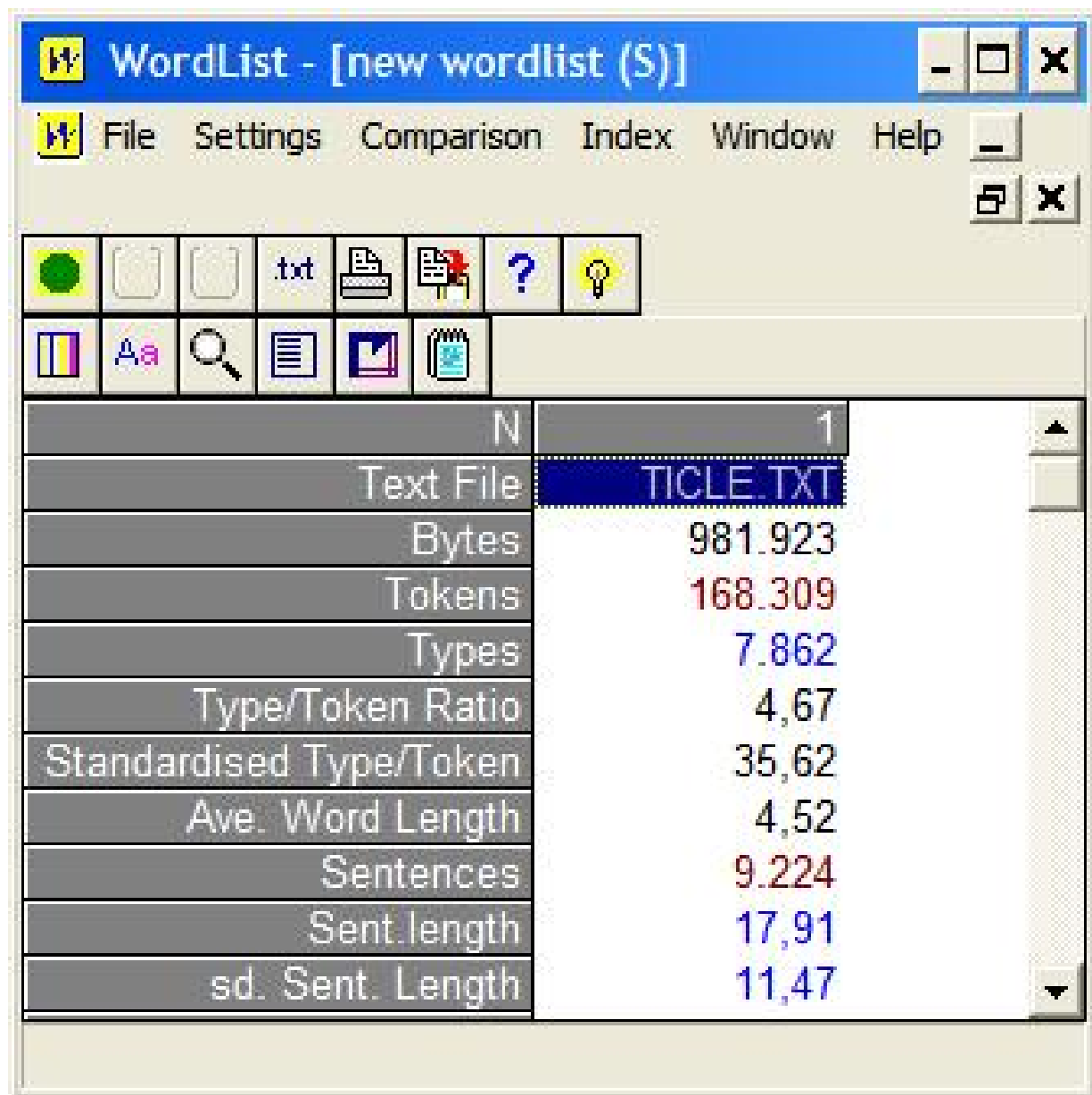


Figure 4.21. Sentence length in the learner corpus.

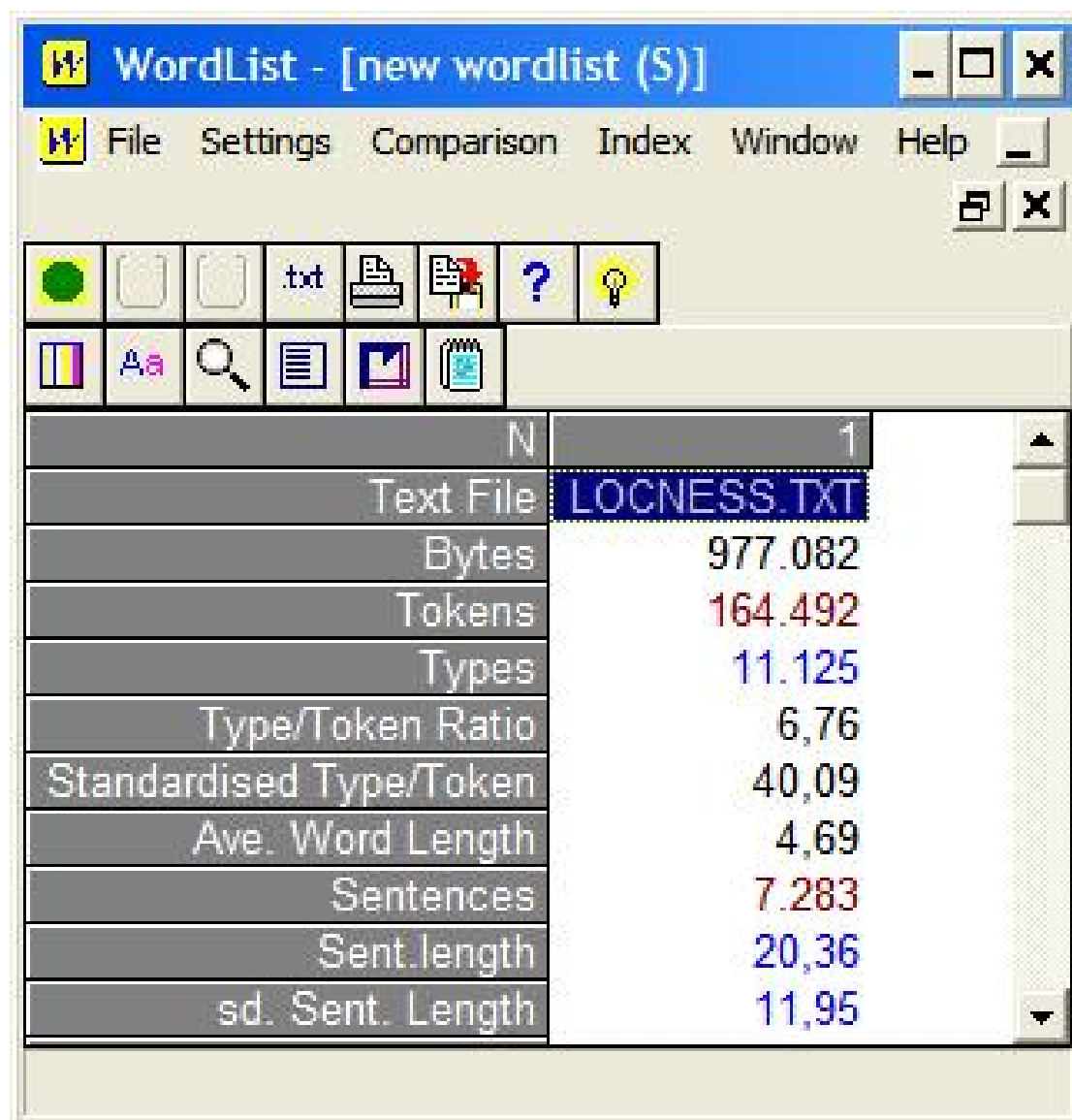


Figure 4.22. Sentence length in the reference corpus.

From a rapid scan of the figures, it becomes apparent that sentence length in the reference corpus (20.36) is longer than that of the learner corpus (17.91). Consequently, this subject calls the long-held erroneous impression among language educators about learners' 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 (4.52) is shorter than that of the reference corpus (4.69). These figures resonate with the findings of previous literature (e.g Dafu: 1994).

## CHAPTER 5

### CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS

#### 5.1. Introduction

This concluding chapter consists of three sections. Section (5.2) “summarizes by reviewing the research questions and findings of the study. Section (5.3) presents the limitations of the study. Section (5.4) provides recommendations for future research.

#### 5.2. Summary

Using empirical methods to examine lexical complexity, and text-profiling in the writing and translation of Turkish 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?

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 Turkish Students of English, (ii) a similar-sized authentic machine-readable reference corpus, and (iii) a number of software programs (e.g., Concordancer, Word list ).

The following 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.

### **5.3. 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 Turkish 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.

#### 5.4. Future Research

Since this is one of the few studies of its kind conducted on the interlanguage lexicology of Turkish students of English as a foreign language via a corpusbased 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. 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 Turkish 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).

## REFERENCES

- Aarts, Jan. 1991. *Intuition Based and Observation Based Grammars*. English Corpus Linguistics, ed. by Karin Aijmer and Bengt. Altenberg, 44-62. London: Longman.
- Aijmer, Karin and Bengt Altenberg. (eds.) 1991. *English Corpus Linguistics*. London: Longman.
- Beaugrande, Robert de. 2001. 'If I were you...': Language Standards and Corpus Data in EFL. *Revista Brasileira de Linguística 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, Susan Conrad & Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge; New York: Cambridge University Press.
- Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt and Company.
- Brooks, Nelson. 1960. *Language and Language Learning*. New York: Harcourt Brace & World.
- Burquest, Donald. 1999. *A Field Guide for Principles and Parameters Theory*. Dallas, SIL International.
- Carter, Ronald. 1987. *Vocabulary: applied linguistic perspectives*. London; Boston: Allen & Unwin.
- 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 Svartvik, Jan, 79-97. Berlin: Mouton de Gruyter.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. M.I.T. Press.

Chomsky, Noam. 1975. *Reflections on Language*. New York: Pantheon

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.

Cumming, A. and Mellow, D. 1996. *An Investigation into the Validity of Written Indicators of Second Language Proficiency*. *Validation 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.cnlbaseinfo/PHD/yangdafueng.htm>

Dulay, Heidi and Mariana Brute. 1974. *Natural sequences in child second language acquisition*. *Language Learning* 24.37-53.

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.

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.

Fries, Charles. 1945. *Teaching and learning English as a foreign language*. Ann Arbor: University of Michigan Press.

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/corporalcorpnavopen.html>* Geeraerts, Dirk. 1997. *Diachronic prototype semantics*. Oxford: Oxford University Press.

Granger S., E. Dagneaux and F. Meunier (2002) *The International Corpus of LearnerEnglish. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires deLouvain.

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.

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.

Haliday, M. A. K. 1989. *Spoken and written language*. Oxford University Press.

Hausser, Roland. 1999. *Foundations of Computational Linguistics, Human-Computer Communication in Natural Language*, 2nd Edition. Berlin, New York: Springer.

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.

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, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hymes, D. 1972. *On communicative Competence*. *Sociolinguistics*, ed. by J. B. Pride and J. Holmes. Harmondsworth; England: Penguin Books.

*Error Analysis: Source, Cause and Significance*. *Error Analysis: Perspectives on Second Language Acquisition*, ed. by Jack C. Richards and M. P. 1974., 189-215. London: Longman.

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.

Kharma, Nayef. 1985. *Advanced Composition in EFL. Abhath AI-Yarmouk* 3.7-22.

Koenig, Jean-Pierre 1999. *Lexical relations. California: CSLI Publications.*

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/>*

Laufer, Batia and Paul Nation. 1995. *Vocabulary size and use: lexical richness in L2 written production. Applied Linguistics* 16. 307-322.

Laufer, Batia and Paul Nation. 1999. *A vocabulary size test of controlled productive ability. Language Testing* 16. 33-51

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, 48 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.

Linnarud, Moira. 1986. *Lexis in composition*. *Lund Studies in English*. Malmö, Sweden: Liber Forlag.

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*. *Leamer 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.

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.

Olsson, M. 1974. *A Study of Errors, Frequencies, Origin and Effects*. Göteborg, Sweden: Pedagogiska Institutionen.

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.

Ramsey, Robert 1981. *A Technique for Interlingual Lexico-Semantic Comparison: The Lexigram*. *TESOL Quarterly* 15. 16-25.

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.

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.

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

Selinker, Larry. 1972. *Interlanguage*. *International Review of Applied Linguistics* 10. 209-31.

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

Sharwood-Smith, Michael. 1994. *Second Language Learning*. London: Longman.

Tribble, Chris. and Chris. Jones. 1990. *Concordances in the Classroom*. London: Longman

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam; Philadelphia: John Benjamins.

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.

## LIST OF APPENDIX

### UCREL CLAWS7 TAGSET

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
ATl	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)
NN02	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)
PNXI	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPHI	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)
VBO	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
VDO	do, base form (finite)
VDD	did
VDG	doing
VDI	do, infinitive (I may do... To do...)
VDN	done
VDZ	does
VHO	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)



VVO	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)

**CURRICULUM VITAE**

**Name:** Fahrettin

**Surname:** Şanal

**Birthdate:** April 4, 1957

**Birthplace:** Osmaniye

**GSM:** (532) 385 05 50

**Adress:** K. İhsaniye Mah. Şehit Ömer Taşer Sk. No: 2/11 Selçuklu – KONYA

**Education**

**1996 – 2007** Çukurova University (Ph.D.)

**1991 – 1994** Selçuk University (E.L.T. Master of Science)

**1989 – 1990** Reading University, The U.K. (E.L.T. Diploma)

**1986 - 1987** Gazi Faculty of Education (E.L.T. B.A)

**1973 – 1976** Gazi Eğitim Enstitüsü (E.L.T. B.A)

**1970 – 1973** Osmaniye High School

**1967 – 1970** Osmaniye Secondary School

**1962 - 1967** Atatürk Elementary School Osmaniye