

147689

M. K. Ü.
INSTITUTE OF SOCIAL SCIENCE
Department of English

147689

THE WRITTEN PERFORMANCE OF THE TURKISH ADVANCED
UNIVERSITY STUDENTS OF ENGLISH WITH REFERENCE TO THE
USE OF CONJUNCTS FROM THE QUANTITATIVE AND
FUNCTIONAL PERSPECTIVES

Dilaver Aif BAYRAKÇI

Supervisor
Asst. Prof. Abdurrahman KİLİMCİ

Degree Sought
MASTER of ARTS

HATAY
April, 2004

MASTER OF ARTS

In The Subject of

ENGLISH LANGUAGE TEACHING

April, 2004

We certify that this dissertation is satisfactory for the award of the degree of Master of Arts.



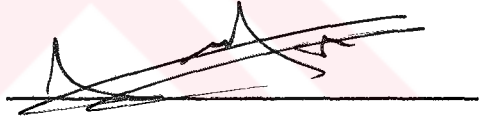
Supervisor

Asst. Prof. Abdurrahman Kilimci



(Member of Examining Committee)

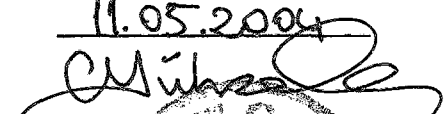
Asst. Prof. Dr. Rıza Öztürk



Asst. Prof. Dr. Cem CAN

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

KOD = 45

11.05.2004

(Director of the Institute)

Prof. Dr. Cemal YÜKSELEN
Sos. Bil. Enst. Müdürü

ACKNOWLEDGEMENTS

I would never have written this thesis without the welcome support and contribution of my advisor, Asst. Prof. Abdurrahman Kilimci who helped me a lot to crystallise my thoughts. But for his encouraging and constructive attitude at all stages, this work would never have taken its present form. He read the work in draft with great patience, and kindly allowed me access to his personal library and offered invaluable advice and insightful suggestions. I will always be grateful to him.

I am also indebted to Asst. Prof. Rıza Öztürk, Head of English Department at Mustafa Kemal University, who encouraged me a lot during this wonderful and sometimes tough experience. I benefited a lot not only from the lectures in a friendly atmosphere during M.A. courses we had, but also with the experiences he provided me in the sense of academic field of study. He has always been kind and supportive.

I would like to thank Asst. Prof. Cem Can for his invaluable comments and suggestions that were very helpful to clarify some points, which escaped my attention both during and in the final stages of the work. I was honoured to have him in my committee.

Many thanks are also due to Prof. Dr. Sylviane Granger, Director of the Centre for English Corpus Linguistics at the Universite Catholiqu de Louvain, Belgium. She kindly provided my supervisor with the native speaker corpus, LOCNESS, which I employed in the study as well.

I also would like to thank the students of both universities, who were kind enough to participate in the study and to provide us with their essays.

Last, but by no means least, I owe a great deal to my wife Aysun and my 3-year-old son Berke. I will always be indebted for her constant, encouraging, marvellous support, for her patience and understanding in times of stress.

Needless to say, I am to blame for any imperfections and faults in the work.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Abstract (Turkish).....	v
Abstract (English).....	vii
List of the Tables.....	ix
List of the Figures.....	x

CHAPTER 1..... 1

INTRODUCTION 1

1.1 Background of the study.....	1
1.2 Statement of the problem.....	1
1.3 Scope of the study.....	3
1.3.1 The semantic roles of conjuncts.....	5
1.4 Aim of the study.....	7
1.5 Operational definitions.....	7
1.6 Research questions.....	8
1.7 Assumptions and limitations.....	8
1.8 Subjects.....	9

CHAPTER 2 9

REVIEW OF LITERATURE..... 9

2.1 What is corpus?	9
2.2 What is corpus linguistics?	12
2.2.1 Benefits of corpus data.....	19
2.2.2 The scope of corpus linguistics.....	20
2.2.3 Some major achievements of computer corpus-based scholarship.....	29

2.3 Types of corpora.....	31
2.3.1 Pre-electronic corpora.....	35
2.3.1.1 Biblical and literary.....	36
2.3.1.2 Lexicographical.....	37
2.3.1.3 Dialect.....	38
2.3.1.4 Language education.....	39
2.3.1.5 Grammatical.....	40
2.3.2 Electronic corpora.....	41
2.3.2.1 General corpora.....	41
2.3.2.2 Specialized corpora.....	42
2.4 Major areas of research in corpus linguistic.....	45
2.4.1 Corpora for lexicography.....	46
2.4.2 Dictionaries as corpora.....	47
2.4.3 Corpora for studying spoken English.....	48
2.4.4 Diachronic corpora.....	48
2.4.5 Corpora for research on language acquisition.....	49
2.4.5.1 Language pedagogy.....	49
2.4.6 Other corpora for special purposes.....	51
2.5 Techniques and procedures used in corpus linguistics.....	54
2.5.1 Issues in corpus design and compilation.....	54
2.5.2 Static or dynamic.....	54
2.5.3 Representativeness and balance.....	56
2.5.4 Size.....	58
2.5.5 Compiling a corpus.....	60
2.5.6 Corpus design.....	60
2.5.7 Planning a storage system and keeping records.....	61
2.5.8 Getting permission.....	61
2.5.9 Data capture, mark up and documentation.....	61
2.6 Application of corpus based analysis.....	64
2.6.1 Applying corpus linguistics to teaching.....	67
2.6.1.1 Syllabus design.....	67
2.6.1.2 Materials development.....	67

2.6.1.3 Classroom activities.....	67
2.6.1.4 Teacher / student roles and benefits.....	68
2.6.1.5 Problematic issues involved.....	69
2.6.1.6 Exploiting a corpus for a classroom activity.....	70
2.7 Corpus-based research on second or foreign language acquisition.....	70
2.7.1 Corpus-based approaches to language teaching.....	70
2.7.1.1 Language teaching methodology.....	70
2.7.1.2 The advantages of doing corpus-based analyses...72	
2.7.2 First language acquisition.....	72
2.7.3 Second language acquisition.....	72
2.7.4 Foreign language acquisition.....	73
2.8 Computer corpus based interlanguage analysis.....	73
2.8.1 Processes.....	74
2.8.2 Contrastive interlanguage analysis.....	75
2.8.3 The role of interlanguage in foreign language teaching.....	77
2.8.4 Automated linguistic analysis.....	78
2.8.4.1 Linguistic software tools.....	78
2.8.4.2 CLC methodology.....	79
2.8.5 The importance of interlanguage analysis in foreign language teaching.....	81
CHAPTER 3	88
METHODOLOGY	88
3.1 Procedure.....	88
3.1.1 Subjects.....	88
3.1.2 Instruments.....	88
3.1.3 Data collection.....	88

CHAPTER 4	89
DATA ANALYSIS	89
4.1.Overall Analysis.....	89
4.2 Individual Analysis.....	91
4.3 Individual comparison of listing conjuncts.....	92
4.4 Individual comparison of summative conjuncts.....	94
4.5 Individual comparison of appositive conjuncts.....	96
4.6 Individual comparison of resultive conjuncts.....	97
4.7 Individual comparison of inferential conjuncts.....	99
4.8 Individual comparison of contrastive conjuncts.....	100
4.9 Individual comparison of transitional conjuncts.....	101
4.10 Overall comparison of conjuncts	102
CHAPTER 5	104
CONCLUSION	104
5.1 Introduction.....	104
5.2 Conclusions.....	104
5.3 Evaluation of the research questions.....	104
5.4 Implications	106
REFERENCES	107
APPENDICES	118
Appendix 1 3 sample argumentative essays by native American University students.....	118
Appendix 2 3 sample argumentative essays by Non-native Turkish advanced University students.....	123

ÖZET**İNGİLİZ DİLİ EĞİTİMİ BÖLÜMÜNDEKİ ÖĞRENCİLERİN,
BAĞLAÇLARI İŞLEVSEL VE NİCELİK AÇISINDAN YAZILI ANLATIMDA
KULLANMA PERFORMANSLARININ ANA DİLİ İNGİLİZCE OLAN
ÖĞRENCİLERİN PERFORMANSLARI ESAS ALINARAK BİLGİSAYAR
DESTEKLİ ORTAMDA KARŞILAŞTIRMALI ANALİZİ****Dilaver Ali BAYRAKÇI****İngiliz Dili Eğitimi Anabilim Dalı Yüksek Lisans****Danışman : Yrd. Doç. Dr. Abdurrahman KİLİMCİ****NİSAN, 2004, 130 sayfa**

Bu çalışmanın amacı öğrencilerin bağlaçları işlevsellik ve nicelik açısından kullanım performanslarının bilgisayarlı dilbilim araçları ile araştırmak ve bunu anadili İngilizce olan üniversite öğrencilerinin dil bilgileri ile karşılaştırmaktır. Çalışma, dilin betimlenmesini amaçlayan yeni bir yaklaşım olan bilgisayarlı dilbilim çerçevesinde öğrenci verilerini bilgisayar yolu ile inceleme yöntemi üzerine odaklanmaktadır.

Öğrencilerin İngilizce dil düzeyini ortaya koymak amacıyla, çalışma, bir bilgisayar programı yardımı ile her bir sözcüğü dilsel işlevine göre etiketlenmiş deney ve kontrol olmak üzere iki veri yığını kullanmaktadır. Deney veri yığını, Çukurova ve Mustafa Kemal Üniversitesi İngiliz Dili Eğitimi Bölümünde son sınıf bilgi seviyesinde bulunan öğrencilerinin yazdığı tartışma ve betimsel nitelikli kompozisyonlardan oluşmaktadır. Kontrol veri yığını ise, Amerikalı üniversite öğrencilerinin tartışma içerikli kompozisyonlarından oluşmaktadır. Kontrol veri yığını esas alarak ve Karşılaştırmalı Ara Dil Analizini benimseyerek yapılan araştırmada, amaç hedef dil ile ara dili (Amerikalı öğrencilerin İngilizce'sinin Türk öğrencilerin İngilizce'si ile karşılaştırılması) karşılaştırarak ilgili bölümde uygulanan müfredatın

öğrencilerin İngilizce'sini akademik yazım normlarına ne kadar yaklaştırdığını araştırmaktır.

Çalışma, bir dizi sözcük ve sözdizimi analizinden elde edilen bulgular ışığında, hem uygulanan müfredat bakımından gelişimsel olarak, hem de ana dili İngilizce olan öğrencilerin performansını esas alarak hangi sözcüklerin öğrenciler tarafından çok sık ya da çok seyrek kullanıldığı, kullanımından kaçınıldığı, resmi yada resmi olmayan kullanım olgusu açısından öğrencilerin yazılı anlatım performansını değerlendirmektedir.

Anahtar Kelimeler: İşlevsel kullanım, nicel kullanım, Çok sık kullanım, Çok seyrek kullanım, İngilizce Dil Düzeyi, Yazılı Anlatım Performansı, Kullanımdan Kaçınma, Resmi Kullanım, Resmi Olmayan Kullanım.



ABSTRACT
**THE COMPUTER CORPUS-BASED QUANTITATIVE AND
FUNCTIONAL INTERLANGUAGE ANALYSIS WITH REFERENCE TO
CONJUNCTS ON THE WRITTEN PERFORMANCE OF TURKISH
STUDENTS FROM THE ENGLISH DEPARTMENT**

Dilaver Ali BAYRAKÇI

Department of English, Master of Arts

Supervisor: Asst. Prof. Abdurrahman KİLİMCİ

April 2004, 130 Pages

The Computer Corpus-Based Quantitative and Functional Contrastive Interlanguage Analysis on the Written Performance of the English Language Learners at Çukurova and Mustafa Kemal Universities with Reference to the Native Speaker Performance.

This study aims to explore learners' interlanguage to determine to what extent their written language reflects the norms of academic writing from both quantitative and functional perspectives through corpus linguistics. After a general review of approaches to learner language and traditional analysis methods, the study focuses on computer learner corpus (CLC) research within corpus linguistics, a new computer based approach to linguistic description.

In order to characterise learners' English, the study utilises two POS (part of speech) tagged corpora, an experimental corpus comprising argumentative and descriptive essays by the Turkish university students of proficiency level and a reference (control) corpus of argumentative essays by the American University students. Taking reference corpus as basis and adopting contrastive interlanguage analysis (CIA), the study, then, investigates to what extent the syllabus in effect has approximated the learners language to academic writing norms through NL vs. IL (comparison

of native language and interlanguage) comparisons carried out both POS tagged and raw corpora.

Finally, in the light of the findings from a serious quantitative and functional analyses, learner groups' performance is evaluated in terms of both the developmental aspect taking into consideration the syllabus in effect and over–and/or underuse phenomena regarding the native speaker performance.

Keywords: Functional performance, Quantitative Performance Linguistic Proficiency, Assumed Linguistic Proficiency, Avoidance, Formal Instruction, Informal Instruction, Overuse, Underuse



LIST OF THE TABLES

1. Overall Analysis.....	89
2. Individual Analysis.....	90
3. Individual comparison of listing conjuncts according to their semantic categories.....	92
4. Individual comparison of summative conjuncts according to their semantic categories.....	94
5. Individual comparison of appositive conjuncts according to their semantic categories.....	96
6. Individual comparison of resultive conjuncts according to their semantic categories.....	97
7. Individual comparison of inferential conjuncts according to their semantic categories.....	98
8. Individual comparison of contrastive conjuncts according to their semantic categories.....	100
9. Individual comparison of transitional conjuncts according to their semantic categories.....	101
10. Overall comparison of conjuncts according to their semantic categories.....	102

LIST OF THE FIGURES

1. The figure of seven conjunctive roles and their subdivisions.....4
2. The figure of five main fields of scholarship in corpus linguistics.....36
3. The total figures of Corpus linguistics studies in the latter half of this century.....45



CHAPTER 1

INTRODUCTION

1.1 Background To The Study

The present study takes the Computer Learner Corpus (CLC) research as its basis. Within the framework of corpus-based approach to linguistic description, CLC adopts Contrastive Interlanguage Analysis (CIA) as its methodology (Granger 1998). Since CLC studies attempts to describe learner English and employ the methods and techniques of corpus linguistics, it is closely related to both corpus linguistics and second language research (SLA). Milton (1998) underlines the significance of interlanguage studies from the CLC perspective and points out that "the collection and study of corpora of interlanguage are powerful and necessary prerequisites to the understanding of the production". Accordingly, this study aims to compile a corpus of learners' writings and carry out a contrastive interlanguage analysis from the perspective of native language (NL) and interlanguage (IL). The outcome of the study is expected to shed light on advanced Turkish university students' interlanguage in terms of the use of conjuncts from both functional and quantitative perspectives.

1.2 Statement of The Problem

According to Granger (1998) adverbial connectors such as however, nevertheless, although, besides and so on - what Quirk et al. (1985) call 'conjuncts' - adopt a cohesive role in the written language. In addition, "connectors can be said to function as cohesive 'signposts' in discourse, helping the listener or reader to relate successive units to each other and thus making sense of the text" (Leech&Starvik 1994:177). Halliday and Hasan (1985) add that conjunctive elements play an indirect cohesive roles because of their specific meanings: they also note that such words are not primarily devices for reaching out into the preceding (or following) text, but

they seem to express certain meanings requiring the presence of other components in the discourse. Halliday and Hasan (1985:4) explain that “ the concept of cohesion is a semantic one; it refers to relations of meaning that exist within the text, and that define it as a text”. They also point out that cohesion, the general meaning of which is embodied in the concept of text, provide ‘texture’, and helps to create text. They define cohesion as a general text-forming relation, or set of such relations, certain of which, when incorporated within a sentence structure. And they add that they are subject to certain restrictions because the grammatical condition of ‘being a sentence’ ensures that the parts go together to form a text anyway. But, they note, the cohesive relations themselves are the same whether their elements are within the same sentence or not (Halliday&Hasan 1985:9).

Celce-Murcia (1990:137) also defines the meaning of conjunction as “ a word or expression that signals the type of link a sentence or clause has with the preceding sentence or text; from a broad semantic perspective, a conjunction expresses an additive, adversative, casual, or sequential tie. Similarly, Granger (1998) points out such cohesive elements are significant in the sense that they connect one idea to the other in a logical way and thus they are commonly called ‘connectors’.

For this reason, in the light of the definitions given it is clear that the less use of the conjuncts not only renders a piece of writing difficult to understand but also causes breakdown between the ideas presented and give a disorganized look to the essays. Granger (1998) notes that coherence and clarity are essential for effective communication. She also adds that connectors such as *but* (to indicate a contrast), *because* (reason), *therefore* (result), *in addition* (listing), *for instance* (exemplification), etc. achieve this by signalling logical or semantic relations between units of discourse. As a result, a student’s knowledge of wide variety conjuncts will make his writing more effective in terms of presenting ideas logically.

We can see from Granger’s research on the use of conjuncts among the Swedish learners of English that, no matter how advance the learners are, they limit themselves to very few conjuncts and she notes that this

gives the idea of repetition or in other words 'overuse' (Granger 1998:81-86). The research suggests that they either avoid using conjuncts such as resultive and contrastive, because they do not feel safe in using them or because they are not taught the significance and the functions of these conjuncts in writing.

In this particular study, taking Granger's (1998) study as a starting point, we conducted a pilot study with the writing samples of advanced Turkish students. Our pilot study results reveal that having a limited repertoire of conjuncts students almost often stick to the same conjuncts and they seem to be unable to restate the same idea in another way.

1.3 Scope of The Study

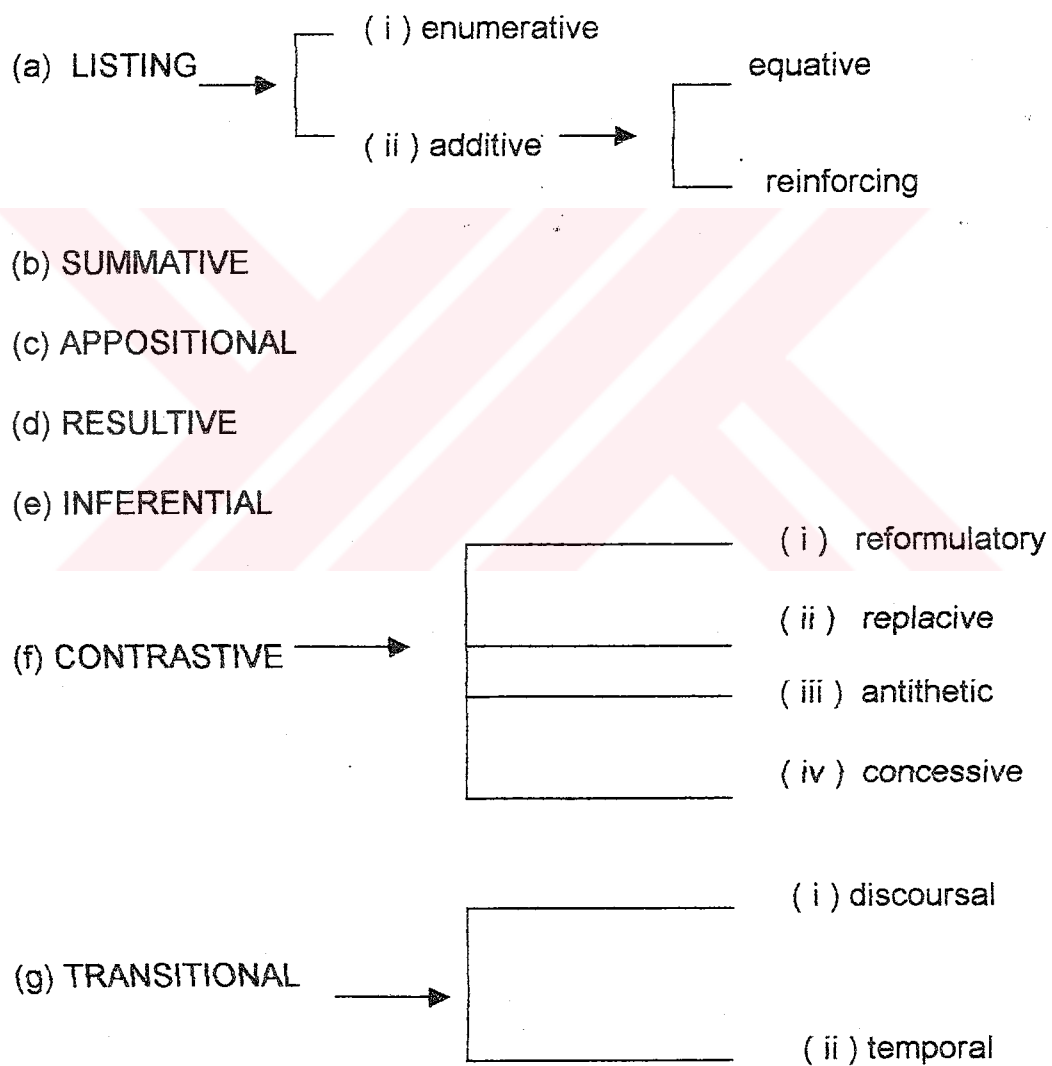
The study focuses on the written performance of the advanced Turkish University students with reference to use of conjuncts from functional and quantitative perspectives. In order to carry out an IL (interlanguage) and NL (nativelanguage) comparison, two comparable corpora were collected: native speaker (NS) and nonnative speaker (NNS) corpora. The NS corpus was made up of the American University students' argumentative essays and, similarly, the NNS corpus was compiled from the argumentative essays written by the advanced Turkish University students in their 3rd and 4th years at both the Çukurova University and Mustafa Kemal University.

The study particularly dwells on the conjuncts used by both NNS and NS in their essays. Both corpora were compared and in terms of the choice of conjuncts employed and their contribution to the argumentative essays written by NNSs and NSs.

Beside providing quantitative information as to the occurrence of the conjuncts in the database, the study also focus on the functional role of conjuncts in the written production of the American university students and advanced Turkish university students of English. Hence, the study is both quantitative and qualitative in this respect.

The conjuncts which were dealt with in the study are based on the seven main conjunctive roles with fairly clear subdivisions in some cases as defined by Quirk et al. (1997:634). Seven conjunctive roles can be distinguished in some cases with fairly clear subdivisions as shown in the following diagram:

fig 1. The roles of conjuncts with their subdivisions



Reference : Quirk et. al. A comprehension of the English Grammar.
Longman 1997. London. p. 634

1.3.1 The Semantic Roles Of Conjuncts

The common conjuncts the study focused on are listed below according to their role classes and subclasses. The list includes all adverb realizations as well as some frequent prepositional phrases and noun phrases except for enumerative conjuncts (which are an open class).

The conjuncts which are rare and used especially in spoken English were not included in the list (Quirk et al. 1997:634-636). The list includes role classes, subclasses and subtypes.

(a) LISTING (Class)

(i) Enumerative (subclass)

first, second, third..., first (ly), secondly, thirdly...,
 one, two, three..., in the first place, in the second place ...
 first of all, second of all, on the one hand ...on the other hand,
 for one thing ...(and) for another (thing), for a start, to begin with,
 to start with,next, then, to conclude, finally, last, lastly, last of all

(ii) Additive (Subclass)

Equative (Subtype)

correspondingly , equally, likewise, similarly, in the same way, by the same token

Reinforcing (Subtype)

again, also, besides, further, furthermore, moreover, in particular, what is more, in addition, above all, on top of it all, to top it (all), to cap it (all)

(b) SUMMATIVE (Class)

altogether, overall, then, therefore, thus, (all) in all; in conclusion, in sum, to conclude, to sum up, to summarize

(c) APPOSSITIVE (Class)

namely, thus, in other words, for example, for instance, that is, that is to say, specifically

(d) RESULTIVE (Class)

accordingly, consequently, hence , now, so,
therefore, thus, as a consequence, in consequence, as a result,
of course; somehow ['for some reason or other']

(e) INFERENTIAL (Class)

else, otherwise, then, in other words, in that case

(f) CONTRASTIVE (Class)**(i) Reformulatory (Subclass)**

better, rather, more accurately, more precisely; alias,
alternatively, in other words

(ii) Replacive (Subclass)

again, alternatively, rather better, worse;
on the other hand

(iii) Antithetic (Subclass)

contrariwise, conversely, instead, oppositely, then;
on the contrary, in contrast, by contrast, by way of
contrast, in comparison, by comparison, by way of
comparison, (on the one hand ...) on the other hand

(iv) Concessive (Subclass)

anyhow, anyway , anyways, besides, else, however,
nevertheless, nonetheless> [also written none the
less], notwithstanding, still, though, yet, in' any case,
in' any event, at' any rate, at' all events, for' all that,
in spite of that, in spite of it all, after all, at the same'
time, on the other hand, all the same, admittedly, of
course, still and all ; that said

g) TRANSITIONAL (Class)**(i) Discoursal (Subclass)**

incidentally, now, by the way, by the by(e)

(ii) Temporal (Subclass)

meantime, meanwhile, in the meantime, in the
meanwhile; originally, subsequently, eventually

1.4 Aim Of The Study

The aim of this study is to find out the written performance of the advanced Turkish University students with reference to use of conjuncts from functional and quantitative perspectives. By carrying out an interlanguage and nativelanguage comparison, it is aimed to detect the choice of conjuncts made use of, their frequency and their contribution to the argumentative essays written by NNSs and NSs. In addition to that it is also concentrated on functional role of conjuncts in written production of the native and nonnative university students.

1.5 Operational Definitions

For the purpose of this study, the terms below will be defined as follows:

LINGUISTIC REPERTOIRE: Learners' cumulative experiences in relation to lexical and syntactic knowledge.

LINGUISTIC PROFICIENCY: Learners' skill to use their stock of lexical and syntactic experiences (Linguistic repertoire) effectively.

ASSUMED LINGUISTIC PROFICIENCY: Learners' knowledge of English with respect to their levels. The higher the levels they are at are, the more proficient they are assumed to be.

AVOIDANCE: Learners' suppressing certain syntactic structures for fear of failure and, as a result, committing an error.

FORMAL INSTRUCTION: This term is interchangeably used with grammar instruction, and means laying emphasis on the grammatical aspects of the language.

UNDERUSE: Students' using certain linguistic items less frequently than the native speakers (NSs)

OVERUSE: Students' using certain linguistic items more frequently than the NSs.

1.6 Research Questions

The research questions the study seeks to clarify are:

Q1. Do advanced Turkish learners of English employ conjuncts to the same extent as native American university students? If they avoid using certain conjuncts, what type of conjuncts do they avoid? And in what frequency do they use them?

Q2. What kind of semantic roles do conjuncts play in NNS and NS argumentative essays? And to what extent do they match in both corpora as far as the semantic relations are concerned?

Q3. What semantic categories of conjuncts characterize both the NNS and NS argumentative essays?

Q4. If certain categories are overused/underused, what might be the reason?

1.7 Assumptions And Limitations

While comparing the two groups of students it was assumed that the Turkish students are all at proficient level. As a result of this, they were all accepted as to have used the conjuncts in a proper way. In addition to that only the conjuncts, which were used most often by the both groups have were taken into consideration. That is to say, top 30. The other rest were excluded from the analyse. The number of the words was to be used in the essays were also limited to 1000 word at the most and 500 words at least for each student.

1.8 Subjects

It is assumed that students in each group have the same linguistic proficiency regardless of such variables as possible difference in their knowledge of English, age, sex, social background, which might result from individual differences. The reason underlying this assumption is that these students have been admitted to the related levels after they have achieved the desired level of proficiency.

CHAPTER 2

REVIEW OF LITERATURE

2.1 What Is Corpus?

According to the linguists such as Quirk, Kennedy, Granger et.al., the concept of carrying out research on written or spoken texts is known to be the field of corpus linguistics. If we make a review of the literature we can see that individual texts are often used for many kinds of literary and linguistic analysis. For example, the stylistic analysis of a poem, or a conversation analysis of a TV talk show can be two of them. In addition to that the shortest description corpus can be made as "A corpus consists of a databank of natural texts, compiled from writing and/or a transcription of recorded speech "(McEnery&Wilson 1996:2).

As a matter of fact, it would be beneficial to have a look at other definitions of corpus by different authorities. There are numerous definitions of corpus, which are similar to each other in meaning, but in principle, Kennedy defines corpus as "In the language sciences a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description "(Kennedy 1998:1). Similarly McEnery and Wilson add some for the description of the corpus as, "we can say that any collection of more than one text can be called a corpus" and they add "corpus is the written text on a particular subject, which can be used as a

basis for linguistics analysis" (McEnery&Wilson 1996:2). In a similar idea Granger notes "A corpus is a body of text assembled according to explicit design criteria for specific purpose" (Granger 1998:7). Krieger (2003) states, "a corpus consists of a databank of natural texts, compiled from writing and/or a transcription of recorded speech". Aston points out that "the corpus primarily constitutes a resource for the learner, either via printed concordances or hands-on concording of corpora" (Aston 1997:51). More to the point, another definition of corpus is that "A corpus is a body of text assembled according to explicit design criteria for a specific purpose" (Atkins and Clear 1992:5). From the immediate definition we can infer that it must be collected very meticulously. What Sinclair notes for the idea is that "the results are only as good as the corpus" (Sinclair 1991:9). In other words it can be said that "the quality of the investigation is directly related to the quality of the data" (Granger 1998:7). In addition to the definitions above, Granger (1998) also exemplifies the notion as:

"To begin with a hypothetical but realistic example, let us suppose that higher education teacher X, in a non-English speaking country, teaches English to her students every week, and every so often sets them essays to write, or other written tasks in English. Now, instead of returning those essays to students with comments and a sigh of relief, she stores the essays (of course with the students' permission) in her computer, and is gradually building up, week by week, a larger and more representative collection of her students' work. Helped by computer tools such as a concordance package, she can extract data and frequency information from this 'corpus', and she can analyse her students' progress as a group in some depth."

Furthermore; Leech (1991) sees the notion corpus as the sole source of evidence in the formation of linguistic theory and clarifies it as - "This was when linguists...regarded the corpus as the sole explicandum of linguistics" (Leech 1991:29). And Sinclair (1996:23) also makes a definition of corpus as "the linguists's corpus is a collection of pieces of language, selected and ordered according to explicit linguistic criteria in order to be used as sample of the language." Crystal states that corpus, plural corpora means a collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. "The main purpose of a corpus is to verify a hypothesis about language - for example, to determine how the usage of a

particular sound, word, or syntactic construction varies" (Crystal 1992:85). In addition to that McArthur (1992) states the word 'corpus' comes from Latin *corpus* body and he also adds similar opinion to that of Crystal's at the same time as the plural form of corpus is usually known as 'corpora'. He briefly continues that corpus is a collection of texts, especially if complete and self-contained. In Granger (1998), he gives *the corpus of Anglo-Saxon verse* as a sample. In order to clarify the notion he asserts that in linguistics and lexicography, a body of texts, utterances or other specimens considered more or less representative of a language, and he adds, they are usually stored as an electronic database. Besides, he reports that currently, computer corpora may store many millions of running words, whose features can be analysed by means of *tagging* (the addition of identifying and classifying tags to words and other formations) and the use of concordancing programs (McArthur 1992:265-266).

When we look up a dictionary we can see that the definition of corpus in Latin is "body", so we can say that a corpus is any body of text. But the term "corpus" when used in the context of modern linguistics it can easily be seen that tends most frequently to have more specific connotations than this simple definition. In the following list McEnery&Wilson (1996) describe the four main characteristics of the modern corpus.

- Sampling and representativeness
- Finite size
- Machine-readable form
- A standard reference

There is often a *tacit* understanding that a corpus constitutes a standard reference for the language variety that it represents. From this, it can be supposed that it will be widely available to other researchers, which is indeed the case with many corpora - e.g. the Brown Corpus, the LOB corpus and the London-Lund corpus. McEnery&Wilson (1996) clarify the advantages of matter as:

- One advantage of a widely available corpus is that it provides a yardstick by which successive studies can be measured. So long as the methodology is made clear, new results on related topics can be directly compared with already published results without the need for re-computation.
- Also, a standard corpus also means that a continuous base of data is being used. This implies that any variation between studies is less likely to be attributed to differences in the data and more to the adequacy of the assumptions and methodology contained in the study.

2.2 What Is Corpus Linguistics?

In recent years, it can be seen that a great amount of research has been put into practice to demonstrate how computers can ease language learning. One specific area on the computer limit, which still remains reasonably open to exploration, is corpus linguistics. Having learned that corpora will revolutionize and boost language teaching, many researchers have become very interested to find out what corpus linguistics and studies can make themselves learn and have to offer the English language teachers and how feasible such an operation would be. This thesis is supposed to address those questions by examining what corpus linguistics is, how it can be applied to teaching English, and some of the issues concerned. Resources are also included both during the account of my thesis and in references section, which will assist anyone who is interested in pursuing this line of study further.

When we make an overall review on corpus linguistics we can say that it first appeared in 1960s. "The computer corpus has infiltrated all fields of language-related research, from lexicography to literary criticism through artificial intelligence and language teaching" (Granger 1998:3). Similarly Leech points out that "corpus linguistics is a new way of thinking about language, which is challenging some of our most deeply-rooted ideas about

language" (Leech 1992:107). He also adds that " it focuses on performance (rather than competence), description (rather than universals) and quantitative as well as qualitative analysis" (Leech 1992:107). Crystal (1992) states that "corpus linguistics deals with the principles and practice of using corpora in language study" (Crystal 1992:85). Mc Arthur adds some to the definition as "corpus linguistics studies data in any such corpus" (McArthur1992:266).

As said by most linguists such as Granger (1998), Kennedy (1998) et al., for many years linguists have been storing compilation and analysis of corpora in computerized databases. Cited compilation and analysis has guided a new art called "corpus linguistics". Kennedy supports the idea as "Over the last three decades the compilation and analysis of corpora stored in computerized databases has led to a new scholarly enterprise known as corpus linguistics" (1998:1). The goal of corpus linguistics is to examine large bodies of texts often with the aid of computers to find out what they have to say about language. Rochester (2001) states that this often involves describing language statistically, and he expresses its bent is definitely empirical, as opposed to Chomskian linguistics, according to him it is more mainstream.

There are various activities, which can be accounted within the scope of corpus linguistics, and as Kennedy (1998) notes in his study, it shows how quantitative analysis can contribute to linguistic description. When we look over the studies have been made so far on the subject we can see that most of them are on English language. The reason lies behind may be explained as its being the most widely taught and learned one both as a second and a foreign language all over the world. The answer to the question "what can be done with corpus?" has the fundamental importance in itself. "Corpus linguistics is not an end in itself but is one source of evidence for improving descriptions of structure and use of languages, and for various applications, including the processing of natural language by machine and understanding how to learn or teach a language" (Kennedy

1998:1). McEnery&Wilson (1996) also define corpus linguistics and its components as in the following order:

- Corpus Linguistics is now seen as the *study of linguistic phenomena* through large collections of *machine-readable texts*: corpora.
- These are used within a number of *research areas* going from the Descriptive Study of the Syntax of a Language to Prosody or Language Learning, to mention but a few.
- The use of *real examples* of texts in the study of language is not a new issue in the history of linguistics.
- Corpus Linguistics has developed considerably in the last decades due to the great possibilities offered by the processing of natural language with *computers*. The availability of computers and *machine-readable text* has made it possible to get data quickly and easily and also to have this data presented in a format suitable for analysis.
- Corpus linguistics is, not the same as mainly obtaining language data through the use of computers. Corpus linguistics is the study and *analysis of data* obtained from a corpus.
- The main task of the corpus linguist is not to find the data but to analyse it. Computers are useful, and sometimes indispensable, *tools* used in this process.

Resources and practices in the teaching of languages and linguistics are seem to be inclined to reflect the separation between the empirical and rationalist approaches. It can be seen that many textbooks contain only made-up examples and can be said that their descriptions are based upon perception or second-hand accounts. Other books, however, are explicitly empirical and use examples and descriptions from corpora or other sources of real life language data.

Corpus examples can be accepted significant in language learning because they expose students to the variety of sentences that they are likely to encounter when using the language in their life experience. Students who

are taught with course books which are written by using traditional methods contain sentences such as "*Jack went to London last year*" are often unable to analyse more complex sentences such as "*Successful businessmen who have long stayed at the summit in the business world know very well that success only comes to those who just will not accept defeat* (from the Spoken English Corpus)" (McEnery&Wilson 1996:4).

Apart from being a source of empirical teaching data, corpora can be used to look critically at existing language teaching materials. Kennedy (1987a, 1987b) has looked at ways of stating quantification and frequency in ESL (English as a second language) textbooks. Holmes (1988) has observed ways of expressing doubt and assurance in ESL textbooks, while Mindt (1992) has looked at future time terms in German textbooks of English. These studies can be said to have similar methodologies. They analyse the relevant structures or/and vocabularies, both in the sample textbooks and in standard English corpora and then they compare their discoveries between the two sets. As we look over the studies carried out on this matter we can see that there were considerable differences between what textbooks are teaching and how native speakers really use language as evidenced in the corpora. Some textbooks put the emphasis over important aspects of usage, or centre on less frequent stylistic choices without regard for more common ones. The general conclusion from these studies is that non-empirically based teaching materials can be misleading and that corpus studies are offered to be used to inform the production of material so that the more common choices of usage are given more attention than those which are less common.

There are many fields related with corpus linguistics. Many fields are becoming more aware of the benefits of this science day by day. "Work relevant for corpus linguistics is being done in many fields, including computer science and artificial intelligence, as well as in various branches of descriptive and applied linguistics" (Kennedy 1998:2). Besides, there are some debatable matters over the validity of a certain corpus in order to be

accepted as proper to be studied on. Corpus linguistics is also a field where undertakings are at a constant increase. Still a good many of features are found day-by-day and added to this field. Kennedy (1998) suggests that because corpus linguistics is a field where activity is increasing very rapidly and where there is as yet no magisterial perspective; he concludes even the very notion of what constitutes a valid corpus can still be controversial. Computer has also brought an incredible pace to this field of research. "The widespread use of the computer corpus has led to the development of a new discipline which has come to be called 'corpus linguistics', a term which refers not just to a new computer-based methodology" (Granger 1998:3). But as Leech puts it, to a new research enterprise, and says that "computer corpus is a new way of thinking about language" (Leech 1992:106). He points the importance of description and performance in corpus linguistics. "With its focus on performance (rather than competence), description (rather than universals and quantitative analysis, it can be seen as contrasting sharply with the Chomskyan approach" (Leech 1992:107).

Kennedy (1998) suggests that it also needs to be understood at the outset that not very use of computers with bodies of text is part of corpus linguistics. He gives an example for that and states as " the aim of Project Gutenberg to distribute 10.000 texts to 100 million computer users by the year 2001 is not in itself part of corpus linguistics although texts included in this ambitious project may conceivably provide textual data for corpus analysis" (Kennedy 1998:2). It is also possible to say as Fillmore does "the two kinds of linguists need each other. Or better, that that the two kinds of linguists, whenever possible, should exist in the same body" (Fillmore 1992:35). One of the utmost uses of the computer can be described as its playing a central role in corpus linguistics. "A first major advantage of computerization is that it liberates language analysts from drudgery and empowers [them] to focus their creative energies on doing what machines do" (Rundell and Stock 1992:14). On this point, Fillmore (1992) implies the benefit of studying corpus as one of the observations he made is that every corpus that he had had a chance to examine, however small, had taught

him facts that he couldn't imagine finding out about in any other way (Fillmore 1992:35).

Furthermore; Granger (1998:3) notes that "More fundamental, however, is the heuristic power of automated linguistic analysis, ie. Its power to uncover totally new facts about language." It is this aspect, rather than "the minoring of intuitive categories of description" (Sinclair 1998:202). Actually studying on corpus linguistics is not new. For many decades linguists have studied it, but there were no computers then and it has been performed by great difficulties. To reach huge size of texts was difficult and took great amount of time and effort. Therefore, it is widely accepted that to come to a reliable conclusion on a research was too tiring and consuming. McEnery&Wison (1996) note that Abercrombie's observations that corpus research is time-consuming, expensive and error-prone are no longer applicable thanks to the development of powerful computers and they add that software which is able to perform complex calculations in seconds, without error.

Computers provided the researchers with a great speed and easiness. Huge size of compilations can be evaluated dependably with the help of computer in a short time and without difficulty. It was also likely to make an unexpected mistake during the course of evaluating them. Kennedy (1998) expresses that corpus linguistics did not begin with the development of computers but he also adds that "there is no doubt that computers have given corpus linguistics a huge boost by reducing much of the drudgery of text-based linguistic description and vastly increasing the size of the databases used for analysis" (Kennedy 1998:2). It can be said that corpus linguistics is not a pointless process made up of just an automatic language description. It has an aim and a very important function as it is so all in other researches, which have been fulfilled in other fields so far, and it has a significant purpose. Kennedy supports the idea, as "It should be made clear, however that corpus linguistics is not a mindless process of automatic language description. As in all the researches being carried out around, linguistics also has a fundamental function and aim to be followed carefully and with great

attention" (Kennedy 1998:2). According to Biber, Conrad & Reppen (1994) there seems to be two general points in relation to the illustrative analyses. First, corpus-based analyses often give us an idea about that earlier conclusions based on perceptions are inadequate or incorrect- that is, the actual patterns of use in large text corpora often run in opposition to our expectations based on perception. Second, corpus-based analyses demonstrate that even the notion of core grammar needs qualification, and it seems to be because "investigation of the patterns of structure and use in large corpora reveals important, systematic differences across registers at all linguistic levels" (Biber, Conrad & Reppen 1994:169). Researchers such as Celce, Halliday, Sinclair et al. make use of the corpora to clarify the confusing and problematic points in many subjects beside, maybe more often nowadays, in language teaching processes. Teachers in secondary and high schools, and lecturers in universities, for example McEnery, Baker and Wilson from Edinburgh University benefit from or/and take the advantage of the results, which they get from the corpora obtained from a specific subject so as to improve their teaching process in their classrooms via sorting out the impediments before them. Linguists, when we have a look at their studies, use corpora to answer questions and solve problems. The use of corpora should be done efficiently and in a proper way. Although some of the instructors who have the assistance of learner's corpora, it seems probable that they still face problems in foreign and second language teaching process. One of the reasons for that can be the lack of analysing the corpora in an effective and fruitful way as well as in an acceptable way in order to have the impediment to be crossed out. Although computer and the sophisticated software developed for corpus-based studies increased and eased the process of analysing it cannot be said enough for the task to be accomplished. In other words, just by depending on the computer itself may not be adequate to sort out the problem. For example, Kennedy (1998) urges that sometimes manual work is needed to accompany to it. Because of that reason, he points out that there are still too many problems to be solved especially in foreign language and in second language teaching. He hints the

idea that linguists are experiencing the use of corpus linguistics widely in order to enhance the process of language teaching. Besides, he adds that to analyse the corpora in an efficient way is also of great importance. Thanks to the blend of manual and computer analysis by some researchers, some of the most revealing insight on language and language use seem to have come from this source. He expresses that it is now possible for researchers with access to a personal computer and off-the-shelf software to do linguistic analysis using a corpus, and to discover facts about a language which have never been noticed or written about previously. He points, the most important skill as not to be able to program a computer or even to manipulate available software, but to be able to ask insightful questions which address real issues and problems in theoretical, descriptive and applied language studies.

Many of the key problems and challenges in corpus linguistics are associated with the following questions conveyed from Kennedy (1998:3-4). The main questions are listed as:

- *How can we best exploit the opportunities which arise from having text stored in machine-retrievable form?
- *What linguistic theories will best help structure corpus-based research?
- *What linguistic phenomena should we look for?
- *What applications can make use of the insights and improved descriptions of languages which come out of this research?

In answering these and other questions he expresses that corpus linguistics has potential to provide solutions and new directions to some of the major issues and problems in the study of human communication. (Kennedy 1998:3-4).

2.2.1 Benefits Of Corpus Data

There are many benefits of corpus data in language studies and some others related to language itself. As conveyed from Kennedy (1998) some of the major ones can be stressed as:

1. Leech (1992) argues that the corpus is a more powerful methodology from the point of view of the scientific method as it is open to objective verification of results
2. Is language production really a poor reflection of language competence as Chomsky really argued? Labov (1969) showed that "the great majority of utterances in all contexts are grammatical". Here, it is not possible to say that all sentences in a corpus are grammatically acceptable, but it seems probable that the Chomsky's (1968:88) claim that performance data is 'degenerate' is an exaggeration (see Ingram 1989:223 for further criticisms of this view).
3. Quantitative data seems to be of use to linguistics. For example, Svartvik's (1966) study of passivisation used quantitative data extracted from a corpora. Elsewhere, all successful approaches to automated part-of-speech analysis rely on quantitative data from corpora. The proof of the pudding is in the eating.

2.2.2 The Scope Of Corpus Linguistics

Brown (1990) states that the field of "corpus linguistics" has explored the use of corpus-based techniques in to a variety of applications such as text retrieval, speech recognition, machine translation and ontology construction. He explicates that in each field the initial corpus-based experiments typically emphasize statistical analysis over linguistic theory, an approach which has led to some remarkable successes. "In machine translation, for example, early statistical approaches demonstrated performance that was competitive with that achieved by contemporaneous linguistically motivated approaches" (Brown *et al.* 1990:79-85). But McEnery (1996) adds that purely statistical approaches also introduce errors that no human would make because the techniques typically exploit term cooccurrence and a fairly complex set of factors actually interact to produce these cooccurrences. They exemplify that the present statistical models to be inadequate to capture some of these interactions, but significant

performance improvements can be achieved when appropriate linguistically motivated constraints are effectively integrated with the statistical analysis. "The experience to date with cross-language text retrieval suggests that similar improvements can be obtained for this aspect of corpus linguistics as well" (McEnery&Wilson 1996:16).

There is now mounting evidence from both corpus linguistics and cross-language text retrieval research that treating both individual words and multi-word phrases as the "terms" which are manipulated can significantly improve the effectiveness of cross-language techniques. This effect was observed by variety of linguists with regard to different aspects. For example van der Eijk (1993) observed this effect with adjacency-based phrases in automatic vocabulary construction experiments, on the other hand Hull and Grefenstette (1996) observed it with dictionary-based phrases in cross-language text recovery experiments, then again Radwan and Fluhr (1995) observed it with dictionary-based phrases in cross-language text retrieval experiments that integrated both dictionary-based and corpus-based techniques. McEnery&Wilson (1996) find it a particularly surprising result since the preponderance of the evidence on text retrieval in a single language indicates that multi-word phrases are of little use. They consider the basis for this effect is that presumably it is translation ambiguity which causes cross-language full-text retrieval systems to achieve lower retrieval effectiveness than their monolingual counterparts, and the use of phrases constrains this translation ambiguity to a significant extent.

Corpus linguistics seems to have also produced several useful tools for designers of cross-language text retrieval systems. According to Kikui (1996) one of them in particular can be needed in almost every cross-language text retrieval application is language identification. He adds that with the notable exception of CL-LSI, cross-language text retrieval techniques typically can require that the language in which the query and each document are expressed be known so that the correct processing can be applied. He concludes that today, language identification techniques with better than 95% accuracy are available.

Krieger (2003) states that the main focus of corpus linguistics can be defined as the discovering of patterns of authentic language use through analysis of actual usage. He argues that the aim of a corpus based analysis is probably not to generate theories of what is possible in the language, such as Chomsky's phrase structure grammar which can generate an infinite number of sentences but which does not account for the probable choices that speakers actually make. And adds that corpus linguistics' only concern is likely to be the usage patterns of the empirical data and what that reveals to us about language behaviour.

Corpus linguistics can be also defined as "compilation of the texts, which is used for the source of a study to make analysis on and finding evidence according to the specific purpose of the researcher. Corpus linguistics, for the most part deals with the depiction and clarification of the nature, and with particular matters such as language acquisition and deviation" (McEnery&Wilson 1993:233). Kennedy defines the notion as "Corpus linguistics is based on bodies of text as the domain of study and the source of evidence for linguistic description and argumentation" (Kennedy 1998:7,8). He adds that it has also embody methodologies for linguistic description in which quantification of the distribution of linguistic items is part of the research activity. It can also be added that corpus linguistics is more related with the performance than competence.

Observation of language has of great importance in the field. Leech notes "the focus of study is on performance rather than competence, and on observation of language in use leading to theory rather than vice versa" (Leech 1992:107). The main focus of corpus linguistics is to discover patterns of authentic language use through analysis of actual usage. Krieger (2003) conveys the aim of a corpus based analysis as it is not to generate theories of what is possible in the language, such as Chomsky's phrase structure grammar which can generate an infinite number of sentences but which does not account for the probable choices that speakers actually make. He adds that corpus linguistics' only concern is the usage patterns of the empirical data and what that reveals to us about language behaviour

According to McEnery&Wilson (1996) it can be deceptive to suggest that corpus linguistics is a theory of language in competition with other theories of language such as transformational grammar, or even more that it is a new or separate branch of linguistics as some other noteworthy linguists do so. In fact, they add, corpus linguistics has a very close relationship with other fields of linguistics that it has the proving speciality in itself for the other studies to make use of. They conclude that every researchers need evidence to exploit in order to put an idea forward they look for a proper and appropriate source to get the suitable information from. Corpus linguistics introduces that kind of information to the researchers as a source of evidence. So far, when we make a brief research it can be seen that they have tried to get the information from other weaker and insufficient kinds of sources such as observing the spoken and written form of languages, nature, rudiments and so on (McEnery&Wilson 1995:259-274)

Kennedy (1998) points the matter as 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. He argues that 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. He suggests, in the case of corpus-based research, the evidence is derived directly from texts. Researchers get the information depending on the data obtained from the texts. "In this sense corpus linguistics differs from approaches to language which depend on introspection for evidence" (Kennedy 1998:7).

Malinowski (1935), in order to clarify the cited idea, wrote about the paradigm shift in his well-known work, *Coral Gardens and their Magic*, which he considered was necessary in the linguistics of the day. Reported from Kennedy (1998), Malinowski suggests that the neglect of the obvious has been fatal to the development of scientific thought. The false conception of language as a means of transfusing ideas from the head of the speaker to that of the listener has, he considers, largely vitiated the

philological approach to language. According to him, the view set forth here is not merely academic. He defines that it compels us, as we shall see, to correlate other activities, to interpret the meaning-text; and this means a new departure in the handling of linguistic evidence. As a conclusion, it will also force us to define meaning in terms of experience and situation.

It can be said that corpus linguists are concerned besides what words, structures or uses are possible in a language, they are also interested in the probability of what is likely to be occurred or what is probable in a language use as well. The linguists' aim at using the information they get on teaching methodology. It can be said that benefiting from corpus linguistics is being widely understood as the time passes by. What is probable or what is necessary or insufficient in language teaching process can be determined through corpus-based linguistic studies and analysing learner's corpora excessively. In other words, both the linguists and the researchers try to come to a conclusion about the needs and the necessities of teaching process. It is possible to find a great deal of evidence from many different kinds of sources in order to justify a theory or a probability on a particular subject. "The use of a corpus as a source of evidence however is not necessarily incompatible with any linguistic theory, and progress in the language sciences as a whole is likely to benefit from a judicious use of evidence from various sources for example from the texts, introspection, elicitation or other types of experimentation as appropriate" (Kennedy 1998:8). We are supposed to make an examination in order to be sure of the facts of any idea or hypothesis. Kennedy (1998) sees judicious evaluation as a vital for a correct conclusion. According to his ideas, any scientific enterprise must be empirical in the sense that it has to be supported or falsified on evidence and, in the final analysis, statements made about language can be based on the introspective judgement of speakers of the language or on a corpus of text. "The difference lies in the richness of the evidence and the confidence we can have in the generalizability of that evidence, in its validity and reliability" (Kennedy

1998:8). He also adds, the evidence should be dependable in order the conclusion to be approved. As a conclusion he suggests, the boundaries, therefore, between corpus-based description and argumentation and other approaches to language description are not rigid, and linguists of varied theoretical persuasions now use corpora for evidence which is complementary to evidence obtained from other sources.

Krieger (2003) notes that one frequently overlooked aspect of language use which is difficult to keep track of without corpus analysis is register. Register consists of varieties of language which are used for different situations. Language can be divided into many registers, which range from the general to the highly specific, depending upon the degree of specificity that is sought. A general register could include fiction, academic prose, newspapers, or casual conversation, whereas a specific register would be sub-registers within academic prose, such as scientific texts, literary criticism, and linguistics studies, each with their own field specific characteristics. Corpus analysis reveals that language often behaves differently according to the register, each with some unique patterns and rules.

In addition to cited usage and benefits of corpus linguistics, it has some other aspects of practice, which differentiates it from the others. Kennedy (1998) points out the concern of corpus linguistics as it is like all linguistics, is concerned primarily with the description and explanation of the nature, structure and use of language and languages and with particular matters such as language acquisition, variation and change. He proposes, corpus linguistics has nevertheless developed something of a life of its own within linguistics, with a tendency sometimes to focus on lexis and lexical grammar rather than pure syntax. He concludes this as partly a result of using methodologies such as concordancing where the contextual evidence available in a single line of wide-carriage computer printout of 130 characters. He also notes that it is sometimes too limited for the analysis of syntax or discourse.

To work in corpus linguistics is nowadays associated with several quite different activities. Scholars working in the field tend to be identified with one or more of them. There seems to be four groups of scholars according to their activities and objectives. The first group of researchers, as conveyed from Kennedy (1998) consists of corpus makers or compilers. These scholars are concerned with the design and compilation of corpora, the collection of texts and their preparation and storage for later analysis. He portrays the second group of researchers to have been concerned with developing tools for the analysis of corpora. And he adds, important contributions to software development especially for the syntactic analysis of corpora have been associated particularly but not exclusively with researchers in computational linguistics. "These researchers have been concerned with the use of corpora to develop, among other things, algorithms for natural language processing and the modelling of linguistic theories" (Kennedy 1998:8). He regards the third group of researchers as descriptive linguists whose main concern has been to make use of computerized corpora to describe reliably the lexicon and grammar of languages, both of the linguistic systems we use and our likely use of those systems. According to him, it is the probabilistic aspect of corpus-based descriptive linguistic studies which especially distinguishes them from conventional descriptive fieldwork in linguistics or lexicography. He suggests, that is, corpus-based descriptive linguistics is concerned not only with what is said or written, where, when and by whom, but how often particular forms are used. He expresses that the measurement of the distribution of words and grammar has encouraged new ways of studying the linguistic basis of varieties of language. And he also adds, corpus provides contexts for the study of meaning in use and, by making available techniques for extracting linguistic information from texts on a scale previously undreamed of, it facilitates linguistic investigations where empiricism is text based. Kennedy (1998:9) describes the fourth area of activity as the one, which has been among the most innovative outcomes of the corpus revolution. And he concludes that it has been the exploitation of corpus-based linguistic description for use in a variety of applications such as language

learning and teaching, and natural language processing by machine, including speech recognition and translation. "At the present time in corpus linguistics, some researchers seem to focus on issues in corpus design, others on methods for text analysis and processing, and still others, probably the majority, on corpus-based linguistic description and the application of such descriptions" (Kennedy 1998:9).

Although the scope of corpus linguistics may be defined in terms of what people do with corpora, Kennedy (1998) suggests that it would be a mistake to assume that corpus linguistics is simply a faster way of describing how a language works, or is about the nature of linguistic evidence. He continues as "Analysis of a corpus by means of standard corpus linguistic research software can and frequently does reveal facts about a language which we might never previously have thought of seeking" (Kennedy 1998:9).

According to Kennedy (1998), corpus linguistics goes beyond the use of corpora as a source of evidence in linguistic description. He adds that it also revives and carries on a concern of some linguists with the statistical distribution of linguistic items in the context of use. "From the 1920s there was, especially in the United States and the United Kingdom, a tradition of word counting in texts in order to discover the most frequent, and arguably therefore the most pedagogically useful, words and grammatical structures for language teaching purposes" (Kennedy 1998:10).

With a corpus-stored computer, it can be said that it is easier to find, sort and count items. It seems to be true for both as a basis for linguistic description and for addressing language-related issues and problems. It should not be seen surprising, therefore, that a wide range of research activities seem to have come to be within the scope of corpus linguistics. According to Kennedy (1998), analyses can contribute to the making of dictionaries, word lists, descriptive grammars, diachronic and synchronic comparative studies of speech varieties, and to stylistic, pedagogical and other applications. "With appropriate software it is easy to study the distribution of phonemes, letters, punctuation, inflectional and

derivational morphemes, words (as variously defined), collocations, instances of particular word classes, syntactic patterns, or discourse structures" (Kennedy 1998:11).

The scope of the corpus linguistics varies. There may be lots of scopes as Kennedy (1998) notes the following:

"The scope and current concerns of a field of scholarship can sometimes be seen or defined through the topics which make up conference programmes and the content of specialist journals. In the 1990s the topics which appear on conference programmes and in journals which cover corpus linguistics include improved ways of annotating corpora, the tagging of parts of speech and the senses of polysemous word forms, improved automatic parsing, identification of collocations, phraseological units and discourse structure, text categorization, research methodology in the face of more and bigger corpora, and the application of this work in lexicography, syntactic description, translation, speech and handwriting recognition, and language teaching. Educational applications are increasingly on the agenda. At Lancaster University in 1994 and 1996 the pedagogical significance of electronic corpora was the subject of conferences on the teaching of linguistics and the teaching of languages."

In March 1993, a Georgetown University Round Table meeting in Washington, DC, on corpus-based linguistics identified the following topics as those in particular need of investigation and dissemination at a time when linguistics was returning to more text-based approaches to language. Kennedy (1998) lists them as in the following:

- 1- the design and development of text-speech corpora
- 2- tools for searching and processing on-line corpora
- 3- critical assessments of on-line corpora and corpus processing tools
- 4- methodological issues in corpus-based analysis applications and results in linguistics and related disciplines, including language teaching, computational linguistics, historical linguistics, discourse analysis and stylistic analysis.

2.2.2 Some Major Achievements Of Computer Corpus-Based Scholarship

We can imagine that with the help of computer the study of corpus has gained great improvement. To begin with, it would be beneficial to give a brief definition of computer corpus but widely accepted at the same time by other linguists such as Kennedy; Granger, Quirk et al. In this respect, Crystal states that "A computer corpus is a large body of machine-readable texts" (Crystal 1992:85). The scope of computer corpus-based scholarship can also be measured by some of its achievements. According to Kennedy (1998) in lexicography the revision of the *Oxford English Dictionary*, its publication in electronic form on CD-ROM and the publication of new learners' dictionaries of English by other major publishers were all based on corpora. The completion of the 100-million-word *British National Corpus* in 1994 set a new standard in corpus design and compilation. Another important international standard set in corpus preparation and formatting has been in the gradual adoption of the Standard Generalized Markup Language (SGML) through the Text Encoding Initiative (TEI) (Kennedy 1998:12). According to Francis (1992), the third addition of Webster's *New International Dictionary* published in 1961 had available a corpus of over 10 million citation slips to validate and illustrate the meanings and uses of the almost half a million headword entries which it contained. "Webster's Third was probably the last major English dictionary to be completed without an electronic database" (Francis 1992:22).

According to Kennedy (1998), some well-known corpora from the beginnings of the computer age are appeared to be the Brown corpus of written American English and the Lancaster-Oslo/Bergen corpus of written British English. The Brown corpus was compiled in the 60's, its British counterpart in the 70's. Both consist of around one million tokens (i.e. words, counted every time they appear). The London-Lund corpus is another corpus of British English created around that time, but this corpus is different from the Brown and the LOB in that it exclusively contains transcripts from spoken material, collected at the Survey of English Usage at University College London. The London-Lund corpus, the Brown corpus, the LOB and other

corpora are now available on CD-ROM as the ICAME collection of English texts. The International Computer Archive of Modern and Medieval English (ICAME), situated at Bergen in Norway, offers a wealth of information on these corpora. Nowadays we shall find modern corpora, which differ from those named above. In the first place, thanks to technological advancements, in particular faster and more powerful computers, the size of modern corpora is vastly greater. The British National Corpus, for example, consists of around 100 million words, i.e. it is a hundred times larger than the Brown corpus! Also, corpus designers today usually try to include as much spoken material as is financially and technically feasible. As we know that creating transcripts of conversations is a time-consuming and expensive process. Three examples of modern corpora are the British National Corpus, which I have just mentioned, the International Corpus of English and the Bank of English, situated at Birmingham University. It is possible to view them as listed by Kennedy (1998:11-67) in the following lines:

- The Bank of English was initiated in 1991 by COBUILD (a division of HarperCollins publishers) and the University of Birmingham. The main purpose of the Bank of English is and has been to provide a textual database for the compilation of dictionaries and for language studies. The Bank of English is a monitor corpus (i.e. new material is constantly added). By now the corpus has got a size of more than 320 million words.
- The British National Corpus was compiled by a consortium of British publishers, of academic institutions such as Oxford University Computing services, Lancaster University's Centre for Computer Research on the English language and the British Library. It is now a 100 million word corpus of modern British English, both written and spoken, including everyday conversations. It is available

on CD-ROM for research purposes; we have got a copy at our department.

- The International Corpus of English (ICE) will ultimately be a collection of 1,000,000 word corpora from each country or region where English is spoken as a first language. The corpus consists of a written and a spoken component. The Survey of English Usage, situated at University College London, is responsible for this project. The home page of the Survey provides information on a variety of research projects, including the International Corpus of English (ICE)

2.3 Types of corpora

In this session we will examine the roles, which corpora may play in the study of language. The importance of corpora to language study can be said to align to the importance of empirical data. Empirical data can enable the linguist to make objective statements, rather than those which are subjective, or based upon the individual's own internalised cognitive perception of language. Empirical data also allows us to study language varieties such as dialects or earlier periods in a language for which it is not possible to carry out a rationalist approach. It is suggested that "the created corpora were supposed to represent proportionally the qualities of the language as a whole ('general' or 'balanced') corpora" (Aston 1997:5).

According to McEnery&Wilson (1996), it is important to note that although many linguists such as Kennedy, Granger et al., may use the term "corpus" to refer to any collection of texts, when it is used here it refers to a body of text which is carefully sampled to be maximally representative of the language or language variety. Corpus linguistics, proper, is supposed to be seen as a division of the activity within an empirical approach to linguistics. They state that although corpus linguistics entails an empirical approach, empirical linguistics is not always supposed to entail the use of a corpus. In

the following pages I intended to consider the roles which corpora use may play in a number of different fields of study related to language. I focused on the conceptual issues of why corpus data are important to these areas, and how they can contribute to the advancement of knowledge in each, providing real examples of corpus use. In view of the huge amount of corpus-based linguistic research, the examples are possibly needs to be selective.

The definition of a corpus as a collection of texts in an electronic database can beg many questions for there are many different kinds of corpora. In some dictionary definitions it is suggested that corpora necessarily consist of structured collections of text specifically compiled for linguistic analysis, that they are large or that they attempt to be representative of a language as a whole. But, according to some of the linguists such as Kennedy (1998), Granger (1998), Quirk (1997) et.al., this is not necessarily so. For example according to Kennedy (1998), not all corpora, which can be used for linguistic research were originally seem to be compiled for that purpose. Historically it is not even the case that corpora are necessarily stored electronically so that they can be machine-readable, although this is nowadays the type. He suggests that electronic corpora consist of continuous text samples taken from whole texts; they can even be made up of collections of citations. At one extreme it is assumed that an electronic dictionary may serve as a kind of corpus for certain types of linguistic research while at the other extreme "it is accepted as a huge unstructured archive of texts may be used for similar purposes by corpus linguists" (Kennedy 1998:3).

Corpora seems to have been compiled for many different purposes, which in turn is assumed to influence the design, size and nature of the individual corpus. But Kennedy defines the reason for collecting corpora as "A major reason for compiling linguistic corpora is to provide the basis for more accurate and reliable descriptions of how languages are structured and used" (Kennedy 1998:88). Some current corpora intended for linguistic research have been designed for general descriptive purposes, that is to say that they have been designed so that they can be examined or investigated to answer questions at various linguistic levels on the prosody, lexis, grammar,

discourse patterns or pragmatics of the language. Other corpora have been designed for specialized purposes such as discovering which words and word meanings should be included in a learners' dictionary; which words or meanings are most frequently used by workers in the oil industry, economics, education, engineering of different kinds; or what differences there are between uses of a language in different geographical social, historical or work-related contexts (Kennedy 1998:4).

A distinction is sometimes made between a corpus and a text archive or text database. The difference is tried to be clarified whereas a corpus designed for linguistic analysis is normally a systematic, planned and structured compilation of text, an archive is a text repository, often huge and opportunistically collected, and normally not structured. It is generally the case, as Leech suggested that "The difference between an archive and a corpus must be that the latter is designed or required for a particular 'representative' function" (Leech 1991:11). It is nevertheless not always easy to see clearly what a corpus is representing, in terms of language variety (Kennedy 1998:4).

Especially it can be frequently seen in today's ELT methodology studies corpora have also been widely used in the teaching of linguistics. Kirk (1994) requires his students in his study to base their projects on corpus data, which they are expected to analyse in the light of a model such as Brown and Levinson's politeness theory or Grice's co-operative principle. According to McEnergy&Wilson (1996), in taking this approach, Kirk can be seen as using corpora not only as a way of teaching students about variation in English but also to introduce them to the main features of a corpus-based approach to linguistic analysis.

A further application of corpora in this field can be stressed is their role in computer-assisted language learning. Recent work carried out by McEnergy and Wilson (1993) at Lancaster University has looked at widely in their study for the role of corpus-based computer software for teaching undergraduates the fundamentals of grammatical analysis. They introduced a software that helps doing the work more efficiently and easier.

McEnergy&Wilson (1996) note that according to its definition, this software 'Cytor' reads in an annotated corpus (either it can be a part-of-speech tagged or parsed) one sentence at a time, hides the annotation and asks the student to annotate the sentence him or herself. Students can call up help in the form of the list of tag mnemonics, a frequency lexicon or concordances of examples. McEnergy, Baker and Wilson (1995) carried out an experiment over the course of a term to determine how effective Cytor was at teaching part-of-speech learning by comparing two groups of students - one who were taught with Cytor, and another who were taught via traditional lecturer-based methods. In general, it was seen that the computer-taught students performed better than the human-taught students throughout the term.

Databases which are made up not of samples, but which constitute an entire population of data, may consist of a single book (e.g. George Eliot's *Middlemarch*) or of a number of works. Kennedy adds that "These corpora may be the work of a single author (e.g. the complete works of Jane Austen) or of several authors (e.g. medieval lyrics), or all the editions of a particular newspaper or magazines in a given year" (Kennedy 1998:4). It is also possible to see that some projects have assembled all the known available texts in a particular genre or from a particular historical period. Some of these databases or text archives are seemed to be very large indeed, and if we make a wide ranged research we can see that they have rarely yet been used as corpora for linguistic research, but it does not mean that they will not be used in the future. As the use of corpora for the researches widens day by day, it is very likely to see them as the subject of a specific research in the following days by any interested researchers. "In many respects it is thus the use to which the body of textual material is put, rather than its design features, which define what a corpus is" (Kennedy 1998:4).

Kennedy (1998) points out that a corpus can constitute an empirical bases not only for identifying the elements and structural patterns which make up the systems we use in a language, but also for planning out our use of these systems. There seems to be various ways of using corpora. A corpus can be analyzed and compared with other corpora or parts of corpora to

study variation. Most importantly, it can be analysed distributionally to show how often particular phonological, lexical, grammatical, discoursal or pragmatic features occur, and also where they occur.

When we check through the early 1980s, it was possible to list on a few fingers the main electronic corpora, which a small band of devotees had put together over the previous two decades for linguistic research. Thus there seemed to be not much detailed studied corpora in that period. Kennedy (1998) notes that "these corpora were available to researchers on a non-profit bases, and were initially available for processing only on mainframe computers. "The development of more powerful micro computers from the mid-1970s and the advent of CD-ROM in the 1980s made corpus-based research more accessible to a much wider range of participants" (Kennedy 1998:4).

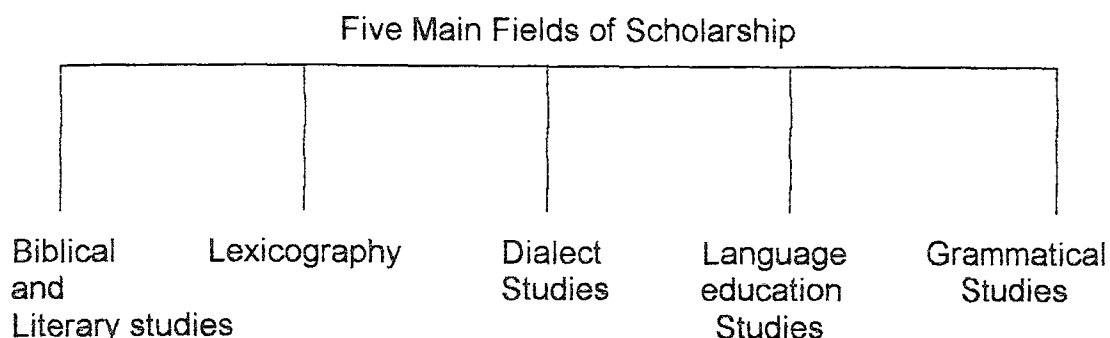
Provided that you take a tour among the sites related with corpus linguistics researches on the Internet or in the libraries, it is easy for a researcher to see that by the 1990s, there were many corpus-making projects in various parts of the world. Lancashire (1991) shows the huge range of corpora, archives and other electronic databases available or being compiled for a wide variety of purposes. Some of the largest corpus projects seem to have been undertaken for commercial purposes can be seen to have compiled by dictionary publishers. That is so probably because other projects in corpus compilation or analysis are on a smaller scale, they do not necessarily become well known. Still on the other hand, "undertaken as part of graduate theses or undergraduate projects, they enabled students to gain original insights into the structure and use of language" (Kennedy 1998:5).

2.3.1 Pre-Electronic Corpora

Corpus-based research is often assumed to have begun in the early 1960s with the availability of electronic, machine-readable corpora. However, before then there seems to have been a considerable tradition of corpus-based linguistic analysis of various kinds, which are occurring in five

main fields of scholarship. Kennedy (1998) assembles them under 5 headings:

Fig.2



Reference: Kennedy G. An Introduction to Corpus Linguistics. Longman 1998. London.p.13)

2.3.1.1 Biblical And Literary

There have been many corpus researches made over the bible. It has been noticed that some words were used more often than the others. The significance of the words frequently used has been taken into consideration while coming to a conclusion about some specific concerns. Some of the major studies and the plot of the researches in this field are mentioned below by Kennedy (1998:13):

"One of the first significant pieces of corpus-based research with linguistic associations involved using the Bible as a corpus on which to base commentaries or criticism. From the 18th century, lists and concordances of words used in the Bible were made in an attempt to show that the various parts of the Bible were factually consistent with each other. Alexander Cruden, a London bookseller, proofreader, morals campaigner and prison reformer, born in Aberdeen in 1701, produced the most famous of these for the Authorized (King James) Version of the Bible. First published in 1736, Cruden's Concordance was a monumental piece of laborious scholarship which went through 42 editions-even before 1879. It included concordances not only for what the author considered to be the major content words in the Bible but also some function words such as how, you, he, once, between (but not on, she or with) and certain collocations such as how long, how many, how much, how much less, how much more, how often, all the nations."

2.3.1.2 Lexicographical

Corpora for lexicography cannot be considered as a new field of study. It is possible to notice many kinds of researches had been accomplished in the past if we make a review of the literature on this subject. Plenty of studies mention about this subject that are available on the Internet nowadays and easy to obtain information through this way. Especially, when we have a discussion which is based on pre-electronic corpora, Francis (1992) reminds us that the use of corpora for lexicography goes back at least to the early 17th century.

Samuel Johnson, built on the work of his pre-decessors over the previous 150 years and recorded on slips of paper a large corpus of sentences from "writers of the first reputation" to illustrate meanings and uses of English words. Johnson worked with some six assistants to assemble over 150,000 illustrative citations for the approximately 40,000 headword entries in his *Dictionary of the English Language*. It is likely that the corpus formed from these sample citations came to well over one million words (Kennedy (1998:14).

We can also notice that the *Oxford English Dictionary (OED)* was similarly corpus-based. When the twelfth and final volume of the OED was published in 1928, as Kennedy reports, "it was the culmination of 71 years of sustained work on a corpus of the canon of mainly literary written English from about AD 1000" (Kennedy 1998:14). As it is known, working manually with text is enormously time-consuming and labour intensive as the first editor of the OED, James Murray, discovered. Kennedy (1998) conveys Murray his successor Henry Bradley, and two editors involved in earlier preparation had all died before the work was completed. This is perfectly a good sample to emphasize how difficult and takes long time to complete a study in those years. Some 2,000 volunteer readers were said to have collected about five million citations totalling perhaps 50 million words to illustrate the meanings and uses of the 414,825 entries which appeared in the dictionary. The total scope of the task managing this vast amount of

information without a computer and the effect on Murray's family can be judged from the commentary on the project written by his grand daughter (Murray 1977). "Over 350 volumes of early manuscripts from *Beowulf* to the 17th century were edited and published for the first time by the Early English Text Society beginning in 1864, initially to make accessible part of the corpus from which the citations for the dictionary were to be extracted" (Kennedy 1998:15).

It is suggested by Kennedy (1998) that subsequent supplements to the *Oxford English Dictionary* were able to make use of several million more citation slips which is conveyed to have been compiled since the first volume was published in 1884. This huge corpus for lexicographical research, and its analysis and description is regarded as one of the greatest pieces of linguistic scholarship in English or any language, and Kennedy states that "it was done without the contribution of speed, accuracy and exhaustiveness, which the computer can bring to the task" (1998:15).

Francis (1992) notes that another great corpus of citations may be noted is Noah's Webster's *An American Dictionary of the English Language*, which was published in 1828. "The third edition of Webster's *New International Dictionary* published in 1961 has available a corpus over 10 million citation slips to validate and illustrate the meanings and uses of the almost half a million headword entries which it contained" (Francis 1992:22).

2.3.1.3 Dialect

Kennedy (1998) informs that alongside the lexicographical work on the OED and Webster's, an interest in linguistic variation in regional dialects also led some 19th century linguists to assemble and interpret corpora of various kinds. Most of the work was on lexical variation in the choice of words for particular concepts, and possible variant forms of particular words, both in spelling and pronunciation. "The *English Dialect Dictionary* (Wright 1898-1905) and *The Existing Phonology of English Dialects* (Ellis, 1889) were two monumental results of specialized corpus-based studies of lexical variation in dialects of the United Kingdom" (Kennedy 1998:15-16).

2.3.1.4 Language Education

When we look through the studies accomplished during the previous century, it can clearly be seen that some of the most influential corpus-based research in the first half of the 20th century had a pedagogical purpose. Thorndike (1921) compiled a corpus of 4.5 million words from 41 different sources to make a word frequency list which was initially intended to lead to better curricula materials for teaching literacy to native speakers of English in the United States. Another work was that during the 1930s, Thorndike's corpus was updated in collaboration with Lorge (Thorndike & Lorge, 1944). "The lexical analysis of this corpus and the published works, which derived from the research of Thorndike and Lorge were enormously influential for the teaching of English in many parts of the world over the next 30 years" (Kennedy 1998:16).

Despite the fact that Thorndike's corpus work had been undertaken with learners of English in United States schools, it can be seen that an earlier corpus was produced in Germany to gather statistical information on the use of words and letters of the German language in order to improve the training of stenographers, who, of course, in an age before dictaphones, tape recorders and photocopiers, had a vital role recording the daily discussions and decisions made in government and business (Kennedy 1998:16). There have been other works those following Thorndike's studies. It is also possible to notice that Thorndike's work was regarded as a pioneering work on this matter. After that study, "other pre-electronic corpora were assembled for teaching purposes in several countries for languages including Dutch, French, German, Italian, Latin, Russian and Spanish" (Kennedy 1998:16). McEnery&Wilson (1996) informs that the studies of child language in the diary studies period of language acquisition research (roughly between 1876 and 1926) were seemed to be based on carefully composed parental documentations recording the child's locutions. It can be seen that these primitive corpora are still used as sources of normative data in language acquisition research today, for example, Ingram

(1978). Corpus collection continued and diversified after the diary studies period: large sample studies covered the period roughly from 1927 to 1957 analysis was gathered from a large number of children with the express aim of establishing models of development. "Longitudinal studies have been dominant from 1957 to the present - again based on collections of utterances, but this time with a smaller (approximately 3) sample of children who are studied over long periods of time, [for example, Brown (1973) and Bloom (1970)]" (Kennedy 1998:16).

According to Kennedy (1998), other corpora, in order to study aspects of the lexico-grammar of English with pedagogical purposes, were compiled in the 1950s and 1960s. The most important one among these was a corpus of about half a million words of written British English assembled in India by H.V. George and his colleagues in Hyderabad from novels, place and non-fiction sources including newspapers. "Some of the analyses carried out on this corpus and other smaller ones compiled for pedagogical purposes in Japan, North America and in Europe" (Kennedy 1998:16).

2.3.1.5 Grammatical

It seems that corpora have long been one of the sources for the compilation of descriptive grammars. In the first half of the 20th century, newspapers and novels were used as a source of examples to illustrate grammatical features or constructions by the major descriptive grammars of English. "The work of Jespersen (1909-49) is probably the best known, but Kruisinga (1931-32) and Poutsma (1926-29) are among those who produced grammars based on somewhat informal corpora of various sizes rather than on introspection" (Kennedy 1998:17).

Afterwards, Fries can be said to have assembled more structured and systematic corpora for grammatical studies in the United States. Fries, in his *American English Grammar* (1940), used a corpus of letters written to the US government by people of different educational and social backgrounds in order to describe the difference of usage among social class people in the

society. "He recorded, for example, the incidence of the past participle *done* used as a preterite among less well-educated correspondents, and noted that they *sung* or *it sunk* were commonly used in standard educated American English of the time" (Kennedy; 1998:17).

The *Survey of English Usage (SEU) Corpus* (Quirk 1968) is considered as the most significant pre-electronic corpus which was assembled particularly for grammatical description. "When Randolph Kirk founded the Survey in 1959, the aim was to collect 200 samples, each of about 5,000 words, representative of spoken and written British English, to form a corpus of one million words which could be used as a basis for describing the grammar and usage of educated adult native speakers of British English" (Kennedy 1998:17).

2.3.2. Electronic Corpora

2.3.2.1 General Corpora

It is possible to say that different kinds of text bodies, which are available in machine-readable form, can be used as corpora for linguistic analysis. These corpora can be compiled in many different ways according to the purpose they were carried out. Their representativeness, organization and format are also important. When we make a research through the corpus linguistics literature, we can see that sometimes several different types of electronic corpora are distinguished.

So far in corpus linguistics studies, some corpora have been brought together basically to make available a text base for indefinite linguistic research. According to Kennedy (1998:19) we call such corpora as "general corpora". The linguists to analyse in order to find answers to particular questions use these corpora, which are consisting of a body of texts. They try to find answers to the questions about the vocabulary, grammar or discourse structure of the language. Some of the questions may be like the ones as: "Are speakers or writers of American English, for example, more likely to say or write *in certain circumstances* rather than *under certain*

circumstances? Is British usage the same? What prosodic, lexical or syntactic devices are used most frequently to initiate turn-bidding interruptions in discourse?" (Kennedy 1998:20). The SEU Corpus can be pointed out to be an example of a general-purpose corpus which has been used especially for research on grammar. "A general corpus is typically designed to be *balanced*, by containing texts from different genres and domains of use including spoken and written, private and public" (Kennedy 1998:20). It is also possible to include texts according to their apparent or possible typicality or influence in a language. By doing that, a greater proportion of samples from wide-circulation newspapers or best-selling novels might be included to our study. In addition to those from some other sources, it can be beneficial to the researcher. "General or balanced corpora are sometimes referred to as *core corpora*, which can be used as a basis for comparative studies" (Kennedy 1998:20).

2.3.2.2 Specialized Corpora

Kennedy (1998) explains the meaning of specialized corpora as the ones which are designed with particular research projects in mind are sometimes called *specialized corpora*. "Corpora assembled by major commercial publishers as sources of word frequency data and citations for the compilation of modern dictionaries are of this kind" (Kennedy 1998:20). Jones adds in his study that specialized corpora have also been assembled to study topics as varied as child language development (Carterette & Jones, 1974) or the English used in Petroleum geology exploration, drilling and refining (Zhu 1989). "Corpora may be even more specialized and consist of examples of people disagreeing with each other in radio interviews, or of teachers' directives in high school classrooms" (Kennedy 1998:20). Leech (1992:112) has described the development of training *corpora* and *test corpora* as specialized corpora to facilitate the building of models of language and of language processing. According to Kennedy (1998), major types of specialized corpora include those compiled for

studies of regional or sociolinguistic variation. *Dialect corpora*, *regional corpora*, *non-standard corpora* and *learners' corpora* come into this category

Spoken language is undoubtedly the commonest form or use of language, nevertheless it can be seen that most of the corpus-based studies of English have so far been based on the analysis of *written corpora* because studying on spoken corpora involves complex phonetic features, is much more time consuming, complicated and expensive to undertake. Kennedy (1998) explains that corpora may also be broadly characterized according to the way they represent a language. On the other hand, he also adds that a corpus can consist of a total statistical population. He exemplifies that a corpus consisting of the complete works of Charles Dickens or of all the edition of the *New York Times* between 1940 and 1990 represents a total population of text. He opposes the idea that the corpus is the representative of an entity. Moreover, he argues that it is that entity. On the other hand, he says that "a one-million-(or 100-million-) word linguistic corpus which seeks to be representative of a language or language variety in a particular year (e.g. written or spoken New Zealand English in 1986) should ideally be sampled systematically from all the spoken and written English produced in that country in that year so that it represents the language as fairly as possible" (Kennedy 1998:20). Kennedy believes that such a *sample-text corpus* (that is the way he names it) is designed to be representative sample of the total population of discourse. He judges that population is not necessarily 'the language as a whole' and claims "texts can be sampled from sub-populations, according to regions, genres, or groups of users (e.g. school-children, women, journalists or immigrants)" (Kennedy 1998:21).

According to Kennedy (1998) sample text corpus may consist of complete texts, which are sampled from a population of complete texts. He conveys them to be sometimes called as a "*full-text corpus*" or he adds it may consist of samples of a specified size taken from complete texts. It seems that particular corpora tend to be suitable for particular types of analysis and as said by Kennedy (1998) some corpora are basically not suitable for certain kinds of research. In order to clarify it, he expresses that for stylistic or

discourse studies, for example, a corpus which consists of 2,000-word samples extracted from many texts may not be able to capture reliable the internal structural characteristics of full texts where introductory and concluding sections may be expected to have different linguistic features. He claims such studies may require the use of full text corpora. Kennedy (1998) points out that corpora can consist of the 'raw' orthographic text of transcribed speech or of writing and contain a minimum of annotations. He adds these annotations that identify such divisions as paragraphs and line numbers. According to him, it is also possible to annotate the original "raw" text or pre-processed linguistically to show the word class of each word in the text by means of a grammatical *tag* or label which is attached to each word. What is more he states, corpora can also be *parsed* to show the sentence structure and the function in the sentences of the different word classes. It can be seen that the *tagging* and *parsing* of corpora can now be undertaken automatically with increasingly high levels of reliability.

From the studying carried out so far we can notice that some corpora have been compiled containing vast amounts of text which are added to, often opportunistically, and which are not necessarily balanced and structured, so that text does not systematically and proportionately come from particular genres or registers. Conveyed from Kennedy (1998) new texts, from daily newspapers and other sources, may actually replace material which was in the corpus earlier. "These *dynamic corpora*, sometimes also referred to as *monitor corpora*, are open-ended language 'banks' which are limited only by the financial resources and technology needed to maintain them" (Kennedy 1998:21). He adds that the types of corpora mentioned so far have been *synchronic corpora*. A synchronic corpus is an attempt to represent a language or a text type at a particular time (e.g. it may contain written texts of American English only published in 1961). He exemplifies, a *diachronic corpus*, on the other hand, represents a language over a period of time. (e.g. it may contain English texts covering the period from about AD 700 to AD 1700) and can be used, among other things, for studying language change.

By means of various technological advances in storage and processing of texts, a significant growth can be observed in corpus-building activity since the mid-1980s. And according to Kennedy (1998), as a result of this, scholars in many language-related fields now can be said to have a wide range of choices for selecting appropriate corpora to seek answers to particular research questions about the nature, structure and use of languages.

2.4 Major Areas Of Research In Corpus Linguistics

According to McEnery&Wilson (1996) the availability of the computerised corpus and the wider availability of institutional and private computing facilities do seem to have encouraged to the stimulation of corpus linguistics. Johansson (1991) shows how corpus linguistics grew during the latter half of this century in the following diagram.

Fig. 3 The total figures of Corpus linguistics studies in the latter half of this century

Date	Studies
To 1965	10
1966-1970	20
1971-1975	30
1976-1980	80
1981-1985	160
1985-1991	320

Reference: Johansson, S. Times change and so do corpora. Longman.1991. London p.17)

2.4.1 Corpora For Lexicography

When we have a look at the history of electronic corpora in corpus linguistics, we can see that from the late 1960s, a number of electronic corpora were compiled for specialized purposes, especially, but not exclusively for lexicographical projects of various kinds. Algeo (1988) reported the compilation of a corpus of about five million words drawn from the 18th century to the present. That five million words had been compiled for studying characteristic Briticisms in the English language. According to Kennedy (1998), the corpus was initially on 110,000 slips but in its computerized form promises to be not only the basis for a dictionary of Briticisms but a source for the diachronic study of British English and a reference source for comparative linguistic and cultural studies especially between British and American English. "The potential value of the corpus can be seen in the light it can throw on such phenomena as the possible source of the discourse item you know" (Kennedy 1998:33).

Kennedy (1998) states that other specialized corpora with similar applied linguistic purpose include the *Jiao Tong University Corpus for English in Science and Technology (JDEST)* and the *Guangzhou Petroleum English Corpus (GPEC)* which is produced in China. Both are designed to facilitate lexical analysis of particular registers, including counts of high frequency words. The *JDEST Corpus* was compiled in the 1980s and consists of about one million words from written English texts in mainly the physical sciences, engineering and technology. The *GPEC* consists of about 411,000 words comprising 700 texts from the petroleum industry from written American and British English sources of the mid-1980s.

In 1993, Stenstörn and Breivik announced the development of a corpus of London teenager language (*COLT*), being a half-million-word corpus of the English of 13-17 years-olds. After its completion in 1994, *COLT* was incorporated into the *British National-Corpus*. Another study on corpus linguistics is the European Corpus Initiative has produced a 93-million-word corpus which includes texts from most of the major European languages and

some others including Chinese, Malay and Japanese. A large corpus of 27 million words of text from Dutch newspapers has been made available for research through the institute for Dutch Lexicography at Leiden University.

According to Kennedy (1998) a corpus of 28 million sentences of written Japanese and much smaller corpus of transcribed spoken Japanese are reported to have been produced for the Japanese Electronic Dictionary Research Institute. Bilingual corpora have also been compiled containing parallel texts from pairs of languages such as English, Finnish, French, Greek, Norwegian, Spanish, Swedish and Welsh.

It is also conveyed from Kennedy (1998) that the first major machine-readable corpus based lexicography project is the *Cobuild* project, which made use of the *Birmingham Collection of English Text*, was a joint venture between a major commercial publisher, Collins, and a research team based in the English Department of the University of Birmingham, hence *Cobuild* (*Collins Birmingham University International Language Database*). By 1997 the size of this monitor corpus was reported to be over 300 million words and growing.

2.4.2 Dictionaries As Corpora

It is interesting to note that while electronic corpora have been used to make dictionaries, at the same time some dictionaries have themselves been used as rather specialized kinds of corpora. Kennedy (1998:35) defines the matter as in the following:

"The development of systems for the automatic processing of natural languages for purposes such as translation requires detailed information about the possible meanings of individual words. At present, published dictionaries are the best practical source of information on the various senses of words. For this reason, machine-readable versions of dictionaries on computer tape or CD-ROM have been used for research on automatic sense disambiguation, among other things. Probably the best-known electronic dictionary used as a corpus in the 1980s was the Longman Dictionary of Contemporary English (LDOCE), which was available on tape for a number of research teams."

2.4.3 Corpora For Studying Spoken English

Although almost all of the corpus linguistics researches were made on written corpora, there are few studies on spoken corpora. It is noted that the London-Lund Corpus of spoken English has been used for lexical, grammatical and discourse analysis as well as for prosodic studies whereas the *Lanchester / IBM Spoken English Corpus (SEC)* (Knowles et al., 1992) has been designed and used most particularly for detailed prosodic research. The corpus consists of about 52,600 words of the spoken standard British English (RP) of adults sampled between 1984 and 1987 from 11 categories including radio news broadcasts, university lectures, religious broadcasts, broadcast fiction, poetry, dialogue and propaganda. "One particular advantage of this corpus is that it is available in several versions including in orthographic transcription, with or without punctuation, grammatically tagged with the CLAWS tagset, parsed, and prosodically transcribed, showing features of stress, intonation and pauses" (Kennedy 1998:36).

2.4.4 Diachronic Corpora

According to Kennedy (1998), the study of language change is most naturally corpus based. It is scarcely surprising therefore that in the mid-1990s some of the most vigorous activity in corpus-based research has been in diachronic studies of English. He adds that with the earliest periods of a language it is possible for corpora to be more or less exhaustive collections of all known written records of the language. He exemplifies the notion as; the *Complete Corpus of Old English* prepared at the University of Toronto consists of 3,022 texts, the entire population of surviving Old English texts. Available now in electronic form, this corpus was published in 1981 as the basis for a definitive *Dictionary of Old English*. "The citations in the *Oxford English Dictionary* constitute a corpus made for lexicographical purposes. This corpus consists of quotations from mainly literary works-covering over eight centuries" (Kennedy 1998:38).

2.4.5 Corpora For Research On Language Acquisition

2.4.5.1 Language Pedagogy

Kennedy (1998) suggests that corpus has so far been vastly used and being used today in language teaching field. Fries and Traver (1940) and Bongers (1947) are the two of the most well known linguists who used the corpus in research on foreign language pedagogy. Indeed, Kennedy (1992) notes that the corpus and second language pedagogy had a strong link in the early half of the twentieth century, with vocabulary lists for foreign learners often being derived from corpora. And McEnery&Wilson (1996) suggest as the word counts which were derived from such studies as Thorndike (1921) and Palmer (1933) were important in terms of defining the target of the vocabulary control movement in second language pedagogy.

The study of language acquisition and development is essentially dependent on transcriptions of interaction between and among children and the people who look after them in natural situations. It can be seen that over the last three decades many rich and important bodies of language acquisition data have been recorded and transcribed for particular purposes. "The *Child Language Data Exchange System* (CHILDES) has since the mid-1980s brought together bodies of mainly child language acquisition data from over 500 children to form a very large database, most of which has been reformatted according to a common set of transcription conventions" (Kennedy;1998:40).

It would be beneficial to give an example of the usage of corpus linguistics and learners' corpora in a language teaching analysis. Conveyed from Kennedy (1998), Green & Hecht (1998) account that since 1979, the language-teacher-training departments of the universities of Munich and York have built up a joint corpus of 'learner language'. They report that the learners are school pupils in French, German, Hungarian, Italian, and Swedish secondary schools at beginners', intermediate, and advanced levels. The corpus consists currently of over 5,000 samples of performance in English as a foreign language, on oral and written communicative tasks

and grammar and vocabulary tests. These foreign language samples are comprehended by native language samples, produced by peer groups in English schools performing the same tasks. The data have been analysed from different points of view. Reported from Kennedy (1998), an earlier paper (Hecht and Green 1989b) looked at the grammatical competence and performance of learners and native speakers. Competence was interpreted as the degree of accuracy achieved when the focus was on the transmission of meaning. It was found that whilst the learners had achieved a good level of competence, there was a sharp fall-off from competence to performance. A comparable group of native speakers, on the other hand, showed only a negligible fall-off.

To paraphrase the matter above it can be said that "the native speakers were much better at 'performing their competence' than learners" (Ellis 1985:197).

It can also be argued by referring to Krashen (1981) as reported by Kennedy (1998) that the learners had recourse to two different grammatical systems in the two tasks. "In the performance task, where the focus was on content, they might be supposed to have drawn largely on *acquired* (implicit) grammar, whereas in the competence task, where their attention was drawn to form, they may have *monitored* their production and corrected it where necessary with the help of *learned* (explicit) rules" (Krashen 1981:4).

According to Kennedy (1998), One-million-word corpus of the English for computer science has been developed at Hong Kong University of Science and Technology. According to Kennedy (1998) it is intended to assist the teaching of English for computer science students in Hong Kong (Fang, 1993; James 1996). "The corpus consists of three 2,000-word samples from each of some 166 English language textbooks used in computer science courses in the early 1990s" (Kennedy 1998:44).

2.4.6 Other Corpora For Special Purposes

On this part of research it can be expressed that in paraphrasing Ecclesiastes, "Of making many corpora there is no end" (Greenbaum 1981:171). In addition to the established general or comparative corpora mentioned earlier, there are increasing numbers of corpora being compiled today for special purposes. Some of them will doubtless be used for other comparative purposes. "Although the word 'corpus' is sometimes is used to label a machine-readable version of even a single book or quite small collections of texts, the majority of these corpus projects for specialized purposes have settled on a corpus size of between about 100,000 and two million words of running text in length" (Kennedy 1998:43).

It is important to have a look at some studies which concern corpus for grammatical description of corpus. The size and their outcomes give us significant information about the researches done in this field. One of them is Nijmegen Corpus. Since the early 1970s a group led by Jan Aarts at the University of Nijmegen has been associated with a number of important corpus building and corpus-analysis projects. It was important because it carried one of their major goals the grammatical description of English. According to Kennedy (1998) the *Nijmegen Corpus* consists of about 132,000 words of British English from 1962 to 1968. Six 20,000-word extracts of written, mainly literary, English from six authors and 12,000 words of transcribed sports commentary make up the corpus. "The *Nijmegen Corpus* is analysed in terms of a very large set of labeled trees or phrase markers, and is intended to be used with the *Linguistic data Base (LDB)* also developed at Nijmegen" (Kennedy 1998:43).

It can be pointed out that "experience with the *Nijmegen Corpus* highlighted various shortcomings in the way in which linguistic knowledge could be formalized for the analysis of texts" (Oostdijk 1991:14). In addition to that "In a sense, work on the 1.5-million-word *Tools for Syntactic Corpus Analysis (TOSCA) Corpus* project at Nijmegen grew out of earlier development work on the smaller *Nijmegen Corpus*. The *TOSCA Corpus* is

linguistically analysed and consist of 75 samples each of 20,000 words from various fiction genres in written British English" (Kennedy 1998:43). According to Kennedy (1998) the samples are larger than in most corpora of this size, in accordance with the compilers' views on the need for samples of a certain minimum size as a foundation for quantitative studies. The texts which make up the *TOSCA Corpus* are intended to be representative of written-to-be-read, published, educated contemporary British English prose produced between 1976 and 1986 (see Oostdijk, 1988a, 1991; van Halteren & Oostdijk, 1993). Forty-five samples come from 21 non-fiction genres including, for example, (auto)biography, history, literary criticism, politics, women's studies, chemistry, economics, physics. There are 30 samples from 9 fiction genres, including, for example, horror, humour, love and romance, and general fiction.

As it is conveyed from Kennedy (1998) one of the most important corpus study is the Longman/Lanchester English Language Corpus (LLELC). The *Longman Corpus Network* is a commercial database consisting of three major corpora, the *Longman/Lanchester English Language Corpus (LLELC)*, the *Longman Spoken Corpus (LSC)*, and the *Longman Corpus of Learners' English (LCLE)*. In the late 1980s, in collaboration with Geoffrey Leech, who had directed the compilation of the *LOB Corpus* at Lancaster University between 1970 and 1976, Della Summers and her team began compiling the *Longman/Lanchester English Language Corpus*. It was intended to be a 'well-balanced' corpus of 20th-century English, covering British, American and other major varieties of native-speaker English in both spoken and written varieties, and to eventually contain up to 50 million words. Between 1991 and 1995 the *British National Corpus* was undoubtedly the most ambitious corpus compilation project yet attempted. The project was established to produce a corpus of about 100 million words of contemporary spoken and written British English. The corpus was designed to be representative of British English as a whole and not just one particular genre, subject field or register. The 4,124 texts in the *BNC* come from 90% written and 10% spoken sources.

According to Greenbaum (1996) the most ambitious corpus project for the comparative study of English worldwide is the *International Corpus of English (ICE)*. In 1988, the late Sidney Greenbaum, Director of the Survey of English Usage at University of College London, proposed the development of a large corpus for comparative study of both spoken and written forms of regional varieties of English throughout the world. The *ICE* project envisages the compilation of up to 20 parallel subcorpora, each consisting of one million words of the English used by adults over the age of 18 who have received formal education through the medium of English to at least the compilation of secondary school, in countries such as the UK, USA, Canada, Australia and New Zealand where English is the dominant or major first language, as well as in the countries such as India, Nigeria, Singapore or regions such as Caribbean where English may be an additional official language or a second language of a significant part of the population. The spoken and written texts are selected from the period 1990-93.

According to Burnard (1988), since the 1970s one of the most remarkable and valuable repositories of machine-readable texts has been the *Oxford text Archive (OTA)*. Based at Oxford University Computing Services, OTA contained by the early 1990s over 2,000 texts or collections of texts from some forty languages including Arabic, Armenian, Coptic, Danish, Dutch, English, Finnish, French, Ffulde, Gaelic, German, Greek, Hebrew, Icelandic, Italian, Kurdish, Latin, Latvian, Malayan, Mayan, Pali, Portuguese, Russian, Sanskrit, Serbo-Croatian, Spanish, Swedish, Turkish and Welsh.

Kennedy (1998) conveys that the *Bell Communications Research Corpus* (also known as the *Bellcore Corpus*) is a very large archive of modern English collected in the USA. It includes about 200 million words of newspaper wire service text, about 50 million words of other journalistic writing, and sundry other bodies of text including the *Brown Corpus* and some English dictionaries.

2.5 Techniques And Procedures Used In Corpus Linguistics

2.5.1 Issues In Corpus Design And Compilation

It is widely accepted that linguistic corpora are intended to be the basis for the analysis and description of the structure. It can also be used for the benefit of languages and for various applications. However, Johansson may very well have been correct in the mid-1990s to have noted that "the verb most frequently collocating with *corpus* is probably *compile*" (Johansson 1994:13). From the 1960s to the 1990s, as a result of a swift outlook to the studies have been carried out so far, it can be noted that the compilation, structure and size of corpora have been the subject of continuing attention among corpus linguists.

It is certain that matters in corpus design and compilation are primarily concerned with the validity and reliability of research based on particular corpus. It also concerns whether that corpus can serve the purposes for which it was intended. "Issues have included whether a corpus should be a static or dynamic sample of a language, how best it can be representative of a language or a genre, how big a corpus should be to be representative or to serve particular purposes, and how big the text samples should be" (Kennedy 1998:60).

2.5.2 Static Or Dynamic

A corpus can be a static collection of texts selected according to some specific factors. It can be typical of the whole language or an aspect of the language at a particular time. "For example, attempt to select text samples from different domains of use of spoken and written English in such a way that the corpus could be taken as synchronically representative of English" (Kennedy 1998:60). A great deal of care should be typically taken in designing the structure of such a corpus. Particular genres with a particular sample size should be included deliberately. "The great grammars of English such as Quirk et al. (1985), which were based on the *SEU Corpus*, took for granted that the corpus was a kind of snapshot of British

English" (Kennedy 1998:60). Such a corpus like a photograph of a setting, aims to capture the main characters of the landscape. On the other hand, even if the design is ideal, obviously not all local varieties can be included in such a static corpus and only certain key genres are to be taken into consideration. Such corpora are a series of static grasp of the small samples from the language. They are even from within texts rather than whole works by authors. The nature of the sample corpus can be such that it freezes the language at a particular point in time but, because of its careful structuring with fixed or in other words certain numbers of texts and text type, it can be used for comparative purposes with similarly prepared corpora. Kennedy notes "Small or large corpora can be static. Even the *British National Corpus* of 100 million words is of this type" (Kennedy 1998:61).

It can be said that the place of static sample corpora seems to be certain for research on high average frequency vocabulary, and for phonological, morphological, syntactic and much discourse-focused research.

"An alternative view of corpus design is that of the dynamic or monitor corpus" (Sinclair 1991:24). A monitor corpus can be said to be more similar to a moving picture than a snapshot. It is so-called because it provides the resources to monitor changing patterns of usage over time. Such a dynamic collection text is continuously growing and changing with the addition of new text samples. Sinclair describes the notion of the monitor corpus as "holding the state of the language at any one time" (Sinclair 1991:25). But, on the other hand, the composition and the size of the monitor corpus is constantly changing and this may therefore make it unsuitable for comparative studies, for example, relative frequency of items in different varieties of the language. Therefore, the monitor corpus can be defined as an entirely different enterprise, from the static sample corpus. The difference is that "the data collection for a monitor corpus is often opportunistic and necessarily not 'balanced'" (Kennedy 1998:61). Another difference is that quantity of text replaces planning of sampling as the main compilation criterion. It needs

expensive resources in computer hardware for capture, storage and processing of text, and the sophisticated software and technical expertise for analysis. Because of that reason, to work with monitor corpora may be less available to individual researchers. It seems likely to be undertaken mainly by large commercial companies and governmental agencies or professional research centres.

2.5.3 Representativeness And Balance

Linked to the issue of whether a corpus should be static or dynamic is the issue of how to achieve valid and reliable grounds for selecting what texts go into a corpus. Questions associated with 'representativeness' and 'balance' are complex and often intractable. Leech (1991) has suggested that a corpus is 'representative' in the sense that findings based on an analysis of it can be generalized to the language as a whole or a specified part of it. The notion of representativeness and balance are, of course, in the final analysis, matters of judgement and can only be approximate.

In the light of the perspectives on variation offered by several decades of research in discourse analysis and sociolinguistics, it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even of a particular genre or subject field or topic. And according to Kennedy, yet it remains a legitimate goal for the compilation of a corpus to be representative of a language. After all, generalizations are an essential part of science and we have no difficulty accepting generalization about the sound systems of a language even when every speaker of the language sounds different. The great dictionaries and grammars of English are all generalizations about the language in this sense (Kennedy 1998:62).

As it is discussed in linguists circle and according to McEnery & Wilson (1996) one of Chomsky's criticisms of the corpus approach was that language is infinite - therefore, any corpus would be skewed. In other words, some utterances would be excluded because they are rare, others which are much more common might be excluded by chance, and alternatively, extremely rare utterances might also be included several times. McEnery &

Wilson (1996) continue as even though today's modern computer technology allows us to collect much larger corpora than those that Chomsky was thinking about, it seems that his criticisms still must be taken seriously. But, this does not mean that we should abandon corpus linguistics, but instead we should try to establish ways in which a much less influenced and representative corpus may be built.

In addition to that we should be interested in creating a corpus that is maximally representative of the variety under examination and also it should provide us with an accurate picture of that variety, as well as their size. According to McEnery&Wilson (1996), we should look for a broad range of authors and genres, which may be considered to "average out" when we take them together and provide a reasonably accurate picture of the entire language population. Kennedy (1998:62) suggests that another issue in corpus design is the balance or weighting between the different sections in a general corpus. He adds that most of the early corpora heavily weighted in favour of written texts or were made up entirely of written texts. With current technology written text is much easier to collect into a corpus, and even in the largest of the structured second-generation sample corpora, the *British National Corpus*, only 10% of the 100 million words is from spoken sources. He notes that the balance between spoken and written texts in the smaller *ICE* is 60% spoken texts and 40% written texts, so this is one of the few corpora with the balance weighted in favour of spoken texts. On the other hand, he goes on, even within a written corpus the question of what genres to include is not straightforward. There is, for example, no comprehensive taxonomy of genres from which to select. Consequently, he states that the issue of balance arises also in corpora that are designed to represent not the language as a whole but one specific domain, genre, topic or subject field, and can be totally avoided only by corpora consisting of everything published in an historical period, the complete works of an author, or any other total population of text.

Within a written corpus, balance is equally seems to be intractable. Sinclair suggested that "for a general written corpus the minimal criteria for

the selection of texts might include the distinction between fiction and non-fiction; book, journal or newspaper; formal or informal; with control of age, gender and origin of the authors" (Sinclair 1991:20).

Many of the most important issues in achieving representativeness in a large corpus seem to have been discussed by Summers (1991), as Kennedy (1998) reports, who noted that even a written corpus of 30 million words is small when compared with the population of written text from which it is sampled. Summers argued in favour of an initial sampling approach "using the notion of a broad range of objectively defined document or text types as its main organizing principle" (Summers 1991:5). Here, it seems significant to express the matter more detailed as in the following statement by Kennedy (1998:63-64):

"This balance of text types can then be modified or fine-tuned on the basis of internal analysis of the corpus. Summers outlines a number of possible approaches to the selection of written texts, including: an 'elitist' approach based on literary or academic merit or 'influentialness'; random selection; 'currency', or the extent to which text is read, thus favouring journalistic texts and current best sellers; subjective judgement of 'typicalness'; availability of text in archives; demographic sampling of reading habits; empirical adjustment of text selection to meet linguistic specifications. A pragmatic approach is of course to use a combination of these approaches and to select from a broad range of sources and text types, taking account of currency and influentialness."

2.5.4 Size

Issues related with how to make a corpus representative or balanced or how to make it suitable for comparative purposes seem to be basically the issues related with the quality of corpus. Sinclair notes "the whole point of assembling a corpus is to gather data in quantity" (1995:21). However, Sinclair (1985) hastens to add that "In practice the size of component tends to reflect the ease or difficulty of acquiring the material." Milton expresses many difficulties of acquiring data and gives useful advice on how to address them. "Is learner data easy to acquire? The answer is definitely 'no'. Even in the most favourable environment" (Milton 1996:235). There are also many related issues associated with the quantity of text. Kennedy (1998) notes that these issues are concerned not only with the total number of

words (tokens) and different words (types) in a corpus, but with how many categories the corpus should contain, how many samples the corpus should contain in each category, and how many words there should be in each sample. "Although questions of size and representativeness affect the validity and reliability of the corpus, it has to be stressed again that any corpus, however big, can never be more than a minuscule sample of all the speech or writing produced or received by all of the users of a major language on even a single day" (Kennedy 1998:66).

Reported by Kennedy (1998), Sinclair (1991:20) suggests, "10-20 million words might constitute a useful small general corpus but will not be adequate for a reliable description of the language as a whole." It was argued that "corpora of finite size were inherently deficient because any corpus is such a tiny sample of a language in use that there can be little finality in the statistics" (Kennedy 1998:67). Sinclair pointed out that "even projected billion-word corpora will show remarkably sparse information about most of a very large word list" (Sinclair 1991:9). On this matter Kennedy (1998:67) points out that the issue is how many tokens of a linguistic item are necessary for descriptive adequacy. Although it is the case that for the descriptive adequacy of low frequency phenomena such as collocations very large corpora are necessary, there is no point in having bigger and bigger corpora if you cannot work with the output. A vast collection of texts is not necessarily a corpus from which generalizations can be made. He adds that a huge corpus does not necessarily 'represent' a language or a variety of a language any better than a smaller corpus. At this stage we simply do not know how big a corpus needs to be for general or particular purposes. He suggests, rather than focusing so strongly on the quantity of data in a corpus, compilers and analysts need also to bear in mind that the quality of the data they work with is at least as important. Studies of many syntactic processes and high frequency vocabulary generally require corpora of between half a million and one million words. Overall corpus size needs to be set against diversity of sources to help achieve representativeness.

Biber (1993b) argued that quantity of data alone can not solve issues in corpus design involving the kinds and the number of texts to be included, the selection of samples from particular texts and the size of each sample. These matters must also be based on theory. Biber argued for a qualitative as well as a quantitative basis for corpus design, and it should encourage continuing use of corpora of the order of one million words for grammatical studies.

2.5.5 Compiling A Corpus

According to Kennedy (1998), there are many linguistic corpora available but these will not always be suitable for the purposes of every prospective user who wish to do corpus-based research. According to the most well known linguists there seems to be three main stages in corpus compilation: corpus design, text collection or capture, and text encoding or mark up.

2.5.6 Corpus Design

Kennedy (1998) notes that when compiling a new corpus, even if it is small, it is sensible for compilers to assume that it may eventually be used for comparative purposes. In order to ensure comparability and compatibility, careful planning and principled ways of selecting texts for a corpus are seen to be highly desirable, and what is more, their ability to be used or referred to by other researchers is also important. He argues that the optimal design of a corpus is highly dependent on the purpose for which it is intended to be used. He gives some advice to the researchers about the compilation of corpus. According to him, the compiler of a corpus should, if possible, have a clear idea of what kinds of analyses are likely to be undertaken and whether they justify the large amount of effort involved in making the corpus. He advocates, in normal circumstances, it is assumed that a purpose-built corpus for particular research is likely to be on a relatively small, finite scale and will be a synchronic or diachronic corpus of spoken and/or written texts. In addition, he adds, the corpus may consist

either of a total population of texts (e.g. all the works of a particular author) or of a sample of texts from a given population. A number of matters need to be kept in mind in deciding on the appropriate type of corpus. The purpose of sampling adequately is so that, on the basis of the sample, generalizations can be made reliably and validly about a population as a whole (Kennedy 1998:70).

2.5.7 Planning A Storage System And Keeping Records

While planning a storage system Kennedy (1998) notes that it is important to keep a catalogue of files and, of course, all material should be backed up and stored securely and separately from the working copies of texts and the working electronic files. "In addition to the storage and cataloguing of texts and their electronic version on computer, it is normally essential to plan to collect and catalogue as much information as possible about the authorship or source of texts. In the case of spoken or written texts, it is important to note when and where the text was recorded or collected" (Kennedy1998:76).

2.5.8 Getting Permission

According to Kennedy (1998), while gathering information on a specific research, corpus compilers, like other users of texts, are strongly expected to follow the same rules and legal requirements as other members of the community. Ethic rules and moral values are significant to be followed. Before texts are copied into a corpus database, compilers must ask for and get the permission of the authors and the publishers who hold the copyright for a work, or the informed consent of individuals whose rights to privacy must be acknowledged.

2.5.9 Data Capture, Mark Up And Documentation

According to Granger (1998), data capture, mark up and documentation are to large extent similar for learner corpora and native corpora. There are few practical hints for the prospective learner corpus

builder. She suggests, of the three methods of data capture – downloading of electronic data, scanning and keyboarding, it is keyboarding that currently seems to be most common in the field of learner corpora. Indeed, she adds, it is the only method for learners' handwritten texts. As a conclusion, the fast-growing number of computers at students' disposal both at home and on school / university premises is improving this situation and researchers can expect to get a higher proportion of material on disk in the near future.

Granger (1998) advocates that both keyboarded and scanned texts contain errors and therefore require significant proofreading. In the case of learner corpora, this stage presents special difficulties. The proofreader has to make sure he edits out the errors introduced during keyboarding or scanning but leaves the errors that were present in the learner text, a tricky and time-consuming task. She notifies, errors such as omissions, additions or homonyms (their /there, it's/its) can only be spotted by careful manual editing. In any case, as errors can escape the attention of even the most careful of proofreaders, she points that it is advisable to keep original texts for future reference. "Using a standard mark-up scheme called SGML (Standard Generalized Markup Language), it is possible to record textual features of the original data, such as special fonts, paragraphs, sentence boundaries, quotations, etc" (Granger 1998:11).

Accordingly, in Granger (1998), Atkins and Clear (1992:9) note that markup insertion is a very time-consuming process and researchers should aim for 'a level of mark-up which maximizes the utility value of the text without incurring unacceptable penalties in the cost and time required to capture the data.

According to Granger (1998), in the case of learner corpora, which tend to contain few special textual features, this stage can be kept to a minimum, although it should not be bypassed. For some types of analysis it would be highly advantageous to have textual features such as quotations, bold or underlining marked up in the learner corpus.

In order to be maximally useful, Granger (1998) suggests that a corpus must be accompanied by relevant documentation. Full details about

the attributes must be recorded for each text and made accessible to the analyst either in the form of an SGML file header included in the text files or stored separately from the text file but linked to it by a reference system. She adds that both methods enable linguists to create their own tailor-made subcorpora, by selecting texts which match a set of predefined attributes, and focus their linguistic search on them. "On this basis, learner corpus analysts are able to carry out a wide range of comparisons: female vs. male learners, French learners vs. Chinese learners, writing vs. speech, intermediate vs. advanced, etc" (Granger 1998:12).

Basically, a corpus is a collection of written and spoken data. However, before we can analyse a corpus with the help of a computer program, this collection must be turned into a computer-readable form. According to Granger (1998:11,19-21), usually this is done by the following means:

- keying texts in scanning
- using texts which are stored in machine-readable form anyway (This is true of the written component of the British National Corpus, for example.)
- using electronic texts available through the internet (check the Gutenberg project for more information)

Since more and more texts are written with the help of personal computers, the last two choices seem to become the standard for written material. As far as spoken corpus material is concerned, this is still an entirely different matter.

According to Kennedy (1998:21) for annotating corpora we can make a list as in the following:

- plain
- tagging (i.e. assigning a word-class marker to each token)
- including information about the text (genre, date of publication, place of publication)
- including information about intonation patterns or pronunciation, e.g. London-Lund Corpus

- Information about the speaker (his/her sex, age, occupation, social and geographical origin)
- parsing, i.e. assigning labelled brackets to each constituent of a sentence
- discourse information of spoken material (laughing, interruptions, hesitations)

2.6 Application of Corpus Based Analysis

Although a corpus can be a new kind of research field involving new methodologies, the use of corpora does not constitute a new or separate branch of linguistics. What is more, corpus linguistics is essentially descriptive linguistics aided by new technology.

Today, It is known that by using computers it is much easier and faster to find, sort, analyse and quantify linguistic features and processes in huge amounts of text. It has had a significant impact in a number of fields in the language sciences. In addition to descriptive studies of phonology, grammar, vocabulary and discourse it can be seen that "corpora of English have been used for lexicography and for research on social, regional, diachronic and stylistic variation, first and second language acquisition, and natural language processing, including tagging and parsing, speech recognition and machine translation" (Kennedy 1998:268-269).

Atkins et al. (1992) classified the potential users of corpora into three main groups as:

- 1- The ones which are interested in the language of texts.
- 2- The ones which are interested in the content of text.
- 3- The ones which are interested in the texts themselves.

It can be said that any such classification certainly has unclear boundaries but, in general, it is a useful way of showing the degree to which corpus-based analysis has effects beyond linguistic description. Via computers, it is predictable that the availability of large bodies of text in electronic form has initiated an incredible increase in corpus-based research on language within computational linguistics. "Computational linguistics is

the interdisciplinary field of scholarship which seeks to study and simulate with computers, the processes and procedures by which we interpret and produce natural language" (Kennedy 1998:269). It can be seen that computational linguists have put the emphasis on the language of texts from two directions. "Initially, extant linguistic descriptions were used as a basis for devising 'knowledge-based' rules for the automatic linguistic analysis of texts. More recently, self-organising approaches to analysis based on probabilities of occurrence in corpora have been used, but both approaches have needed each other" (Kennedy 1998:269).

As Atkins et al. have noted, research suggests that "self-organising statistical techniques gain much in effectiveness when they act on the output of grammatically-based analysis" (Atkins et al. 1992:14).

According to some of the applied linguists such as Kennedy, Granger, Halliday et al. who were concerned with language teaching, "it can be said that a corpus cannot only be one source of authentic text to which learners of a language might be exposed, but also the corpus can be analysed to discover the relative weight which might be given to lexical or grammatical items and processes in curricula and teaching materials" (Kennedy 1998:269).

The description by the linguists who study on corpus may be useful in order to enlighten their characteristics for a person who means to make a research in this field. Chafe defines a corpus linguist as:

"A linguist who tries 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"

(Chafe 1992:96)

It has previously been reminded that for lexicography to use corpora has of great importance. That we can make use of it for identifying types, for the development of new senses for types, and for the relative frequency of use of these different senses. In a machine-readable corpus there are numerous research topics, which are potentially as various and wide

ranging as the facts about a language and the use of that language. It can be valuable and useful for students to be invited to study on some questions for example, about the behaviour of particular words or structures.

As suggested in Kennedy (1996b), there may be some research questions, which are likely to be central in corpus-based descriptive studies. For instance the questions may be similar to the following ones: What are the linguistic units, patterns, systems or processes in the language, genre or text and how often, when, where, why and with whom are they used? According to Kennedy (1998:276), the texts may be either spoken or written; it is seen that corpus-based studies have characteristically focused on 4 main types of analysis and description:

- 1- Word-based studies which explore the ecology of lexis both in terms of the occurrence and frequency of occurrence of items. Work by Renouf and Sinclair (1991), Kjellmer (1994) and others suggests that the expansion of lexical studies to include collocations is likely to be one of the most innovative and productive areas of corpus-based research.
- 2- Studies of the co-occurrence of grammatical word-class tags as expressions of syntactic patterning and as the bases for quantitative studies of the use of syntax.
- 3- Studies of the co-occurrence of groups of linguistic items or processes to show by means of factor analysis the linguistic characteristics of genre.
- 4- Studies of the structure of discourse, especially of spoken interaction, and of the bases of cohesion in spoken and written texts.

Biber (1998) has pioneered the use of this approach to establish text types, which differs from the usual domain or topic-based approaches.

Kennedy (1998) notes that systematic coverage is one of the most challenging and potentially most significant fields contributing to and making use of corpus-related research. The impact of corpora on computational linguistics can be estimated if we consider the extent to which the content

and direction of research within the field has changed within less than a decade.

2.6.1 Applying Corpus Linguistics To Teaching

According to Barlow (2002), there are three subjects in which corpus linguistics can be applied to teaching. These can be counted as syllabus design, materials development, and classroom activities.

2.6.1.1 Syllabus Design

According to Barlow (2002), the syllabus organizes the teacher's decisions regarding the focus of a class with respect to the students' needs. Frequency and register information could be quite helpful in course planning choices. By conducting an analysis of a corpus which is relevant to the purpose a particular class, the teacher can determine what language items are linked to the target register.

2.6.1.2 Materials Development

Krieger (2003) conveys from Barlow (2002) that the developer uses initiative in order to develop materials according to the students' needs. With the help of a corpus, a materials developer could work on real examples which provide students with an opportunity to discover features of language use. In this situation, the materials developer could conduct the analysis or simply use a published corpus study as a reference guide.

2.6.1.3 Classroom Activities

According to Barlow (2002), these can be performed on student-conducted language analyses. In these activities, the students use a concordancing program and a deliberately chosen corpus in order to make their own discoveries about language use. The teacher can guide a programmed investigation which will lead to expected results or can have the students do it on their own, leading to less predictable findings. This exemplifies data driven learning, which encourages learner autonomy by training students to draw their own conclusions about language use.

2.6.1.4 Teacher/Student Roles And Benefits

Krieger (2003) conveys from Barlow (1992) that the teacher would act as a research facilitator rather than the more traditional imparter of knowledge. The benefit of such student-centered discovery learning can be that the students are given access to the facts of authentic language use, which comes from real contexts rather than being constructed for pedagogical purposes. By this way, they are challenged to construct generalizations and note patterns of language behaviour. Although this kind of study is not expected to present immediate quantifiable results, studying concordances can make students more aware of language use. Richard Schmidt (1990), who is a proponent of consciousness-raising, argues that what language learners become conscious of -- what they pay attention to, what they notice influences and in some ways determines the outcome of learning. According to Willis (1998), students may be able to determine:

- the potential different meanings and uses of common words
- useful phrases and typical collocations they might use themselves
- the structure and nature of both written and spoken discourse
- that certain language features are more typical of some kinds of text than others

According to Barlow (1992) a corpus and concordancer can be used to:

- compare language use--student/native speaker, standard English/scientific English, written/spoken
- analyze the language in books, readers, and course books
- generate exercises and student activities
- analyze usage--when is it appropriate to use obtain rather than get?
- examine word order

2.6.1.5 Problematic Issues Involved

According to Krieger (2003), there are several challenges in employing the use of a corpus for the purpose of teaching. One of them is corpus selection. For some teaching purposes, any large corpus can be proper. But the teacher needs to make sure that the corpus is useful for the particular teaching context and is representative of the target list. Another option can be to construct a corpus, especially when the target list is highly specific. This can be done by using a textbook, course reader, or a bunch of articles which the students have to read or they should have the representativeness of what they have to read. A corpus does not need to be large in order to be effective. The fundamental criteria can be that of relevance to the students. For example, it is highly expected to be selected with the learning objectives of the class in mind, matching the purpose for learning with the corpus.

He argues that related to the issue of corpus selection is that of corpus bias, which can cause frustration for the teacher and student. The reason for this is that the data can be misleading; if one uses a very large general corpus, it may obscure the register variation which reveals important contextual information about language use. The pitfall to be avoided is that a corpus may tell us more about itself than about language use. Another obstacle to confront can be the comprehensibility issue: if you use concordancing in a class, it can be quite difficult for the students (or even the teacher) to understand the data that it provides. Lastly, the issue of learning style differences--for some students, discovery learning is simply not the optimal approach. All of these points reinforce the caveat that careful consideration is required before a new technology is introduced in the classroom, especially one, which has not been thoroughly explored and streamlined.

2.6.1.6 Exploiting a Corpus for a Classroom Activity

Krieger (2003) points out that although corpora may sound reasonable in theory, applying it to the classroom can be challenging because the information it provides appears to be so confusing. For this reason, the teacher is expected to harness a corpus by filtering the data for the students. Although some support having students conduct their own analyses, at present it is possible for them to see corpora's greatest potential as a source for materials development. Susan Conrad (2000) implies that materials writers take register specific corpus studies into account. Biber, Conrad and Reppen (1998) emphasize the need for materials writers to acknowledge the frequency which corpus studies reveal of words and structures in their materials design.

2.7 Corpus-Based Research On Second Or Foreign Language Acquisition

2.7.1 Corpus-Based Approaches To Language Teaching

Kennedy (1998) briefly accounts the approaches as second language teaching theories and practises are certainly seemed to be influenced by facts and opinions about the conditions necessary for successful language learning. These can be counted as: "the background, attitudes and goals of particular learners, the techniques and procedures which are believed to facilitate language learning most efficiently, and what we believe needs to be learned in order to be a language user" (Kennedy 1998:280).

2.7.1.1 Language Teaching Methodology

We can say that in addition to its impact on the linguistic content of language pedagogy, corpus-based research also has the potential to have a direct effect on language teaching methodology. As we have seen, for over three decades the analysis and description of corpora has accumulated facts about lexis, grammar and discourse at a time when language teachers were

being encouraged to expose learners to authentic spoken discourse or written texts and to the negotiation of meaning with interlocutors. The details of analysed wordlists, or statistics on the relative frequency of items or processes in the language provided by corpus analysis, have been in a sense out of step with typical meaning-based approaches to language teaching. Corpora are supposed to be used with care for pedagogical purposes. It is highly accepted that it should inform instruction rather than determining it. It is thought significant to be seen as this way so as to avoid the risk of a return to prescriptivism. Frequency of occurrence in corpora should be only one of the criteria, which is used to influence instruction. Sometimes according to the objectives of the learners, less frequent items or processes in a language may require more attention than the most frequent, just because they are known to be learning problems with a wide range of uses. The idea that should be kept in mind is that the facts about language and language use, which emerge from corpus analyses, should never be allowed to become a burden or trouble for pedagogy.

When we look over the researches carried out in this field it is possible to notice that in spite of an obvious increase in the use of corpora during the 1990s, it would be misleading to suggest that the use of corpora has become part of the mainstream in language teaching theory and practice. Partially the reason for that is, meaning-focused instruction has been widely accepted, and it seems to have been a reluctance by many teachers to return to instruction with a central focus on the forms of language. There appears to be already signs that the fruits of corpus-based research may help these approaches bring together for language teaching. According to Kennedy (1998:288-290) a second reason perhaps why the potential contribution of corpus linguistics is yet to be realized in pedagogy is because of an apparent need for users of corpora to have computing skill. With newly available corpora and software, however, doing corpus analysis can be said to become easier by the day.

2.7.1.2 The Advantages Of Doing Corpus-Based Analyses

When we make a review of the studies related to corpus linguistics, it probably offers a more objective view of language than that of introspection, intuition and anecdotes. John Sinclair (1998) pointed out that this is because speakers do not have access to the subliminal patterns, which run through a language. A corpus-based analysis can investigate almost any language patterns. These can be lexical, structural, lexico-grammatical, discourse, phonological and morphological. And it does that often with very specific agendas such as discovering male versus female usage of tag questions, children's acquisition of irregular past participles, or counterfactual statement error patterns of Japanese students and so on. Krieger (2003) suggests that with the proper analytical tools, an investigator can find out not only the patterns of language use, but also the extent to which they are used, and the contextual factors that influence variability. For example, one could examine the past perfect to see how often it is used in speaking versus writing or newspapers versus fiction. Or one might want to investigate the use of synonyms like begin and start or big/large/great to determine their contextual preferences and frequency distribution.

2.7.2 First Language Acquisition

The study of language acquisition and development is suggested to be crucially depended on transcriptions of interaction between and among children and caregivers in natural situations. "Over the last decades it can be noted that many rich and important bodies of language acquisition data have been recorded and transcribed for particular purposes" (Kennedy; 1998:40).

2.7.3 Second Language Acquisition

According to Perdue (1993), it is noticeable that corpora have also been brought together for research on second or foreign language acquisition. Kennedy (1998:42) notes that these include the *European Science Foundation Second Language Databank (ESFSLDB)* of transcribed speech collected for the longitudinal study of the learning of Dutch, English,

French, German or Swedish by adult immigrants from different language backgrounds including Punjabi, Spanish, Turkish, Finnish, Italian and Moroccan Arabic.

2.7.4 Foreign Language Acquisition

There are some studies on the characteristics of English as a foreign language. And these studies of characteristics are expected to be facilitated by several important corpora. For example, Grangers (1993) notes that an *International Corpus of Learners' English* (ICLE) is being developed at the Catholic University of Louvain in Belgium. According to Kennedy (1998), the corpus is compiled from written texts produced by advanced learners of English from a number of countries including Belgium, China, France, Germany, the Netherlands and Sweden. According to Kennedy (1998), the texts come from 500-word student essays. He argues that each learner variety has a minimum of 400 essays, thus making up national subcorpora of approximately 200,000 words each. "Study of the general characteristics of Interlanguage, the transitional forms used by learners of a second language, should also benefit from this corpus" (Kennedy 1998:42).

Warren (1992) conveys that whereas the ICLE is designed primarily for academic research, a large corpus of learners' English has been developed for academic research, lexicography and for the preparation of commercial language teaching materials. "The *Longman Corpus of Learners' English* (LCLE) totals approximately 10 million words and is made up of samples of written English from sources including examination answers, letters, reports, diaries and student essays from learners of English of over eight different levels of proficiency from over 160 different language backgrounds" (Kennedy 1998:42).

2.8 Computer Corpus Based Interlanguage Analysis

As said by McEnery and Wilson (1996), it is a common belief that corpus linguistics was abandoned entirely in the 1950s, and then adopted once more almost as suddenly in the early 1980s. This is simply untrue, and

does a disservice to those linguists who continued to pioneer corpus-based work during this interregnum.

For example, they go on, Quirk (1960) planned and executed the construction of his ambitious Survey of English Usage (SEU) which he began in 1961. In the same year, Francis and Kucera began work on the now famous Brown corpus, a work which was to take almost two decades to complete. These researchers were in a minority, but they were not universally regarded as peculiar and others followed their lead. In 1975 Jan Svartvik started to build on the work of the SEU and the Brown corpus to construct the London-Lund corpus.

During this period the computer slowly started to become the mainstay of corpus linguistics. Svartvik computerised the SEU, and as a consequence produced what some, including Leech (1991) still believe to be to this day an unmatched resource for studying spoken English.

McEnery and Wilson (1996) also note that the availability of the computerised corpus and the wider availability of institutional and private computing facilities do seem to have provided a spur to the revival of corpus linguistics.

2.8.1 Processes

According to McEnery&Wilson (1996), considering the marriage of machine and corpus, it seems worthwhile to consider in slightly more detail what these processes that allow the machine to aid the linguist are. The computer has the ability to search for a particular word, sequence of words, or perhaps even a part of speech in a text. So if we are interested, say, in the usages of the word *however* in the text, we can simply ask the machine to **search** for this word in the text. The computer's ability to **retrieve** all examples of this word, usually in context, is a further aid to the linguist.

They argue that the machine can find the relevant text and display it to the user. It can also **calculate** the number of occurrences of the word so that information on the frequency of the word may be gathered. We may then be interested in **sorting** the data in some way - for example, alphabetically on

words appearing to the right or left. We may even sort the list by searching for words occurring in the immediate context of the word. We may take our initial list of examples of *however* presented in context (usually referred to as a **concordance**), and extract from this another list, say of all the examples of *however* followed closely by the word *we*, or followed by a punctuation mark.

McEnery&Wilson (1996) conclude that the processes described above are often included in a **concordance program**. This is the tool most often implemented in corpus linguistics to examine corpora. Whatever philosophical advantages we may eventually see in a corpus, it is the computer which allows us to exploit corpora on a large scale with speed and accuracy.

2.8.2 Contrastive Interlanguage Analysis

A learner corpus based on clear design criteria lends itself particularly well to a contrastive approach. Not a contrastive approach in the traditional sense of CA (Contrastive Analysis), which compares different languages, but in the totally new sense of "comparing / contrasting what non-native and native speakers of a language do in a comparable situation" (Pery-Woodley 1990:143). This new approach, which Selinker (1989:285) calls a 'new type of CA' and which some refer to as CIA-Contrastive Interlanguage Analysis-lies at the heart of CLC-based studies. James sees this new type of comparison as a particularly apt basis for a "quantificational contrastive typology of a number of English IIs" (James 1994:14).

According to Granger (1998:12) CIA involves two major types of comparison:

- 1- NL vs. IL, i.e. comparison of native language and interlanguage;
- 2- IL vs. IL, i.e. comparison of different interlanguages.

Since the two types of comparison have different points, it might be useful to examine them separately. To begin with, Granger (1998) suggests that IL / IL comparisons aim to discover the characteristics of non-nativeness of learner language. It could be possible to study at all levels of

proficiency, however, it would be more beneficial to study at the most advanced ones, these characteristics are not only supposed to involve plain errors, but also the differences in the frequency of use of certain words, phrases or structures, in terms of their being overused and underused. She argues that before CLCs became available, work on learner production data had focused mainly on errors, but now SLA specialists can also investigate quantitatively distinctive features of interlanguage (i.e. overuse /underuse), a brand new field of study which has important suggestions for language teaching. In order to clarify the notion she gives the example of writing textbooks and electronic tools such as grammar checkers, even those designed for non-native speakers, advice learners against using the passive and suggest using the active instead. She concludes that a recent study of the passive in native and learner corpora however, shows that learners underuse the passive and that they are thus not in need of this type of inappropriate advice.

When we have a look at the NL /IL comparisons we notice that they require a control corpus of native English and they are widely and easily available. Granger (1998) suggests that a corpus such as ICA (International Corpus of English) even provides a choice of standards: British, American, Australian, Canadian, etc. But, on the other hand, she urges that one factor which analysts should never lose sight of is the comparability of text type. That the reason for this she explains, because many language features are style-sensitive, it is essential to use control corpora of the same genre. As demonstrated in Granger and Tyson (1996:p.23), she includes, a comparison of the frequency of three connectors- *therefore*, *thus* and *however*- in the LOB, a corpus covering a variety of text types, and ICLE, which only includes argumentative essay writing, leads to a completely distorted view. At this point, we can conclude that it fails to bring out the underuse of these connectors by learners which clearly appears when a comparable corpus of native speaker argumentative essays is used instead. Granger (1998) adds, the corpus used in this case was the LOCNESS' (Louvain Corpus of Native English Essays), a 300,000- word corpus of

essays written by British and American university students. In this respect, there seems to be another perspective that this corpus has the advantage of being directly comparable to ICLE, on the other hand, it has the inconvenience of being relatively small and containing student, in other words, non-professional writing. As a conclusion of this, criticism can be levelled against most control corpora. While comparing the corpora there are some matters to be kept in mind as Granger (1998) points out, each has its limitations and the important thing is to be aware of them and make an informed choice based on the type of investigation to be carried out.

The second type of comparison – IL vs. IL – may involve comparing IIs of the same language or of different languages. The main objective of IL / IL comparisons according to Granger (1998), can be said, is to gain a better insight into the nature of interlanguage. She suggests, by comparing learner corpora or subcorpora covering different varieties of English (different in terms of age, proficiency level, L1 background, task type, learning setting, medium, etc.), it is possible to evaluate the effect of these variables on learner output. Lorenz, for example, draws interesting conclusions on the effect of age / proficiency on written output by comparing German learners of English from two different age groups with a matched corpus of native English students (Granger 1998:12-14).

2.8.3 The Role Of Interlanguage In Foreign Language Teaching

Because of its very nature Nicke (1995) argues that *interlanguage* (IL) is always been an important issue in connection with language acquisition even before the term and its quasi-synonyms were coined. The disputed characteristics and definition of IL have sometimes made it appear mysterious. "After a short discussion of its evolution and some of its theoretical aspects, its practical direct and indirect relevance for ELT including the socio-linguistic phenomenon of international varieties of English (EIL) also the subject *Language awareness* is described" (Nicke 1995:1).

2.8.4 Automated Linguistic Analysis

2.8.4.1 Linguistic Software Tools

One of the main advantages of computer learner corpora seem to be that they can be analysed with a wide range of linguistic tools, from simple ones, which merely search, count and display, to the most advanced ones, which provide sophisticated syntactic and /or semantic analysis of data. According to Granger (1998) these programs can be applied to large amounts of data, thus allowing for a degree of empirical validation that has never been available to SLA researchers before.

According to Granger (1998), text retrieval programs- commonly referred to as 'concordancers'- are almost certainly the most widely used linguistic software tools. Initially quite rudimentary, they have undergone tremendous improvement over the last few years and the most recent programmes, such as WordSmith, for example, have reached a high degree of sophistication, enabling researches to carry out searches which they could never hoped to do manually. They can count words, word partials and sequences of words and sort them in a variety of ways. They also provide information of how words combine with each other in the text. Finally "they can also carry out comparisons of entities in two corpora and bring out statistically significant differences, a valuable facility for CIA-type research" (Granger; 1998:14-15). It is illustrated by the linguists that CLC-based concordances can be used to discover patterns of Error, which in turn can be converted into useful hints for learners in ELT dictionaries or grammars.

Granger (1998) suggests that the value of computerization, however, goes far beyond that of quick and efficient manipulation of data. Using the appropriate electronic tools, SLA researches can also enrich the original corpus data with linguistic annotations of their choice. This type of annotation can be incorporated in three different ways: automatically, semi-automatically or manually. They can be briefly described as:

Parts of speech (POS) tagging as suggested by Granger (1998) is a good example of fully automatic annotation. She adds, POS taggers assign

a tag to each word in a corpus, which indicates its word-class membership. The interest of this sort of annotation for SLA researchers seems to be obvious. For example, it enables them to conduct selective searches of particular parts of speech in learner productions.

Semi-automatic annotation tools are thought to enable researchers to introduce linguistic annotations interactively, using the categories and templates provided or by loading their own categories. Granger (1998) argues that if software does not exist for a particular type of annotation, researchers can always develop and insert their own annotation manually. She believes that this is the case for error tagging. And thus, she concludes, once error taxonomy has been drawn up and error tags inserted into the text files, the learner corpus can be produced.

2.8.4.2 CLC Methodology

One issue of importance in CLC methodology according to Granger (1998) is how to approach learner corpora. Research can be hypothesis-based or hypothesis-finding. Using the traditional-based approach, the analyst starts from a hypothesis based on the literature on SLA research and uses the learner corpus to test his hypothesis. The advantage of this approach as quoted from Granger (1998) is that the researcher knows where he is going, which greatly facilitates interpretation of the results. "The disadvantage is that the scope of the research is limited by the scope of the research question" (Granger, 1998:14).

The other approach is defined as follows by Scholfield in more exploratory, as it is conveyed from Granger (1998), "hypothesis-finding" research, "the researcher may simply decide to gather data, e.g. of language activity in the classroom, and quantify everything he or she can think of just to see what emerges" (Scholfield 1995:24). This type of approach seems to be particularly well suited to CLCs, since the analyst simply has to feed the data into a text analysis program and wait to see what comes out. This approach seems potentially very powerful since it can help us gain totally new insights into learner language. However, according to Granger (1998), it

is potentially a very dangerous one. She argues that SLA specialists should avoid falling prey to what I would call 'so what?' syndrome, which unfortunately affects a number of corpus linguistics studies. "With no particular hypothesis in mind, the corpus linguist may limit his investigation to frequency counts and publish the 'results' without providing any interpretation for them. 'So what?' are the words that immediately come to mind when one reads such articles." (Granger 1998:15-16).

There seems to be no way however, in which one approach is better, in absolute terms, than the other. Depending on the topic and the availability of appropriate software, the analyst will tend to use one or the other or combine the two. Moreover, as Granger (1998) conveys from Stubbs "The linguist always approaches data with hypotheses and hunches, however vague" (Stubbs 1996:47). Therefore, we can say that the CLC enables SLA researchers to approach learner data with a mere hunch and let the computer do the rest.

Granger (1998:16) points out another important methodological issue as the role of the statistical-quantitative approach in computer-aided analysis. As said by Granger, in the field of learner corpora, the notion of statistical significance should be weighed against that of pedagogical significance. We can say that a teacher who is analysing his learners' output with the help of computer techniques may well come up with highly interesting new insights based on quantitative information which may in itself not to be statistically significant but which nevertheless has value within a pedagogical framework. "A much greater danger than not being 'statistically significant', is to consider figures as an end in themselves rather than a means to an end" (Granger 1998:16).

Finally, a computerized approach can be said to have linguistic limitations. It seems to suit better to some aspects of language than others and SLA researchers should not limit their investigations to what the computer can do. Granger (1998) points out that ideally suited for the analysis of lexis and to some extent grammar, it is much less useful for discourse studies, and she conveys from Virtanen (1996:162) that "many

textual and discoursal phenomena of interest are harder to get at with the help of existing software, and a manual analysis of the texts then seems the only possibility". She also suggests that SLA researchers should never hesitate to adopt a manual approach in lieu of, or to complement, a computer-based approach. As Ball remarks, conveyed from Granger (1998) "given the present state of the art, automated methods and manual methods for text analysis must go hand in hand" (Ball 1994:295).

Granger (1998) suggest that by offering more accurate descriptions of learner language than have ever been available before, computer learner corpora will help researchers to get more of the facts right. They will contribute to SLA theory by providing answers to some yet unresolved questions such as the exact role of transfer. And in a more practical way, "they will help to develop new pedagogical tools and classroom practices which target more accurately the needs of the learner" (Granger 1998:18). We should keep in mind that a corpus or a computer cannot do all the work for us. Even though we have got powerful computers and sophisticated concordance programs, it is still our job to formulate intelligent and sensible research questions. The quality of the material the computer offers we should depend on the quality of the questions we have asked. Besides, we should not think that corpus linguistics encompasses just the counting of words.

2.8.5 The Importance Of Interlanguage Analysis in Foreign Language Teaching

Some time ago one would have perhaps hesitated to include the IL-phenomenon among the topics of a conference. But today, when we look at the programme of a conference about corpus linguistics or rather related with '**Interlanguage analysis**' or/and '**Contrastive studies**' it can easily be seen that several papers dealing with it, apparently interpreting the IL-phenomenon also in terms of a problem connected with languages in contact. Nickel (1989) states that the logical reason for this is certainly that in spite of all disputes concerning the definition and nature of IL there is still agreement that it can only arise when more than one language is acquired.

"Strictly speaking, it would also apply to dialect speakers trying to acquire an official standard of their language within L1-acquisition. But, naturally, it is normally only applied in FL- and SL-teaching" (Nickel 1989:298) & (Hammerly 1991:148). , "where it crops up particularly in connection with topics like Contrastive Analysis (CA), Error Analysis (EA) or with FL- and SL-learning in general" (Sridhar; 1981:207-241).

"In spite of many publications on IL there still "remain many unsolved problems requiring much more research" (Young 1988:418). Here I intend to discuss neither the theoretical nor empirical aspects of the phenomenon, but addresses itself, directly and indirectly, to FLT- issues and discusses only briefly some of the most important aspects of IL, as far as they are relevant for the topic.

It is undeniable that the popularity of IL studies has been growing during the past decades. While Dulay, Burt and Krashen state that "the study of learners' errors had been a primary focus of L2-research during the last decade" (Krashen et al.;1982:139-140). The causes for this enhance in popularity are multiple: "continued research in fields like CA, EA, but also general interest in language acquisition, very often in connection with psycho- and socio-linguistic concepts" (Chomsky's universals). The popularity is reflected in the growing number of symposia, and also in articles and monographs. Though the journal *Interlanguage Studies Bulletin* (ISB) ceased publication, other journals connected with language acquisition have continued its heritage. IL's popularity is also reflected in its rich terminology. Though L.Selinker (1992) rightly disagrees as:

With the widespread conclusion that there is synonymy of the three concepts 'transitional competence', 'approximative systems' and IL... because each of the three makes different theoretical claims about the nature of the SLA process (L.Selinker1992:221).

The fact still remains that this quasi-synonymy, to which a term like 'idiosyncratic dialect' could be added, certainly demonstrates the interest in the varying degrees of learners' competences and performances. Even the term IL has been interpreted in many ways ranging from more to less

fossilized versions, giving preference to one or the other of Selinker's original five central processes "(language transfer, transfer of training, strategies of second-language learning, strategies of second-language communication and over-generalization of target language linguistic material)" (Selinker 1972:209-230). "The heterogeneity and partial overlapping of these processes have been sometimes criticized" (Lander 1981:56-71). In any case variation is the most important characteristic of IL. "...the interlanguage hypothesis stands or falls on how adequately variation may be accounted for" (Yong 1988:282).

Undoubtedly, however, the main difference lies in the attitude towards the transfer phenomenon, which some, like Hammerly, Shridar, Nickel and many other colleagues in the field of CA, consider still of importance, though by no means constituting the only factor (Nickel/Wagner 1968, Vol.6/3:255), (Gass/Selinker 1983:7), (Selinker 1992:5,237).

Without any doubt, apart from transfer, almost inevitable fossilization plays a very important part in connection with characteristics of the IL-hypothesis, "a reasonable theoretical study" (Selinker 1992:246): less than 5% of FL-learners are estimated to reach native-speaker competence. Even SL-learners' pronunciations, in spite of a long period of immersion in the target-language country, some of them famous politicians, artists and scientists, support Selinker's supposition, also based on his personal linguistic experience, that it seems to be extremely difficult, if at all possible, to avoid a certain 'accent'. "While children under certain conditions, and again mostly in connection with SL, may acquire a native-like accent, possibilities in the case of adult learners seem to be much more limited" (Neufeld 1987:322).

It is obvious that all these factors concern teachers and learners in different direct and indirect ways. "It must be made clear that there are striking differences between FL- and SL- teaching due to factors like motivation, intensity of teaching and learning, important pragmatic factors, social setting, and native vs. non-native teachers, to mention only a few

factors, which also explain differences concerning results within research on errors" (Nickel 1989:298).

Teachers will certainly be only too pleased to hear that errors, including fossilizations, are unavoidable, their qualities and quantities depending upon many social and individual factors like age, motivation and ambition of teachers and students, to mention only some of them. "Hammerly's expectations are very high, hardly achievable and if so, in connection with SL rather than with FL teaching" (Nickel 1995:80). Teachers should also realize that they are users of IL themselves, unless they are native speakers, and should discuss it quite frankly with at least advanced students. Simplifying strategies, a well-known factor within the IL-phenomenon, should not only be accepted but also taught as survival strategies, particularly in communicative processes in order to help in connection with the well-known backsliding processes under stress-conditions, where again contrastive interference seems to be particularly strong and a great deal of understanding, patience and tolerance is needed on the teacher's part.

Since some of the errors are also due to wrong teaching strategies methodological changes, including the use of the mother tongue, may profit from insights into the IL-phenomenon, and particularly into the transfer complex. German learners of English, for instance, have great difficulty in learning correctly the problematic continuous vs. non-continuous tenses (EF vs. SF). Here the problem of sequencing becomes obvious. "If the SF is first introduced because of its formal similarity to German, the later introduction of the EF may lead to under representation. If the contrary is done, overrepresentation may occur" (Zydatiss 1979, Vol.3:25-45). Though there is no didactically safe way out of this dilemma one might think of the two forms together with some cognitive insights into their formations and also different structural values within the English system perhaps concentrating more on one of the two forms. This would also correspond to the modern renaissance of the interest in grammar due to cognitive psychology, but would, at the same time, also fit into so called pragmatic approaches, since the habitual as

well as the non habitual functions are pragmatically needed at a relatively early point of time. This is one of many examples where due to wrong teaching strategies, one of the central processes of IL-production, errors may occur. Methodological changes as described above, cognitive insights, but also well-dosed drills including exercises like, for instance, well planned translations from L1 into L2 may help as remedial strategies.

Nickel notes that fossilization on a higher level very often happens also when a certain communicative level has been reached, motivation begins to decrease and a certain *plateau*-effect starts to operate. "All teachers have had this frustrating experience. While highly motivated learners want to go on in fields like subtle periphery grammar, synonymy and stylistics, the unmotivated ones hardly make any progress" (Nickel 1995:4-5).

Error evaluation, a highly subjective and complex field, should also profit from IL-insights. Though the distinction between errors (competence) and mistakes (performance), as once made by Corder, is not always easy to define, it is still useful from a pedagogical point of view. "The problematic evaluation of errors is undoubtedly primarily a pedagogical difficulty, acceptability and communicability, to mention only a few ones" (Legenhausen 1975:110).

According to Nickel (1985) the problem of native norms is also an important factor, accounting, among other reasons, for the differences in the marking of errors by different evaluators. "In this context it is worth mentioning that very often native speakers, especially of more tolerant linguistic backgrounds like English, turn out to be more tolerant than non-native speakers" (Nickel 1985, Vol. 67-68:153).

Nickel (1986) also states that it is needless to say IL's do not consist only of errors, but also of a normally overwhelming quantity of correct items, which should be given credit in the form of encouragement in order to motivate students. "In many school systems all over the world there is unfortunately still more emphasis given to the negative than to the positive aspect" (Nickel 1986:71-79).

"The IL-phenomenon also has a non-native normative aspect. While some 'fossilized' versions of varieties of English like, for instance, Indian English have been recognized as being institutionalised by some" (Kachru 1983:31-57), though not by all (Quirk 1988:237), "there is still some hesitancy in recognizing nationally conditioned accents like, for instance, the French one. Here again insights into the strong interference, particularly between the two phonological systems in connection with the IL-product 'French English' should, within certain limitations, lead to more tolerance, too, within certain limits" (Quirk et al 1985:27). Smith (1985) suggests that the realistic assessment of different varieties of English of particularly the FL-type have led to the not undisputed concept of EIL, which is understood as the use of relatively high-level IL's, employed for international communicative purposes, particularly, but not exclusively, among non-native speakers. It is mainly, again, the accent problem which arises here. On the whole, of course diversions should not be too strong in order to guarantee clear communication.

Selinker (1992) states that as has been demonstrated, IL has many more or less direct applications to FLT, helping people to understand difficult learning processes better and hence also to judge and evaluate IL-phenomena with more understanding. He continues that among several of learners' rights there is also the right of using IL's and hence making errors. Though striving for a native speaker's competence should in theory be the objective of FLT, teachers should be aware of all the limitations and expect 'perfection', if at all, only from a minority. In his opinion, taking the realistic existence of different types of IL's into account, however, does not mean embracing absolute fatalism. "In spite of a certain degree of inevitability, especially in connection with FLT, it is still true that some learners fossilize much less than others and then there is, of course, also the possibility for very good teachers to postpone or bypass some effects of fossilization" (Selinker.1992:252).

According to Garret/James (1991) apart from these more or less direct, though modest and limited, implications for FLT, IL-insights could also

be discussed in the interdisciplinary subject language awareness. Trim (1992:9) also suggests the same opinion on the matter. Nickel (1995) continues as in connection with the analysis of errors through the learners themselves or their peers, their corrections and explanations, problems and processes, as described further above in connection with language acquisition, could be discussed and explained. "The creative aspect of IL could also be shown by pointing to universal processes like, for instance, analogy" (Nickel 1995:6). He clarifies the notion as in the following:

"Observations of errors committed by speakers of other languages within the learner's culture, but also by very young mother-tongue learners, would show them the natural and necessary function of errors in a very complicated learning process. Needless to say, this kind of enlightenment should start as soon as possible, but, of course, in a correspondingly well-apportioned and adapted way. An age which tries to explain to young people the most complicated problems in, for instance, technology and ecology, should also find ways of explaining some of the fascinating, though by no means hitherto sufficiently explored, awe-inspiring processes of language acquisition. Tolerance towards errors committed by others will also lead to a better understanding among human beings and make learners recognize the wisdom of the old Latin adage: errare humanum est."

As far as faulty linguistic forms are concerned, Finocchiaro (1982) might tend to support the following view as:

"The world, our countries, our communities will survive with faulty pronunciation and less than perfect grammar, but can we be sure they will continue to survive without real communication, without a spirit of community, indeed without real communion among peoples?"

Undoubtedly the relationship between theory and teaching has always been complex and "establishing clear and usable categories from Interlanguage work...is a task which still needs to be attempted" (Brumfit 1984:323). According to Nickel (1995) nevertheless FLT, because of its own status and responsibility, has the right to tap, though cautiously, theories and descriptions whenever they help the teaching- and learning- process. "As with 'pedagogical grammars' simplifications of all kinds, controversial theories and even sometimes inconsistencies are justified as long as they serve the good purpose of increasing the learner's motivation and of teaching foreign languages efficiently" (Nickel 1995:7).

CHAPTER 3

METHODOLOGY

3.1 Procedure

3.1.1 Subjects

Data were collected from both Turkish and American university students. Only the advanced Turkish university students were chosen from among those who are in their fourth year after the Michigan Placement Test was administered. The American university students are all native speakers of English and were chosen from the LOCHNESS database collected at various American Universities.

3.1.2 Instruments

In this study, WordSmith Tools (Scott, 1997) was used for the lexical and syntactic analysis. Chi square (X^2) test was employed for the statistical analyses. Threshold level was set at $p < 0.05$ to determine the items overused and underused by Turkish learners.

3.1.3 Data Collection

Two comparable corpora were compiled: the NS and NNS corpora. The NS corpus consists of the argumentative essays of the American University students and the NNS corpus is made up of the argumentative essays of the advanced Turkish University students in the fourth year, who study English at the ELT departments of both the University of Çukurova and Mustafa Kemal University.

CHAPTER 4

DATA ANALYSIS

In this section, the occurrences of conjuncts in the NNS corpus are compared with that in the NS corpus to determine the NNS deviation from and approximation to the NS norm. To achieve this aim, along with the NS corpus used as a yardstick in the comparisons, the findings presented in Longman Spoken and Written English (LSWE) Corpus (Biber et al, 2000) are also employed to check the representativeness of the NS corpus and to make interpretations regarding the semantic functions and register appropriacy of the conjuncts under investigation (875-890)

Table 1. Overall comparison of conjuncts in the NNS and the NS corpora

Overall	NNS		NS		X ²	O/U
	f	%	f	%		
	1.278	0.84	805	0.54	100.0	+

4.1 Overall Analysis

As is seen in Table 1, the overall frequency of the conjuncts in NNS writing is considerably higher than that of NS writing. The very high X² value indicates that conjuncts are significantly overused in general. What could be the reason for the heavy use of conjuncts by the NNSs? Could Turkish learners be neglecting other important devices for creating textual cohesion such as coordinators and subordinators and therefore using conjuncts more than NSs? That is, could they be expressing their thoughts with short and simple sentences and, therefore, be feeling a need for employing more conjuncts to state the relationship between the units of discourse. Or could they be transferring their knowledge in L1 in the choice of such conjuncts in their writing. An answer to the second questions is beyond the scope of this study due to the lack of a corpus of academic prose in Turkish. In this

respect, considering the first question posed above, the interpretations as regards the occurrences of conjuncts in the NNS corpus are geared to to what extent such linguistic items are appropriate in terms of academic prose and what semantic meanings they add to learner's writing when compared with the NS corpus, and whether the overuse of conjuncts arises from the neglect of other cohesive devices. At this point, the individual comparison of these conjuncts in terms of their semantic categories will be more revealing.

Table 2. The occurrences of top 30 conjuncts across NNS and NS corpora

CONJUNCTS	NNS		NS		X ²	O/U
	f	%	f	%		
1 However	140	0.09	176	0.12	4.5	-
2 For example	136	0.09	52	0.03	35.3	+
3 Of course	82	0.05	25	0.02	28.4	+
4 So	79	0.05	74	0.05		=
5 Moreover	78	0.05	5		61.3	+
6 On the other hand	75	0.05	19	0.01	31.3	+
7 Therefore	73	0.05	81	0.05		=
8 For instance	59	0.04	11		30.8	+
9 Also	53	0.03	30	0.02	5.5	+
10 First of all	47	0.03	8		25.6	+
11 Further	39	0.03	3		28.6	+
12 Furthermore	39	0.03	3		28.6	+
13 In conclusion	38	0.03	9		16.2	+
14 In addition	35	0.02	7		16.9	+
15 As a result	32	0.02	7		14.4	+
16 Besides	27	0.02	0		24.6	+
17 Consequently	25	0.02	5		11.7	+
18 Secondly	23	0.02	1		18.0	+
19 Otherwise	20	0.01	9			=
20 To sum up	19	0.01	0		16.8	+
21 Firstly	19	0.01	1		14.2	+
22 That is	19	0.01	3		10	+
23 Then	18	0.01	6		4.9	+
24 First	12		1		7.5	+
25 Thirdly	11		0		8.9	+
26 Again	11		2		4.8	+

27 In other words	11		4			=
28 On the contrary	10		2		4.0	+
29 Eventually	9		22	0.01	4.8	-
30 Yet	6	0.02	63	0.04	15.2	-

4.2 Individual Analysis

As is obvious from the individual comparisons of the most frequent 30 conjunctions in Table 2 most of the conjuncts seem to be more favorite with the learners in the NNS corpora as opposed to the NSs with a few exceptions. The striking difference between the NS and the NNS corpora is particularly obvious in the use of *moreover*. The considerably high X^2 value (61.3) indicates that this conjunct is highly significantly overused in the NNS in contrast to the NS corpus. This conjunct is followed by such overused conjuncts as *moreover*, for example, for instance, on the other hand, further, furthermore, of course, first of all, besides, secondly, in addition, to sum up, in conclusion, as a result, firstly, consequently, that is, thirdly, first, also, then, again, on the contrary. It is also remarkable that, of these conjuncts, to sum up, besides and thirdly are not used by NS students in their essays.

As for the underused conjuncts, yet, eventually, however and therefore are not as common in the NNS corpus as in the NS. On the other hand, conjuncts such as so, in other words, otherwise and therefore are observed approximately in the same frequency in both corpora. In this respect, the NNSs do not show variation from the NS standard in the use of these conjuncts.

The investigation of conjuncts in the LSWE corpus reveals that although both conversation and academic prose mostly employ single adverbs to realize linking adverbials (conjuncts), academic prose displays greater structural diversity in the use of linking adverbials. That is, adverb phrases (e.g. Even so), prepositional phrases (e.g. For example), finite clauses (e.g. That is), and non-finite clauses (e.g. to conclude) commonly occur in academic prose. In general, the heavy reliance of the NNSs on linking adverbials when developing arguments or marking the connection

between the specific information and their point make them deviate from the native norm although almost all of these conjuncts characterize academic prose with exception of *then* and *so*. While *so* occurs as frequently in the NNS corpus as in the NS corpus, *then* particularly common in conversational discourse occur more frequently in the NNS corpus, which renders the NNS writing speech like. In order to shed light on the use of conjuncts in both the NNS and the NS writing, the following comparisons will focus on the semantic categories of conjuncts.

Table 3. Individual comparison of **Listing Conjuncts** according to their semantic categories

	CONJUNCTS	NNS		NS		X ²	O/U
		f	%	f	%		
Enumerative	First of all	47	0.03	8		25.6	+
	Secondly	23	0.02	1		18.0	+
	Firstly	19	0.01	1		14.2	+
	Then	18	0.01	6		4.9	+
	First	12		1		7.5	+
	Thirdly	11		0		8.9	+
Reinforcing	Moreover	78	0.05	5		61.3	+
	Also	53	0.03	30	0.02	5.5	+
	Further	39	0.03	3		28.6	+
	Furthermore	39	0.03	3		28.6	+
	In Addition	35	0.02	7		16.9	+
	Besides	27	0.02	0		24.6	+
	Again	11		2		4.8	+

4.3 Individual Comparison Of Listing Conjuncts

According to the X² values presented in Table 3, the conjuncts are more frequent with the learners in the NNSs as opposed to the NSs with no exceptions. The striking difference between the NS and the NNS corpora is particularly obvious in the use of *moreover* in the reinforcing category and *first of all* in the enumerative one. The considerably high X² values for first of

all (25.6) and *moreover* (61.3) indicate that these conjuncts are highly significantly overused in the NNS in contrast to the NS corpus. When we take the other conjuncts used by NNS into consideration, we can obviously see that, *first of all* is followed by such overused conjuncts as *secondly*, *firstly*, *thirdly*, *first* and *then*. And the case is the same for the conjuncts in the reinforcing category. The most frequent conjunct *moreover* is followed by the other linking adverbials such as *also*, *further*, *furthermore*, *in addition*, *besides* and *again*. Of these conjuncts, *further*, *furthermore*, *besides*, *in addition* are highly significantly overused by NNS. What is also striking is that *thirdly* and *besides* have never been used by NS.

According to Biber et al (2000), listing can be viewed as a sort of basic language function which the conjuncts are used to give a particular structure or orientation to. In this respect, Quirk et al (1997) note that the structure is supposed to show order by having items performing an ENUMERATIVE function. Quirk et al (1997) also point out that the enumerative function is expected to assign numerical labels to the items listed in our formal or informal speeches during the course of life. According to Quirk et al (1997) the reinforcing subtype of additive conjuncts, on the other hand, characteristically measure an item as adding greater weight to a preceding one. Findings in LSWE corpus reveal that enumerative/ additive (reinforcing) adverbials quite commonly occur in academic prose than conversation (Biber et al, 2000: 885).

Why do Turkish learners use conjuncts more heavily than native speakers? A closer look at the NNS writing might be revealing. Excerpts from the writings of the NS and the NNS show that the Turkish learners' writing is characterized by short sentences in contrast to the NS writing made up of relatively longer, more complex sentences. The following are the excerpts from both the NS and NNS corpora.

NNS e.g. 1: In this example, it can be seen only two bad sides of money. *First of all*, it makes people selfish. They only think themselves and they don't want to share their belongings with others who really need it. Secondly, you see the example of a miserable man who can do anything to have something to eat.

NS e.g. 1: They respond that, *first of all*, this statement can be used against teaching of evolution. There is a scientific theory that states evolution happened under conditions that do not exist today (because life does not originate naturally, as it supposedly did back then); thus, scientists cannot create the same conditions they assume were the original ones.

NNS e.g. 2: The court will decide whether the patient should die or not. So that, the application of euthenasia is not in our hands. You can just decide whether want to die or not and want him/her to die. *Moreover*, as it is not commonly used there are no special laws for euthenasia. If it is done by the patient himself/herself it is called suicide and if it is done by an another person especially by a doctor it is called as a murder.

NS e.g. 2: Twenty-five to forty percent of welfare recipients have learning disabilities, and sixteen percent have substance abuse problems. Thirty-five percent have physical disabilities or have a disabled person in their household. *Moreover*, since sixty-seven percent have never married and forty percent have three or more children, many lack the support of a spouse and have difficulty finding affordable childcare <R>.

Table 4. Individual comparison of Summative Conjuncts according to their semantic categories

	Conjuncts	NNS		NS		X ²	O/U
		f	%	f	%		
Summative							
	In conclusion	38	0.03	9		16.2	+
	To sum up	19	0.01	0		16.8	+

4.4 Individual Comparison Of Summative Conjuncts

Quirk et al (1997) summative conjuncts precedes an item which is to be looked at in relation to all the items that have gone before. Although the

summative conjuncts are of comparatively low frequency across NNS and NS corpora, two of these conjuncts, *in conclusion* and *to sum up*, are again significantly overused by the NNS. While *in conclusion* occurs only 9 times in the NS corpus, *to sum up* is never preferred by the NSs.

Biber et al. (2000) note that in an academic prose, prepositional phrases are relatively common as linking adverbials. Although summative adverbials are quite common in academic prose in LSWE corpus, they are not frequent in the NS writing. It is the NNSs who tend to use these conjuncts in their writing. Two conjuncts in conclusion and to sum up are particularly prominent in the NNS writing to introduce an item which embodies the points made before. Again the reason for the overuse of this category seems to be the NNSs' attempt to connect short and simple sentences in contrast to the NSs' more complex clauses already combined with subordinate conjunctions.

NNS e.g. 3: It is claimed that animal testing is unreliable because animals are different from human, but all mammals have the same basic organs. Moreover, they share common physiology with humans. *In conclusion*, the discussion between for and against animal testing is still going on.

NS e.g. 3: The idea of a nuclear war is practically non-existent today. Arms sales have gone down so that nations may spend this money in more worthwhile areas. *In conclusion*, as the turn of the century slowly approaches, I look back with a feeling of accomplishment as well as anticipation as to what lies ahead.

Table 5. Individual comparison of **appositive conjuncts** according to their semantic categories

Appositive	NNS			NS		X ²	O/U
	Conjuncts	f	%	f	%		
For example	136	0.09	52	0.03	35.3	+	
For instance	59	0.04	11		30.8	+	
That is	19	0.01	3		10	+	

4.5 Individual Comparison Of Appositive Conjuncts

From the table above, it is clear that the NNS corpus presents a heavy use of appositive conjuncts in comparison to that of the NS. The use of *for example* by NNS (136) is highly significantly overused, being approximately three times that of NS (52). The case is no different in the use of *for instance* by NS (11), which is only one fifth of the NNS' (59). The conjunct *that is* strikes attention with its few occurrences in both corpora although it is again overused by the NNS.

Biber et al (2000:881) state that appositional linking adverbials occur by far the most commonly in academic prose "as connectors for examples that support more general claims, and with restatements that clarify previous statements." This may be one of the reasons that the both groups tend to use them in their writings relatively more frequently than the other type of conjuncts discussed so far.

The following are two excerpts from the NNs and NS. It is noteworthy that what proceeds the conjunct in the NNS writing is a simple clause, while the same conjunct is proceeded by a more complex clause (ing-clause) constructed with structural links such as Wh-words. This may also account

for the more frequent use of conjuncts by the NNSs.

NNS e.g. 4 : In recent years money has become the most important thing in the world. You can do nothing without it. **For example**, you cannot eat or buy what you want, you cannot go anywhere you want, you cannot own even a needle. In short, you cannot live without it.

NS e.g. 4 : Upon realizing that homosexuals are not the only possible carriers of the disease, but drug abusers, people who interact w/ many unknown sexual partners, as well as dentists who "forget to put on their plastic gloves;" people may begin to think before acting. **For example**, drugs are a grave matter in their own, however, most people do not think in terms of drugs as a possible suicide.

Table 6. Individual comparison of resultive conjuncts according to their semantic categories

Resultive	Conjuncts	NNS		NS		X ²	O/U
		f	%	f	%		
	Of course	82	0.05	25	0.02	28.4	+
	So	79	0.05	74	0.05		=
	Therefore	73	0.05	81	0.05		=
	As a result	32	0.02	7		14.4	+
	Consequently	25	0.02	5		11.7	+

4.6 Individual Comparison Of Resultive Conjuncts

When resultive conjuncts are taken into consideration, the immediate difference is that certain conjuncts in the category occur as commonly in the NNS as in the NS corpus in comparison to the conjunct classes examined so far. *So* and *therefore* stand out as frequent conjuncts in both corpora, the difference indicating no significant variation between NNS and NS. The NNSs employ the conjunct *so* 79 times and *therefore* 73, which is approximately the same case in the NS' corpus (74 and 81, respectively). On the other hand, *of course*, *consequently* and *as a result* are overused by the NNS, *of course* being the one most significantly overused.

Biber et al, (2000:877) point out that "linking adverbials in the result category show that the second unit of discourse states the result or consequence of the preceding discourse". Findings in LSWE corpus (Biber et al, 2000) display that a large proportion of linking adverbials in the semantic category of result are used in both academic prose and conversation.

As is obvious in the excerpts below, these conjuncts serve as linking words between relatively shorter clauses in the NNS corpus, which, of course, necessitates the use of these conjuncts "for developing arguments or signaling the connection between specific information and an author's point" (Biber et al, 2000:881)

NNS e.g. 5 : If committing suicide is not an illegal or wrong thing, assisting suicide should be admitted in the same way. *Of course*, people who are depressed or who feel they are a weight on their families should be counseled and helped to live.

NS e.g. 5 : Another "discovery" of this past century has been feminism, which has had a significant impact on the lives of both men and women. *Of course*, the issue of feminism is not a new one, but it is just within the past century that much progress has been made towards the equality of women (with the important exception of women winning the right to vote in the 1800s).

Table 7. Individual comparison of **inferential conjuncts** according to the semantic categories

	Conjuncts	NNS		NS		X ²	O/U
		f	%	f	%		
Inferential	Otherwise	20	0.01	9			=
	In other words	11		4			=

4.7 Individual Comparison Of Inferential Conjuncts

From the table 7 it is seen that inferential conjuncts frequent are not frequent in both NNS and NS writing. Not only the NNSs but also the NSs focus merely on two conjuncts in their argumentative essays. Although the NNSs use both linking adverbials more than NNS do, there is no significant difference between the two groups with respect to X^2 analysis. In other words, neither conjunct is over/underused by the NNS learners. Biber et al, (2000:881) state that all semantic categories of linking adverbials help develop arguments or show the relation between specific information and an author's point in academic prose. It is noticeable that both only one "single linking adverbial" (otherwise) and one "prepositional phrase" (in other words) are employed in the NNS and NS writing. That these conjuncts do not take place among the most common linking adverbials in conversation and academic prose (Biber et al, 2000:887) may account for the fact that they do not occur frequently in the NNS and NS data.

The two excerpts from the NNS and NS data illustrate the use of "in other words" by a student from each group.

NNS e.g. 6: You haven't warned me about the danger; *in other words*, you didn't care about me.

NS e.g. 6: Another scientist, Solomon, also agrees with Forst's conclusion, but uses different reasoning. In 1975 he stated, <*>. *In other words*, Solomon believes some criminals may just want to die, and then murder so they can end up on death row.

Table 8. Individual comparison of **contrastive conjuncts** according to their semantic categories

	Conjuncts	NNS		NS		X ²	O/U
		f	%	f	%		
Concessive	However	140	0.09	176	0.12	4.5	-
	On the other hand	75	0.05	19	0.01	31.3	+
	Yet	6	0.02	63	0.04	15.2	-
Antithetic	On the contrary	10		2		4.0	+

4.8 Individual Comparison Of Contrastive Conjuncts

Table 8 illustrates that concessive conjuncts occur very frequently across the NNS and NS corpora. X² values indicate that *however* and *yet* are underused, while *on the other hand* is overused. Obviously, the NNSs use *on the other hand* to the neglect of other concessive conjuncts *however* and *yet*. However, in the antithetic category, *on the contrary* is infrequent in both corpora. Although rare, this conjunct is overused by the NNSs.

Following are examples of the usage of contrastive conjuncts from the NNS and NS data.

NNS e.g. 7: The patient who prefers euthanasia is thought to be committed suicide and the doctor who applies euthanasia to his/her patient is thought to be murderer and punished according to the law of 'killing people'; **however**, in some countries, 'euthanasia' is included in law; thus, people can choose to die. There are some organizations for euthanasia. Their aim is to defend the rights of volunteers of euthanasia.

NS: e.g. 7 : Proponents of prayer in public schools believe that a religious infusion is needed to balance the lack of values and the increasing rate of violence in society. The opponents hold, **however**, that prayer in public schools would destroy the separation of church and state, and that prayer will not be able to end the ills of society.

NNS e.g. 8 : As a first step, physical features built up differences between women and men. Also this affects the sex equality. The man and the women have same basic physical features. Like that they have head, to legs, arms, hair, eyes, ears, lungs, stomach etc. **On the contrary** these similarities there are certain differences. Such that the men have stronger musculars strong arms and legs.

NS e.g. 8 : PKD is one of the most common human genetically determined diseases. <*>. This sounds exciting to those individuals with the disease but it does not eliminate the disease. **On the contrary**, it gives the disease more opportunity to affect the generations to come because the individuals with the disease survive longer and can pass the disease on to more individuals

Table 9. Individual comparison of **transitional conjuncts** according to their semantic categories

	Conjuncts	NNS		NS		X ²	O/U
		f	%	F	%		
Temporal	Eventually	9		22	0.01	4.8	-

4.9 Individual Comparison Of Transitional Conjuncts

Amongst the transitional class conjuncts, only temporal subclass is observed in the writings of the NNSs and NSs. As is obvious from Table 9, the only conjunct *eventually* is underused by NNS. It has the frequency of 9 for NNS and 22 for NS.

Quirk et al (1997) note that temporal (transition) conjuncts such as *now, by the way, meantime, meanwhile, eventually* and etc., serve to shift attention to another topic or to a temporally related event. The low frequency of temporal conjuncts is also compatible with the corpus findings of Biber et al, who state that transition adverbials are rare in all registers (2000:880). The following excerpts illustrate the usage of the temporal conjuncts in both corpora.

NNS e.g 9: It was useful for that times because the area they were living was too small they could see the smoke and understood the message but after that they began to live different places and the places where they live began to be enlarged so they used pigeons in these times.

Eventually day by day the number of the people increase so that the placements of them also become bigger and bigger as a result of this development people need new communication aids.

NS e.g. 9: Not only that but since your body is now mostly empty of bacteria, the resistant strains multiply and make it more difficult to kill off the next time you get an infection.

Eventually the antibiotic no longer works and the doctor must use an alternative, perhaps stronger, antibiotic treatment.

Table 10. Overall comparison of conjuncts according to their semantic categories

Conjunctive Roles	NNS		NS		X ²	O/U
	f	%	f	%		
RESULTIVE	331	0.22	239	0.16	13.1	+
CONCESSIVE	293	0.19	289	0.19		=
REINFORCING	244	0.16	64	0.04	101.2	+
APPOSITIONAL	224	0.15	69	0.05	78.5	+
SUMMATIVE	70	0.05	22	0.01	23.2	+
ENUMERATIVE	36	0.02	51	0.03		=
INFERENCEAL	32	0.02	14		6.0	+
EQUATIVE	13		11			=
ANTITHETIC	11		8			=
TEMPORAL	9		22	0.01	4.8	-
REPLACIVE			3			=

4.10 Overall Comparison Of Conjuncts

The overall comparison of conjuncts in Table 1. has indicated that conjuncts are extremely overused by the NNSs. However, the comparison

according to semantic categories presents a different picture. *Resultive, reinforcing, appositional, summative* and *inferential* categories are all overused by NNS. However, there is no significant difference between the NNSs and NSs in terms of *concessive, enumerative, equative, antithetic* and *replacive* categories. The only category, which is underused is the *temporal*. The most striking overuse (X^2 101.2) results from the reinforcing conjuncts (moreover, also, further etc.). This category is followed by appositional (for example, etc.) (X^2 78.5), summative (to sum up, etc.) (X^2 23.2) and resultive (of course, so, etc) (X^2 13.1).



CHAPTER 5

CONCLUSIONS AND IMPLICATIONS

5.1 Introduction

In this section, the research questions are evaluated in the light of the findings obtained from the comparisons of the NNS and NS corpora. Also, the outcome of the study is discussed in terms of English learning and teaching.

5.2 Conclusions

The study has revealed that computer learner corpora can successfully be utilized to determine and diagnose the troublesome areas of overuse/underuse which may not be easily pinpointed otherwise. The study has shown that Turkish learners decidedly resort to putting sentences together by frequently using conjuncts. Samples from the NNS data has showed that the NNS writing is characterized by shortness of clauses and less embedding in contrast to the dense presentation of information in embedded clauses in the NS writing. In this section the research questions are evaluated. These research questions focus on whether conjuncts investigated are underused or overused. Also, the research questions explore the role of these conjuncts in the NNS and NS data from a functional perspective.

5.3 Evaluation Of The Research Questions

Q1. Do advanced Turkish learners of English employ conjuncts to the same extent as native American university students? If they avoid using certain conjuncts, what type of conjuncts do they avoid? And in what frequency do they use them?

The study has revealed that the NNSs use conjuncts more frequently than the NSs. However, the more detailed analysis has revealed that the NNSs have invariably overused all semantic categories except the temporal

one (Table 10). Yet, the concessive category has presented an interesting distribution in that *however* and *yet* have been significantly underused, while *on the other hand* has been significantly overused (Table 8). While the NSs have employed *however* and *yet* to highlight contrasting information, the NNSs have preferred *on the other hand*. The most common linking adverbials (conjuncts) in conversation and academic prose in the LSWE corpus per million words include only *however* and *yet*, both having far more occurrences in academic prose, particularly *however*.

Q2. What kind of semantic roles do conjuncts play in NNS and NS argumentative essays? And to what extent do they match in both corpora as far as the semantic relations are concerned?

Resultive, reinforcing, appositional, summative and inferential categories are all overused by NNS. The overuse of these categories suggest that the NNSs often tend to add items of discourse to one another, reformulate and exemplify their statements, sum up the information in the proceeding discourse and show the result or consequence of the proceeding discourse. This pattern of use in the NNS writing renders it strikingly different from the NS academic prose.

Q3. What semantic categories of conjuncts characterize both the NNS and NS argumentative essays?

The NNS writing is characterized by the heavy use of resultive, reinforcing, appositional, summative and inferential categories. Of the contrastive conjuncts, the concessive category particularly characterize the NS argumentative essays. This category is underused by the NNSs with the exception of *on the other hand*, which is significantly overused.

Q4. If certain categories are overused/underused, what might be the reason?

The excerpts from the learner data has indicated that short and simple clauses characterize the NNS writing. That's why, the NNSs seem to be in constant need for connecting units of discourse to achieve a textual flow.

5.4 Implications

The present study has provided concrete evidence and useful statistics as to the use of conjuncts beyond intuitions by means of NL (native language) vs. IL (interlanguage) comparison. The analyses have shown that the English of advanced Turkish learners share a number of features which make it different from NS language in the use of conjuncts. The study has found that advanced Turkish learners of English either overuse or underuse conjuncts, which leads to stylistic deviations.

What seems to be a major problem for the Turkish learners is their lack of register awareness. The main pedagogical conclusion to be drawn from this study is that Turkish EFL students need to be exposed to a wide range of registers and to a more extensive training in argumentative or expository writing.

REFERENCES

ABERCROMBIE, D. (1963), **Studies in Phonetics and Linguistics**, London: Oxford University Press.

AARTS, J. and MEIJS, W. (eds) (1986), **Corpus Linguistics II**, Amsterdam: Rodopi Press.

AIJMER, K. and ALTENBERG, B. (eds) (1991), **English Corpus Linguistics**, London: Studies in Honour of Jan Svartvik, Longman.

ALTENBERG, B. (1984), "Causal linking in spoken and written English", **Studia Linguistica Texts** 38, pp.20-69.

ANTAKI, C. and NAJI, S. (1987), "Events explained in conversational "because" statements", **British Journal of Social Psychology** 26, pp. 119-126.

ATKINS, B. T. S. and LEVIN, B. (1995), "Building on a corpus: a linguistic and lexicographical look at some near-synonyms", **International Journal of Lexicography** 8:2, 85-114.

ALTENBERG, Bengt & GRANGER, Sylviane (2001), "The grammatical and lexical patterning of make in native and non-native student writing", **Applied linguistics**, Vol. 22, No. 2, pp. 173-194

ASTON, Guy (1997), **Enriching the Learning Environment**, London: Longman

BARLOW, Michael (1992), **Using Concordance Software in Language Teaching and Research**, Kasugai, Japan: LLAJ & IALL Press.

BIBER, Douglas & CONRAD, Susan (2001), "Corpus based research in TESOL", **TESOL Quarterly**, Vol. 35, No. 2, pp. 331-335.

BIBER, Douglas & CONRAD, Susan & REPPEN, Randi (1998), **Corpus linguistics: Investigating Language Structure and Use**, Cambridge.

BLOOM, L. (1970), **Language Development, Form and Function in Emerging Grammars**, Cambridge: MA-MIT Press.

BOAS, F. (1940), **Race, Language and Culture**, New York: Macmillan Press.

BONGERS, H. (1947), **The History and Principles of Vocabulary Control**, Worden: Wocopi.

BROWN, R. (1973), **A First Language: The Early Stages**, Cambridge: Harvard University Press.

BROWN, P. F.; COCKE, J.; PIETRA, S. A. D.; PIETRA, V. J. D.; JELINEK, F.; LAFFERTY, J.D.; MERCER, R. L.; and ROOSSIN, P. S. (1990), "A Statistical Approach To Machine Translation", **Computational Linguistics** 16(2):79-85.

BRUMFIT, Christopher (1984), **Theoretical Implications of Interlanguage Studies For Language Learning**. Interlanguage ed. C. Cramer A. Davies A.P.R. Howatt, 312-323. Edinburgh: Edinburgh University Press.

BALLESTEROS, L., and CROFT, B. (1996), "Dictionary methods for cross-lingual information retrieval", **Proceedings of the 7th International DEXA Conference on Database and Expert Systems**, 791-801. [http://ciir.cs.umass.edu/\[0\]info/psfiles/\[0\]irpubs/ir.html](http://ciir.cs.umass.edu/[0]info/psfiles/[0]irpubs/ir.html), (23 July 2002)

BALLESTEROS, L., and CROFT, W. B. (1997), "Phrasal translation and query expansion techniques for cross-language information retrieval", In **AAAI Symposium on Cross-Language Text and Speech Retrieval**. American Association for Artificial Intelligence. To appear. [http://www.ee.umd.edu/\[0\]medlab/filter/sss/\[0\]papers/ballesteros.ps](http://www.ee.umd.edu/[0]medlab/filter/sss/[0]papers/ballesteros.ps). (23 July 2002)

CHOMSKY, N. (1964), **Formal Discussion**, in Bellugi, U and Brown R. (eds.) *The Acquisition of Language*. Monographs of the Society for Research in Child Development 29. pp 37-9

CHOMSKY, N. (1965), **Aspects of the Theory of Syntax**, Cambridge, MA: MIT Press.

CHOMSKY, N. (1968), **Language and Mind**, Harcourt Brace, New York.

CONRAD, Susan (2000), "Will corpus linguistics revolutionize grammar teaching in the 21st century?" **TESOL Quarterly** Vol. 34, pp. 548-560

DAVIS, M. W., and OGDEN, W. C. (1997), Implementing cross-language text retrieval systems for large-scale text collections and the world wide web. In **AAAI Symposium on Cross-Language Text and Speech Retrieval**. American Association for Artificial Intelligence. To appear. [http://www.ee.umd.edu/\[0\]medlab/filter/sss/papers/\[0\]davis.ps](http://www.ee.umd.edu/[0]medlab/filter/sss/papers/[0]davis.ps).

DAVIS, M. (1996), New experiments in cross-language text retrieval at NMSU's Computing Research Lab. In Harman, D. K., ed., **The Fifth Text REtrieval Conference (TREC-5)**. NIST. To appear.

DUMAIS, S. T.; LETSCHE, T. A.; LITTMAN, M. L.; and LANDAUER, T. K. (1997), Automatic cross-language retrieval using latent semantic indexing. In **AAAI Symposium on Cross-Language Text and Speech Retrieval**.

American Association for Artificial Intelligence. To appear.
[http://www.ee.umd.edu/\[0\]medlab/filter/sss/\[0\]papers/dumais.ps](http://www.ee.umd.edu/[0]medlab/filter/sss/[0]papers/dumais.ps).

DUMAIS, S. T.; LANDAUER, T. K.; and LITTMAN, M. L. (1996), Automatic cross-linguistic information retrieval using latent semantic indexing. In Grefenstette, G., ed., **Working Notes of the Workshop on Cross-Linguistic Information Retrieval**. ACMSIGIR. [http://superbook.bellcore.com/std/\[0\]papers/SIGIR96.ps](http://superbook.bellcore.com/std/[0]papers/SIGIR96.ps)

DUMAIS, S. T. (1994), Latent Semantic Indexing (LSI): TREC-3 report. In Harman, D., ed., **Overview of the Third Text REtrieval Conference**, 219--230. NIST. [http://potomac.ncsl.nist.gov/\[0\]TREC/](http://potomac.ncsl.nist.gov/[0]TREC/).

DULAY, H. / BURT, M. / KRASHEN, S. (1982), **Language Two**. New York & Oxford: Oxford University Press.

FILLMORE, Charles J. 1992, "Corpus linguistics" or "Computer-aided armchair linguistics", in: Svartvik, Jan. (ed.) **Directions in Corpus Linguistics**. Berlin/New York, 35)

FREDERKING, R.; MITAMURA, T.; NYBERG, E.; and CARBONELL, J. (1997), Translingual information access. In **AAAI Symposium on Cross-Language Text and Speech Retrieval**. American Association for Artificial Intelligence. To appear. [http://www.ee.umd.edu/\[0\]medlab/filter/sss/\[0\]papers/frederking.ps](http://www.ee.umd.edu/[0]medlab/filter/sss/[0]papers/frederking.ps).

FOX, Gwyneth (1998), Using corpus data in the classroom, In Brian Tomlinson (Ed.) **Materials development in language teaching**, Cambridge

FINOCCHIARO, Mary. (1982). "Reflections on the past, the present, and the Future". **The language Teacher** 20(3). 4-9.

FRIES, C. and TRAVER, A. (1940), **English Word Lists: A Study of their Adaptability and Instruction**, Washington, DC: American Council of Education.

GARRETT, P., JAMES, C., eds. (1991), **Language Awareness in the Classroom**. London: Longman.

GASS, Susan & LARRY Selinker, eds. (1983), **Language Transfer in Language Learning**. Rowley: Newbury House Publishers.

GRANGER, Sylviane (ed) (1998), **Learner English on Computer**. Longman, London.

GILARRANZ, J.; GONZALO, J.; and VERDEJO, F. (1997), An approach to conceptual text retrieval using the eurowordnet multilingual semantic

database. In **AAAI Symposium on Cross-Language Text and Speech retrieval**. American Association for Artificial Intelligence. To appear. [http://www.ee.umd.edu/\[0\]medlab/filter/sss/\[0\]papers/gilarranz.ps](http://www.ee.umd.edu/[0]medlab/filter/sss/[0]papers/gilarranz.ps).

GARNHAM, A., SHILLOCK, R., BROWN, G., MILL, A. and CUTLER, A. (1981), "Slips of the tongue in the London-Lund corpus of spontaneous conversation", **Linguistics** 19: 805-17.

HAYASHI, Y.; KIKUI, G.; and SUSAKI, S. (1997), Titan: A cross-linguistic search engine for the www. In **AAAI Symposium on Cross-Language Text and Speech Retrieval**. American Association for Artificial Intelligence. To appear. [http://www.ee.umd.edu/\[0\]medlab/filter/sss/\[0\]papers/hayashi.ps](http://www.ee.umd.edu/[0]medlab/filter/sss/[0]papers/hayashi.ps).

Hlava, M. M. K.; HAINEBACH, R.; BELONOGOV, G.; and KUZNETSOV, B. (1997), Cross-language retrieval - English/Russian/French. In **AAAI Symposium on Cross-Language Text and Speech Retrieval**. American Association for Artificial Intelligence. To appear. [http://www.ee.edu/\[0\]medlab/filter/sss/\[0\]papers/hlava.ps](http://www.ee.edu/[0]medlab/filter/sss/[0]papers/hlava.ps).

HULL, D. A., and GREFENSTETTE, G. (1996), "Experiments in multilingual information retrieval". In **Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. [http://www.xerox.fr/\[0\]grenoble/mltt/people/\[0\]hull/papers/sigir96.ps](http://www.xerox.fr/[0]grenoble/mltt/people/[0]hull/papers/sigir96.ps).

HAMMERLY, Hector. (1991), **Fluency and Accuracy**. (Multilingual Matters, 73). Clevedon & Philadelphia: Multilingual Matters LTD.

HARRIS, Z. (1951), **Methods in Structural Linguistics**, Chicago: University of Chicago Press.

HOCKETT, C (1948), "A note on structure", **International Journal of American Linguistics** 14: 269-71.

HALLIDAY, M. (1991), **Corpus studies and probabilistic grammar**, in Aijmer and Altenberg 1991, pp 30-43.

HOFLAND, K. and JOHANSSON, S. (1982), **Word Frequencies in British and American English**, Bergen: Norwegian Computing Centre for the Humanities.

HOLMES, J. (1988), "Doubt and certainty in ESL textbooks", **Applied Linguistics** 9: 21-44.

HOLMES, J. (1994), "Inferring language change from computer corpora: some methodological problems", **ICAME Journal** 18: 27-40.

INGRAM, D. (1978), **Sensori-motor development and language acquisition**, in Lock, A (ed) *Action, Gesture and Symbol: The Emergence of Language*, pp261-90, London: Academic Press.

INGRAM, D. (1989), **First Language Acquisition**, Cambridge University Press, Cambridge.

JOHANSSON, S. (1991), "Times change and so do corpora", in Aijmer and Altenburg (eds.) **English corpus linguistics: studies in Honour of Jan**

JOHANSSON, S. and NORHEIM, E. (1988), "The subjunctive in British and American English", **ICAME Journal** 12: 27-36.

JOHANSSON, S. and STENSTRÖM, A-B. (eds) (1991), **English Computer Corpora: Selected Papers and Research Guide**, Berlin: Mouton de Gruyter.

KAMEYAMA, M. (1997), Information extraction across linguistic barriers. In **AAAI Symposium on Cross-Language Text and Speech Retrieval**. American Association for Artificial Intelligence. To appear. [http://www.ee.umd.edu/\[0\]medlab/filter/sss/papers/\[0\]kameyama.ps](http://www.ee.umd.edu/[0]medlab/filter/sss/papers/[0]kameyama.ps).

KIKUI, G. (1996), Identifying the coding system and language of on-line documents on the internet. In **Sixteenth International Conference of Computational Linguistics (COLING)**. International Committee on Computational Linguistics. [http://isserv.tas.ntt.jp/chisho/\[0\]paper/9608KikuiCOLING.ps.Z](http://isserv.tas.ntt.jp/chisho/[0]paper/9608KikuiCOLING.ps.Z).

KACHRU, Braj. (1983), **The other Tongue**. Oxford: Pergamon.

KENNEDY, G. (1987), "Expressing temporal frequency in academic English", **TESOL Quarterly** 21: 69-86.

KENNEDY, G. (1987), "Quantification and the use of English: a case study of one aspect of the learner's task", **Applied Linguistics** 8: 264-86.

KENNEDY, G. (1992), "Preferred ways of putting things", in Svartvik J. (ed) **Directions in Corpus Linguistics**, pp 335-73, Berlin: Mouton de Gruyter.

KENNEDY, G. (1998), **An Introduction to Corpus Linguistics**, Longman, London.

KIRK, J. (1994), **Teaching and language corpora: the Queen's approach**, in Wilson and McEnery 1994, pp 29-51.

KJELLMER, G. (1986), **The lesser man: observations on the role of women in modern English writings**", in Arts and Meijs 1986, pp 163-76.

- KRIEGER, Daniel (2003), **The Internet TESL Journal**, Vol. IX, .
No. <http://iteslj.org/> <http://iteslj.org/Articles/Krieger-Corpus.html>. 3 March 2003
- KYTÖ, M., Rissanen, M. and Wright, S. (eds) (1994), **Corpora across the Centuries**, Amsterdam, Rodopi.
- LANDAUER, T. K., and LITTMAN, M. L. (1990), Fully automatic cross-language document retrieval using latent semantic indexing. In **Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research**. Waterloo Ontario: UW Centre for theNewOEDandTextResearch. 31--38. [http://www.cs.duke.edu/mlittman/\[0\]docs/x-lang.ps](http://www.cs.duke.edu/mlittman/[0]docs/x-lang.ps).
- LANDAUER, T. K., and LITTMAN, M. L. (1991), "A statistical method for language-independent representation of the topical content of text segments". In **Proceedings of the Eleventh International Conference: Expert Systems and Their Applications**, volume 8, 77--85.
- LEECH, Geoffrey (1997), "Teaching in language corpora": a convergence, In Gerry Knowles, Tony Mcenery, Stephen Fligelstone, Anne Wichman, (Eds.) **Teaching and language corpora** . Longman pp. 1-22
- LABOV, V. (1969), "The logic of non-standard English", **Georgetown Monographs on Language and Linguistics** 22.
- LEECH, G. (1991), **The state of the art in corpus linguistics**, in Aijmer K. and Altenberg B. (eds.) **English Corpus Linguistics: Studies in Honour of Jan Svartvik**, pp 8-29. London: Longman.
- LEECH, G. (1992), **Corpora and theories of linguistic performance**, in Svartvik, J. **Directions in Corpus Linguistics**, pp 105-22. Berlin: Mouton de Gruyter.
- LEECH, G. and FALLON, R. (1992), "Computer corpora - what do they tell us about culture?", **ICAME Journal** 16: 29-50.
- LEECH, G. and Short, M. (1981), **Style in Fiction**, London: Longman.
- LEITNER, G. (1991), **The Kolhapur corpus of Indian English: intravarietal description and/or intervareital comparison**, in Johansson and Stenström 1991, pp 215-32.
- MILTON, J. (1998), **Exploring L1 and interlanguage corpora in the design of an electronic language learning and production environment**. In S. Granger (ed), **Learner English on Computer**. Longman, London.

McENERY, A. and WILSON, A. (1993), "The role of corpora in computer-assisted language learning", **Computer Assisted Language Learning** 6(3): 233-48.

McENERY, A., BAKER, P. and WILSON, A. (1995), "A statistical analysis of corpus based computer vs traditional human teaching methods of part of speech analysis", **Computer Assisted Language Learning** 8(2/3):259-74.

McENERY, T. and WILSON, A. (1996), **Corpus Linguistics**. Edinburgh University Press.

MARCUS, R. S. (1994), Intelligent assistance for document retrieval based on contextual, structural, interactive Boolean models. In *RIAO 94 Conference Proceedings, Intelligent Multimedia Information Retrieval Systems and Management*, volume 2, 27-43. Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaire (C.I.D.).

MEIJS, W. (ed) (1987), **Corpus Linguistics and Beyond**, Amsterdam: Rodopi.

MINDT, D. (1991), **Syntactic evidence for semantic distinctions in English**, in Aijmer and Altenburg 1991, pp 182-96.

MINDT, D. (1992), **Zeitbezug im Englischen: eine didaktische Grammatik des englischen Futurs**, Tübingen: Gunter Narr.

MYERS, G. (1991), **Pragmatics and corpora**, talk given at **Corpus Linguistics Research Group**, Lancaster University.

MCCARTHY, Michael & CARTER, Ronald (2001), "Size isn't everything: spoken English, corpus, and the classroom". **TESOL Quarterly** Vol. 35, No. 2, pp. 337-340

MINDT, Dieter (1997), **Corpora and the teaching of English in Germany**, In Gerry Knowles, Tony Mcenery, Stephen Fligelstone, Anne Wichman, (Eds.) *Teaching and language corpora* . Longman pp. 40-50

NEUFELD, Gerald. (1987), **On the Aquisition of Prosodic and Articulatory Features in Adult language Learning**. Interlanguage Phonology ed. By Georgette loup & Steven H. Weinberger, 321-332. Cambridge: Newbury House Publishers.

NICKEL, G./ Wagner, K. H. (1968), "Contrastive Linguistics and Language Learning". **IRAL** 6(3). 233-255.

NICKEL, Gerhard. (1985), "How Native can (or Should) a Non-Native Speaker be?". *ITL* 67-68. 141-160.

NICKEL, G./ Stalker, J. (1986), **Problems of Standardization and Linguistic Variation in present-Day English**. Heidelberg: Julius Gross.

NICKEL, Gerhard. (1989), "Some Controversies in present-Day Error Analysis: Contrastive vs. Non-Contrastive Errors". *IRAL* 27 (4). 293-305.

NICKEL, Gerhard. (1992), 'Contrastive versus Non-Contrastive Errors: A Controversy in Error Analysis (EA)'. *South African Journal of Linguistic* 10 (4). 229-234.

NICKEL, Gerhard. (1992), **The role of Simplification in Connection with English as FL**. OmSprak og Utdanning. Universitetsforlaget Oslo 163-175.
Quirk, Randolph. et alii 1985. *A Comprehensive Grammar of English*. New York: Longman.

NATION, I.S.P (2001), **Learning vocabulary in another language**. Cambridge

O'CONNOR, J. and ARNOLD, G. (1961) **Intonation of Colloquial English**, London: Longman.

OOSTDIJK, N. and de HAAN, P. (1994a), "Clause patterns in modern British English: a corpus-based (quantitative) study", *ICAME Journal* 18: 41-79.

OOSTDIJK, N. and de HAAN, P. (eds) (1994b), **Corpus Based Research into Language**, Amsterdam: Rodopi.

OARD, D. W., and DORR, B. J. (1996a), Evaluating cross-language text filtering effectiveness. In Grefenstette, G., ed., **Proceedings of the Cross-Linguistic Multilingual Information Retrieval Workshop**. ACM SIGIR. [http://www.ee.umd.edu/\[0\]medlab/filter/papers/sigir96.ps](http://www.ee.umd.edu/[0]medlab/filter/papers/sigir96.ps).

OARD, D. W., and DORR, B. J. (1996b), **A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19**, University of Maryland, Institute for Advanced Computer Studies. [http://www.ee.umd.edu/\[0\]medlab/\[0\]filter/\[0\]papers/mlir.ps](http://www.ee.umd.edu/[0]medlab/[0]filter/[0]papers/mlir.ps).

OARD, D. W. (1996), "Adaptive Vector Space Text Filtering for Monolingual and Cross Language Applications". **Ph.D. Dissertation**, University of Maryland, College Park. [http://www.ee.umd.edu/\[0\]medlab/filter/papers/thesis.ps.gz](http://www.ee.umd.edu/[0]medlab/filter/papers/thesis.ps.gz).

OARD, D. W. (1997a), "Adaptive filtering of multilingual document streams". In Submitted to **RIAO 97**.

PALMER, H. (1933), "Second Interim Report on English Collocations", Tokyo: **Institute for Research in English Teaching**.

PEITSARA, K. (1993), **On the development of the by-agent in English**, in Rissanen, Kytö and Palander-Collin 1993 pp 217-33.

POLLITT, A. S., and ELLIS, G. (1993), Multilingual access to document databases. In **21st Annual Conference Canadian Society for Information Science**, 128—140

QUIRK, R. (1960), "Towards a description of English usage", **Transactions of the Philological Society**, pp 40-61.

QUIRK, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985), **A Comprehensive Grammar of the English Language**, London, Longman.

QUIRK, Randolph. (1988), **The Questions of Standards in the International use of English**. Language Spread and Language Policy ed. P. H. Lowenberg, 229-241. Washington D.C.: Georgetown University Press.

QUIRK et. al., (1997) **A comprehension of the English Grammar**. Longman. London.

RADWAN, K., and FLUHR, C. (1995), Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval. In **Fourth Annual Symposium on Document Analysis and Information Retrieval**, 121--136.

RISSANEN, M. (1989), Three problems connected with the use of diachronic corpora, **ICAME Journal** 13: 16-19.

RISSANEN, M., KYTÖ, M. and Palander-Collin, M. (eds) (1993), **Early English in the Computer Age**, Berlin, Mouton de Gruyter.

SELINKER, Larry. (1972), "Interlanguage". **IRAL** 10(3). 209-230.

SELINKER, Larry. (1992), **Rediscovering Interlanguage**. London & New York: Longman.

SHARWOOD Smith, Michael. (1994), "Second Language Learning": Theoretical Foundations. London & New York: Longman. Smith, Larry. (1985). EIL versus ESL/EFL: What's the Difference and what Difference does the Difference make?. **English Teaching Forum** 23(4). 2-6.

SAMPSON, G. (1992), "Probabilistic parsing", in Svartvik, J. *Directions in Corpus Linguistics*, pp 425-47. Berlin: Mouton de Gruyter. Svartvik, J. (1966) *On Voice in the English Verb*, The Hague: Mouton.

SVARTVIK, London, Longman. pp 305-14. KADING, J. (1879), *Haufigkeitswörterbuch der deutschen Sprache*, Steglitz: privately published.

SHERIDAN, P.; WECHSLER, M.; and SCHÄUBLE, P. (1997), Cross-language speech retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence. To appear. [http://www.ee.umd.edu/\[0\]medlab/filter/sss/\[0\]papers/she_ridan.ps](http://www.ee.umd.edu/[0]medlab/filter/sss/[0]papers/she_ridan.ps).

SOERGEL, D. (1997), Multilingual thesauri in cross-language text and speech retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence. To appear. [http://www.ee.umd.edu/\[0\]medlab/filter/sss/\[0\]papers/soe_rgel.ps](http://www.ee.umd.edu/[0]medlab/filter/sss/[0]papers/soe_rgel.ps).

SRIDHAR, S. N. (1981), *Contrastive Analysis, Error Analysis and Interlanguage: Three Phases of a Goal*. Contrastive Linguistics and the Language Teacher ed. By Jacek Fisiak, 207-241. Oxford & New York: Pergamon Press.

SINCLAIR, John (1998), *Corpus evidence in language description*, In Gerry Knowles, Tony Mcenery, Stephen Fligelstone, Anne Wichman, (Eds.) *Teaching and language corpora*, Longman pp. 27-39

STEVENS, Vance (1995), Concordancing with language learners: Why? When? What? *CAELL Journal* Vol 6, No. 2 pp. 2-10.

STEVENS, Vance (1991), Classroom concordancing: Vocabulary materials derived from relevant, authentic text. *English for Specific Purposes* Vol. 10, pp. 35-46.

SCHREUDER, R. and Kerkman, H. (1987), *On the use of a lexical database in psycholinguistic research*, in Meijs 1987, pp 295-302.

STENSTÖM, A-B. (1984), "Discourse items and pauses", Paper presented at Fifth ICAME Conference, Windermere. Abstract in *ICAME News* 9 (1985): 11.

STENSTÖM, A-B. (1987), "Carry-on signals in English Conversation", in *Meijs 1987*, pp 87-119.

TRIM, John. (1992), "Language Teaching in the perspective of the Predictable Requirements of the Twenty-First Century". *AILA* 9.7-20.

THORNDIKE, E. (1921), **A Teacher's Wordbook**, New York: Columbia Teachers College.

THURSTON, Jennifer & CANDLIN, Christopher (1998), "Concordancing and the teaching of the vocabulary of academic English". **English for Specific Purposes** Vol. 17, No. 3, pp. 267-280

TOTTIE, G. (1991), **Negation in English Speech and Writing: A study in Variation**, San Diego: Academic Press.

VAN DER EIJK, P. (1993), "Automating the acquisition of bilingual terminology". In **Sixth Conference of the European Chapter of the Association for Computational Linguistics**, 113--119.

WILLIS, Jane (1998), **Concordances in the classroom without a computer**, In Brian Tomlinson (Ed.) **Materials development in language teaching**, Cambridge

WILSON, A. (1992), "The Usage of Since: A Quantitative Comparison of Augustan, Modern British and Modern Indian English", **Lancaster Papers in Linguistics** 80.

WILSON, A and McENERY, A. (eds) (1994), "Corpora in Language Education and Research": A Selection of Papers from Talc94, Unit for Computer Research on the **English Language Technical Papers 4** (special issue), Lancaster University.

YOUNG, R. (1988), "Variation and the Interlanguage Hypothesis". **Studies in Second Language Acquisition** 10. 281-301.

ZYDATISS, Wolfgang. (1979), "Learning Problem Expanded Form – A Performance Analysis". **Studies in Descriptive Linguistics** 3. 25-45.

APPENDICES

Appendix 1 3 sample argumentative essays by native American University students.

<ICLE-US-IND-0024.1>

The first thing I would like to address is the fact that this phrase is incorrect: "The Love of Money Is the Root of All Evil". I believe this corrected statement to be true. However, since the original subject is <*> I will have to disagree with this statement.

If money were the root of all evil, we all would have to be evil because we all have money. Some have more than others; some have less than others. Most money is attained in honest ways: through employment, investments, inheritances, and many other lucrative means. When money is accumulated by these means, it seems incorrect to say that money is the root of all evil. If one is attaining money honestly, how can honestly be associated with evil?

It is true to say that money is the root of some evil. People have lied, cheated, betrayed, stolen, and even murdered to attain money. These actions are evil and because someone wanted money, or wanted more money than they had, they resorted to evil to attain it.

Other things as well can be the root of evil: attitudes, morals (lack of), greed; there is an endless list of things that can become the root of evil. The love of money can indeed produce an attitude of dishonesty, a lack of morals, and greed.

If money was the root of all evil, then the many agencies, groups and programs that we have in America and around the world would not be able to function without the donations of money from those who are able to give. I have to agree that many times these worth-while causes are sometimes being used to satisfy an individual's greed; but I am sure that most if not all

funds donated are used for the purposes intended in those agencies and programs which are reputable.

In conclusion, the root of all evil is the choice of the individual: will he choose to be morally decent, or will he choose to be evil?

<ICLE-US-MICH-0003.1>

I feel the invention of the computer has significantly changed people's lives. This, realistically, has been quite a new invention, and in its short life span has been able to change the world significantly. Not to mention how it will change people's life styles in the future. These changes resulting from the invention of the computer, have been, I feel both bad and good.

Technologically, computers have improved people's lives drastically. No longer are the times when people must write a paper manually, or even with a simple typewriter. Personal computers are becoming quite common, and therefore less expensive and easier to own. Computers have also made live transpire much quicker than before. Gone are the days when a person who wanted money from a bank had to wait until it was open. We now have ATM, automatic teller machines that are able to proceed with the same transactions as a teller would do. It seems these days EVERYTHING is run by computers! We are able to complete our tasks much better and much more efficiently. Basically every facet of life has been affected by computers bringing about positive results.

While computers have brought about improvements, in my opinion, they also have "impersonalized" the world. Jobs are being taken away from humans who are not "fast enough". Interaction with humans in some places is extinct. Yes, it might be more efficient to have a computer make your transaction at a bank, but I miss the human interaction between people. In addition, it is also very dangerous to place all of one's trust in a machine. It is a common fact machines do fail! Sure, a car with a computer that tells the driver when an on-coming obstacle is approaching is safe, but what happens when the driver relies too much on the computer and not enough on their

own brain? People soon stop thinking, and allow machines to do the work for them. I feel afraid when our society relies so heavily on a bunch of wires. Computers are not able to express and truly feel emotions, and therefore are incapable, in some instances to function. With the invention of the computer comes many changes from the lifestyle of the past to changes in the lifestyle for the future. Some of these changes will save people's lives, such as advances in medical technology, safety, etc. But I fear that our world will become a very impersonal place. I hope that we will be able to determine what improvements should be made and what should just remain "slow".

<ICLE-US-SCU-0012.2>

Sometimes when people suffer from extreme illnesses they look for other options as an easy way out. When the pain becomes unbearable and the medication stops working some people would rather die than to continue to live in pain. Those patient's who desire to end their lives do not wish to do so on their own. They prefer to have an easy and painless death, and so they choose a form of suicide which is referred to as euthanasia. Euthanasia is a "mercy killing" for the purpose of putting an end to extreme suffering. Many of the cases concerning euthanasia in the past few years have been performed or assisted by a doctor known as Dr. Kevorkian. There has been much controversy over assisted suicides, especially those performed with the assistance of a doctor. Should it be legalized to allow doctors to assist in suicides of terminally ill patients, or should they be bound by their Hippocratic Oath?

There are several different forms of euthanasia. One form of euthanasia is when a person's life is in a permanently unconscious state, sometimes referred to as a "vegetable." When a person is in this state, if they do not have a living will stating their wishes, the family has the right to make the decision concerning whether the patient should remain on life support or not after they have consulted with the physician concerning the matter. There

are very few legal objections to this form of euthanasia. It is widely practiced in hospitals all over America. Many people leave in their will expressed wishes to not be kept alive on life support systems. When the physician determines that there is no chance of a patient's recovery, it is their duty to carry out the patient's wishes and disconnect life support. Another type of euthanasia is when a patient requests that the doctor provide no more treatment. If medical conduct to end the patient's life is prohibited, the patient is allowed maximum opportunity to change his mind and demand treatment. The patient declining treatment normally remains alive for a period and thereby receives some opportunity to change one's mind or to eliminate any mistake on the physicians part in comprehending the patient's wishes. A third type of euthanasia is self-inflicted by one's ending of one's life by means of suicide. This type of euthanasia is occurring more frequently because it is seen as the easiest way out.

William F. May wrote in his essay, The Right to Die and the Obligation to Care: Allowing to Die, Killing for Mercy, and Suicide that there are four major cultural and social forces opposed to the right to die. The first force is the view among Christians that suicide is <*>. The second force states that self-destruction is not only immoral, unnatural, and sinful, but illegal. A third social force pushing against the right to die is the spectacular success of modern technology. The advancing of technology has led to many breakthroughs in medicine. Illnesses that were once thought fatal are being cured. The fourth societal force is the medical profession.

May's four forces give us much insight into modern societies outlook on euthanasia which is an issue that has been debated for a long time. These forces have led to no concrete laws thus far. The main type of euthanasia being debated at the moment is physician assisted suicide. Physicians are helping patients end their own life by using their medical knowledge. The most publicized case in the United States is the case of Jack Kevorkian. Dr. Kevorkian developed a device in October 1989 that would end one's life quickly and painlessly. He assisted a 54 year old Alzheimer's disease patient in committing suicide in June 1990. In December, Kevorkian

was charged with first-degree murder, but his charge was dismissed due to the fact that Michigan has no law against assisted suicide. Even though his charge was dropped, he was ordered not to help anyone else commit suicide or to give advice about it. However, in February 1991, he violated the court order by giving advice about the preparation of the drug he invented to a terminally ill cancer patient. He still did not obey the court order and still instructed two Michigan women how to commit suicide. Finally, in October 1991, additional murder charges were lodged against him, but in July 1992 the charges were dismissed once again by Judge David Breck. David Breck stated that <*>. He also expressed his belief that physician-assisted suicide remains an alternative for patients experiencing Re unmanageable pain. This view has struck the heart of the medical community.

All physicians are bound by the Hippocratic Oath. Each physician must pledge to adhere to the Hippocratic oath while performing their duties as a physician to society. While taking the Hippocratic Oath, every physician must state this: <*>. These words of the oath are altered by many people when contradicting the right to death by choice. Physicians use this statement from the oath as an argument against death by choice.

An advocate for euthanasia named Joseph Fletcher made an effort to find contradictions in the oath. He said that the oath promised two things: <*>. Doctors do not have the authority morally or ethically to make the decision to allow a patient to die. One's expertise as a physician does not give one the ability to decide the fate of the patient. A doctor might be asked to administer an injection which would bring about an easy and painless death for a patient who is terminally ill. In this case, he must be guided by his conscience. It is not the right of a doctor to decide to initiate or suggest euthanasia as an option, but it is one's option as a physician to cooperate with a patient's decision.

Appendix 2 3 sample argumentative essays by Non-native Turkish advanced University students

ARGUMENTATIVE ESSAY
“CHEATING IN SCHOOL”

THE OUTCOME OF MANY MISTAKES: COPYING

Why some children choose studying while others choose copying? For a long time, the statistics show. That the reason for copying is something related only a child's own psychology, but new investments indicate copying is an outcome of many other reasons about school systems that give damage to a child psychology.

When a child comes to school from its own small world, it comes across with both physical, cultural realities of the real world and is to struggle against all of them. From now then, the child has some rules to obey and a school programme to learn. Nevertheless, the world outside is so far away from the one in the school which is systematic and organized. In such a complex contrary, the child is face to face many strict problems and the scientists suggest that “the teachers must teach that the school is a place where the pupils can think freely and discover every belief and every opinion”, on the other hand, teachers are far away from this suggestion. So, they have wrong opinions and teaching styles that damage a child's emotions and lead him or her even to copy.

First of all, most teachers argue that education should be applied in strict plans. It is doubtless that planning takes an important part in teaching process, but the teachers are unsuccessful in answering their pupils needs and interests because of that plans. In fact, students needs and interests must be the main function of an education system.

Nextly, education aims at teaching the students how to desalinize themselves on their own, but it does not create a suitable environment to apply this aim. In many schools, many rules have been designed and turned in to prohibition, the reasons of which are unknown. What is more, the education system does not help the students to get the main reasons of these. For example, the rules may be discussed in classrooms and be decided with the participation of students. Whereas in schools, pupils that memorize and obey the teacher's rules are accepted as hardworking, while creative, questioning pupils are blamed for their disorderness in respect and duty.

Furthermore, in schools academic success is thought as more important than the others talents. Some artistic subjects such as physical education, art and music are ignored since they are thought to lessen the academic programme. Bloom indicates " No academic subject is superior to another."

In addition to these, educating in schools must be more active and noisier. Generally, it is the teacher who speaks and who is the most active in a class. The opposite scene, the pupils involving and speaking, making noise is considered as a sign of indiscipline.

Likely, the children are motivated to compete rather than cooperation. In this competition while one wins, others lost. This leads envious and hateful emotions. So the education becomes winning rather than learning. Of course the life has some competitive ways, but life throughout is not a competition. For these reasons education must not be forgotten that competing kills the emotions of cooperation, love and friendship.

According to researches, many of the programmes and classroom activities are designed for adults logic not for a child's logic. This makes learning more difficult for such small learners.

Consequently, these faults and many others lead mental disordernesses an damage pupils characters and cause to behave wrongly such as copying. According to patterson "Instead of learning, passing class has become the purpose. Sometimes, the pupils create a study method with

the help of which s/he can pass the class without learning and the others choose copying." This a moral decline in a child character. In fact, this trick and hypocritical behaviour is not found in one who has strong respect, honest and empathy. As discussed, the chance of giving these good features are in the hands of teachers and the only thing to be considered is a more efficient teaching-learning process. Faults, here copying, are not only on outcome of something, the total function of the system is to be considered.

ANIMAL TESTING ALTERNATIVES

Animal testing is a methodology and biochemical assay to ensure products safety and effectiveness there by protecting consumers health. There is no such a personal care which is required by law to be tested on animals On the other hand there are laws regulations and policies governing animal testing such as the animal Welfare act and the public health service policy on humane care and use of laboratory animals Moreover many co operations and companies discuss that there are many alternative methods instead of animal testing.

In 1959 "the principles of humane experimental technique" was published in London defining the concept of animal testing alternatives as the there R's Refinement reduction and Replacement .The only viable choice for a true animal rights supporter is the replacement of animals used in test the other two are morally wrong.

In 1998 visionary group launches protect by the center for alternatives to animal testing (CAAT) which affiliated with the division of toxicological sciences in the department of environ mental health sciences of the Johns Hopkins university created an extraordinary gathering of business people. In this project Henry Spira animal rights pioneer Katherine Stitzel of Procter Gamble Karin Hakanson of Levi Strauss and Nelson Garnett of the National Institutes of Health Hans Ahr of Bayer AG Vera Baumans from Utrecht

university in the Netherland and Uyra Barker of Mary Kay Cosmetics take place.

This remarkable gathering was doing a remarkable thing imagining a new world a world without animal testing. At the end of the meeting the group agreed on that the three Rs will be taken in reality and it will be created on three primary areas.

- 1- Science and technology
- 2- Communication
- 3- Animal Welfare

In Science and technology the group proposed developing methods for interpreting data from new technologies organizing a series of forums for industry managers scientists to learn about new technology and how to use it. In communication area the group recommended on establishing databases on the Alternatives to animal testing Web site.

In animal Welfare, the group planned to begin to effect changes by supporting the development of better technology for measuring the parameters of pain, distress, illness, etc. and promoting a philosophy of "zero-based animal use" by educating scientists.

So, many large companies and co-operations had an agreement on animal testing alternatives. Before that agreement, some single companies have been creating a fine world without animal testing. For example Revlon Cosmetics, US Food and Drug Administration (FDA) and the Soap and Detergent Association followed and still has been following the imagine of a world without animal testing.

As an example US FDA announced in the Federal Register of February, 19, 1998, the Limulus Amebocyte Lysate (LAL) test is an end product endotoxin test for human injectable drugs, animal injectable drugs and medical devices. Moreover, guidelines inform that LAL can be used an alternative to the rabbit pyrogen test.

At present many other alternatives created to reduce or replace animal

testing, but it is still in the early stages of development. The specific tests are:

-Eytex : produced by the National Testing Corp. In Palm Springs California. It is an in-vitro (test-tube) procedure that measures eye irritancy via a protein alteration system. This is used by Avon instead of cruel Draize eye irritancy test.

-Skintex: an in-vitro method to assess skin irritancy that uses pumpkin rinol to mimic the reaction of foreign substance on human skin.

In this part, there must be told that Eytex and Skintex can measure 5000 different materials.

-Epipack: produced by Clonetics in San Diego, California; uses cloned human tissue to test potentially harmful substance.

-Neutral Red Bioassay: developed at Rockefeller University and promoted by Clonetics, cultured human cells that are used to compute the absorption of a water soluble dye to measure relative toxicity.

-Testskin: produced by Oranogenesis in Cambridge. Test skin uses human skin grown sterile plastic bag and uses for measuring irritancy (this method is used by Avon, Amway and Estee Lauder.)

-TOPKAT: is a computer software program that measures toxicity, mutagenicity and teratogenicity. It is produced by Health Design in New York and used by the US. Army, the Environmental Protection Agency and the Food and Drug Administration.

-Agarose Diffusion Method; Tests for toxicity of plastic and synthetic devices used in medical devices such as heart valve artificial joints and intravenass lines.

Progress toward the widespread use of alternatives to animal testing will continue to gain strength as awareness of and support for alternatives is made known. As consumers, we can make a difference in the lives of innocent animals by purchasing only products deemed not cruelty-free (not animal testing) and writing to the companies that still do animal testing and letting them know why you not purchase their products.

Mohandas K.Gandhi commanded on the rights on animals and animal

testing in his autobiography, "The story of My Experiments". "To remind the life of the lamb is no less precious than that of a human being. I should be unwilling to take the life of the lamb for the sake of human body."

VALUES AND CONSEQUENCES OF SCHOOL INTERACTION

School interaction plays a vital role in our lives. Today we have all accepted that education for all is a condition for a well-functioning society, new values and also the growth of society takes places through education.

Firstly, education has a social function. Entry to preschool or school can be daunting experience for young children with their own language background. Because, the setting is quiet different from the home environment that the children have come from and people around them look different, speak in a different language. They need to feel confident in themselves and their families in order to get used to the school environment. It is essential that children have emotional security if they are to grow up as confident healthy people who can take responsibility for themselves and others. Young children with low self-esteem may experience later difficulties with learning.

The school is one of those places where children and young people learn most about the ground rules in our society. They meet children from different social classes with different ethnic backgrounds and with different family backgrounds. At school that young people establish a broader and more balanced interaction with their surroundings. Teachers have an important role to play in ensuring that children develop a positive self-esteem.

The development of a positive self-esteem is supported when children are acknowledgement and acceptance of race, ethnicity, religion and language, socioeconomic level and ability.

"students placed in low ability classroom groups or tracks, where they know they are perceived as low achievers, are not challenged to do their best. Since higher order thinking skills are developed in high ability groups and basic skills in lower ability groups, this system of sorting and labeling students is slowly contributing to a class-based society that could eventually become as rigid as any in the world. (Benham-Tye, 1984)"

"Heterogeneous grouping does not mean that teachers should teach to the slowest student in the group (Benham-Tye, 1984). If done correctly, heterogeneous grouping has the advantage of being more truly democratic. It brings together and provides a common learning experience to students with different backgrounds, interests, cultures and plans for the future. Benham-Tye (1984) notes that the content, teaching methods, classroom climate and teacher-student interaction of heterogeneous classrooms resemble average and upper track classes. Cooperating learning strategies which utilize small heterogeneous groups for instruction and learning have been found to result in high achievement for students at all previous "tracking" levels (Slavin, 1986; Kogan, 1989)."

"some major differences between field dependent and field independent learners are presented. (Howard, 1987)

Educators who want to reach relational, field sensitive youth will succeed by utilizing activities that facilitate social interaction and promote the use of higher order thinking skills. Research suggests that involving the class in lively group discussion, group projects and the telling of stories and personal experiences is more effective than passive, non-social drill and practice activities. Learning should begin with the larger picture that is directly related to the life experiences of the learner. Giving students a sense of a particular activity and how that activity relates to something in their life experiences—present and future—can be a strong motivating factor for relational and field sensitive learners. Educators will find that personal compliments, praise, enthusiasm and even hugs will work wonders in promoting with the self-image and the achievement of this youth."

In addition to that children can learn respect when they are respected by others, teachers, parents and colleagues. A good indication of a positive philosophy of respect for others should be demonstrated by the teachers themselves in their interactions with other staff and children. Teachers need to demonstrate equal respect for everybody. The behaviour shown towards others can be disrespectful.

Secondly, we must accept that school interaction as an instructor and a guide for the future. The natural impulses of children and young people do not always correspond with the habits of the surrounding society. Education makes the students find out what they are going to do in their future life and in which situations they will feel the need to use their knowledge. Furthermore, education functions provide better opportunities in order to find a job.

Every new generation reinterprets the values that exist in our society. At the same time every older generation warns the young against the consequences of changing those values. The education system must ensure that people become aware of the possible results of new values.

Finally, educational institutions have always played a central role as conveyers of value concepts and the growth of society.